The Normal Approximation

Gregory M. Shinault

Overview

Goals for this Lecture

1. Learn how to use the Normal distribution to approximate the Binomial distribution, and when it is useful.
2. Learn how to derive the standard formula for confidence intervals of proportions.
3. Learn how to use the *continuity correction* to make a more accurate approximation.

This material corresponds to section 4.1-4.3 of the textbook.

Introduction

The binomial distribution is incredibly useful in applications. Today we learn to approximate it with the normal distribution for several reasons.

1. Binomial coefficients can be computationally expensive (and unstable).
2. Limit theorems are where we see universal behavior arise from randomness.

Big Idea

Suppose $X \sim \text{Bin}(n, p)$.
Set $\mu = \mathbb{E}X = np$ and $\sigma^2 = \text{Var}(X) = np(1-p)$.

If n is a large number then

$$X \overset{d}{\approx} Y \sim N(\mu, \sigma^2).$$

Alternately,

$$\frac{X - \mathbb{E}X}{SD(X)} \overset{d}{\approx} Z \sim N(0, 1).$$

Formal Statement

de Moivre-Laplace Theorem (1738): Suppose $S_n \sim \text{Bin}(n, p)$ is a sequence of random variables for a fixed value of p.

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \Phi(b) - \Phi(a).$$

Typical Use

Generally we will have $X \sim \text{Bin}(n, p)$ and want to compute $\mathbb{P}(k \leq X \leq \ell)$.

$$\begin{aligned} \mathbb{P}(k \leq X \leq \ell) &= \mathbb{P}\left(\frac{k-np}{\sqrt{np(1-p)}} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{\ell-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{\ell-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Proof

Interactive Explanation

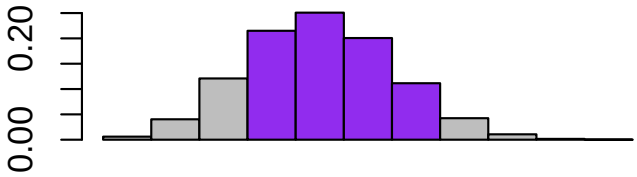
I wrote a simple app to illustrate how closely the Normal distribution fits the Binomial distribution: `http://shinault.shinyapps.io/BinomialApprox`

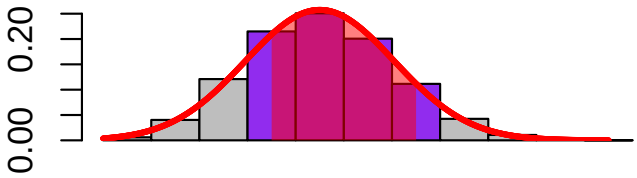
You can experiment with a web app to see how well the Gaussian PDF fits the Binomial PMF for various choices of the parameters n and p.

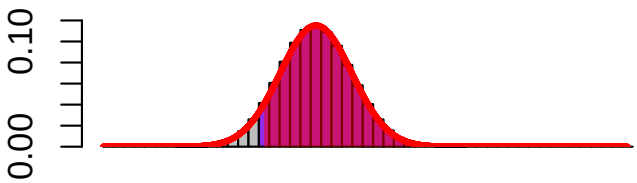
Proof Idea

Use Stirling's formula $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ and Riemann sums.

This is a limit theorem, so the primary mathematical ideas used are from analysis (Math 421 or Math 521). We will not cover the formal proof in lecture. Instead, we will look at some pictures to illustrate the ideas of the proof.


0.00
0.20


0.00
0.20



Practical Concerns

When is n big enough?

Rule of Thumb: If $Var(X) = np(1-p) > 10$, the normal approximation to the binomial distribution should be fairly accurate.

Example

This is Wisconsin, so we ask 100 people if the prefer mozzarella sticks or cheese curds. The probability a random person will prefer cheese curds is 0.6. Approximate the probability that at least 70 people prefer cheese curds.

Solution in R

```
SampleSize <- 100
Prob <- 0.6
MeanX <- SampleSize * Prob
SDX <- sqrt(SampleSize * Prob * (1 - Prob))
Actual <- pbinom(100, size = SampleSize, prob = Prob) -
    pbinom(69, size = SampleSize, prob = Prob)
Approx <- pnorm((100 - MeanX)/SDX) - pnorm((70 -
    MeanX)/SDX)
c(Actual, Approx)
```

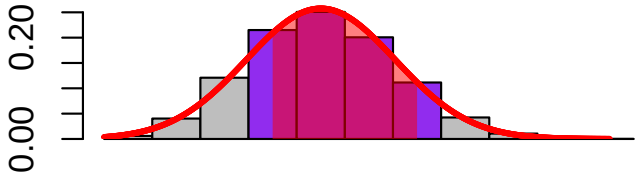
```
## [1] 0.02478282 0.02061342
```

Example (Confidence intervals)

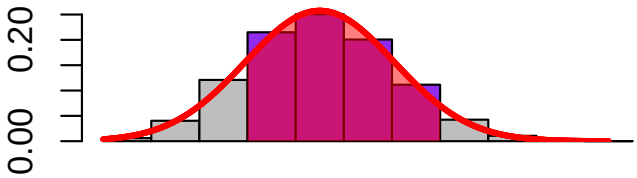
Suppose we want to predict the next election. To do so we take a public opinion poll and let $\hat{p}$ denote the fraction of people who will vote for the Democratic candidate.
How many people must we survey in order to be 95% certain that $\hat{p}$ will be within 0.005 of the fraction of people who will vote for the Democratic candidate for the whole population, p?

Continuity Correction

Missing The Right and Left Edges of Rectangles!



The Fix: Extend the Integral



Formula for Continuity Correction

For integers a, b and $X \sim \text{Bin}(n, p)$, we can use the improved approximation

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a - 1/2 \leq X \leq b + 1/2) \\ &\approx \Phi\left(\frac{b+1/2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-1/2-np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

Example

This is Wisconsin, so we ask 100 people if the prefer mozzarella sticks or cheese curds. The probability a random person will prefer cheese curds is 0.6. Approximate the probability that at least 70 people prefer cheese curds, using the continuity correction this time.

Solution in R

```
SampleSize <- 100
Prob <- 0.6
MeanX <- SampleSize * Prob
SDX <- sqrt(SampleSize * Prob * (1 - Prob))
Actual <- pbinom(100, size = SampleSize, prob = Prob) -
    pbinom(69, size = SampleSize, prob = Prob)
Approx <- pnorm((100 - MeanX)/SDX) - pnorm((70 -
    MeanX)/SDX)
```

```
ApproxImp <- pnorm((100 + 0.5 - MeanX)/SDX) -
    pnorm((70 - 0.5 - MeanX)/SDX)
c(Actual, Approx, ApproxImp)
```

```
## [1] 0.02478282 0.02061342 0.02623975
```

Example

We buy 2 gross of eggs for the Lion's Club pancake breakfast. In most packages the probability each egg is broken is 0.05. What is the approximate probability that less than 10 of our eggs are broken? Be certain to use the continuity correction.

Solution in R

```
SampleSize <- 2 * 144
Prob <- 0.05
MeanX <- SampleSize * Prob
SDX <- sqrt(SampleSize * Prob * (1 - Prob))
Actual <- pbinom(9, size = SampleSize, prob = Prob)
Approx <- pnorm((9 - MeanX)/SDX) - pnorm((0 -
    MeanX)/SDX)
ApproxCor <- pnorm((9 + 0.5 - MeanX)/SDX) - pnorm((0 -
    0.5 - MeanX)/SDX)
c(Actual, Approx, ApproxCor)
```

```
## [1] 0.08619932 0.07209659 0.09258931
```

The Wrap Up

Summary

1. If n is big, then Bin(n, p) is approximately the same distribution as $N(np, np(1-p))$.
2. Our rule of thumb for using this approximation is that Var$(X) > 10$.
3. The continuity correction can increase the accuracy significantly. However, you are only required to use it if explicitly requested in this course.
4. For this course, the most important application of the normal approximation is in deriving formulas for confidence intervals for proportions.

Next Step

What do we do if the rule of thumb says not to use the normal approximation?

Finer Points

The proof of the de Moivre-Laplace Theorem is provided in your textbook. I strongly recommend reading it.

Just how accurate is the normal approximation? If you are interested in precise statements about the error in this approximation, look up the Berry-Esseen inequality.