PROFESSOR: Good afternoon, everybody. Welcome to Lecture 8. So we're now more than halfway through the lectures.

All right, the topic of today is sampling. I want to start by reminding you about this whole business of inferential statistics. We make references about populations by examining one or more random samples drawn from that population.

We used Monte Carlo simulation over the last two lectures. And the key idea there, as we saw in trying to find the value of pi, was that we can generate lots of random samples, and then use them to compute confidence intervals. And then we use the empirical rule to say, all right, we really have good reason to believe that 95% of the time we run this simulation, our answer will be between here and here.

Well, that's all well and good when we're doing simulations. But what happens when you to actually sample something real? For example, you run an experiment, and you get some data points. And it's too hard to do it over and over again.

Think about political polls. Here was an interesting poll. How were these created? Not by simulation. They didn't run 1,000 polls and then compute the confidence interval. They ran one poll-- of 835 people, in this case. And yet they claim to have a confidence interval. That's what that margin of error is. Obviously they needed that large confidence interval.

So how is this done? Backing up for a minute, let's talk about how sampling is done when you are not running a simulation. You want to do what's called probability sampling, in which each member of the population has a non-zero probability of being included in a sample.

There are, roughly speaking, two kinds. We'll spend, really, all of our time on something called simple random sampling. And the key idea here is that each member of the population has an equal probability of being chosen in the sample so there's no bias.

Now, that's not always appropriate. I do want to take a minute to talk about why. So suppose we wanted to survey MIT students to find out what fraction of them are nerds-- which, by the way, I consider a compliment. So suppose we wanted to consider a random sample of 100 students. We could walk around campus and choose 100 people at random. And if 12% of them were nerds, we would say 12% of the MIT undergraduates are nerds-- if 98%, et cetera.

Well, the problem with that is, let's look at the majors by school. This is actually the majors at MIT by school. And you can see that they're not exactly evenly distributed. And so if you went around and just sampled 100 students at random, there'd be a reasonably high probability that they would all be from engineering and science. And that might give you a misleading notion of the fraction of MIT students that were nerds, or it might not.

In such situations we do something called stratified sampling, where we partition the population into subgroups, and then take a simple random sample from each subgroup. And we do that proportional to the size of the subgroups. So we would certainly want to take more students from engineering than from architecture. But we probably want to make sure we got somebody from architecture in our sample.

This, by the way, is the way most political polls are done. They're stratified. They say, we want to get so many rural people, so many city people, so many minorities-- things like that. And in fact, that's probably where the election recent polls were all messed up. They did a very, retrospectively at least, a bad job of stratifying.

So we use stratified sampling when there are small groups, subgroups, that we want to make sure are represented. And we want to represent them proportional to their size in the population. This can also be used to reduce the needed size of the sample. If we wanted to make sure we got some architecture students in our sample, we'd need to get more than 100 people to start with. But if we stratify, we can take fewer samples. It works well when you do it properly. But it can be tricky to do it properly. And we are going to stick to simple random samples here.

All right, let's look at an example. So this is a map of temperatures in the United States. And so our running example today will be sampling to get information about the average temperatures. And of course, as you can see, they're highly variable. And we live in one of the cooler areas.

The data we're going to use is real data-- and it's in the zip file that I put up for the class-- from the US Centers for Environmental Information. And it's got the daily high and low temperatures for 21 different American cities, every day from 1961 through 2015. So it's an interesting data set-- a total of about 422,000 examples in the dataset. So a fairly good sized dataset. It's fun to play with.

All right, so we're sort of in the part of the course where the next series of lectures, including today, is going to be about data science, how to analyze data. I always like to start by actually looking at the data-- not looking at all 421,000 samples, but giving a plot to sort of give me a sense of what the data looks like. I'm not going to walk you through the code that does this plot. I do want to point out that there are two things in it that we may not have seen before.

Simply enough, I'm going to use numpy.std to get standard deviations instead of my own code for it. And random.sample to take simple random samples from the population. random.sample takes two arguments. The first is some sort of a sequence of values. And the second is an integer telling you how many samples you want. And it returns a list containing sample size, randomly chosen distinct elements.

Distinct elements is important, because there are two ways that people do sampling. You can do sampling without replacement, which is what's done here. You take a sample, and then it's out of the population. So you won't draw it the next time. Or you can do sampling with replacement, which allows you to draw the same sample multiple times-- the same example multiple times.

We'll see later in the term that there are good reasons that we sometimes prefer sampling with replacement. But usually we're doing sampling without replacement. And that's what we'll do here. So we won't get Boston on April 3rd multiple times-- or, not the same year, at least.

All right. So here's the histogram the code produces. You can run it yourself now, if you want, or you can run it later. And here's what it looks like. The daily high temperatures, the mean is 16.3 degrees Celsius. I sort of vaguely know what that feels like. And as you can see, it's kind of an interesting distribution. It's not normal. But it's not that far, right? We have a little tail of these cold temperatures on the left. And it is what it is. It's not a normal distribution. And we'll later see that doesn't really matter.

OK, so this gives me a sense. The next thing I'll get is some statistics. So we know the mean is 16.3 and the standard deviation is approximately 9.4 degrees. So if you look at it, you can believe that.

Well, here's a histogram of one random sample of size 100. Looks pretty different, as you might expect. Its standard deviation is 10.4, its mean 17.7. So even though the figures look a little different, in fact, the means and standard deviations are pretty similar. If we look at the population mean and the sample mean-- and I'll try and be careful to use those terms-- they're not the same. But they're in the same ballpark. And the same is true of the two standard deviations.

Well, that raises the question, did we get lucky or is something we should expect? If we draw 100 random examples, should we expect them to correspond to the population as a whole? And the answer is sometimes yeah and sometimes no. And that's one of the issues I want to explore today.

So one way to see whether it's a happy accident is to try it 1,000 times. We can draw 1,000 samples of size 100 and plot the results. Again, I'm not going to go over the code. There's something in that code, as well, that we haven't seen before. And that's the ax.vline plotting command. V for vertical. It just, in this case, will draw a red line-- because I've said the color is r-- at population mean on the x-axis. So just a vertical line. So that'll just show us where the mean is. If we wanted to draw a horizontal line, we'd use ax.hline. Just showing you a couple of useful functions.

When we try it 1,000 times, here's what it looks like. So here we see what we had originally, same picture I showed you before. And here's what we get when we look at the means of 100 samples. So this plot on the left looks a lot more like it's a normal distribution than the one on the right. Should that surprise us, or is there a reason we should have expected that to happen?

Well, what's the answer? Someone tell me why we should have expected it. It's because of the central limit theorem, right? That's exactly what the central limit theorem promised us would happen. And, sure enough, it's pretty close to normal. So that's a good thing.

And now if we look at it, we can see that the mean of the sample means is 16.3, and the standard deviation of the sample means is 0.94. So if we go back to what we saw here, we see that, actually, when we run it 1,000 times and look at the means, we get very close to what we had initially. So, indeed, it's not a happy accident. It's something we can in general expect.

All right, what's the 95% confidence interval here? Well, it's going to be 16.28 plus or minus

1.96 times 0.94, the standard deviation of the sample means. And so it tells us that the confidence interval is, the mean high temperature, is somewhere between 14.5 and 18.1.

Well, that's actually a pretty big range, right? It's sort of enough to where you wear a sweater or where you don't wear a sweater. So the good news is it includes the population mean. That's nice. But the bad news is it's pretty wide.

Suppose we wanted it tighter bound. I said, all right, sure enough, the central limit theorem is going to tell me the mean of the means is going to give me a good estimate of the actual population mean. But I want it tighter bound. What can I do?

Well, let's think about a couple of things we could try. Well, one thing we could think about is drawing more samples. Suppose instead of 1,000 samples, I'd taken 2,000 or 3,000 samples. We can ask the question, would that have given me a smaller standard deviation? For those of you who have not looked ahead, what do you think? Who thinks it will give you a smaller standard deviation? Who thinks it won't? And the rest of you have either looked ahead or refused to think. I prefer to believe you looked ahead.

Well, we can run the experiment. You can go to the code. And you'll see that there is a constant of 1,000, which you can easily change to 2,000. And lo and behold, the standard deviation barely budges. It got a little bit bigger, as it happens, but that's kind of an accident. It just, more or less, doesn't change. And it won't change if I go to 3,000 or 4,000 or 5,000. It'll wiggle around. But it won't help much. What we can see is doing that more often is not going to help.

Suppose we take larger samples? Is that going to help? Who thinks that will help? And who thinks it won't? OK. Well, we can again run the experiment. I did run the experiment. I changed the sample size from 100 to 200. And, again, you can run this if you want. And if you run it, you'll get a result-- maybe not exactly this, but something very similar-- that, indeed, as I increase the size of the sample rather than the number of the samples, the standard deviation drops fairly dramatically, in this case from 0.94 0.66. So that's a good thing.

I now want to digress a little bit before we come back to this and look at how you can visualize this-- Because this is a technique you'll want to use as you write papers and things like that-- is how do we visualize the variability of the data? And it's usually done with something called an error bar. You've all seen these things here. And this is one I took from the literature. This is plotting pulse rate against how much exercise you do or how frequently you exercise.

And what you can see here is there's definitely a downward trend suggesting that the more you exercise, the lower your average resting pulse. That's probably worth knowing. And these error bars give us the 95% confidence intervals for different subpopulations.

And what we can see here is that some of them overlap. So, yes, once a fortnight-- two weeks for those of you who don't speak British-- it does get a little bit smaller than rarely or never. But the confidence interval is very big. And so maybe we really shouldn't feel very comfortable that it would actually help.

The thing we can say is that if the confidence intervals don't overlap, we can conclude that the means are actually statistically significantly different, in this case at the 95% level. So here we see that the more than weekly does not overlap with the rarely or never. And from that, we can conclude that this is actually, statistically true-- that if you exercise more than weekly, your pulse is likely to be lower than if you don't.

If confidence intervals do overlap, you cannot conclude that there is no statistically significant difference. There might be, and you can use other tests to find out whether there are. When they don't overlap, it's a good thing. We can conclude something strong. When they do overlap, we need to investigate further.

All right, let's look at the error bars for our temperatures. And again, we can plot those using something called pylab.errorbar. Lab So what it takes is two values, the usual x-axis and y-axis, and then it takes another list of the same length, or sequence of the same length, which is the y errors. And here I'm just going to say 1.96 times the standard deviations. Where these variables come from you can tell by looking at the code. And then I can say the format, I want an o to show the mean, and then a label. Fmt stands for format.

errorbar has different keyword arguments than plot. You'll find that you look at different ways like histograms and bar plots, scatterplots-- they all have different available keyword arguments. So you have to look up each individually. But other than this, everything in the code should look very familiar to you.

And when I run the code, I get this. And so what I've plotted here is the mean against the sample size with errorbars. And 100 trials, in this case. So what you can see is that, as the sample size gets bigger, the errorbars get smaller. The estimates of the mean don't necessarily get any better.

In fact, we can look here, and this is actually a worse estimate, relative to the true mean, than the previous two estimates. But we can have more confidence in it. The same thing we saw on Monday when we looked at estimating pi, dropping more needles didn't necessarily give us a more accurate estimate. But it gave us more confidence in our estimate. And the same thing is happening here. And we can see that, steadily, we can get more and more confidence.

So larger samples seem to be better. That's a good thing. Going from a sample size of 50 to a sample size of 600 reduced the confidence interval, as you can see, from a fairly large confidence interval here, ran from just below 14 to almost 19, as opposed to 15 and a half or so to 17. I said confidence interval here. I should not have. I should have said standard deviations. That's an error on the slides.

OK, what's the catch? Well, we're now looking at 100 samples, each of size 600. So we've looked at a total of 600,000 examples. What has this bought us? Absolutely nothing. The entire population only contained about 422,000 samples. We might as well have looked at the whole thing, rather than take a few of them. So it's like, you might as well hold an election rather than ask 800 people a million times who they're going to vote for. Sure, it's good. But it gave us nothing.

Suppose we did it only once. Suppose we took only one sample, as we see in political polls. What can we can conclude from that? And the answer is actually kind of surprising, how much we can conclude, in a real mathematical sense, from one sample. And, again, this is thanks to our old friend, the central limit theorem.

So if you recall the theorem, it had three parts. Up till now, we've exploited the first two. We've used the fact that the means will be normally distributed so that we could use the empirical rule to get confidence intervals, and the fact that the mean of the sample means would be close to the mean of the population.

Now I want to use the third piece of it, which is that the variance of the sample means will be close to the variance of the population divided by the sample size. And we're going to use that to compute something called the standard error-- formerly the standard error of the mean. People often just call it the standard error. And I will be, alas, inconsistent. I sometimes call it one, sometimes the other.

It's an incredibly simple formula. It says the standard error is going to be equal to sigma, where sigma is the population standard deviation divided by the square root of n, which is going to be the size of the sample. And then there's just this very small function that implements it. So we can compute this thing called the standard error of the mean in a very straightforward way.

We can compute it. But does it work? What do I mean by work? I mean, what's the relationship of the standard error to the standard deviation? Because, remember, that was our goal, was to understand the standard deviation so we could use the empirical rule.

Well, let's test the standard error of the mean. So here's a slightly longer piece of code. I'm going to look at a bunch of different sample sizes, from 25 to 600, 50 trials each. So getHighs is just a function that returns the temperatures. I'm going to get the standard deviation of the whole population, then the standard error of the means and the sample standard deviations, both. And then I'm just going to go through and run it. So for size and sample size, I'm going to append the standard error of the mean. And remember, that uses the population standard deviation and the size of the sample. So I'll compute all the SEMs. And then I'm going to compute all the actual standard deviations, as well. And then we'll produce a bunch of plots-- or a plot, actually.

All right, so let's see what that plot looks like. Pretty striking. So we see the blue solid line is the standard deviation of the 50 means. And the red dotted line is the standard error of the mean. So we can see, quite strikingly here, that they really track each other very well. And this is saying that I can anticipate what the standard deviation would be by computing the standard error.

Which is really useful, because now I have one sample. I computed standard error. And I get something very similar to what I get of the standard deviation if I took 50 samples and looked at the standard deviation of those 50 samples. All right, so not obvious that this would be true, right? That I could use this simple formula, and the two things would track each other so well. And it's not a coincidence, by the way, that as I get out here near the end, they're really lying on top of each other. As the sample size gets much larger, they really will coincide.

So one, does everyone understand the difference between the standard deviation and the standard error? No. OK. So how do we compute a standard deviation? To do that, we have to look at many samples-- in this case 50-- and we compute how much variation there is in those 50 samples.

For the standard error, we look at one sample, and we compute this thing called the standard error. And we argue that we get the same number, more or less, that we would have gotten had we taken 50 samples or 100 samples and computed the standard deviation. So I can avoid taking all 50 samples if my only reason for doing it was to get the standard deviation. I can take one sample instead and use the standard error of the mean. So going back to my temperature-- instead of having to look at lots of samples, I only have to look at one. And I can get a confidence interval. That make sense? OK.

There's a catch. Notice that the formula for the standard error includes the standard deviation of the population-- the standard deviation of the sample. Well, that's kind of a bummer. Because how can I get the standard deviation of the population without looking at the whole population? And if we're going to look at the whole population, then what's the point of sampling in the first place?

So we have a catch, that we've got something that's a really good approximation, but it uses a value we don't know. So what should we do about that? Well, what would be, really, the only obvious thing to try? What's our best guess at the standard deviation of the population if we have only one sample to look at? What would you use? Somebody? I know I forgot to bring the candy today, so no one wants to answer any questions.

AUDIENCE: The standard deviation of the sample?

PROFESSOR: The standard deviation of the sample. It's all I got. So let's ask the question, how good is that? Shockingly good. So I looked at our example here for the temperatures. And I'm plotting the sample standard deviation versus the population standard deviation for different sample sizes, ranging from 0 to 600 by one, I think.

So what you can see here is when the sample size is small, I'm pretty far off. I'm off by 14% here. And I think that's 25. But when the sample sizes is larger, say 600, I'm off by about 2%. So what we see, at least for this data set of temperatures-- if the sample size is large enough, the sample standard deviation is a pretty good approximation of the population standard deviation.

Well. Now we should ask the question, what good is this? Well, as I said, once the sample reaches a reasonable size-- and we see here, reasonable is probably somewhere around 500-- it becomes a good approximation. But is it true only for this example? The fact that it happened to work for high temperatures in the US doesn't mean that it will always be true.

So there are at least two things we should consider to asking the question, when will this be true, when won't it be true. One is, does the distribution of the population matter? So here we saw, in our very first plot, the distribution of the high temperatures. And it was kind of symmetric around a point-- not perfectly. But not everything looks that way, right?

So we should say, well, suppose we have a different distribution. Would that change this conclusion? And the other thing we should ask is, well, suppose we had a different sized population. Suppose instead of 400,000 temperatures I had 20 million temperatures. Would I need more than 600 samples for the two things to be about the same?

Well, let's explore both of those questions. First, let's look at the distributions. And we'll look at three common distributions-- a uniform distribution, a normal distribution, and an exponential distribution. And we'll look at each of them for, what is this, 100,000 points.

So we know we can generate a uniform distribution by calling random.random. Gives me a uniform distribution of real numbers between 0 and 1. We know that we can generate our normal distribution by calling random.gauss. In this case, I'm looking at it between the mean of 0 and a standard deviation of 1. But as we saw in the last lecture, the shape will be the same, independent of these values.

And, finally, an exponential distribution, which we get by calling random.expovariate. Very And this number, 0.5, is something called lambda, which has to do with how quickly the exponential either decays or goes up, depending upon which direction. And I'm not going to give you the formula for it at the moment. But we'll look at the pictures. And we'll plot each of these discrete approximations to these distributions.

So here's what they look like. Quite different, right? We've looked at uniform and we've looked at Gaussian before. And here we see an exponential, which basically decays and will asymptote towards zero, never quite getting there. But as you can see, it is certainly not very symmetric around the mean.

All right, so let's see what happens. If we run the experiment on these three distributions, each of 100,000 point examples, and look at different sample sizes, we actually see that the difference between the standard deviation and the sample standard deviation of the population standard deviation is not the same.

We see, down here-- this looks kind of like what we saw before. But the exponential one is really quite different. You know, its worst case is up here at 25. The normal is about 14. So that's not too surprising, since our temperatures were kind of normally distributed when we looked at it. And the uniform is, initially, much better an approximation.

And the reason for this has to do with a fundamental difference in these distributions, something called skew. Skew is a measure of the asymmetry of a probability distribution. And what we can see here is that skew actually matters. The more skew you have, the more samples you're going to need to get a good approximation. So if the population is very skewed, very asymmetric in the distribution, you need a lot of samples to figure out what's going on. If it's very uniform, as in, for example, the uniform population, you need many fewer samples. OK, so that's an important thing. When we go about deciding how many samples we need, we need to have some estimate of the skew in our population.

All right, how about size? Does size matter? Shockingly-- at least it was to me the first time I looked at this-- the answer is no. If we look at this-- and I'm looking just for the uniform distribution, but we'll see the same thing for all three-- it more or less doesn't matter. Quite amazing, right?

If you have a bigger population, you don't need more samples. And it's really almost counterintuitive to think that you don't need any more samples to find out what's going to happen if you have a million people or 100 million people. And that's why, when we look at, say, political polls, they're amazingly small. They poll 1,000 people and claim they're representative of Massachusetts.

This is good news. So to estimate the mean of a population, given a single sample, we choose a sample size based upon some estimate of skew in the population. This is important, because if we get that wrong, we might choose a sample size that is too small. And in some sense, you always want to choose the smallest sample size you can that will give you an accurate answer, because it's more economical to have small samples than big samples.

And I've been talking about polls, but the same is true in an experiment. How many pieces of data do you need to collect when you run an experiment in a lab. And how much will depend, again, on the skew of the data. And that will help you decide.

When you know the size, you choose a random sample from the population. Then you compute the mean and the standard deviation of that sample. And then use the standard deviation of that sample to estimate the standard error. And I want to emphasize that what you're getting here is an estimate of the standard error, not the standard error itself, which would require you to know the population standard deviation. But if you've chosen the sample size to be appropriate, this will turn out to be a good estimate.

And then once we've done that, we use the estimated standard error to generate confidence intervals around the sample mean. And we're done. Now this works great when we choose independent random samples. And, as we've seen before, that if you don't choose independent samples, it doesn't work so well. And, again, this is an issue where if you assume that, in an election, each state is independent of every other state, and you'll get the wrong answer, because they're not.

All right, let's go back to our temperature example and pose a simple question. Are 200 samples enough? I don't know why I chose 200. I did. So we'll do an experiment here. This is similar to an experiment we saw on Monday.

So I'm starting with the number of mistakes I make. For t in a range number of trials, sample will be random.sample of the temperatures in the sample size. This is a key step. The first time I did this, I messed it up. And instead of doing this very simple thing, I did a more complicated thing of just choosing some point in my list of temperatures and taking the next 200 temperatures. Why did that give me the wrong answer? Because it's organized by city. So if I happen to choose the first day of Phoenix, all 200 temperatures were Phoenix-- which is not a very good approximation of the temperature in the country as a whole.

But this will work. I'm using random.sample. I'll then get the sample mean. Then I'll compute my estimate of the standard error by taking that as seen here. And then if the absolute value of the population minus the sample mean is more than 1.96 standard errors, I'm going to say I messed up. It's outside. And then at the end, I'm going to look at the fraction outside the 95% confidence intervals.

And what do I hope it should print? What would be the perfect answer when I run this? What fraction should lie outside that? It's a pretty simple calculation. Five, right? Because if they all were inside, then I'm being too conservative in my interval, right? I want 5% of the tests to fall outside the 95% confidence interval.

If I wanted fewer, then I would look at three standard deviations. Instead of 1.96, then I would expect less than 1% to fall outside. So this is something we have to always keep in mind when we do this kind of thing. If your answer is too good, you've messed up. Shouldn't be too bad, but it shouldn't be too good, either. That's what probabilities are all about. If you called every election correctly, then your math is wrong.

Well, when we run this, we get this lovely answer, that the fraction outside the 95% confidence interval is 0.0511. That's exactly-- well, close to what you want. It's almost exactly 5%. And if I run it multiple times, I get slightly different numbers. But they're all in that range, showing that, here, in fact, it really does work.

So that's what I want to say, and it's really important, this notion of the standard error. When I talk to other departments about what we should cover in 60002, about the only thing everybody agrees on was we should talk about standard error. So now I hope I have made everyone happy. And we will talk about fitting curves to experimental data starting next week. All right, thanks a lot.