CIS 2033, Spring 2017[1]

Instructor: David Dobor

April 17, 2017

[1] Supplemental reading: Bertsekas Textbook Chapter 7.

An important application of the central limit theorem is in the approximate calculation of the binomial probabilities.

Normal approximation to the binomial

Here is what is involved. We start with random variables X_i that are independent. Each of the X_is has the same distribution: they're all Bernoulli with parameter p.

We add n of these Bernoulli random variables. The resulting random variable, $S_n = X_1 + \ldots + X_n$, has a binomial PMF with parameters n and p. We also know its mean, it is np, and we know its variance, it is $np(1-p)$.

- X_i: independent, Bernoulli(p); $\quad 0 < p < 1$
- $S_n = X_1 + \cdots + X_n$: Binomial(n, p)
  - mean np, variance $np(1-p)$

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 the following random variable, which we denote by Z_n,

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}, \tag{1}$$

which is a standardized version of S_n (we subtract the mean of S_n and divide by its standard deviation), then the random variable Z_n has a CDF that approaches the CDF of a standard normal, as n goes to infinity.

- CDF of $\dfrac{S_n - np}{\sqrt{np(1-p)}} \longrightarrow$ standard normal

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 $S_n \leq 21$.

Now, in this case, we can calculate it exactly using the binomial formula:

$$\sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} = 0.8785$$

Now, let us proceed using the central limit theorem. We are interested in the probability $\mathbf{P}(S_n \leq 21)$, but we will use the fact about the CDF of the related random variable Z_n.

(1) The first step is to calculate np, which is 18.

(2) The second step is to calculate this denominator in (1), which evaluates to $\sqrt{36 \cdot 0.5 \cdot (1 - 0.5)} = 3$.

(3) The next step is to take the event of interest, $S_n \leq 21$, and rewrite it in terms of the random variable Z_n, which according to the central limit theorem is approximately standard normal:

$$\begin{aligned} \mathbf{P}(S_n \leq 21) &= \mathbf{P}\Big(\frac{S_n - 18}{3} \leq \frac{21 - 18}{3} \Big) \\ &= \mathbf{P}(Z_n \leq 1) \approx \Phi(1) = 0.8413. \end{aligned}$$

Here we looked at the tables for the normal distribution and find the answer to be 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. Let us see how this can be done.

	.00	.01	.02	.03	.04
0.0	.5000	.5040	.5080	.5120	.5160
0.1	.5398	.5438	.5478	.5517	.5557
0.2	.5793	.5832	.5871	.5910	.5948
0.3	.6179	.6217	.6255	.6293	.6331
0.4	.6554	.6591	.6628	.6664	.6700
0.5	.6915	.6950	.6985	.7019	.7054
0.6	.7257	.7291	.7324	.7357	.7389
0.7	.7580	.7611	.7642	.7673	.7704
0.8	.7881	.7910	.7939	.7967	.7995
0.9	.8159	.8186	.8212	.8238	.8264
1.0	.8413	.8438	.8461	.8485	.8508
1.1	.8643	.8665	.8686	.8708	.8729

Figure 1: A fragment of the standard normal table. The circled quantity is used in the calculations shown on the left.

The 1/2 correction for integer random variables

We just approximated the probability $\mathbf{P}(S_n \leq 21)$ and found its approximate value to be ≈ 0.8413. We now make an observation that

$$\mathbf{P}(S_n \leq 21) = \mathbf{P}(S_n < 22), \qquad \text{because } S_n \text{ is integer}$$

Why is that? S_n 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, meaning that the events $S_n \leq 21$ and $S_n < 22$ are one and the same event. It must be, then, that the probabilities of these events are equal.

Now, instead of using the central limit approximation to calculate the probability $\mathbf{P}(S_n \leq 21)$, as we did in the previous section, let us follow the same procedure but try to calculate the $\mathbf{P}(S_n < 22)$ probability instead:

$$\begin{aligned}\mathbf{P}(S_n < 21) &= \mathbf{P}\Big(\frac{S_n - 18}{3} < \frac{22-18}{3}\Big) \\ &= \mathbf{P}(Z_n < 1.33) \approx \Phi(1.33) = 0.9082\end{aligned}$$

Now, we compare the value we just obtained, 0.9082, with the exact answer for this problem, which is 0.8785. We see that we again missed it.

Using the approximation of $\mathbf{P}(S_n \leq 21)$ to the true quantity 0.8785 gave us an underestimate of it; using the approximation of $\mathbf{P}(S_n < 22)$ to the true quantity 0.8785, we obtained an overestimate.

The true value is somewhere in the middle. This suggests that we may want to do something that combines these two alternative choices.

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 that has the same mean and variance, as shown in Figure 3.

When we calculate probabilities, if we want to find the discrete probability that $S_n \leq 21$, which is the sum of the heights of the bars to the left of, and including, 21 in the PMF shown in Figure 4, we look at the area under the normal PDF from 21 and below. It is shaded in red in Figure 4.

In the alternative approach, when we use the central limit theorem to approximate the probability of the event $S_n < 22$, we look at the

	.00	.01	.02	.03	.04	.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
0.1	.5398	.5438	.5478	.5517	.5557	.5596
0.2	.5793	.5832	.5871	.5910	.5948	.5987
0.3	.6179	.6217	.6255	.6293	.6331	.6368
0.4	.6554	.6591	.6628	.6664	.6700	.6736
0.5	.6915	.6950	.6985	.7019	.7054	.7088
0.6	.7257	.7291	.7324	.7357	.7389	.7422
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749
1.2	.8849	.8869	.8888	.8907	.8925	.8944
1.3	.9032	.9049	.9066	.9082	.9099	.9115
1.4	.9192	.9207	.9222	.9236	.9251	.9265

Figure 2: A fragment of the standard normal table. The circled quantity is used in the calculations shown on the left.

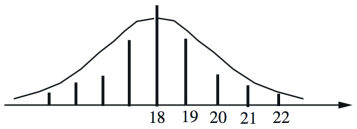

Figure 3: The binomial PMF and the normal PDF.

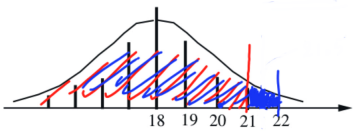

Figure 4: The red shaded region is to the left of 21, the blue shaded region is to the left of 22.

event of falling below 22. This means that we're looking at the blue shaded area from 22 and lower in Figure 4.

So in one approach, the particular region between 21 and 22 is not used in the calculation; in the second approach, it is used in the calculation. Should it be used or not?

It makes more sense to use only part of the solid blue region shown if Figure 5 and assign it to the calculation of the probability of being at 21 or less. Namely, we can take the mid point, 21.5, and calculate the area under the normal PDF only going up to 21.5.

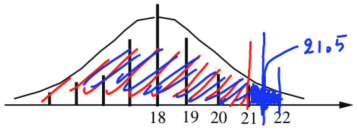

Figure 5: Only half of the solid blue region between 21 and 22 will be used when applying the 1/2 correction to our previous calculations.

What this amounts to is looking at the event $S_n \leq 21.5$. Now, this event is, of course, identical to both of the events $S_n \leq 21$ and $S_n < 22$, because again, S_n 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 one way or another.

Now,

$$\begin{aligned}\mathbf{P}(S_n \leq 21.5) &= \mathbf{P}\left(Z_n \leq \frac{21.5 - 18}{3} \right) \\ &\approx \Phi(1.17) = 0.8790\end{aligned}$$

Here 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 of the other two choices. 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.

	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292

Figure 6: A fragment of the standard normal table. The blue circled quantity is used in the calculations shown on the left.

Approximating individual probabilities of the binomial PMF: De Moivre - Laplace Approximation

Let us try to approximate, as an example, the probability that S_n takes a value of exactly 19. We will write the event that S_n is equal to 19 as the event that S_n lies between 18.5 and 19.5.

$$\begin{aligned}
\mathbf{P}(S_n = 19) &= \mathbf{P}(18.5 \leq S_n \leq 19.5) \\
&= \mathbf{P}\Big(\frac{18.5 - 18}{3} \leq Z_n \leq \frac{19.5 - 18}{3} \Big) \\
&= \mathbf{P}(0.17 \leq Z_n \leq 0.5) \\
&\approx \Phi(0.5) - \Phi(0.17) \\
&= 0.6915 - 0.5675 = 0.124
\end{aligned}$$

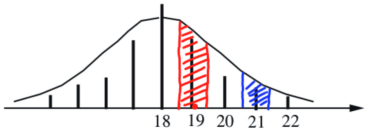

Figure 7: The red shaded region between 18.5 and 19.5 under the normal curve approximates the probability that the binomial random variable takes a value of 19. Similarly, the blue shaded region between 20.5 and 21.5 approximates the probability that the binomial random variable takes a value of 21.

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 S_n takes a value of 21, we would consider the area under the normal PDF from 20.5 to 21.5.

The exact answer,

$$0.1251 = \binom{36}{19} \left(\frac{1}{2}\right)^{36},$$

that can be obtained by using the binomial probability formula, 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. In fact, there are theoretical results 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.