Homework1

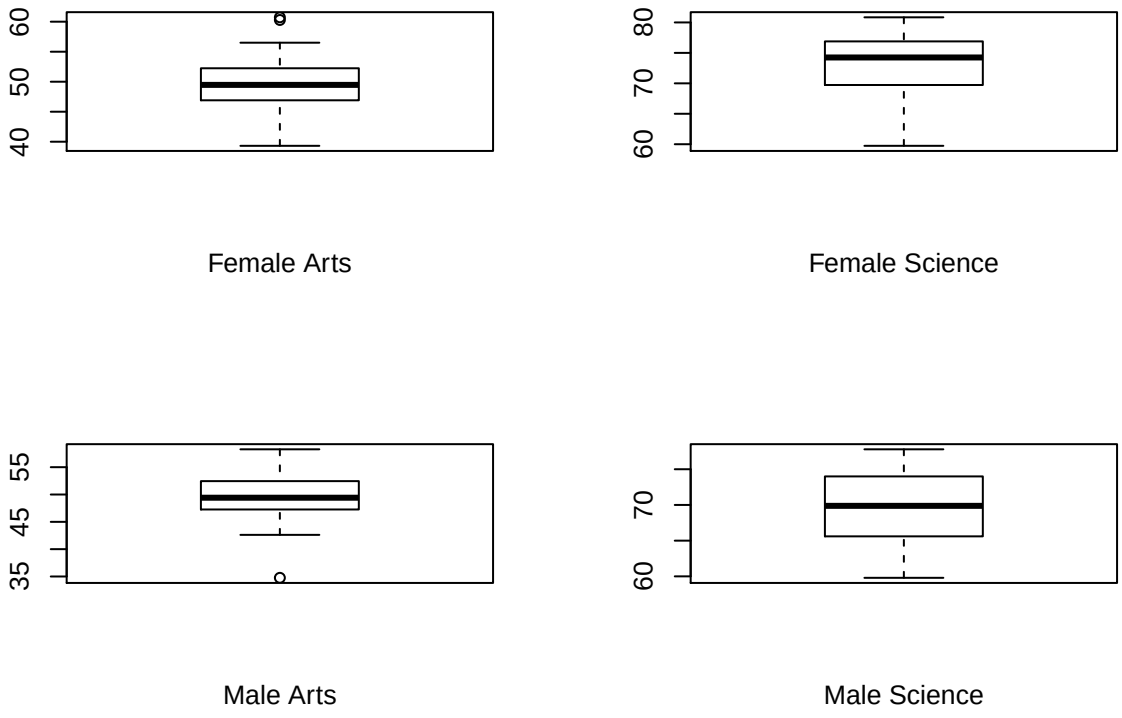
Nitish Neelagiri

September 7, 2015

This is an R Markdown document for Homework 1 submission of STAT 645 - Applied Biostatistics and Data Analysis course I completed as part of MS-MIS degree at Texas A&M

1)For the income by degree and gender datasetin the file "inc_deg_data.csv": (a) Make side by side box plots of income for different gender

```
setwd("/Users/Nitish/Data")
incData <- read.csv("inc_deg_data.csv")
attach(incData)
par(mfrow = c(2,2))
boxplot(income[gender == 0 & degree == 0], xlab = "Female Arts")
boxplot(income[gender == 0 & degree == 1], xlab = "Female Science")
boxplot(income[gender == 1 & degree == 0], xlab = "Male Arts")
boxplot(income[gender == 1 & degree == 1], xlab = "Male Science")
```



(b) Report mean, median, standard deviation, first and third quantiles of income

```
mean(income)
```

```
## [1] 60.62618
```

```r
median(income)
```

```
## [1] 60.04209
```

```r
sd(income)
```

```
## [1] 11.88252
```

```r
quantile(income)
```

```
##       0%      25%      50%      75%     100%
## 34.75845 49.45450 60.04209 71.42238 80.84025
```

2) Set your random seed to be 101. Create a 100x5 matrix of random realizations from the standard normal distribution.

(a) Report the column means. Demonstrate how you would do this using

(b) the apply function (ii) vector / matrix arithmetic

```r
set.seed(101)
randomMatrix <- matrix(rnorm(500, 0, 1), nrow = 100, ncol = 5)
```

Using apply

```r
apply(randomMatrix, 2, mean)
```

```
## [1] -0.037191100 -0.042002372  0.002070399 -0.025466513 -0.211317178
```

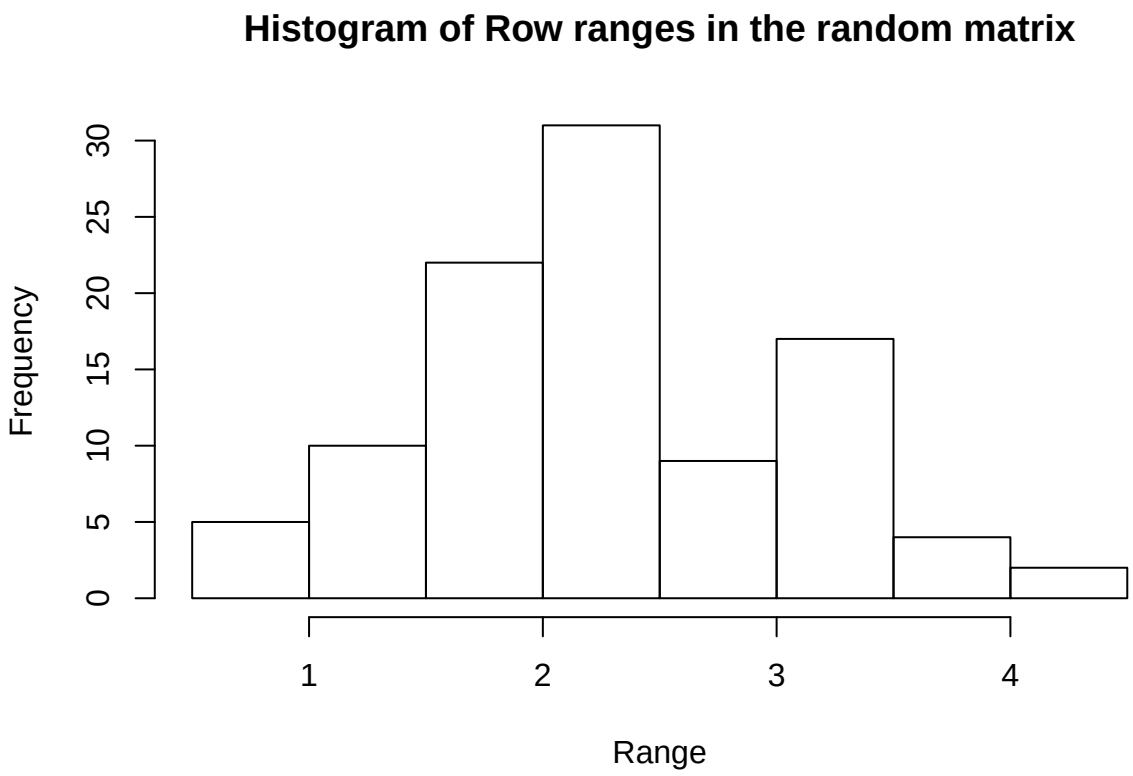
Using colmeans (ii)

```r
colMeans(randomMatrix)
```

```
## [1] -0.037191100 -0.042002372  0.002070399 -0.025466513 -0.211317178
```

(b)Make a histogram of row ranges in the matrix

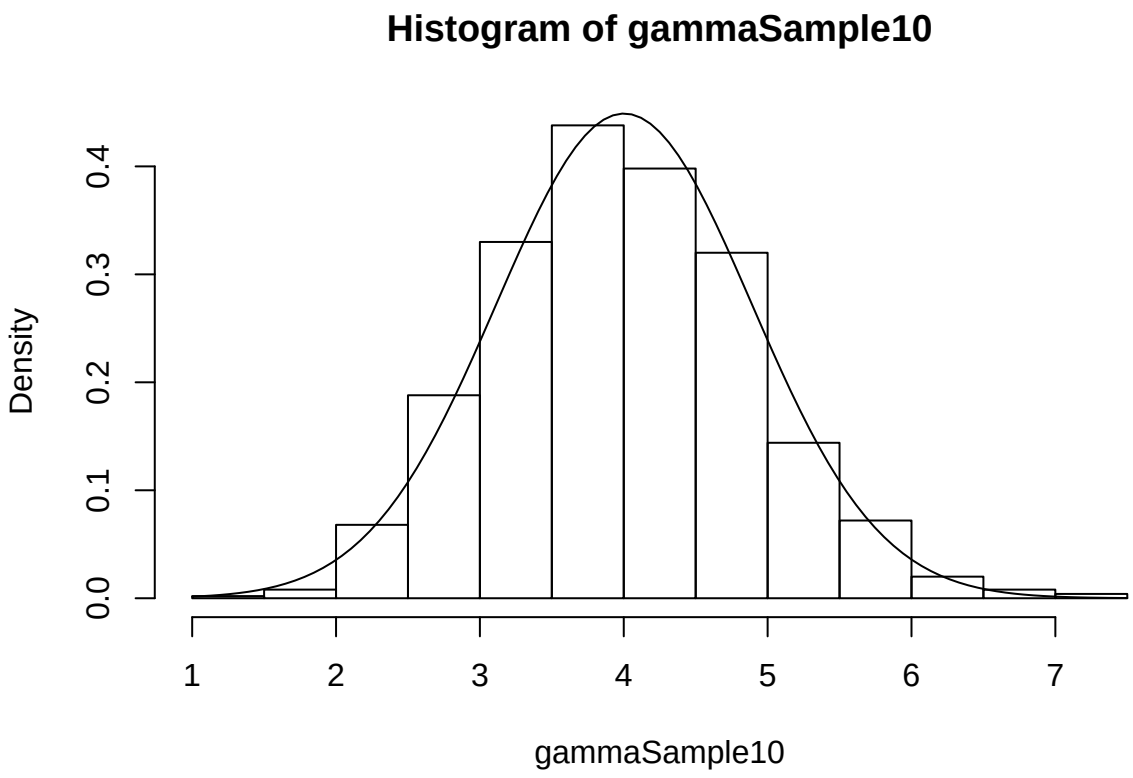
```r
hist(apply(randomMatrix, 1, function(x) max(x) - min(x)), xlab = "Range", ylab = "Frequency", main = "H
```



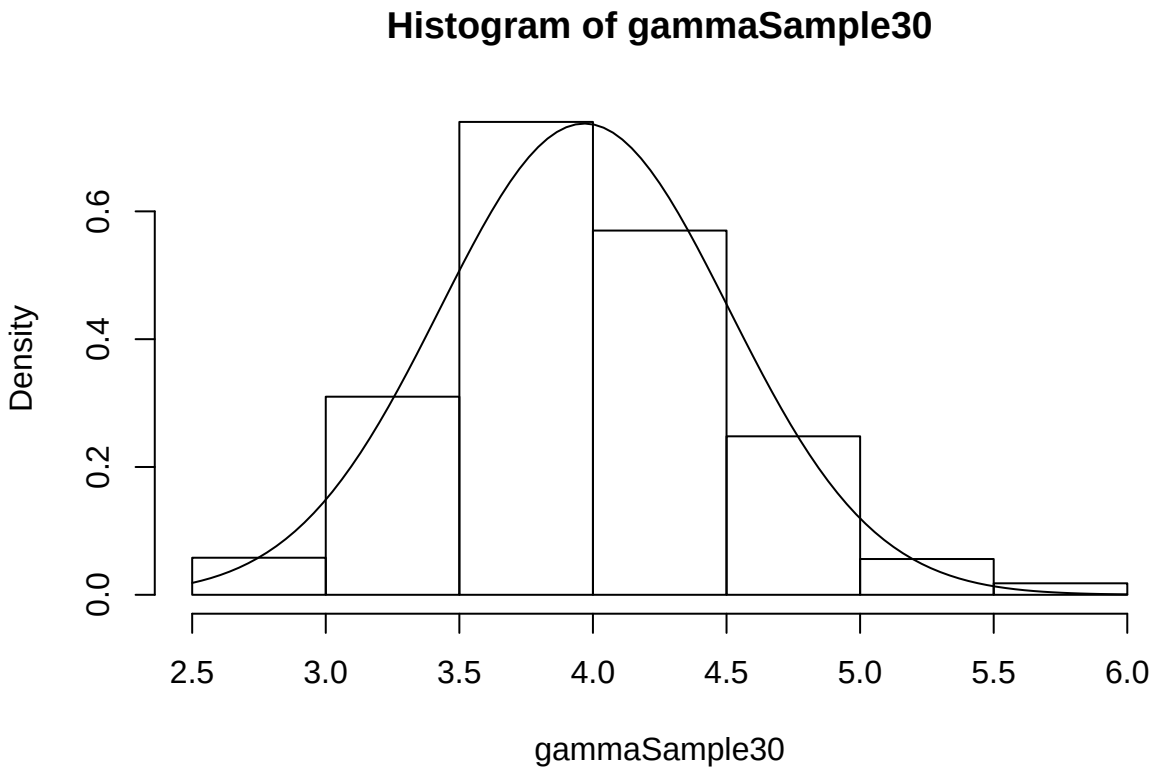
Analysis: The histogram shows a slightly right skewed nature of row ranges. Most of the rows have a range between 2.0 and 2.5

3) Consider the gamma distribution with shape and scale parameters both equal 2; Simulate samples of size n = 10,30,90 from this distribution, repeating B = 1000 times. For each simulated dataset, compute the sample mean. For each sample size draw a probability histogram. Overlay the normal curve that would apply if the central limit theorem could be assumed to hold.

```
gammaDist1000 <- rgamma(1000, shape = 2, scale = 2)

gammaSample10 <- replicate(1000, {
    sampleMeans10 <- sample(gammaDist1000, 10)
    mean(sampleMeans10)
})

gammaSample30 <- replicate(1000, {
    sampleMeans30 <- sample(gammaDist1000, 30)
    mean(sampleMeans30)
})

gammaSample90 <- replicate(1000, {
    sampleMeans90 <- sample(gammaDist1000, 90)
    mean(sampleMeans90)
})

hist(gammaSample10, probability = TRUE)
curve(dnorm(x, mean=mean(gammaSample10), sd=sd(gammaSample10)), add=TRUE)
```




```
hist(gammaSample30, probability = TRUE)
curve(dnorm(x, mean=mean(gammaSample30), sd=sd(gammaSample30)), add=TRUE)
```



```r
hist(gammaSample90, probability = TRUE)
curve(dnorm(x, mean=mean(gammaSample90), sd=sd(gammaSample90)), add=TRUE)
```