R Assignment 3

Jonathan De Los Santos

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

Plagiarism: You are not allowed to share your write-up with your peers. It’s okay to advise your peers about how to solve a problem, but you never share your own write-up.

Problem 1: 40 points

Problem 2: 40 points

Format: 20 points

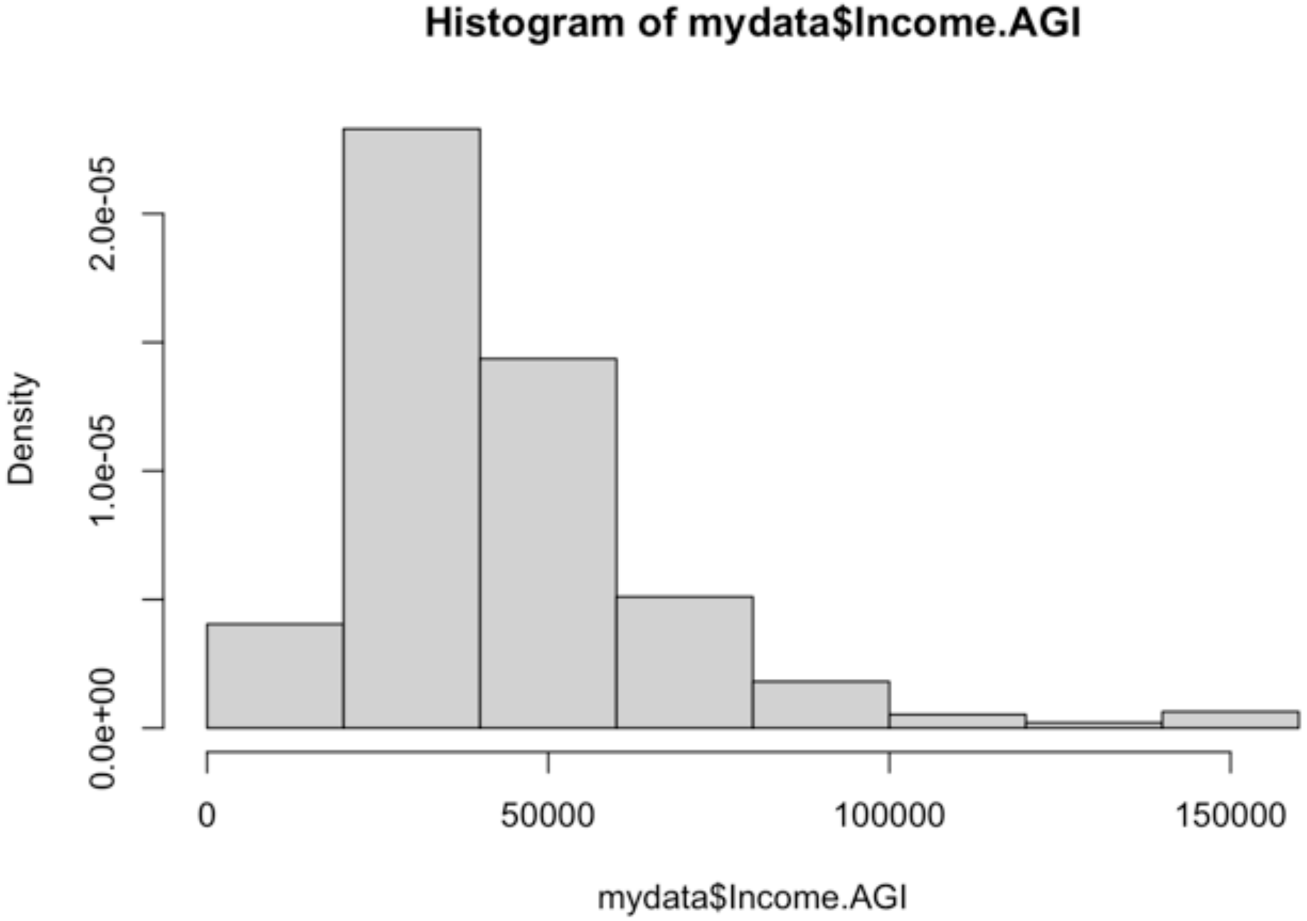
Problem 1 (40 points)

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)

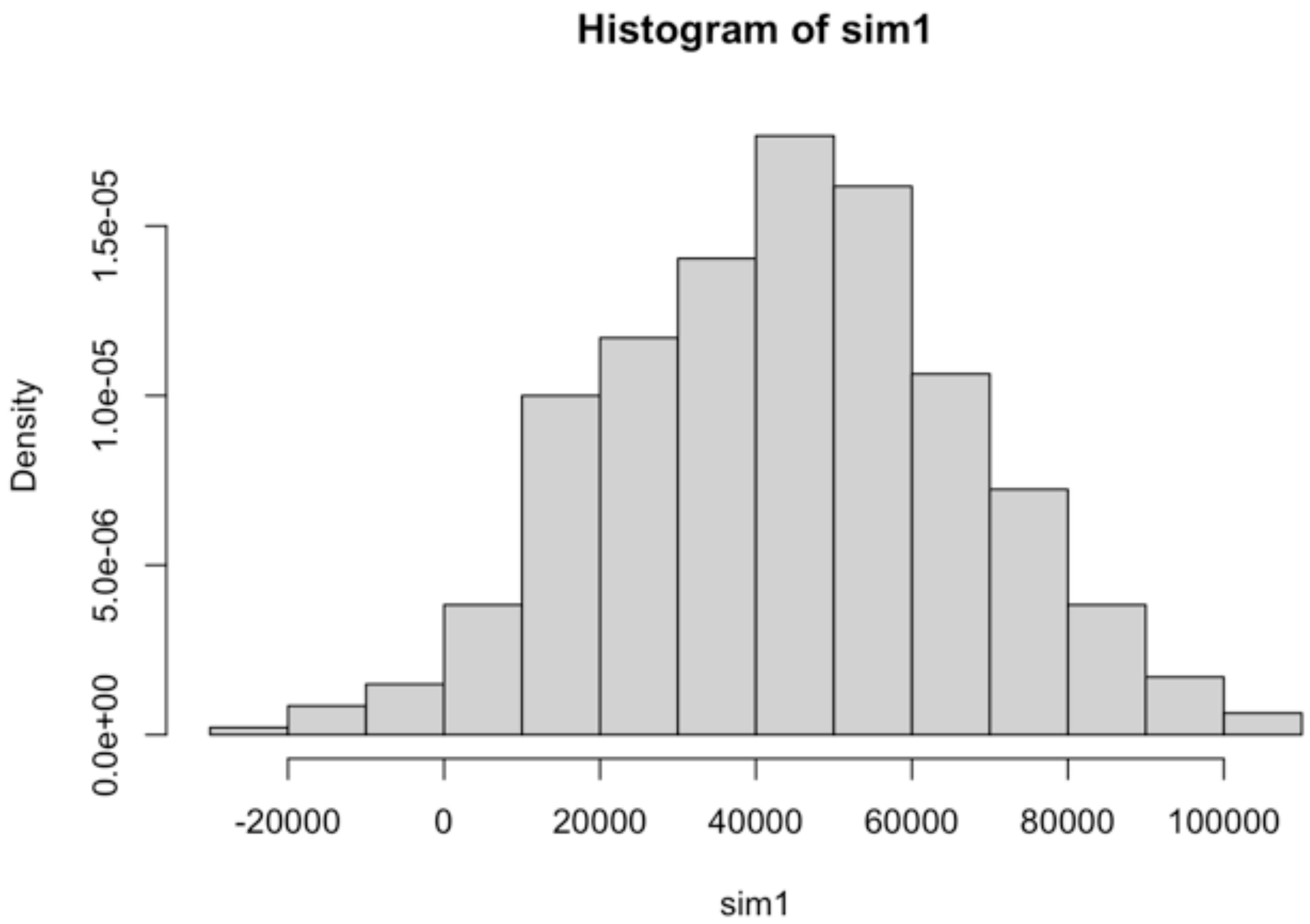
```
hist(mydata$Income.AGI, freq = F)
```



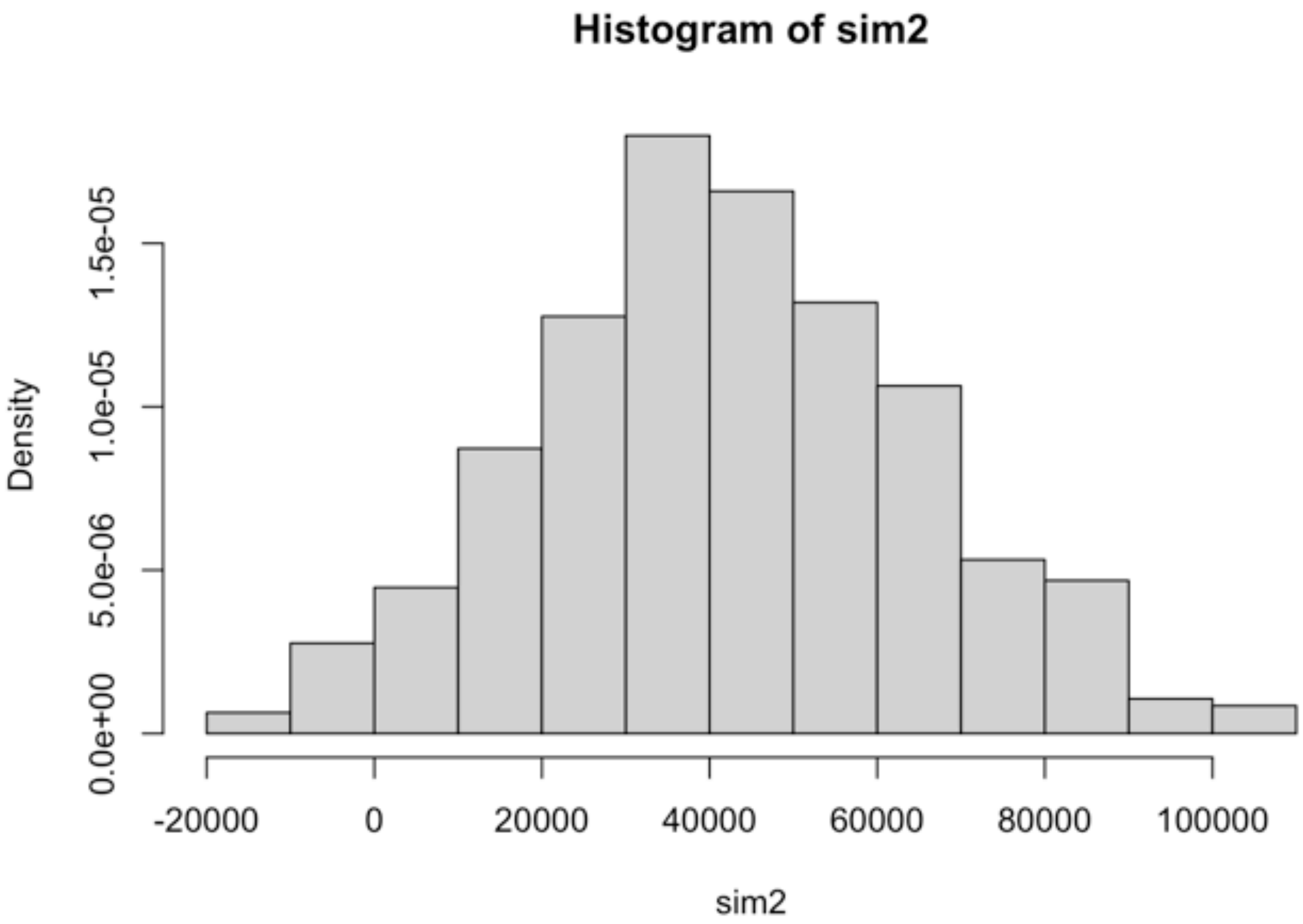
The AGI in this dataset are not normally distributed, there is an obvious skew to the right.

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)

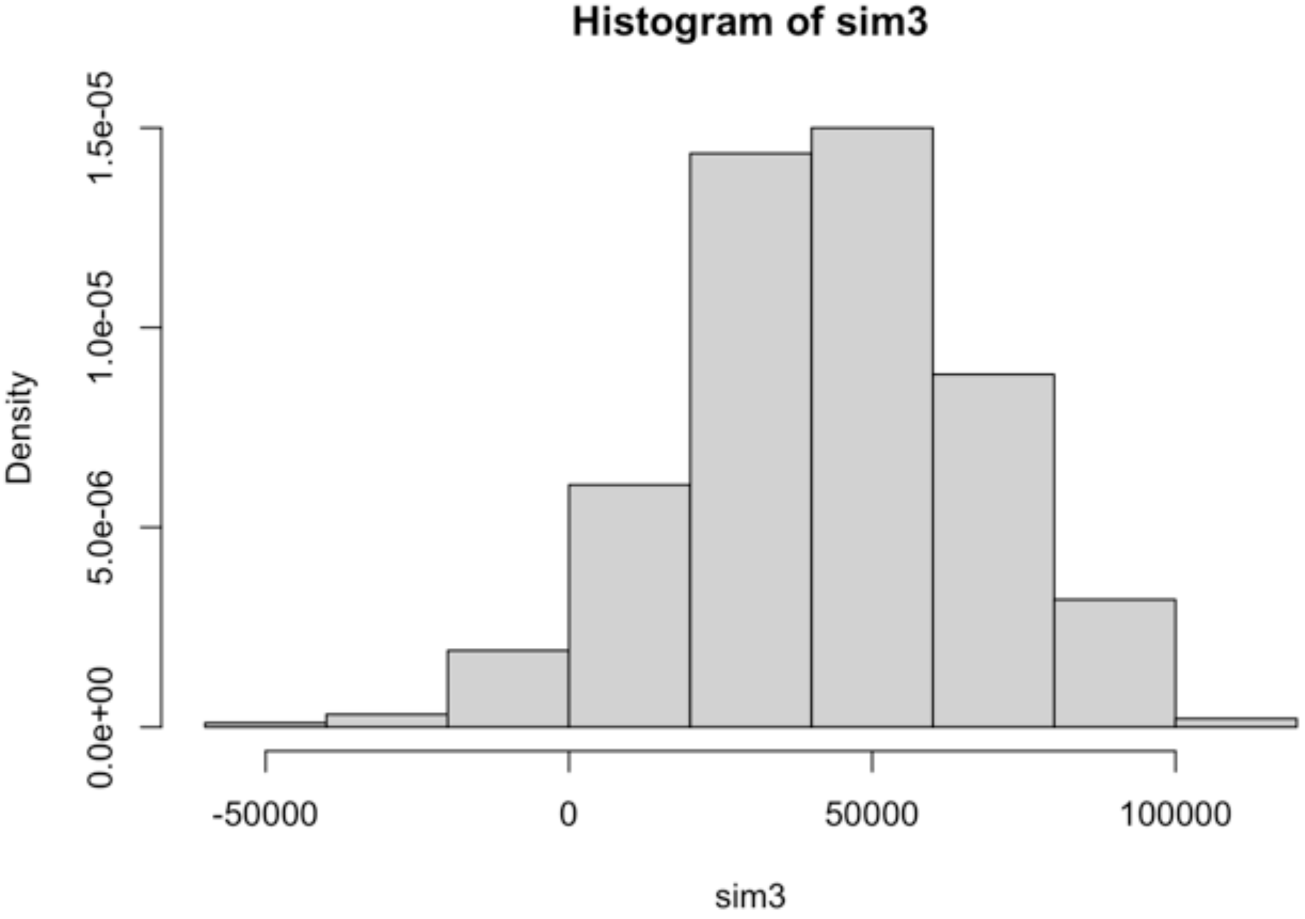
```
# Hint: Find the mean() and sd() of the mydata$Income.AGI, then simulate normal insta
nces by sim1 <- rnorm(470, mean, sd), sim2 <- rnorm(470, mean, sd), and .... then rep
ort hist() of sim1, sim2, .....

# Store mean, sd, and length in variables for use in simulations
mean <- mean(mydata$Income.AGI)
sd <- sd(mydata$Income.AGI)
size <- length(mydata$Income.AGI)

# Create simulations
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)

# Print histogram of each simulation
hist(sim1, freq = F)
```



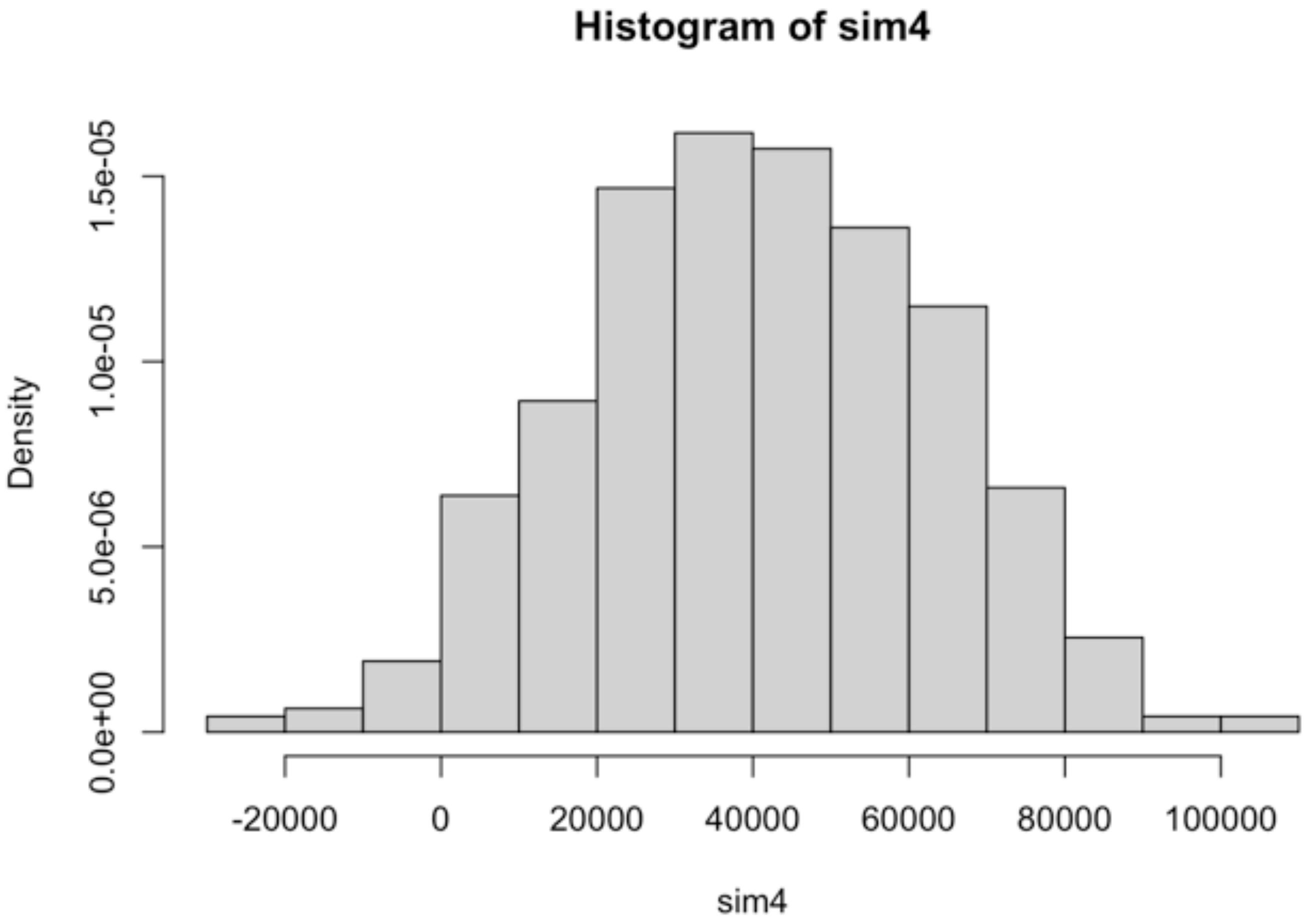
```
hist(sim2, freq = F)
```



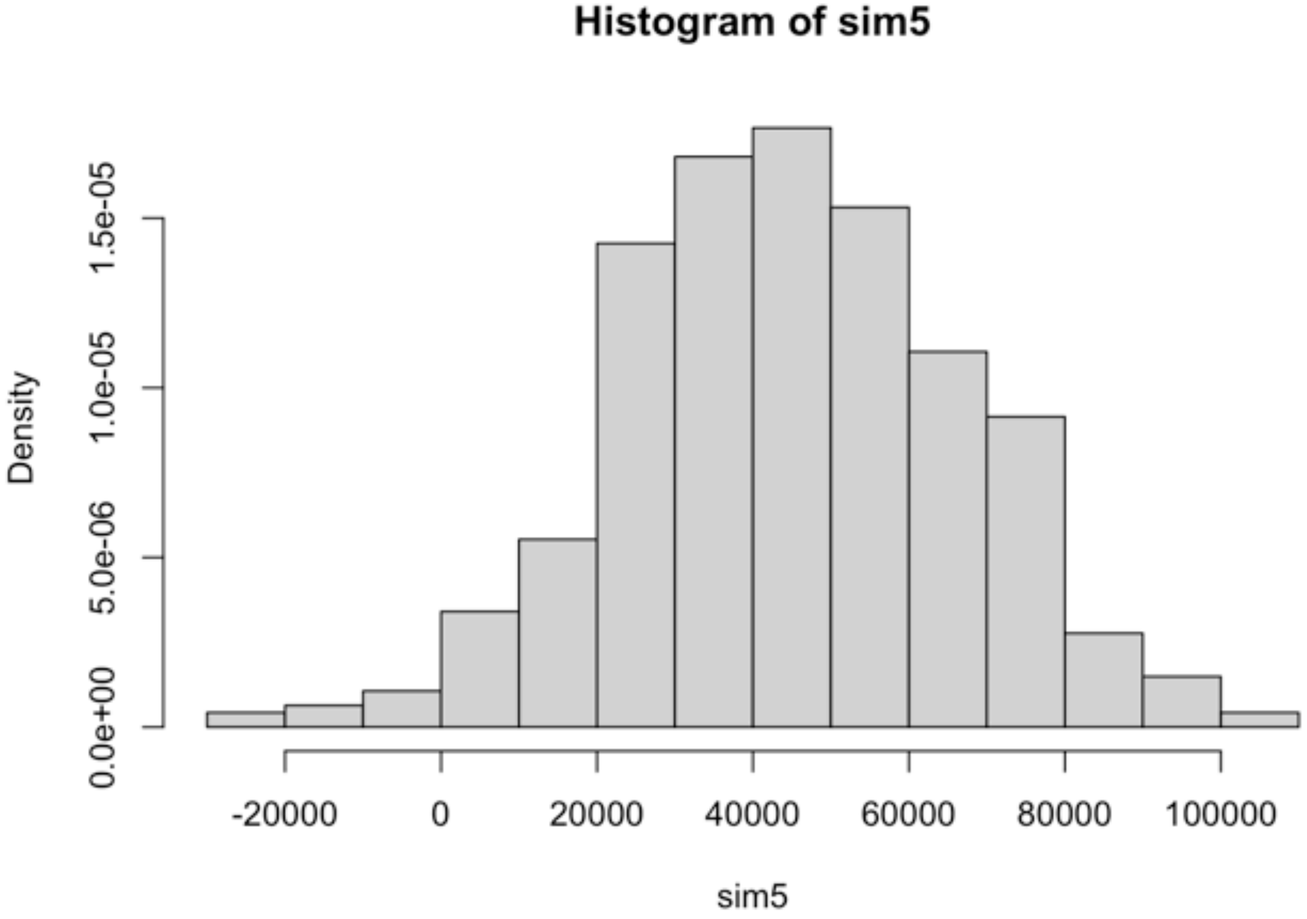
```
hist(sim3, freq = F)
```



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



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

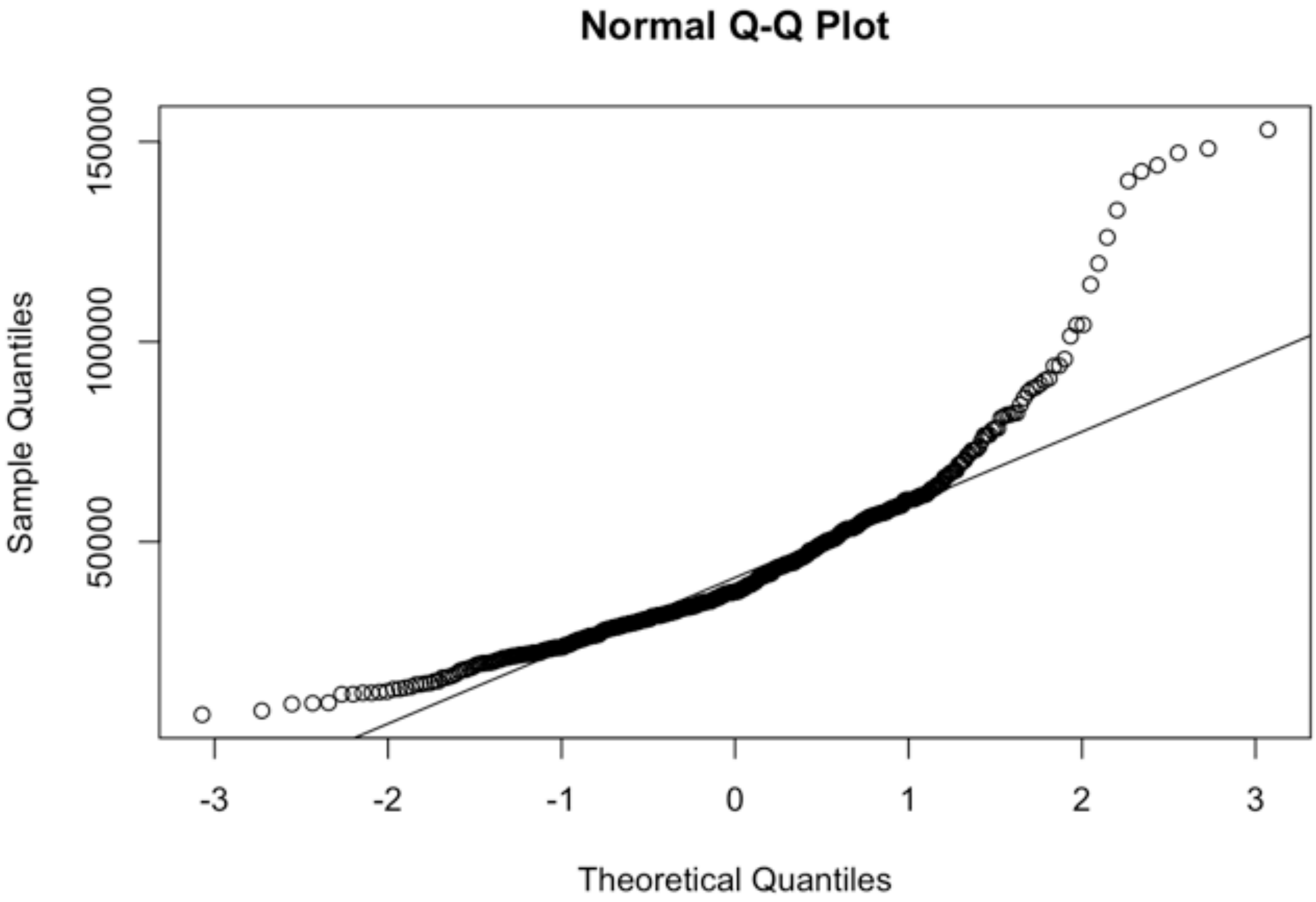


c.How many of the histograms you made in part b looks similar to the histogram of the original data (part a) (in terms of skewness, normal bell shape, and the overall appearance, from your points of view)? (2 points)

In these simulations, none of these histograms look like the original data. They appear much more normal with very limited skew.

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)

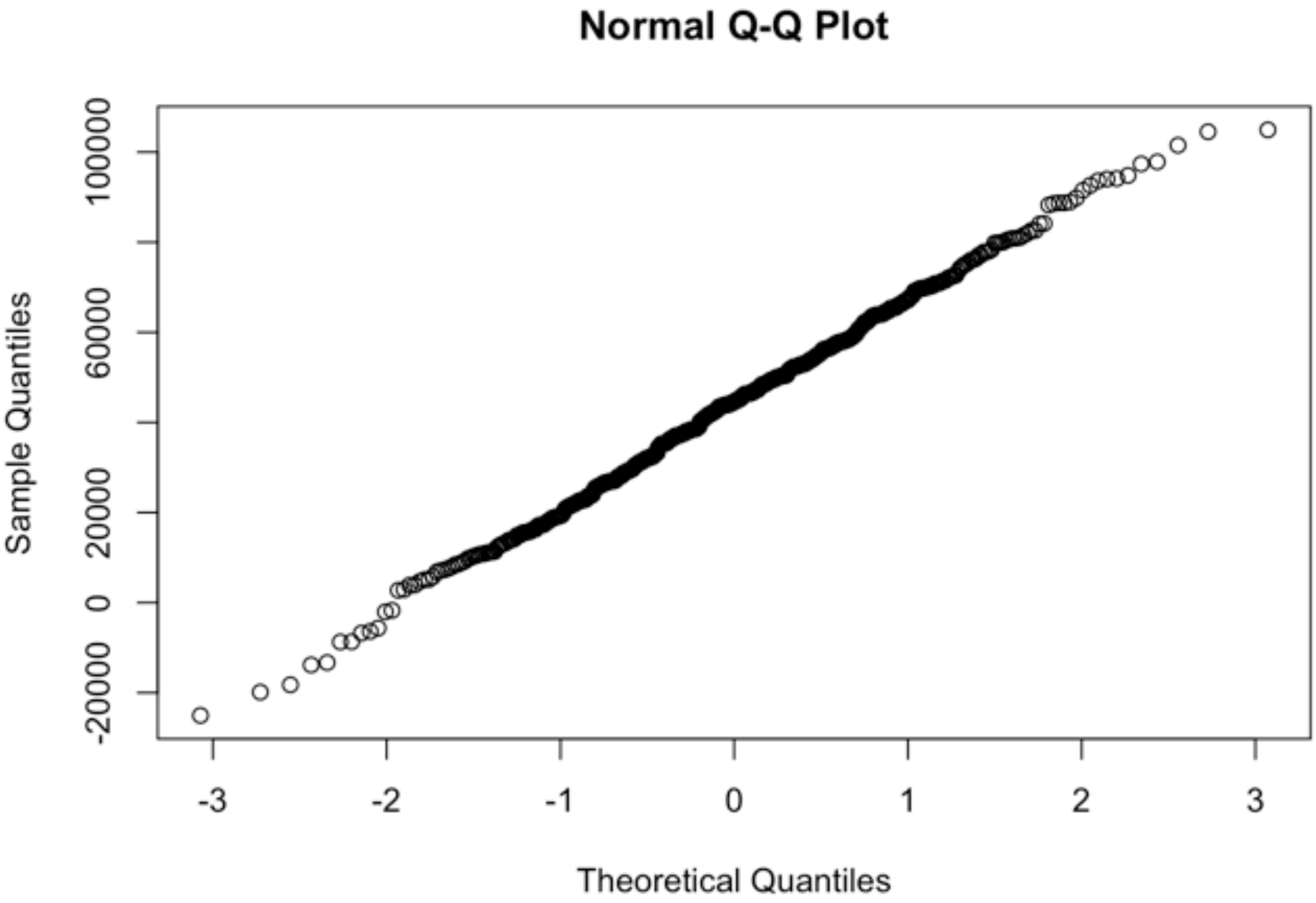
```
# Hint: You can use qqnorm and qqline functions. I did this in the lecture.
qqnorm(mydata$Income.AGI)
qqline(mydata$Income.AGI)
```



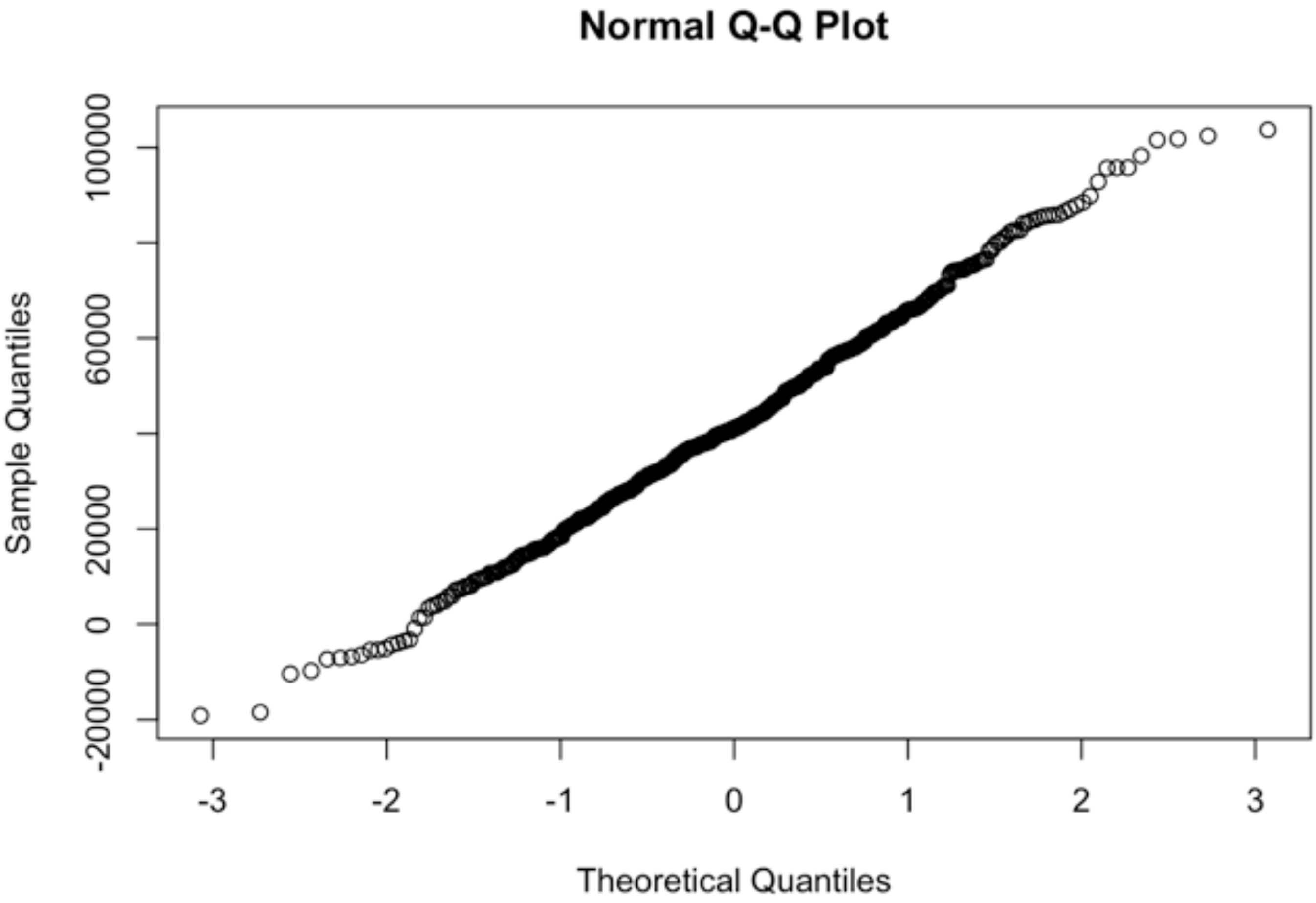
The normal quantile plot of the dataset does not appear to be normal, it deviates significantly from the straight line we’d expect from a normal normal distribution

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

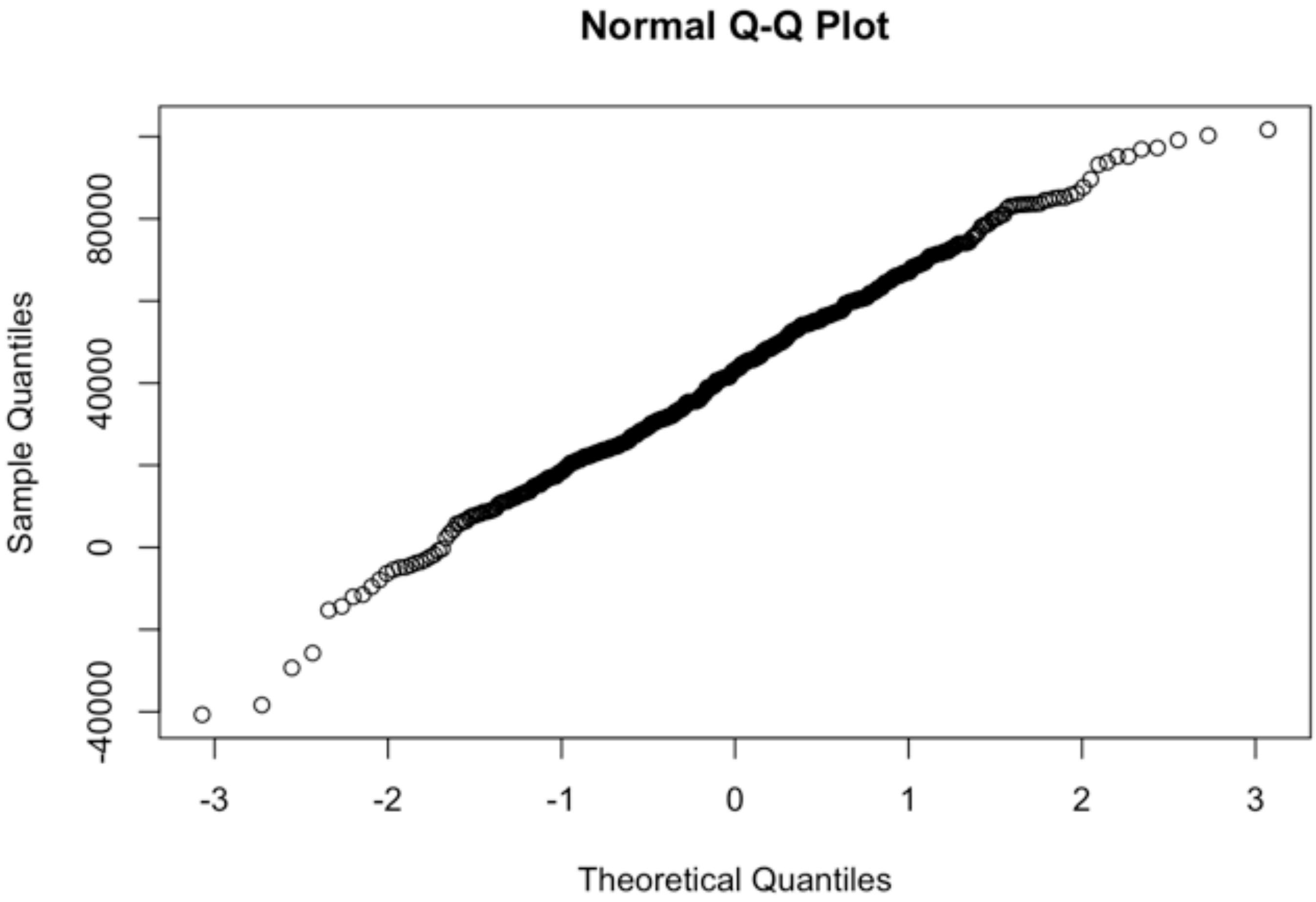
```
# Same process as part b, just need to use qqnorm and qqline instead of hist. Do no t
forget to discuss your findings.
qqnorm(sim1)
```


Normal Q-Q Plot
Sample Quantiles
-20000
0
20000
60000
100000
Theoretical Quantiles
-3
-2
-1
0
1
2
3

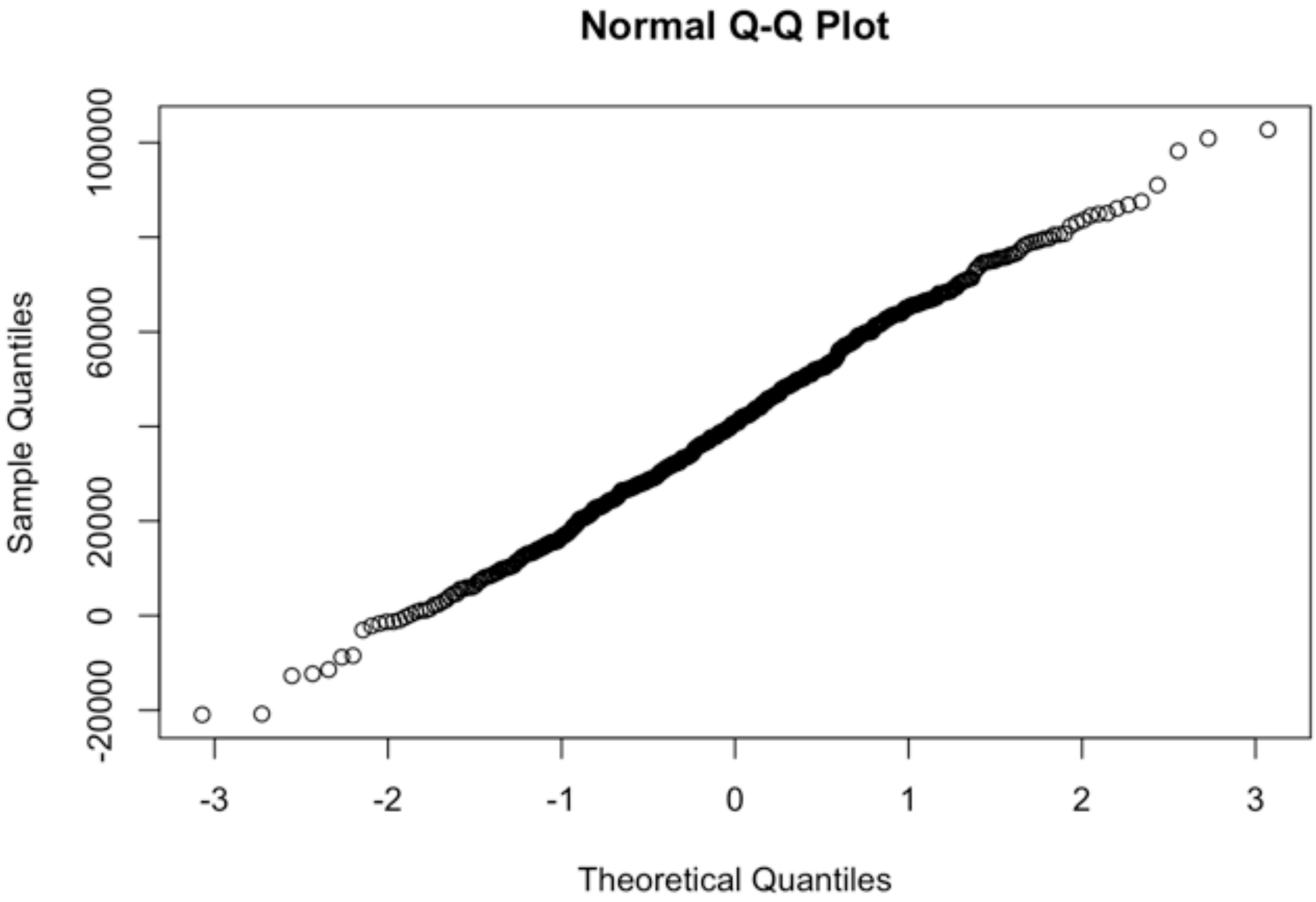
```
qqnorm(sim2)
```


Normal Q-Q Plot
Sample Quantiles
-20000
0
20000
60000
100000
Theoretical Quantiles
-3
-2
-1
0
1
2
3

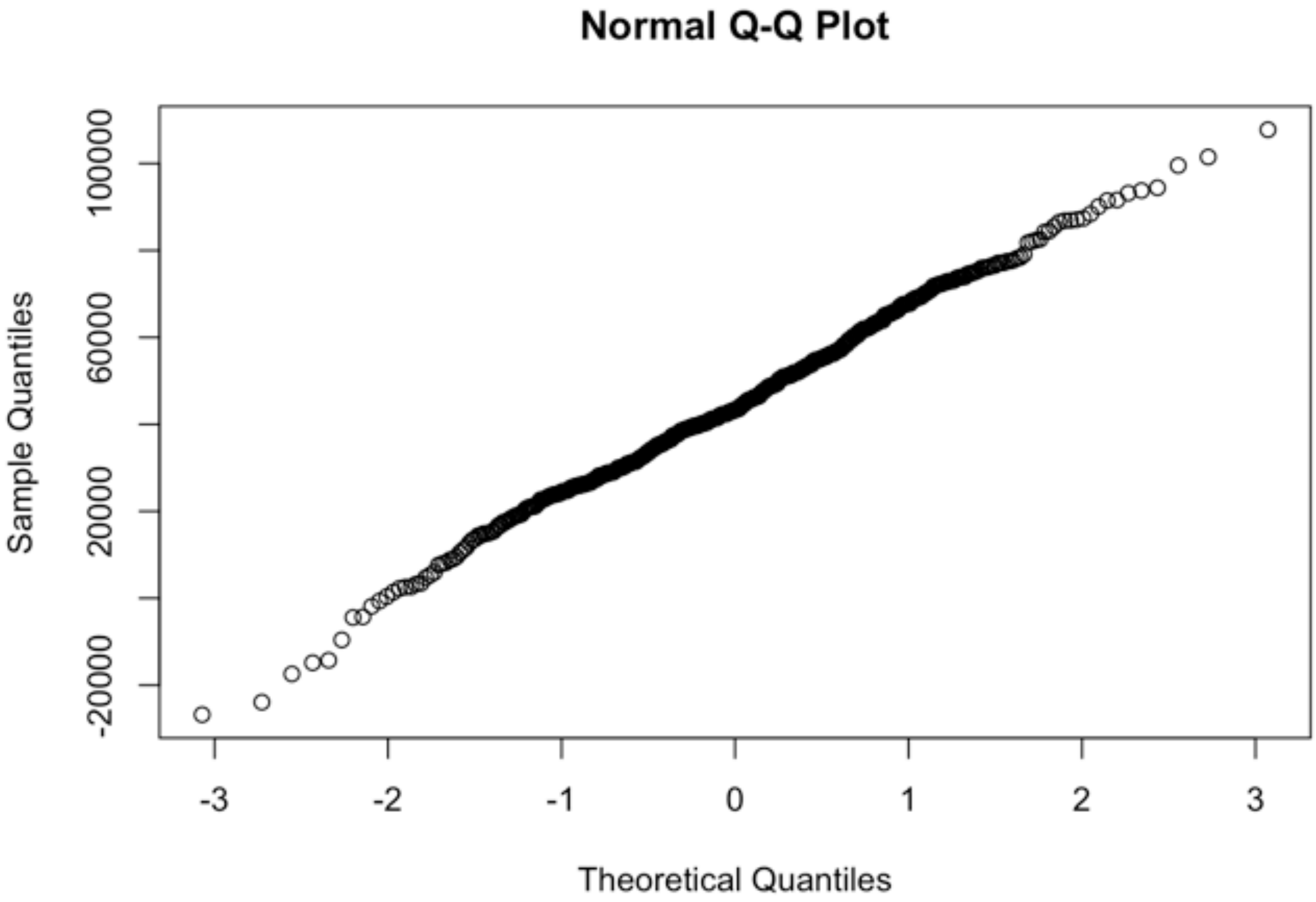
```
qqnorm(sim3)
```


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

```
qqnorm(sim4)
```


Normal Q-Q Plot
Sample Quantiles
-20000
0
20000
60000
100000
Theoretical Quantiles
-3
-2
-1
0
1
2
3

```
qqnorm(sim5)
```



These normal quantile plots all look very close to the line we'd expect in a normal distribution. Some rise slightly above it, but not enough that I could call them not normal by sight.

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 them dip significantly below the line as seen in the q-q plot of the dataset. In the simulations I produced, none of them look like the original data.

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)

My simulations are also all not distributed similarly to the original data, therefore my p-value is 0 and I can also reject the null hypothesis and conclude the original data is not normally distributed.

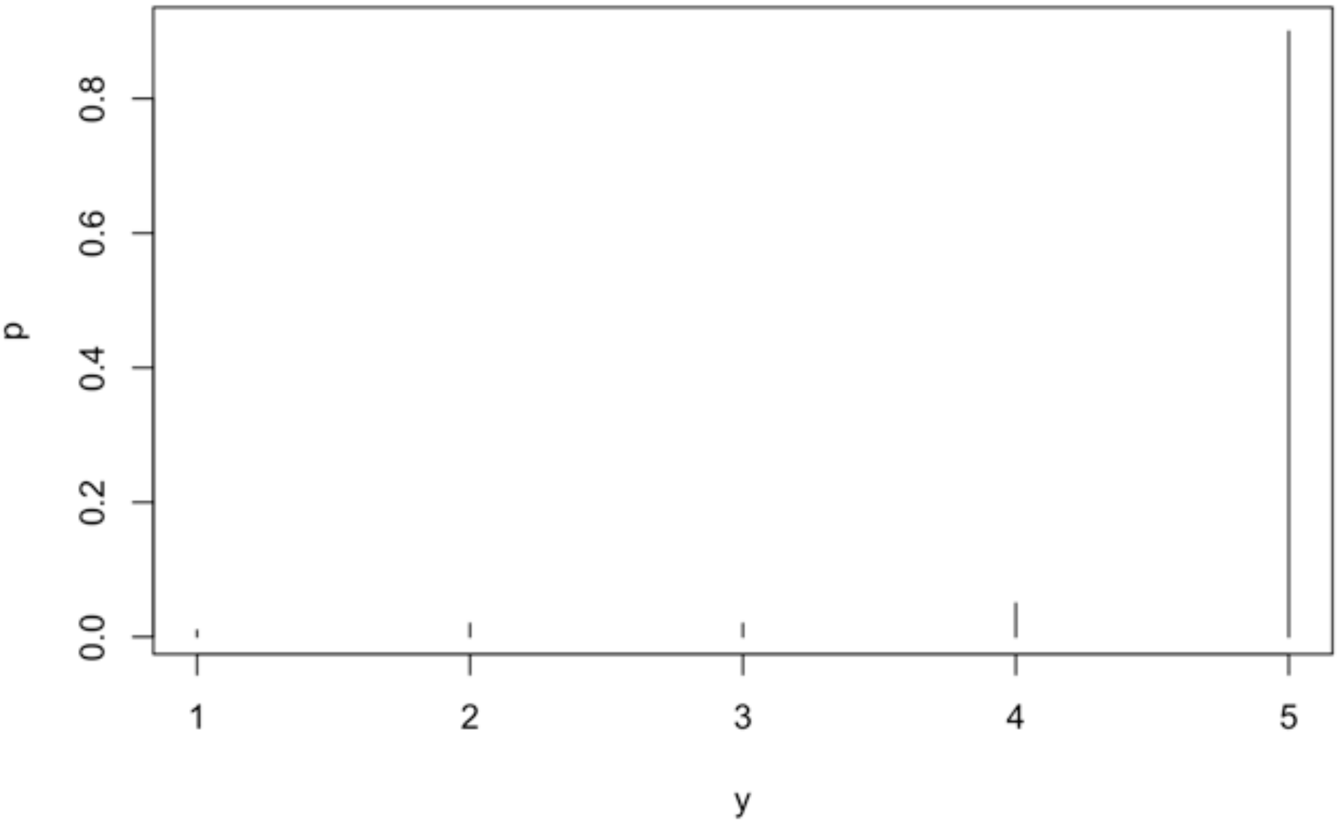
Problem 2 (40 points)

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)

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

# Plot probabilities of y
plot(y, p, type = 'h')
```



Y is not normally distributed.

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

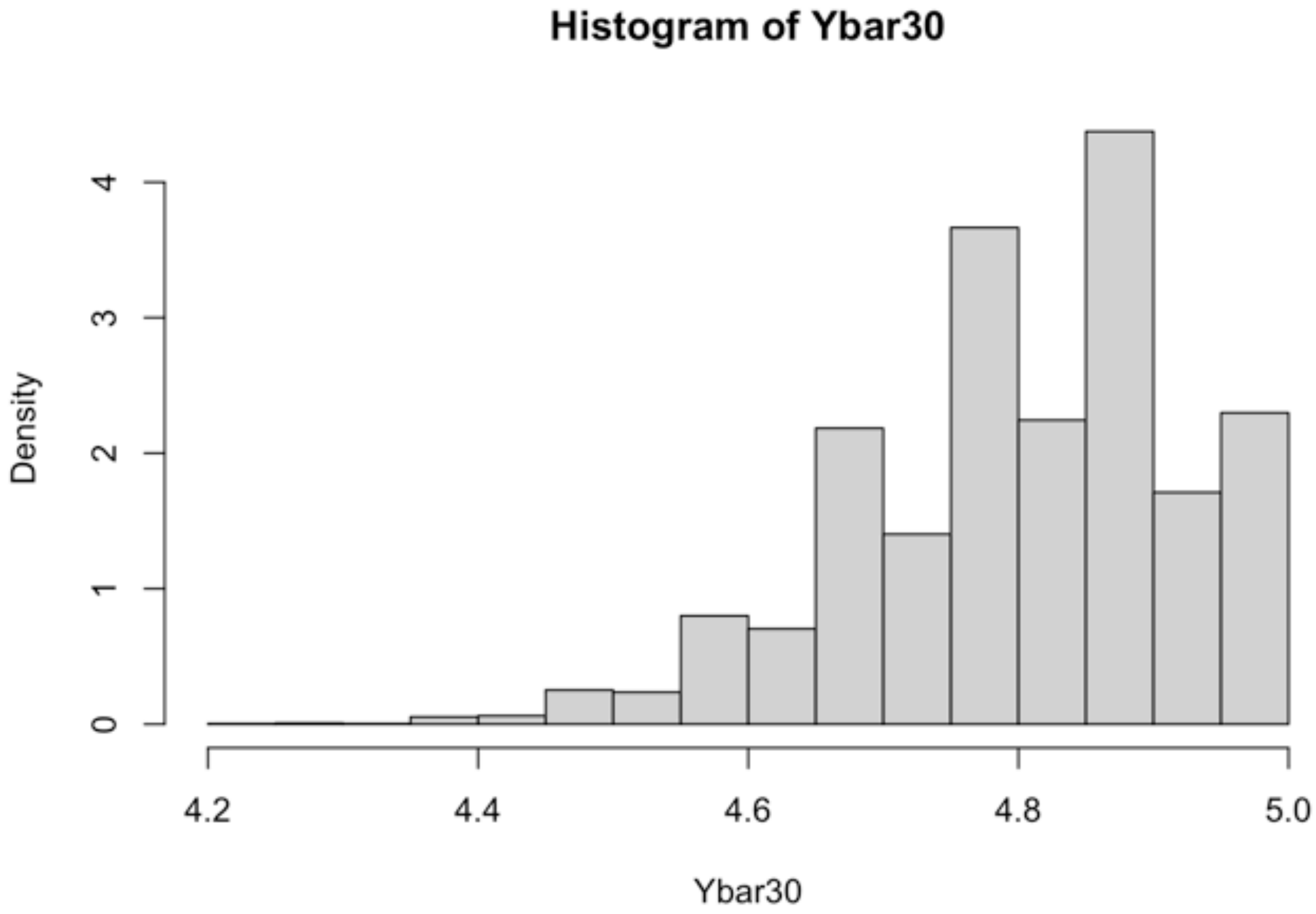
```
y = c(1, 2, 3, 4, 5)
p = c(.01, .02, .02, .05, .90)

# Define n, the simulation size, and multiply to create sample size
n = 30
simSize = 10000
ntotal = n*simSize

# Create simulation, store in sim
sim <- sample(y, ntotal, p, replace = T)

# Create matrix from sim dividing values into 30 columns
simMatrix <- matrix(sim, ncol = 30)

# Get means of rows and store in Ybar30 (30 for sample size)
Ybar30 <- rowMeans(simMatrix)
head(Ybar30)
```

```
## [1] 4.833333 4.766667 4.966667 4.833333 4.700000 4.633333
```

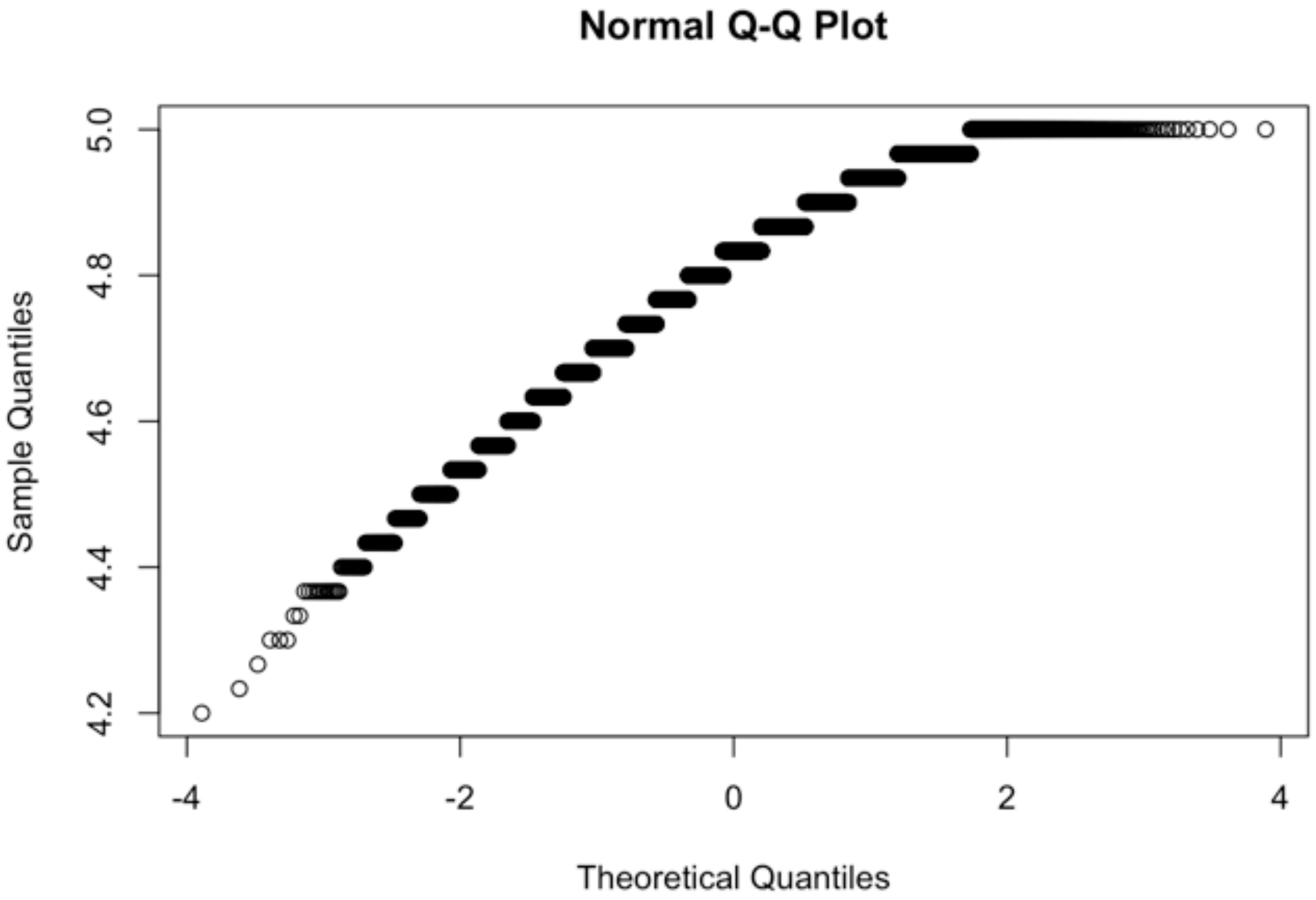
The row means aggregate around 4.8 which can be viewed in Ybar30

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)

```
hist(Ybar30, freq = F)
```



```
qqnorm(Ybar30)
```



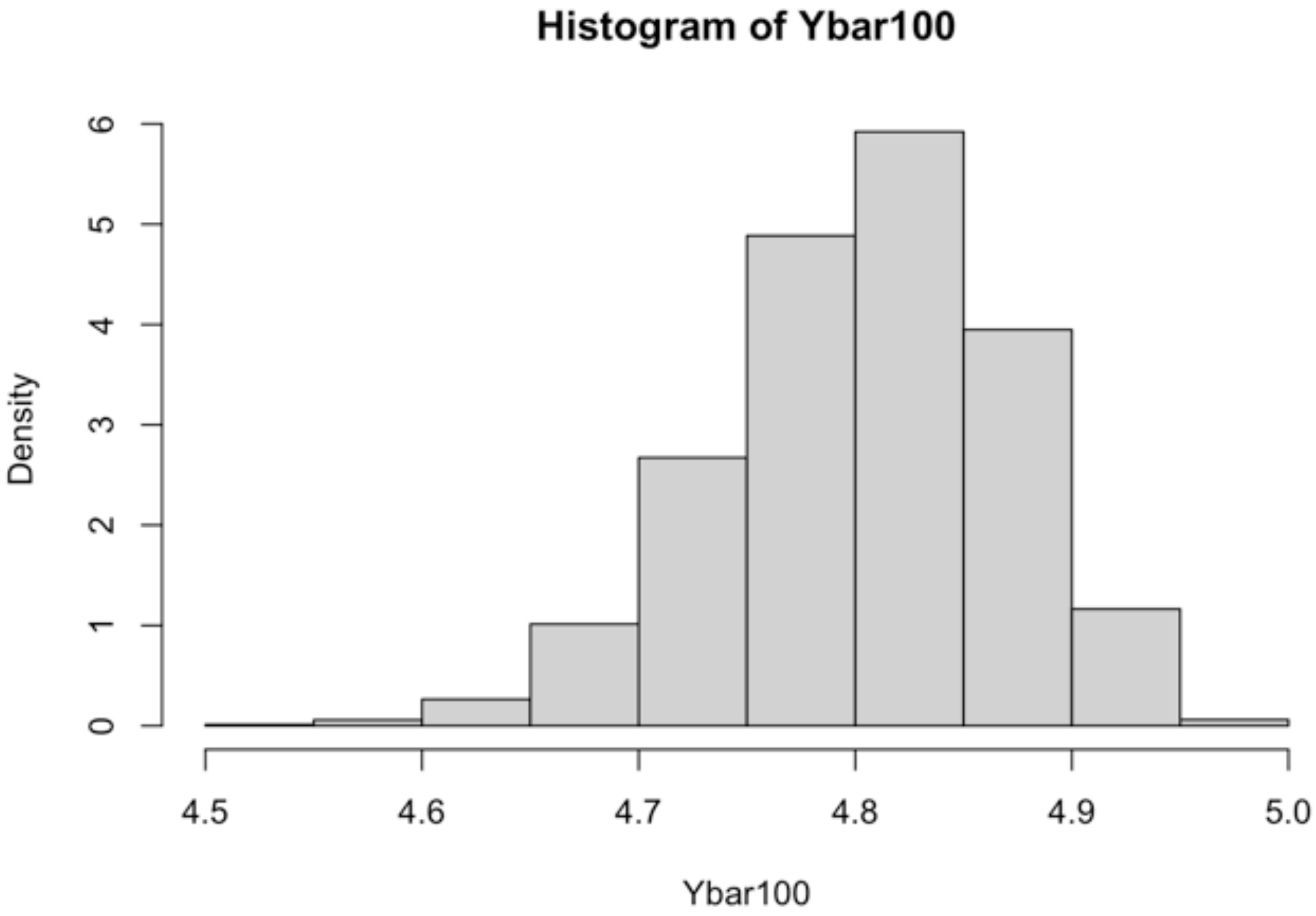
This distribution is not normal. However, this n is not large enough to confirm CLT and we should increase it to be sure.

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)

```
# Define n, the simulation size, and multiply to create sample size
n = 100
simSize = 10000
ntotal = n*simSize

# Create simulation, store in sim
sim <- sample(y, ntotal, p, replace = T)

# Create matrix from sim dividing values into n columns
simMatrix <- matrix(sim, ncol = n)

# Get means of rows and store in Ybarn (n depending on sample size)
Ybar100 <- rowMeans(simMatrix)
head(Ybar100)
```

```
## [1] 4.80 4.80 4.71 4.66 4.84 4.68
```

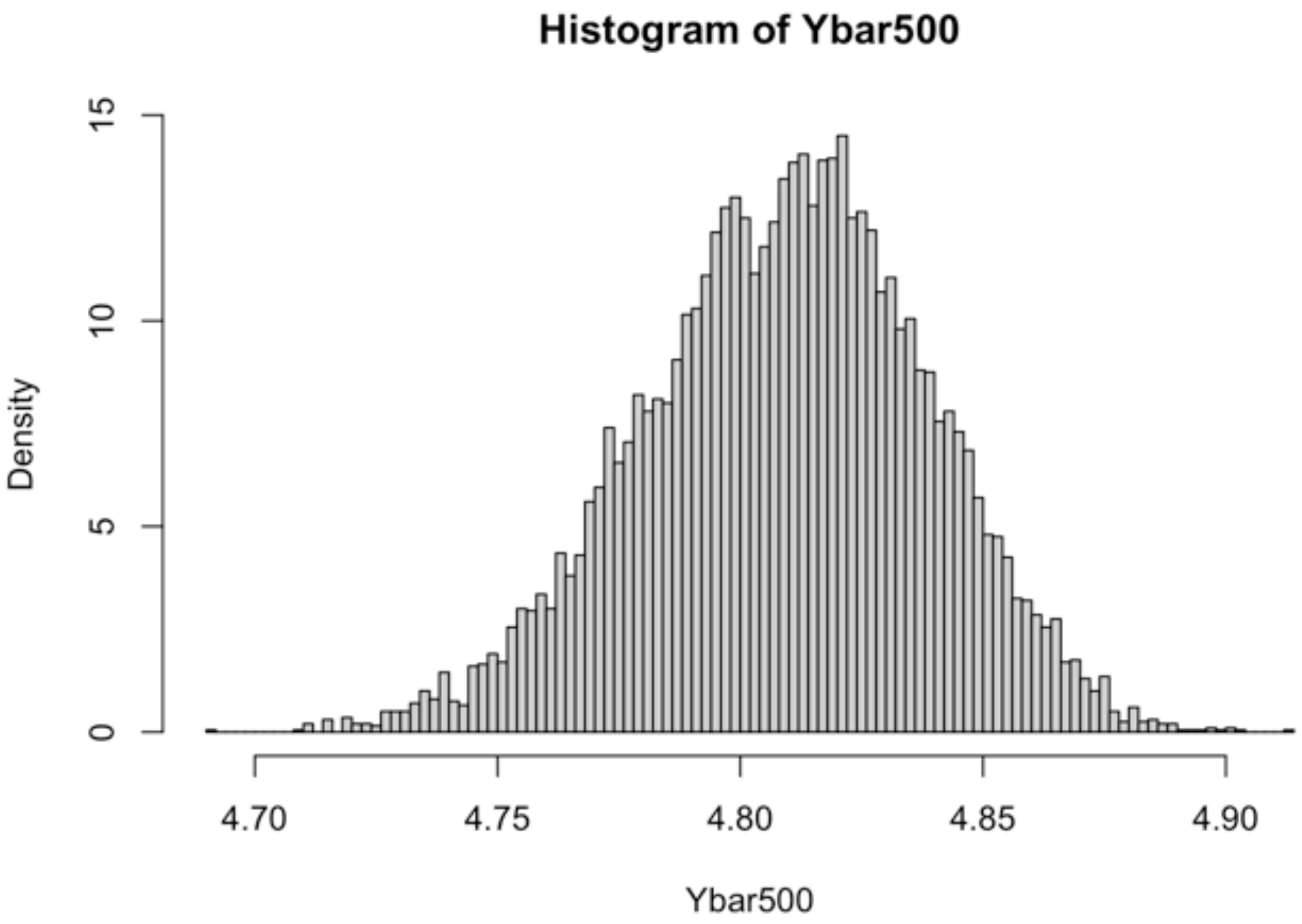
```
hist(Ybar100, freq = F)
```



The histogram is approaching a normal distribution, but still skews left.

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

```
n = 500
simSize = 10000
ntotal = n*simSize

# Create simulation, store in sim
sim <- sample(y, ntotal, p, replace = T)

# Create matrix from sim dividing values into n columns
simMatrix <- matrix(sim, ncol = n)

# Get means of rows and store in Ybarn (n depending on sample size)
Ybar500 <- rowMeans(simMatrix)

# Add breaks to see more precision
hist(Ybar500, freq = F, breaks = 100)
```



At n = 500 we can finally see a normal distribution, with a slight left tail. This is still different from the heavily left skewed histogram we started with. By adding breaks we can see more precision.