Week 6
Video Transcripts

Video 1 (6:42): Motivation of Central Limit Theorem

What we're going to do is, we're going to talk about sampling. So, we're going to return to statistics, and in particular, we're going to use data to do estimation. And in that context, we're going to talk about two things. What's the accuracy of our estimate, and how big of a sample do we need in order to achieve the required accuracy? Before I do that, I want to show you some pictures of something, that we are going to be using. It is called the central limit theorem. So, here is a motivating application which I'm not going to solve exactly, but I want to illustrate using the data, what it...what it talks about. Suppose, we wanted to pick how much money you want to put in an ATM in order to satisfy the demand for—from people working into the—to the ATM overnight. So, let's say, you have an assessment of how many people are going to show up, and now you want to pick how much–how many dollars you want to put into the machine. So, one way to do it is you and collect data, and this table, that I'm showing you here has data from an ATM location, in the upper west side of Manhattan. The data is a little bit dated. It is about 15 years old, but it was actual data. Now we're going. Now, one way to look at it is to say, look the mean–the average is about 125 dollars, the standard deviation is about 55 dollars. So, these are the numbers that you get from this table, so maybe you know, I can use these numbers to compute that. What I am going to do next, I'm going to show you some—the histograms of this data, and a few versions of it, and then I'm going to move on. All right! So, first thing we can do is, take this table and plot the histogram. What you realize is most people get a 100 dollars, or used to get 100 dollars, some people get 200 dollars, and then people get sort of amounts all over. Right! Now, is this normally distributed? The answer is no. Right! Clearly, it has multiple peaks, there are gaps in the data. This doesn't quite look normal to us. Now, what I would like to do is I would like to pick how many dollars I want to put to the ATM overnight. So, the reality is, I don't actually care about how much money one guy gets from the ATM. But what I really care is how much money many people together get from the ATM. So, what I could do is, I could go back to my data, and I can randomly pick people, pair them up, let's say, so now, I have two guys together, I sum their amount, so I have the amount that two people get together, and then I can plot a histogram for these amounts. And I can do it for twos and fives and tens and until I get sort of how many people, let's say 100 people get together, or 200 people get together, which is what I am kind of interested in. If I group them in twos, I get this histogram. So, I want to make two observations—three observations. First one, is...is getting closer to looking like something that is bell-shaped, like a normal distribution. The average for '2' people is twice the average for '1' person. Right! So, that should be '2' times '125' dollars, which is about '250'. Right! Is equal to '250', but the...the meaning is a little bit different. Okay! So, that's observation number two. Observation number three, all right, is that the standard deviation from '2' visitors— '2' visits to the ATM is larger than the standard deviation from '1' visitor, but is not twice as large, is square root of '2' larger from '1' visitor. So, '1'. Okay! So, the standard deviation grows– the mean grows, the standard deviation also grows, but

the standard deviation doesn't grow as fast. Right! And why does it grow, as square root of '2' versus '2'? Well, what's happening is, the first person may actually get a very high amount, the next person may also get a high amount, but may also get a low amount, so there is going to be some cancellation between two people, and as a result, they sort of–these overall standard deviation, when I look at people like pairs will grow, but it will grow a little bit more slowly. You can go back to the stuff that we did, in the past lectures, in fact, and confirm to yourself that this would be the case. Right! That...that the standard deviation grows like the square root of the number of people that I'm including in my sample. If I include '5' people, which not that many, you realize that this is close to normal. The mean for '5' is '5' times the mean for '1', and standard deviation from '5' people is square root of '5' times the standard deviation from one person. There is still a skewness to the right, but the reality is, if I keep averaging, if I keep playing this game, this will all disappear, and we will actually start to steer into a normal distribution. Okay! And this is actual data.

Video 2 (6:01): Central Limit Theorem

One could ask why is this happening, and the answer to that question is called the central limit theorem. So, in particular, the central limit theorem says the following. It says that, if I have data, these are random variables, and they all have the same distributions. So, these are like all iid., independent draws from the same kind of pool of visitors to the ATM. All right! So, if they're all coming from the same population, and I sum them up and they're all independent, then when I look at the sum, these will be like the total amount requested from the ATM by many people. Then this goes to a normal distribution. Okay! And what did we say before? We said that the mean from 'n' visitors is equal to 'n' times the mean of '1' visitor. So, this is the first thing that we said. And the next thing that we said before is the standard deviation from 'n' visitors is square root of 'n' and the standard deviation from '1' visitor. Okay! And...and...and from that, we get that the variance for 'n' people is 'n' times the variance from '1' person, which is this. Right! The variance of one guy sigma squared, so if I sum 'n' of them, I get 'n' times sigma squared. So, this is really exactly what we were getting in the picture. If I look at the sums, I will get the normal distribution, it will be centered around where I think it is 'n' times the '1' guy average. And if I look at the variance, it will be 'n' times the variance of '1' person. And if I look at the standard deviation, it will be square root of 'n' times the standard deviation of '1' person, the dollar amount requested by '1' guy. Now, what I'm going to do next is I'm going to use exactly this result, but I'm going to do the following. I'm not going to look at the sum. Okay! But I'm going to look at the sample mean. So, I'm going to take the sum, let's say, I'm going to sum up…up the dollar amount requested by 50 people, I'm going to sum it all up and then I'm going to divide by 50. And if you think about it, this is my sample estimate from 'n' observations for the mean. Okay! So, I don't know exactly what's the average amount of the entire population of visitors to the ATM, I just collected some data, I look at that, and I get some estimate, 'X' bar. Reality is, if I take instead of this data set sample, I get a slightly different sample, I may get a slightly different estimate, okay, but this should be close. And...and if I plot a histogram of these estimates, if I take this sum divided by 'n' what do I know? I know for a fact that this will be normally distributed. Let me just delete this line, so that it doesn't block us. It will be normally distributed. If I divide by 'n', the mean will become mu, so this is the kind of mean of the underlying

population. And the variance will be sigma squared over 'n' or...or else the standard deviation of 'X' bar, which is what I'm really after, is sigma. If I take this thing and divide it by 'n', I'm going to get sigma over squared of 'n'. All right! So, in...in...in...in crude terms, what...what do I know about 'X' bar? 'X' bar will be correct estimate or unbiased estimate on average. All right! And its error-which is captured by the standard deviation- its error goes down like '1' over square root of 'n'. So, the fact that the error will go down is intuitive. If I get more data, I will get a better estimate and the error from the truth will be smaller. Okay! Now, the question is, how is my accuracy improving as my sample size is improving. And the answer to that is, like '1' over square root of 'n'. So, I'm going to use exactly these results to go back and do estimation. So, I will, in fact, skip through this result, but you can try it by yourself and compute how much money you should put into the ATM. Let's say, if you were planning for '225' visitors. This is an illustration of how the central limit theorem works for other distributions. So, if I sum normals, I know I get the normal, always. Right! We saw that before. But if I sum different things like the histogram for dollar amounts requested by the ATM, it takes a little bit of time until I get to the normal, but I always get to the normal.

Video 3 (6:40): Sampling

All right! So now, I'm going to use exactly that. And in particular, I'm going to start from data and I'm going to use statistics, to let's say, estimate parameters that I'm going to use in my models. So, in the past we've seen examples where, let's say, we had estimates for the average demand for a product, or the average return of a fund. And, the question is, where is these numbers coming from? Well, they come from data. All right! So now, I'm going to go to the data, estimate the number, and in addition to the estimate, I'm going to actually start talking about how accurate is the estimate. So, that's what we're going to do for the remainder of this session. So, here is the data that I'm going to be looking at. So, this is a...a study that Kaiser Family Foundation did. And, they...they collect a sample from school-age kids, and they measure how much time do they spend in front of, sort of, online media. So, that could be TV, computers, video games, and the like, or media, in general— printing, print as well. Now—and, what we have here is numbers about how much time do they spend over time. And, this is how much time they spent across the different categories. One thing that is impressive or shocking, for some of us that have kids, is that these numbers are very high. So, this says that, accounting for multitasking—the fact that you may be doing two things at the same time, kids–school-age kids, in 2009 in the survey, that is higher right now, were spending about seven and a half hours per day, in front of these–watching TV or listening to their music or playing in their computer, or playing video games, etc. You also see that some of–some things are increasing. Video games has been increasing, computers has been increasing, TVs have been increasing, music and audio has been increasing, everything has been increasing, apart from print. So—all right! So, what I want to do is I want to look at these numbers, for example, this number seems to be significantly higher than five years earlier. The question is, "How accurate is this estimate of seven and a half hours?" Right! How much data do we need to achieve a desired accuracy...accuracy? So, these are the questions that I want to answer. If you...if you...if you want, you know, we're...we're going to analyze the survey, the results of a survey, and we're going to talk about the accuracy. And, the other thing that we may want to do is we want to put ourselves in the...in the...in the shoes of

somebody that wants to design a survey. And, before they actually go and conduct it, they want to decide how many people they need to include in their survey, when they conduct it. So, there is some information here, which I'm not going to read in detail. You can read it offline. It says that about '702' people that were included in the survey, and this is a random sample of school-age kids across the United States. They actually kept detailed diaries of how much time they spent in front of each of these media per day. And, this is the data that was used in order to compute the results. So, this is the...this is the important number here. All right! So, I'm going to use that number later on. So, let's...let's keep going on. So, what did they do? They...they used '702' people that were completing these media diaries, then I looked at the average. I look at—each of these '702' people gave me a number. They said, "I spent, you know, nine hours and 42 minutes, three hours and 17 minutes, etc." I took all of these and I averaged them out, and I got this number '458' minutes. This is '7' hours and '38', right, so, I'll...I'll write it in minutes. So, I get this number, and I know this number is much bigger than what was the case five years earlier. And, perhaps, I want to...I want to ask myself whether '702' kids is sufficient for me to get a good enough accuracy in my estimate. For example, I may want to ask myself, what's the probability that the rest of it is off from the truth, by more than '15' minutes? Okay! So, all right! So, here–so, that's...that's it, right, so, that's...that's the question. So, here's what we did. We collected these [inaudible]. I am going to assume that the person that picked these people to respond, did it carefully, so that I'm sampling throughout the country, and I'm not picking, let's say, kids that are—all have the same exam tomorrow and as a result, they're—they're studying for their exam. So, they're going to be independent. And, because I'm sampling the entire country accurately, they're going to be—I'm going to think of them as what I call identically distributed. So, they're random samples from the U.S. population, and it's representative. All right! So, that's the set up. And, these are the setup–these are the assumptions of the central limit theorem.

Video 4 (13:03): Accuracy of Sample Mean for Fixed Sample Size

And now, what I'm going to do is, I'm going to try to estimate the true population—average number of minutes that kids, school-age kids spend in front of these media outlets. And, population is all U.S. school-age, eight to 18 year-old kids. So, that's my population. All right! So, if I could ask every single kid in the United States to keep a detailed diary, then I would know the truth, but that is too expensive, and...and too laborious, and...and a waste of everybody's time, so what I do is I draw a sample. So, I draw a sample of '702' kids, I get their...I get their answers, I average them out, and then I get an estimate. And, this estimate presumably is going to have some kind of error. Right! Because I didn't ask everybody, and had I asked a different group of '702' people, maybe I would have gotten a slightly different answer, and the question is by how much. Right! So, I want to...I want to estimate what is the probability that my estimate 'X' bar is off, it's either too big, plus the probability that my estimate is too small. Right! So, what is the probability that my error—I'm either too high in my estimate or too low in my estimate, and my...my–I'm I'm willing to live that is, sort of, some notion of tolerance. If you get it within '15' minutes, I'm happy. All right! So, what do we know? Well, we know, from the central limit theorem, that 'X' bar is normally distributed, it's centered around mu, and the variance is sigma squared, over 'n'. Right! Or, the standard deviation of 'X' bar, which I'm going to give a different name—I'm going

to also call it the standard error of 'X' bar, is sigma, over square root of 'n', that's sigma over square root of '702'. All right! So, I'm going to use that to compute that probability. So, here is what we have. Right! Let's...Let's draw the picture. So, the distribution for 'X' bar, it's centered at the truth mu, which is unknown, and I want to figure out what is the probability that by lack of a draw, so to speak, I got something that is—these are...these are–this is the probability that 'X' bar is bigger than mu plus '15'. Right!

That's this probability. Now, so I know that distribution is normal. I know it's centered at the mean, so I just need to compute that probability. So, what I'm going to do is I'm going to figure out what's that distance. I'm going to measure this... measure this distance in standard deviations of 'X' bar. Right! That's what we always do. So, I'm going to say that's the probability that 'Z', the standard normal, is bigger than—so this distance here, this is mu plus '15'. Right! So, that distance is '15' minutes. And, I need to divide that by the standard deviation of 'X' bar. The standard deviation of 'X' bar is sigma, divided by square root of '702'. So, I'm...I'm, literally, almost done here. My...my problem—and I'm going to put it in red, is nobody told me, what's sigma. Right! So, unfortunately here, I either need to go back to the data to compute sigma, or I need to guess it. So, let me just do on a side here. We need to guess sigma. And, that sounds a little bit like, you know, why...why on earth are we doing that? So, there are two responses to that. If I had the data, I would go to the data and compute sigma, okay! But...but I personally don't have the data right now. The second thing is most often this calculation is being done before you collect the data, and at that point in time, you need to guess it. Right! So, an intelligent guess would be interesting to come about. And, what I want you to do is, literally, pause the video and think for three minutes about how you would guess that number in, let's say, hours, or number of minutes. How big is the standard deviation? So, what—standard deviation of what? So, people are going to give me responses, a 100 minutes, a 1,000 minutes, 400 minutes, 500 minutes, and I want to compute the standard deviation of that. How heterogeneous is the population, in terms of how much more or less time they spend in front of their computers, their iPod, and video games, and...and...and the like. All right! And...and...and, once you're done thinking, let's resume. All right So, I'm going to resume now. So, here is how...how I would guess. So, it's kind of tricky to guess sigma. All right! But what we could do, as a first step, is we could guess the range of responses. And, you know, I want to...I want to us add the identifier here, reasonable range of responses. So, how much time do you spend in front of these, per day? Or, what would be the range? Well, the range may be anywhere between '0', you know, I don't watch TV, I don't have a computer, never listen to the music, don't play videogames, don't read magazines or watch movies, I just study, all right, perfect, or do sports, or, so '0'. And then, perhaps we need to figure out some maximum. All right! So, maximum is clearly 24 hours, but maybe you know, you need to sleep a little bit, you know, do some other stuff with your life. I don't know, so at...at...at...at maximum could be, I don't know, 12... 12 hours, maybe '18' hours. I mean, that's like a... that's a very large upper bound, okay! That...that says that you spent 18 hours per day in front of these media outlets, I mean, you know, probably there—a reasonable number would have been 12, but let's be conservative. Right! Now, so I have the range now, and the next thing I'm going to say, is like...I'm going to say, like most data will be within plus or minus '3' sigma, oh sorry, plus or minus '3' sigma of the mean, which implies that the range is roughly '6' sigma wide, which from that implies that sigma is roughly '18' over '6', that's 3 hours. I'm sorry, '180' minutes. Well, let me just write it in. That's '3' hours

or '180' minutes. Okay! So, that's...that's the estimate. And, if instead of '18' hours we have 12 hours there, then that estimate would get 12 over six, that would be 120 minutes. Okay! So, that's—so, what do I know now? Well, it's not that I trust my estimate, but I know that that number should not be less than, I don't know, 75 or 80, and is not going to be bigger than, you know, '180' or 200. Okay! So, if I solve the problem with '180', I would get a conservative estimate of this probability of...of being–of having too big of an error, and if I solve it for, I don't know, 100 or 80, I would get that sort of more optimistic estimate. So now, I'm going to take this number and stick it in there. All right! So, I'm going to get the probability that 'Z' is bigger than '15', over '180', over square root of '702'. All right! And, that's the probability that 'Z' is bigger—if you do this calculation, '2.21'. All right! That is this probability. I need to, at the end, multiply it by two, that would be the probability that you're too small. So, let me just quickly show you what that number is. So, I have a table here. So, '2.21' would be here. All right! The probability that 'Z' is bigger than '2.21' is '1' minus '0.9864', and that is, essentially, about '1.4%', and then I need to multiply that by two. So, that number is '1.4%'. And, that is a pretty conservative estimate. Right! Because I put a sigma there that is huge. If sigma was smaller, then this number would go down. So, the probability that the arrow is too high is about twice that amount. Just write that in...in...in red. So, the probability that the error, in absolute value, exceeds '15' minutes is twice that probability which is '1.4%', which is '2.8%', which is low. Right! So, I'm pretty confident that the estimate is pretty robust.

Video 5 (7:05): Sample Size Determination

So, we want to estimate the population mean, we draw a sample, the sample needs to be selected accurately, and there are experts that can do that. Then, we take this sample, we apply the central limit theorem. The central limit theorem tells us that 'X' bar is normal and it's centered around the truth, and let's say, the standard error for 'X' bar looks like, sigma over square root of 'n'. I take these things and then, I compute what's the accuracy. When I compute the accuracy of my estimate, I have some kind of tolerance. Right! So, for example, plus or minus 15 minutes was okay. So, if you're within plus or minus 15 minutes, I'm okay. But, if you're bigger than that, the error is bigger than that, then I want to know about it. So, you...you give me a tolerance and then I go to the...to the central limit theorem, and I use it to compute the probability of exceeding that tolerance. The other thing I could have done, is I could have said, "Look! I want you to be within 15 minutes with that target probability, and now, I want you to compute how many people you need to include in the sample." And, if you think about that, that's the first question we answer before running the survey. So, we go–before we go and collect the data, we have to think about what we're estimating, and we need to decide how much data do we need. So, let's do that calculation for a second, and let's do it for the Kaiser example. So here, I want to figure out the following. I want to find the sample size, I'm willing to live with 15 minutes worth of error, but I want the probability of that to be less than or equal to '5%'. So, what do I know? I have this picture. The...the person that is going to run the survey says, "Look! If you happen to be too high or too low. I don't want that to happen too often. I want this probability to be '2.5%' and I want this probability to be '2.5%'." I plotted the normal distribution because I know that the distribution of 'X' bar is...is...is normal from the central limit theorem. So, what do I know? For the tail probability, to be '2.5%', this distance needs to be

how much, right, well, it needs to be '1.96' standard errors of 'X' bar. Okay! That's the only way...the only way you're going to have '2.5%' tail availability is, if you're '1.96' standard errors away. Now, this distance is '15' minutes, times one point—that needs to be equal to '1.96' times, sigma—I'll put '180' the square root of 'n'. This formula, here, is equal to the standard error. And now, I can solve for 'n'. Right! And in particular, 'n' will be '1.96' squared, times '180', over '15' squared, and that's '553'. I want to rewrite that formula. So, this says that 'n' is some kind of multiplier or—that depends on the target probability, e.g., '5%' squared, times sigma of population, divided by tolerance squared. Okay! So—and...and that formula works in every application, you have the standard deviation of...of...of the population or of the data that're going to be collecting, you divide it by the tolerance that you're willing to...to have, you square this, and you multiply that by the corresponding multiplier that you get from...from the 'Z' table squared. Okay! And, that's how you pick the sample size when you run—start surveys of...of that...of that type. All right! So, I will end at this point. As a summary, we talked about the central limit theorem. We talked about how to use it, to assess the estimation– the accuracy of estimates that we get from data, so, how does the accuracy depend on how much data we have, what's the probability that the error will be too high. And then, the third thing is if I know how accurate I want to be, let's compute how many people I need in my survey. All right! That's it. Thank you.

Video 6 (7:27): Polling Examples

We're going to continue our discussion about sampling. The difference today is going to be that we will go and collect data, and...and we're going to ask people a yes or no question, and then we're going to try to assess what's the accuracy of the—of this type of polling, where we estimate what fraction of people support 'X' versus 'Y' or believe something versus something else. So, we're going to talk about the—how do we do sampling or polling, what's the accuracy of these polls, how much—many people do we need to survey, etc. All right! So, let's...let's move on. So, here's the example I picked. So, this...this showed up recently in the press, and it was done by the Yale Foundation on climate change, and essentially, this is a poll that has been going on for a very long time, and...and to check whether Americans believe that climate change is, let's say man-made or potentially—or mostly man-made or...or influenced. And...and then, they also ask subsequent questions about, what do you think that this is going to affect our lives, do you think this is going to happen in ten years, 20 years, 50 years, etc. So, I'm just going to focus on one question and the question is, whether people believe that global warming or our climate change is mostly caused by, sort of, human activity. All right! Now, here is the, sort of, evolution of the responses of that poll. We have data for this one. All right! So, I'm going to go and just look at these numbers, and... and in some sense, I'm going to go and try to talk a little bit about the accuracy of this number. So, we polled some people. We got that '58%' of Americans, adults—these are 18 plus year-old Americans believe that, sort of, that the...the climate is changing, and this is influenced mostly by human activity. But the question is, how many people, how many respondents, right, how accurate, etc. right! If you remember last time, when we estimated how much time do students spend in front of the different type of media outlets, we said, you know, we're willing to live with an estimate that is plus or minus 15 minutes. So, you know, the...the estimator was about four and a half hours. So, plus or minus 15 minutes would've been, sort of, pretty reasonable. So here, we've got a '58%'. So, one

question would be, is it '58%' plus or minus two, plus or minus 5, plus or minus 10? What's the tolerance that we get within that estimate? All right! So, that's...that's the example that I'm going to work through with you. Let me just give you one slide to see how do these numbers differ in the US, relative to how people feel about them in other countries. So, this is a study that appeared a couple of months earlier, and it's a commissioned...it's a commissioned study in Europe. It focused on four countries, France, Germany, Norway, and the...and the UK, and they had, sort of—it's a big report that you can actually find it online. Main thing that I focus here was, sort of, this question which is, whether we believe that climate change is predominantly caused by natural processes, or partially natural and partially man-made, and then man-made, etc. You...you can think of this as being, sort of, the response where one would argue that the climate change is primarily driven by the natural evolution of the climate and the environment, and you can see that these numbers, when you sum them all up, these are small. Right! These are of the order of ten to 15%, whereas these numbers together are obviously the remainder. So, it...it seems like there's a difference in perception in the US than in other countries. I would like to return to the US poll now, which we're going to analyze. And in particular, I'm going to analyze this number here. Right! So, we said that—we went to the people and we said, "Look! I want you to tell us, what...what do you think that's the main cause of climate change? And in particular, is it mostly human caused? Or, is it mostly due to the natural change in the environment that happens over time?" All right! And what we got is we got that '58%' believe that it's mostly caused by humans. And then—I want to—I will skip through that, I want to focus on the fine print, and in particular the fine print is here, all right, '1,266' adults plus or minus '3 percentage points', with a '95%' confidence level. Okay! So, I want to go back now, and redo this poll, taking into account that it's '1,266' people, and I want to somehow verify these claims that we have in the report. In particular, one takeaway message from that is that with about 1,000 individuals, you get pretty reasonable accuracy, it seems, in these polls. All right! Plus or minus '3 percentage points' with '95%'confidence is a pretty reasonable degree of accuracy. And as a rule of thumb, when a major newspaper—at least in this country, publishes a poll or the results of a poll in their front page, you would expect that, you can guess immediately, that they had about 1,000 individuals included. They will never actually have 5,000, 10,000, or...or 50,000 individuals included. So, typical polls will have sample sizes of that...of that order of magnitude. In fact, as a...as a verification of that, if you look at the European study, independent of the American study, you'll realize that they also had about 1,000 individuals or respondents, in each country. So that they run it in four countries and they had 1,000 people respond in each of these locations.

Video 7 (8:39): Estimating Proportions

So, let's see what is going on. All right! So, in particular, we went and we asked '1,266' individuals, and...and we asked them whether they think that climate change is mostly human-caused, or was mostly caused by a change in the natural environment. Presumably, there were some other responses because these numbers don't add up to one. And we want to...we want to see how polling works, and then how this number relates to the plus or minus '3%' points, and then talk a little bit about whether a number of that magnitude is sufficiently large. Okay! So, in some ways, what we're going to do now, is review of something that you did earlier on when you talk about probability and the binomial

distribution. Let's see how this is going to work out. All right! So, what do we have? There is a...there is a ground truth. This is unknown, of course. And this is the—let's say fraction of all American adults, that believe that the climate change is mostly human-caused. So, this is the ground truth. Right! So, if we were to go and ask every single individual, and they told us what they believe, then we would have this true fraction. Instead of that, we'll go and collect a sample with '1,266' individuals. And each of these individuals will give us a response. And their response will be let's say a 'Yes', which we'll write it as a '1', or a 'No' which we can write it as a '0'. And this response is a random...random variable. All right! And my sample is representative in the sense that I've sampled all geographic areas, all demographic groups, all socio-economic groups, and... and...and...and the like. And I want us to convince ourselves of the following things. So, once I—when I'm about to go and grab a person and ask them that question, in the absence of knowing anything about the individual, if you were to ask me, what do you think is the probability they're going to say yes. Well, I would say, look, I have no idea about that individual. That individual should be representative of the American population. And as a result, they will probably say yes, with probability 'p'. I know I don't know that number, but, you know, that's my...that's my guess. And they will say no with probability '1' minus 'p'. All right! And then what I'm going to do is I'm going to say, look, I have this spreadsheet, and I'm going to make an estimate, and I'm going to denote that estimate by 'p' hat, and my estimate is simple. I'm just going to go into the spreadsheet and I'm going to count the yeses and divide by '1,266'. So, I'm going to say how many people said "yes", and I'll divide that by '1,266'. And in this sort of notation that I've adopted so far, this will look like the summation of the 'X_i's, because the yeses are ones divided by 'n'. Now, I wrote it like that because I hope that most people recognize that this object, when you sum up independent random variables that have the same distribution identically distributed, then you're going to get through the central limit theorem something on the other side that is normally distributed. So, that's what we're going to use again. Now, as a warm-up, let me just do two calculations. So, first of all, and this is a review from your probability portion when you talked about Bernoulli random variables in the binomial. So, the expected value of each of these responses, I'm going to look at its outcome, and multiply by the probability. So, one outcome is I'm going to get a '1', and it's going to happen as probability 'p'. The other outcome is I'm going to get a '0'. It's going to happen with probability '1' minus 'p'. So, the inexpectation, the number that comes back is a 'p', when I average out. Now, the variance, which is the other thing that is important here, is I'm going to sum up again over all outcomes. So, for each outcome, I'm going to see, how far is the outcome? So, the '1' or the '0' from the mean, which is 'p'. I'm going to square it, and then I'm going to multiply it by the probability. So, I'm going to go '1' minus 'p'. So, this is the outcome, minus the...minus the mean, squared, times the probability that I'm going to see this outcome, plus the outcome '0' minus its own, minus the mean 'p' squared, times its own probability '1' minus 'p'. And that altogether is 'p' times '1' minus 'p'. All right! And now, I'm just going to use the central limit theorem. So, the central limit theorem said that, let's say for 'X-bar', we had that this is going to be normal mu sigma squared over 'n'. All right! So, this was what we had before. Now, just so that we're clear, mu is 'p', and sigma squared is 'p', '1' minus 'p'. All right! So, I'm going to get that 'p' hat is normally distributed. It's centered at 'p', so, it's correct on average. And the variance of that estimator is 'p', '1' minus 'p' over 'n'. I like to work with the standard deviation of 'p' hat because that's the thing that we use in the calculations. So, the standard deviation of 'p' hat, which we also called the standard error of 'p' hat, is the square root of

'p', '1' minus 'p' over 'n'. All right! Now, how large should 'n' be in order for us to be able to use that formula? Well, this is sort of a...a...a crude rule of thumb. So, 'n', 'p' times '1' minus 'p', bigger than '9'. So, you know, let's say if 'p' is a half, I would say 'n' times a half times a half, so, 'n' times a quarter bigger than '9'. So, that would be 'n' bigger than 36. If 'n' is point one, you may need about a hundred samples, so, something...something small. All right! So, we have that...that's 'p' hat. And 'p' hat minus 'p'. So, if I...if I subtract 'p', that would be normal, so, that's the error. It's centered at '0'. And the standard error is this number. That's it. So, polling is the same thing as what we did last time in...in our sampling lecture. And we can use the central limit theorem again to assess the accuracy of the poll.

Video 8 (8:07): Accuracy of a Poll of a Given Sample Size

So, let's do the calculation. So, I want to do the following thing. I have here the distribution of 'p' hat. It's centered at the truce. And let's say my margin of error here was plus or minus '3%' points. So, I would like to figure out what's the probability that I'm either too high, which is this, or too low, which is that. So, I want to figure out that probability, taking into account that I have '1,266' people—individuals in the sample. So, the probability that my error in absolute value exceeds '.03', this is symmetric, so, this is going to be twice the probability that 'p' hat is bigger than 'p' plus '.03'. So, when we have a normal distribution, what we always do is we go, and measure that distance, in standard deviation. So, let's just do that. So, that distance is '.03' percentage points. And, what I need to do is I need to divide by the standard deviation of 'p' hat. So, this is 'p', '1' minus 'p', over '1,266'. So, I'm almost done. Right! I'm almost done. But I have a problem. And the...and the problem that I have is, you know, 'p' is still in this expression. And...and that's a...that's a problem because 'p' is unknown. Right! The only reason I'm doing the poll, is to figure out what that number is. So now, we need to talk about what should we do. So, we need to really guess a reasonable value for 'p'. There are really two guesses that one could suggest. The first one would be to use 'p' hat itself, which was '.58'. So, that's the first thing we could do. So, we look—we...we know the answer is close to '.58'. May not—maybe it's not exactly right, but let's take that number there. That doesn't seem like a...a big error. So, that's...that's one thing we can do. And we'll do—we will use that, not now, but in about 10 minutes. The second thing we can do, we can use a conservative estimate for 'p'. And that would be 'p' equals a half. And why did I say conservative? Well, I said conservative because the variance looks like this. This is as a function of 'p'. So, when 'p' is '0', there is no variance, okay! And when 'p' is '1', there is no variance. And the maximum variance is achieved at '.5'. So, this is really the most random. And this is a case where my standard error gets—is the worst. So, if I could tell you my accuracy, under that very conservative estimate, then the real accuracy would only be better than that. So, what I'm going to do is I'm going to use 'p' equals a half. And, you kind of always use 'p' equals a half before you collect the data because there is nothing else you could have done. And when the poll has several questions, they may differ in terms of their estimated probability speed. There is...there is one small caveat related to that. And...and...and that is, if I was about to go and estimate something, Suppose, I...I'm about to run a...a survey and I don't know exactly what the right value of 'p' hat is, and I say, just use a half–a half is conservative, and this is true. But, in some cases, it may be too conservative. For example, if I was estimating the probability that people are going to click on an ad if...if I show them the—sort of the logo of the firm on the top left

corner versus the top right corner, and I claim that by moving the logo, you're more likely to click on the ad, but that...that number that I'm going to try to estimate is going to be very small. It's going to be of the order of 1%, half a percent. So, half a percent would be around here. So, using this variance estimate instead of this, may be overly conservative. So, you know, sometimes we may want to tweak that rule a little bit. All right! So, that's all I wanted to say about that. So, I'm going to use 'p' equals a half, and I'm going to stick it in here and I'm going to have '.03' over a half, and a half over '1266' root, and that's twice the probability that 'Z' is bigger. So, this number here, if you evaluate it, is about '2.1', '2.13'. And this probability times two is about '.04'. So, what is that? What did we get? We got that the...the probability, therefore, the probability that the error exceeds '.03' is around point—is around '4%'. So, the probability that the error is about '.03' is actually a little bit better than this. Right! So, they're saying you're within plus or minus '3'. '95%' of the time, we computed that to be about '96%'of the time.

Video 9 (6:59): Sample Size Selection for a Poll of a Given Target Accuracy

This is the—this is a calculation you can do after you collected the data. Most of the times, you actually do this calculation before you collect the data. In...in the sense, that you want to go and start asking people questions. But, before you start, you want to know how many people do you need to have in the poll in order for this thing to be pretty reasonable, and you can write a big report and...and you can report anything and...and people will not going to complain about the accuracy. So, I could flip the question, and I could say, "Listen, this is your target accuracy, and this is the probability, and, you know, tell me how big of a sample size you need." So, I can do that similarly to what we did last time. So, I want to be to—'p' plus '.02', 'p' minus '.02', and I told you, let's say that this probability is going to be '.025', and that probability will be the same, all right! Now, in order to have a tail probability of '.025', what we know from our discussion of the normal distribution is that this distance, all right, this distance, which is '.02' need to be equal to '1.96' standard errors of 'p' hat. So, that's the only way you're going to get two and a ½% in the tail. You need to be '1.96' standard deviations away. So,'.02' is equal to '.96', then I have this 'p', one minus 'p' over 'n' formula. I have no idea what 'p' is so, I'm going to use my conservative estimate in this formula. So, I'm going to replace it by a half, times, a half over 'n', and I'm going to solve for 'n'. All right! So, if I solve for 'n', I'm going to get '1.96' squared, all right, times—so, let's see—so, we're going to have this, which is a fourth, right, and now, we're going to have '1' '.02' squared, and that is about '2400' individuals, all right! So, the sample size, for a poll...for a poll—that's an o, is typically '1.96' squared, times a fourth—the fourth coming...coming from this, times '1' over the margin of error squared. So, if instead of '2%', you wanted '1%', right, if I wanted to make the margin of error half the size, I need to square that number. So, I would need—'n' will have to be four times larger, all right! So, every time you...you want to—you're trying to make the your...your margin of error half the size at the same level of probability, at the same probability level, then your...your sample size gets four times bigger. Or, if I wanted to make it ten times more accurate, I would need 100 times, ten squared, the sample size. So, let's summarize what...what we've done. So, we talked about sampling. So, we saw what we do when we estimate the mean, that was the, sort of, number of minutes in front of different media...media...media outlets per day–number of minutes per day, per student. And, this was the proportion example that we discussed. There are two things here, right! First of all, what's your

permissible error? So, this is the plus or minus 15 minutes or the plus or minus point point o three. And then, what's the probability of exceeding that error? Right! So, what's the probability either that you're going to satisfy that tolerance or that you're going to fall outside that? And, we typically want to answer two questions. You give me a sample, and a sample size 'n', and a target accuracy, and I compute this probability of, sort of, falling outside it, or—right, or you tell me how good your poll and your accuracy want to be, and I want probability, and I go, and compute what's the sample size that you need. And, this is sort of typical in all surveying, all sampling, all polling. All right! And, the main tool, as we saw, is really the central limit theorem.

Video 10 (5:14): Confidence Intervals for the Estimated Proportion

Now, I want to finish off by doing one very simple thing. So, when we...we use data, and we got an estimate, let's say that was 'p' hat. And, what I want to do now is, I want to say—make statements like, look 'p' hat, '.58', let's say, which was '58%' of the American adults believe that climate change is mostly human caused, but, you know, we...we don't really believe that it's—the right answer is exactly '58%'. So, maybe what we want to quote to the reader is an interval. And, we can say something like, "I am '95%' confident that the... the truth, the ground truth lies in an interval. So, that's typical and—of how people present that, and the way we do it is we...we're going to use, let's say, our point estimate, he—in...in our case, 58%. And then, I want to put some kind of margin of error, okay! And, what I want to discuss right now is how do we do the margin of error calculation, and then, sort of, end. So, let's see how to do this. So, what do I know? I want to construct a confidence interval, and I want to be, let's say, '95%' confident. So, if I want to be '95%' confident, I want to have two and a ½% probability on each tail. So, I guess what I'm saying here is this is negative '1.96'. This is positive '1.96'. So, to be '95%' confident, which is the same thing as saying, '2.5%' on each tail, right, we need to be '1.96' standard errors away, okay! So, this distance needs to be '1.96'. So, the confidence interval is going to be 'p' hat—so, the truth is going to be within the interval that looks like, '.58' plus or minus '1.96' standard errors, and the standard error is '.58', times '1' minus '.58', over '1266', right! So, this is a standard error of our estimate. So, this is essentially the accuracy of our estimate or the standard deviation of our estimate. So, when I give you a confidence interval, I say, "Look, my best guess is '.58', but then I'm going to add to it '1.96' standard errors." Last thing to talk about here is, if I wanted to be more accurate or less accurate, so if I change the confidence level, then, you know, the point estimate doesn't change, the standard error doesn't change, right, this formula. The only thing that is going to change is this.
The multiplier depends on the confidence level. Now, if I want to be less confident, so let's say, I want to give you a 90% confidence level, or if I want to be less confident, I can quote you a narrower interval, right! So, if I wanted this tail probability to be '.05' and this tail probability to be '.05', my multiplier would have been '1.645'. So, as confidence goes up, so, if you want to be more confident, your multiplier itself goes up, which is intuitive. If you want me to be correct more time, the easiest thing for me to do is just quote you a wider interval, all right! And, I do that by changing the multiplier.