PROFESSOR: So, if you remember, just before the break, as long ago as it was, we had looked at the problem of fitting curves to data. And the example we had seen, is that it's often possible, in fact, usually possible, to find a good fit to old values. What we looked at was, we looked at a small number of points, we took a high degree polynomial, sure enough, we got a great fit. The difficulty was, a great fit to old values does not necessarily imply a good fit to new values. And in general, that's somewhat worrisome.

So now I want to spend a little bit of time I'm looking at some tools, that we can use to better understand the notion of, when we have a bunch of points, what do they look like? How does the variation work? This gets back to a concept that we've used a number of times, which is a notion of a distribution. Remember, the whole logic behind our idea of using simulation, or polling, or any kind of statistical technique, is the assumption that the values we would draw were representative of the values of the larger population. We're sampling some subset of the population, and we're assuming that that sample is representative of the greater population. We talked about several different issues related to that.

I now want to look at that a little bit more formally. And we'll start with the very old problem of rolling dice. I presume you've all seen what a pair of dice look like, right? They've got the numbers 1 through 6 on them, you roll them and something comes up. If you haven't seen it, if you look at the very back, at the back page of the handout today, you'll see a picture of a very old die. Some time from the fourth to the second century BC. Looks remarkably like a modern dice, except it's not made out of plastic, it's made out of bones. And in fact, if you were interested in the history of gambling, or if you happen to play with dice, people do call them bones. And that just dates back to the fact that the original ones were made that way. And in fact, what we'll see is, that in the history of probability and statistics, an awful lot of the math that we take for granted today, came from people's attempts to understand various games of chance.

So, let's look at it. So we'll look at this program. You should have this in the front of the handout. So I'm going to start with a fair dice. That is to say, when you roll it, it's equally probable that you get 1, 2, 3, 4, 5, or 6. And I'm going to throw a pair. You can see it's very simple. I'll take d 1, first die is random dot choice from vals 1. d 2 will be random dot choice from vals 2. So I'm going to pass it in two sets of possible values, and randomly choose one or the other, and then return them. And the way I'll conduct a trial is, I'll take some number of throws, and two different kinds of dice. Throws will be the empty set, actually, yeah. And then I'll just do it. For i in range number of throws, d 1, d 2 is equal to throw a pair, and then I'll append it, and then I'll return it. Very simple, right? Could hardly imagine a simpler little program.

And then, we'll analyze it. And we're going to analyze it. Well, first let's analyze it one way, and then we'll look at something slightly different. I'm going to conduct some number of trials with two fair die. Then I'm going to make a histogram, because I happen to know that there are only 11 possible values, I'll make 11 bins. You may not have seen this locution here, Pylab dot x ticks. That's telling it where to put the markers on the x-axis, and what they should be. In this case 2 through 12, and then I'll label it. So let's run this program. And here we see the distribution of values. So we see that I get more 7s than anything else, and fewer 2s and 12s. Snake eyes and boxcars to you gamblers. And it's a beautiful distribution, in some sense. I ran it enough trials. This kind of distribution is called normal. Also sometimes called Gaussian, after the mathematician Gauss. Sometimes called the bell curve, because in someone's imagination it looks like a bell.

We see these things all the time. They're called normal, or sometimes even natural, because it's probably the most commonly observed probability distribution in nature. First documented, although I'm sure not first seen, by deMoivre and Laplace in the 1700s. And then in the 1800s, Gauss used it to analyze astronomical data. And it got to be called, in that case, the Gaussian distribution. So where do we see it occur? We see it occurring all over the place. We certainly see it rolling dice. We see it occur in things like the distribution of heights. If we were to take the height of all the students at MIT and plot the distribution, I would be astonished if it didn't look more or less like that. It would be a normal distribution. A lot of things in the same height. Now presumably, we'd have to round off to the nearest millimeter or something. And a few really tall people, and a few really short people.

It's just astonishing in nature how often we look at these things. The graph looks exactly like that, or similar to that. The shape is roughly that. The normal distribution can be described, interestingly enough, with just two numbers. The mean and the standard deviation. So if I give you those two numbers, you can draw that curve. Now you might not be able to label, you couldn't label the axes, because how would you know how many trials I did, right? Whether I did 100, or 1,000 or a million, but the shape would always be the same. And if I were to, instead of doing, 100,000 throws of the dice, as I did here, I did a million, the label on the y-axis would change, but the shape would be absolutely identical. This is what's called a stable distribution. As you change the scale, the shape doesn't change.

So the mean tells us where it's centered, and the standard deviation, basically, is a measure of statistical dispersion. It tells us how widely spread the points of the data set are. If many points are going to be very close to the mean, then the standard deviation is what, big or small? Pardon? Small. If they're spread out, and it's kind of a flat bell, then the standard deviation will be large. And I'm sure you've all seen standard deviations. We give exams, and we say here's the mean, here's the standard deviation. And the notion is, that's trying to tell you what the average score was, and how spread out they are. Now as it happens, rarely do we have exams that actually fall on a bell curve. Like this. So in a way, don't be deceived by thinking that we're really giving you a good measure of the dispersion, in the sense, that we would get with the bell curve.

So the standard deviation does have a formal value, usually written sigma, and it's the estimates of x squared minus the estimates of x, and then I take all of this, the estimates of x, right, squared. So I don't worry much about this, but what I'm basically doing is, x is all of the values I have. And I can square each of the values, and then I subtract from that, the sum of the values, squaring that. What's more important than this formula, for most uses, is what people think of as the -- why didn't I write it down -- this is interesting. There is a some number -- I see why, I did write it down, it just got printed on two-sided. Is the Empirical Rule. And this applies for normal distributions. So anyone know how much of the data that you should expect to fall within one standard deviation of the mean? 68. So 68% within one, 95% of the data falls within two, and almost all of the data within three. These values are approximations, by the way. So, this is, really 95% falls within 1.96 standard deviations, it's not two. But this gives you a sense of how spread out it is. And again, this applies only for a normal distribution. If you compute the standard deviation this way, and apply it to something other than a normal distribution, there's no reason to expect that Empirical Rule will hold. OK people with me on this? It's amazing to me how many people in society talk about standard deviations without actually knowing what they are.

And there's another way to look at the same data, or almost the same data. Since it's a random experiment, it won't be exactly the same. So as before, we had the distribution, and fortunately it looks pretty much like the last one. We would've expected that. And I've now done something, another way of looking at the same information, really, is, I printed, I plotted, the probabilities of different values. So we can see here that the probability of getting a 7 is about 0.17 or something like that. Now, since I threw 100,000 die, it's not surprising that the probability of 0.17 looks about the same as 17,000 over here. But it's just a different way of looking at things. Right, had I thrown some different number, it might have been harder to visualize what the probability distribution looked like. But we often do talk about that. How probable is a certain value?

People who design games of chance, by the way, something I've been meaning to say. You'll notice down here there's just, when we want to save these things, there's this little icon that's a floppy disk, to indicate store. And I thought maybe many of you'd never seen a floppy disk, so I decided to bring one in. You've seen the icons. And probably by the time most of you got, any you ever had a machine with a quote floppy drive? Did it actually flop the disk? No, they were pretty rigid, but the old floppy disks were really floppy. Hence they got the name. And, you know it's kind of like a giant size version. And it's amazing how people will probably continue to talk about floppy disks as long as they talk about dialing a telephone. And probably none of you've ever actually dialed a phone, for that matter, just pushed buttons . But they used to have dials that you would twirl. Anyway, I just thought everyone should at least see a floppy disk once. This is, by the way, a very good way to get data security. There's probably no way in the world to read the information on this disk anymore.

All right, as I said people who design games of chance understand these probabilities very well. So I'm gonna now look at, show how we can understand these things in some other ways of popular example. A game of dice. Has anyone here ever played the game called craps? Did you win or lose money? You won. All right, you beat the odds. Well, it's a very popular game, and I'm going to explain it to you. As you will see, this is not an endorsement of gambling, because one of the things you will notice is, you are likely to lose money if you do this. So I tend not, I don't do it.

All right, so how does a game of craps work? You start by rolling two dice. If you get a 7 or an 11, the roller, we'll call that the shooter, you win. If you get a 2, 3, or a 12, you lose. I'm assuming here, you're betting what's called the pass line. There are different ways to bet, this is the most common way to bet, we'll just deal with that. If it's not any of these, what you get is, otherwise, the number becomes what's called the point. Once you've got the point, you keep rolling the dice until 1 of o things happens. You get a 7, in which case you lose, or you get the point, in which case you win. So it's a pretty simple game. Very popular game. So I've implemented. So one of the interesting things about this is, if you try and actually figure out what the probabilities are using pencil and paper, you can, but it gets a little bit involved. Gets involved because you have to, all right, what are the odds of winning or losing on the first throw? Well, you can compute those pretty easily, and you can see that you'd actually win more than you lose, on the first throw. But if you look at the distribution of 7s and 11s and 2s, 3s, and 12s, you add them up, you'll see, well, this is more likely than this. But then you say, suppose I don't get those. What's the likelihood of getting each other possible point value, and then given that point value, what's the likelihood of getting that before a 7? And you can do it, but it gets very tedious.

So those of us who are inclined to think computationally, and I hope by now that's all of you as well as me, say well, instead of doing the probabilities by hand, I'm just going to write a little program. And it's a program that took me maybe 10 minutes to write. You can see it's quite small, I did it yesterday. So the first function here is craps, it returns true if the shooter wins by betting the pass line. And it's just does what I said. Rolls them, if the total is 1 or 11, it returns true, you win. If it's 2, 3, or 12, it returns false, you lose. Otherwise the point becomes the total. And then while true, I'll just keep rolling. Until either, if the total gets the point, I return true. Or if the total's equal 7, I return false. And that's it. So essentially I just took these rules, typed them down, and I had my game.

And then I'll simulate it will some number of bets. Keeping track of the numbers of wins and losses. Just by incrementing 1 or the other, depending upon whether I return true or false. I'm going to, just to show what we do, print the number of wins and losses. And then compute, how does the house do? Not the gambler, but the person who's running the game, the casino. Or in other circumstances, other places. And then we'll see how that goes.

And I'll try it with 100,000 games. Now, this is more than 100,000 rolls of the dice, right? Because I don't get a 7 or 11, I keep rolling. So before I do it, I'll as the easy question first. Who thinks the casino wins more often than the player? Who thinks the player wins more often than the casino? Well, very logical, casinos are not in business of giving away money.

So now the more interesting question. How steep do you think the odds are in the house's favor? Anyone want to guess? Actually pretty thin. Let's run it and see. So what we see is, the house wins 50, in this case 50.424% of the time. Not a lot. On the other hand, if people bet 100,000, the house wins 848. Now, 100,000 is actually a small number. Let's get rid of these, should have gotten rid of these figures, you don't need to see them every time. We'll keep one figure, just for fun. Let's try it again. Probably get a little different answer. Considerable different, but still, less than 51% of the time in this trial. But you can see that the house is slowly but surely going to get rich playing this game.

Now let's ask the other interesting question. Just for fun, suppose we want to cheat. Now, I realize none of you would never do that. But let's consider using a pair of loaded dice. So there's a long history, well you can imagine when you looked at that old bone I showed you, that it wasn't exactly fair. That some sides were a little heavier than others, and in fact you didn't get, say, a 5 exactly 1/6 of the time. And therefore, if you were using your own dice, instead of somebody else's, and you knew what was most likely, you might do better. Well, the modern version of that is, people do cheat by putting little weights in dice, to just make tiny changes in the probability of one number or another coming up. So let's do that. And let's first ask the question, well, what would be a nice way to do that? It's very easy here. If we look at it, all I've done is, I've changed the distribution of values, so instead of here being 1, 2, 3, 4, 5, and 6, it's 1, 2, 3, 4, 5, 5, and 6. I snuck in an extra 5 on one of the two dice. So this has changed the odds of rolling a 5 from 1 in 6 to roughly 3 in 12. Now 1/6, which is 2/12, vs. 3/12, it's not a big difference. And you can imagine, if you were sitting there watching it, you wouldn't notice that 5 was coming up a little bit more often than you expected. Normally. Close enough that you wouldn't notice it.

But let's see if, what difference it makes? What difference do you think it will make here? First of all, is it going to be better or worse for the player? Who thinks better? Who thinks worse? Who thinks they haven't a clue? All right, we have an honest man. Where is Diogenes when we we need him? The reward for honesty. I could reward you and wake him up at the same time. It's good. All right, well, let's see what happens. All right, so suddenly, the odds have swung in favor of the player. This tiny little change has now made it likely that the player win money, instead of the house. So what's the point? The point is not, you should go out and try and cheat casinos, because you'll probably find an unpleasant consequence of that. The point is that, once I've written this simulation, I can play thought experiments in a very easy way. So-called what if games. What if we did this? What if we did that? And it's trivial to do those kinds of things. And that's one of the reasons we typically do try and write these simulations. So that we can experiment with things. Are there any other experiments people would like to perform while we're here? Any other sets of die you might like to try?

All right, someone give me a suggestion of something that might work in the house's favor. Suppose a casino wanted to cheat. What do you think would help them out? Yeah?

STUDENT: Increase prevalence of 1, instead of 5?

PROFESSOR: All right, so let's see if we increase the probability of 1, what it does? Yep, clearly helped the house out, didn't it? So that would be a good thing for the house. Again, you know, three key strokes and we get to try it. It's really a very nice kind of thing to be able to do. OK, this works nicely. We'll get normal distributions. We can look at some things.

There are two other kinds of distributions I want to talk about. We can get rid of this distraction. As you can imagine, I played a lot with these things, just cause it was fun once I had it. You have these in your handout. So the one on the upper right, is the Gaussian, or normal, distribution we've been talking about. As I said earlier, quite common, we see it a lot. The upper left is what's called a, and these, by the way, all of these distributions are symmetric, just in this particular picture. How do you spell symmetric, one or two m's? I help here. That right? OK, thank you. And they're symmetric in the sense that, if you take the mean, it looks the same on both sides of the mean. Now in general, you can have asymmetric distributions as well. But for simplicity, we'll here look at symmetric ones.

So we've seen the bell curve, and then on the upper left is what's called the uniform. In a uniform distribution, each value in the range is equally likely. So to characterize it, you only need to give the range of values. I say the values range from 0 to 100, and it tells me everything I know about the uniform distribution. Each value in that will occur the same number of times. Have we seen a uniform distribution? What have we seen that's uniform here? Pardon?

STUDENT: Playing dice.

PROFESSOR: Playing dice. Exactly right. Each roll of the die was equally likely. Between 1 and 6. So, we got a normal distribution when I summed them, but if I gave you the distribution of a single die, it would have been uniform, right? So there's an interesting lesson there. One die, the distribution was uniform, but when I summed them, I ended up getting a normal distribution.

So where else do we see them? In principle, lottery winners are uniformly distributed. Each number is equally likely to come up. To a first approximation, birthdays are uniformly distributed, things like that. But, in fact, they rarely arise in nature. You'll hardly ever run a physics experiment, or a biology experiment, or anything like that, and come up with a uniform distribution. Nor do they arise very often in complex systems. So if you look at what happens in financial markets, none of the interesting distributions are uniform. You know, the prices of stocks, for example, are clearly not uniformly distributed. Up days and down days in the stock market are not uniformly distributed. Winners of football games are not uniformly distributed. People seem to like to use them in games of chance, because they seem fair, but mostly you see them only in invented things, rather than real things.

The third kind of distribution, the one in the bottom, is the exponential distribution. That's actually quite common in the real world. It's often used, for example, to model arrival times. If you want to model the frequency at which, say, automobiles arrive, get on the Mass Turnpike, you would find that the arrivals are exponential. We see with an exponential is, things fall off much more steeply around the mean than with the normal distribution. All right, that make sense? What else is exponentially distributed? Requests for web pages are often exponentially distributed. The amount of traffic at a website. How frequently they arrive. We'll see much more starting next week, or maybe even starting Thursday, about exponential distributions, as we go on with a final case study that we'll be dealing with in the course.

You can think of each of these, by the way, as increasing order of predictability. Uniform distribution means the result is most unpredictable, it could be anything. A normal distribution says, well, it's pretty predictable. Again, depending on the standard deviation. If you guess the mean, you're pretty close to right. The exponential is very predictable. Most of the answers are right around the mean. Now there are many other distributions, there are Pareto distributions which have fat tails, there are fractal distributions, there are all sorts of things. We won't go into to those details.

Now, I hope you didn't find this short excursion into statistics either too boring or too confusing. The point was not to teach you statistics, probability, we have multiple courses to do that. But to give you some tools that would help improve your intuition in thinking about data. In closing, I want to give a few words about the misuse of data. Since I think we misuse data an awful lot. So, point number 0, as in the most important, is beware of people who give you properties of data, but not the data. We see that sort of thing all the time. Where people come in, and they say, OK, here it is, here's the mean value of the quiz, and here's the standard deviation of the quiz, and that just doesn't really tell you where you stand, in some sense. Because it's probably not normally distributed. You want to see the data. At the very least, if you see the data, you can then say, yeah, it is normally distributed, so the standard deviation is meaningful, or not meaningful. So, whenever you can, try and get, at least, to see the data.

So that's 1, or 0. 1 is, well, all right. I'm going to test your Latin. Cum hoc ergo propter hoc. All right. I need a Latin scholar to translate this. Did not one of you take Latin in high school? We have someone who did. Go ahead.

STUDENT: I think it means, with this, therefore, because of this.

PROFESSOR: Exactly right. With this, therefore, because of this. I'm glad that at least one person has a classical education. I don't, by the way. Essentially what this is telling us, is that correlation does not imply causation. So sometimes two things go together. They both go up, they both go down. And people jump to the conclusion that one causes the other. That there's a cause and effect relationship. That is just not true. It's what's called a logical fallacy. So we see some examples of this. And you can get into big trouble.

So here's a very interesting one. There was a very widely reported epidemiological study, that's a medical study where you get statistics about large populations. And it showed that women, who are taking hormone replacement therapy, were found to have a lower incidence of coronary heart disease than women who didn't. This was a big study of a lot of women. This led doctors to propose that hormone replacement therapy for middle aged women was protective against coronary heart disease. And in fact, in response to this, a large number of medical societies recommended this. And a large number of women were given this therapy. Later, controlled trials showed that in fact, hormone replacement therapy in women caused a small and significant increase in coronary heart disease. So they had taken the fact that these were correlated, said one causes the other, made a prescription, and it turned out to be the wrong one.

Now, how could this be? How could this be? It turned out that the women in the original study who were taking the hormone replacement therapy, tended to be from a higher socioeconomic group than those who didn't. Because the therapy was not covered by insurance, so the women who took it were wealthy. Turns out wealthy people do a lot of other things that are protective of their hearts. And, therefore, are in general healthier than poor people. This is not a surprise. Rich people are healthier than poor people. And so in fact, it was this third variable that was actually the meaningful one. This is what is called in statistics, a lurking variable.

Both of the things they were looking at in this study, who took the therapy, and who had a heart disease, each of those was correlated with the lurking variable of socioeconomic position. And so, in effect, there was no cause and effect relationship. And once they did another study, in which the lurking variable was controlled, and they looked at heart disease among rich women separately from poor women, with this therapy, they discovered that therapy was not good. It was, in fact, harmful. So this is a very important moral to remember. When you look at correlations, don't assume cause and effect. And don't assume that there isn't a lurking variable that really is the dominant factor. So that's one statistical, a second statistical worry.

Number 2 is, beware of what's called, non-response bias. Which is another fancy way of saying, non-representative samples. No one doing a study beyond the trivial can sample everybody or everything. And only mind readers can be sure of what they've missed. Unless, of course, people choose to miss things on purpose. Which you also see. And that brings me to my next anecdote.

A former professor at the University of Nebraska, who later headed a group called The Family Research Institute, which some of you may have heard about, claimed that gay men have an average life expectancy of 43 years. And they did a study full of statistics showing that this was the case. And the key was, they calculated the figure by checking gay newspapers for obituaries and news about stories of death. So they went through the gay newspapers, took a list of everybody whose obituary appeared, how old they were when they died, took the average, and said it was 43. Then they did a bunch of statistics, with all sorts of tests, showing how, you know, what the curves look like, the distributions, and the significance. All the math was valid. The problem was, it was a very unrepresentative sample. What was the most unrepresentative thing about it? Somebody?

STUDENT: Not all deaths have obituaries. PROFESSOR: Well, that's one thing. That's certainly true. But what else? Well, not all gay people are dead, right? So if you're looking at obituaries, you're in fact only getting -- I'm sure that's what you were planning to say -- sorry. You're only getting the people who are dead, so it's clearly going to make the number look smaller, right? Furthermore, you're only getting the ones that were reported in newspapers, the newspapers are typically urban, rather than out in rural areas, so it turns out, it's also biased against gays who chose not come out of the closet, and therefore didn't appear in these. Lots and lots of things with the problems.

Believe it or not, this paper was published in a reputable journal. And someone checked all the math, but missed the fact that all of that was irrelevant because the sample was wrong. Data enhancement. It even sounds bad, right? You run an experiment, you get your data, and you enhance it. It's kind of like when you ran those physics experiments in high school, and you've got answers that you knew didn't match the theory, so you fudged the data? I know none of you would have ever done that, but some people been known to do. That's not actually what this means. What this means is, reading more into the data than it actually implies. So well-meaning people are often the guiltiest here.

So here's another one of my favorites. For example, there are people who try to scare us into driving safely. Driving safely is a good thing. By telling holiday deaths. So you'll read things like, 400 killed on the highways over long weekend. It sounds really bad, until you observe the fact that roughly 400 people are killed on any 3-day period. And in fact, it's no higher on the holiday weekends. I'll bet you all thought more people got killed on holiday weekends. Well, typically not. They just report how many died, but they don't tell you the context, say, oh, by the way, take any 3-day period. So the moral there is, you really want to place the data in context. Data taken out of context without comparison is usually meaningless.

Another variance of this is extrapolation. A commonly quoted statistic. Most auto accidents happen within 10 miles of home. Anyone here heard that statistic? It's true, but what does it mean? Well, people tend to say it means, it's dangerous to drive when you're near home. But in fact, most driving is done within 10 miles of home.

Furthermore, we don't actually know where home is. Home is where the car is supposedly garaged on the state registration forms. So, data enhancements would suggest that I should register my car in Alaska. And then I would never be driving within 10 miles of home, and I would be much safer. Well, it's probably not a fact. So there are all sorts of things on that. Well, I think I will come back to this, because I have a couple more good stories which I hate not to give you. So we'll come back on Thursday and look at a couple of more things that can go wrong with statistics.