Lab_2

Abhinav Garg

1/19/2017

Part 1

1. Load the Data

```
setwd("/Users/abhi/Documents/UW/Courses/Winter_Quarter_17/INFX_573/Week_3/Lab_2
")
ratings <- read.csv("ratings.csv")
movies <- read.csv("movie.titles.csv")
dim(ratings)
```

```
## [1] 100004      5
```

```
users <- unique(ratings$userId)
d <- ratings[users,]
dim(d)
```

```
## [1] 671   5
```

```
#Mean
mean(ratings$rating)
```

```
## [1] 3.543608
```

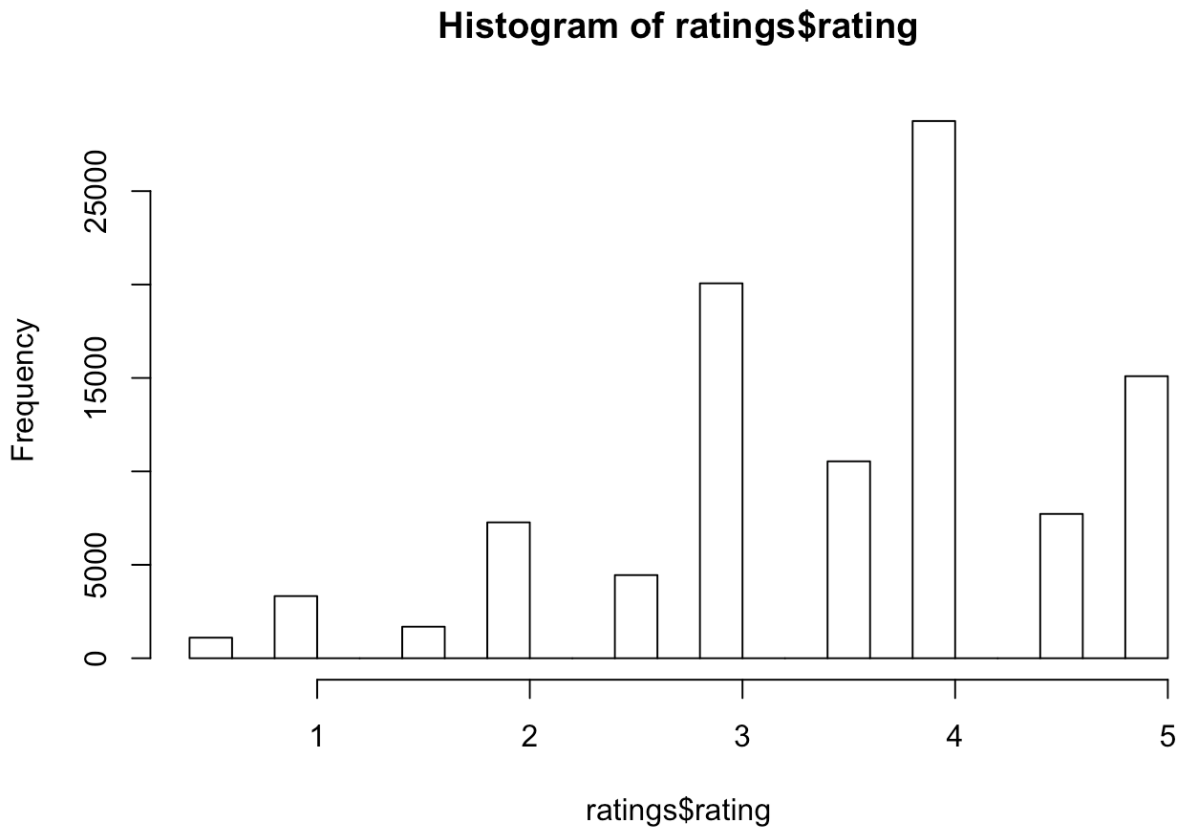
```
#Median
median(ratings$rating)
```

```
## [1] 4
```

```
#Mode
Mode <- function(x) {
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}
Mode(ratings$rating)
```

```
## [1] 4
```

```
#Plot histogram

hist(ratings$rating)
```



2. Link the two datasets using movieid

```
ix <- match(ratings$movieId, movies$movieId)

head(ratings$movieId)
```

```
## [1]   31 1029 1061 1129 1172 1263
```

```
head(movies$movieId[ix])
```

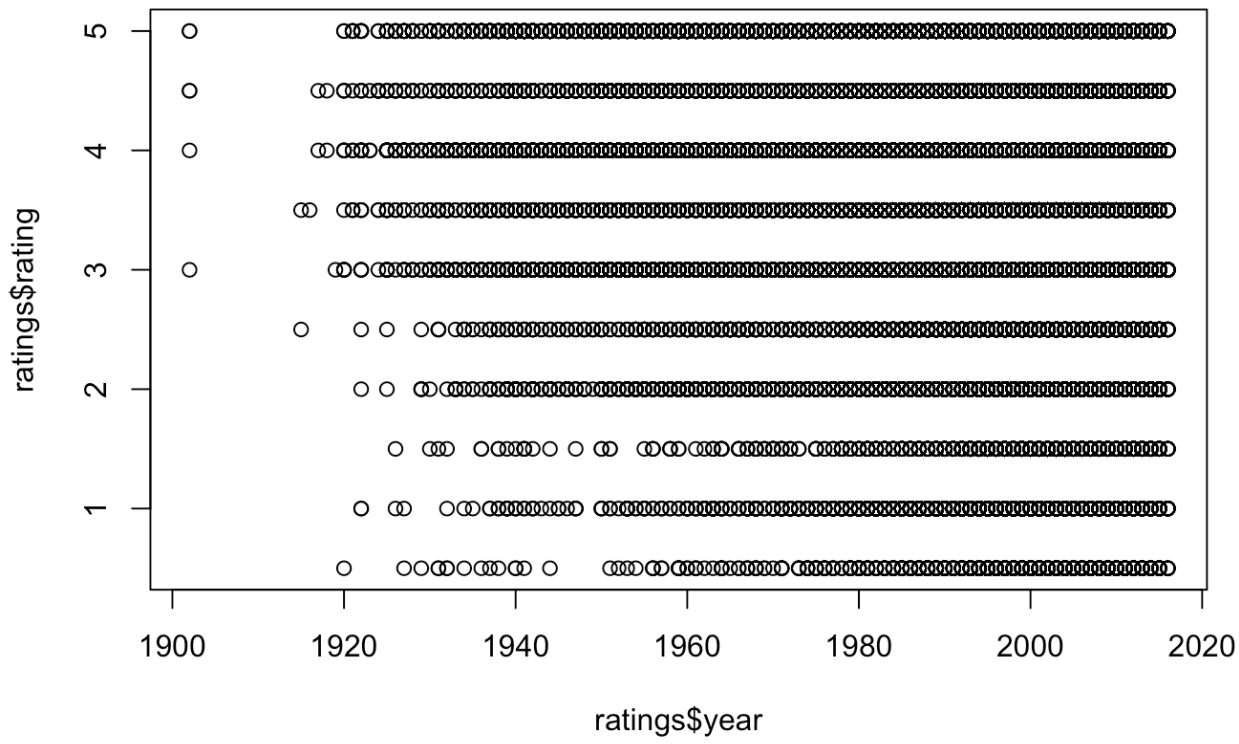
```
## [1]   31 1029 1061 1129 1172 1263
```

```
temp <- merge(ratings, movies, by = 'movieId')
head(temp)
```

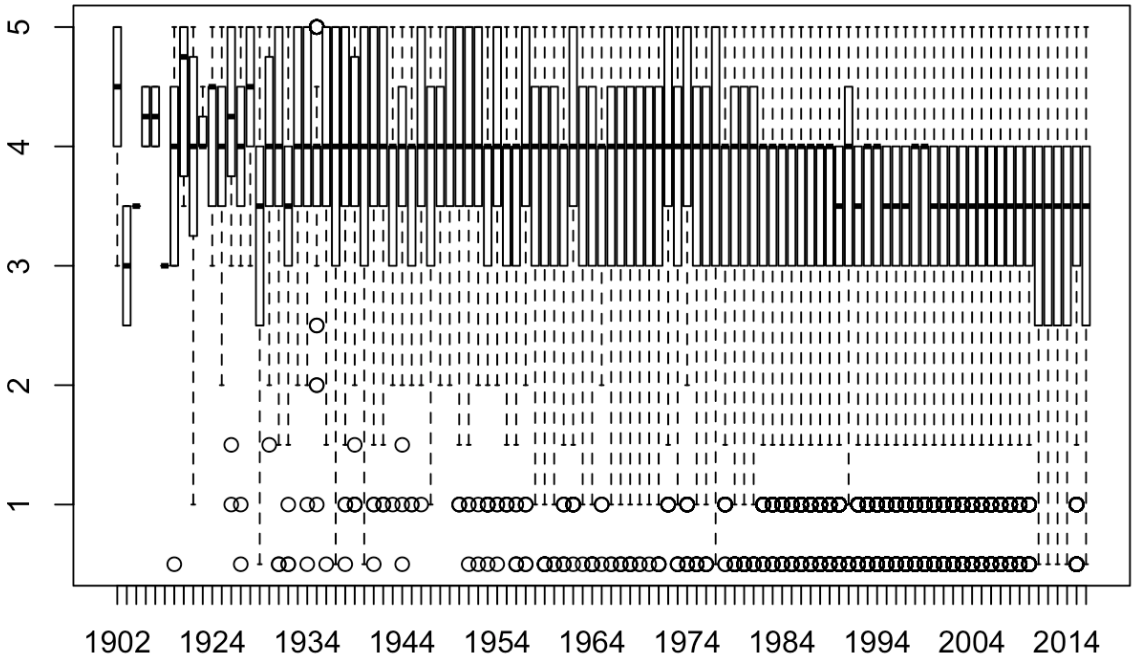
```
##   movieId userId rating year.x                                      genre
## 1       1    136    4.5   1995 Adventure|Animation|Children|Comedy|Fantasy
## 2       1     43    4.0   1995 Adventure|Animation|Children|Comedy|Fantasy
## 3       1    428    5.0   1995 Adventure|Animation|Children|Comedy|Fantasy
## 4       1    241    3.0   1995 Adventure|Animation|Children|Comedy|Fantasy
## 5       1    390    4.0   1995 Adventure|Animation|Children|Comedy|Fantasy
## 6       1    329    5.0   1995 Adventure|Animation|Children|Comedy|Fantasy
##              title                                       genres year.y
## 1 Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy   1995
## 2 Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy   1995
## 3 Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy   1995
## 4 Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy   1995
## 5 Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy   1995
## 6 Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy   1995
```

3. Exploring Relationships I

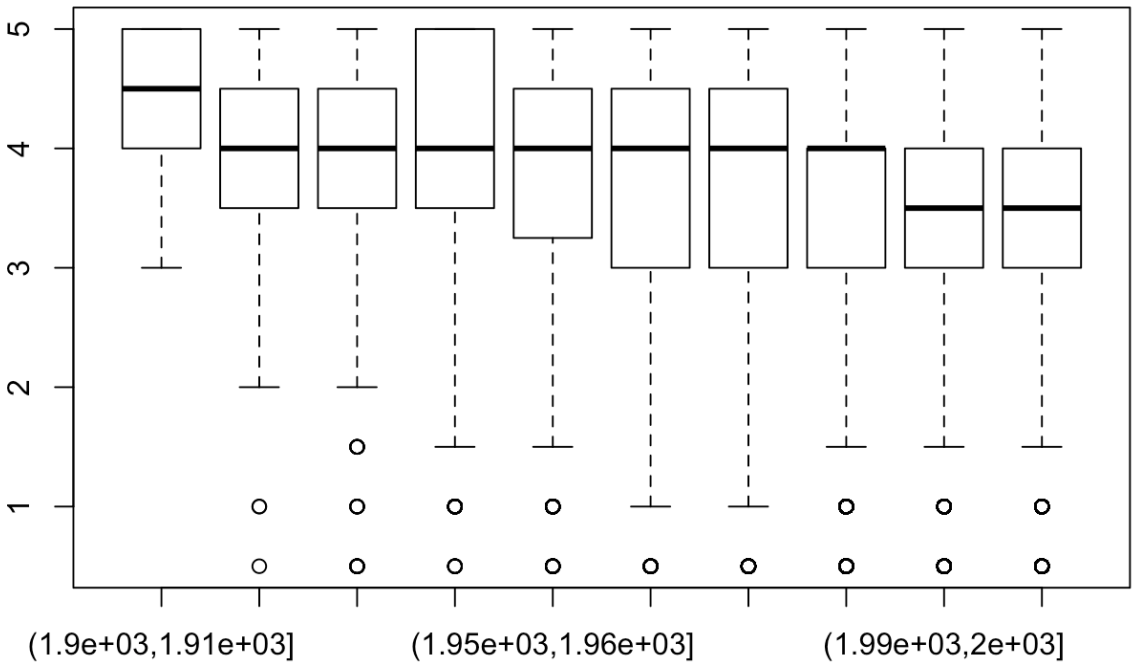
```
plot(ratings$year, ratings$rating)
```




```
boxplot(rating ~ year, data = ratings)
```

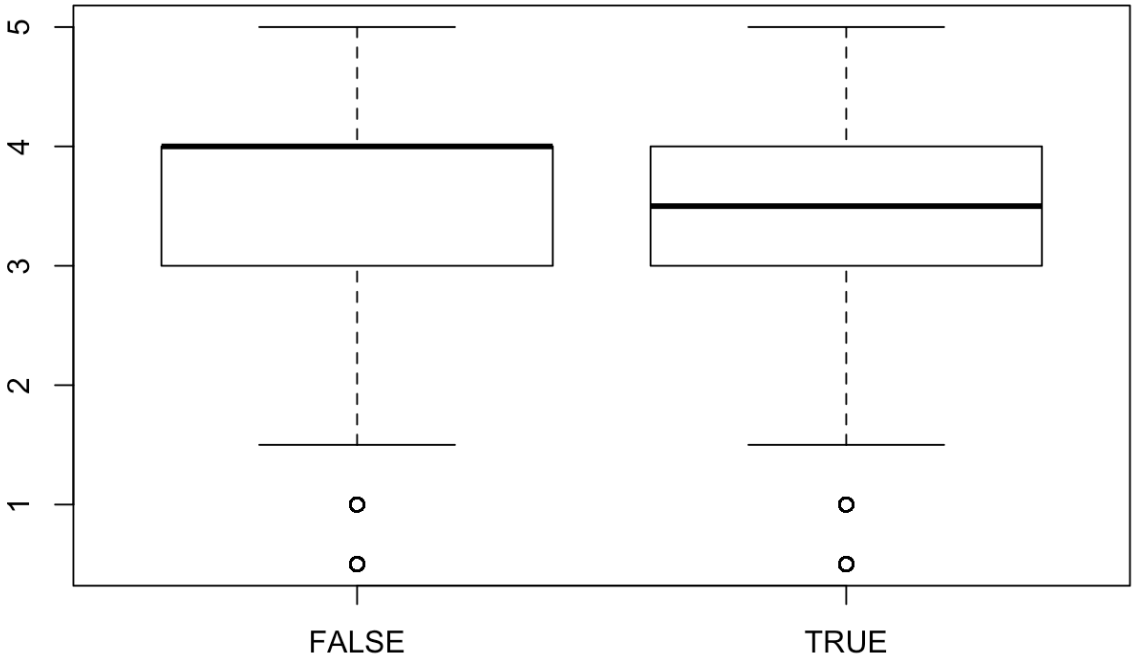



```
boxplot(rating ~ cut(year,breaks = 10), data = ratings)
```



4. Exploring Relationships II

```
# a) Do the ratings vary by genre? Create a 'comedy' column and draw a box plot
 of ratings for comedy versus others:

ratings$comedy <- rep(F, nrow=ratings)
ratings$comedy[grep("comedy",ratings$genre,ignore.case = T)] <- T

boxplot(rating ~ comedy, data=ratings)
```




```
# b) Run a t-test to see if the differences in ratings for comedy versus non-co
medy

t.test(ratings$rating, ratings$comedy)
```

```
##
##  Welch Two Sample t-test
##
## data:  ratings$rating and ratings$comedy
## t = 859.33, df = 140320, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.156148 3.170579
## sample estimates:
## mean of x mean of y
## 3.5436083 0.3802448
```

5. Extra credit

```
sorted_ratings <- ratings[order(-ratings$rating),]
movie_popularity <- aggregate(ratings$rating,by= list(unique.values = sorted_ra
tings$movieId),FUN = length)

movie_popularity <- movie_popularity[order(-movie_popularity$x),]

top_ten_id<- head(movie_popularity$unique.values, n = 10)
top_ten_id
```

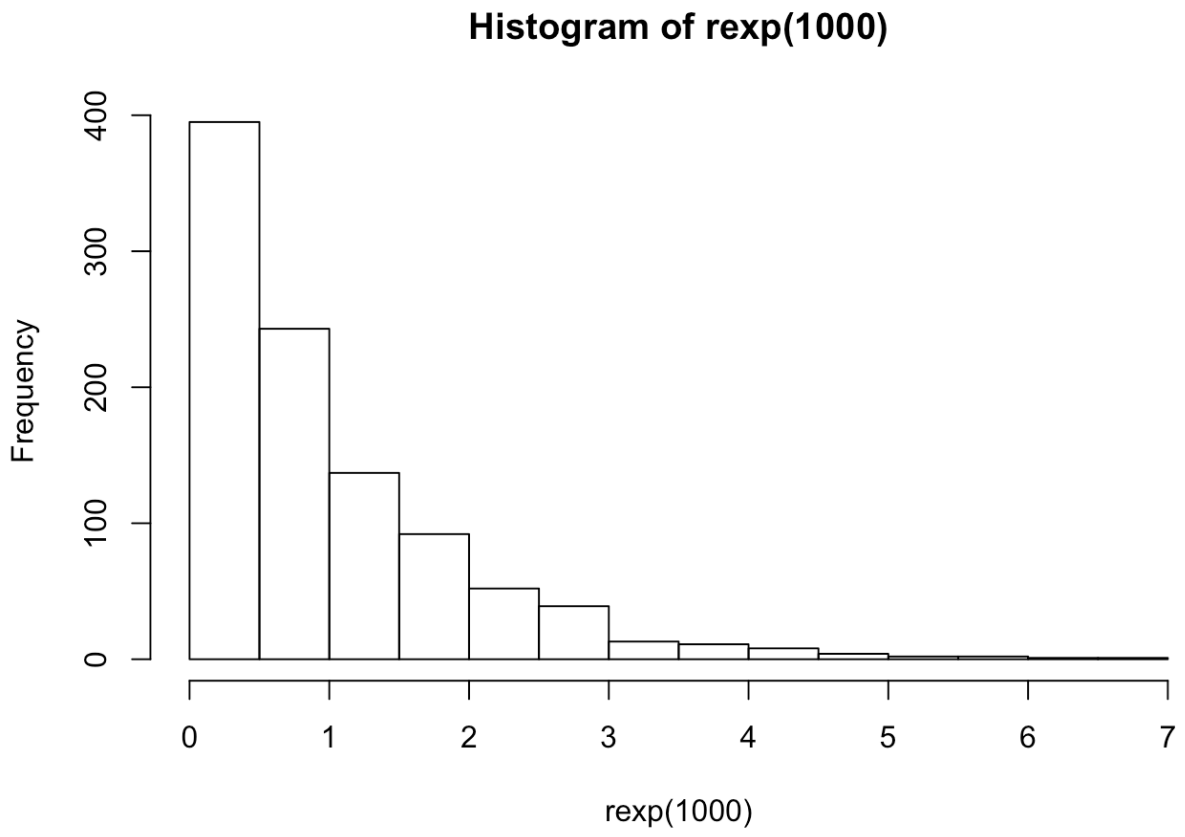
```
##  [1]  356  296  318  593  260  480 2571    1  527  589
```

```
movies[movies$movieId %in% top_ten_id,]$title
```

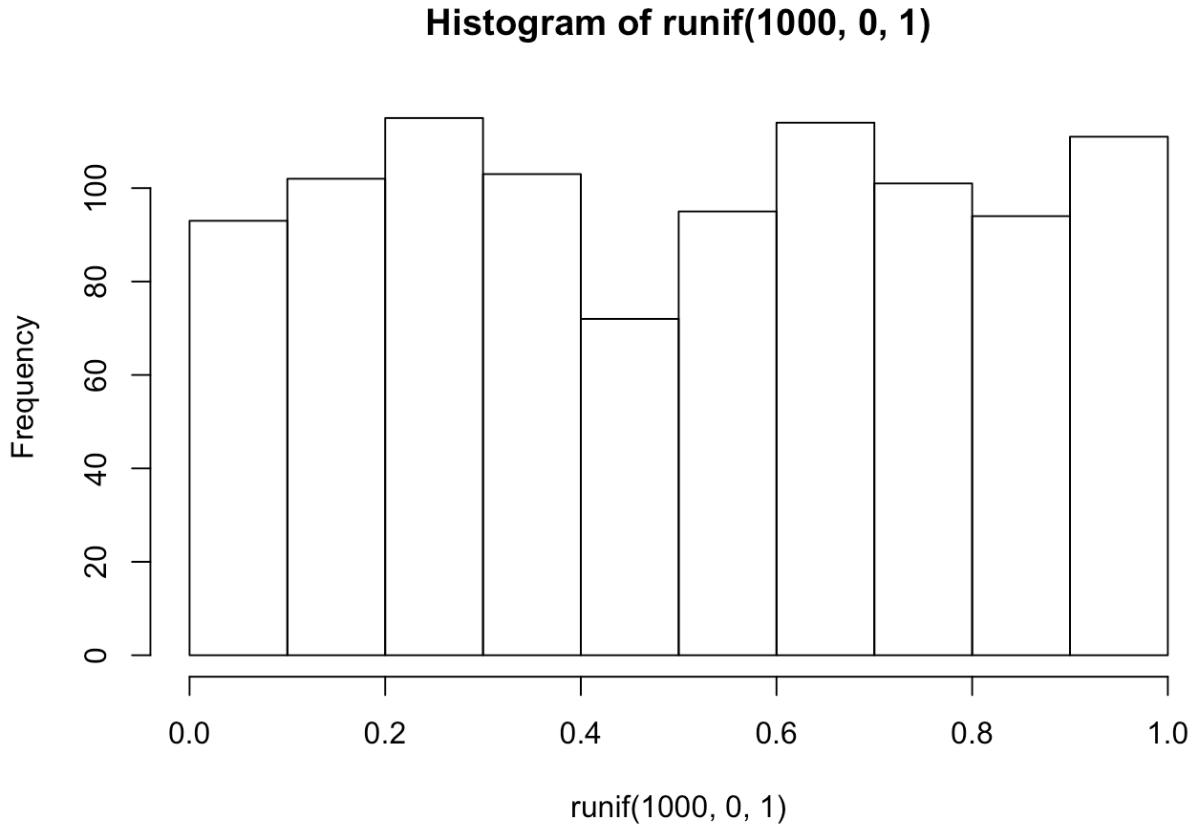
```
##  [1] Toy Story (1995)
##  [2] Star Wars: Episode IV - A New Hope (1977)
##  [3] Pulp Fiction (1994)
##  [4] Shawshank Redemption, The (1994)
##  [5] Forrest Gump (1994)
##  [6] Jurassic Park (1993)
##  [7] Schindler's List (1993)
##  [8] Terminator 2: Judgment Day (1991)
##  [9] Silence of the Lambs, The (1991)
## [10] Matrix, The (1999)
## 9123 Levels: ¡Three Amigos! (1986) ... Zulu (2013)
```

Part 2

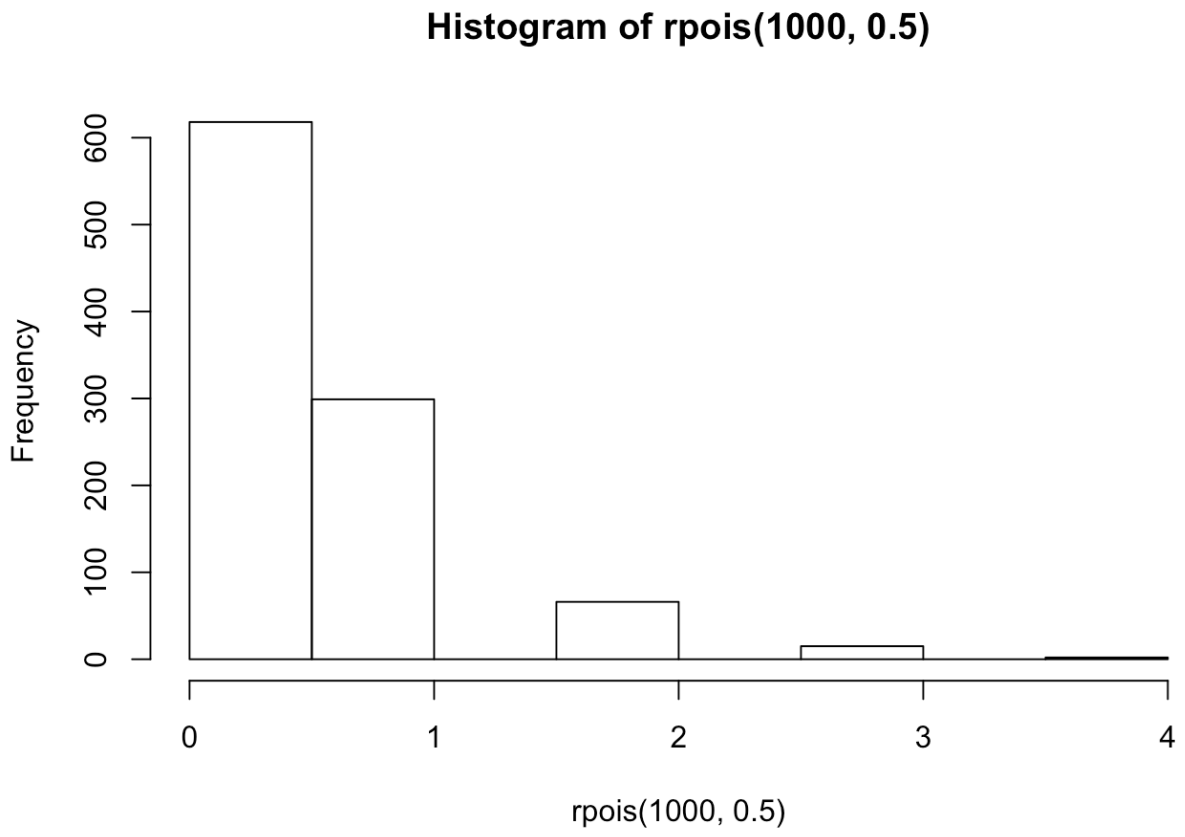
```
# Examine several distributions
hist(rexp(1000))
```



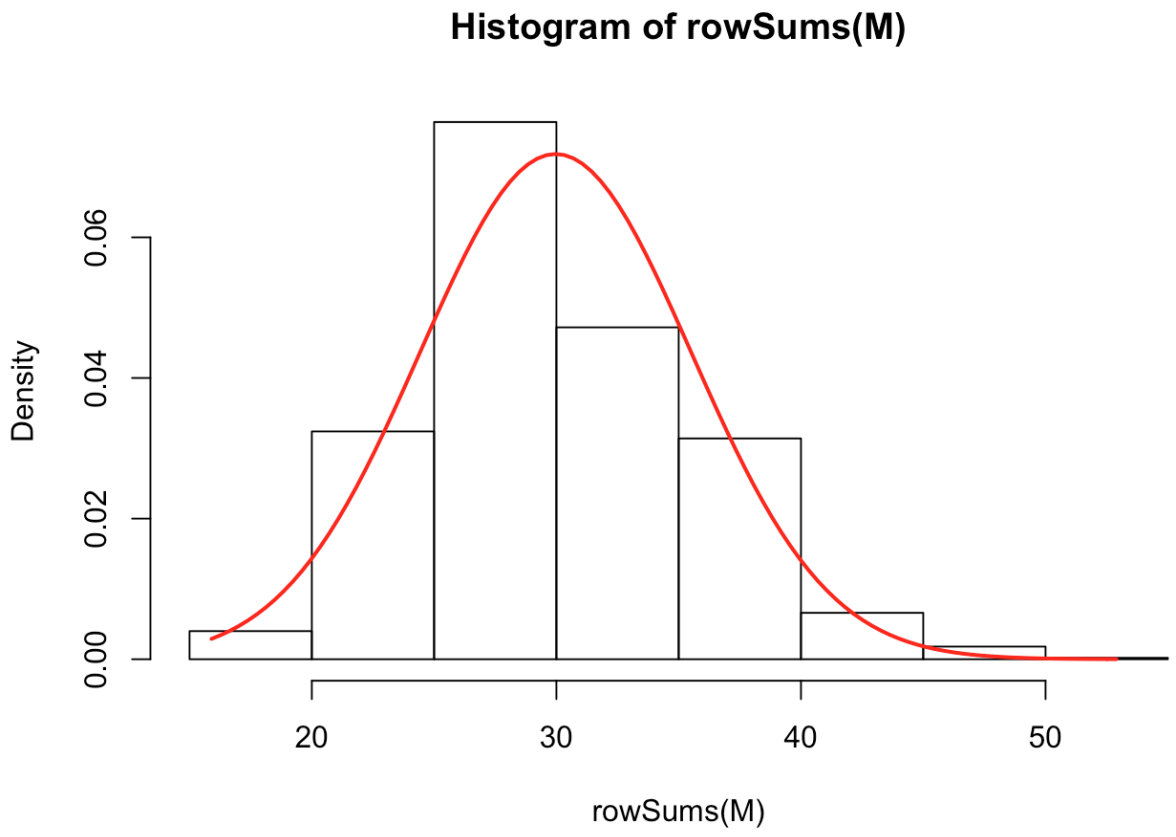

```
hist(runif(1000,0,1))
```



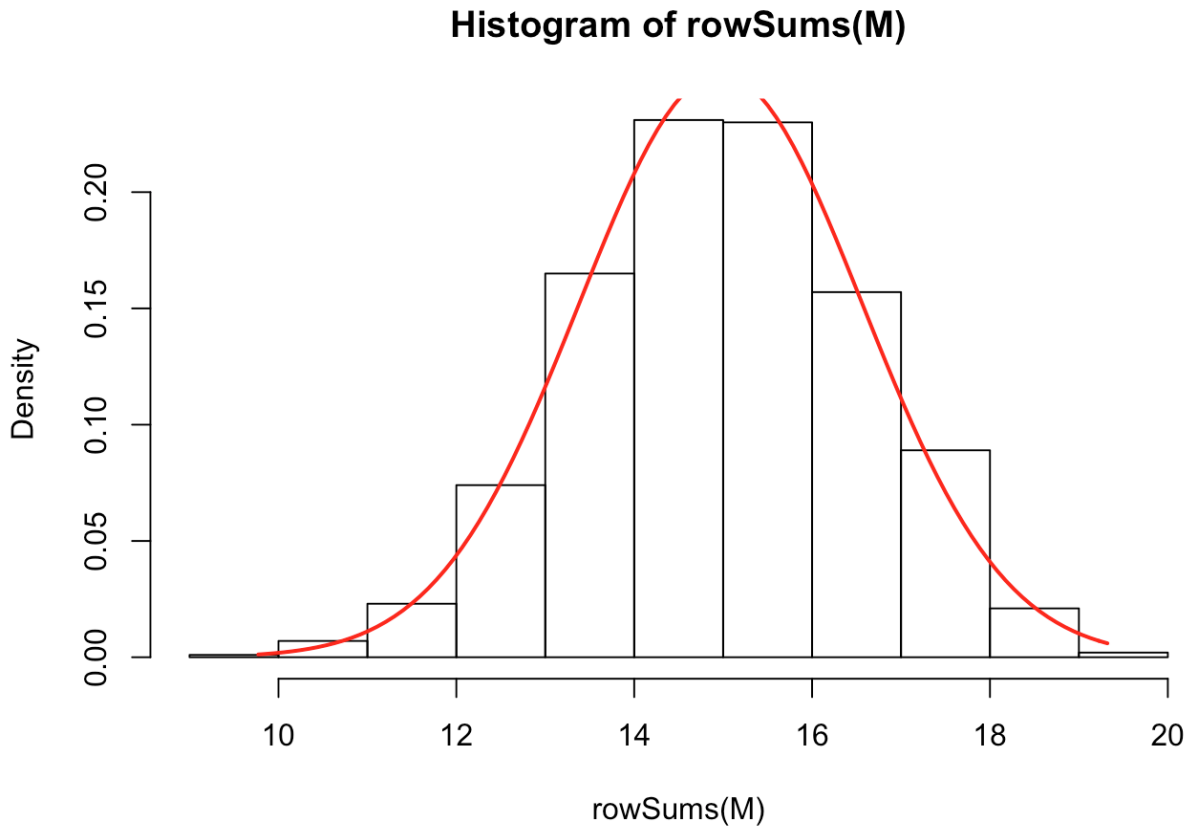

```
hist(rpois(1000,0.5))
```



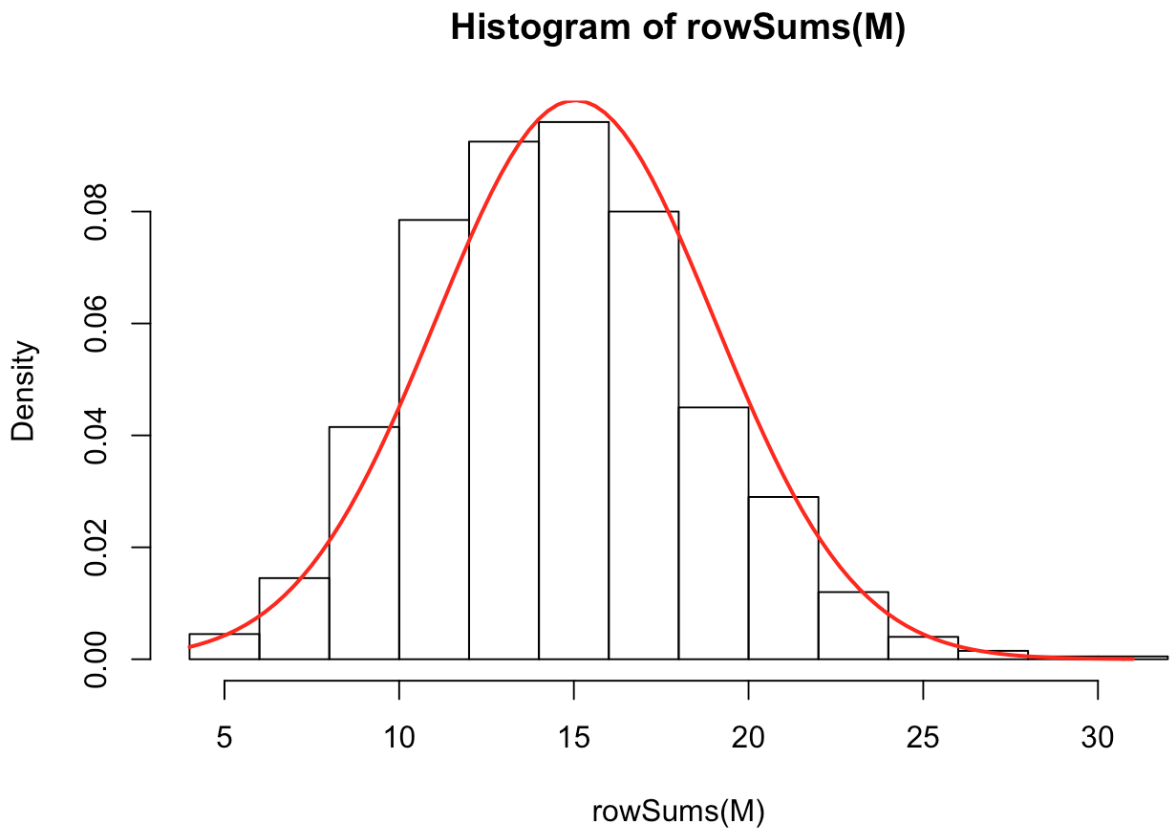

```
# Look at the distributions of sums of these samples

N <- 1000 # number of exponential draws
n.samp <- 30 # number of sums to take
M <- matrix(NA, nrow=N, ncol=n.samp) # create an empty matrix to fill with samp
les
for(j in 1:n.samp) M[,j] <- rexp(N) #generate the samples
hist(rowSums(M), freq = F) # plot a histogram of the sums across rows of our ma
trix M

curve(dnorm(x, mean(rowSums(M)), sd(rowSums(M))), min(rowSums(M)),
max(rowSums(M)), add=T, col="red", lwd=2)
```



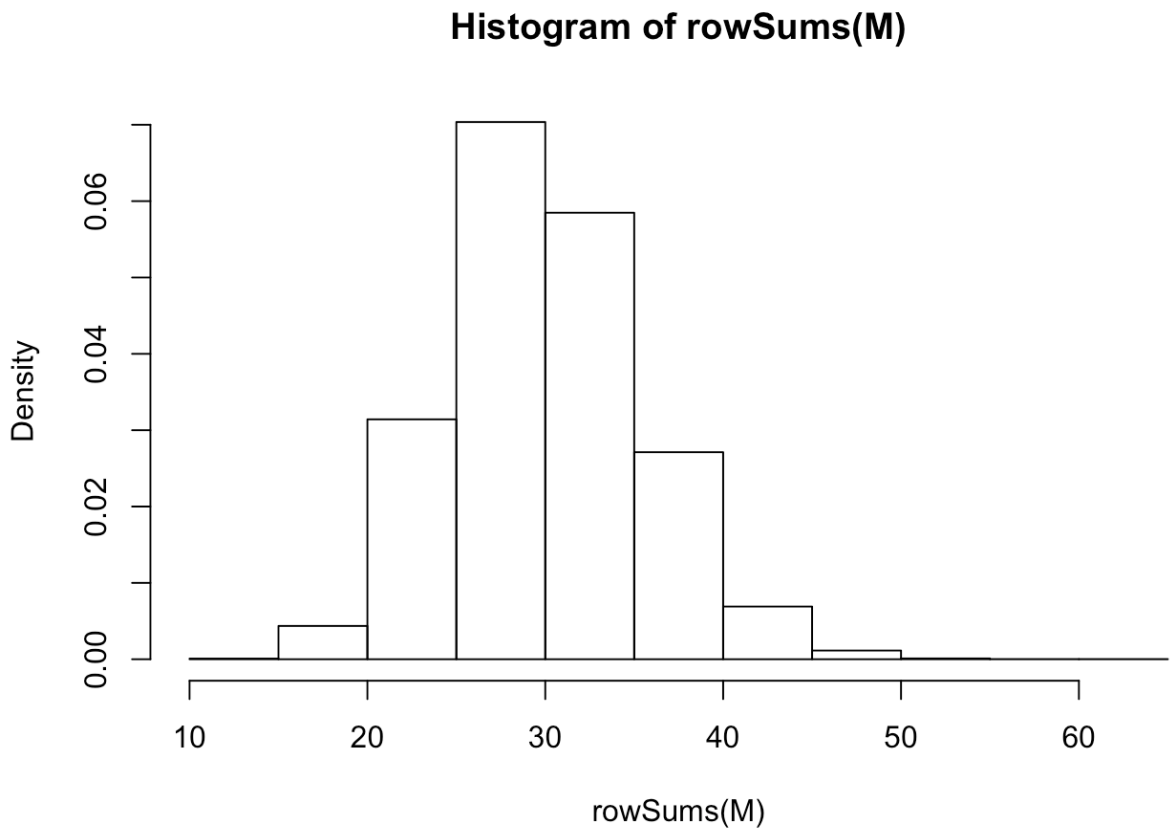

```
# For Uniform

for(j in 1:n.samp) M[,j] <- runif(N,0,1) #generate the samples
hist(rowSums(M), freq = F) # plot a histogram of the sums across rows of our ma
trix M

curve(dnorm(x, mean(rowSums(M)), sd(rowSums(M))), min(rowSums(M)),
max(rowSums(M)), add=T, col="red", lwd=2)
```



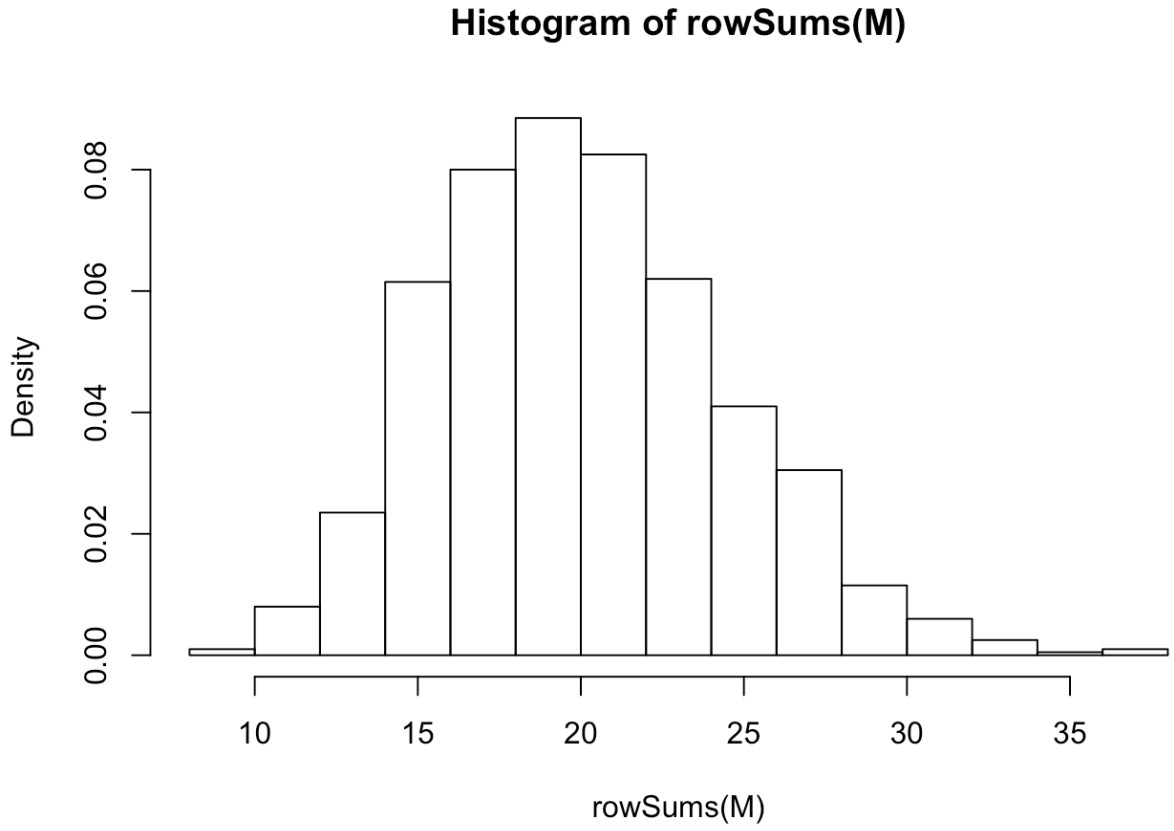

```
# For Poisson

for(j in 1:n.samp) M[,j] <- rpois(N,0.5) #generate the samples
hist(rowSums(M), freq = F) # plot a histogram of the sums across rows of our ma
trix M

curve(dnorm(x, mean(rowSums(M)), sd(rowSums(M))), min(rowSums(M)),
max(rowSums(M)), add=T, col="red", lwd=2)
```




```
# Further Explorations
# by increasing the number of samples we see an increase in the number of bins
N <- 10000 # number of exponential draws
n.samp <- 30 # number of sums to take
M <- matrix(NA, nrow=N, ncol=n.samp) # create an empty matrix to fill with samp
les
for(j in 1:n.samp) M[,j] <- rexp(N) #generate the samples
hist(rowSums(M), freq = F)
```




```
# by changing the number of samples
# changing the number of sums to take
N <- 1000 # number of exponential draws
n.samp <- 20 # number of sums to take
M <- matrix(NA, nrow=N, ncol=n.samp) # create an empty matrix to fill with samp
les
for(j in 1:n.samp) M[,j] <- rexp(N) #generate the samples
hist(rowSums(M), freq = F)
```



}

By decreasing the number of sums we see that the density is more focussed towards the center