PROFESSOR: Good morning. This is the second of two lectures that I'm retaping in the summer because we had technical difficulties with the lectures that were taped during the academic term. I feel I need to tell you this for two reasons. One, as I said before, the room is empty. And so when I say something hilarious and there's no laughter, it's because the room is empty. And if I'm not asking the students questions during the lecture, it's because there are no students.

The other important thing I want you to understand is that I do own more than one shirt and more than one pair of pants. And the reason I'm dressed exactly the way I was for lecture 13 is I gave lecture 13 five minutes ago, even though this is lecture 15. So again, here I am. And I apologize for the uniformity in my clothing.

OK, on lecture 14, which came between 13 and 15, at the end of it I was talking about flipping coins and ended up with the question how can we know when it's safe to assume that the average result of some finite number of flips is representative of what we would get if we flipped the same coin many more times. In principle, an infinite number of times.

Well, we might flip a coin twice, get 1 heads and 1 tails and conclude that the true probability of getting a head is exactly 0.5. Turns out-- assume I have a fair coin-- this would have been the right conclusion.

Just because we have the right answer it doesn't mean our thinking is any good. And in fact, in this case our reasoning would have been completely faulty because if I flipped it twice and had gotten 2 heads, you might have said oh, it's always heads. But we know that wouldn't have been right.

So the question I want to pose at the start of today's lecture is quite simply how

many samples do we need to believe the answer? So how many samples do we need to look at before we can have confidence in the result.

Fortunately, there's a very solid set of mathematics that lets us answer this question in a good way. At the root of all of it is the notion of variance. Variance is a measure of how much spread there is in the possible outcomes.

Now in order to talk about variance, given this definition, we need to have different outcomes, which is why we always want to run multiple trials rather than say one trial with many flips. In fact, you may have wondered why am I not-- if I could end up flipping the coin a million times, why would I do multiple trials adding up to a million rather than 1 trial of a million?

And the reason is by having multiple trials, each of which give me a different outcome, I can then look at how different the outcomes of the different trials are and get a measure of variance. If I do 10 trials and I get the same answer each time, I can begin to believe that really is the correct answer. If I do 10 trials and get 10 wildly different answers, then I probably shouldn't believe any one of those answers, and I probably shouldn't even think I can average those answers and believe the mean is a real answer because if I run an 11th trial maybe I'll get something totally different yet again.

We can formalize this notion of variance in a way that should be familiar to many of you. And that's the concept of a standard deviation. Something I, in fact, already showed you when we looked at the spread of grades on the first quiz this semester.

Informally, what the standard deviation is measuring is the fraction of values that are close to the mean. If many values are close to the mean, the standard deviation is small. If many values are relatively far from the mean, the standard deviation is relatively large.

If all values are the same, then the standard deviation is 0. In the real world that essentially never happens. We can write a formula for this. Fortunately, it's not all about words.

And we can say the standard deviation of x, where x is a set of trials-- sigma is usually used to talk about that-- is equal to the square root of 1 over the absolute value of the length of x. So that's 1 over the number of trials times the summation of the value of each trial, little x and big X, of x minus mu squared, where mu is the mean. And as I said, that's the cardinality of x.

Well, so that's a formula. And those of you are majoring in math are going to love that. But for those of you who are more computationally oriented, I recommend you just take a look at the code. So here's an implementation of the standard deviation.

So the standard deviation of x is equal-- start by getting the mean of x, which is by summing x and dividing it by the length of x. Then I'm just going to sum all the values in x and do the computation. So that code and that formula are the same thing.

All right, now we know what standard deviation means. What are we going to do with it? We're going to use it to look at the relationship between the number of samples we've looked at and how much confidence we should have in the answer.

So we'll do this again looking at a bunch of code. So I've got this function flip plot, which doesn't quite fit in the screen, but that's OK. It's not very interesting in the details.

What it does is it runs multiple trials of some number of coin flips and plots a bunch of values about the relative frequency of heads and tails and also the standard deviation of each. So again nothing very exciting, in the code I'm just going to keep track for all these trials.

The minimum and the maximum exponent. I'm using that so I can run a lot of trials quickly. The mean ratios, the differences, and the standard deviations for exponent in range. Minimum exponent to maximum exponent plus 1. I'm going to build an x-axis.

So this is going to be the number of flips. And then for the number of flips I'm going

to run a bunch of tests and get the ratios of heads to tails and the absolute difference between heads and tails. And then, I'm going to do a bunch of plotting.

And again, what I want you to notice is when I'm doing the plotting, I'm going to label the axes and put some titles on. And I'm also going to use semilog because given that I'm looking at different powers, it would compress everything on the left if I would just use linear.

All right. Let's run it. Actually, let's comment out the code we need to run it. So I'm going to call flip plot with a minimum exponent of 4, a maximum exponent of 20. That's pretty high. And I'm going to run 20 trials.

This could take a little while to run, but not too long. And it'll give us some pretty pictures to look at. Give me a chance to have a drink of water.

I know the suspense is killing you as to what these plots are going to look like. Here they are. All right. So if we look at plot one, that's the ratio of heads to tails. And as you can see, it bounces around in the beginning. When we have a small number of flips, the ratio moves a bit. But as I get to a lot of flips out here, 10 to the 5th, 10 to the 6th, what we're seeing is it begins to stabilize.

We don't get much difference. Kind of interesting where it's stabilizing. Maybe not where we'd expect it. I would have guessed it would stabilize a little closer to one than it did as I got out here. And maybe I have an unfair coin.

That's the problem with running these experiments in real time that I can't necessarily get the answer I want. But for the most part, actually, it looks much better on my screen than it does on your screen. In fact, in my screen, it looks like it's very close to 1.

I don't know. I guess there's some distortion here. Think 1.

And if we look at the standard deviation of the ratio of heads to tails, what we're seeing is that's also dropping from somewhere up here around 10 to the 0 down to 10 to the minus 3. And it's dropping pretty steadily as I increase the number of trials.

That's really what you would hope to see and expect to see that not the number of trials, the number of flips. Sorry. As I flip more coins, the variance between trials should get smaller because in some sense, randomness is playing a less important role.

The more random trials you do, the more likely you are to get something that's actually representative of the truth. And therefore, you would expect the standard deviation to drop.

All right. Now, what we're saying here is because the standard deviation is dropping, not only are we getting something closer to the right answer but perhaps more importantly, we have better reason to believe we're seeing the right answer.

That's very important. That's where I started this lecture. It's not good enough to get lucky and get the correct answer. You have to have evidence that can convince somebody that really is the correct answer. And the evidence here is the small standard deviation.

Let's look at a couple of the other figures. So here's Figure (3). This is the mean of the absolute difference between heads and tails. Not too surprisingly-- we saw this in the last lecture-- as I flip more coins, the mean difference is going to get bigger. That's right. We expect the ratio to get smaller, but we expected the mean difference to get bigger.

On the other hand, let's look at Figure (4). What we're looking here is this difference in the standard deviations. And interestingly, what we're seeing is the more coins I flip, the bigger the standard deviation is.

Well, this is kind of interesting. I look at it, and I sort of said that when the standard deviation is small, we think that the variance is small. And therefore, the results are credible. When the standard deviation is large, we think the variance is large. And therefore, the results are maybe incredible.

Well, I said that a little bit wrong. I tried to say it right the first time. What I have to

ask is not is the standard deviation large or small, but is it relatively large a relatively small?

Relative to what? Relative to the mean. If the mean is a million, and the standard deviation is 20, it's a relatively small standard deviation. If the mean is 10, and the standard deviation is 20, then it's enormous.

So it doesn't make sense. And we saw this. We looked at quizzes. If the mean score on Quiz 1 were 70 and the standard deviation were 5, we'd say OK, it's pretty packed around the mean. If the mean score were 10, which maybe is closer to the truth, and the standard deviation were 5, then we'd say it's not really packed around the mean. So we have to always look at it relative or think about it relative to that.

Now the good news is we have, again, a mathematical formula that lets us do that. Get rid of all those figures for the moment. And that formula is called the coefficient of variation.

For reasons I don't fully understand, this is typically not used. People always talk about the standard deviation. But in many cases, it's the coefficient of variation that really is a more useful measure.

And it's simply the standard deviation divided by the mean. So that let's us talk about the relative variance, if you will. The nice thing about doing this is it lets us relate different datasets with different means and think about how much they vary relative to each other.

So if we think about it-- usually we argue in that if it's less than 1, we think about that as low variance. Now there should be some warnings that come with the coefficient of variation. And these are some of the reasons people don't use it as often because they don't want to bother giving the warning labels.

If the mean is near 0, small changes in the mean are going to lead to large changes in the coefficient of variation. They're not necessarily very meaningful. So when the mean is near 0, the coefficient of variation is something you need to think about with several grains of salt. Makes sense. You're dividing by something near 0, a small

change is going to produce something big.

Perhaps more importantly, or equally importantly-- and this is something we're going to talk about later-- is that unlike the standard deviation, the coefficient of variation cannot be used to construct confidence intervals. I know we haven't talked about confidence intervals yet, but we will shortly.

All right. By now, you've got to be tremendously bored with flipping coins. Nevertheless, I'm going to ask you to look at one more coin flipping simulation. And then, I promise we'll change the topic. And this is to show you some more aspects of the plotting facilities in PyLab.

So I'm going to just flip a bunch of coins, run a simulation. You've seen this a zillion times. And then, we'll make some plots.

And this is really kind of the interesting part. What I want you to notice about this-- let's take a look at here. So now we have been plotting curves. Here we're going to plot a histogram.

So I'm going to give a set of values, a set of y values. In this case the fraction of heads. And a number of bins in which to do the histogram.

So let's look a little example first here independent of this program. Oops. Wrong way. So I'm going to set l, a list, equals 1, 2, 3, 3, 3, 4. And then, I'm just going to plot a histogram with 6 bins. And then show it.

I've done something I'm not supposed to do. I just know title. There's no x-label. No y-label. That's because this is totally meaningless. I just wanted to show you how histograms work. And what you'll see here is that it's shown that I've got three instances of this value, of 3, and one of everything else. And it's just giving me essentially a bar chart, if you will.

Again many, many plotting capabilities you'll see on the website. This is just a simple one. One I like to use and use fairly frequently.

Some other things I want to show you here is I'm using xlim and ylim. So what we could do here is this is setting the limits of the x and y-axis, rather than using defaults saying the lowest value should be this thing, the variable called xmin, which I've computed up here. And the highest ymin.

What you'll see if we go up a little bit-- so I'm getting the fraction of heads1 and computing the mean 1, and the standard deviation 1. Then I'm going to plot a histogram of the way we looked at it.

And then what I'm going to do is say xmin and xmax is pyLab.xlim. If you call xlim with no arguments, what it will return is the minimum x value and the minimum y value of the current plot, the current figure.

So now I stored the minimum x values and the maximum x values to the current one. And I did the same thing for y. And then going to plot the figure here.

Then I'm going to run it again. I'm going to run another simulation, getting fracHeads 2, mean 2, standard deviation 2. Going to plot the histograms. But then I'm going to set, for the new one, the x limit of this to the previous ones that I saved from the previous figure.

Why am I doing that? Because I want to be able to compare the two figures. As we'll see when we have our lecture on how to lie with data, a great way to fool people with figures is to subtly change the range of one of the axes. And then you look at things and wow, that's really different or they're really the same. When in fact, neither conclusion is true. It's just that they've been normalized to either look the same or look different. So it's kind of cheating.

And then we'll do it. So now let's run it and see what we get. I don't need this little silly thing first. Let's see. It's going to take a long time, maybe.

That's one way to fill up a lecture. Just run simulations that take a long time to run. Much easier to prepare than actual material. But nevertheless, shouldn't take forever.

I may have said this before. I have two computers. I have a fast one that sits at my desk that I use to prepare my lectures and a slower one that I use to give the lectures. I should probably be testing all these things out on the slow computer before making you wait. But really, it's going to stop. I promise.

Ah. All right. So we'll look at these. So Figure (1) has got 100,000 trials of 10 flips each. And Figure (2), 100,000 trials of a 1,000 flips each. And let's look at the two figures side by side. Make them a little smaller so we can squeeze them both in.

So what have we got here? Notice if we look at these two plots, the means are about the same. 0.5 and 0.499. Not much difference.

The standard deviations are quite different. And again, you would expect that. A 100 flips should have a lot higher standard deviation than a 1,000 flips. And indeed, it certainly does. 0.15 is a lot smaller than 0.05.

So that tells us something good. It says, as we've discussed, that these results are more credible than these results. Not to say that they're more accurate because they're not really. But they're more believable. And that's what's important.

Notice also the spread of outcomes is much tighter here than it is here. Now, that's why I played with xlim. If I used the default values, it would not have looked much tighter when I put this up on the screen because it would have said well, we don't have any values out here. I don't need to display all of this. And it would have then about the same visual width as this. And therefore, potentially very deceptive when you just stared at it if you didn't look carefully at the units on the x-axis.

So what I did is since I knew I wanted to show you these things side by side and make the point about how tight the distribution is, I made both axes run the same length. And therefore, produce comparable figures.

I also, by the way, used xlim and ylim if you look at the code, which you will have in your handout, to put this text box in a place where it would be easy to see. You can also use the fault of best, which often puts it in the right place. But not always.

The distribution of the results in both cases is close to something called the normal distribution. And as we talk about things like standard deviation or a coefficient of variation, we are talking about not just the average value but the distribution of values in these trials.

The normal distribution, which is very common, has some interesting properties. It always peaks at the mean and falls off symmetrically. The shape of the normal distribution, so I'm told, looks something like this. And there are people who imagine it looks like a bell. And therefore, the normal distribution is often also called the bell curve.

That's a terrible picture. I'm going to get rid of it. And indeed, mathematicians will always call it this. This is often what people use in the non-technical literature. There was, for example, a very controversial book called "The Bell Curve," which I don't recommend reading.

OK. So this is not a perfect normal distribution. It's not really exactly symmetric. We could zoom in on this one and see if it's better. In fact, let me make that larger. And then we'll zoom in on it. Now that we're not comparing the two, we can just zoom in on the part we care about.

And you can see again it's not perfectly symmetric. But it's getting there. And in fact, the trials are not very big. Only a 1,000 flips. If I did 100,000 trials of a 100,000 flips each, we wouldn't finish the lecture. It'd take too long. But we'd get a very pretty looking curve. And in fact, I have done that in the quiet of my office. And it works very nicely.

And so in fact, we would be converging here on the normal distribution. Normal distributions are frequently used in constructing probabilistic models for two reasons.

Reason one, is they have nice mathematical properties. They're easy to reason about for reasons we'll see shortly. That's not good enough. The curve where every value is the same has even nicer mathematical properties but isn't very useful. But

the nice thing about normal distributions is -- many naturally occurring instances.

So let's first look at what makes them nice mathematically and then let's look at where they occur. So the nice thing about them mathematically is they can be completely characterized by two parameters, the mean and the standard deviation.

Knowing these is the equivalent to knowing the entire distribution. Furthermore, if we have a normal distribution, the mean and the standard deviation can be used to compute confidence intervals. So this is a very important concept. One that you see all the time in the popular press but maybe don't know what it actually means when you see it.

So instead of estimating an unknown parameter-- and that's, of course, all we've been doing with this whole probability business. So you get some unknown parameter like the probability of getting a head or a tail, and we've been estimating it using the various techniques. And typically, you've been estimating it by a single value, the mean of a set of trials.

A confidence interval instead allows us to estimate the unknown parameter by providing a range that is likely to contain the unknown value. And a confidence that the unknown value lies within that range. It's called the confidence level.

So for example, when you look at political polls, you might see something that would say the candidate is likely to get 52% of the vote plus or minus 4%. So what does that mean? Well, if somebody doesn't specify the confidence level, they usually mean 5%.

So what this says is that 95% percent of the time-- 95th confidence interval-- if the election were actually conducted, the candidate would get somewhere between 48% and 56% of the vote. So 95% percent of the time, 95% percent of the elections, the candidate would get between 48% and 56% of the votes. So we have two things, the range and our confidence that the value will lie within that range.

When they make those assumptions, when you see something like that in the press, they are assuming that elections are random trials that have a normal distribution.

That's an implicit assumption in the calculation that tells us this.

The nice thing here is that there is something called the empirical rule, which is used for normal distributions. They give us a handy way to estimate confidence intervals given the mean and the standard deviation. If we have a true normal distribution, then roughly speaking, 68% of the data are within the one standard deviation above the mean. And 95% percent within two standard deviations. And almost all, 99.7%, will fall within three.

These values are approximations. They're not exactly right. It's not exactly 68 and 95. But they're good enough for government work. So we can see this here. And this is what people use when they think about these things.

Now this may raise an interesting question in your mind. How do the pollsters go about finding the standard deviation? Do they go out and conduct a 100 separate polls and then do some math? Sort of what we've been doing. You might hope so, but that's not what they do because it's expensive. And nobody wants to do that.

So they use another trick to estimate the standard deviation. Now, you're beginning to understand why these polls aren't always right. And the trick they use for that is something called the standard error, which is an estimate of the standard deviation.

And you can only do this under the assumption that the errors are normally distributed and also that the sample population is small. And I mean small, not large. It's small relative to the actual population.

So this gets us to one of the things we like about the normal distribution that in fact, it's often an accurate model of reality. And when people have done polls over and over again, they do discover that, indeed, the results are typically normally distributed. So this is not a bad assumption. Actually, it's a pretty good assumption.

So if we have p, which is equal to the percentage sample. And we have n, which is equal to the sample size, we can say that the standard error, which I'll write SE, is equal to the formula p times 100-- because we're dealing with percentages-- minus p divided by n to the 1/2, the square root of all of this.

So if for example, a pollster were to sample 1,000 voters, and 46% of them said that they'll vote for Abraham Lincoln-- we should be so lucky that Abraham Lincoln were running for office today-- the standard error would be roughly 1.58%. We would interpret this to mean that in 95% percent of the time, the true percentage of votes that Lincoln would get is within two standard errors of 46%.

I know that's a lot to swallow quickly. So as always, we'll try and make sense of it by looking at some code. By now, you've probably all figured out that I'm much more comfortable with code than I am with formulas.

So we're going to conduct a poll here. Not really, we're going to pretend we're conducting a poll. n and p. We'll start with no votes. And for i in range n, if random.random is less than p over 100, the number of votes will be increased by 1. Otherwise, it will stay where it was and will return the number of votes.

Nothing very dramatic. And then, we'll test the error here. So n equals 1,000, p equals 46, the percentage of votes that we think Abraham Lincoln is going to get. We'll run 1,000 trials. Results equal that. For t in range number of trials results.append, I'll run the poll.

And we'll look at the standard deviation, and we'll look at the results. And we'll print the fraction of votes and the number of polls. All right, let's see what we get when we do this.

Well, pretty darn close to a normal distribution. Kind of what we'd expect. The fraction of votes peaks at 46%. Again what we'd expect. But every once in while, it gets all the way out here to 50 and looks like Abe might actually win an election. Highly unlikely in our modern society. And over here, he would lose a lot of them.

If we look here, we'll see that the standard deviation is 1.6. So it turns out that the standard error, which you'll remember we computed using that formula to be 1.58-- you may not remember it because I said it and didn't write it down-- is pretty darn close to 1.6.

So remember the standard error is an attempt to just use a formula to estimate what the standard deviation is going to be. And in fact, we use this formula, very simple formula, to guess what it would be. We then ran a simulation and actually measured the standard deviation, no longer a guess. And it came out to be 1.6. And I hope that most of you would agree that that was a pretty good guess.

And so therefore because, if you will, the differences are normally distributed, the distribution is normal. It turns out the standard error is a very good approximation to the actual standard deviation. And that's what pollsters rely on and why polls are actually pretty good. So now the next time you read a poll, you'll understand the math behind it. In a subsequent lecture, we'll talk about some ways they go wrong that have nothing to do with getting the math wrong.

Now, of course, finding a nice tractable mathematical model, the normal distribution, is of no use if it provides an inaccurate model of the data that you care about. Fortunately, many random variables have an approximately normal distribution. So if for example, I were doing a real lecture and I had 100 students in this room, and I were to look at the heights of the students, we would find that they are roughly normally distributed. Any time you take a population of people and you look at it, it's quite striking that you do end up getting a normal distribution of the heights. You get a normal distribution of the weights.

Same thing will be true if you look at plants, all sorts of things like that. I don't know why this is true. It just is.

What I do know is that-- and probably this is more important-- many experimental setups, and this is what we're going to be talking about going forward, have normally distributed measurement errors. This assumption was used first in the early 1800s by the German mathematician and physicist Carl Gauss. You've probably heard of Gauss, who assumed a normal distribution of measurement errors in his analysis of astronomical data.

So he was measuring various things in the heavens. He knew his measurements of where something was were not 100% accurate. And he said, well, I'll bet it's equally

likely it's to the left of where I think it is or the right as where I think it is. And I'll bet the further I get from its true value, the less likely I am to guess that's where it is. And he invented at that time what we now call the normal distribution. Physicists insist today still on calling it a Gaussian distribution. And it turned out to be a very accurate model of the measurement errors he would make.

If you guys are in a chemistry lab, or a physics lab, or a bio lab, mechanical engineering lab, any lab where you're measuring things, it's pretty likely that the mistakes you will make will be normally distributed. And it's not just because you were sloppy in the lab. Actually, if you were sloppy in the lab, they may not be normally distributed. If you're not sloppy in the lab, they'll be normally distributed.

It's true of almost all measurements. And in fact, most of science assumes normal distributions of measurement errors in reaching conclusions about the validity of their data. And we'll see some examples of that as we go forward. Thanks a lot. See you next time.