STAT S 520
HOMEWORK 8
RAMPRASAD BOMMAGANTY(rbommaga)

Problem SET A :

1.

a. Given that n = 400 and α = 0.05
Also, $\bar{x}$ = 3.194887 and s^2 = 104.0118

The test to be used in this situation is the ***Student's 1 - sample t test.***

The value of the test statistic is as follows:

T = $$\frac{3.194887 - 0}{\sqrt{\frac{104.0118}{400}}}$$

Therefore, test statistic T = 6.265334

b. Both third option and fourth option give the best approximation of the significance probability.
Third option gives a significance probability of 0.1050907, while the fourth option gives the probability of 0.1054576.

c. TRUE. The **p** value is less than α.

2. Given,

n=20, 1- α = 0.95
Therefore, α = 0.05

If we assume some value to be the median and consider the value c, which will be the minimum of the number of values greater than median and the rest, we can use the pbinom function calculate the significance probability:

For c value 4: 1-2*pbinom(4,20,0.5) = 0.9881821
For c value 5 : 1-2*pbinom(5,20,0.5) = 0.9586105

For c value 6: 1-2*pbinom(6,20,0.5) = 0.8846817
Among the above three trials, for c value = 5, we get the correct confidence coefficient of 0.9586.

Thus the interval here is x_6 to x_{14}.

After running the sort operation on the data, the interval can be estimated from 238 to 251.

PROBLEM SET B:

1. The four samples are as follows:

   RCODE:

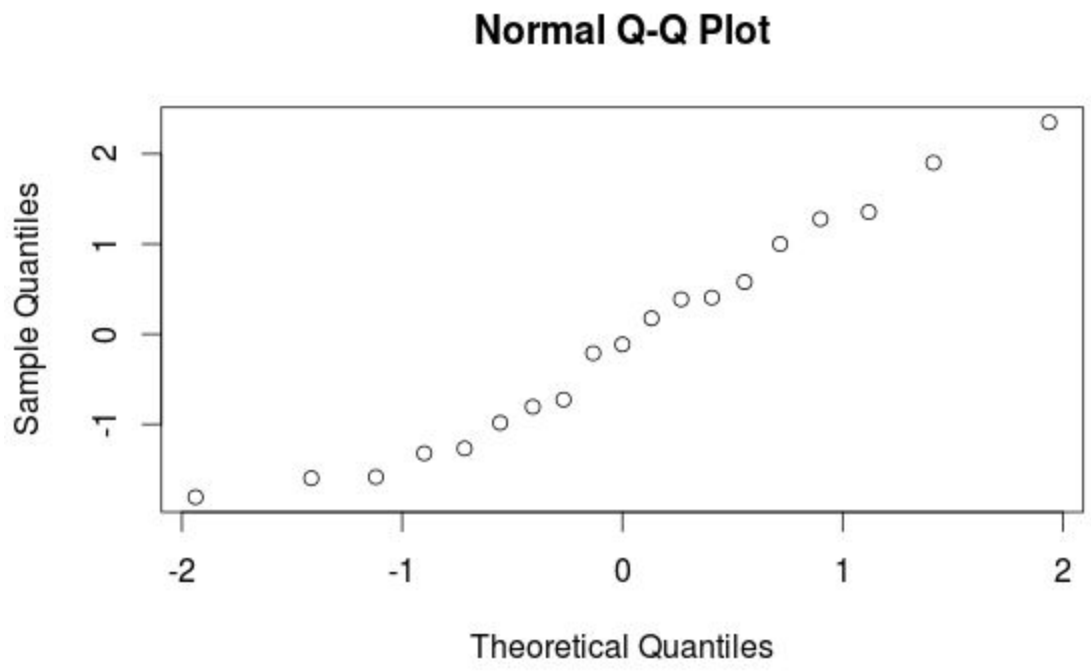
```
> sample1 = rnorm(19)
> print(sample1)
 [1] -1.3198858  0.1766393 -1.5833071 -1.5968682 -0.1132856  1.2771352 -0.2101412
0.4063501
 [9]  0.3870672  2.3490488 -1.8065075 -0.9826248  0.9997275 -0.8049705  1.3543361
0.5789444
[17] -0.7249064  1.9012337 -1.2671781

> sample2 = rnorm(19)
> print(sample2)
 [1]  1.270449757 -0.273843404  0.976915832  1.723334114  0.023854600  0.009242885
-0.102151168
 [8]  0.265095227  0.398130571 -0.475268185 -0.742403536 -0.253585939 -0.920581855
1.113599949
[15]  1.330148428  0.152338415 -2.385705703 -0.325096075 -1.671188966

> sample3 = rnorm(19)
> print(sample3)
 [1]  0.59462384  1.22990257 -0.58158856 -0.36398017 -1.04874570  0.10251881
-0.52118894
 [8]  0.20926016  0.60763298 -0.44877453  0.36564099  0.63986639 -0.04409405
-0.29774454
[15] -1.24841287 -0.20798291 -1.60818597  2.00128467 -1.01969658

> sample4 = rnorm(19)
> print(sample4)
 [1]  1.48525060  0.50926855 -0.99229591  0.71184783 -1.19215740  0.03595572
0.54136071
```

[8] -0.66722184 1.53515113 -1.42512722 0.05943145 1.22660779 0.89888252
-0.60458853
[15] 0.35998600 -0.02109047 -0.54317301 0.23689711 0.22088848

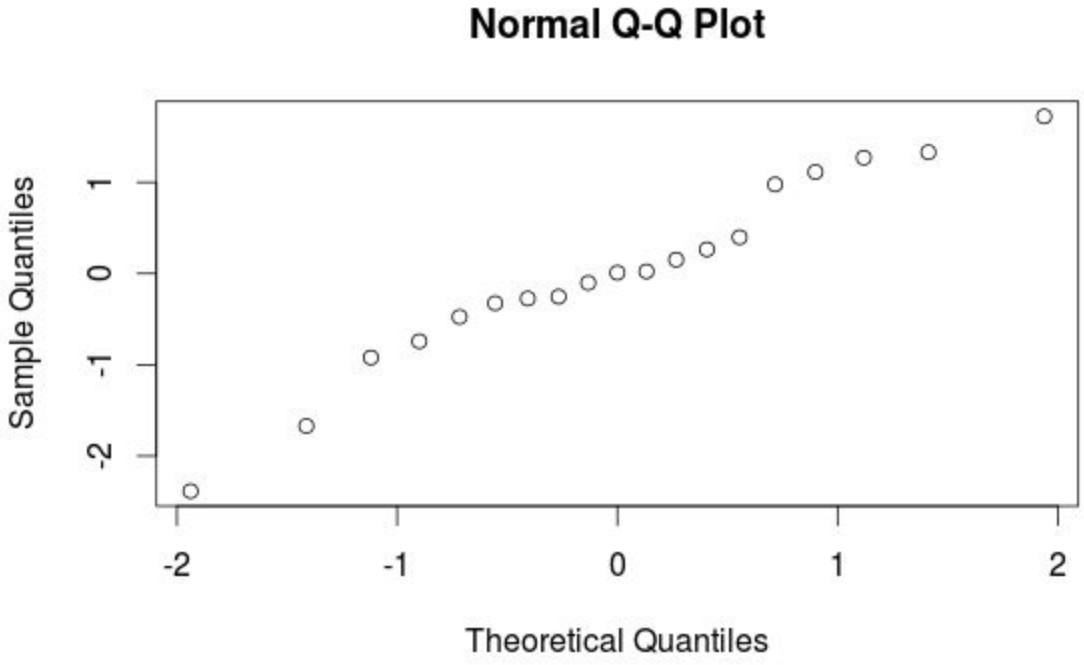
b.

The normal probability plots are as follows:

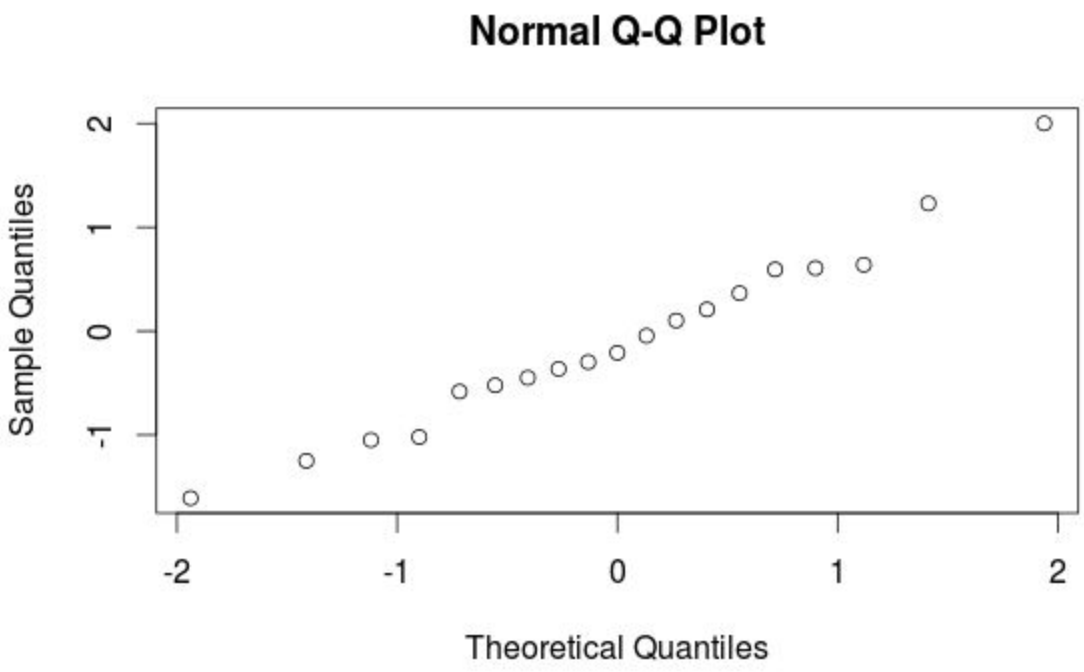
qqnorm(sample1)



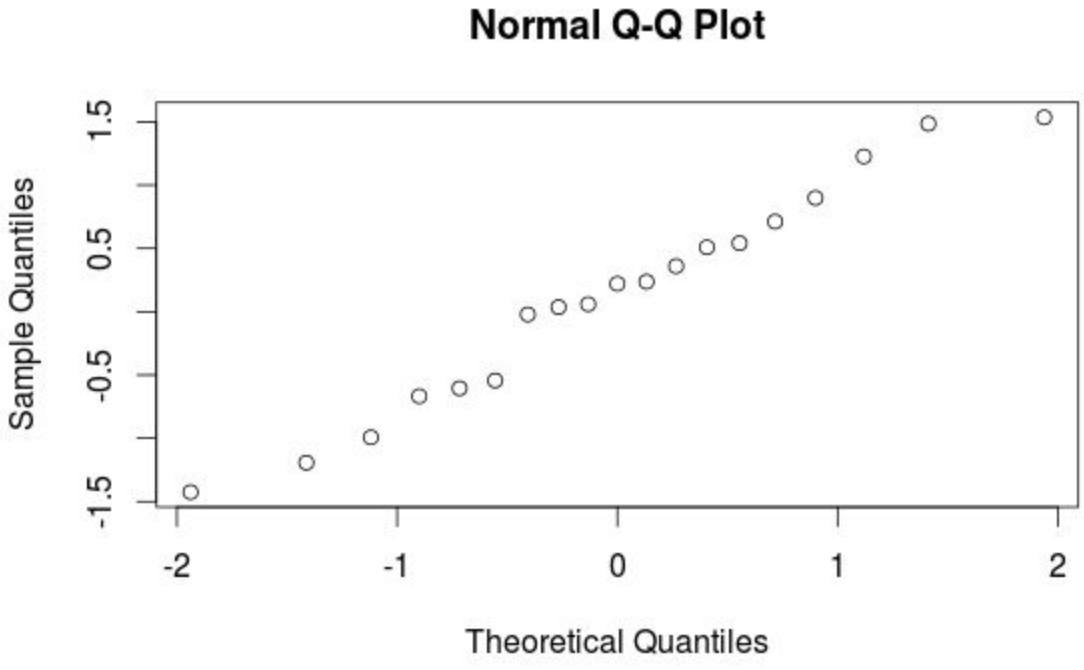
qqnorm(sample2)



qqnorm(sample3)



qqnorm(sample4)



The above plots seem to be similar to the normal probability plot from the figure 10.1 with similar tails.

c. Since $\vec{x}$ values have been taken from the final column of table 10.2, the R code for computation is as follows:

```
log2CPA <- c(1.1402, -1.8658, 0.8520, -1.8251, 0.8530, -0.0589, -1.6554, -1.7599, -1.4330, -1.3853, 2.9794, 2.4919, 2.1601, 2.2670, -0.5479, -0.7164, 0.6462, -0.8365, 1.1997)

print(log2CPA)

result = IQR(log2CPA)/ sd(log2CPA)
print(result)
```

The value of the result is 1.590833 i.e, the ratio of interquartile range to the standard deviation of $\vec{x}$.

For sample1,

```
result = IQR(sample1)/ sd(sample1)
print(result)
```

Ratio = 1.532275

For sample2,

```
result = IQR(sample2)/ sd(sample2)
print(result)
```

Ratio = 1.054153

For sample3,

```
result = IQR(sample3)/ sd(sample3)
print(result)
```

Ratio = 1.167196

For sample4,

```
result = IQR(sample3)/ sd(sample3)
print(result)
```

Ratio = 1.390343

No, it cannot really be said if the distribution was taken from a normal distribution. The reason being only one out of the four simulations are even remotely close to the observed data and we might need more than just four simulations to comment on the distribution of the observed data.

2.

```
statistic <- (mean(log2CPA)-0)/ (sd(log2CPA)/sqrt(19))
print(statistic)
```

The value of statistic = 0.3545188

```
> p = 1 - pt(statistic, df = 18)
> print(p)
[1] 0.3635348

> mean(log2CPA) - qt(0.95, df= 18)* sd(log2CPA) / sqrt(19)
[1] -0.5131008

> mean(log2CPA) + qt(0.95, df= 18)* sd(log2CPA) / sqrt(19)
[1] 0.7768166
```

Since p is much greater than α, we do not reject the null hypothesis. The two sided confidence interval with a confidence coefficient of 0.90 is [-0.513 0.776].

The interval width here = 1.289 is similar to the interval width from Section 10.4 = 1.028.

3.

```
> sorted_log2CPA <- sort(log2CPA)
> print(sorted_log2CPA)
 [1] -1.8658 -1.8251 -1.7599 -1.6554 -1.4330 -1.3853 -0.8365 -0.7164 -0.5479 -0.0589  0.6462
[12]  0.8520  0.8530  1.1402  1.1997  2.1601  2.2670  2.4919  2.9794
```

If y is considered as the number of values that are greater than 0, then y = 9.

Now,

```
> p <- 1 - pbinom(8,19,0.5)
> print (p)
[1] 0.6761971
```

Since p value is very high we do not reject the null hypothesis.

Given that 1 - α = 0.90 i.e, a 90% confidence interval.

By experimenting with k values of 4,5,6 and 7 we get :

```
[1] 0.8329315
> 1 - 2*pbinom(5,19,0.5)
[1] 0.9364319
> 1 - 2*pbinom(4,19,0.5)
[1] 0.9807892
> 1 - 2*pbinom(7,19,0.5)
[1] 0.6407166
```

We select k =5 and hence determine : x_6 to x_{14}
Hence, the interval ranges from -1.3853 to 1.1402.

This interval width = 2.5255 is much larger than 1.028, the interval width from Section 10.4

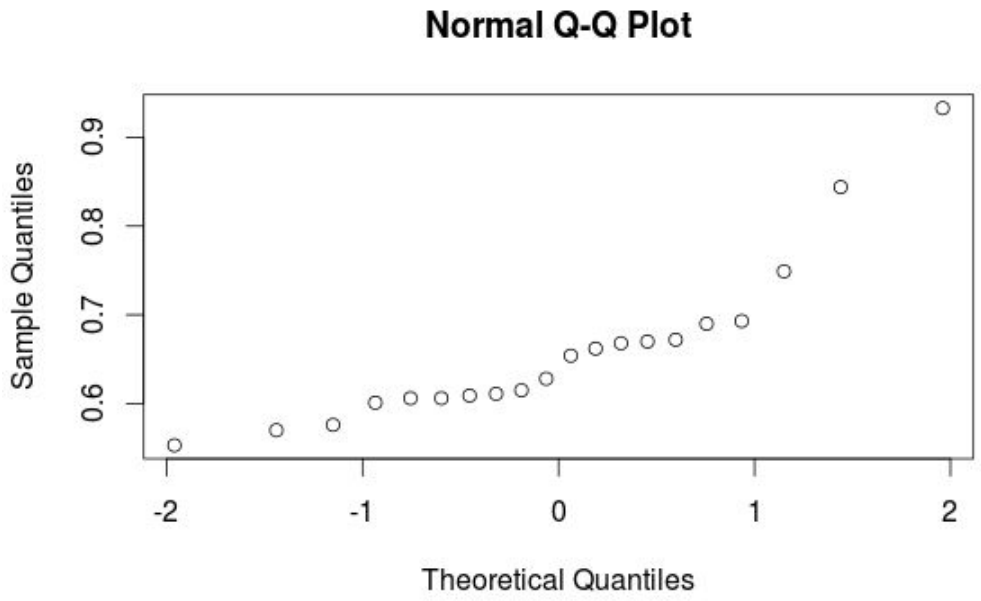
PROBLEM SET D:

1.

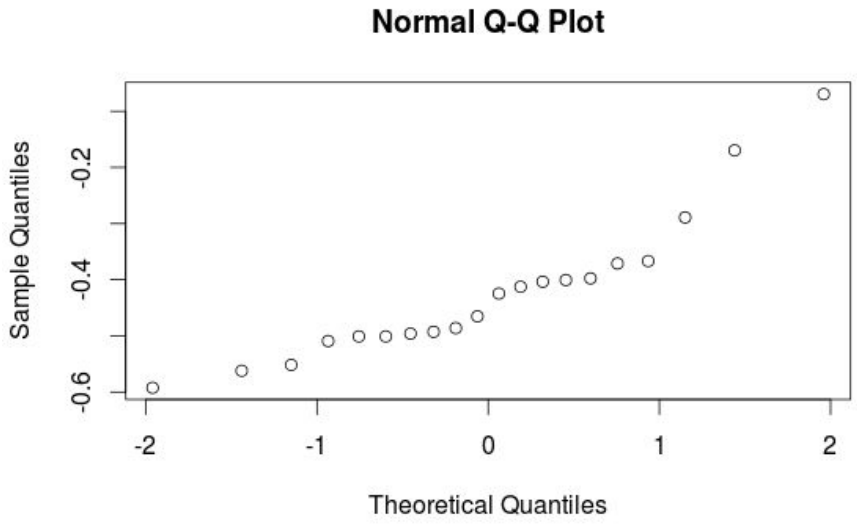
The R code is as follows:

```
> ratio <-
c(0.693,0.662,0.690,0.606,0.570,0.749,0.672,0.628,0.609,0.844,0.654,0.615,0.668,0.601,0.57
6,0.670,0.606,0.611,0.553,0.933)
> print(ratio)
 [1] 0.693 0.662 0.690 0.606 0.570 0.749 0.672 0.628 0.609 0.844 0.654 0.615 0.668 0.601
0.576
[16] 0.670 0.606 0.611 0.553 0.933
> log_ratio <- log(ratio)
> print(log_ratio)
 [1] -0.36672528 -0.41248972 -0.37106368 -0.50087529 -0.56211892 -0.28901630
-0.39749694
```

[8] -0.46521511 -0.49593701 -0.16960278 -0.42464793 -0.48613301 -0.40346711
-0.50916034
[15] -0.55164762 -0.40047757 -0.50087529 -0.49265832 -0.59239728 -0.06935008
> qqnorm(ratio)



> qqnorm(log_ratio)



Neither of the above plots looks like the one of a normal distribution. But if we had to strictly choose between any one of the above, the log_ratio plot seems a little better in comparison to ratio plot.

2.

Let us consider that we picked the logged ratio data. The hypotheses are :

$H_0 = \mu = \log(0.618034)$

$H_1 = \mu \neq \log(0.618034)$

Using Student's 1-sample t-test we get the following:

RCODE:

```
> t_statistic <- (mean(log_ratio)-log(0.618034))/
(sd(log_ratio)/sqrt(length(log_ratio)))
> print (t_statistic)
[1] 2.020006

> p = 2*(1 - pt(abs(t_statistic),df = length(log_ratio)-1))

> print(p)
[1] 0.05771066
```

Thus, the value of p = 0.577. Thus, we do not reject the null hypothesis.

3.

Applying trial and error runs for finding the appropriate k value such that the confidence interval is 90%.

R CODE:

```
> qbinom(0.05,length(ratio),0.5)
[1] 6
> 1 - 2*pbinom(6,length(ratio),0.5)
[1] 0.8846817
> 1 - 2*pbinom(5,length(ratio),0.5)
[1] 0.9586105
> 1 - 2*pbinom(7,length(ratio),0.5)
[1] 0.736824
```

We can choose k = 6 because it is the closest to 90% confidence.

Now,

```
> sort(ratio)
 [1] 0.553 0.570 0.576 0.601 0.606 0.606 0.609 0.611 0.615 0.628 0.654 0.662
0.668 0.670 0.672
[16] 0.690 0.693 0.749 0.844 0.933
```

The confidence interval for the population median is as follows: 0.606 to 0.672

Problem No: 4:

a.

Given that the distribution of the change in test scores is a normal distribution with mean = 6.5 and sd = 12.
The sample size = 61

$H_0 = \mu \geq 0$
$H_1 = \mu < 0$

RCODE:

```
> z = (6.5 - 0)/ (12/sqrt(61))
> print (z)
[1] 4.230552
>
> p <- pnorm(z)
> print(p)
[1] 0.9999883
```

Since, p value is much greater than α, we cannot reject the null hypothesis.

b.

Yes, the study proves that live reggae music improves the math scores. The significance probability is much greater than α, thereby eliminating any possibility to reject the null hypothesis that there are negative changes in test scores. This implies that test scores are positive on the average and hence reggae music has contributed to an increase in test scores.