STAT S 520
HOMEWORK 9
RAMPRASAD BOMMAGANTY(rbommaga)

PROBLEM SET B:

3.

a. The experimental unit is the aerobics student .
b. The experimental units were drawn from a single population. Only one unit was drawn from each population. This is a 1 sample problem.
c. There are two measurements that are taken on each experimental unit i.e, each aerobics student. One of the measurement is the number of watts expended by the student after static stretching and the other one is the number of watts expended by the student after dynamic stretching.
d. If we consider X_i to be the random variable for the number of watts expended for dynamic stretching and Y_i to be the random variable for the number of watts expended for stretching stretching then,

$D_i = X_i - Y_i$, where the random variable D is the difference of the two random variables X and Y.

The parameter of interest for this problem is μ where $\mu = ED_i$, since it is a 1-sample problem.

e. The appropriate null hypothesis and alternative hypothesis are as follows:

$H_0 : \mu \leq 0$
$H_1 : \mu > 0$

PROBLEM SET C:

1.

a. The experimental units are the heavy middle aged men.
b. The experimental units were drawn from two populations. One population consisted of heavy middle aged men who were characterized by urgency, aggression and ambition. The other population consisted of heavy middle aged men who were non-competitive, more relaxed and less hurried. One unit was drawn from each population. It is a 2 - sample problem.
c. One measurement was taken from each experimental unit i.e, each heavy middle aged man. The measurement was the cholesterol level from persons belonging to both the populations.

d. If we consider X_i to be the cholesterol level measurement from the population displaying Type A behavior - urgency, aggression and ambition and Y_j to be the cholesterol measurement from the population displaying Type B behavior - non-competitive, more relaxed and less hurried, then the parameter of interest to be considered is Δ, where $\Delta = \mu_1 - \mu_2$ where $\mu_1 = EX_i$ and $\mu_2 = EY_j$.

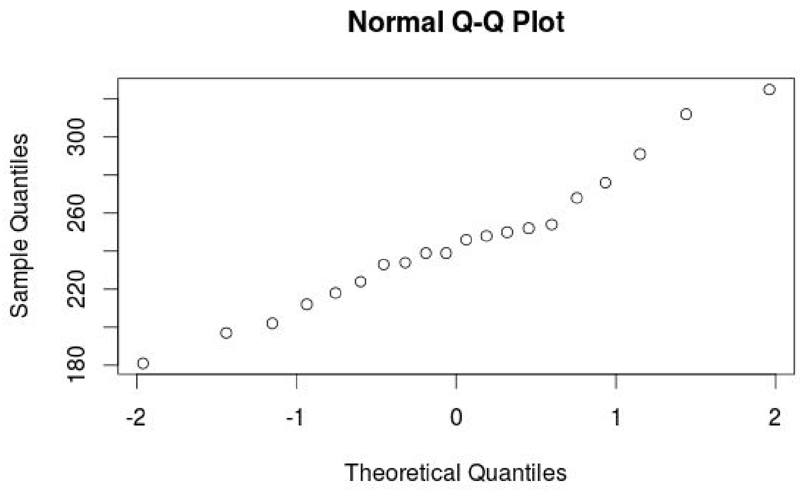
e. The appropriate null and alternative hypothesis are as follows:

$H_0 : \Delta \leq 0$
$H_1 : \Delta > 0$

2. The R code is as follows:

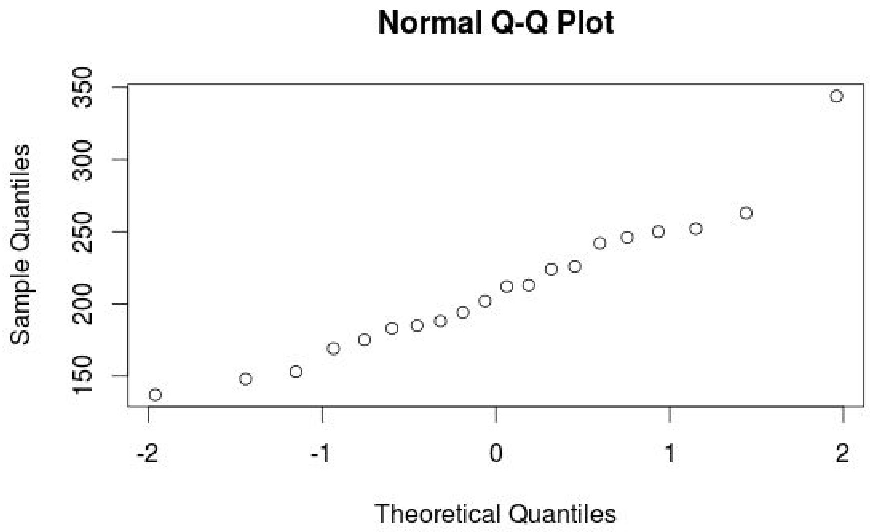
```
> dataset <- scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/cholesterol.dat")
Read 40 items
> print(dataset)
 [1] 233 291 312 250 246 197 268 224 239 239 254 276 234 181 248 252 202 218 212
325 344 185
[23] 263 246 224 212 188 250 148 169 226 175 242 252 153 183 137 202 194 213
> len <- length(dataset)
> print(len)
[1] 40
> x <- dataset[1:20]
> y <- dataset[21:40]
> qqnorm(x)
```

The normal probability plot for x is as follows:



```
> qqnorm(y)
```

The normal probability plot for y is as follows:



By looking at the normal probability plots above, it can be said that both the samples might have been drawn from normal distributions. But it cannot be said with near certainty because in the first plot, even though the plot stays a straight line for the most part, it does bend at one point. Similarly in the second plot, it stays more or less straight but there is a huge outlier at the end.

In conclusion, it might have been from a normal distribution, barring a few anomalies.

3. Given that the α value is 0.05

    a.

    The R code is as follows:

```
> delta_hat <- mean(x)-mean(y)
> tw = (delta_hat-0)/sqrt((var(x)/len_x) + (var(y)/len_y))
> numerator <- (var(x)/len_x + var(y)/len_y) ^ 2
> denom <- ((var(x)/len_x)^2/len_x-1) + ((var(y)/len_y)^2/len_y-1)
> v = numerator/denom
> print(v)
[1] 37.35923
> p = 1 -pt(tw, df = v)
> print(p)
[1] 0.007283272
```

The significance probability is 0.007 which is less than 0.05. Hence, we reject the null hypothesis.

b. Given that $1-\alpha = 0.90$.

Therefore, $1 - \alpha/2 = 0.95$.

The R code is as follows:

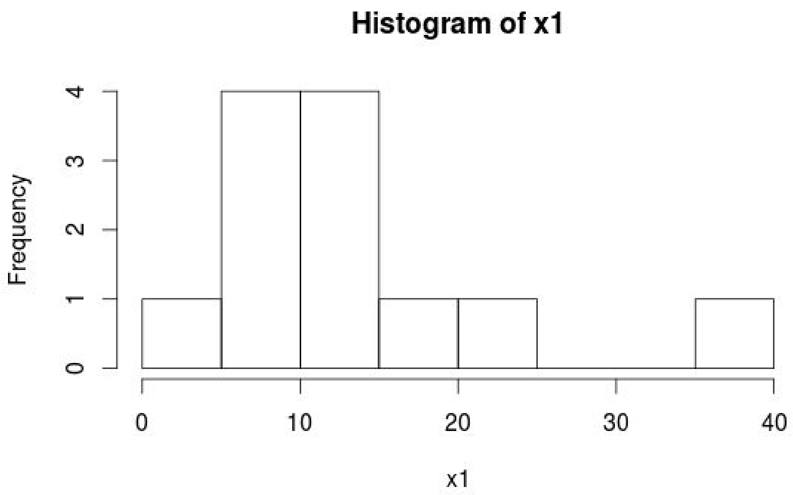
```
> num2 <- sqrt(numerator)
> qt = qt(0.95, df = v)
> int1 <- delta_hat+(qt*sqrt(num2))
> int2 <- delta_hat-(qt*sqrt(num2))
> print(int1)
[1] 57.62645
> print(int2)
[1] 11.87355
```

Therefore, the confidence interval lies between 11.87355 and 57.62645.

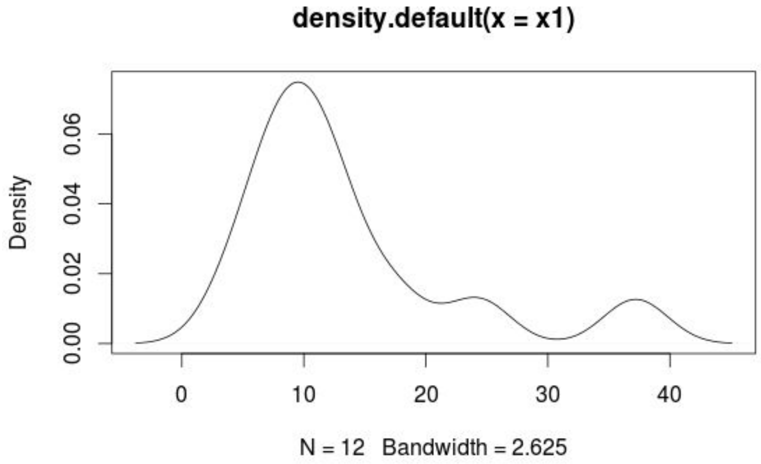
PROBLEM SET D:

1. The R code is as follows:

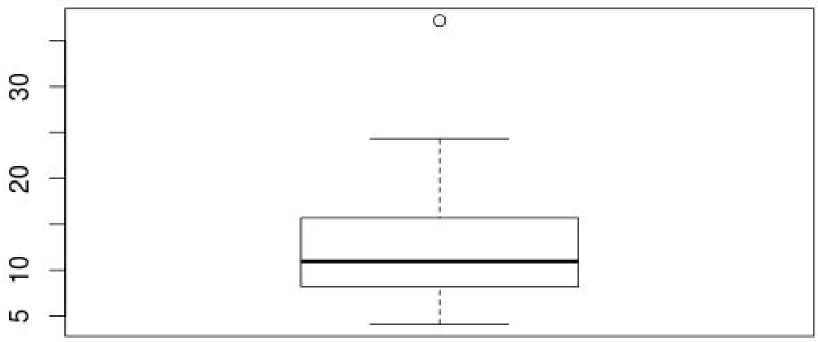
```
> dataset2 <- scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/globulin.dat")
Read 24 items
> print(dataset2)
 [1]  4.1  6.3  7.8  8.5  8.9 10.4 11.5 12.0 13.8 17.6 24.3 37.2 11.5 12.1 16.1 17.8 24.0 28.8
[19] 33.9 40.7 51.3 56.2 61.7 69.2
> x1 <- dataset2[1:12]
> y1 <- dataset2[13:24]
> hist(x1)
```



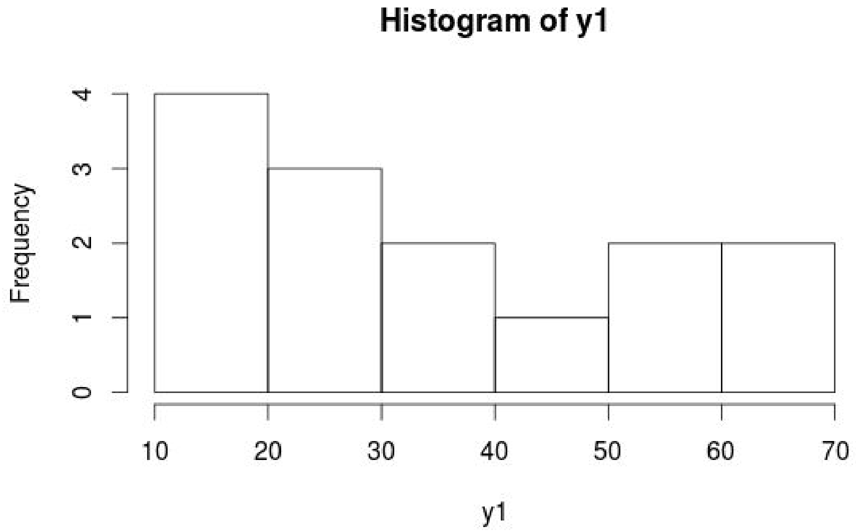
```
> plot(density(x1))
```



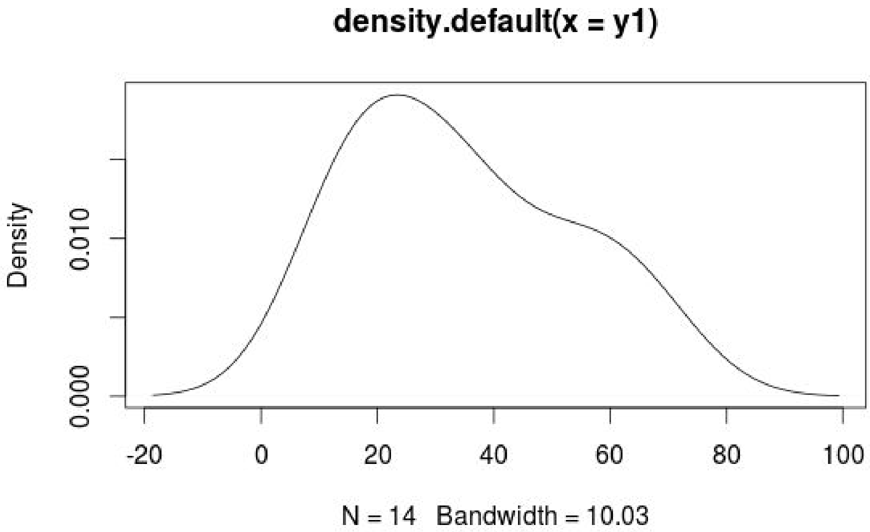
```
> boxplot(x1)
```



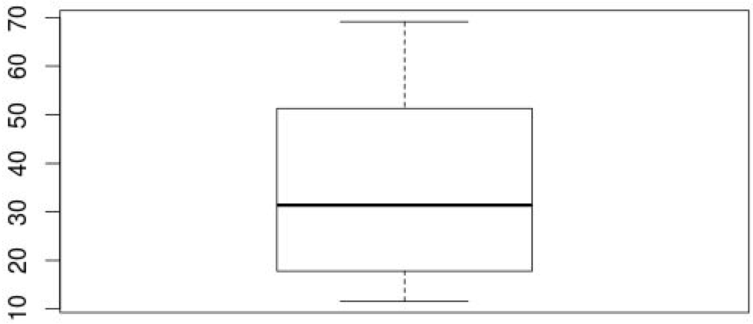
```
> hist(y1)
```



```
> plot(density(y1))
```



```
> boxplot(y1)
```

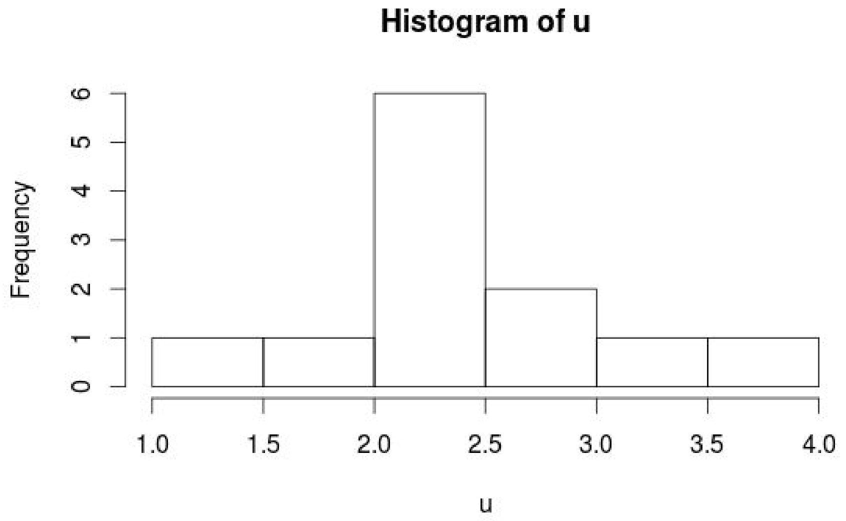


From the above plots for both the control group and the diabetic patients group, it can be said that the samples do not appear to be samples from symmetric distributions. From the plots, it can be observed that the figures are slightly right skewed and not symmetric.

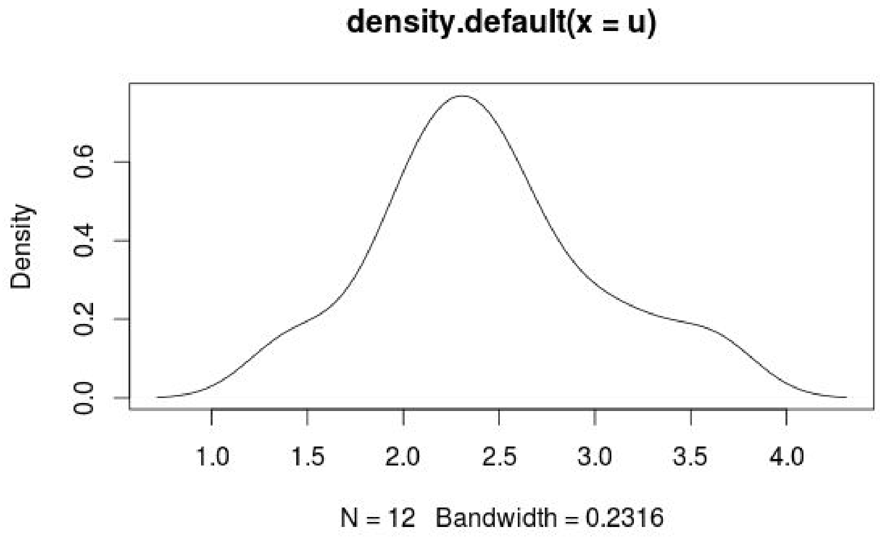
2.

The R code is as follows:

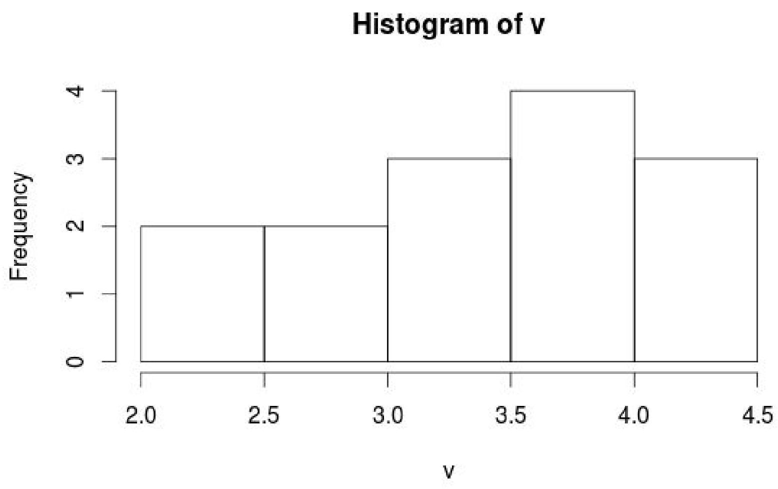
```
> u <- log(x1)
> v <- log(y1)
> hist(u)
```



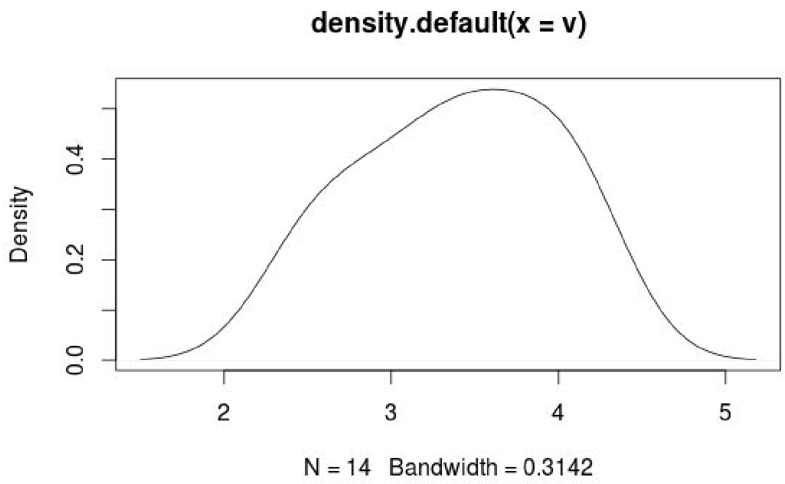
```
> plot(density(u))
```



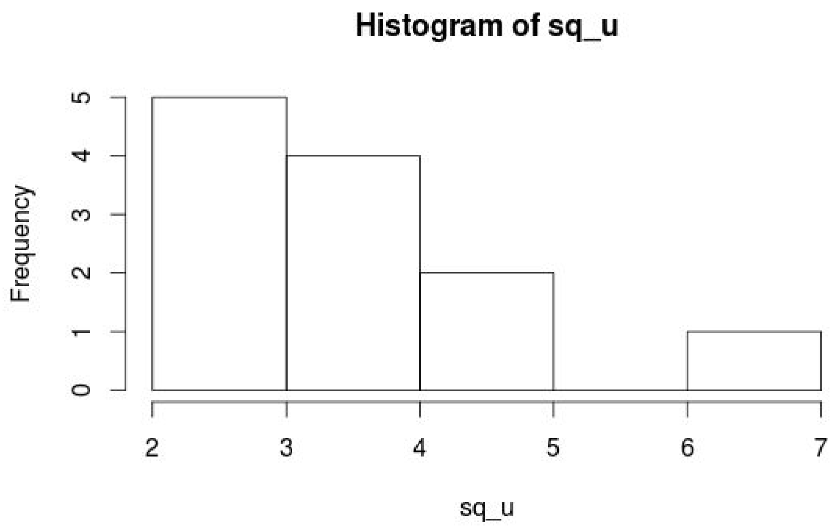
```
> hist(v)
```



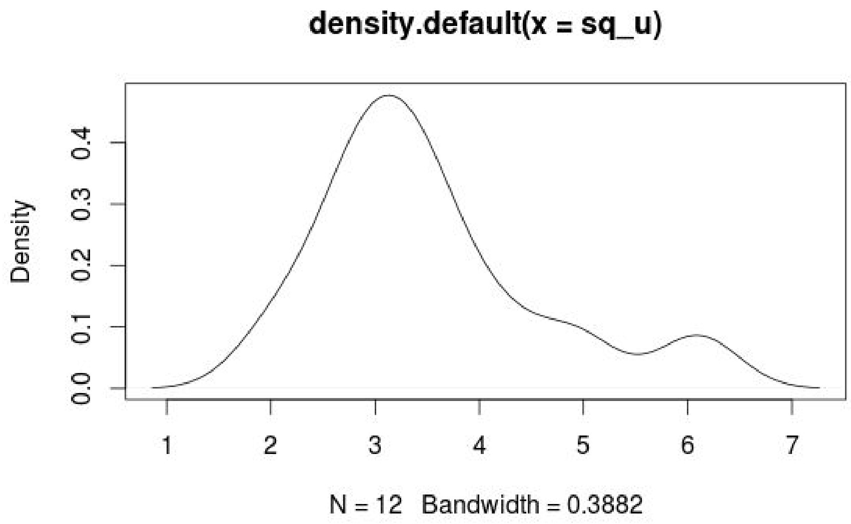
```
> plot(density(v))
```



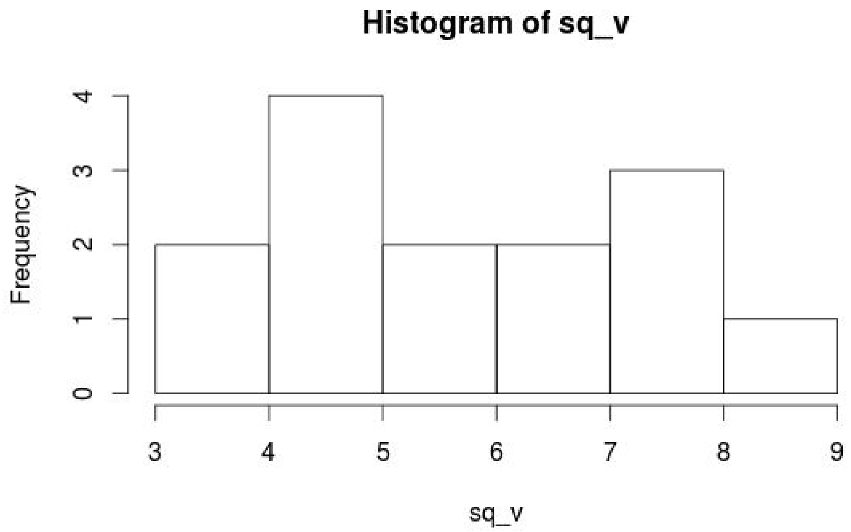
```
> sq_u <- sqrt(x1)
> sq_v <- sqrt(y1)
> hist(sq_u)
```



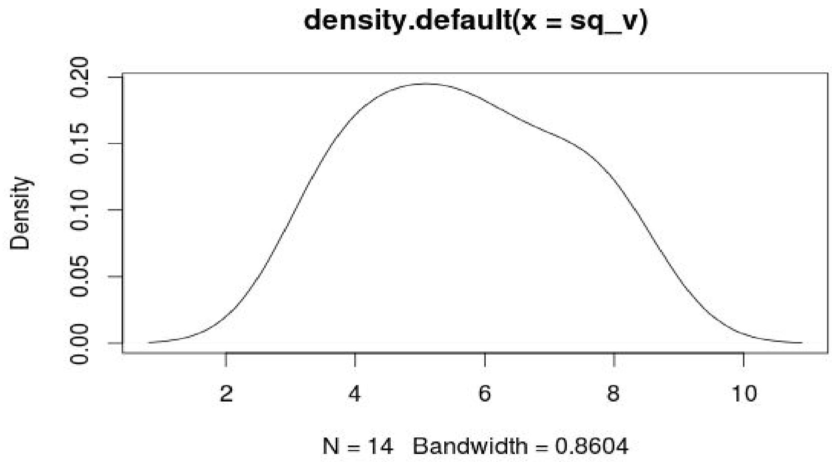
```
> plot(density(sq_u))
```



```
> hist(sq_v)
```



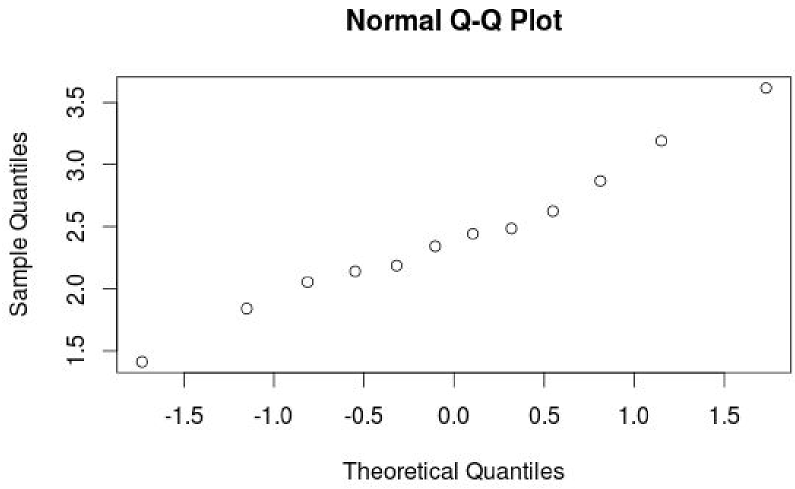
```
> plot(density(sq_v))
```



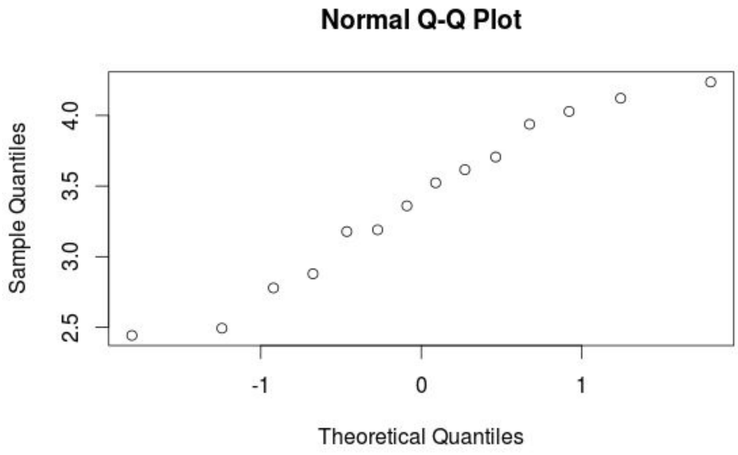
From the density plots, it looks like the samples undergoing the square root transformation have a better chance of being from symmetric distributions. Although both of them might be from symmetric distributions, I prefer the samples which have undergone square root transformation.

3. The R code is as follows:

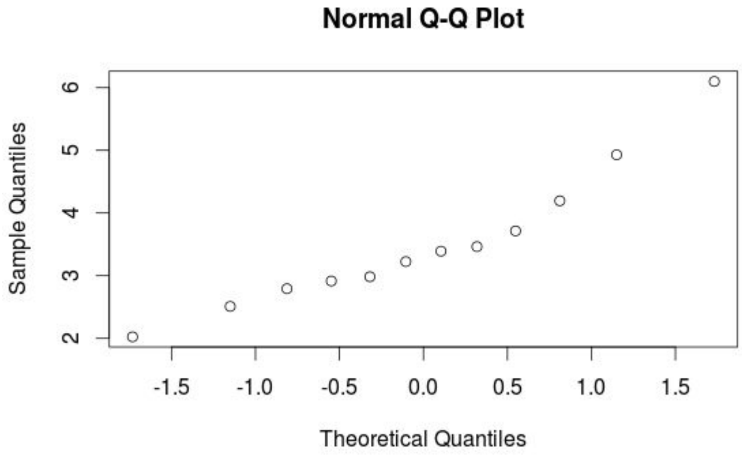
```
> qqnorm(u)
```



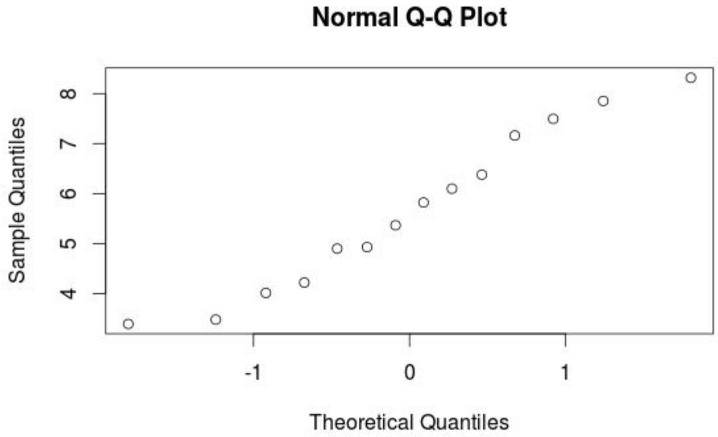
```
> qqnorm(v)
```



```
> qqnorm(sq_u)
```



```
> qqnorm(sq_v)
```



From the above normal probability plots, the transformed samples do appear like they were taken from a normal distribution. Among the two transformations, square root transformation does a better job of representing a straight line, especially in the case of sq_v plot.

4.

The researchers have claimed that diabetic patients have increased urinary β-thromboglobulin excretion. Since we have picked the square root transformation as the better approximation of normal distributions, let us pick square root transformation to apply the Welch's approximate t -test.

The R code is as follows:

```
> delta = mean(sqrt(y1)) - mean(sqrt(x1))
> tw = (delta)/sqrt(var(sq_v)/12 + (var(sq_u))/12)
> print(tw)
```

```
[1] 3.838114
> numerator = (var(sq_v)/12 + var(sq_u)/12) ^ 2
> denom = ((var(sq_v)/12)^2)/11 + ((var(sq_u)/12)^2)/11
> v = numerator/denom
> print(v)
[1] 19.50769
> p = 2*(1 - pt(abs(tw),df=v))
> print(p)
[1] 0.001065808
```

Since the p value is less than 0.05, we reject the null hypothesis which states that there is no increase in the value of thromboglobulin excretion. Therefore, the alternative claim that diabetic patients have increased urinary β- thromboglobulin excretion cannot be rejected.

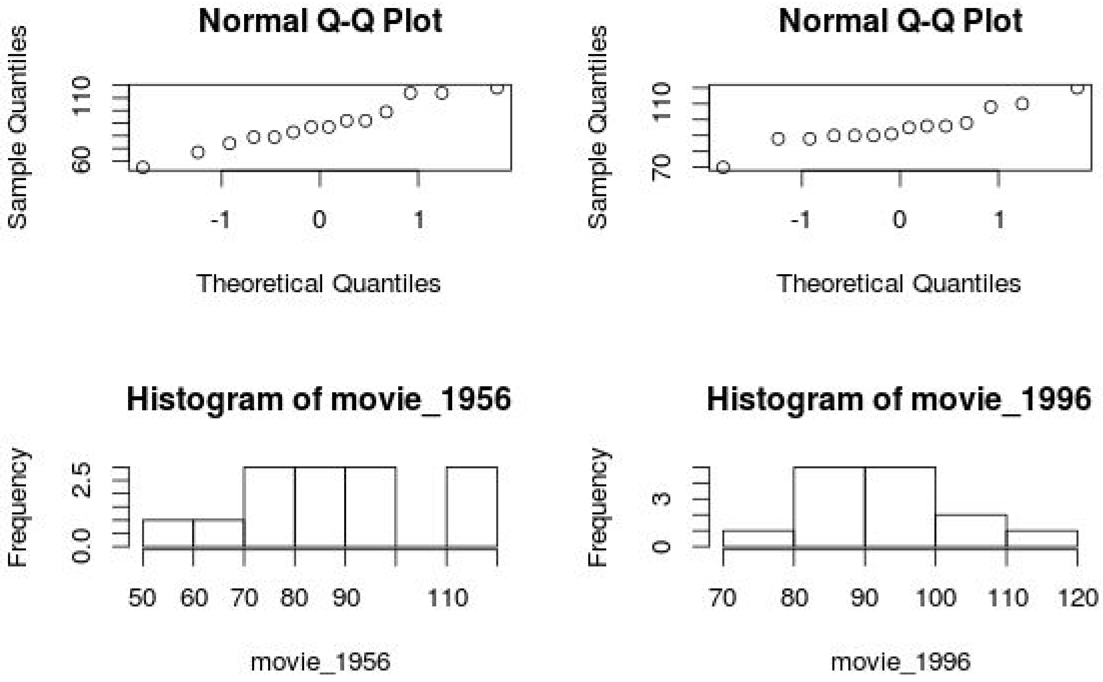
PROBLEM SET E:

3.

Let us first determine if the samples are from normal distributions:

The R code is as follows:

```
movie_data <- scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/films.dat")
movie_1956 <- movie_data[1:14]
movie_1996 <- movie_data[15:28]
par(mfrow = c(2,2))
qqnorm(movie_1956)
qqnorm(movie_1996)
hist(movie_1956)
hist(movie_1996)
```



From the above plots it can seen that the samples are not from a normal distribution.

Now, let us apply the Welch's 2-sample t-test:

The R code :

```
> t.test(movie_1996,movie_1956)
```

The observations are as follows:

```
        Welch Two Sample t-test

data:  movie_1996 and movie_1956
t = 1.105, df = 22.395, p-value = 0.2809
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.623821 18.480963
sample estimates:
mean of x mean of y 
 95.00000  88.57143 
```

Since the p value is greater than α , we cannot reject the null hypothesis that movies in 1996 were more or less of the same length that of 1956.

We can also apply the Wilcox test which focuses on general family shift to determine the same:

R Code:

```
> wilcox.test(movie_1996,movie_1956, conf.int = TRUE)
```

The observations are as follows:

```
        Wilcoxon rank sum test with continuity correction

data:  movie_1996 and movie_1956
W = 126, p-value = 0.2058
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -4.000047 17.999975
sample estimates:
difference in location
             7.000011
```

Since the p-value is greater than α, we cannot reject the null hypothesis here as well. Wilcox test essentially conveys the same information as that of the Welch's 2-sample t-test. One interesting observation to note is that the sample size of 14 each is too small to determine accurate results. In conclusion, the tests discourage the film buff's impression.