We now revisit the polling problem that we have started earlier. When we first looked at that problem, we used the Chebyshev inequality to obtain certain bounds and numerical results. What we want to do now is instead to use a central limit theorem-type approximation, which we hope that it will be more accurate and more informative.

Let us remind ourselves of the setting. We want to estimate a certain number, p, which is the fraction of the population that will vote yes in a certain referendum. And we estimate p by picking a sample out of the population.

We pick n people. We pick them randomly, uniformly over the population and independently. For each one of the people in the sample, we ask them if they will vote to yes or no, and then we record their answers in Bernoulli random variables, Xi. So by the assumptions that we have made, these Xi's are independent Bernoulli random variables, and their mean is equal to p.

We count how many X's were equal to 1. That's the number of yeses. We divide by n, and that gives us the fraction in the population that have responded yes. This is the sample mean of the X's. And we use this sample mean to estimate the unknown fraction p.

We would like the error in our estimation to be small, that is the difference between the sample mean and the true value p to be small, less, let's say, than one percentage point. Now there's no way of guaranteeing that this spec will be met with certainty, unless we sample almost everyone in the population. But what we can do instead is to ask that these specifications are violated with only a small probability.

So we look at the probability that our estimation error is larger than what we want. This is the case that we do not meet the specs, and we would like this probability to be small. One possible question is what the value of n should be in order to meet the specs. But in order to do any calculations, we first need a way of approximating this probability.

We will do that using the central limit theorem. The central limit theorem involves this standardized version of the random variable Sn, where Sn stands for the sum of the X's. We know that this random variable is approximately normal. And what we want to do now is to take this event and rewrite it in an

equivalent way but which involves this random variable Zn.

Let us start. First, we note that here we a mu and a sigma, so we should know what these are. For a Bernoulli random variable, the mean is what we already wrote down, and sigma is the square root of p times 1 minus p.

Now let's look at this event. Mn is the same as Sn/n, by definition. And we can write p in this form, minus n times p divided by n. And we want this quantity to be larger than or equal to 0.01. So this event here is identical to that event up there.

This starts to look like this expression. p is the same as Mu. But there is a little bit of a difference in the denominator terms. So let's see what we can do.

Let's take this same event but multiply both sides of the inequality by a square root of n. This causes this denominator term to become just square root of n, and we get a square root of n term in the numerator on the other side. This is an equivalent description of the event.

Now we can multiply both sides of this inequality by sigma-- actually the denominators on both sides by sigma-- and we obtain this equivalent representation. But now we notice that here we do have the random variable Zn that we wanted. And so we managed to express this event in terms of the random variable Zn. In particular what we have is that this probability is the same as the probability that the absolute value of Zn is larger than or equal to 0.01 square root of n divided by sigma.

Then we can use the central limit theorem approximation to approximate this probability by the corresponding probability where we now use a standard normal random variable instead of the Zn random variable. So here, Z stands for a standard normal random variable with mean 0 and variance equal to 1.

Let us now continue on a new slide so that we have some working space. And here is the result that we have derived so far. If somebody gives us the value of n, we would like to be able to calculate this probability using this approximation. However, there's a slight difficulty because sigma is a function that depends on p, and it is not known.

However, as we discussed when we first started the polling problem, we do know that sigma is always less than or equal to 1/2. And this suggests that we could use here the worst-case value of the standard

deviation, replace sigma by 1/2 and instead look at this probability here.

How are these two probabilities related? Which direction does the inequality go? A sketch will be useful here.

Z is a standard normal, and it's centered at 0. Somewhere here, we have a value of 0.02 square root n. And somewhere further out, we have the value of 0.01 square root n divided by sigma.

Why are these two values ordered this way? Since sigma is less than 1/2, 1 over sigma is bigger than 2. So this expression here is bigger than this expression there.

Since the inequality goes this way, now we can compare these two events. This event, that Z is larger in absolute value than this number, is the probability of this tail of the distribution. And we will have a similar probability from the other end of the tail of the distribution.

Here we're talking about the probability of being larger than or equal to this number, which would correspond only to this part of the tail and, similarly, a small part of the tail from the other side. The blue event is smaller than the red event. This is the probability of the blue event, so it's going to be no larger than the probability of the red event.

Now if somebody gives us a value of n, we should be able to calculate this probability. How do we calculate it? The probability that the absolute value is above a certain number is equal to the probability of this tail plus the probability of that tail. But because of the symmetry of the normal distribution, this is twice the probability of each one of the tails.

What is the probability of this tail? It's 1 minus the probability of whatever is below that. So it's 1 minus. And the probability of being below that, this is the standard normal CDF evaluated at 0.02 square root n. So we do have now an expression for the desired probability, or at least a bound for it, which is expressed in terms of the standard normal CDF.

If somebody gives you a value of n, you can plug in here. If n is 10,000, then square root of n is 100. And this number becomes equal to 2. And so in this case, what we obtain is that the probability of interest is less than or equal to 2 times 1 minus Phi of 2.

Now we invoke the standard normal table. From the normal table, we obtain that this quantity is equal to twice 1 minus 0.9772, which evaluates to 0.046. So if we use 10,000 people in our sample, then we

will get an accuracy of one percentage point with very high probability. The probability that we do not meet the specification so that the accuracy that we get is worse than one percentage point, that probability is quite small. It's 0.046. That is 4 and something percent.

This is pretty good. And suppose that your boss now tells you, I only want the probability of not meeting the specs to be 5%. You look at this result, and you say, with 10,000, I achieved a probability of a large error that's less than 5%. This means that I probably have some leeway and that I can reduce the size of my sample.

What could the size of the sample be and still meet those specs? What we're trying to do here is that we have this approximation for the probability of interest, and we want to set this probability to a value of 0.05. Then we want to ask, what is the value of n that will result in this particular probability of not meeting the specs?

Now we can do the algebra. And we find that this corresponds to requiring that phi of 0.02 square root n to be equal to 0.975. What's the interpretation of this? We want to choose n so that the probability of the two tails is 5%. This means that we want this probability here to be 2 and 1/2 percent.

This means that the probability of whatever is to the left of this number should be 0.975, including the tail. This means, again, that we have to look at the standard normal table and ask, what's the value for which the CDF is equal to 0.975? So we look around, and we find 0.975 to be here, and it corresponds to 1.96.

This tells us that 0.02 square root n should be equal to 1.96. Then we solve for n, and we find that the value of n is 9,604, which is indeed some reduction from the 10,000 that we had originally.

How does this relate to the real world? When you read newspapers about polls, you will never see sample sizes that are about 10,000. You will usually see sample sizes of the order of 1,000, sometimes even smaller.

How can they do that? Well, they can do that because the specs that they impose are not as tight as the specs that we have here. Usually, they tell you that the results are accurate within three percentage points, let's say, instead of one percentage point. And by moving from 0.01 to 0.03, and if you repeat those calculations, you will find that the sample size of about 1,000 will actually do.