Lab 8

Sean DeVries

March 6, 2017

In today's lab we will discuss the Normal distribution and show you quick ways to use R to answer these types of problems.

The Central Limit Theorem

For today we are going to use the lipids.txt dataset from the course website. Here we are going to look at the triglyceride levels of the 3026 adults in the study.We are going to visually look at this variable to see if it follows a normal distibution.
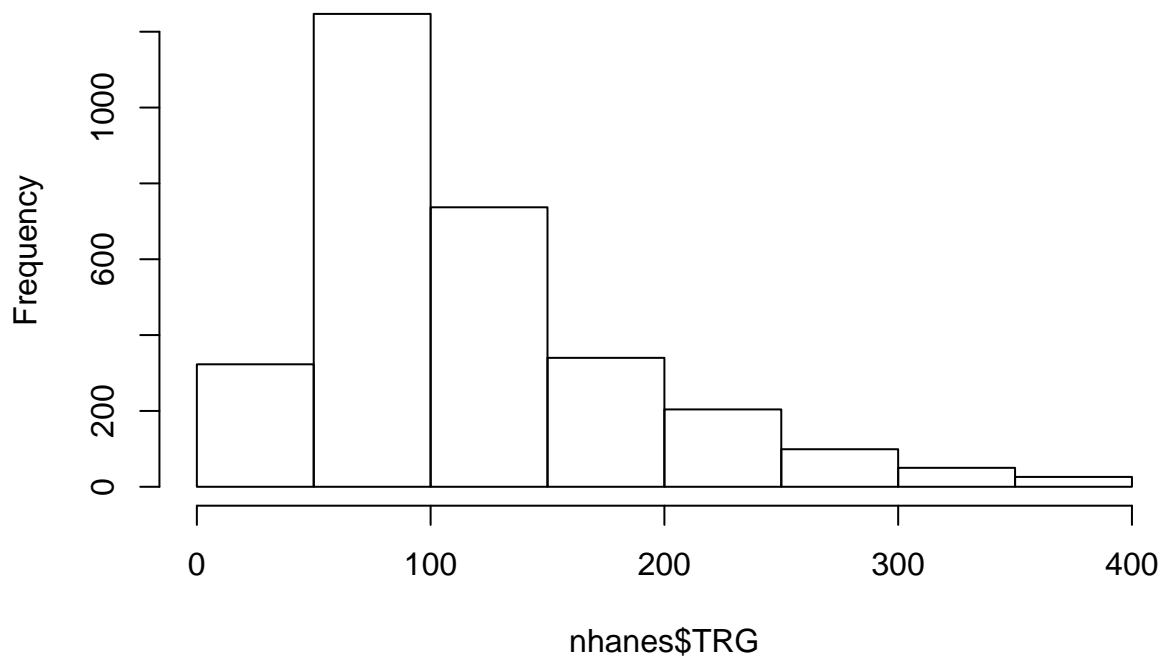
```
nhanes <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lipids.txt")
```

Lets begin by looking at a histogram of the data. Does it look normal to you? Is there a skewness to the data?

```
x <- seq(0,max(nhanes$TRG), len=101)

hist(nhanes$TRG)
```


Histogram of nhanes$TRG

From looking at the histogram we can see that the data is skewed right and do not follow a normal distribution very closely. Which can make sense to us when we use what we have learned in other labs to see that the mean of the triglyceride levels is 116.9 mg/dl which the AMerican Heart Association defines as being a high level.

Now lets say that we could not sample 3026 women and instead obtained a much smaller sample size. R lets us simulate this scenario for when we get a much smaller sample size. In this example we will do a scenario where we only get 10, 50, and 100 people. By using the sample() function everyone in this room can get a unique dataset which is just made by picking out observations from the data.

```
sample10 <- sample(nhanes$TRG, 10)
sample50 <- sample(nhanes$TRG, 50)
sample100 <- sample(nhanes$TRG, 100)
```

But how accurate are these samples? Lets begin by looking at the means of the three samples to see how they compare to that of the overall mean from the original dataset.

```
mean(sample10)
```

```
## [1] 127.2
```

```
mean(sample50)
```

```
## [1] 107.62
```

```
mean(sample100)
```

```
## [1] 123.81
```

As you should see the larger the sample the closer we get to that actual mean of the data.

Binomial Simulation

In R we can use the function rbinom() which takes its arguments to obtain a sample of outcomes from the binomial. An example of this would be flipping a coin, or any other binomial outcome experiment. The rbinom function takes three main arguments, (number of times the binomial will be ran, sample size or number of trials in each binomial, and the probaility of a success in each individual trial). For example if we flipping a coin we would have the following rbinom function for only 1 n value.

```
rbinom(n=1, size = 10, prob=.5)
```

```
## [1] 5
```

If you run this numerous times you will see that you can obtain different numbers and if you understand the normal distribution you can understand that most times you will get a value between 4 and 6, but it is the randomness that allows us to get different values around the class. This can be seen when I change the n value as shown below.

Repeating the experiment

Above we ran just one experiment, but lets say we want to run this same experiment numerous times. This is done by changing that n value in the rbinom function.

```r
rbinom(n=50, size =10, prob =.5)
```

```
##  [1] 3 5 6 7 5 4 5 6 4 4 5 3 5 2 3 6 5 8 4 4 4 4 7 6 8 5 3 6 6 5 4 6 5 8 9
## [36] 9 6 4 3 6 6 6 3 7 2 6 6 5 3 2
```

As the binomial distribution predicts you can see that many of us get many 4,5,and 6's and get an occasion 1 or 9. We can save these results by saving them to a variable name and then many statistics can be found using this variable.

Class Excerices

To show you all how the central limit theorem works we are going to have everyone in the class run a sample(nhanes$TRG, 10). Then each person will come up to the TA's computer and put the mean and sd of their 10 samples into the class spreadsheet. The TA's will then save the document as a CSV file and will upload the file using read.csv. We will then look at the average of all of the classes means and standard deviations to see how they compare to that of the actual data.

Questions to do as a class

Now by yourself, run a rbinom function with as many n values as you like and save it to a variabel and then create a histogram of it. Is this histogram normally distributed? why or why not?

What is the way to find average number of heads in each experiment if we do an n value of what you did above?

Now look at the standard deviation and standard error of each.

From all of these and by looking at a histogram of your data and the means and standard deviations we can see how the central limit theorem applies to the binomial distribution. This is because the standard deviations and standard errors you all found will be close to the $\sqrt{(np(1-p))}$ or $\sqrt{(p(1-p)/n)}$.

Summary/Review Questions

Using R can you answer these questions, dont look at the answers, I included the code so that you can study from it:

A local pharmacy discovers the average amount of people that come in on any given day follows closely to a normal distriubtion with a mean of 150 with a standard deviation of 22. To them a "busy" day is a day in which they help 180 people. A "slow" day is a day in which they help 100 people or less.

A. On a random Tuesday what is the probability of the day being a "busy" day?

B. What is the probability all seven days of the week are "slow" days?

C. Using the binomial funciton we learned last week, what is the probability that three days are "busy" days given the answer to A?

Answers:

A)

```r
pnorm(180, mean = 150, sd = 22, lower.tail = FALSE)
```

```
## [1] 0.08634102
```

B)

```r
pnorm(100, mean = 150, sd = 22, lower.tail = TRUE)^7
```

```
## [1] 2.694716e-14
```

C) Make sure to look at the p-value here for the answer.

```r
binom.test(3,7,p = pnorm(180,mean = 150, sd = 22, lower.tail = FALSE))
```

```
## 
##  Exact binomial test
## 
## data:  3 and 7
## number of successes = 3, number of trials = 7, p-value = 0.01727
## alternative hypothesis: true probability of success is not equal to 0.08634102
## 95 percent confidence interval:
##  0.09898828 0.81594843
## sample estimates:
## probability of success
##              0.4285714
```