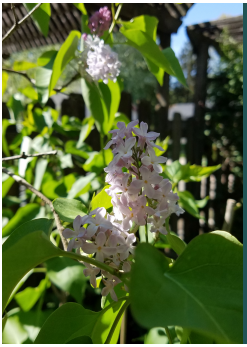


Module C: Understanding the Data-Generating Process

Slide Deck C10:

Approximating the Binomial

The section in which we learn about how we can use a continuous distribution to approximate a discrete one. Here, we are also given a peek into the Central Limit Theorem (CLT) and why it is so important to statistical inference.

Today's Objectives

By the end of this slidedeck, you should

- describe the Binomial distribution in detail
- describe the Normal distribution in detail
- explain how the Normal distribution can be used to approximate the Binomial distribution
- calculate approximate Binomial probabilities using the Normal distribution
- calculate exact Binomial probabilities
- explain why approximating the Binomial is useful, *even if* we have a computer to calculate those probabilities exactly

The Meaning of Approximating Distributions

Since this slidedeck covers how we can approximate the Binomial distribution using the Normal, we need to start by asking one important question:

- What does it mean for one distribution to "approximate" a different distribution?

The answer is:

- the cumulative probabilities are close

That is, if X_1 and X_2 are from different distributions that are "approximately" the same, then for all values of x:

$$\mathbb{P}[\ X_1 \leq x\ ] \approx \mathbb{P}[\ X_2 \leq x\ ]$$

Of course, there is a lot of "statistical" detail hidden in the "$\approx$" sign. How close is "close enough"? That is a question of precision. It is left up to the scientist for what precision is needed. Different applications have different needs.

Approximating the Binomial with the Normal

In this slidedeck, we will be approximating the Binomial distribution with the Normal distribution. To begin, let

$$X_1 \sim \mathcal{B}in(n, p)$$
$$X_2 \sim \mathcal{N}(\mu, \sigma^2)$$

To achieve our goals, we would like to determine some rules on n, p, μ, and σ^2 to ensure that

$$\mathbb{P}[\ X_1 \leq x\ ] \approx \mathbb{P}[\ X_2 \leq x\ ]$$

Surprisingly: It is not as difficult as it appears to get a good "first-order approximation," as the following illustrates. The first step is to review what we know about the two distributions.

The Binomial Distribution

What do we know about the Binomial distribution? What is its statistical definition?

If

$$X_1 \sim \mathcal{B}in(n, p)$$

then:

1. the number of trials, n, is known
2. each trial has two possible outcomes, 'success' and 'not success'
3. the success probability, p, does not change from trial to trial
4. the trials are independent
5. the random variable is the number of success in those n trials

The Binomial Distribution

This definition of the Binomial distribution leads to

- $\mathbb{E}\left[X_1\right] = np$
- $\mathbb{V}\left[X_1\right] = np(1-p)$

From this, it is natural to use these parameters to define the approximate Normal distribution.

Let us see how well

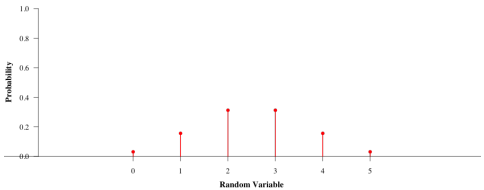
$$X_2 \sim \mathcal{N}\big(\mu = np;\ \sigma^2 = np(1-p)\big)$$

works as an approximation.

We will do this by comparing the cumulative distribution functions.

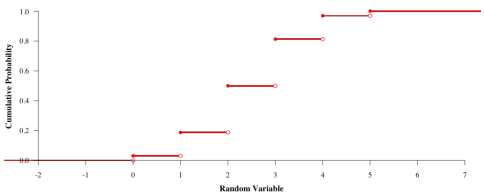
The Binomial pmf

Here is a graphic of the probability mass function (pmf) of the $Bin(5,\ 0.500)$ in red. The pmf provides probabilities for a discrete distribution.



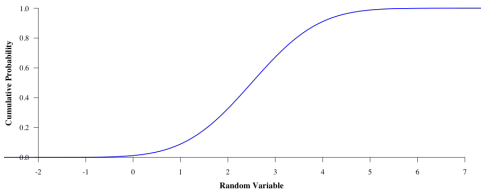
The Binomial CDF

Here is a graphic of the CDF of the $Bin(5,\ 0.500)$ in red and of the $\mathcal{N}(2.50, 1.25)$ in blue. This will help us see how well the Normal can approximate the Binomial.



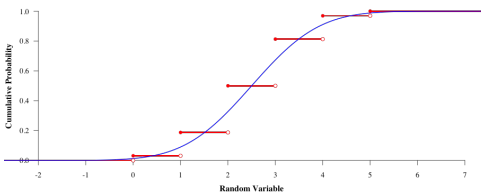
The Normal CDF

Here is a graphic of the CDF of the $\mathcal{B}in(5,\ 0.500)$ in red and of the $\mathcal{N}(2.50, 1.25)$ in blue. This will help us see how well the Normal can approximate the Binomial.



Comparing the Binomial and Normal CDFs

Here is a graphic of the CDF of the $\mathcal{B}in(5,\ 0.500)$ in red and of the $\mathcal{N}(2.50, 1.25)$ in blue. This will help us see how well the Normal can approximate the Binomial.



Illustrating the Effect of Sample Size on the Approximation

During class, you saw an animated graphic of the difference in the CDFs of the Binomial and Normal — as we increase the sample size, n, while holding $p = 0.500$.

It would be helpful to write what you learned about it in this space.

Several Results and Conclusions

From this brief experiment, we can come to one conclusion:

- The approximation is better for larger values of n

With additional exploration, we could come to a second conclusion:

- The approximation is better for values of p close to 0.500

Combining these two observations leads to the following "rule of thumb"

- The Normal distribution is "sufficiently" close to the Binomial distribution if both the expected number of successes and the expected number of failures is at least 10:
  $np \geq 10$ and $n(1-p) \geq 10$.

Note: The approximation is improved by using a "**continuity correction**" of 0.500. (added if $\leq$, subtracted if $\geq$, both if $=$). Let us see some examples...

Example 1: Cheating

Example

Let us examine a certain Mathematics course taught by a certain professor at a certain midwestern college in the United States. On average, 2% of the students cheat on their examinations, thus making their transcripts all but useless as a means of illustrating their knowledge.

Use the continuity correction factor estimate the probability that fewer than two students cheat on the midterm in that mathematics course with 50 students.

In other words, if X is the number of students cheating in the Math midterm examination, and

$$X \sim \mathcal{B}in(50, 0.02)$$

then calculate $\mathbb{P}[\ X < 2\ ]$.

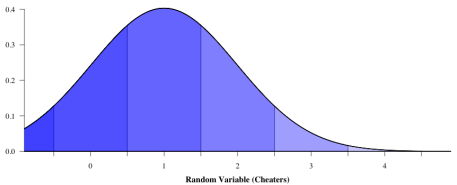
Example 1: Cheating

Solution steps

- First, what is the Normal distribution that best approximates this Binomial?
  - $X \sim \mathcal{B}in(50, 0.02)$
  - Thus, $\mu = np = 1$ and $\sigma^2 = np(1-p) = 0.98$
  - Thus, $X \sim Y \sim \mathcal{N}(\mu = 1, \sigma^2 = 0.98)$.
- Next, we determine the probability statement.
  - The problem is asking for $\mathbb{P}[\ X < 2\ ]$
  - The continuity correction adds and/or subtracts 0.500 to the value, **whichever includes the region**.
  - Thus, $\mathbb{P}[\ X < 2\ ] \approx \mathbb{P}[\ Y \leq 1.5\ ]$

Example 1: Cheating

The approximate distribution:



Example 1: Cheating

Solution...

Therefore, using the continuity correction, the probability of fewer than two students cheating on their Mathematics test is

$$\texttt{pnorm(1.5, m=1, s=sqrt(0.98))} = 0.6932474$$

We can calculate the correct probability using the computer so we can compare how well the approximation did:

$$\texttt{pbinom(1, size=50, prob=0.02)} = 0.7357714$$

Example 2: Births

Example

While it is not entirely accurate, for the purposes of this problem, let us assume that the probability of a girl being born is 50%. Estimate the probability of more than 55 girls being born in 100 births using a continuity correction.

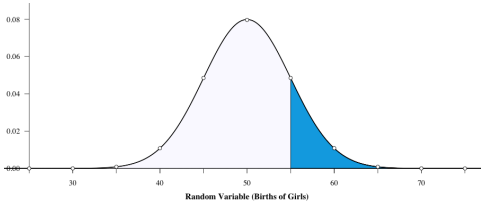
In other words, if X is the number of girls born in those 100 births, then

$$X \sim \mathcal{B}in(100, 0.500)$$

and we are to calculate $\mathbb{P}[\ X > 55\ ]$.

Example 2: Births

The approximate distribution:



Example 2: Births

Solution

Here are the probability calculations:

$$\begin{aligned} & X \sim \mathcal{B}in(100, 0.500) \\ \implies \quad & Y \sim \mathcal{N}(\mu = 50, \sigma^2 = 25) \end{aligned}$$

$$\begin{aligned} \mathbb{P}[\, X > 55 \,] &\approx \mathbb{P}[\, Y \geq 55.5 \,] \\ &= 1 - \mathbb{P}[\, Y \leq 55.5 \,] \\ &= 0.1356661 \end{aligned}$$

Thus, there is about a 13.6% probability of more than 55 girls will be born in 100 births.

Example 2: Births

Solution...

This last calculation came from R's command

```
1 - pnorm(54.5, m=50, s=5)
```
$= 0.1356661$

The real answer is

```
1 - pbinom(55, size=100, prob=0.500)
```
$= 0.1356265$

Errors:

The absolute error is just	0.000 04	"estimate − actual"
The relative error is only	0.029%	"(estimate − actual) / actual"

Example 3: The Stats Examination

Example

After many hours of studying for your statistics examination, you believe that you have a 90% probability of answering any given question correctly. Your test includes 50 true/false questions. Assuming that your estimate is the true probability that you will answer any question correctly, use a Normal distribution to estimate the probability that you will miss no more than 4 questions.

In other words, if X is the number of problems you get **wrong** on this examination, then

$$X \sim \mathcal{B}in(50,\ \mathbf{0.100})$$

and we are to calculate $\mathbb{P}[\ X \leq 4\ ]$.

Example 3: The Stats Examination

Solution Here are the probability calculations:

$$\begin{aligned} & X \sim \mathcal{B}in(50,\ 0.100) \\ \Longrightarrow \qquad & Y \sim \mathcal{N}(\mu = 5,\ \sigma^2 = 4.5) \end{aligned}$$

$$\begin{aligned} \mathbb{P}[\ X \leq 4\ ] &\approx \mathbb{P}[\ Y \leq 4.5\ ] \\ &= 0.4068319 \end{aligned}$$

Thus, the probability that you miss no more than four problems is 40.7%:

- That is, the probability that you will earn at least an A- is 40.7%.
- I leave it as an exercise to show that the probability of earning an A is 25.0%.

Example 3: The Stats Examination

Solution...

This last calculation came from R's command

```
pnorm(4.5, m=5, s=sqrt(4.5))
```
$= 0.4068319$

The real answer is

```
1 - pbinom(4, size=50, prob=0.100)
```
$= 0.4311984$

Errors:

The absolute error is just	-0.024 367	"estimate − actual"
The relative error is only	-5.650%	"(estimate − actual) / actual"

Example 4: Toothpaste

Example

Many toothpaste commercials advertise that "3 out of 4 dentists recommend their brand of toothpaste." Use a Normal distribution to estimate the probability that in a random survey of 400 dentists, *exactly* 300 will recommend Brand F toothpaste.

In this calculation, let us assume that the commercials are correct, and therefore, there is a 75% chance that any specific dentist will recommend Brand F toothpaste.

In other words, if X is the number of dentists recommending Brand F toothpaste, then

$$X \sim \mathcal{B}in(400, 0.750)$$

and we are to calculate $\mathbb{P}[\ X = 300\ ]$.

Example 4: Toothpaste

Solution

Here are the probability calculations:

$$\begin{aligned} & X \sim \mathcal{B}in(400, 0.750) \\ \implies \quad & Y \sim \mathcal{N}(\mu = 300, \sigma^2 = 75) \end{aligned}$$

$$\begin{aligned} \mathbb{P}[\ X = 300\ ] &\approx \mathbb{P}[\ 299.5 < Y \leq 300.5\ ] \\ &\approx \mathbb{P}[\ Y \leq 300.5\ ] - \mathbb{P}[\ 299.5 \leq Y\ ] \\ &\approx 0.0460403 \end{aligned}$$

Thus, the probability that *exactly* 300 randomly-selected dentists recommend Brand F toothpaste is approximately 0.0460403.

Example 4: Toothpaste

Solution...

This last calculation came from R's command

```
pnorm(300.5, 300, sqrt(75)) - pnorm(299.5, 300, sqrt(75))
```
$= 0.046\,040$

The real answer is

```
dbinom(300, size=400, prob=0.75)
```
$= 0.046\,024$

Errors:

The absolute error is just	0.000 016	"estimate − actual"
The relative error is only	0.034%	"(estimate − actual) / actual"

Today's Summary

Now that we have concluded this lecture, you should be able to

- describe the Binomial distribution in detail
- describe the Normal distribution in detail
- explain how the Normal distribution can be used to approximate the Binomial distribution
- calculate approximate Binomial probabilities using the Normal distribution
- calculate exact Binomial probabilities
- explain why approximating the Binomial is useful, *even if* we have a computer to calculate those probabilities exactly

Useful R Functions

Allow me to remind you of the following functions dealing with the Binomial and Normal distributions. Remember that the CDFs (!) are important in approximations.

- `dbinom(x, size, prob)` is the pmf, $\mathbb{P}[\ X = x\ ] = p$
- ! `pbinom(x, size, prob)` is the CDF, $\mathbb{P}[\ X \leq x\ ] = p$
- `qbinom(p, size, prob)` is the quantile function, $\mathbb{P}[\ X \leq x\ ] = p$
- `rbinom(n, size, prob)` generates n random values from this distribution

- `dnorm(x, m,s)` is the density function, $f(x)$
- ! `pnorm(x, m,s)` is the CDF $F(x) = \mathbb{P}[\ X \leq x\ ]$
- `qnorm(p, m,s)` is the quantile function $\mathbb{P}[\ X \leq x\ ] = p$
- `rnorm(n, m,s)` generates n random values from this distribution

Supplemental Activities

The following are supplements for the topics covered today.

- SCA 5 is for discrete distributions like the Binomial.
- SCA 6 is for continuous distributions like the Normal.

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCA, **Laboratory Activity B** is helpful for learning how to handle some discrete distributions and **Laboratory Activity C** is helpful for learning how to handle some continuous distributions. Lab C actually illustrates the Central Limit Theorem, which is central to why the Normal can be used to approximate the Binomial.
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding some distributions — and the Central Limit Theorem:

- Hawkes Learning: Section 6.5
- Intro to Modern Statistics: Section 13.1
- R for Starters: Appendix B.3, Appendix C
- Wikipedia: Binomial Distribution, Normal Distribution, Central Limit Theorem