Lesson 2.3

The Normal Approximation to the Binomial Distribution

Introduction

An important application of the Central Limit Theorem deals with approximating the sampling distributions of functions of count data.

Suppose that $Y_1, Y_2, \cdots, Y_n$ is a random sample from a Bernoulli(p) distribution; that is, $Y_i = 1$, if the i^{th} trial is a "success", and $Y_i = 0$, otherwise. Recall that the probability mass function of the Bernoulli random variable is

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y},\ y = 0, 1 \\ 0, \text{ otherwise} \end{cases}$$

Hence, the sample $Y_1, Y_2, \cdots, Y_n$ is a string of zeros and ones, where $P(Y_i = 1) = p$, for each i. In the Bernoulli model,

$$\begin{aligned} E(Y) &= p \\ V(Y) &= p(1-p) \end{aligned}$$

Also, we know that

$$U = \sum_{i=1}^{n} Y_i \sim binom(n, p).$$

Define the sample proportion

$$\hat{p} = \frac{1}{n}U = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Note that $\hat{p}$ is an average of iid values of 0 and 1, thus the CLT must apply. That is, for large n,

$$\hat{p} \sim AN\left(p, \frac{pq}{n}\right), \, q = 1 - p$$

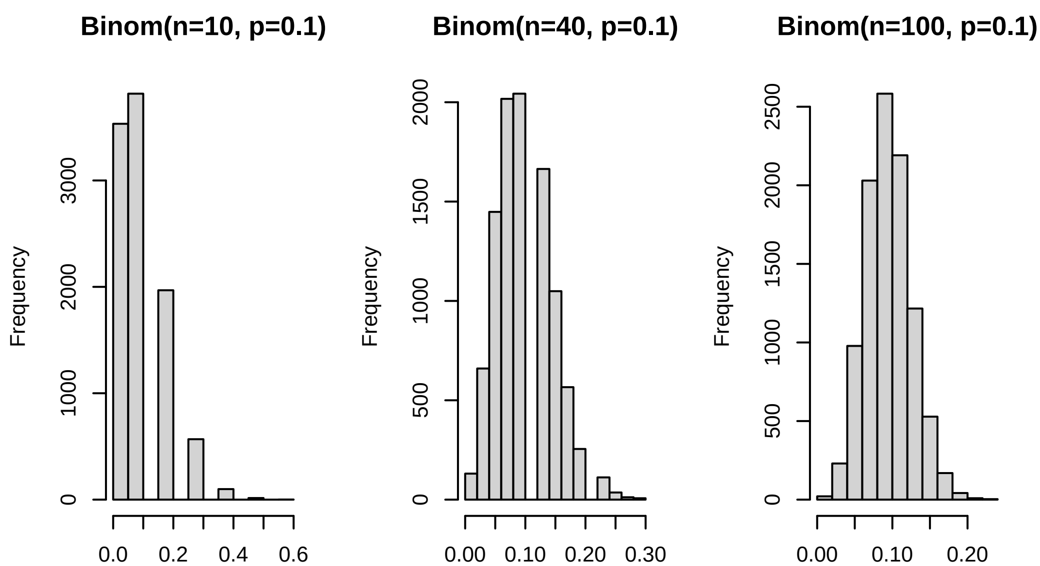
Equivalently, we say

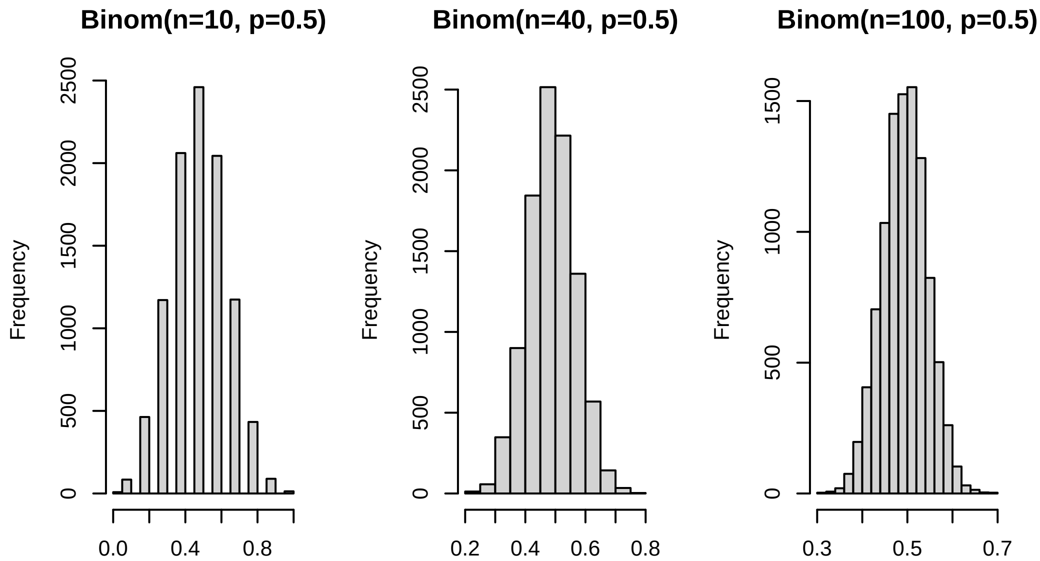
$$U = \sum_{i=1}^{n} Y_i \sim AN(np, \frac{pq}{n}), \, q = 1 - p$$

<u>How good is this approximation?</u>

Since we are sampling a "binary" population, one might wonder how well the normal distribution approximates the true sampling distribution of $\hat{p}$. The approximation is **best** when

1. n is large (in other words the approximation improves as n increases), and
2. p is close to 0.5





Rules of Thumb:

One can feel comfortable using the normal approximation as long as both np and $n(1-p)$ are larger than 10. Other guidelines have been proposed in the literature (instead of 10 they suggest 5). This is just a guideline.

Example 2.3.1

Suppose that Y has a binomial distribution with $n = 30$ and $p = 0.4$. Find the exact probabilities that $Y \leq 8$ and $Y = 8$ and compare these to the corresponding values found by using the normal approximation.

SOLUTION

a. Using R we obtain

$$P(Y \leq 8) = 0.094$$
$$P(Y = 8) = 0.0505$$

b. We check first the conditions: $np = 30 \times 0.4 = 12 > 10$ and $n(1-p) = 30 \times (1-0.4) = 18 > 10$, hence, we can apply the normal approximation. So we assume that $Y \sim AN(\mu, \sigma^2)$, where $\mu = np = 12$ and $\sigma^2 = np(1-p) = 7.2$.

$$
\begin{aligned}
P(Y \leq 8) &\approx P\left(Z \leq \frac{8.5 - 12}{\sqrt{7.2}}\right) \\
&= P(Z \leq -1.30) \\
&= 0.0968
\end{aligned}
$$

Meanwhile,

$$
\begin{aligned}
P(Y = 8) &\approx P(7.5 \leq Y \leq 8.5) \\
&= P\left(\frac{7.5 - 12}{\sqrt{7.2}} \leq Z \leq \frac{8.5 - 12}{\sqrt{7.2}}\right) \\
&= P(-1.68 \leq Z \leq -1.30) \\
&= P(Z \leq -1.30) - P(Z \leq -1.68) \\
&= 0.0503
\end{aligned}
$$

Example 2.3.2

Previous studies have found that 75% of adults use the internet on a regular basis. A researcher believes this percentage has recently increased. He conducts a survey and discovers that 2144 out of 2824 adults surveyed use the internet on a regular basis. Assuming 75 actually is the correct percentage, what's the probability of seeing at least this many users out of 2824 users?

SOLUTION: [Left as a classroom exercise]

Example 2.3.3

Six percent of people are universal blood donors (i.e., they can give blood to anyone without it being rejected). A hospital needs 10 universal donors to donate blood, so they conduct a blood drive. If 200 volunteers donate blood, what is the probability tht the number of universal donors is at least 10?

SOLUTION: [Left as a classroom exercise]