INFX 573 Lab: Central Limit Theorem

Shuyang Wu

October 25th, 2016

Collaborators:

Don't forget to list the full names of your collaborators!

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week5a_lab.Rmd` file from Canvas. Open `week5a_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week5a_lab.Rmd`.

2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.

3. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.

4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, rename the R Markdown file to `YourLastName_YourFirstName_lab5a.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

   In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
library(gridExtra)
```

Problem 1: Simulating Data in R

R can easily generate random samples from many different probability distributions. Here, you will use this functionality to explore the Central Limit Theorem by performing a simulation experiment.

Hint: You might find out more about the distribution functions in R using the help files, `?distributions`.

Step 1: Pick your favorite probability distribution.

- What distribution did you choose? Bernoulli Distribution
- What are the parameters that characterize the distribution you chose? Mean, variance, skewness
- Describe a situation in which you would expect to see this distribution in real-world data. Survival status

Hint: Look at OpenIntro Statistics Chapter 3.

Hint: Each distribution function in R has an associated function to generate random deviates, e.g. `rbinom` for the Binomial distribution.

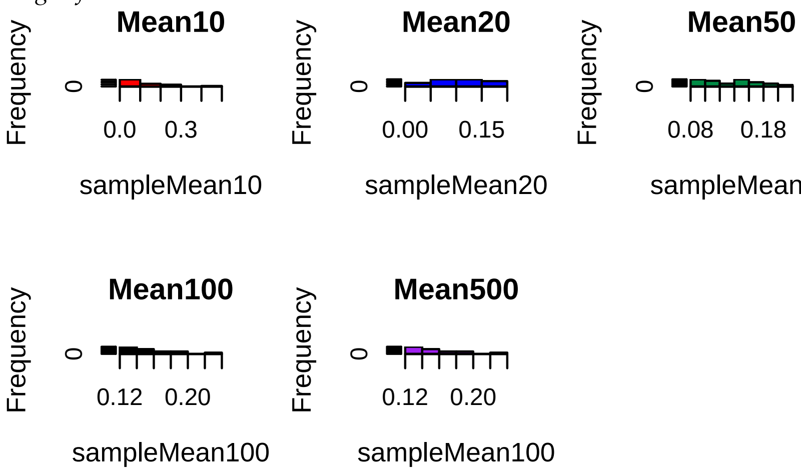
```
sample10 <- NULL
sample20 <- NULL
sample50 <- NULL
sample100 <- NULL
sample500 <- NULL
a <- c(10, 20, 50, 100, 500)
for (i in 1:100) {
    n = sample(a, 1)
    if (n == 10) {
        sample10 <- cbind(sample10, rbinom(n,
            1, 0.15))
    } else if (n == 20) {
        sample20 <- cbind(sample20, rbinom(n,
            1, 0.15))
    } else if (n == 50) {
        sample50 <- cbind(sample50, rbinom(n,
            1, 0.15))
    } else if (n == 100) {
        sample100 <- cbind(sample100, rbinom(n,
            1, 0.15))
    } else {
        sample500 <- cbind(sample500, rbinom(n,
            1, 0.15))
    }
}
```

Step 2: Choose a value for each parameter in the distribution (e.g. the mean and variance for the Normal distribution). Use the random generation function for this distribution to construct 100 random samples of sample sizes `n = 10, 20, 50, 100, 500`*.*

```
# par(mfrow=c(2,5)) ?gridExtra
sampleMean10 <- colMeans(sample10)
sampleMean20 <- colMeans(sample20)
sampleMean50 <- colMeans(sample50)
sampleMean100 <- colMeans(sample100)
sampleMean500 <- colMeans(sample500)
# Plot them together
par(mfrow = c(2, 3))
```

```
hist(sampleMean10, col = "red", main = "Mean10",
    cex.lab = 1.1)
hist(sampleMean20, col = "blue", main = "Mean20",
    cex.lab = 1.1)
hist(sampleMean50, col = "springgreen4", main = "Mean50",
    cex.lab = 1.1)
hist(sampleMean100, col = "black", main = "Mean100",
    cex.lab = 1.1)
hist(sampleMean100, col = "purple", main = "Mean500",
    cex.lab = 1.1)
```

Step 3: Compute the sample mean for each of the 100 random samples. Construct a visualization showing the distribution of the sample mean for each case (i.e. probability distribution and sample size pair). You might find the following code helpful for showing multiple plots at once.



- What is the true population mean for the distribution?

```
sum <- sum(sampleMean10 * 10) + sum(sampleMean20 *
    20) + sum(sampleMean50 * 50) + sum(sampleMean100 *
    100) + sum(sampleMean500 * 500)
popMean <- sum/(10 * 21 + 20 * 15 + 50 * 25 +
    100 * 20 + 500 * 10)
```

Population mean = 0.227

Hint: Most distributions are characterized by parameters related to the mean and variance.

- What patterns do you see in the distribution of the sample mean as the sample size n increases?
- How does this simulation experiment demonstrate the Central Limit Theorem?