IMT 573 Lab: Central Limit Theorem

Naga Soundari Balamurugan

October 25th, 2018

Collaborators: Jayashree Raman, Hye Kim

Instructions:

1. Download the `week5b_lab.Rmd` file from Canvas. Open `week5b_lab.Rmd` in RStudio and supply your solutions to the assignment by editing `week5b_lab.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. If you are using more than just a standard function that you found from another source, please credit the source in the comments.
4. Collaboration on labs is encouraged, but students must turn in an individual assignments. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_Lab5b.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

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

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

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?

  Negative Binomial Distribution

- What are the parameters that characterize the distribution you chose?

  x: vector of (non-negative integer) quantiles. size: target for number of successful trials, or dispersion parameter (the shape parameter of the gamma mixing distribution). Must be strictly positive, need not be integer. prob: probability of success in each trial. 0 < prob <= 1. mu: alternative parametrization via mean, usually multiplied by 2^(1 + x). n: number of observations

- Describe a situation in which you would expect to see this distribution in real-world data.

An immigrant to United States getting the H1B visa approved on second chance.

Hint: Look at OpenIntro Statistics Chapter 3.

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`.

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

```
# Negative binomial Distributions
negBinomDist_10 <- rnbinom(n = 100, size = 10, mu = 10)
negBinomDist_10
```

```
##   [1]  9  4  9 19  9  9 10  8  9 14  8 15  6 18 14  7  3  6 17  9 15 13  6
##  [24] 11  8  8 10  4 12  6  6  7  7  3 10 11  7  7 18 14  8 17 11 10  4 11
##  [47]  9 10 14 17 10 10  8  9 11  9 16 14  9 12 15 13 10 11  9  6 12 16 11
##  [70] 16 10  5 12  9  8  2  7  7 12 12  5 16 11 10 22 10  9  4 14  3 10  9
##  [93] 16  8 13 10  3  6 10  7
```

```
negBinomDist_20 <- rnbinom(n = 100, size = 20, mu = 10)
negBinomDist_20
```

```
##   [1]  6 13 10 14  5  6 12  9  6  6  6  8 12 10  9  3 14 12 11 11 12  9 11
##  [24] 16 12 10  9  4  8 11  6  6 12  6 11 10  9  7  7  1 13  7  7 11 10  8
##  [47]  8 11  6 12  7 15  6 10 13  9  9  4  7  8 17  3  9  7 18 13  7 10  9
##  [70]  7  8  7 14  8  6 19  7 11 14  7 14  4  7 15 11  6  9  7  8  8 19 11
##  [93] 10 10 13 13  9  7 16 24
```

```
negBinomDist_50 <- rnbinom(n = 100, size = 50, mu = 10)
negBinomDist_50
```

```
##   [1]  5 14 12 10  5  6 12 12  4 11  9 18 10 11  6  8 10 14  3 10  7  9 21
##  [24]  7 12  3 13  7 10  9  9 11 15  9  8  8 10 11  6 12  8 22 10  6 10  5
##  [47] 10 15 11  8  7 11 11  7  5  4 12  7  8 12 17 15 11 10  7  9 13 11 13
##  [70] 19 14  8 12 11  9  7  5  2  7 22 12  9 10 12 14 11  9 11  8 16  6 12
##  [93] 12  9 10 13  9  7 12 10
```

```
negBinomDist_100 <- rnbinom(n = 100, size = 100, mu = 10)
negBinomDist_100
```

```
##   [1]  9 17  9 12 16 15  7  8  4 14 11  7  6  7 13 11  9 12 12  4  6  8 13
##  [24] 10  9 13 12 12 11 16  8 11  9 16 11 10  7 17  7 13 11 12 12  9  6 12
##  [47] 12  9  7 13  9  8  4 10  8 11 10  3 13 10 10  7 12 13 11 12 11 13  8
##  [70] 12  8 14  9 13  9 13 13  8 12 12  6 12  7  7 14  9  5 11  5 10  7 11
##  [93]  9  7 14 14  7  8 10  7
```

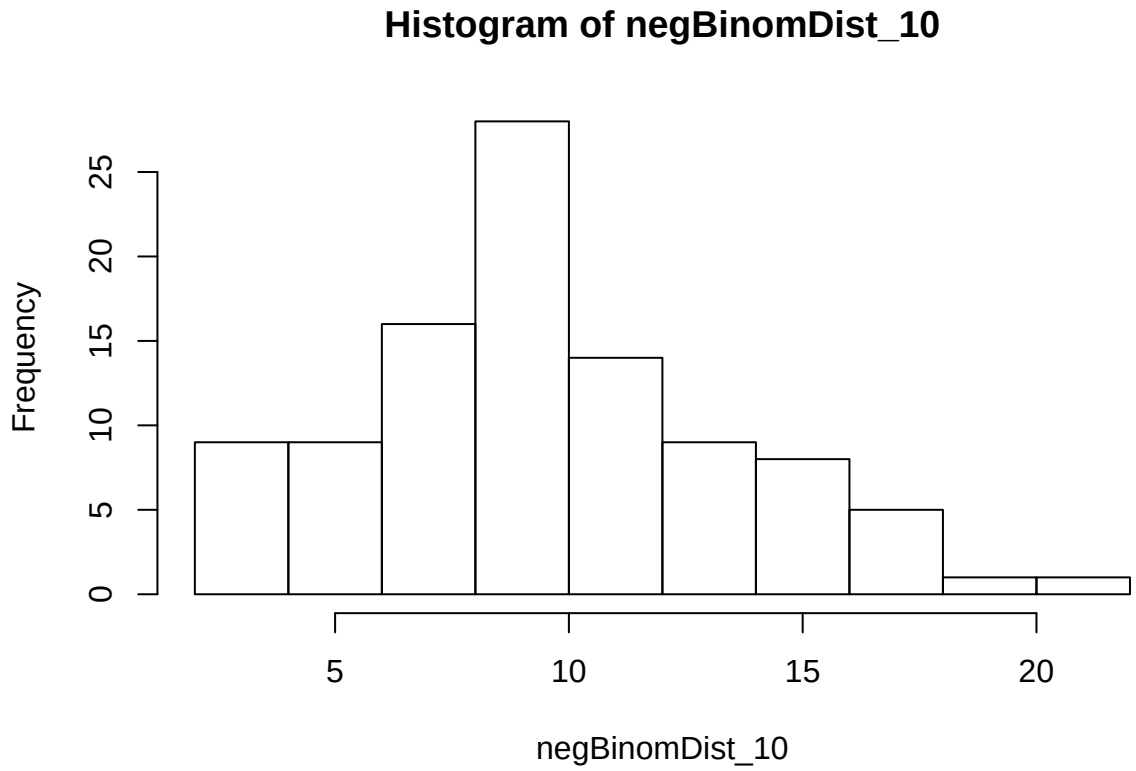
```
negBinomDist_500 <- rnbinom(n = 100, size = 500, mu = 10)
negBinomDist_500
```

```
##   [1]  4 11 14  8 14 14 10 13  8 11 11 13 15 11 10 16  7  9 10 10 16  9 11
##  [24] 10 10 11  9  7  8  8  8 14  8 10  8  9  9  9  5 13  9  9  6 10 12 10
##  [47] 13 13 13 13 11  9  8 14  9  6 12 11  9 10 11  7 11 11 12 13 13 11 11
##  [70]  7  9  6  9  7  5 15  7 10 15  9  8 10  7 10  8  7  7 10 10  9 14  9
##  [93] 10 11 11  6  8  4 12  8
```

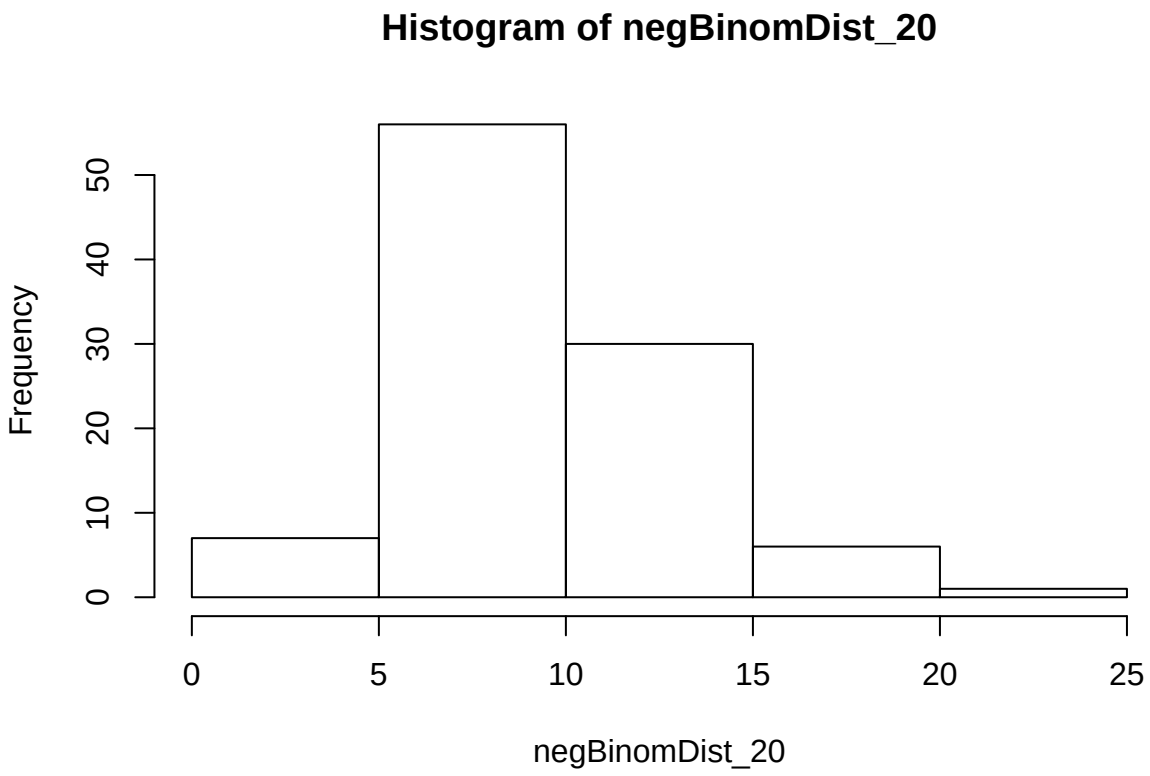
Step 3: Compute the sample mean for each of the 100 random samples. Construct a visualization showing the distribution of the sample mean for each case (i.e. probability distribution and sample size pair).

Hint: We've often used 'ggplot' for plotting in this class. The ggplot function expects a dataframe as input. If you input something other than a dataframe, it tries to coerce the data into a dataframe using the 'fortify()' function. If you've created vectors of data rather than dataframe/s you might want to just use the 'hist()' function instead of 'ggplot'.

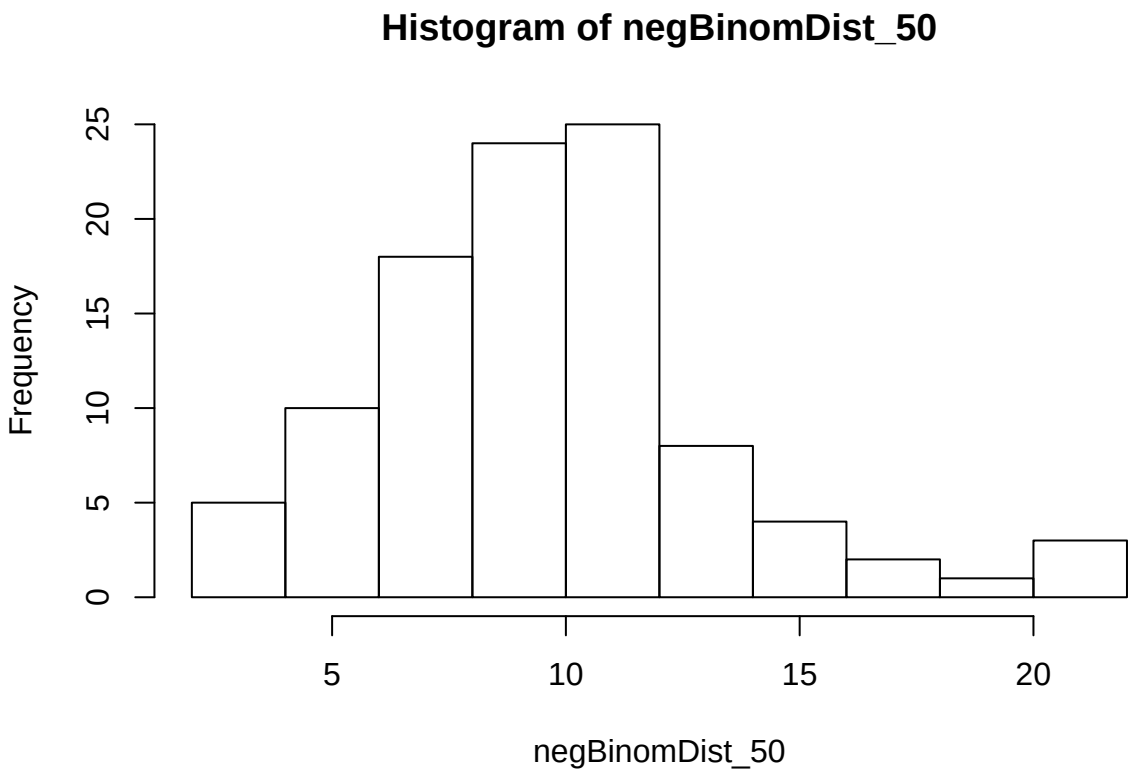
```
# EDIT ME!

# distList <- list(negBinomDist_10, negBinomDist_20, negBinomDist_50, negBinomDist_100,
#                  negBinomDist_500)

# meandf <- lapply(distList, "mean")
# meandf


hist(negBinomDist_10)
```



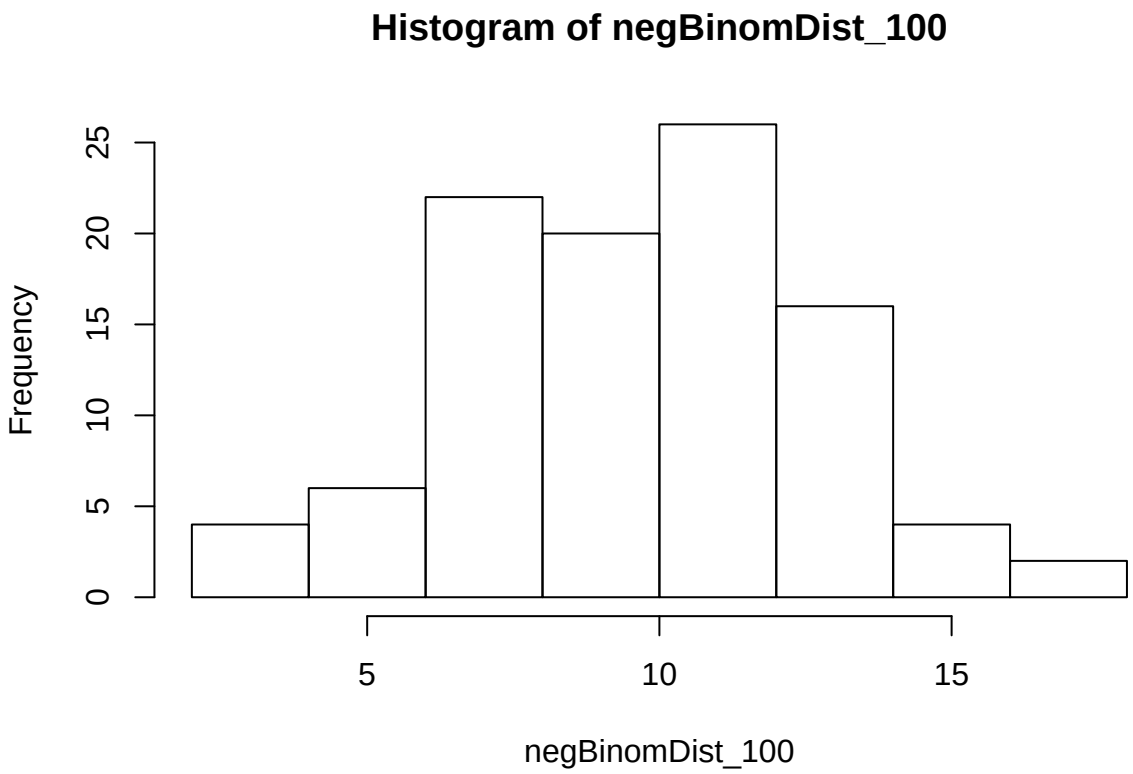

```
hist(negBinomDist_20)
```



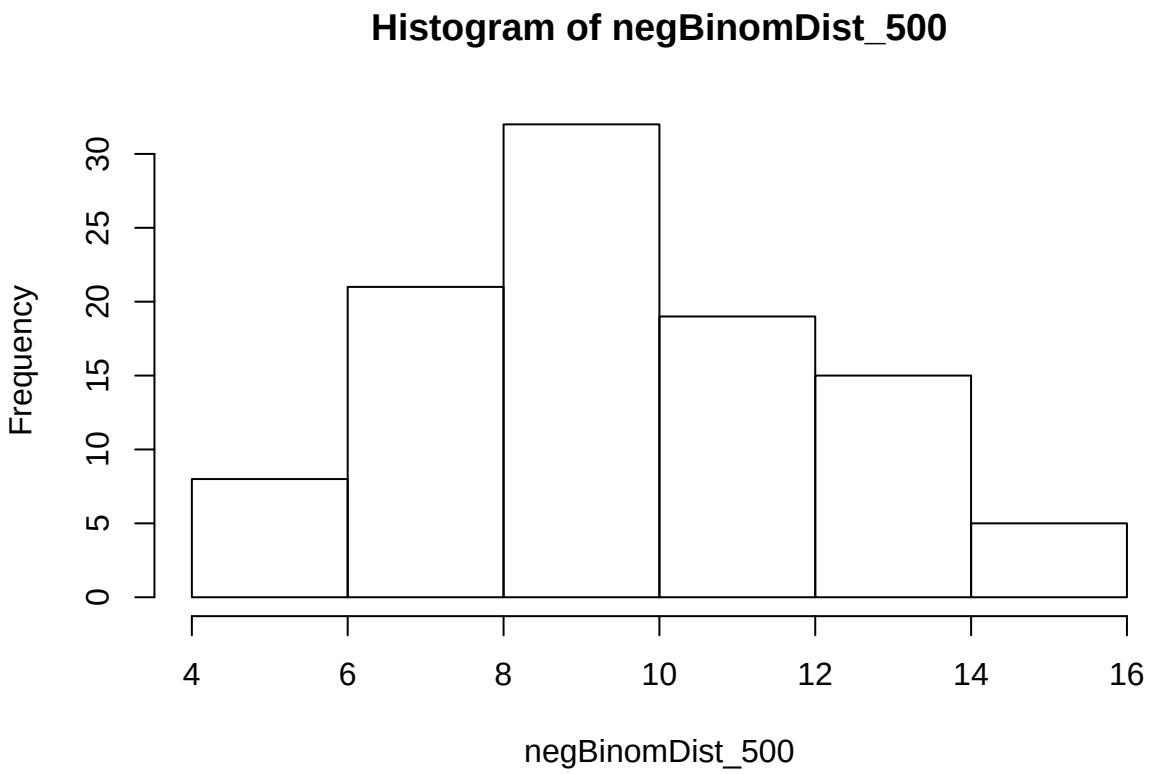

```
hist(negBinomDist_50)
```




```
hist(negBinomDist_100)
```




```
hist(negBinomDist_500)
```



- What is the true population mean for the distribution?

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

The true population mean is centered at 10 for all the distributions as we have set it to that value.

- What patterns do you see in the distribution of the sample mean as the sample size n increases?

  As the sample size n increases, the mean also increases. They both are positively correlated.

- How does this simulation experiment demonstrate the Central Limit Theorem?

  As the central limit theorem states, the distribution becomes more like a normal distribution as the sample size increases.