R Assignment 2

Alireza Sheikh-Zadeh, PhD

Document format: Follow the instructions given on the web page. Always review your solution word document before submission.

Problem 1: 40 points

Problem 2: 40 points

Format: 20 points

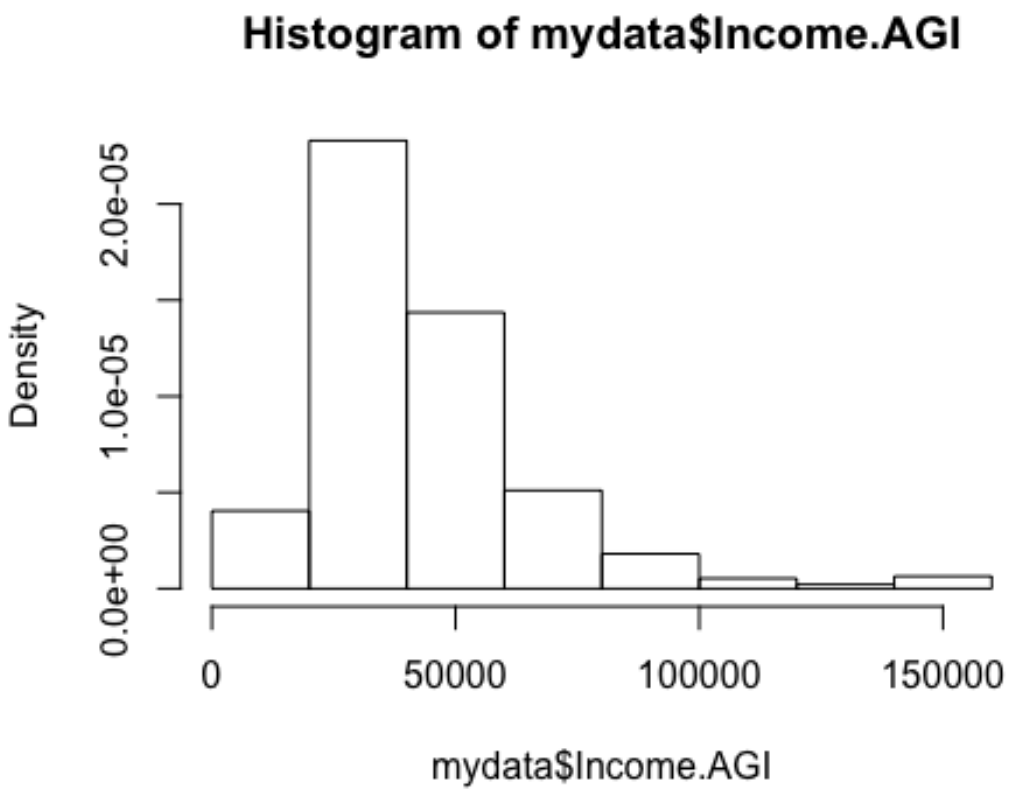
Problem 1 (40 points)

Use the charitable contributions data set:

```
mydata <- read.csv("http://tiny.cc/charitabletax")
```

a. The Adjusted Gross Income (AGI) of a taxpayer is given in the variable "Income.AGI." Display the histogram of the "Income.AGI" data showing the relative frequency in the y-axis. Does it look normally distributed (from your subjective point of view, Yes or No)? (5 points)

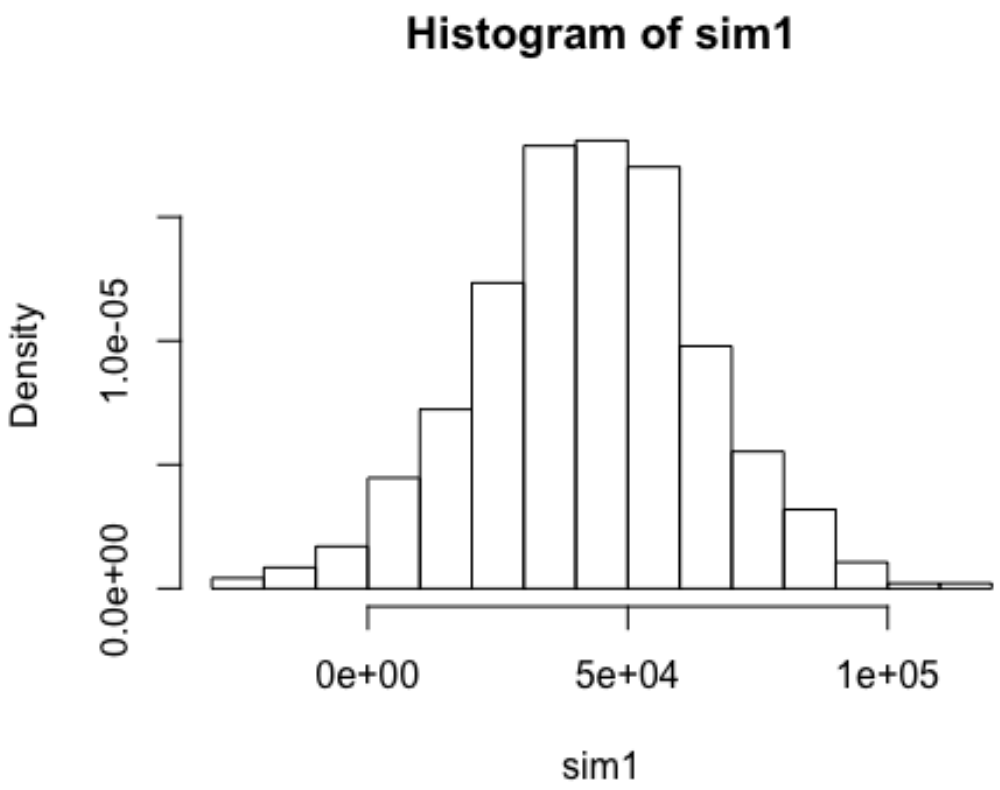
```
# No, this looks right skewed.
hist(mydata$Income.AGI, freq = F)
```



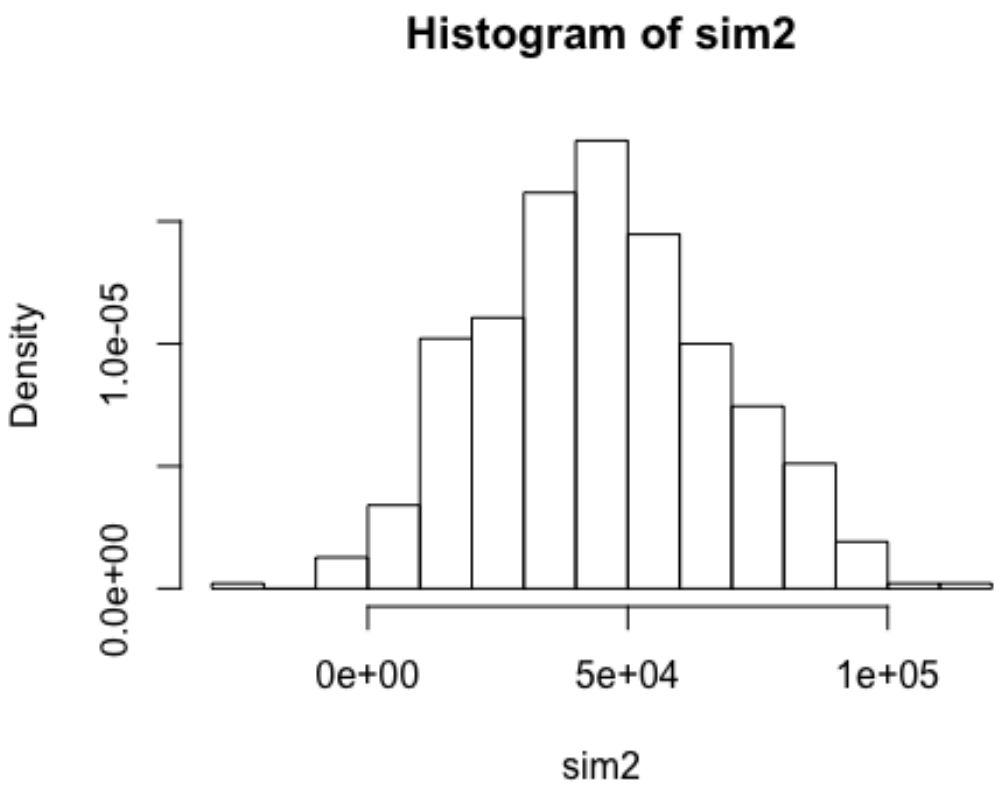

```
library(e1071)
skewness(mydata$Income.AGI)

## [1] 1.804498
```

b. Simulate five times from a normal distribution having the same mean, standard deviation and sample size (n = 470) as for the "Income.AGI" data, and name these simulated data as sim1, sim2, sim3, sim4, and sim5. Then construct the histogram of each simulated data. (10 points)

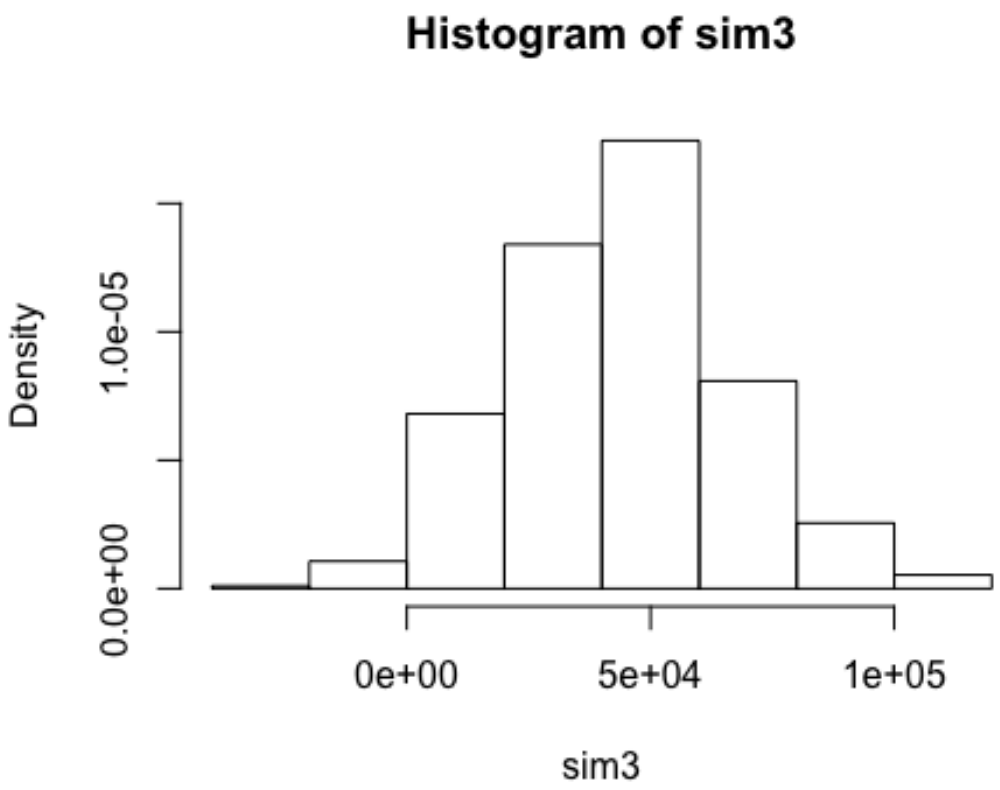
```
# 5 Simulations and 5 histograms of those simulations.

mean <- mean(mydata$Income.AGI)
sd <- sd(mydata$Income.AGI)
size <- length(mydata$Income.AGI)
sim1 <-rnorm(size, mean, sd)
sim2 <-rnorm(size, mean, sd)
sim3 <-rnorm(size, mean, sd)
sim4 <-rnorm(size, mean, sd)
sim5 <-rnorm(size, mean, sd)
hist(sim1, freq = F)
```



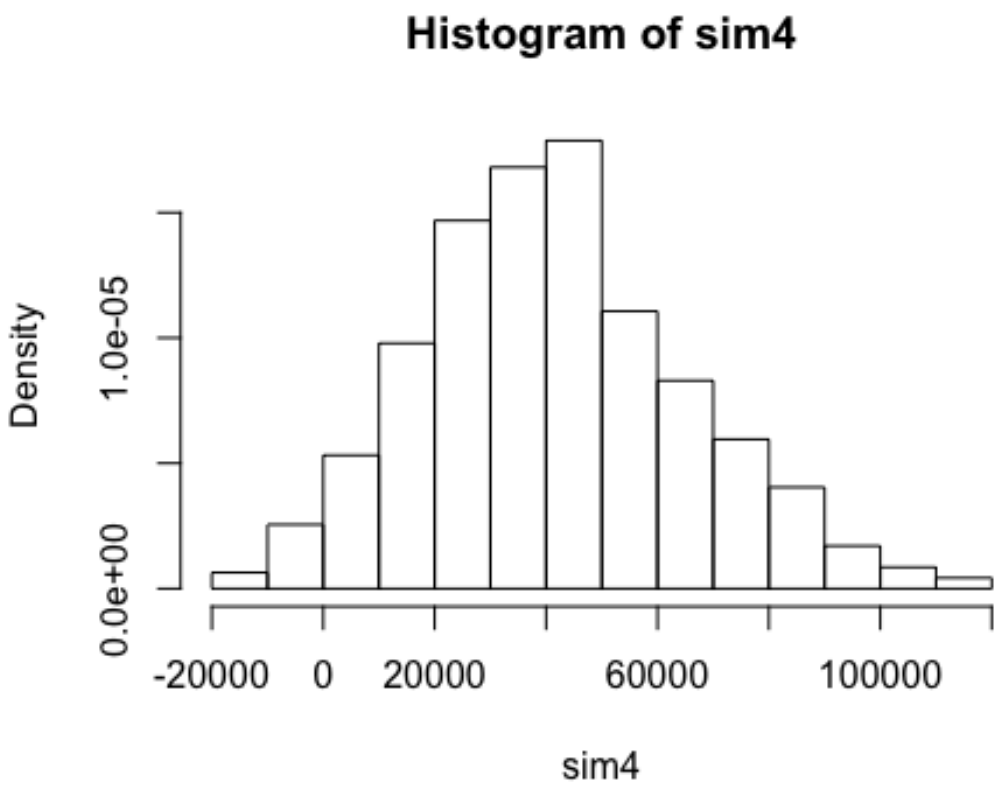

```r
hist(sim2, freq = F)
```



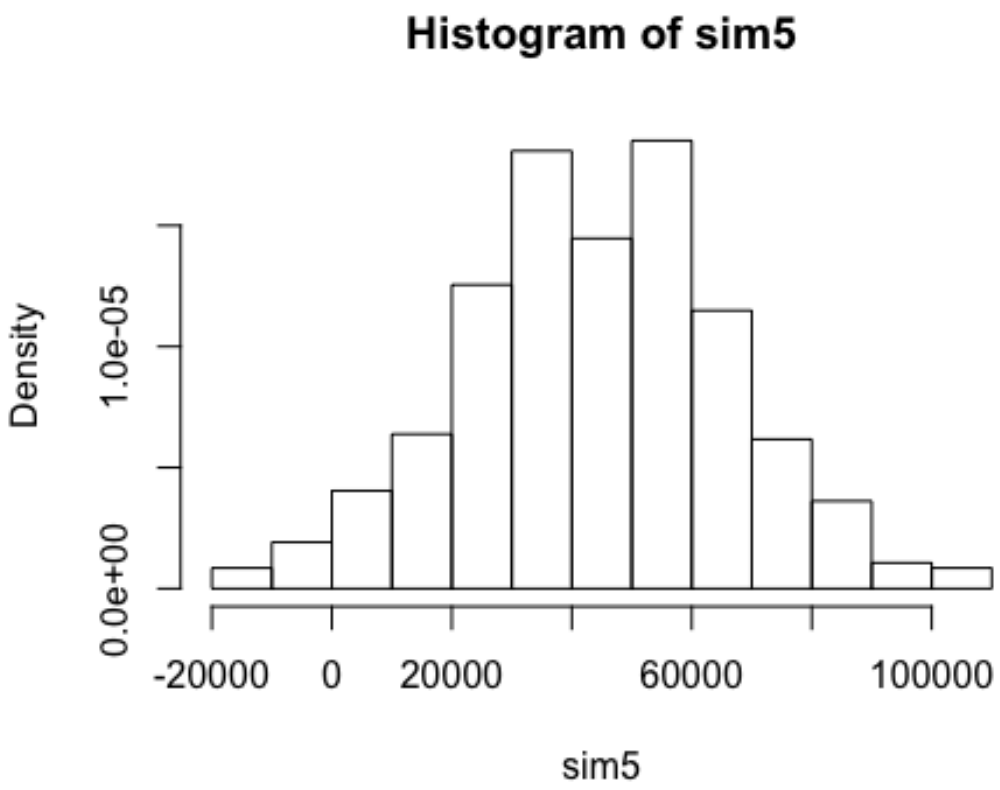

```r
hist(sim3, freq = F)
```




```r
hist(sim4, freq = F)
```




```
hist(sim5, freq = F)
```

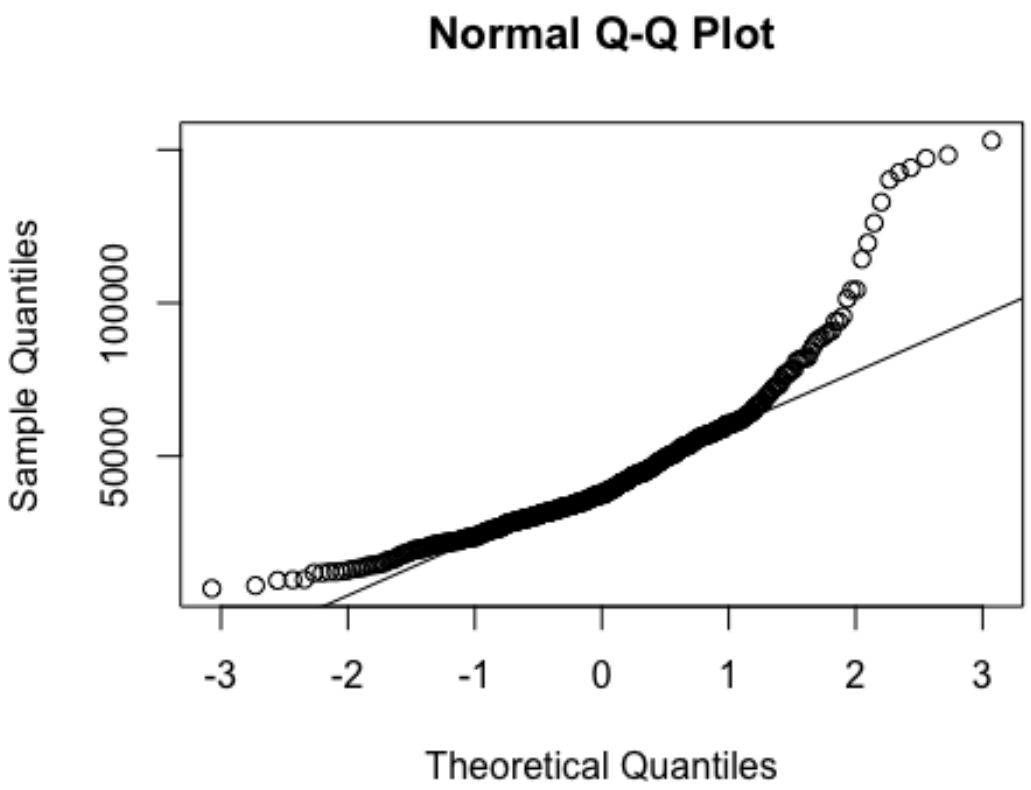


c.How many of the histograms you made in part b looks similar to the histogram of the original data (part a) (in terms of skewness, normal bell shape, and the overall appearance, from your points of view)? (2 points)

```
# None of them looks similar to the orgiinal. The five histrogram simulations
demonstrate more symmetry while the original shows a skewness.
```

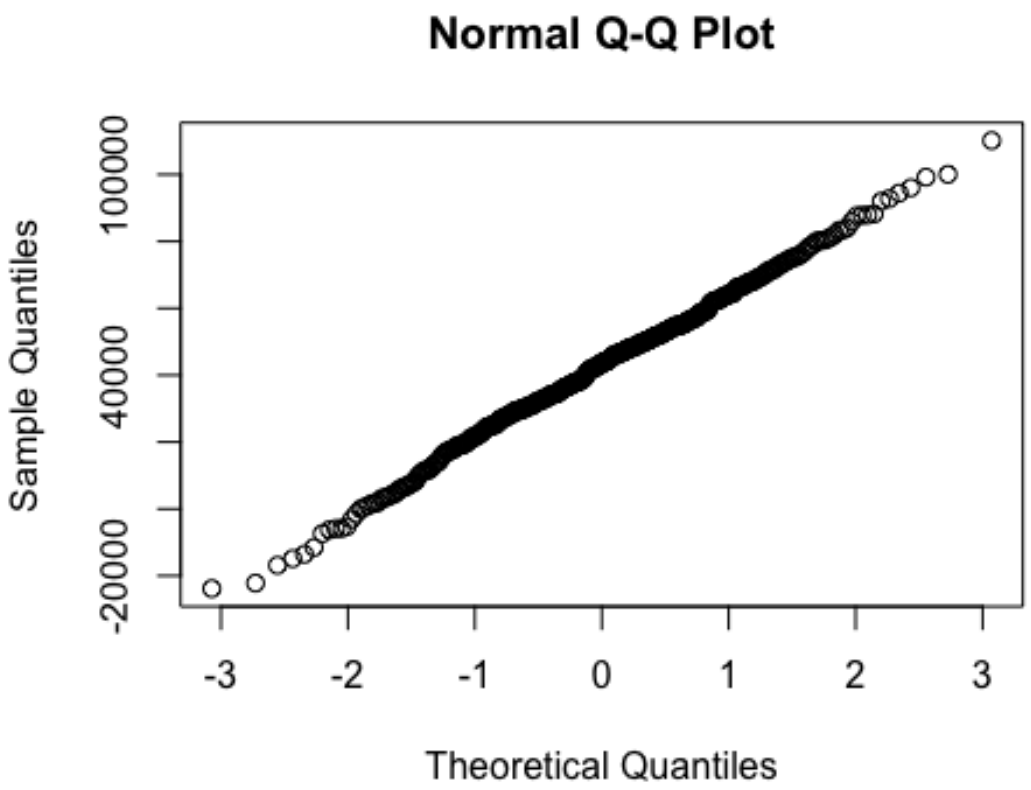
d. Display the normal q-q plot of the "Income.AGI" data. Interpret the q-q plot for analyzing the normality of the data. (5 points)

```
# This displays the qqnorm and qqline for analyzing.
qqnorm(mydata$Income.AGI)
qqline(mydata$Income.AGI)
```

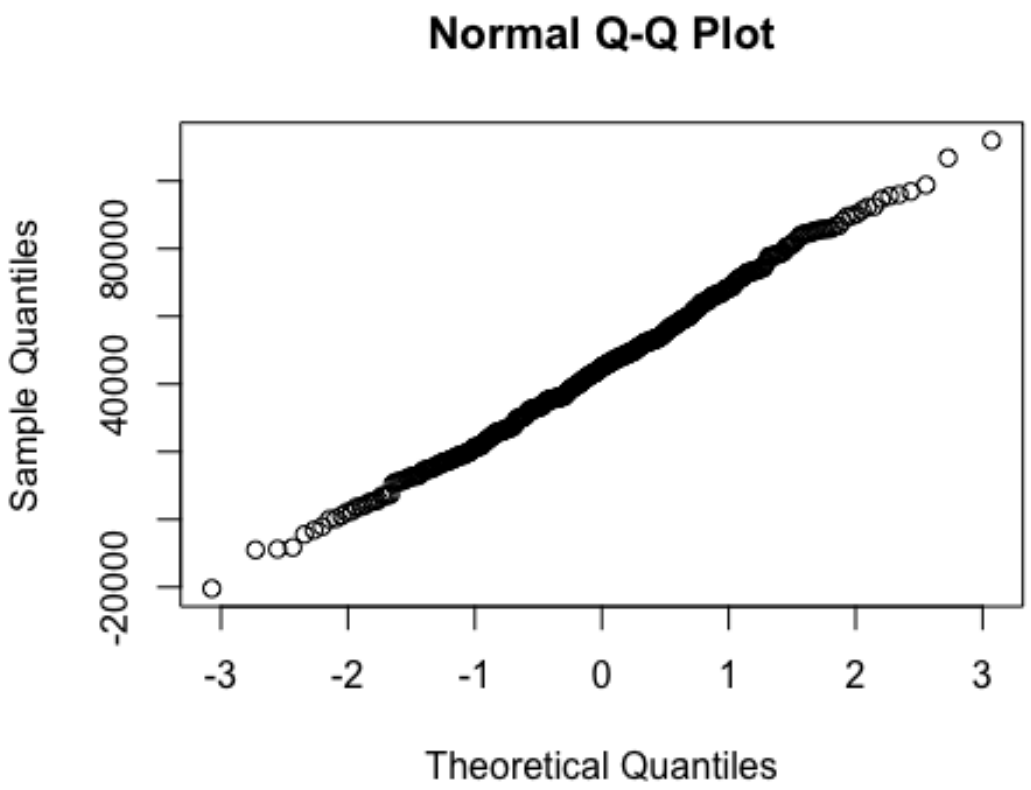


e. Construct the q-q plot of the simulated data sets: sim1, sim2, sim3, sim4, sim5. (10 points)

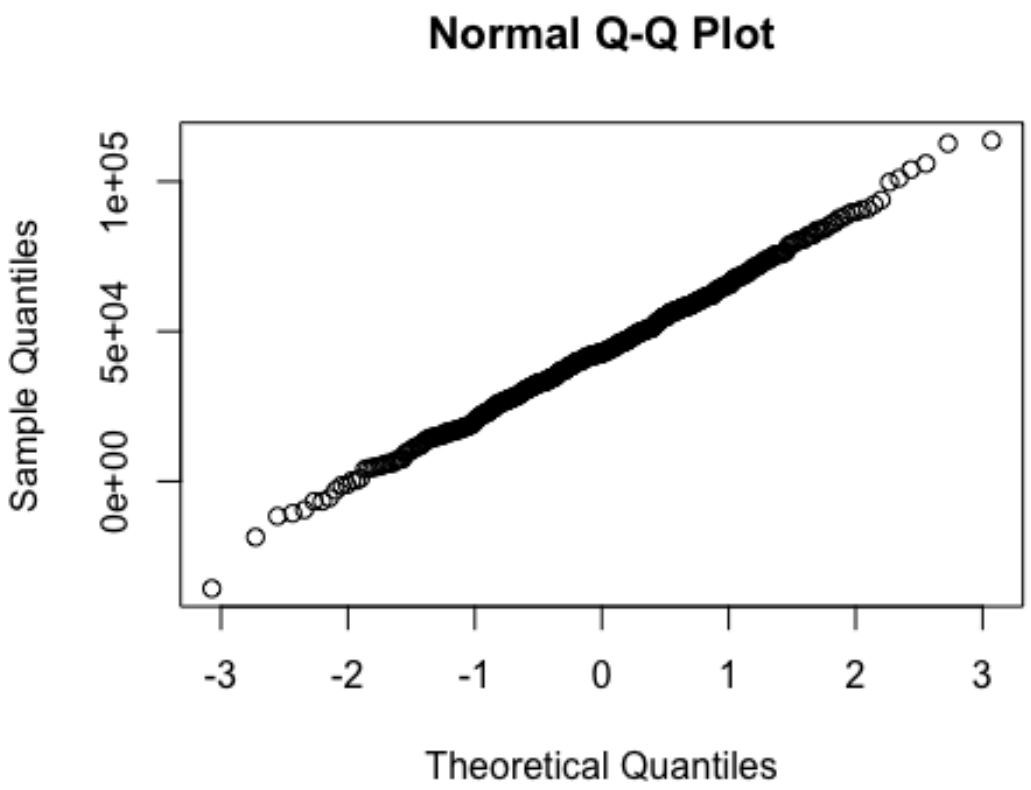
```
# This is the q-q plot of simulations 1-5
qqnorm(sim1)
```


Normal Q-Q Plot
Sample Quantiles
-20000
40000
100000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

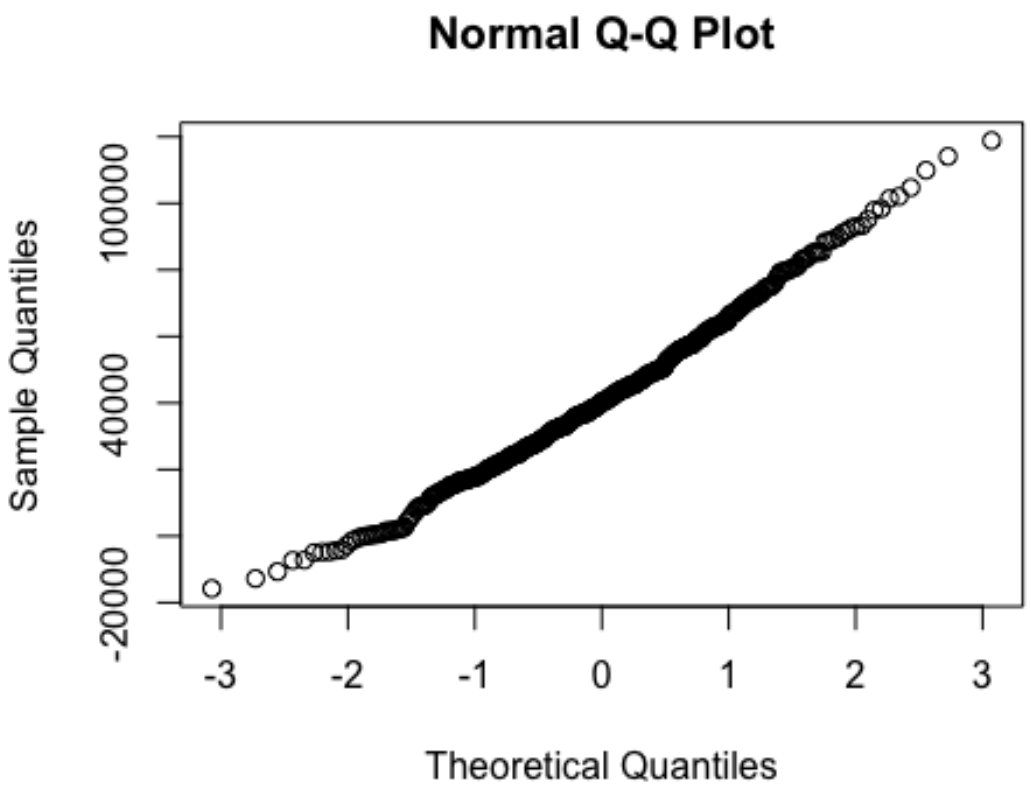
```
qqnorm(sim2)
```


Normal Q-Q Plot
Sample Quantiles
-20000
40000
80000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

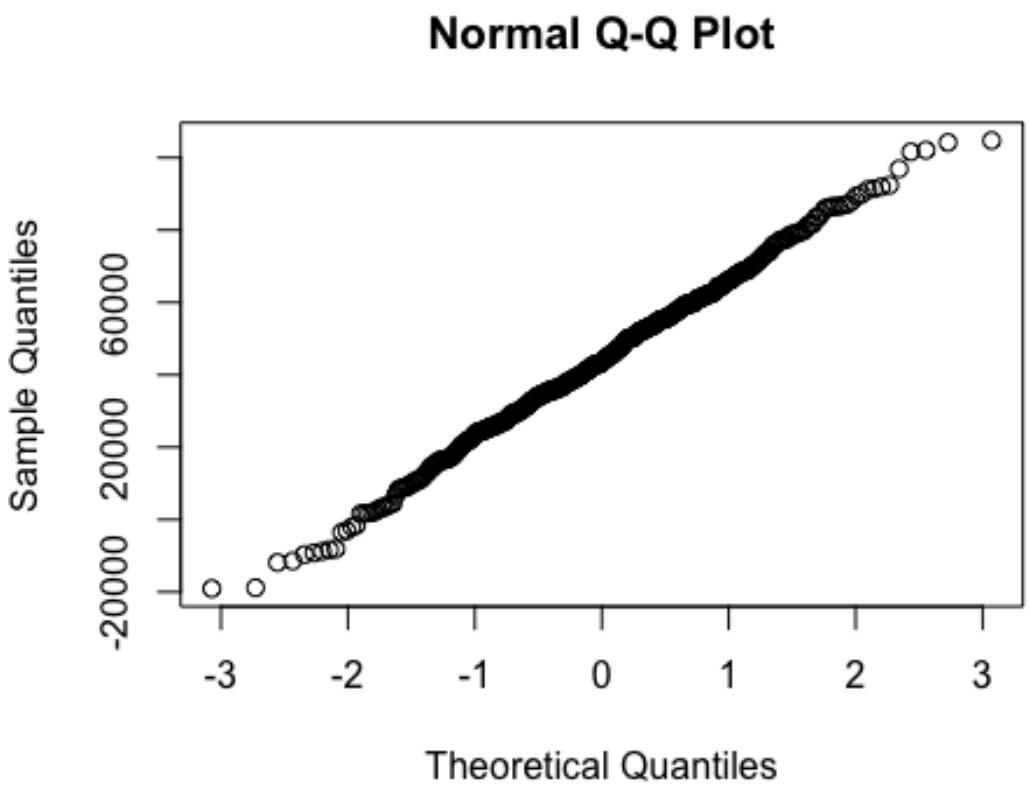
```
qqnorm(sim3)
```


Normal Q-Q Plot
Sample Quantiles
0e+00
5e+04
1e+05
-3
-2
-1
0
1
2
3
Theoretical Quantiles

```
qqnorm(sim4)
```


Normal Q-Q Plot
Sample Quantiles
-20000
40000
100000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

```
qqnorm(sim5)
```



f. How many of the q-q plots you made in part e looks similar to the q-q plot of the original data (part d)? Overall after reviewing your answer to part c and f, the distribution of how many of the simulated data sets looks similar to the original data (zero or one or … or 5)? (3 points)

```
# None of the simmulated q-qplot diagrams are similar to the original. Zero
of the simulations look similar to the original.
```

g. In this problem, we implicitly practiced the notion of hypothesis testing. Here, our null hypothesis is that the data is normally distributed, that's why in part b, we simulate data by normal distribution (rnorm using the same mean, sd, and size as the original data). The alternative hypothesis is that the data is not normally distributed. By looking at your answer to part f, you can estimate the probability that your data is similar to null hypothesis or not (This is called P-value; you learn it later!). For example, based on my evidence, zero of the simulated data are similarly distributed with the original data, then the p-value will be 0/5 = 0, which is less than 0.05, so we reject the null hypothesis and conclude that there is enough evidence that the data is not normally distributed. What is your p-value and conclusion? (5 points)

```
# Zero of the simulated data are similarly distributed with the original
Data. The p-value is 0.
```

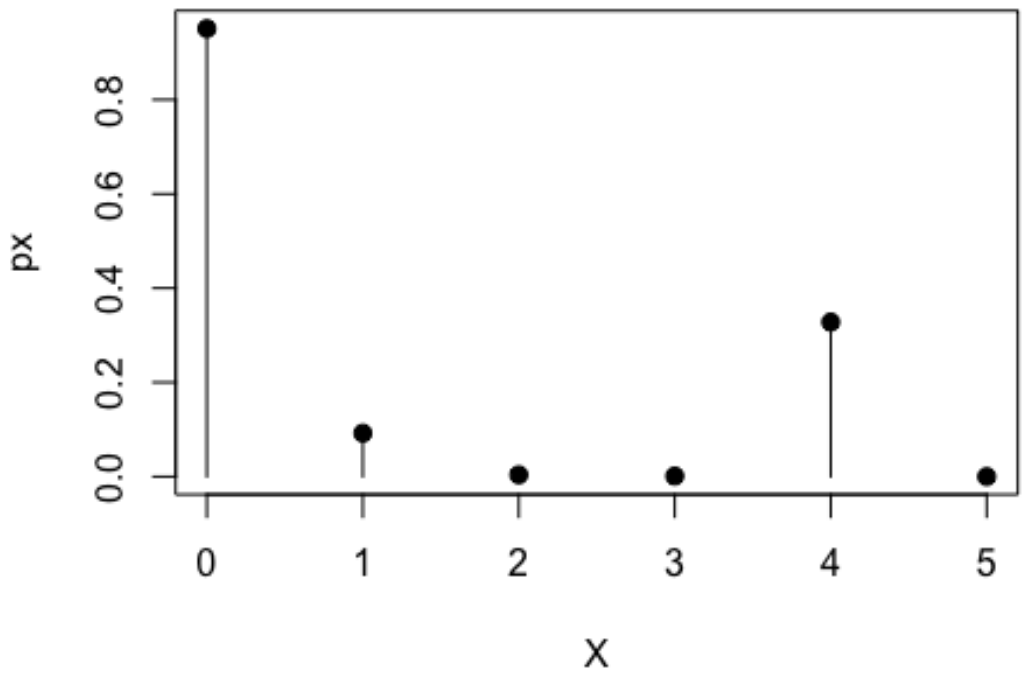
Problem 2 (40 points)

Assume the following distribution is the true distribution that produces observable customer satisfaction data. These satisfaction data are obtained by sending emails to loyal customers. There is a link in the email to a survey, which customers can access and enter their survey data, should they decide to participate. No incentives (coupons, discounts, etc.) are given to encourage the customers to fill out this survey. The data are the respondent's answer to the question, "On your recent visit to our store, were you satisfied (overall) with your shopping experience? Answer"1" means "definitely unsatisfied," answer "5" means "highly satisfied" and all other answers are intermediate.

Satisfaction, Y= y	p(y)
1	0.01
2	0.02
3	0.02
4	0.05
5	0.90
Total	1.0

a. Create a needle plot for the distribution of *Y*. Is *Y* normally distributed? (5 points)

```
# No, its not normally distributed.

y = c(1, 2, 3, 4, 5) # and
p = c(.01, .02, .02, .05, .90)

X <- 0:5
px <- dbinom(X, size = 5 , prob = p )

plot(X, px, type = 'h')
points(X,px, pch=19)
```

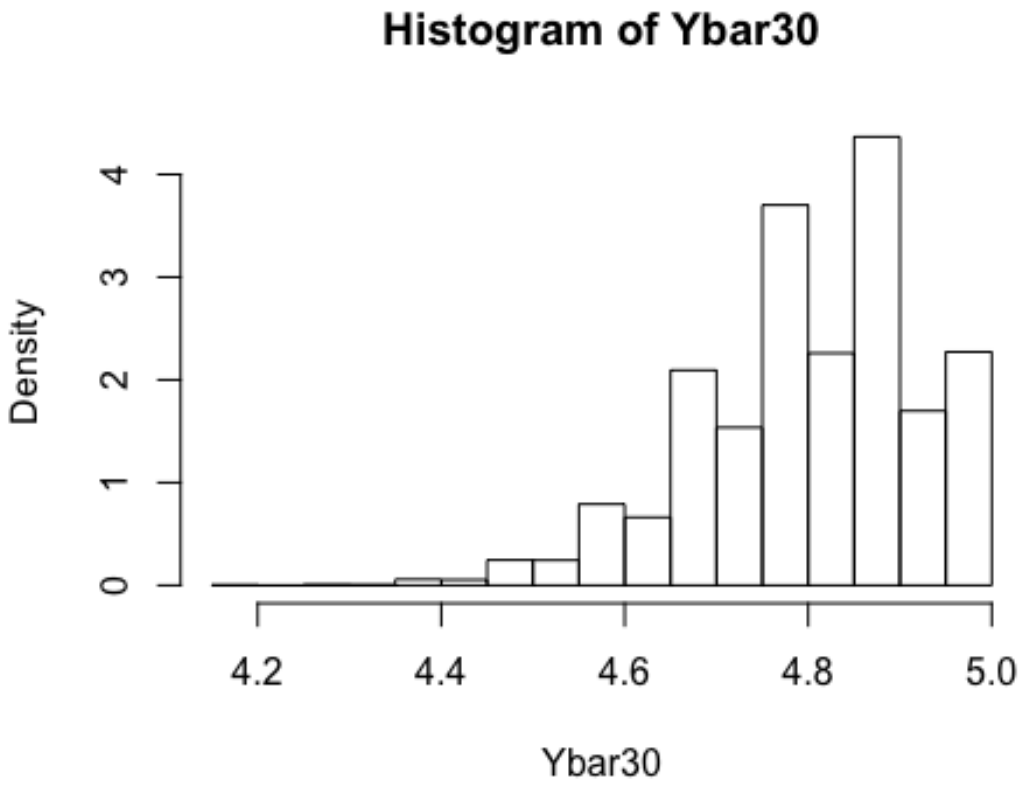


b. Simulate 10000 samples of satisfaction data, and the size of each sample is n = 30 customers. Then calculate the mean of each sample ($\overline{Y}$) and save it as an object called Ybar30. (10 points)

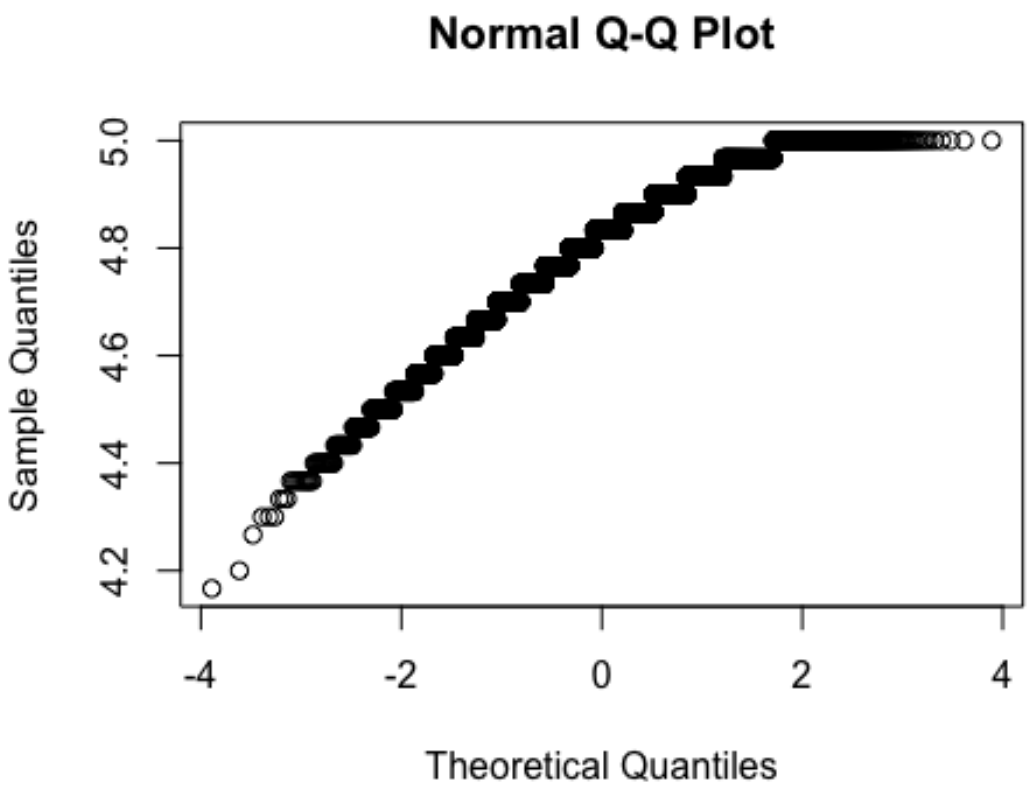
```
# Hint:
# You cab set n=1000 and NSIM = 10000
# then
# y = c(1, 2, 3, 4, 5) # and
# p = c(.01, .02, .02, .05, .90)
# just follow the same process as I explained in the video for the life
insurance example.
n = 30
simSize = 10000
ntotal = n*simSize

sim <- sample(y, ntotal, prob = p, replace = T)


simMat <- matrix(sim, ncol = n)

Ybar30 <- rowMeans(simMat)
```

c. Based on the central limit theorem, if the sample size will be large enough, the distribution of the sample mean will be normally distributed regardless of the distribution of the original random variable. Graph the histogram and q-q plot of Ybar30 that you produced in a. Is the distribution of Ybar30 normal? Is the sample size of n = 30 large enough to confirm the CLT? (5 points)

```
# No it is not normal and a sample size of 30 is not enough to confirm CLT
hist <-hist(Ybar30, freq = F)
```




```
qqnorm(Ybar30)
```

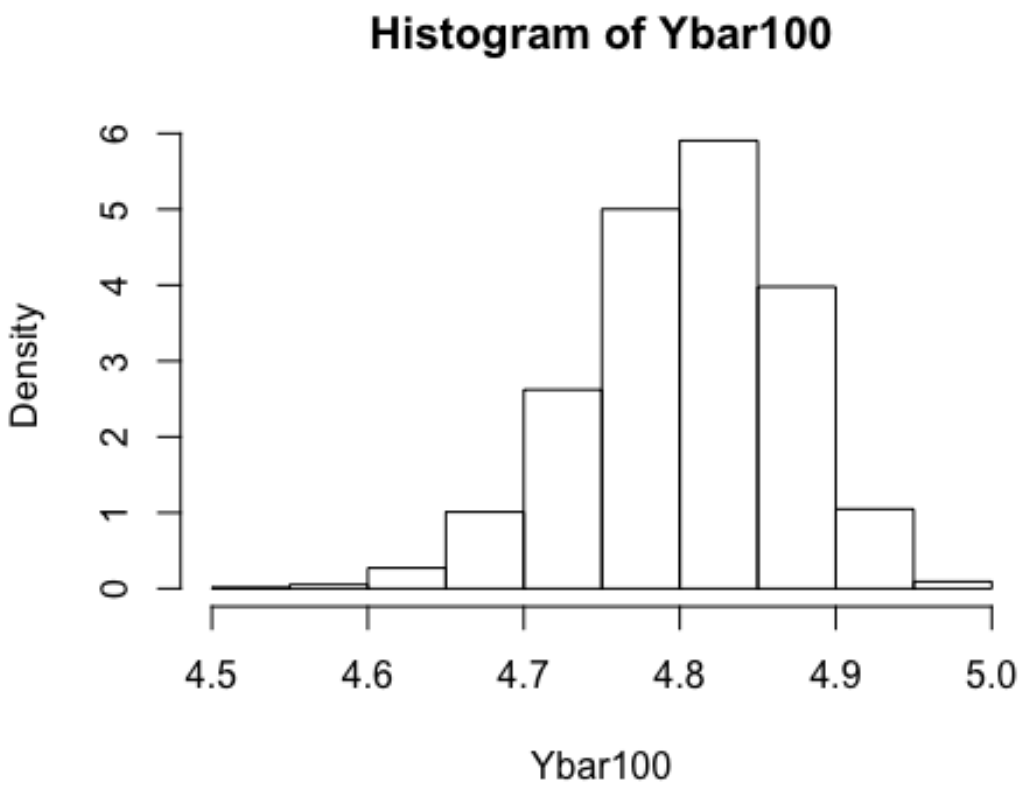



```
head(hist$density)
```

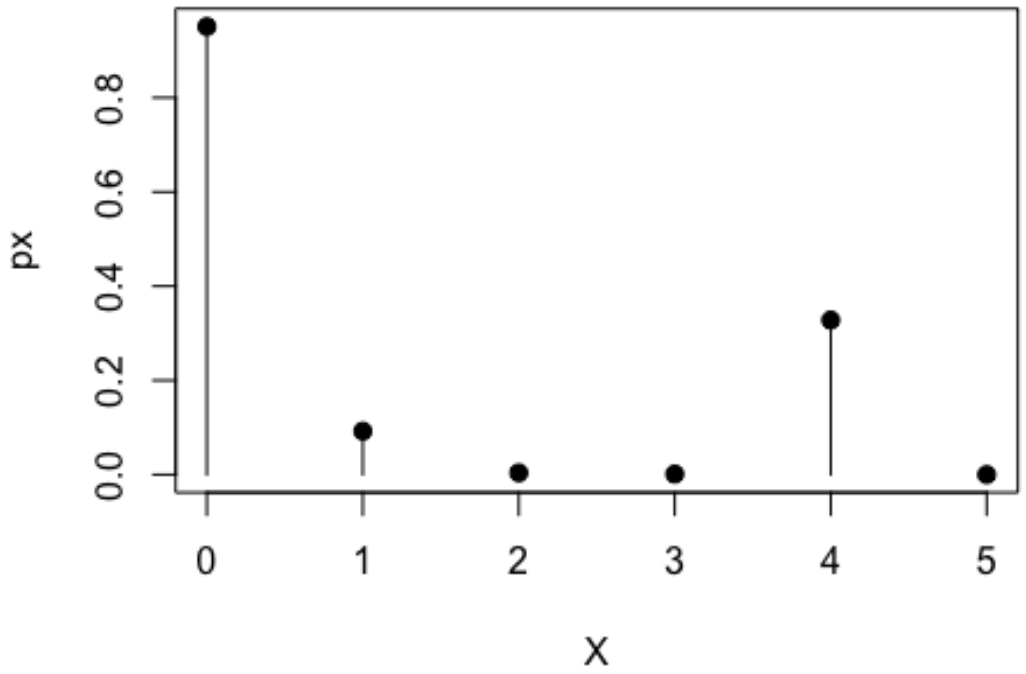
```
## [1] 0.004 0.000 0.008 0.006 0.058 0.054
```

d. repeat part a and b for n = 100. Save the sample means into an object called Ybar100. Is n = 100 large enough to confirm the CLT? (10 points)

```
# It is getting closer to comfirming, but it's not able to confirm CLT and
not normally distributed.
n = 100
simSize = 10000
ntotal = n*simSize
sim <- sample(y, ntotal, prob = p, replace = T)
simMat <- matrix(sim, ncol = n)
Ybar100 <- rowMeans(simMat)
hist <-hist(Ybar100, freq = F)
```

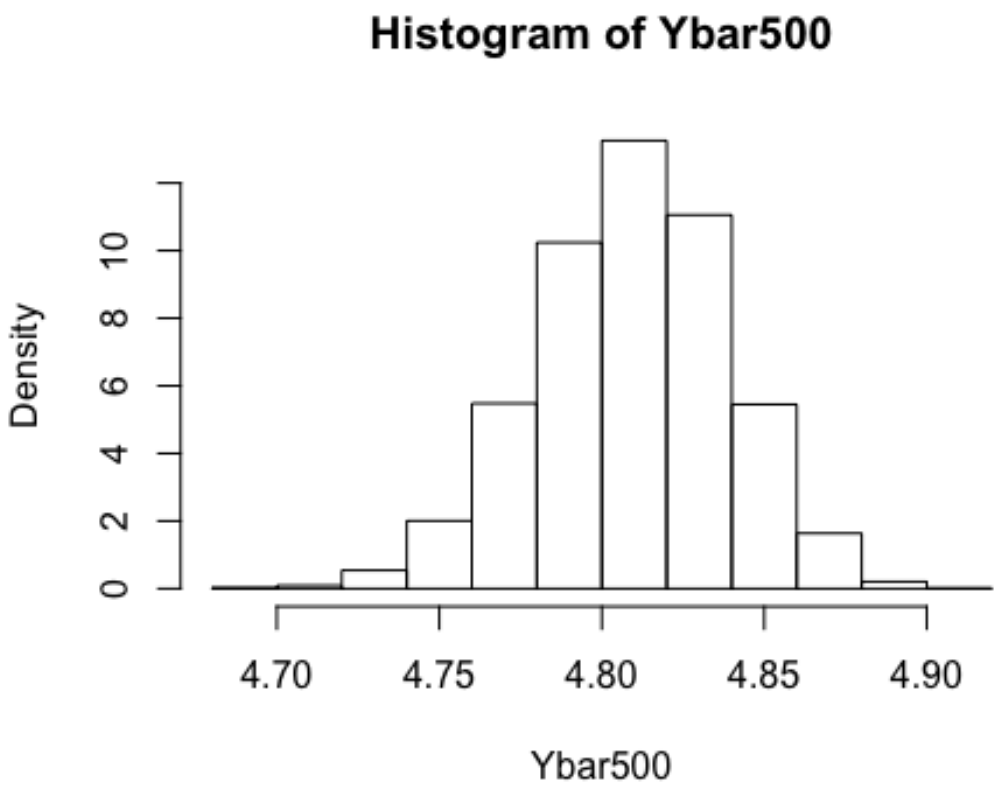



```r
X <- 0:5
px <- dbinom(X, size = 5 , prob = p )

plot(X, px, type = 'h')
points(X,px, pch=19)
```

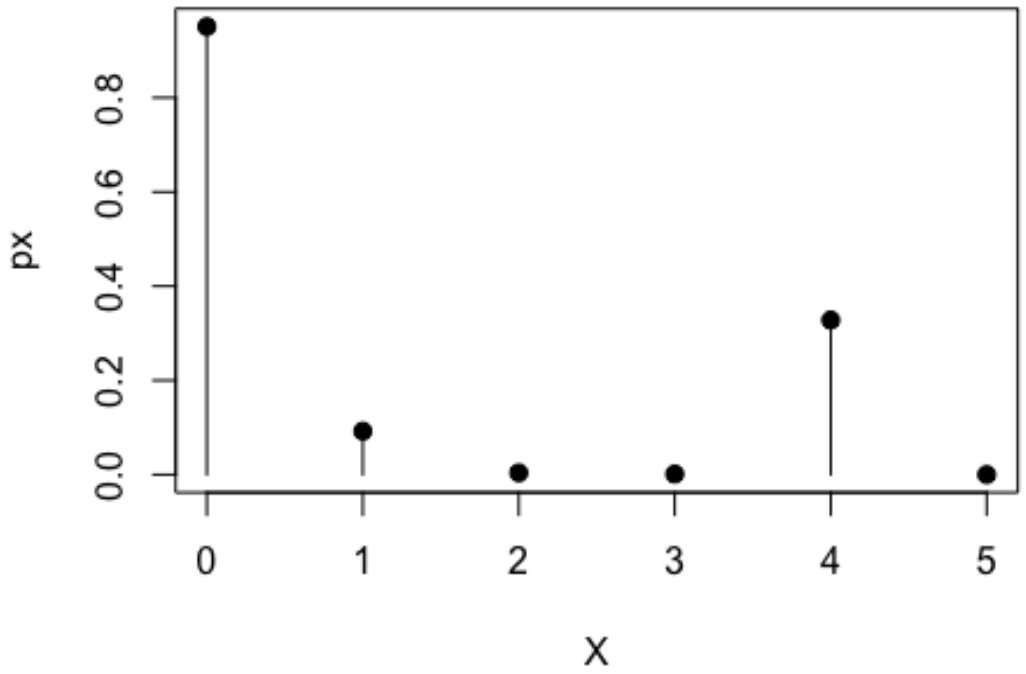


e. Choose an appropriate n, and repeat part a and b to confirm the appropriate sample size for the CLT? (10 points)

```
# Yes this confirms CLT and is distributed normally.
n = 500
simSize = 10000
ntotal = n*simSize
sim <- sample(y, ntotal, prob = p, replace = T)
simMat <- matrix(sim, ncol = n)
Ybar500 <- rowMeans(simMat)
hist <-hist(Ybar500, freq = F)
```




```r
y = c(1, 2, 3, 4, 5) # and
p = c(.01, .02, .02, .05, .90)

X <- 0:5
px <- dbinom(X, size = 5 , prob = p )

plot(X, px, type = 'h')
points(X,px, pch=19)
```


px
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
X