CHAPTER 9

Normal approximation to the binomial

A special case of the *central limit theorem* is the following statement.

Theorem 9.1 (Normal approximation to the binomial distribution)

If S_n is a binomial variable with parameters n and p, Binom (n, p), then

$$\mathbb{P}\left(a \leqslant \frac{S_n - np}{\sqrt{np(1-p)}} \leqslant b\right) \xrightarrow[n\to\infty]{} \mathbb{P}(a \leqslant Z \leqslant b),$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0,1)$.

This approximation is good if $np(1-p) \geqslant 10$ and gets better the larger this quantity gets. This means that if either p or $1-p$ is small, then this is valid for large n. Recall that by Proposition 6.1 np is the same as $\mathbb{E}S_n$ and $np(1-p)$ is the same as $\operatorname{Var} S_n$. So the ratio is equal to $(S_n - \mathbb{E}S_n)/\sqrt{\operatorname{Var} S_n}$, and this ratio has mean 0 and variance 1, the same as a standard $\mathcal{N}(0,1)$.

Note that here p stays fixed as $n \to \infty$, unlike in the case of the Poisson approximation, as we described in Proposition 6.3.

Sketch of the proof. This is usually not covered in this course, so we only explain one (of many) ways to show why this holds. We would like to compare the distribution of S_n with the distribution of the normal variable $X \sim \mathcal{N}\left(np, \sqrt{np\,(1-p)}\right)$. The random variable X has the density

$$\frac{1}{\sqrt{2\pi np\,(1-p)}} e^{-\frac{(x-np)^2}{2np(1-p)}}.$$

The idea behind this proof is that we are interested in approximating the binomial distribution by the normal distribution in the region where the binomial distribution differs significantly from zero, that is, in the region around the mean np. We consider $\mathbb{P}\left(S_n = k\right)$, and we assume that k does not deviate too much from np. We measure deviations by some small number of standard deviations, which is $\sqrt{np\,(1-p)}$. Therefore we see that $k - np$ should be of order $\sqrt{n}$. This is not much of a restriction since once k deviates from np by many standard deviations, $\mathbb{P}\left(S_n = k\right)$ becomes very small and can be approximated by zero. In what follows we assume that k and $n-k$ of order n.

We use Stirling's formula is the following form

$$m! \sim \sqrt{2\pi m} e^{-m} m^m,$$

where by $\sim$ we mean that the two quantities are asymptotically equal, that is,their ratio tends to 1 as $m \to \infty$. Then for large n, k and $n-k$

$$\begin{aligned}
&\mathbb{P}\left(S_n = k\right) = \frac{n!}{k!\left(n-k\right)!}p^k\left(1-p\right)^{n-k} \\
&\sim \frac{\sqrt{2\pi n}e^{-n}n^n}{\sqrt{2\pi k}e^{-k}k^k\sqrt{2\pi\left(n-k\right)}e^{-(n-k)}\left(n-k\right)^{n-k}}p^k\left(1-p\right)^{n-k} \\
&= \left(\frac{p}{k}\right)^k\left(\frac{1-p}{n-k}\right)^{n-k}n^n\sqrt{\frac{n}{2\pi k\left(n-k\right)}} = \left(\frac{np}{k}\right)^k\left(\frac{n\left(1-p\right)}{n-k}\right)^{n-k}\sqrt{\frac{n}{2\pi k\left(n-k\right)}}.
\end{aligned}$$

Now we can use identities

$$\begin{aligned}
\ln\left(\frac{np}{k}\right) &= -\ln\left(1+\frac{k-np}{np}\right), \\
\ln\left(\frac{n\left(1-p\right)}{n-k}\right) &= -\ln\left(1-\frac{k-np}{n\left(1-p\right)}\right).
\end{aligned}$$

Then we can use $\ln\left(1+y\right) \sim y - \frac{y^2}{2} + \frac{y^3}{3}, y \to 0$ to see that

$$\begin{aligned}
&\ln\left(\left(\frac{np}{k}\right)^k\left(\frac{n\left(1-p\right)}{n-k}\right)^{n-k}\right) = k\ln\left(\frac{np}{k}\right) + \left(n-k\right)\ln\left(\frac{n\left(1-p\right)}{n-k}\right) \\
&\sim k\left(-\frac{k-np}{np} + \frac{1}{2}\left(\frac{k-np}{np}\right)^2 - \frac{1}{3}\left(\frac{k-np}{np}\right)^3\right) \\
&+ \left(n-k\right)\left(\frac{k-np}{n\left(1-p\right)} + \frac{1}{2}\left(\frac{k-np}{n\left(1-p\right)}\right)^2 + \frac{1}{3}\left(\frac{k-np}{n\left(1-p\right)}\right)^3\right) \\
&\sim -\frac{\left(k-np\right)^2}{2np\left(1-p\right)}.
\end{aligned}$$

Thus

$$\left(\frac{np}{k}\right)^k\left(\frac{n\left(1-p\right)}{n-k}\right)^{n-k} \sim e^{-\frac{(k-np)^2}{2np(1-p)}}.$$

Now we use our assumption that $k-np$ should be of order $\sqrt{n}$ to see that

$$\begin{aligned}
&k-np \approx \sqrt{n}, \\
&n-k \approx n\left(1-p\right) - \sqrt{n}, \\
&k\left(n-k\right) \approx n^2p\left(1-p\right),
\end{aligned}$$

so

$$\sqrt{\frac{n}{2\pi k\,(n-k)}} \sim \frac{1}{\sqrt{2\pi np\,(1-p)}}.$$

□

Example 9.1. Suppose a fair coin is tossed 100 times. What is the probability there will be more than 60 heads?

Solution: $np = 50$ and $\sqrt{np(1-p)} = 5$. We have

$$\mathbb{P}(S_n \geqslant 60) = \mathbb{P}((S_n - 50)/5 \geqslant 2) \approx \mathbb{P}(Z \geqslant 2) \approx 0.0228.$$

Example 9.2. Suppose a die is rolled 180 times. What is the probability a 3 will be showing more than 50 times?

Solution: Here $p = \frac{1}{6}$, so $np = 30$ and $\sqrt{np(1-p)} = 5$. Then $\mathbb{P}(S_n > 50) \approx \mathbb{P}(Z > 4)$, which is less than $e^{-4^2/2}$.

Example 9.3. Suppose a drug is supposed to be 75% effective. It is tested on 100 people. What is the probability more than 70 people will be helped?

Solution: Here S_n is the number of successes, $n = 100$, and $p = 0.75$. We have

$$\begin{aligned}\mathbb{P}(S_n \geqslant 70) &= \mathbb{P}((S_n - 75)/\sqrt{300/16} \geqslant -1.154)\\ &\approx \mathbb{P}(Z \geqslant -1.154) \approx 0.87.\end{aligned}$$

(The last figure came from a table.)

When $b - a$ is small, there is a correction that makes things more accurate, namely replace a by $a - \frac{1}{2}$ and b by $b + \frac{1}{2}$. This correction never hurts and is sometime necessary. For example, in tossing a coin 100 times, there is positive probability that there are exactly 50 heads, while without the correction, the answer given by the normal approximation would be 0.

Example 9.4. We toss a coin 100 times. What is the probability of getting 49, 50, or 51 heads?

Solution: We write $\mathbb{P}(49 \leqslant S_n \leqslant 51) = \mathbb{P}(48.5 \leqslant S_n \leqslant 51.5)$ and then continue as above.

In this case we again have

$$
\begin{aligned}
&p = 0.5,\\
&\mu = np = 50,\\
&\sigma^2 = np(1-p) = 25,\\
&\sigma = \sqrt{np(1-p)} = 5.
\end{aligned}
$$

The normal approximation can be done in three different ways:

$$\mathbb{P}(49 \leqslant S_n \leqslant 51) \approx \mathbb{P}(49 \leqslant 50 + 5Z \leqslant 51) = \Phi(0.2) - \Phi(-0.2) = 2\Phi(0.2) - 1 \approx 0.15852$$

or

$$\mathbb{P}(48 < S_n < 52) \approx \mathbb{P}(48 < 50 + 5Z < 52) = \Phi(0.4) - \Phi(-0.4) = 2\Phi(0.4) - 1 \approx 0.31084$$

or

$$\mathbb{P}(48.5 < S_n < 51.5) \approx \mathbb{P}(48.5 < 50 + 5Z < 51.5) = \Phi(0.3) - \Phi(-0.3) = 2\Phi(0.3) - 1 \approx 0.23582$$

Here all three answers are approximate, and the third one, **0.23582**, is the most accurate among these three. We also can compute the precise answer using the binomial formula:

$$
\begin{aligned}
\mathbb{P}(49 \leqslant S_n \leqslant 51) &= \sum_{k=49}^{51} \tbinom{100}{k} \left(\frac{1}{2}\right)^{100} = \frac{37339688790147532337148742857}{158456325028528675187087900672}\\
&\approx 0.2356465655973331958...
\end{aligned}
$$

In addition we can obtain the following normal approximations

$$\mathbb{P}(S_n = 49) \approx \mathbb{P}(48.5 \leqslant 50 + 5Z \leqslant 49.5) = \Phi(-0.1) - \Phi(-0.3) = \Phi(0.3) - \Phi(0.1) \approx 0.07808$$

$$\mathbb{P}(S_n = 50) \approx \mathbb{P}(49.5 \leqslant 50 + 5Z \leqslant 50.5) = \Phi(0.1) - \Phi(-0.1) = 2\Phi(0.1) - 1 \approx 0.07966$$

$$\mathbb{P}(S_n = 51) \approx \mathbb{P}(50.5 \leqslant 50 + 5Z \leqslant 51.5) = \Phi(0.3) - \Phi(0.1) \approx 0.07808$$

Finally, notice that

$$0.07808 + 0.07966 + 0.07808 = 0.23582$$

which is the approximate value for $\mathbb{P}(49 \leqslant S_n \leqslant 51) \approx \mathbb{P}(48.5 < 50 + 5Z < 51.5)$.

Continuity correction

If a continuous distribution such as the normal distribution is used to approximate a discrete one such as the binomial distribution, a *continuity correction* should be used.

For example, if X is a binomial random variable that represents the number of successes in n independent trials with the probability of success in any trial p, and Y is a normal random variable with the same mean and the same variance as X. Then for any integer k we have that $\mathbb{P}(X \leqslant k)$ is well approximated by $\mathbb{P}(Y \leqslant k)$ if $np(1-p)$ is not too small. It is better approximated by $\mathbb{P}(Y \leqslant k + 1/2)$ as explained at the end of this section. The role of $1/2$ is clear if we start by looking at the normal distribution first, and seeing how we use it to approximate the binomial distribution.

The fact that this approximation is better based on a couple of considerations. One is that a discrete random variable can only take on only discrete values such as integers, while a continuous random variable used to approximate it can take on any values within an interval around these specified values. Hence, when using the normal distribution to approximate the binomial, more accurate approximations are likely to be obtained if a continuity correction is used.

The second reason is that a continuous distribution such as the normal, the probability of taking on a particular value of a random variable is zero. On the other hand, when the normal approximation is used to approximate a discrete distribution, a continuity correction can be employed so that we can approximate the probability of a specific value of the discrete distribution.

For example, if we want to approximate $\mathbb{P}(3 \leqslant X \leqslant 5) = \mathbb{P}(X = 3 \text{ or } X = 4 \text{ or } X = 5)$ by a normal distribution, it would be a *bad* approximation to use $\mathbb{P}(Y = 3 \text{ or } Y = 4 \text{ or } Y = 5)$ as the probability of Y taking on 3, 4 and 5 is 0. We can use *continuity correction* to see that

$$\mathbb{P}(3 \leqslant X \leqslant 5) = \mathbb{P}(2.5 \leqslant X \leqslant 5.5)$$

and *then* use the normal approximation by $\mathbb{P}(2.5 \leqslant Y \leqslant 5.5)$.

Below is a table on how to use the continuity correction for normal approximation to a binomial.

Binomial	Normal
If $\mathbb{P}(X = n)$	use $\mathbb{P}(n - 0.5 < X < n + 0.5)$
If $\mathbb{P}(X > n)$	use $\mathbb{P}(X > n + 0.5)$
If $\mathbb{P}(X \leqslant n)$	use $\mathbb{P}(X < n + 0.5)$
If $\mathbb{P}(X < n)$	use $\mathbb{P}(X < n - 0.5)$
If $\mathbb{P}(X \geqslant n)$	use $\mathbb{P}(X > n - 0.5)$

9.1. Exercises

Exercise 9.1. Suppose that we roll 2 dice 180 times. Let E be the event that we roll two fives no more than once.

(a) Find the exact probability of E.
(b) Approximate $\mathbb{P}(E)$ using the normal distribution.
(c) Approximate $\mathbb{P}(E)$ using the Poisson distribution.

Exercise 9.2. About 10% of the population is left-handed. Use the normal distribution to approximate the probability that in a class of 150 students,

(a) at least 25 of them are left-handed.
(b) between 15 and 20 are left-handed.

Exercise 9.3. A teacher purchases a box with 50 markers of colors selected at random. The probability that marker is black is 0.6, independent of all other markers. Knowing that the probability of there being more than N black markers is greater than 0.2 and the probability of there being more than N + 1 black markers is less than 0.2, use the normal approximation to calculate N.

9.2. Selected solutions

Solution to Exercise 9.1(A): The probability of rolling two fives in a particular roll is $\frac{1}{36}$, so the probability that we roll two fives no more than once in 180 rolls is

$$p = \begin{pmatrix} 180 \\ \\ 0 \end{pmatrix} \left(\frac{35}{36}\right)^{180} + \begin{pmatrix} 180 \\ \\ 1 \end{pmatrix} \left(\frac{1}{36}\right) \left(\frac{35}{36}\right)^{179} \approx .0386.$$

Solution to Exercise 9.1(B): we are interested in the number of successes to be 0 or 1, that is, $\mathbb{P}\left(0 \leqslant S_{180} \leqslant 1\right)$. Since the binomial is integer-valued, we apply the continuity correction and calculate $\mathbb{P}\left(-0.5 \leqslant S_{180} \leqslant 1.5\right)$ instead. We find that the expected value is $\mu = 180 \cdot p = 5$ and the standard deviation is $\sigma = \sqrt{180p(1-p)} \approx 2.205$. Now, as always, we convert this question to a question about the standard normal random variable Z,

$$\begin{aligned} \mathbb{P}\left(-0.5 \leqslant S_{180} \leqslant 1.5\right) = \mathbb{P}\left(\frac{-0.5-5}{2.205} \leqslant Z \leqslant \frac{1.5-5}{2.205}\right) &= \mathbb{P}\left(-2.49 < Z < -1.59\right) \\ &= \left(1 - \Phi\left(1.59\right)\right) - \left(1 - \Phi\left(2.49\right)\right) \\ &= \left(1 - 0.9441\right) - \left(1 - 0.9936\right) = 0.0495. \end{aligned}$$

Solution to Exercise 9.1(C): We use $\lambda = np = 5$ (note that we found this already in (b)!). Now we see that

$$\mathbb{P}(E) \approx e^{-5}\frac{5^0}{0!} + e^{-5}\frac{5^1}{1!} \approx 0.0404.$$

Solution to Exercise 9.2: Let X denote the number of left-handed students in the class. We use Theorem 9.1 with $X \sim \operatorname{Binom}(150, 0.1)$ below. Note that $np = 15$.

(a) $\mathbb{P}(X \geqslant 25) = \mathbb{P}\left(\frac{X-15}{\sqrt{13.5}} \geqslant \frac{10}{\sqrt{13.5}}\right) \approx 1 - \Phi(2.72) \approx 0.00364$. Note that in this approximation we implicitly use that the tail of this probability distribution is small, and so instead of a two-sided interval we just used a one-sided interval.

We can see that the result is really close to the two-sided estimates as follows.

$$\begin{aligned} &\mathbb{P}(150 \geqslant X \geqslant 25) = \mathbb{P}\left(\frac{135}{\sqrt{13.5}} \geqslant \frac{X-15}{\sqrt{13.5}} \geqslant \frac{10}{\sqrt{13.5}}\right) \\ &\approx \Phi(36.74) - \Phi(2.72) \approx 0.00364. \end{aligned}$$

Finally, with the *continuity correction* the solution is

$$\begin{aligned} &\mathbb{P}(150 \geqslant X \geqslant 25) = \mathbb{P}(150.5 \geqslant X \geqslant 24.5) \\ &= \mathbb{P}\left(\frac{135.5}{\sqrt{13.5}} \geqslant \frac{X-15}{\sqrt{13.5}} \geqslant \frac{9.5}{\sqrt{13.5}}\right) \approx \Phi(36.87) - \Phi(2.59) \approx 0.00480. \end{aligned}$$

(2) Similarly to the first part

$$\begin{aligned}\mathbb{P}(15 \leqslant X \leqslant 20) &= \mathbb{P}(14.5 < X < 20.5) \\ &= \Phi\left(\frac{5.5}{\sqrt{13.5}}\right) - \Phi\left(\frac{-0.5}{\sqrt{13.5}}\right) \approx \Phi(1.5) - 1 + \Phi(0.14) \approx 0.4889.\end{aligned}$$

Solution to Exercise 9.3: Let X denote the number of black markers. Since $X \sim \operatorname{Binom}(50, 0.6)$ we have

$$\mathbb{P}(X > N) \approx 1 - \Phi\left(\frac{N-30}{2\sqrt{3}}\right) > 0.2 \text{ and } \mathbb{P}(X > N+1) \approx 1 - \Phi\left(\frac{N-29}{2\sqrt{3}}\right) < 0.2.$$

From this we deduce that $N \leqslant 32.909$ and $N \geqslant 31.944$ so that $N = 32$.