We will now consider a very practical application of the weak law of large numbers, and the calculations associated with it. The application has to do with polling. There's a certain referendum that's going to take place. We're close enough to the day of the referendum so that voters have made up their minds, and there is a fraction p of the population that represents the voters that are going to vote yes. But the referendum has not yet taken place, and you want to find out, to predict or estimate what p actually is.

What you do is that you go ahead, and you select at random a number of people out of the population. And for each person, you record their answer, whether they intend to vote yes, or whether they intend to vote no. When we say that the people are randomly selected, what we mean is that we choose them uniformly from the population. And since there's a fraction p that will vote yes, this means that this random variable will be 1 with probability p, and therefore, the expected value of Xi is equal to p.

In addition, we assume that we select those people independently. Now note that if we select people independently, there's always a chance that the first person polled will be the same as the second person polled, something that we do not really want to happen. However, if we assume that the population is very large, or even idealize the situation by assuming that the population is infinite, then this is never going to happen, and this will not be a concern.

So how do we proceed? We look at the results that we got from the people that we polled. We count how many said yes, divide by n, and this gives us the fraction of yeses in the sample that we have obtained. And this is a pretty reasonable estimate for the unknown fraction p, the fraction of yeses in the overall population.

Now perhaps your boss has asked you to find out the exact value of p. What should your response be? Well, there is no way to calculate p exactly on the basis of a finite and random poll.

Therefore, there is going to be some error in our estimation of p. Then, perhaps your boss comes back and says, OK, then try to give me an estimate of p which is very accurate. I would like you to come up with an estimate which is correct within one percentage point. Can you do this for me?

Your answer might be, OK, let me try polling 10,000 people, and see if I can guarantee for you such a small error. But after you think about the situation a little more, you realize that there is no way of guaranteeing such a small error with certainty. What if your unlucky, and the people that you poll happen to be not representative of the true population?

So you come back to your boss and you say, I cannot guarantee with certainty that the error is going to be small, but perhaps I can guarantee for you that the error that I get is small with high probability. Or alternatively, I'm going to guarantee for you that the probability that we get an error that's bigger than that is very small. So how small is it going to be? Let's try to derive a bound on this probability of an error larger than one percentage point.

Using the calculations that we carried out when we derived the weak law of large numbers, we know that this probability of a large error is bounded above by this quantity. What is this quantity? Sigma squared is the variance of the random variable that we're sampling. And since this is Bernoulli, this variance is p times 1 minus p, and then we divide by n, which in our case is 10 to the fourth times epsilon squared. Epsilon is 10 to the minus 2, so here we have 10 to the minus 4.

OK, but now, what is this quantity? This quantity depends on p, and we do not know what p is. However, if you take this expression, and plot it as a function of p, what you obtain is a plot of this type. And the maximum happens when p is equal to 1/2, in which case we get a value of 1/4.

That is, the variance of the Bernoulli is, at most, 1/4. And therefore, we obtain this bound here where the denominator terms have disappeared because they're equal to 1. So you tell your boss, if I sample 10,000 people, then the probability of an error more than the one percentage point is going to be less than 25%.

At which point, your boss might reply and say, well, a probability of a large error of 25% is too big. This is unacceptable. I would like you to have a probability of error that's less than 5%.

So suppose now that we want to change this, and make it only a 5% error-- 5% or less. How are you going to proceed? Well, you have this quantity here, this upper bound, which we know to be less than or equal to 1/4 divided by n times epsilon squared, which is, in our case, 10 to the minus 4.

We would like this quantity to be less than or equal to 5%, which is 5/10 to the second power. And after you solve this inequality, you find that this is equivalent to taking n larger than or equal to 10 to the sixth. And then the five together with that four gives us a denominator of 20. And this number is equal to 50,000.

So at this point, you can go back to your boss and tell them that one way of guaranteeing that the probability of a large error is less than or equal to 5% is to take n equal to 50,000. So 50,000 will suffice to achieve the desired specs. Notice that the desired specs have two parameters. One is the accuracy that you want, and the other is the confidence with which the accuracy is going to be achieved.

Now 50,000 is a pretty large number. If you notice the results of polls, the way that they are presented in newspapers, they usually tell you that there's an accuracy of plus or minus three percentage points, not one percentage point. That helps things a little.

It means that you can do with a somewhat smaller sample size. And then, there's another effect. Our calculation here was based on this inequality, which is the Chebyshev inequality. But the Chebyshev inequality is not that accurate. It turns out that if we use more accurate estimates of this probability, we will find that actually much smaller values of n will be enough for our purposes.