Lab 2: Extra Credit

STAT 310, Spring 2021

The Central Limit Theorem (CLT) states that the sampling distribution for $\hat{p}$ approximately follows a normal distribution centered around the population proportion p, and with standard error $\sqrt{p(1-p)/n}$. That is:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

In this lab, we will verify the CLT using a computer simulation in R.

We can think in terms of an election. Suppose that a population consists of 1 million voters, and that 60% percent of the population will vote for a certain candidate, so $p = 0.6$ is the population proportion.[1] Now, if we take a random sample of $n = 100$ voters, we would expect that the sample proportion $\hat{p}$ of those 100 voters that support the candidate to be close to 0.6, but not exactly 0.6 because of random sampling error. To create a sampling distribution, we would repeatedly take random samples of size $n = 100$, and compute the sample proportion from each sample.

Here are the steps for simulating a sampling distribution:

1. Create a vector containing 600,000 ones and 400,000 zeros, which represents the population of 1 million voters that support (coded as 1) and do not support (coded as 0) the candidate.
2. Take a random sample of size $n = 100$ voters from the population. Then compute the sample proportion of those voters that support the candidate.
3. Repeat the previous step 5,000 times to generate 5,000 sample proportions.
4. Make a histogram of the 5,000 sample proportions.

[1] In reality, we usually don't know the value of the population proportion p, but for the purpose of the simulation we will assume that it is known.

Here's the code to simulate one sample proportion in R:

```
pop_size = 10^6 # population size
n = 100 # sample size

# make population with 600,000 ones and 400,000 zeros
population = c(rep(1, 0.6*pop_size), rep(0, 0.4*pop_size))

# take random sample of size 100
samp = sample(population, size = n)
# compute sample proportion
phat = sum(samp) / n
phat
```

```
## [1] 0.59
```

In this first simulation, we see that the sample proportion is 0.59, which has an error of -0.01.

Each time we run this code we will get a different value for the sample proportion. For example:

```
# take random sample of size 100
samp = sample(population, size = n)
# compute sample proportion
phat = sum(samp) / n
phat
```

```
## [1] 0.62
```

In this second simulation, we see that the sample proportion is 0.62, which as an error of +0.02.

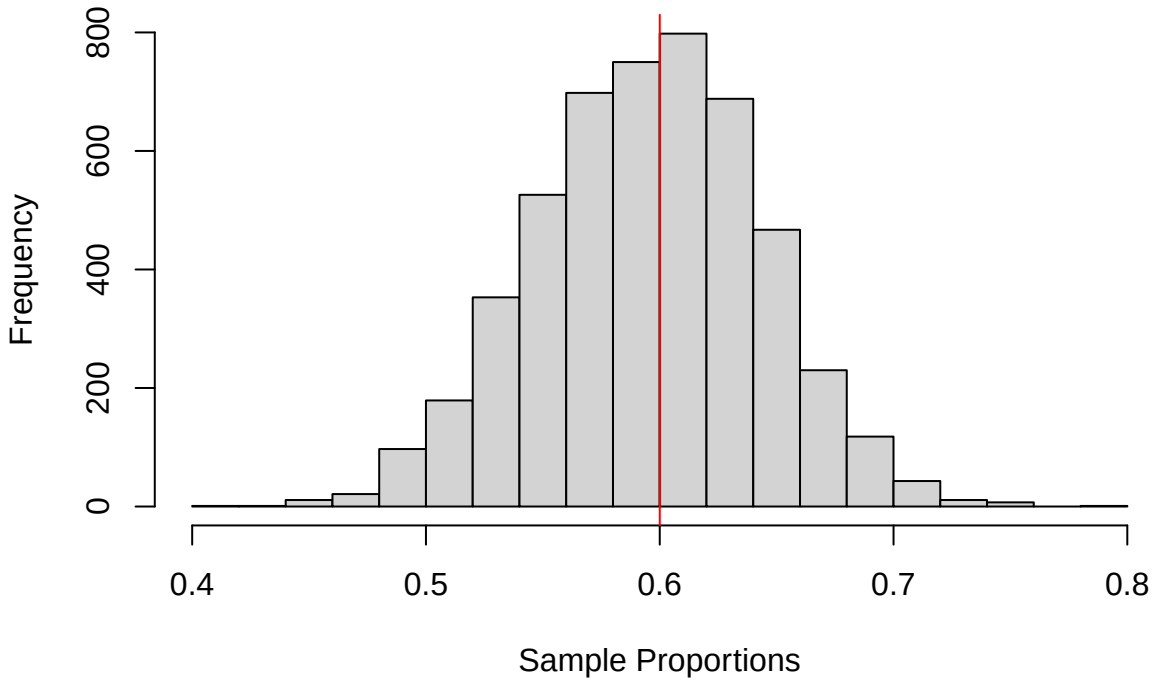
To simulate the sampling distribution, we need to run this code 5,000 times. This would be tedious to do manually. So instead we use a coding construct called a `for loop`, which will repeat code as many times as we want. The `for loop` is one of the most important constructs in computer science. Every programming language has its own version of the `for loop`.

Here's the code which uses a `for loop` to generate the sampling distribution:

```
pop_size = 10^6 # population size
n = 100 # sample size
population = c(rep(1, 0.6*pop_size), rep(0, 0.4*pop_size))
phats = c() # initialize vector for sample proportions
for(i in 1:5000) {
  samp = sample(population, size = n)
  phats[i] = sum(samp) / n
}
```

The vector `phats` contains the 5,000 sample proportions. Next we plot the histogram:

```
hist(phats, xlab = "Sample Proportions", main = "")
abline(v = 0.6, col = "red") # make vertical line at p=0.6
```



Amazingly, the histogram looks like a normal distribution centered around the population proportion $p = 0.6$. This verifies the CLT!

Moreover, if we take the standard deviation of the 5,000 sample proportions we get

```
sd(phats)
```

```
## [1] 0.04829453
```

which is equal to SE $= \sqrt{p(1-p)/n} = \sqrt{0.6(0.4)/100} = 0.049$. This verifies the SE formula from the CLT!

Extra Credit Assignment (6 points)
Due: Friday, March 12

Directions: Your answers should be formatted using a word processor (e.g., Microsoft Word, Google Docs). For each exercise, copy and paste the graphs from RStudio to your document (you do not need to copy and paste the code since it is already provided). Then convert your final document to PDF format.

Exercise 1. Run the code below to simulate the sampling distribution when the sample size $n = 50$:

```
pop_size = 10^6 # population size
n = 50 # sample size
population = c(rep(1, 0.6*pop_size), rep(0, 0.4*pop_size))
phats = c() # initialize vector for sample proportions
for(i in 1:5000) {
  samp = sample(population, size = n)
  phats[i] = sum(samp) / n
}
```

Next, make the histogram of the 5000 sample proportions:

```
hist(phats, xlim = c(0.3, 0.9))
```

Here `xlim` specifies the range for the x-axis.

Exercise 2. Run the code below to simulate the sampling distribution when the sample size is $n = 1000$:

```
pop_size = 10^6 # population size
n = 1000 # sample size
population = c(rep(1, 0.6*pop_size), rep(0, 0.4*pop_size))
phats = c() # initialize vector for sample proportions
for(i in 1:5000) {
  samp = sample(population, size = n)
  phats[i] = sum(samp) / n
}
```

Next make the histogram of the 5000 sample proportions:

```
hist(phats, xlim = c(0.3, 0.9))
```

Exercise 3. How did increasing the sample size from $n = 50$ to $n = 1000$ affect the shape of the sampling distribution?