MITx Video | 6.041x | Spring 2014 | Video transcript | L19-5: Normal approximation to the binomial

An important application of the central limit theorem is in the approximate calculation of the binomial probabilities. Here is what is involved. We start with random variables-- Xi-- that are independent. And they have the same distribution. They're all Bernoulli with parameter p.

We add n of those random variables, and the resulting random variable, Sn, we know that it has a binomial PNF with parameters n and p. We also know its mean, and we do know its variance.

What the central limit theorem tells us, in this case, since we're dealing with the sum of independent identically distributed random variables, is the following. If we take this random variable here that we have been denoting by Zn, which is a standardized version of Sn-- we subtract the mean of Sn and divide by the standard deviation-- this random variable has a CDF that approaches as n goes to infinity, the CDF of a standard normal.

So let us use what we now know to calculate some probabilities. Let us fix some parameters. n is 36. p is 0.5. And we wish to calculate the probability that Sn is less than or equal to 21.

Now, in this case, we can calculate it exactly using the binomial formula. The probability of being less than or equal to 21 is the sum of the probabilities of all the numbers from 0 to 21. And this is the probability of obtaining a number k. And by calculating this expression, we obtain this number, which is the exact answer.

Now, let us proceed using the central limit theorem. We are interested in this probability, but we will use the fact about the CDF of this related random variable. So the first step is to calculate n times p, which is 18. The second step is to calculate this denominator here, which in our case evaluates to 3.

Now, since we know something about the CDF of this random variable, what we need to do is to take this event and rewrite it in terms of this random variable. So we have the event of interest, which is that Sn is less than or equal to 21. This is the same as the event that Sn minus 18 is less than or equal to 21 minus 18. And it's the same as this event here, where we divide both sides by 3.

Now, what we have here is the probability that this random variable Zn is less than or equal to 1. But now, Zn is approximately a standard normal, so we can use here the CDF of the standard normal distribution, which is Phi of 1. And at this point, we look at the tables for the normal distribution. We'll find this entry here. And this gives us an answer of 0.8413.

This is a pretty good approximation of the exact answer, which is 0.8785. But it is not a great approximation. It is off by about four percentage points. Can we do better than that? It turns out that we can get a better approximation. And let us see how this can be done.

Recall that we approximated this probability using the central limit theorem and found this numerical value. But we make an observation that this probability is equal to this probability here. Why is that? Sn is an integer random variable. Therefore, if I tell you that it is strictly less than 22, I'm also telling you that it is 21 or less.

Therefore, this event here is the same as that event here. And therefore, their probabilities are the same. So instead of using the central limit approximation to calculate this probability, let us follow the same procedure but try to calculate this probability here.

And this probability here is equal to the probability that Sn minus-- we subtract the mean, divide by the standard deviation of Sn-- is strictly less than 22 minus 18 divided by 3, which is the probability that the random variable that we denote by Zn, which is this expression here, is strictly less than 22 minus 18 over 3. And this is 1.33.

Now, at this point, we pretend that Zn is a standard normal random variable-- the probability that the standard normal is less than a certain number. This is the standard normal CDF evaluated at that number. And then we look up at the normal tables at 1.33 and we find this value of 0.9082.

Now, we compare this value with the exact answer for this problem. And we see that we again missed it. Using this approximation to this quantity gave us an underestimate of this number. Now, we obtained an overestimate. The true value is somewhere in the middle. So this suggests that we may want to do something that combines these two alternative choices here.

But before doing that, it's good to understand what exactly have we be doing all along. What we're doing is the following. We have the PMF of the binomial centered at 18, which is the mean. It's a discrete random variable. But when we use the central limit theorem, we pretend that the binomial is normal, but while we keep the same mean and variance.

Now, when we calculate probabilities, if we want to find the discrete probability that Sn is less than or equal to 21, which is the sum of these probabilities, what we do is we look at the area under the normal PDF from 21 and below. In the alternative approach, when we use the central limit theorem to approximate the probability of this event, we go to 22, and we look at the event of falling below 22. This means that we're looking at the area from 22 and lower.

So in one approach, this particular region is not used in the calculation. That's what we did here. But in the second approach, it was used in the calculation. Should it be used or not?

It makes more sense to use only part of this solid region and assign it to the calculation of the probability of being at 21 or less. Namely, we can take the mid point here, where the mid point is at 21.5, and calculate the area under the normal PDF only going up to 21.5.

What this amounts to is looking at this particular event here. Now, this event is, of course, identical to this event that we have been considering, because again, Sn is a discrete random variable that takes integer values. But when we approximate it by a normal, it does make a difference whether we write the event this way or that way.

So here, we're going to obtain the probability that the standardized version of Zn is less than. We follow the same calculation, but now we have 21.5 minus 18 divided by 3. And this number here is 1.17. And using the central limit theorem calculation, this is the CDF of the standard normal evaluated at 1.17, which we can go and look up in the normal table to find the value of 0.8790.

And now, we notice that this value is remarkably close to the true value. It is much better as an approximation that what we obtained using either this choice or that choice. And since this approximation is so good, we may consider even using it to approximate individual probabilities of the binomial PMF. Let's see what that takes.

Let us try to approximate, as an example, the probability that Sn takes a value of exactly 19. So what we will do will be to write the event that Sn is equal to 19 as the event that Sn lies between 18.5 and 19.5. In terms of the picture that we were discussing before, what we are doing, essentially, is to take the area under the normal PDF that extends from 18.5 to 19.5 and declare that this area corresponds to the discrete event that our binomial random variable takes a value of 19.

Similarly, if we wanted to calculate approximately the value of the probability that Sn takes a value of

21, we would consider the area under the normal PDF from 20.5 to 21.5. So let us now continue with this approach. We do the usual calculations, which is to express this event in terms of standardized values. That is, we subtract throughout the mean of Sn and divide by standard deviation.

So what we obtain here is the standardized version of Sn. And that has to be, now, less than or equal to 19.5 minus 18 divided by 3, which is the probability that our standardized random variable lies between 0.17 and 0.5.

And now, if we pretend that Zn is a standard normal random variable, which is what the central limit theorem suggests, this is going to be equal to the probability that the standard normal is less than or equal to 0.5 minus the probability that it is less than 0.17. And if we look up those entries in the normal tables, what we find is an answer of 0.6915 minus this number, which evaluates to 0.124.

And what is the exact answer if we were to use the binomial probability formulas? The exact answer is remarkably close to what we obtained in our approximation. This example illustrates a more general fact that this approach of calculating individual entries of the binomial PMF gives very accurate answers. And in fact, there are theorems, there are theoretical results to this effect, that tell us that this way of approximating-- asymptotically, as n goes to infinity and in a certain regime-- does give us very accurate approximations.