ISL Fall 2015 Assignment I. (70+30) pts.

NAME(s):

LOKESHWAR REDDY INJA (16207239/ LIRK9)
SANTHOSH KUMAR GATTU (16211118/ SG6N6)
SANTOSH KUMAR (16207971/ SK7Z9)

You may submit this assignment in groups of upto 3 each. Write your names on this sheet and include it as the cover page for your submission.
The objective of this assignment is to practice using R. The last section of Chapter 2 of the textbook includes a short introductory tutorial on R – the corresponding code is provided on the textbook website (http://www-bcf.usc.edu/~gareth/ISL/). Your submission should include both your code as well the answers to the questions.
Electronic submission on Blackboard is due latest by 11 pm on Wednesday, Sep 16th. Upload only one submission per group. Submissions received after the deadline will be graded only for effort for a maximum of 70% of the total grade (Refer to class syllabus for detailed grading policy). **State any assumptions you make, justify your answers, show intermediate steps and explain your results for maximum credit.** All answers should be in your own words with any sources you refer to cited at the appropriate places. Any knowledge you acquire from the Internet should be written in your own words and be appropriately referenced. Copying and pasting from the Internet, each other or any other source will not count as your effort (Refer to class syllabus for detailed policy on plagiarism).

Q1. (10) Choose any three probability distributions.

a. For each distribution, generate a random sample of size s and plot its histogram.
b. Repeat 'a' for increasing values of s. How does the shape of the histogram change with increasing s?
c. For each distribution, generate n random samples of size s each and compute the mean m for each sample.
d. Plot a histogram for the values of m. How does the shape of the histogram change with increasing n?

Q2. (10) Chapter 2: Question 7.

Q3. (10) Chapter 2: Parts c-g of Question 10.

Choose any three probability distributions.

a. For each distribution, generate a random sample of size s and plot its histogram.

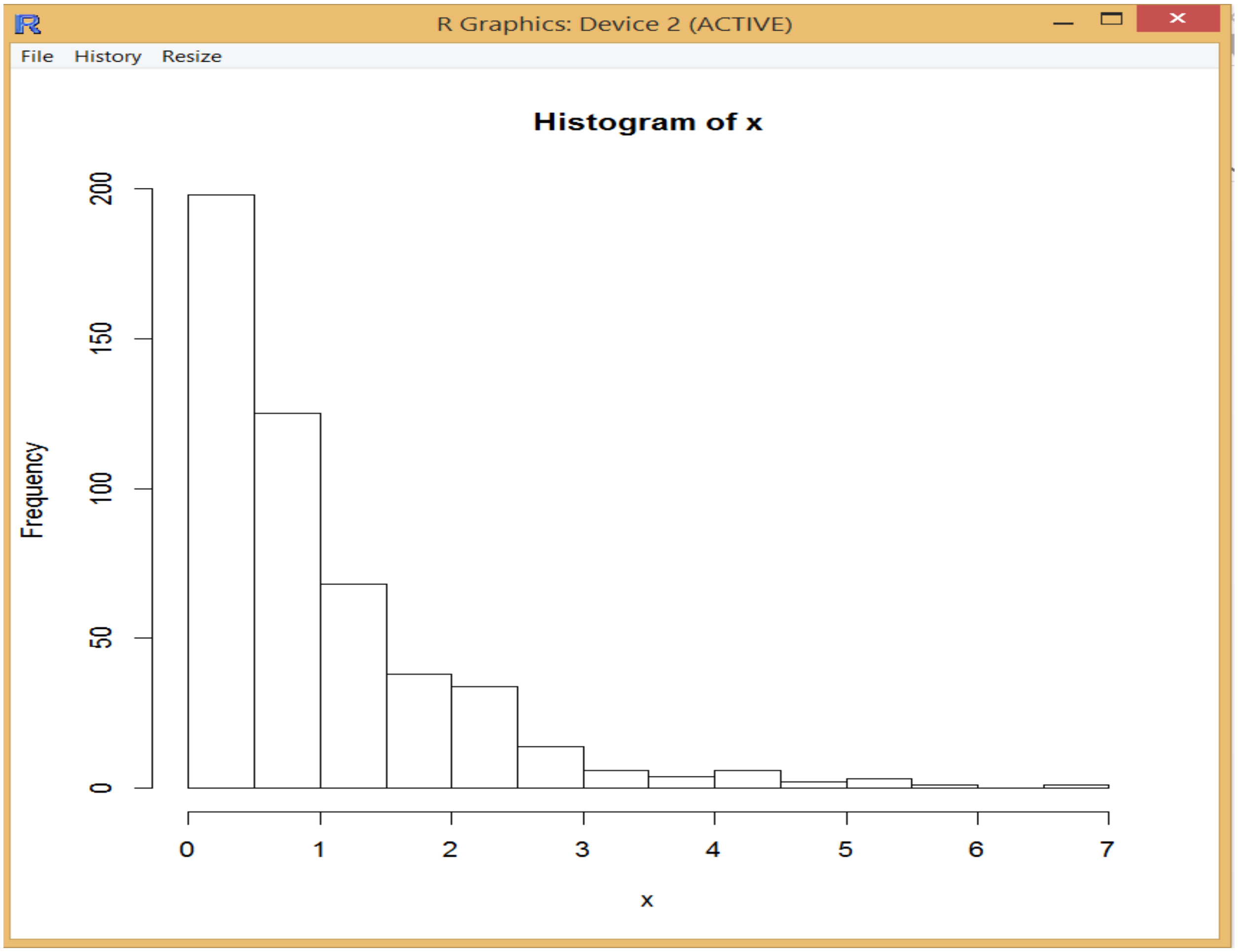
Sol

Exponential Distribution:

Code:

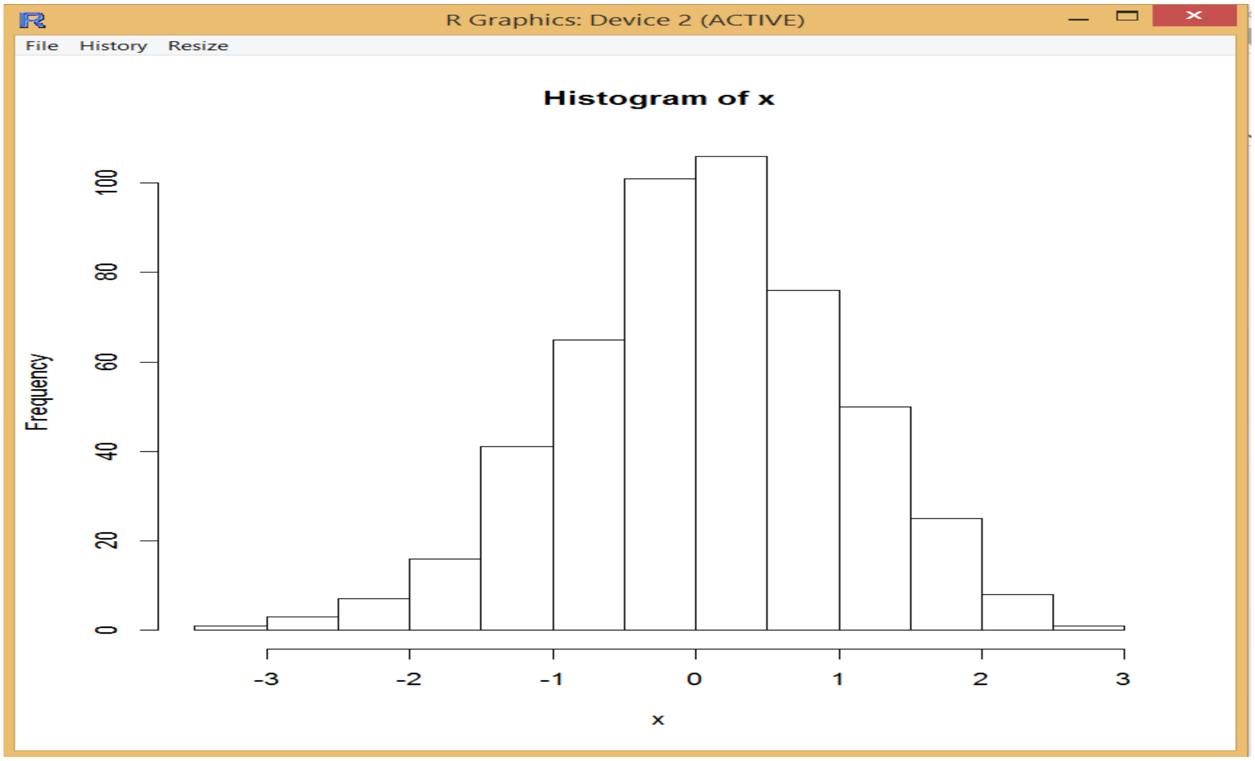
```
> x <- rexp(500, rate=1)
> hist(x)
```

Histogram:



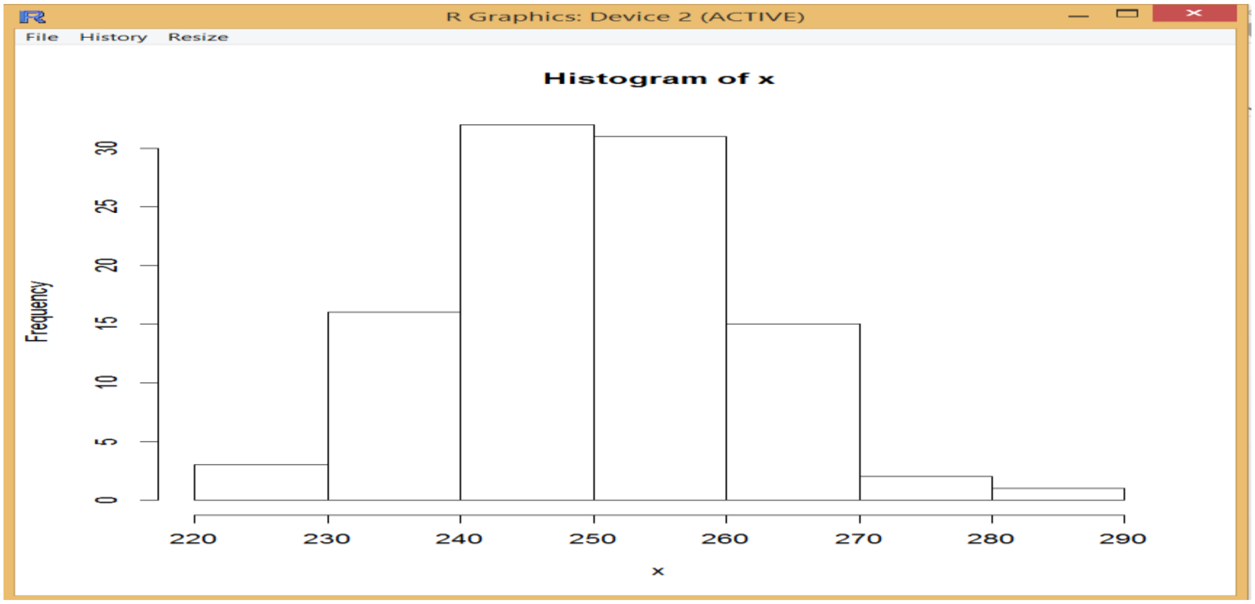
Normal Distribution:

```
X <- rnorm(500)
Hist(x)
```



Binomial Distribution:

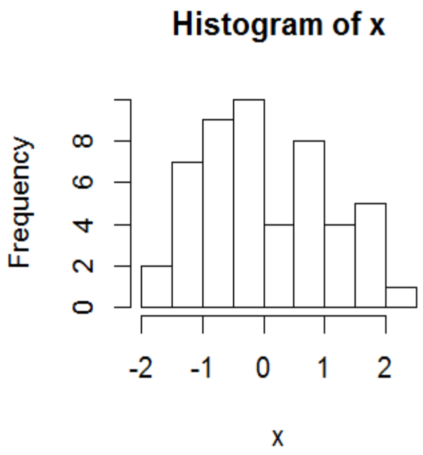
```
X <- rbinom(100, 500, 0.5)
hist(x)
```



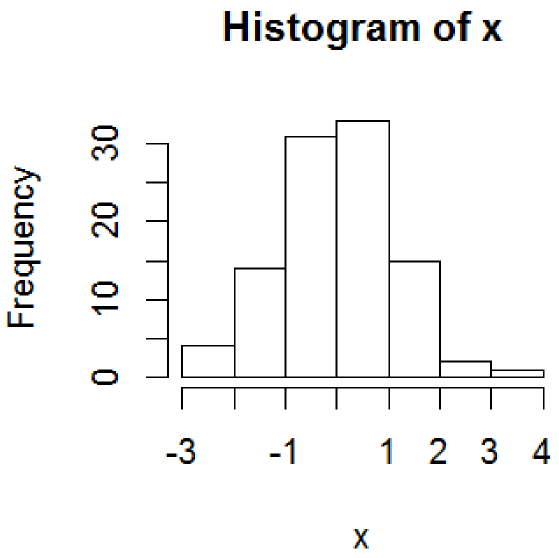
b. Repeat 'a' for increasing values of s. How does the shape of the histogram change with increasing s?

Random Normal distribution

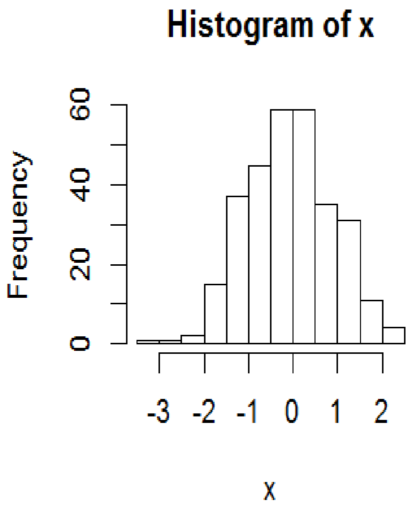
rnorm(50)



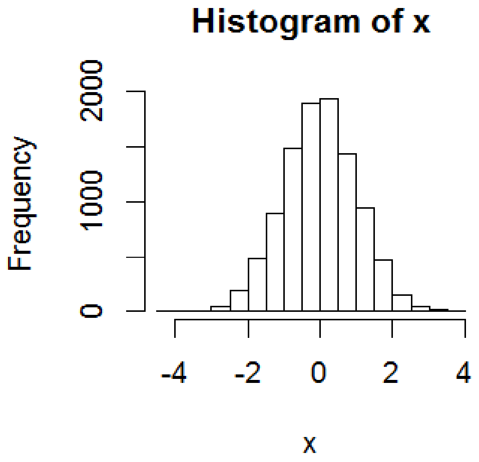
rnorm(100)



rorm(300)



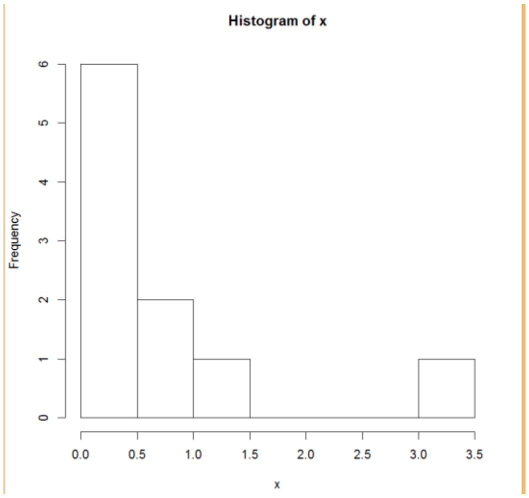
rnorm(10000)



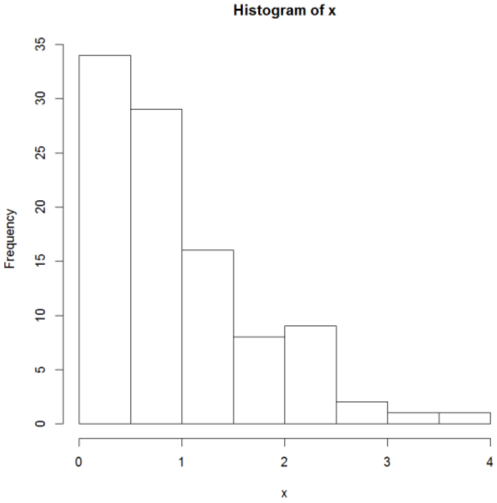
From the above plots, it is clear that with increase in number of samples 's' variance is decreasing and we get more precise Normal distribution curve.

Exponential Distribution:

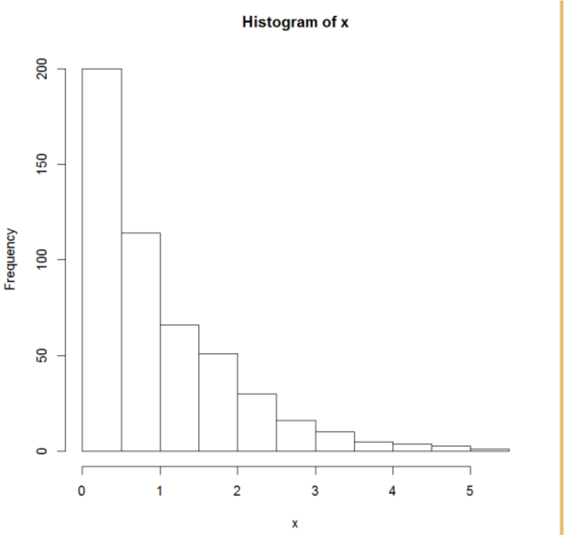
```
> x <- rexp(10, rate=1)
> hist(x)
```



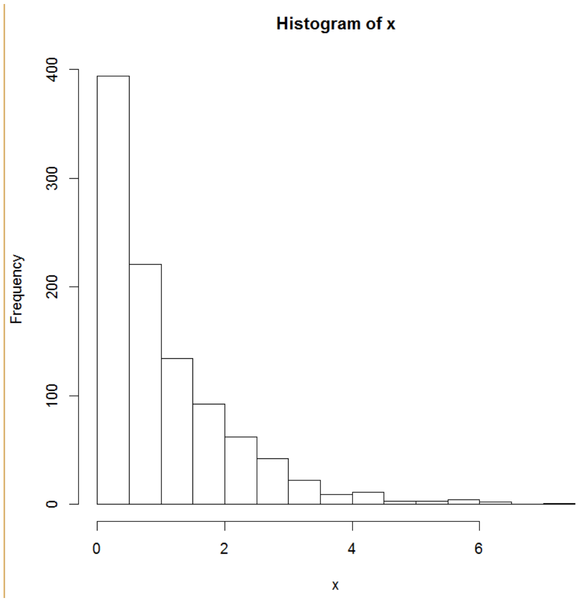
```
> x <- rexp(100, rate=1)
> hist(x)
```



```
> x <- rexp(500, rate=1)
> hist(x)
```



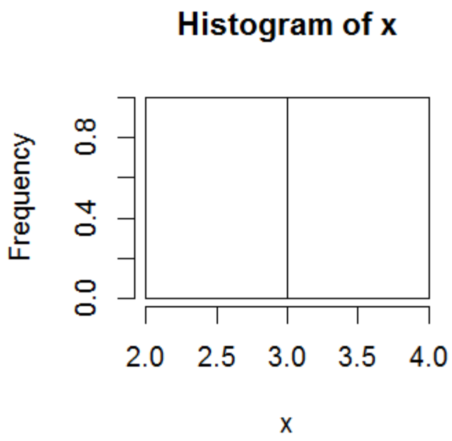
```
x <- rexp(1000, rate=1)
> hist(x)
```



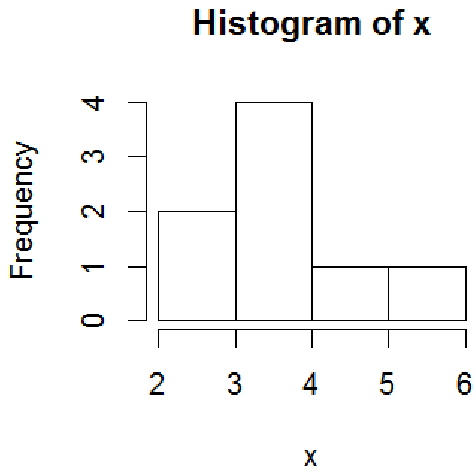
As the sample space‘s’ increases, the exponential curve becomes more steep.

Binomial Distribution:

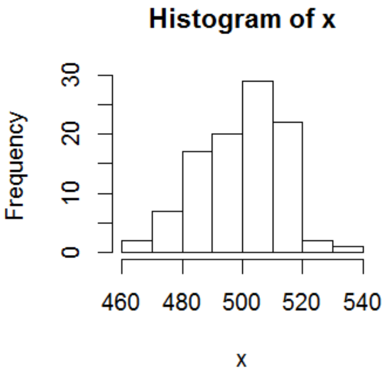
X <- rbinom(2,10,0.5)



x <- rbinom(8,10,0.5)



X <- rbinom(100,1000,0.5)



As the sample space‘s’ increases Binomial distribution follows Normal distribution curve

c. For each distribution, generate n random samples of size s each and compute the mean m for each sample.

Sol.

Normal Distribution

Code:

```
➢ ynorm = NULL
➢ for(i in 1:300){ ynorm[i]=mean(rnorm(i))}
➢ for(i in 1:300){ print(ynorm1[i])}
```

Exponential Distribution:

Code:

```
➢ yexp = NULL
➢ for(i in 1:300){ yexp[i]=mean(rexp(i,rate=1))}
➢ for(i in 1:300){ print(yexp1[i])}
```

Binomial Distribution:

Code:

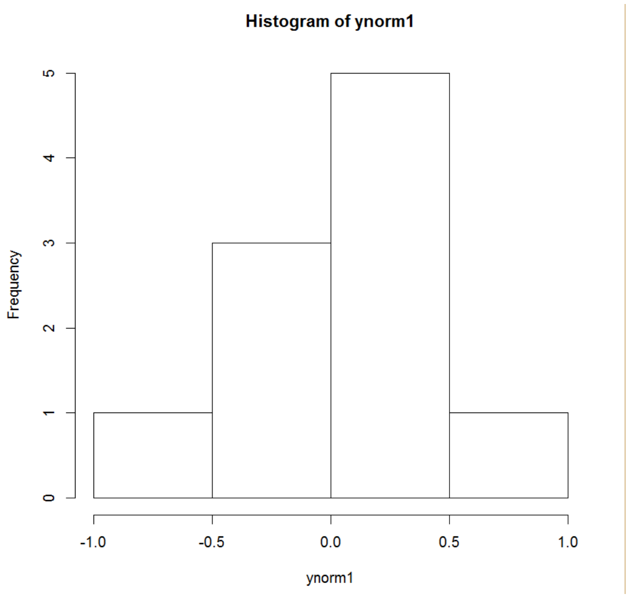
```
➢ ybnm = NULL
➢ for(i in 1:300){ ybnm[i]=mean(rbinom(i,i+50,0.5))}
➢ for(i in 1:300){ print(ynorm1[i])}
```

d. Plot a histogram for the values of m. How does the shape of the histogram change with increasing n?

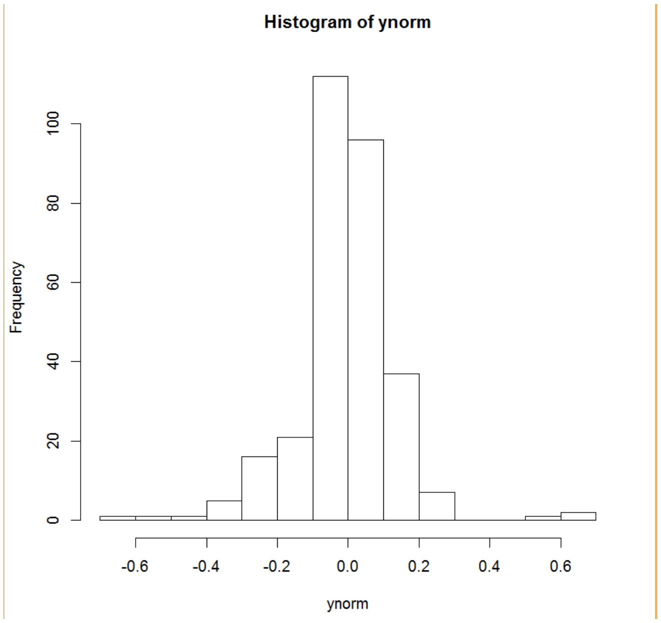
Normal Distribution:

Code:

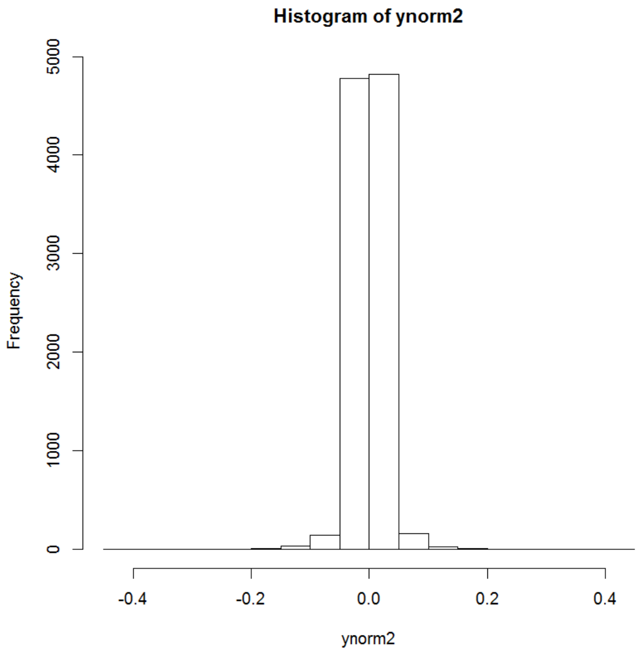
```
> ynorm1 = NULL
> for(i in 1:10){ ynorm1[i]=mean(rnorm(i))}
> hist(ynorm1)
```



```
> ynorm = NULL
> for(i in 1:300){ ynorm[i]=mean(rnorm(i))}
> hist(ynorm)
```



```
> ynorm2 = NULL
> for(i in 1:10000){ ynorm2[i]=mean(rnorm(i))}
> hist(ynorm2)
```

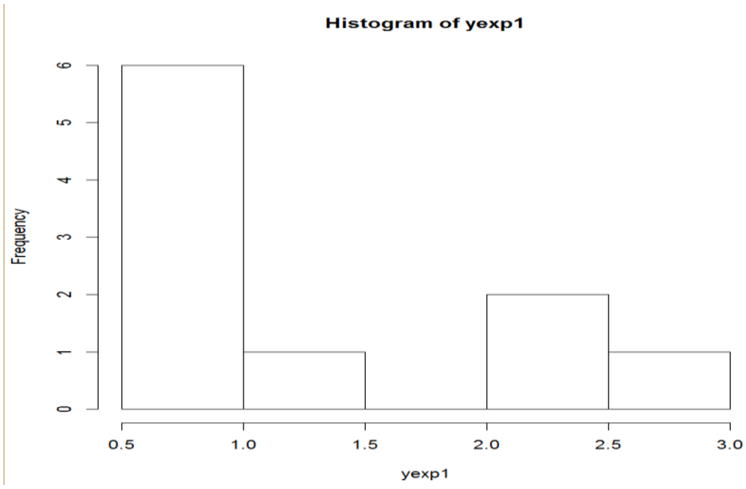


From the above plots, it is clear that with increase in number of samples ‘s’ variance is decreasing and we get more precise Normal distribution curve.

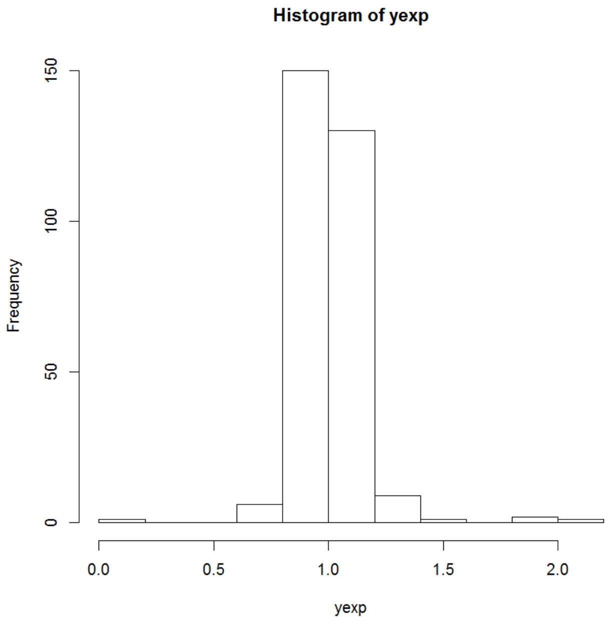
Exponential Distribution:

Code:

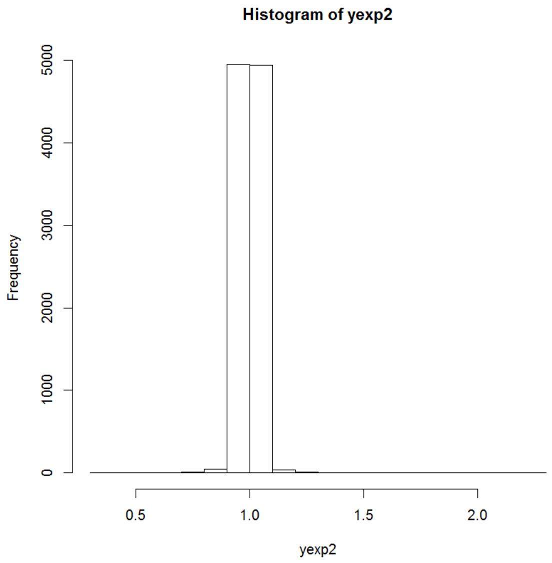
```
> yexp1 = NULL
> for(i in 1:10){ yexp1[i]=mean(rexp(i,rate=1))}
> hist(yexp1)
```



```
> yexp = NULL
> for(i in 1:300){ yexp[i]=mean(rexp(i,rate=1))}
> hist(yexp)
```



```
> yexp2 = NULL
> for(i in 1:10000){ yexp2[i]=mean(rexp(i,rate=1))}
> hist(yexp2)
```

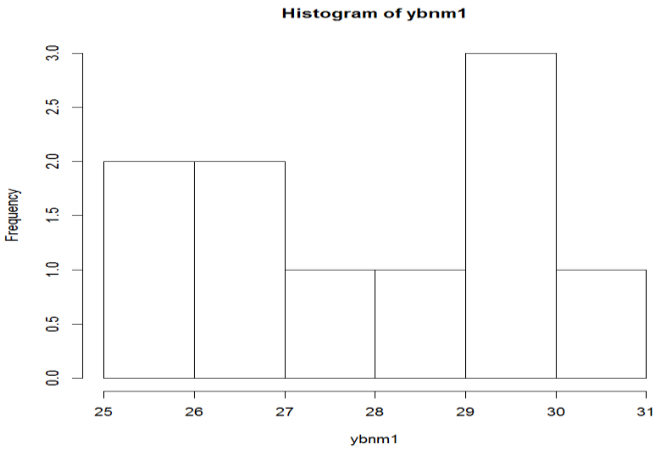


The means of Exponential Distribution follows Normal distribution curve as the sample increases.

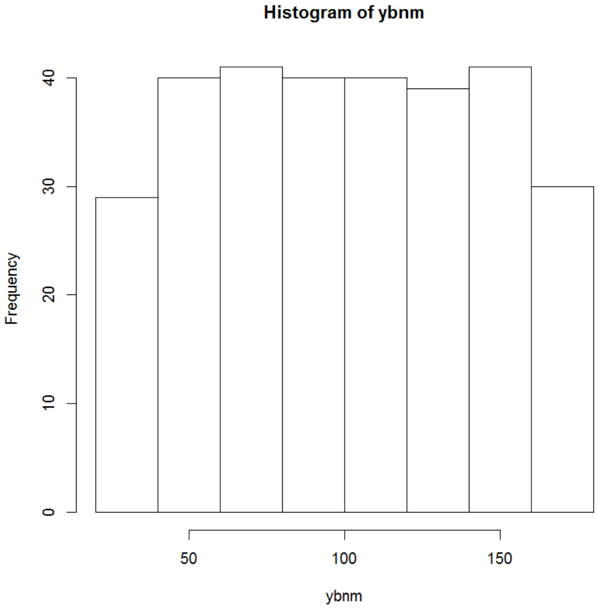
Binomial Distribution:

Code:

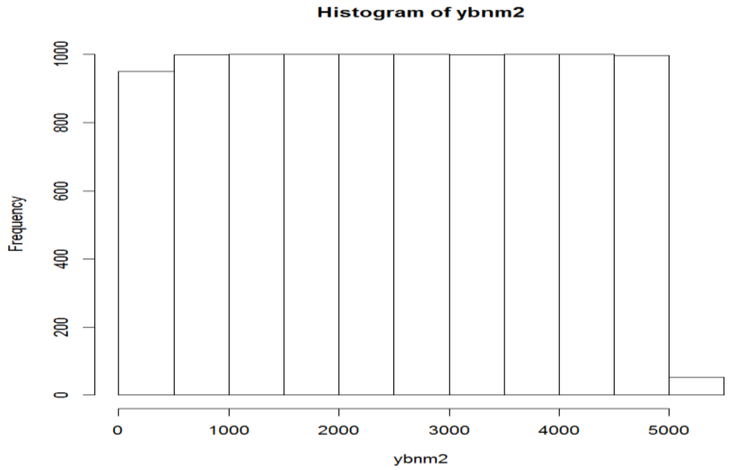
```
> ybnm1 = NULL
> for(i in 1:10){ ybnm1[i]=mean(rbinom(i,i+50,0.5))}
> hist(ybnm1)
```



```
> ybnm = NULL
> for(i in 1:300){ ybnm[i]=mean(rbinom(i,i+50,0.5))}
> hist(ybnm)
```



```
> ybnm2 = NULL
> for(i in 1:10000){ ybnm2[i]=mean(rbinom(i,i+50,0.5))}
> hist(ybnm2)
```



From the above plots it is clear that the mean values are equally distributed as sample space increases.

Q. 2)
Sol

a. Euclidean distance between each observation and the test point, X1,X2, X3

   d(a,b) = Sqrt((p1-q1)^2+(p2-q2)^2+(p3-q3)^2)

   1) Sqrt((0-0)^2+(0-3)^2+(0-0)^2)
      = sqrt(9)
      =3
   2) Sqrt((0-2)^2+(0-0)^2+(0-0)^2)
      = sqrt(4)
      =2
   3) Sqrt((0-0)^2+(0-1)^2+(0-3)^2)
      = sqrt(10)
      =3.16
   4) Sqrt((0-0)^2+(0-1)^2+(0-2)^2)
      = sqrt(5)
      =2.24
   5) Sqrt((0-(-1))^2+(0-0)^2+(0-1)^2)
      = sqrt(2)
      =1.414
   6) Sqrt((0-1)^2+(0-1)^2+(0-1)^2)
      = sqrt(3)
      =1.732

b. Prediction with k=1 is **Green** because 5th observation is nearest to the test point among all six observations.

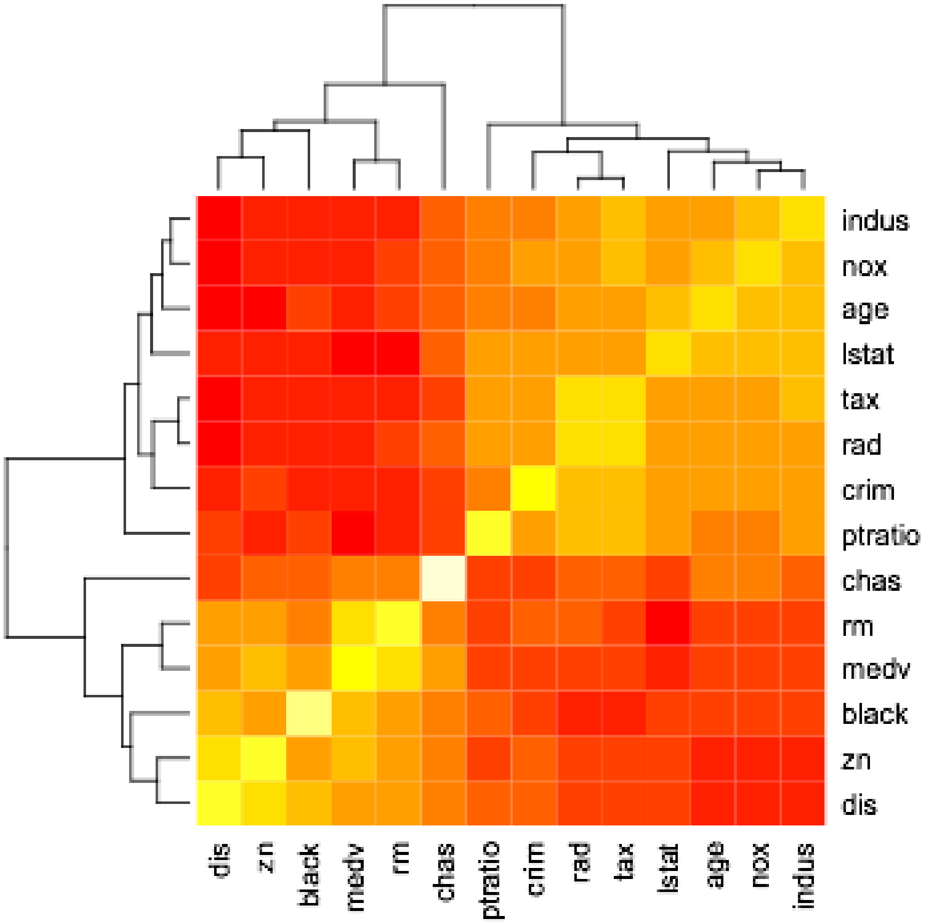
c. Prediction with k=3 is **Red** because here we need to take 3 nearest neighbors to the test point which are 2nd observation- Red, 5th observation- Green and 6th observation- Red (It consists of 2 Red points and 1 Green point), resulting in estimated probability of 2/3 for the red class and 1/3 for the green class.

d. As K value increases the Bayes decision boundary becomes less flexible (linear) hence for Bayes decision boundary to be highly nonlinear, K value should be small.

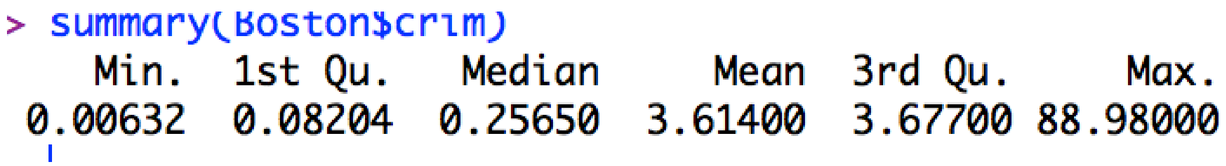
10.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

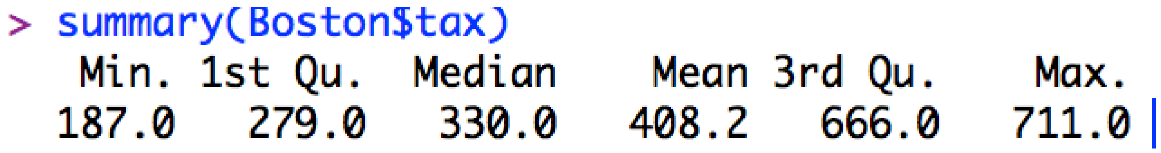
```
heatmap(cor(Boston, use=“pairwise.complete.obs"))
```



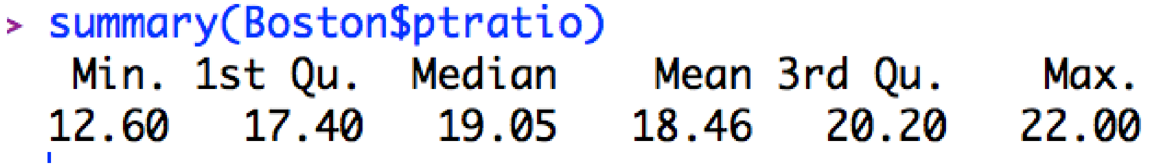
(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



```
> summary(Boston$crim)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
 0.00632  0.08204  0.25650  3.61400  3.67700 88.98000 
```

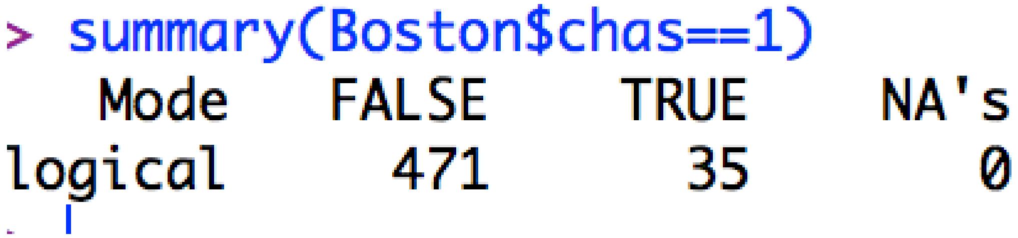


```
> summary(Boston$tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  187.0   279.0   330.0   408.2   666.0   711.0 
```



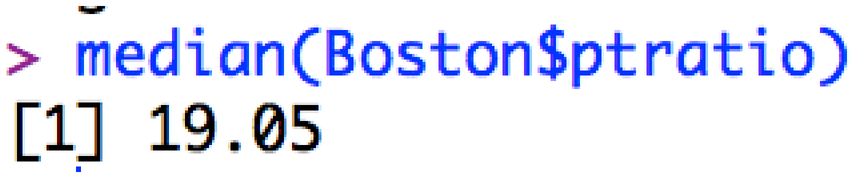
```
> summary(Boston$ptratio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.60   17.40   19.05   18.46   20.20   22.00 
```

e) How many of the suburbs in this data set bound the Charles river?



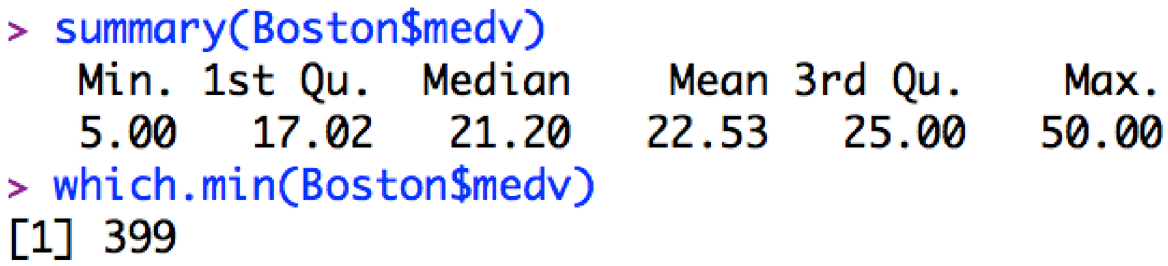
```
> summary(Boston$chas==1)
   Mode   FALSE    TRUE    NA's 
logical     471      35       0 
```

(f) What is the median pupil-teacher ratio among the towns in this data set?



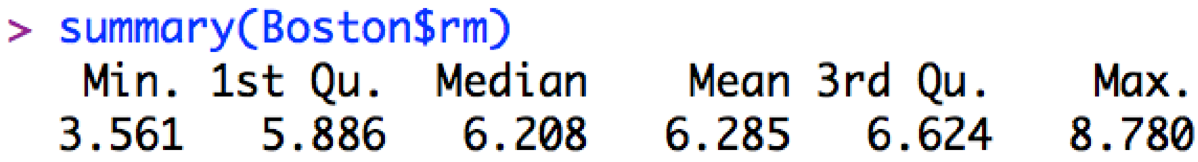
```
> median(Boston$ptratio)
[1] 19.05
```

(g) Which suburb of Boston has lowest median value of owner- occupied homes?



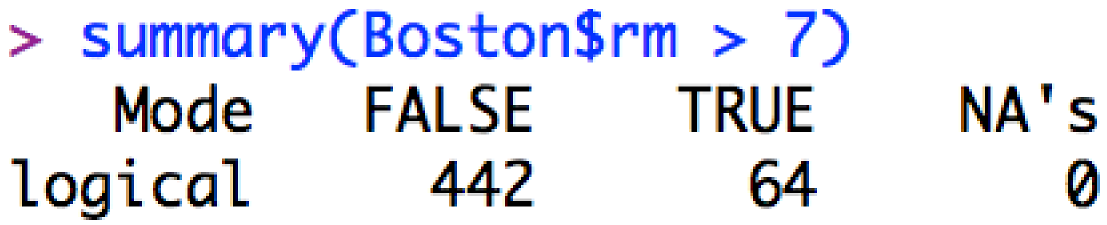
```
> summary(Boston$medv)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   17.02   21.20   22.53   25.00   50.00
> which.min(Boston$medv)
[1] 399
```

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling?



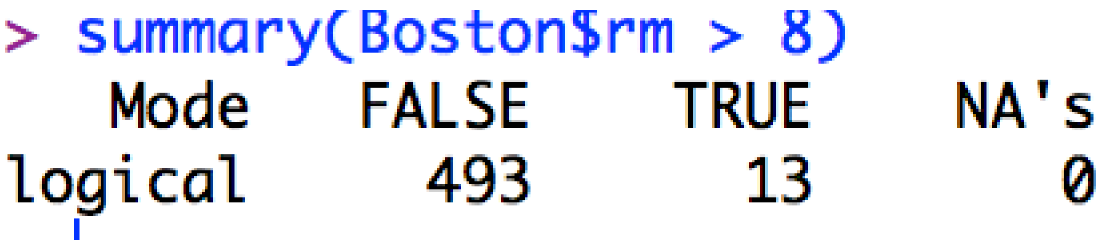
```
> summary(Boston$rm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.561   5.886   6.208   6.285   6.624   8.780
```

There are 64 suburbs with more than 7 rooms per dwelling.



```
> summary(Boston$rm > 7)
   Mode   FALSE    TRUE    NA's
logical     442      64       0
```

There are 13 suburbs with more than 8 rooms per dwelling



```
> summary(Boston$rm > 8)
   Mode   FALSE    TRUE    NA's
logical     493      13       0
```