PROBLEM SET 10

VINAY KUMAR RANGANATH BABU

(10 points.) We wish to study the effects of four different rat feeds (fruit, carbs, meat, andmixed) on rat weight gain. 140 rats, each kept in a separate cage, are randomly split into fourequal groups of 35. Each group is put on a different feed for a month. The amount of weight gained during the month is measured for each rat. The sample means and sample standarddeviations of the results (all in grams) are:

• Fruit: mean 83.5, standard deviation 16.9

• Carbs: mean 92.3, standard deviation 14.6

• Meat: mean 88.6, standard deviation 14.2

• Mixed: mean 99.4, standard deviation 14.1

Figure 1 (on the next page) shows normal quantile plots of the data.
(a) What are the assumptions of the analysis of variance F-test? Are the data here approximately consistent with these assumptions?

Assumptions:

All samples are independent

All variances are equal

Population is normal

- Looking at the sd it shows they are nearly same. Also for the experiment sake wew are good to accept that. Also the qq plot says that the populations are nearly normal as it is a straight line approx. the samples are kept in different cage and so we can count for the assumption to be independent of each other.

(b) Complete the analysis of variance table below.

```
>  grandmean = (83.5+92.3+88.6+99.4)/4

> grandmean

[1] 90.95

> SSB = 35*((83.5 - grandmean)^2 + (92.3 - grandmean)^2 + (88.6 -grandmean)^2 + (99.4 -
grandmean)^2)

> SSB

[1] 4698.75
```

```
> between.df = 4-1

> between.ms = SSB/between.df

> between.ms

[1] 1566.25

> SSW = 34*(16.9^2 + 14.6^2 + 14.2^2 + 14.1^2)

> within.df = 140-4

> within.ms = SSW/within.df

> within.ms

[1] 224.805

> SSB

[1] 4698.75

> SSW

[1] 30573.48

> F = between.ms/within.ms

> F

[1] 6.967149

> pvalue = 1 - pf(F, df1=3, df2 =136)

> pvalue

[1] 0.0002140835
```

Variation	SumOfSquare	DF	Mean Square	F	P-Value
Between	4698.75	3	1566.25	6.967149	0.0002140835
Within	30573.48	136	224.805		
Total	35272.23	139	1791.055		

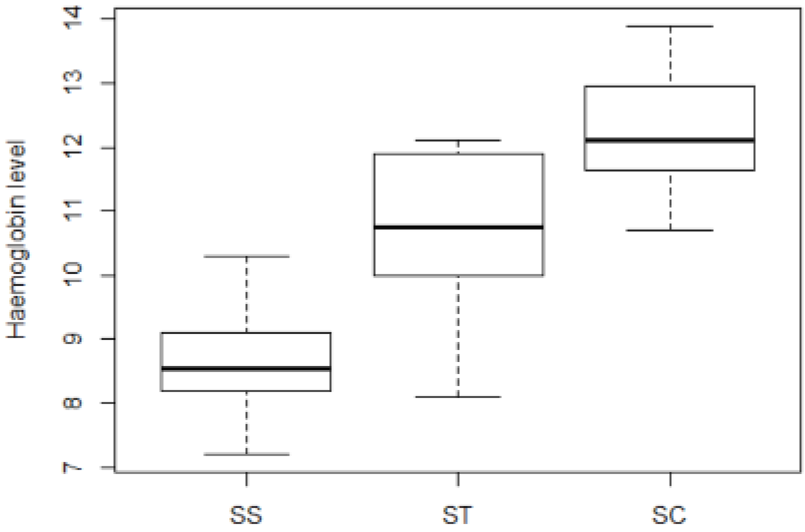
(c) What can you conclude from the analysis of variance? (Give a substantive conclusion, not just “reject” or “don’t reject.

Since p-value is so small we reject the Null hypothesis , the assumption that the weight gained for different cage is not same.

```
> v <- scan(filename)
Read 41 items
> SS <- v[1:16]
> ST <- v[17:26]
> SC <- v[27:41]
```

Let’s write code in R to get the boxplot:

```
> boxplot(SS, ST, SC, range=0,
+ names=c("SS","ST","SC"),
+ ylab="Hemoglobin level")
```



Looking at the boxplots, it does not seem likely that the average hemoglobin levels of the three groups are the same.

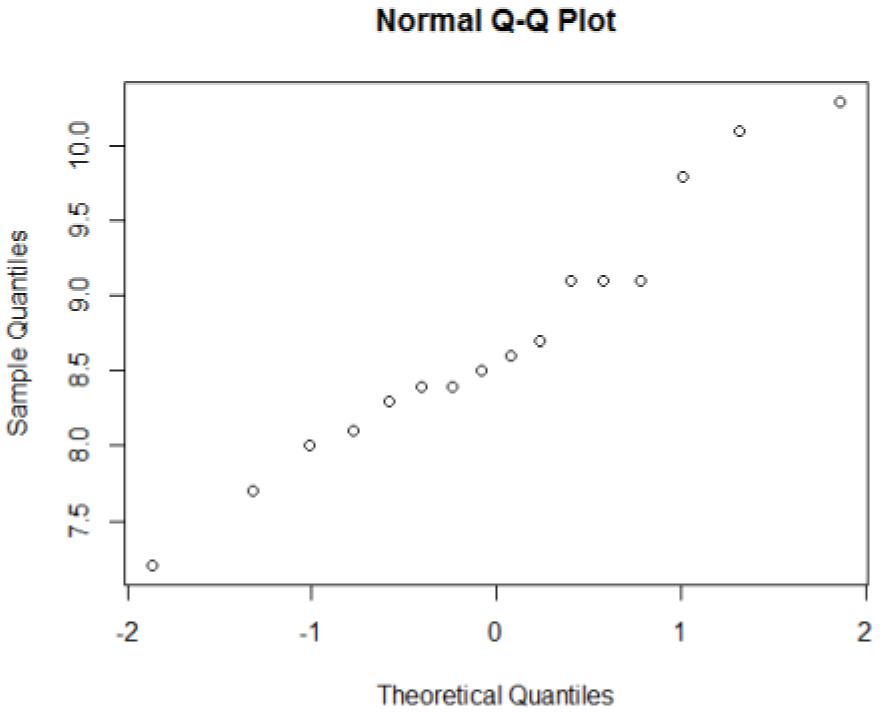
To check for equal variances, let’s calculate the sample SDs:

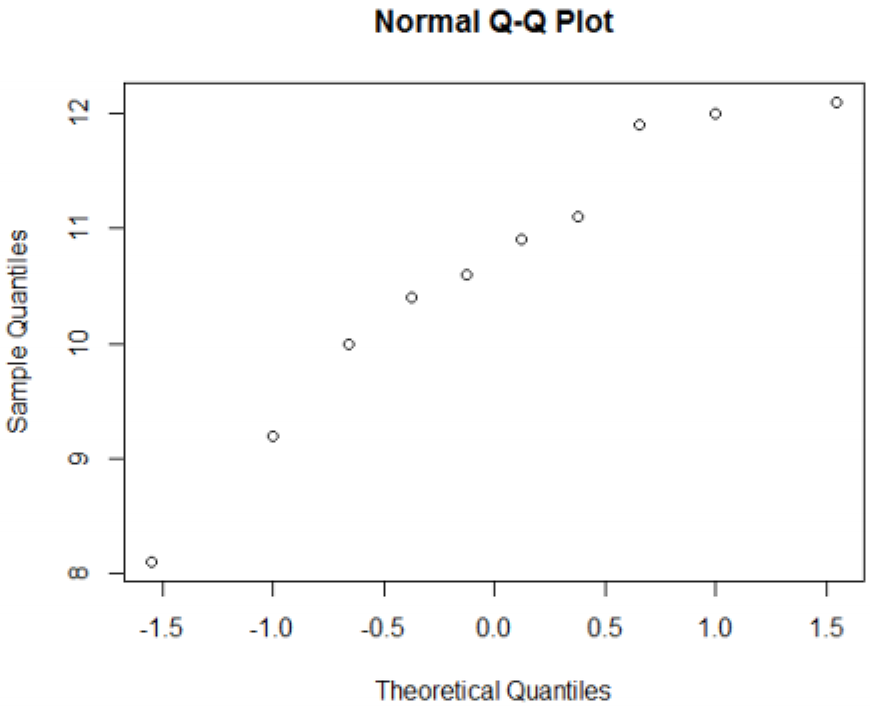
```
> sd(SS)
[1] 0.844492
> sd(ST)
[1] 1.284134
```

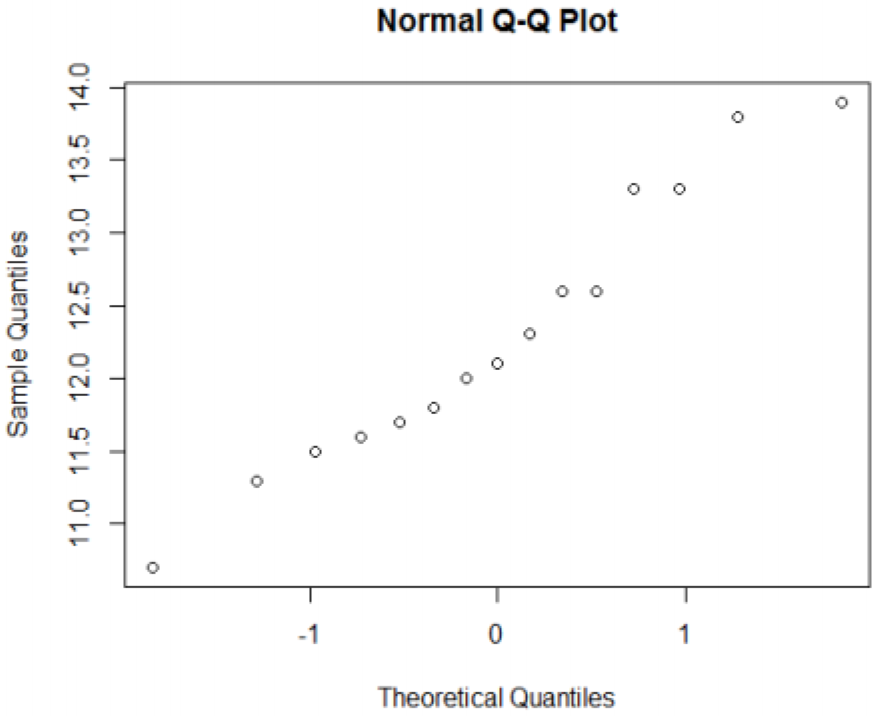
```
> sd(SC)
[1] 0.9418826
```

Values are almost close to each other so we can go ahead.

For second rule let’s check the normality:







Doesn’t exactly looks like a straight line for the purpose of normality.

Part- 2:

```
> SC = c(7.2 ,7.7, 8.0, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7,
+ 9.1, 9.1, 9.1, 9.8, 10.1, 10.3)
> ST = c(8.1 ,9.2, 10.0, 10.4, 10.6, 10.9 ,11.1, 11.9 ,12.0, 12.1)
> SC = c(10.7 ,11.3, 11.5, 11.6, 11.7, 11.8, 12.0, 12.1, 12.3, 12.6,
+       12.6, 13.3, 13.3, 13.8, 13.9)
> SS = c(7.2 ,7.7, 8.0, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7,
+       + 9.1, 9.1, 9.1, 9.8, 10.1, 10.3)
> gm = (mean(SS)*length(SS) + mean(ST)*length(ST) +mean(SC)*length(SC))/41
> gm
[1] 10.49268
> SSB = length(SS)*((mean(SS) - gm)^2) + length(ST)*((mean(ST) - gm))^2 +
length(SC)*((mean(SC) - gm)^2)
> SSB
[1] 99.8893
> between.df = 3-1
> between.ms = SSB/between.df
> between.ms
[1] 49.94465
> SSW = (length(SS)-1)*var(SS) + (length(ST)-1)*var(ST) + (length(SC)- 1)*var(SC)
> SSW
[1] 37.9585
> within.df = length(SS)+length(ST)+length(SC) - 3
> within.df
[1] 38
> within.ms = SSW / within.df
> F = between.df/within.ms
> F
[1] 2.002187
> F = between.ms/within.ms
> F
[1] 49.99926
> p = 1 - pf(F, df1=between.df, df2=within.df)
> p
[1] 2.281786e-11
```

Variation	SumOfSquare	DF	Mean Square	F	P-Value
Between	99.889	2	49.94465	49.99926	2.281786e-11
Within	37.958	38	0.9989		
Total	137.848	40	50.94356		

Such a small p –value indicates to reject the Null Hypothesis it means the sickle cell disease doesn't have the same mean hemoglobin

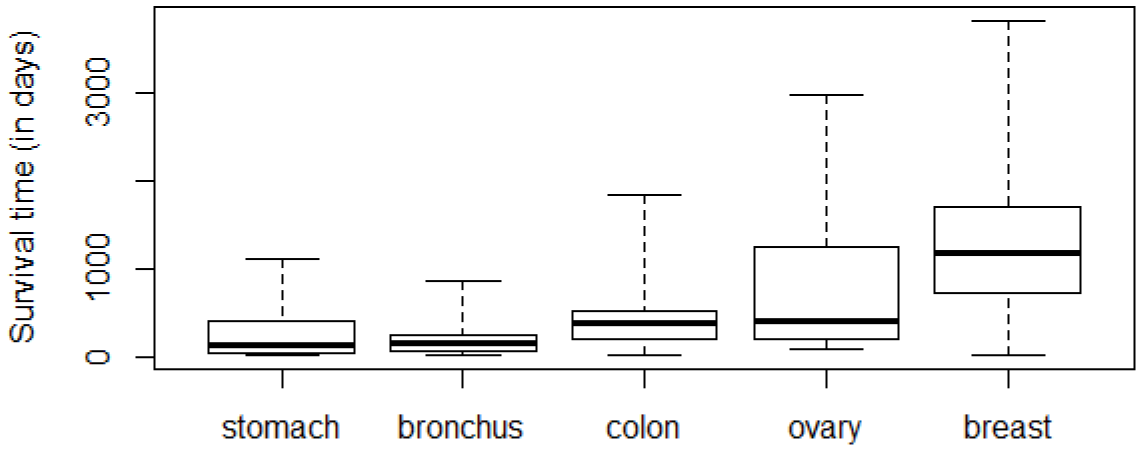
Problem Set C A total of *N* = 64 patients with *k* = 5 types of advanced cancer were treated with ascorbate. The resulting survival times (in days) are displayed in Table 12.4.7

1. Construct two side-by-side boxplots, one of _*x*1, . . . ,_*x*5, one of _*y*1, . . . ,_*x*5. Which data, the observed survival times or the transformed survival times, more nearly satisfy the ANOVA assumptions of normality

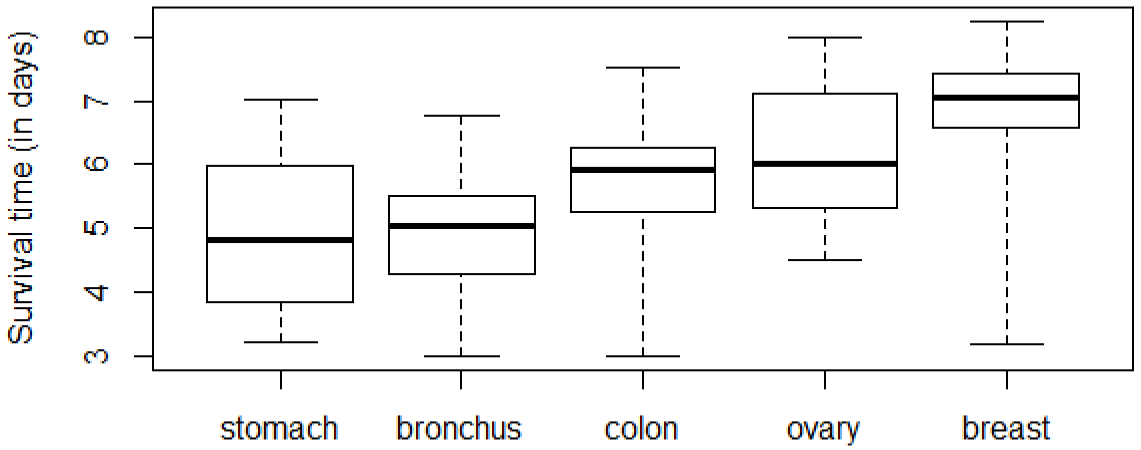
and homoscedasticity? Explain.

```
> Stomach = c(124,42,25,45,412,51,1112,46,103,876,146,340,396,223,138)
> Bronchus = c(81,461,20,450,246,166,63,64,155,859,151,166,37,163,101)
> Stomach = c(124,42,25,45,412,51,1112,46,103,876,146,340,396)
> Brinchus = c(81,461,20,450,246,166,63,64,155,859,151,166,37,223,138,72,245)
> Bronchus = c(81,461,20,450,246,166,63,64,155,859,151,166,37,223,138,72,245)
> Colon = c(248,377,189,1843,180,537,519,455,406,365)
> Colon = c(248,377,189,1843,180,537,519,455,406,365,942,776,372,163,101,20,283)
> Ovary = c(1234,89,201,356,2970,456)
> Breast = c(1235,24,1581,1166,40,727,3808,791,1804,3460,719)
```

```
> log.stomach = log(Stomach)
> log.bronchus = log(Bronchus)
> log.colon = log(Colon)
> log.ovary = log(Ovary)
> log.breast = log(Breast)
> boxplot(Stomach, Bronchus, Colon, Ovary, Breast, range=0,
+ names=c("stomach","bronchus","colon", "ovary", "breast"),
+ ylab="Survival time (in days)")
```



```
> boxplot(log.stomach, log.bronchus, log.colon, log.ovary, log.breast, range=0,
+ names=c("stomach","bronchus","colon", "ovary", "breast"),
+ ylab="Survival time (in days)")
```



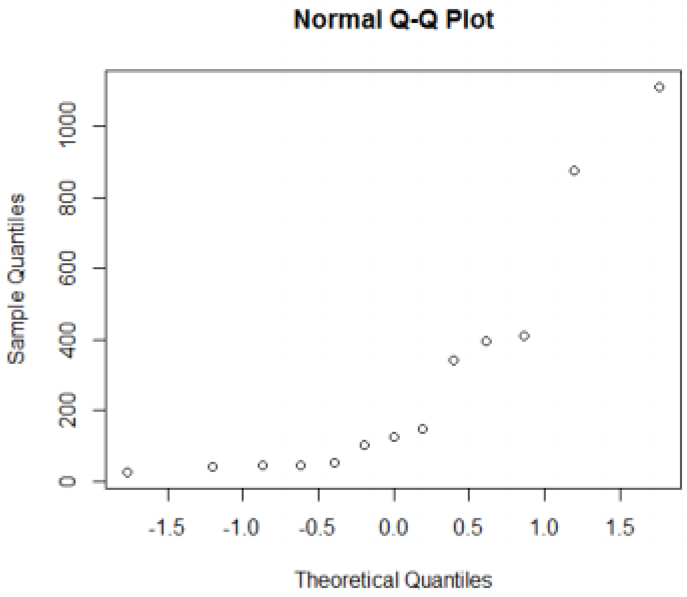
The mean for the log seems to be close to each other as compared to the actual plot

```
> sd(Stomach)
[1] 346.3096
> sd(Bronchus)
[1] 209.8586
> sd(Ovary)
[1] 1098.579
> sd(Colon)
[1] 427.1686
> sd(Breast)
[1] 1238.967
```

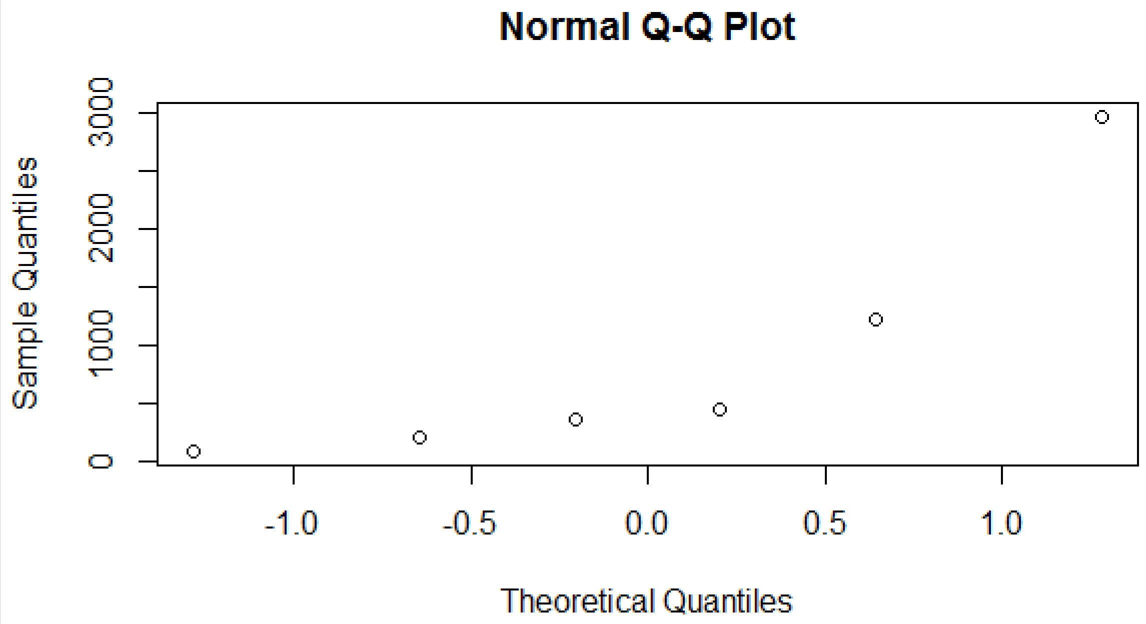
```
> sd(log.colon)
[1] 0.9974766
> sd(log.stomach)
[1] 1.250207
> sd(log.breast)
[1] 1.647755
> sd(log.bronchus)
[1] 0.9534041
> sd(log.ovary)
[1] 1.256931
```

Again the SD of the log are more closer to each other than the actual plot, So, we can say that the log transformed data confirms to the homoscedasticity approximation.

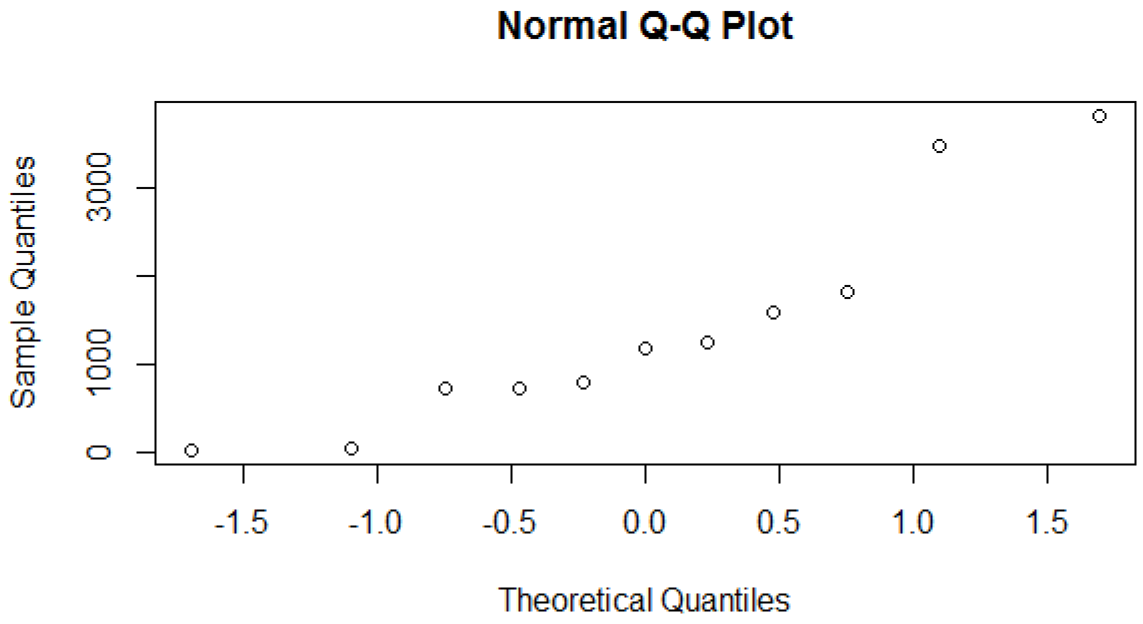
```
> qqnorm(Stomach)
```



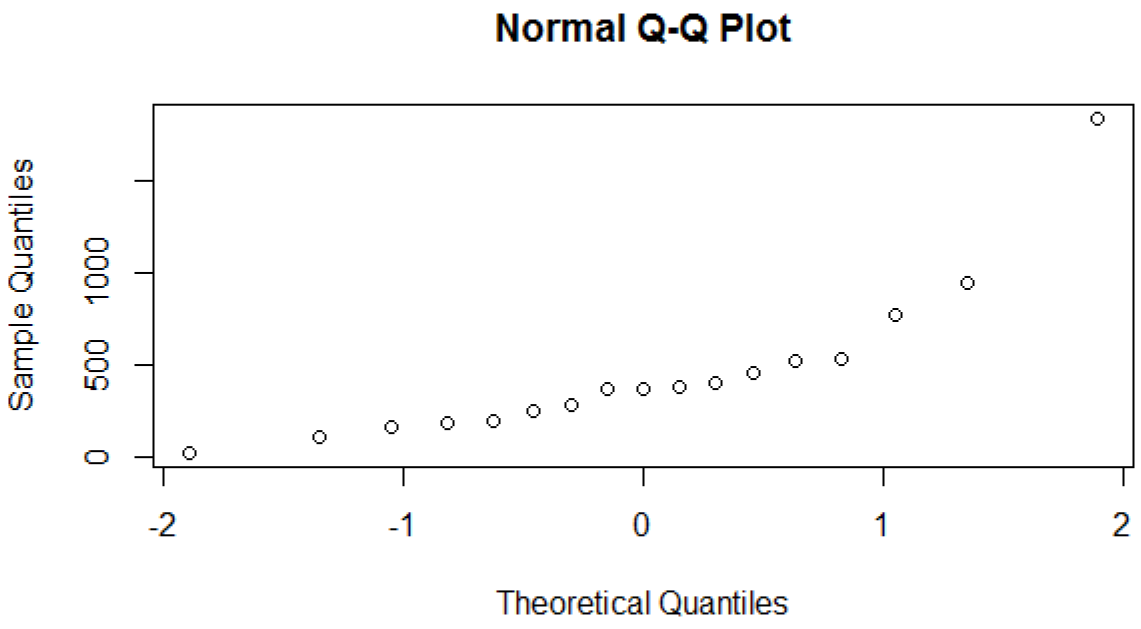
```
> qqnorm(Ovary)
```



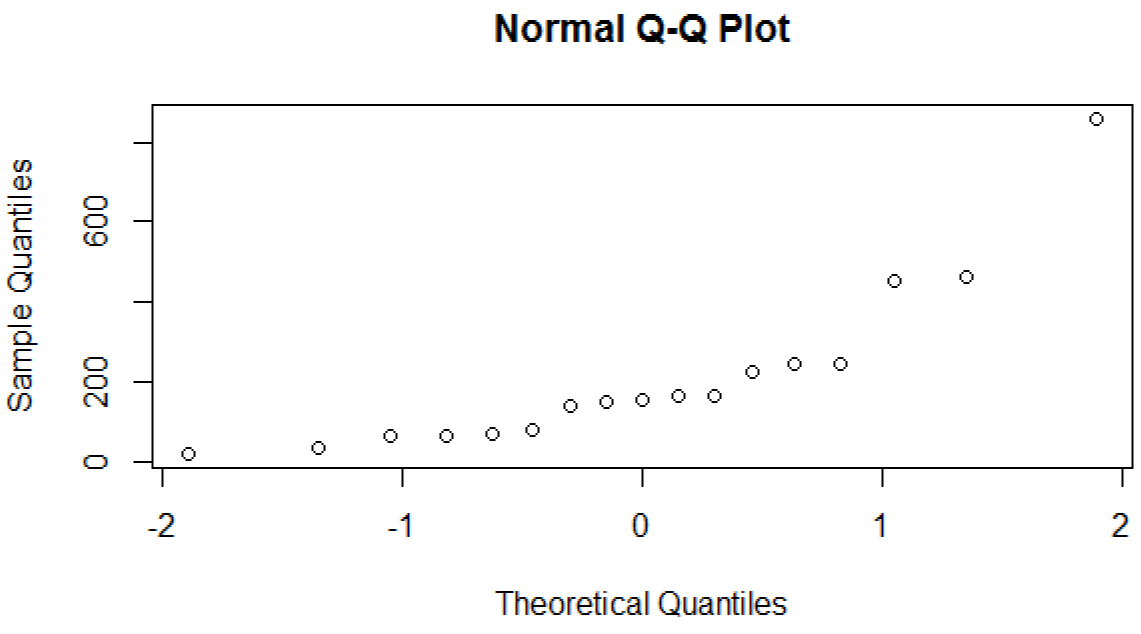
```
> qqnorm(Breast)
```



```
> qqnorm(Colon)
```

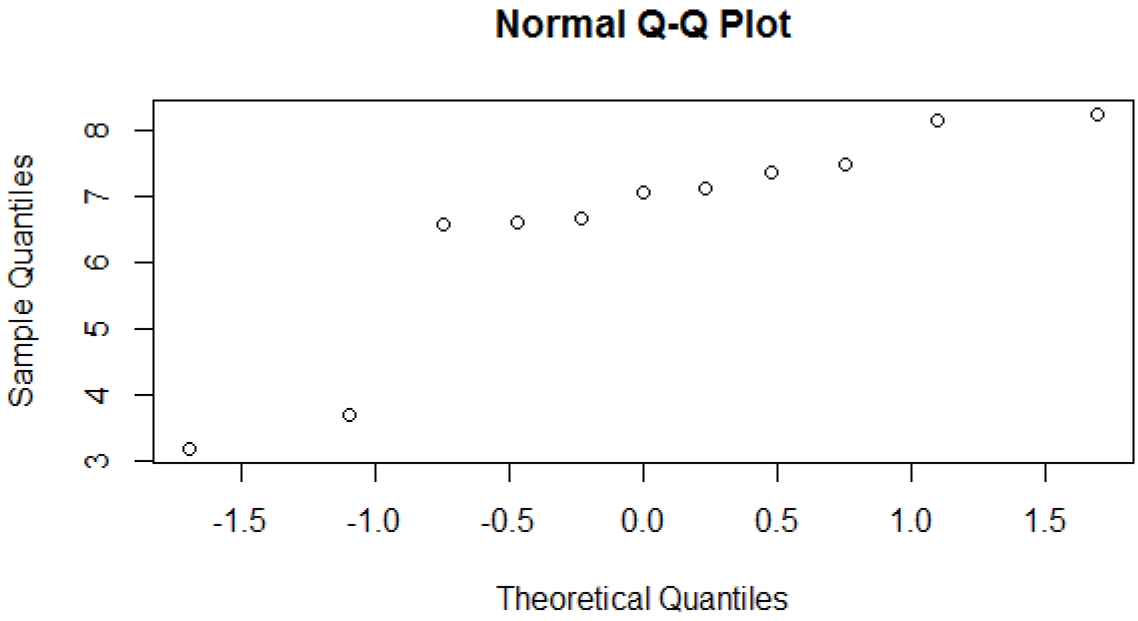


```
> qqnorm(Bronchus)
```

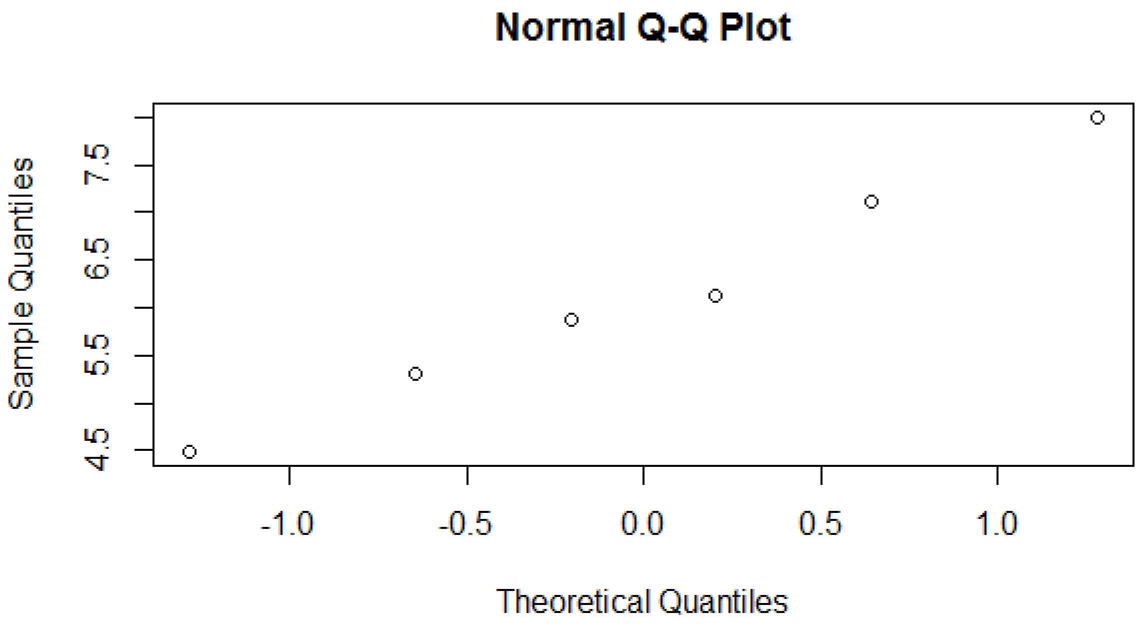


The above QQ-plots definitely do not confirm to the normality assumption. We need to try the log plot.

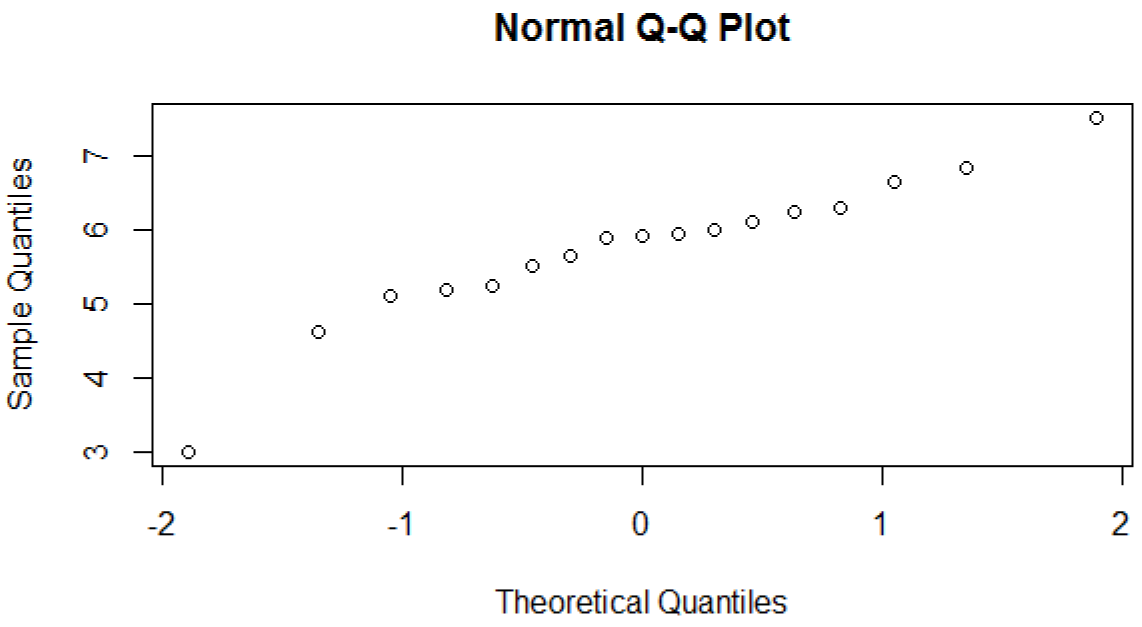
```
> qqnorm(log.breast)
```



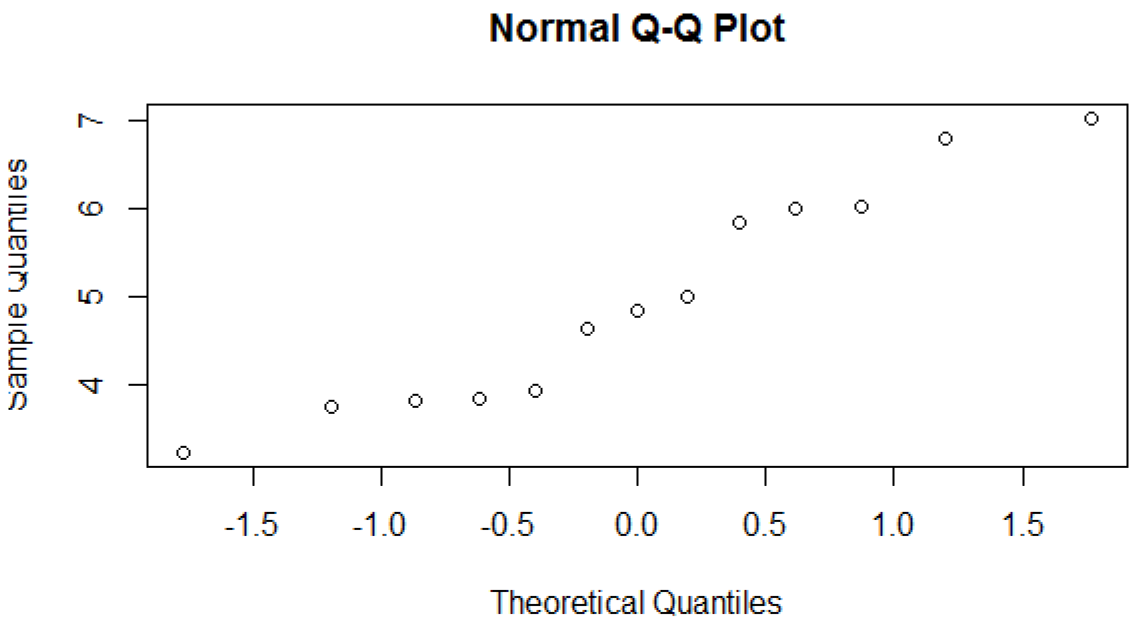
```
> qqnorm(log.ovary)
```



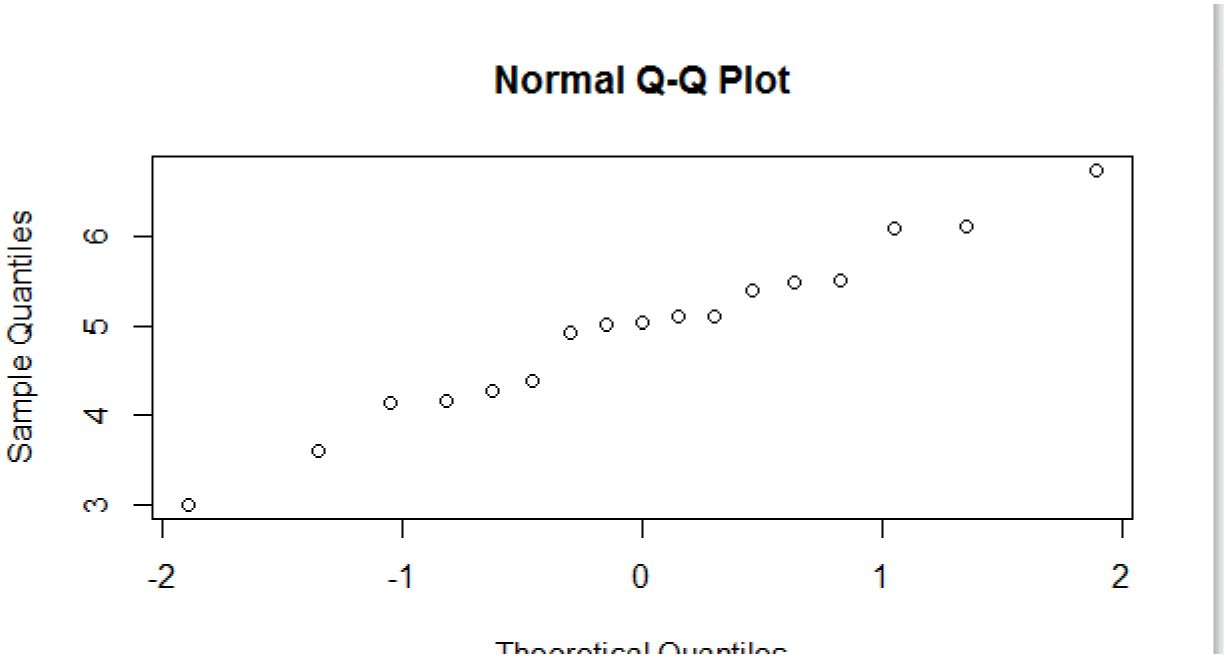
```
> qqnorm(log.colon)
```



```
> qqnorm(log.stomach)
```



```
> qqnorm(log.bronchus)
```



Looking at the QQ-plots of the log transformed data, we can say that they are closer to normality than the original data. The samples are very small and assuming normality should not be so straightforward.

For performing the ANOVA test our best option is to go with the log of the actual plot.

H0 : mu1=mu2=mu3=mu4=mu5

H1: not equal

```
> gm = (mean(log.stomach)*length(log.stomach) + mean(log.bronchus)*length(log.bronchus) +
mean(log.colon)*length(log.colon) + mean(log.ovary)*length(log.ovary) +
mean(log.breast)*length(log.breast))/64

> gm

[1] 5.555785

> SSB = length(log.stomach)*((mean(log.stomach) - gm)^2)
+length(log.bronchus)*((mean(log.bronchus) - gm))^2 +length(log.colon)*((mean(log.colon) -
gm)^2) +length(log.ovary)*((mean(log.ovary) - gm)^2) +length(log.breast)*((mean(log.breast) -
gm)^2)

> SSB

[1] 24.48656

> between.df = 4

> between.ms = SSB/between.df

> between.ms

[1] 6.121639

> SSW = (length(log.stomach)-1)*var(log.stomach) + (length(log.bronchus)-1)*var(log.bronchus)
+ (length(log.colon)-1)*var(log.colon) +(length(log.ovary)-1)*var(log.ovary) + (length(log.breast)-
1)*var(log.breast)

> SSW

[1] 84.26959

> within.df = 59

> within.ms = SSW/within.df

> within.ms

[1] 1.428298

> F = between.ms/ within.ms

> pvalue = 1 - pf(F, df1=4, df2=59)
> pvalue
[1] 0.004122
```

Variation	SumOfSquare	DF	Mean Square	F	P-Value
Between	24.487	4	6.122	4.2285967	0.004122
Within	84.269	59	1.428		
Total	108.756	63	7.5499		

Such low value of p rejects the Null hypothesis, it states that mean of log survival time is not same across all organs affected. But its hard to conclude on such evidences as the sample size is too small .