Assignment 2, EDDA 2017

Martin de la Riva and Kieran O'Driscoll, group 23

24 April 2017

Exercise 1

1.

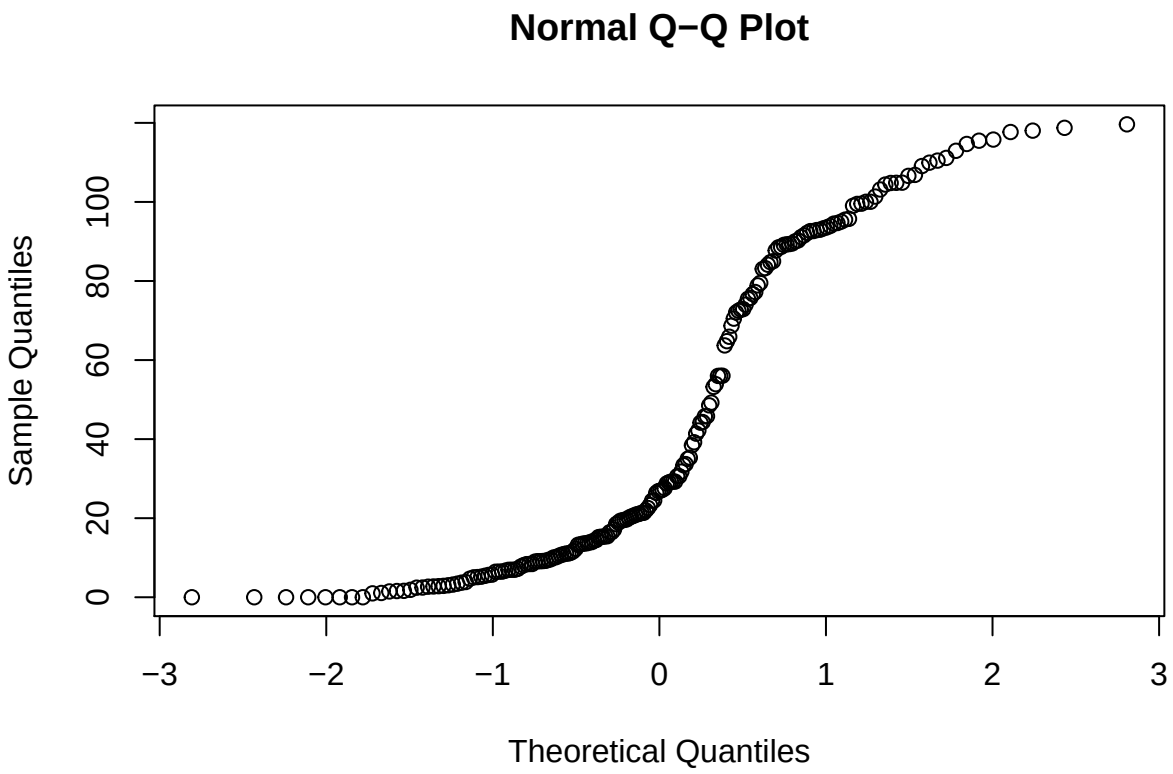
For this exercise we first create an exponential distribution of a significant amount of elements with $\lambda = 0.035$. With this, we get a precise value fo what the median of such distribution is.

```
telephone=read.table("telephone.txt",header=TRUE)

# exponential distribution of 20000 elements
exp_d = rexp(20000, rate=0.035)
exp_d_median = median(exp_d)
```

Then, we perform a Wilcoxon test to see if the median is close to the previously mentioned distribution. In order to perform this test, we first have to assume that the data is a random sample from a symmetric population with a certain median m. By looking at the QQ-plot of the dataset, We can make the symmetric assumption and therefore go on with our Wilcoxon test.

```
# Test if data may come from a Symmetric population
qqnorm(telephone[,"Bills"])
```



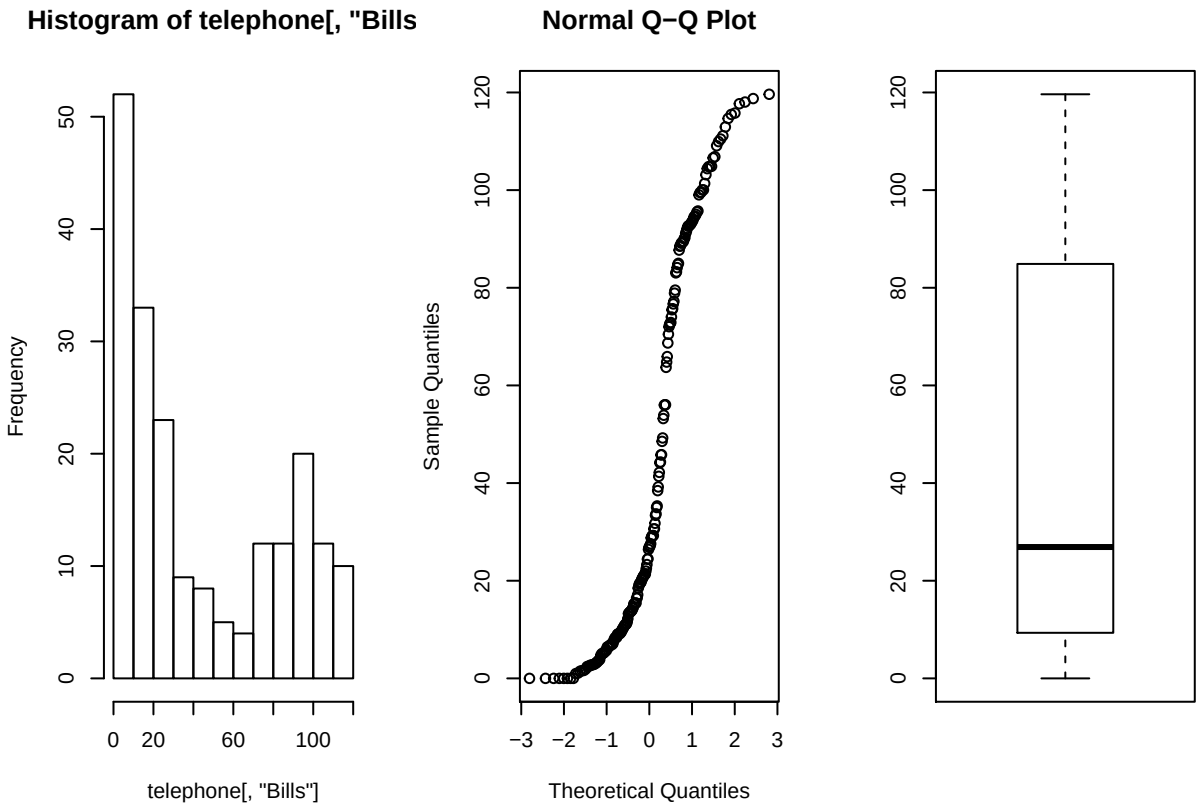

```r
# Rejected, median is not close to exponential median
wilcox.test(telephone[,"Bills"], mu = exp_d_median)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  telephone[, "Bills"]
## V = 14575, p-value = 3.376e-08
## alternative hypothesis: true location is not equal to 19.84154
```

This results in a really low p-value, therefore rejecting the hypothesis that the data stems from an exponential distribution with rate $\lambda = 0.035$.

2.

```r
par(mfrow=c(1,3))
hist(telephone[,"Bills"])
qqnorm(telephone[,"Bills"])
boxplot(telephone[,"Bills"])
```



It can be seen that most people pay around 0 to 20$. If you wanna be competitive you should start pricing around that value. 30-70$ is an unsuccesfull pricing, as it has the lowest clients, while it starts to increase at more than 70$.

In Conclusion:
Either offer a cheap service of less than 30$; or create a high quality service with a price lower than 80$, so you win clients that are used to pay more.

Exercise 2

Firstly, we have different measures in the datasets. We need to transform these values in order to have the 3 datasets with the same measures and criteria.
We will use the 1879-1882 criteria, which is in km/s, and substracting 299000 to it. In order to do this we will have to transform light3 into this criteria. We know light3 was calculated measuring how many microseconds it takes to perform 7442 km, and afterwards substracting 24.8 milliseconds.

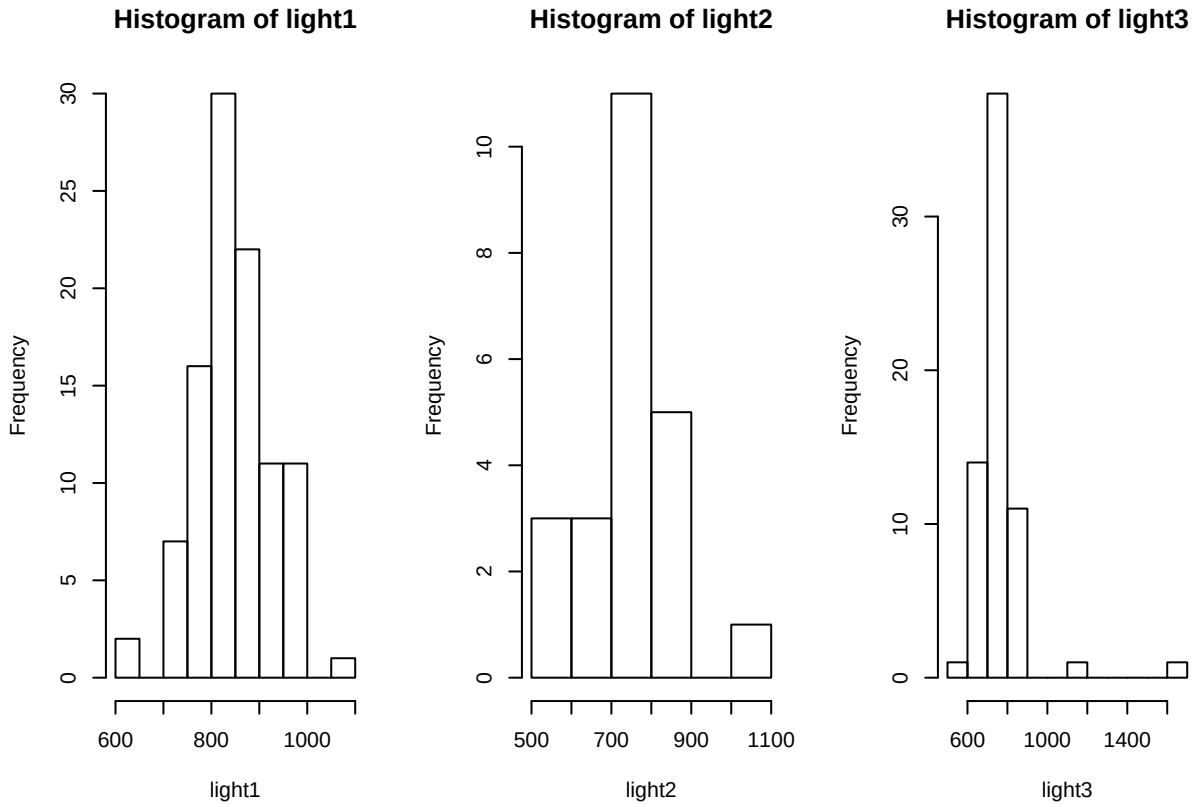
NOTE: We assumed the assignment means milliseconds when talking about Newcomb measurements (24.8 substracted), as after multiplying by 1000 you do get microseconds, and otherways the measurements don't match by any means the speed of light.

```
light1 = scan("light1879.txt")
light2 = scan("light1882.txt")
light3 = scan("light.txt")
```

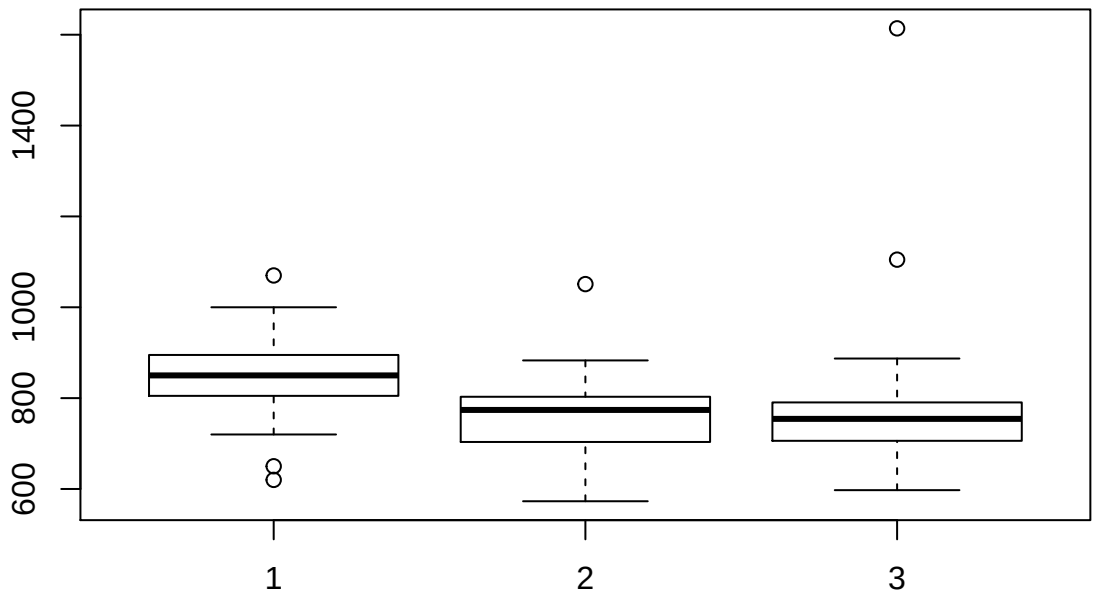
```r
# Transform light3 to 1879-1882 criteria
light3 = (light3/1000)+24.8 # milliseconds
light3 = (light3/1000) # seconds to perform 7442 km
light3 = (7442/light3) # in 1 second it will perform x km (transform to km/s)
light3 = light3 - 299000 # final 1879-1882 criteria
```

1.

```r
par(mfrow=c(1,3))
hist(light1)
hist(light2)
hist(light3)
```




```r
par(mfrow=c(1,1))
boxplot(light1, light2, light3)
```



Michelson second experiment coincides with Newcomb measurements, while Michelson's first experiment (light1) differs from the others.
Although histograms weren't as helpful as we expected them to be, the Boxplots were really useful to see the similarities and differences between the datasets.

2.

```
t.test(light1)$conf.int
```

```
## [1] 836.7226 868.0774
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(light2)$conf.int
```

```
## [1] 709.8976 802.5372
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(light3)$conf.int
```

```
## [1] 731.9112 795.8250
## attr(,"conf.level")
## [1] 0.95
```

NOTE: We maintained 1879-1882 criteria (km/sec - 299000). We also tried with only km/sec, but the confidence intervals where exactly the same although adding 299000 to them.

3.

Again, 2 and 3 coincide (light3 interval is inside light2 confidence interval values), while 1 differs, having a confidence interval with higher values, and being completely outside of the two other ones.

4.

```
most_precise_speed_of_light = 299792.458 - 299000
most_precise_speed_of_light
```

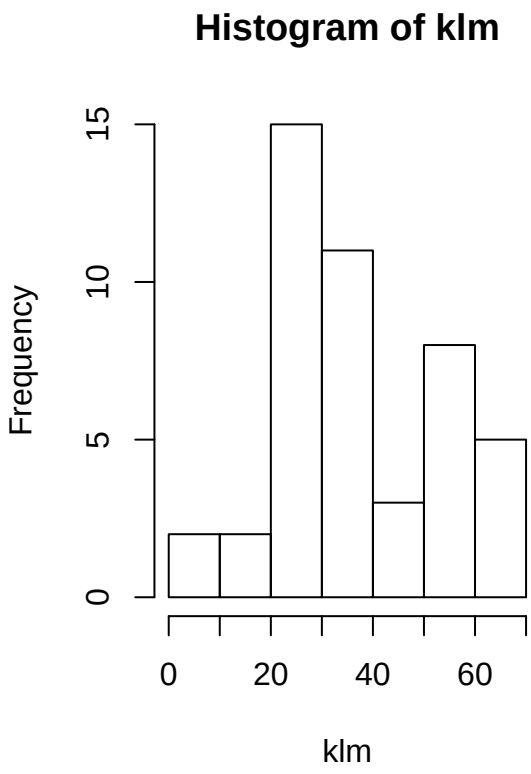
```
## [1] 792.458
```

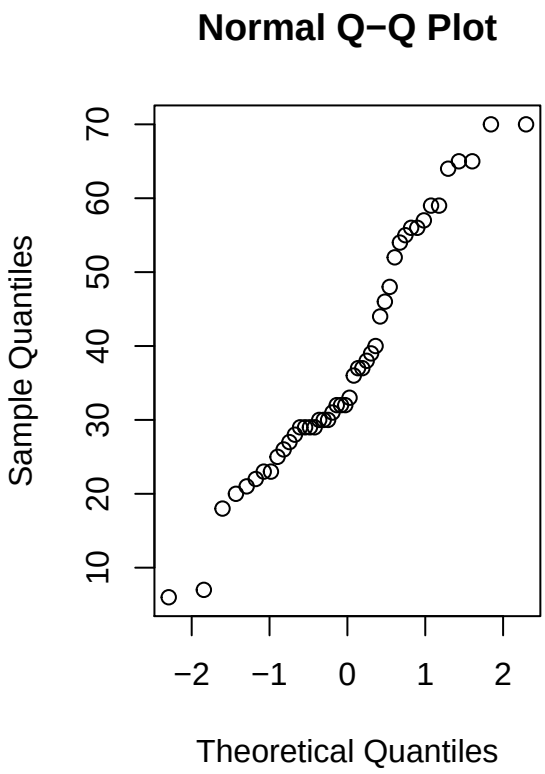
The most precise actual speed of light shows that Michelson's first experiment measures were clearly off, while his second and Newcomb's were much closer to the current accurate solution.

Exercise 3

1.

```
klm = scan("klm.txt")
# First, we check distribution of the data.
klm <- klm[klm<71]
par(mfrow=c(1,2))
hist(klm)
qqnorm(klm)
```





Assuming the maximum delivery duration of the parts is 70 days, we first ignore outliers that are out of that assumption.
In order to test the median we can know two methods, a sign test or a Wilcoxon test. Looking at the histogram and QQ-plot, we can assume the data to stem from a symmetric population with a certain median. Therefore, we can use Wilcoxon test, which is preferred as it is based on more information about the dataset.
In order to test wether the median is 32 or less, we will firstly test if it's equal to 32, and if it is rejected, we will check the lower values to see if it can have a lower median.

```
wilcox.test(klm,mu=32)
```

```
## Warning in wilcox.test.default(klm, mu = 32): cannot compute exact p-value
## with ties
```

```
## Warning in wilcox.test.default(klm, mu = 32): cannot compute exact p-value
## with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  klm
## V = 638.5, p-value = 0.04627
## alternative hypothesis: true location is not equal to 32
```

We get a low p-value, and therefore the hypothesis that the median is equal to 32 is rejected. When

testing lower values, it can be seen how the p-value decreases, and therefore we also reject the hypothesis that the median of the population may be lower than 32.

```
wilcox.test(klm,mu=31)[[3]]
```

```
## Warning in wilcox.test.default(klm, mu = 31): cannot compute exact p-value
## with ties
```

```
## Warning in wilcox.test.default(klm, mu = 31): cannot compute exact p-value
## with zeroes
```

```
## [1] 0.02354406
```

```
wilcox.test(klm,mu=30)[[3]]
```

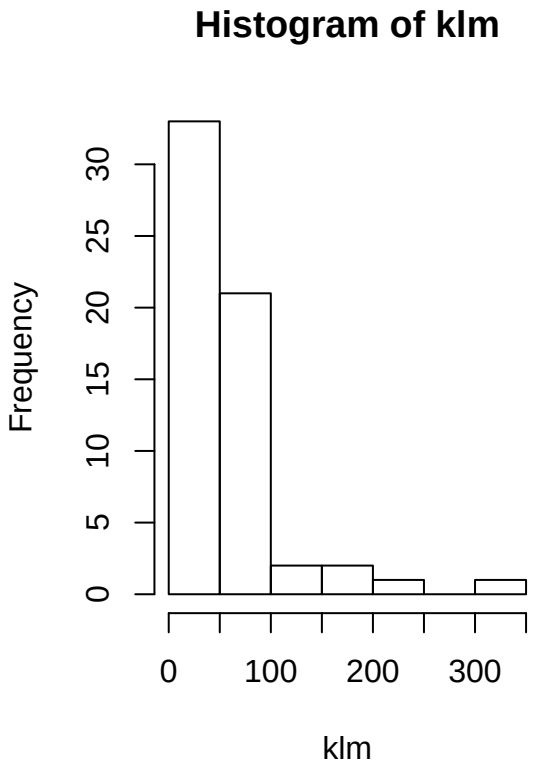
```
## Warning in wilcox.test.default(klm, mu = 30): cannot compute exact p-value
## with ties
```

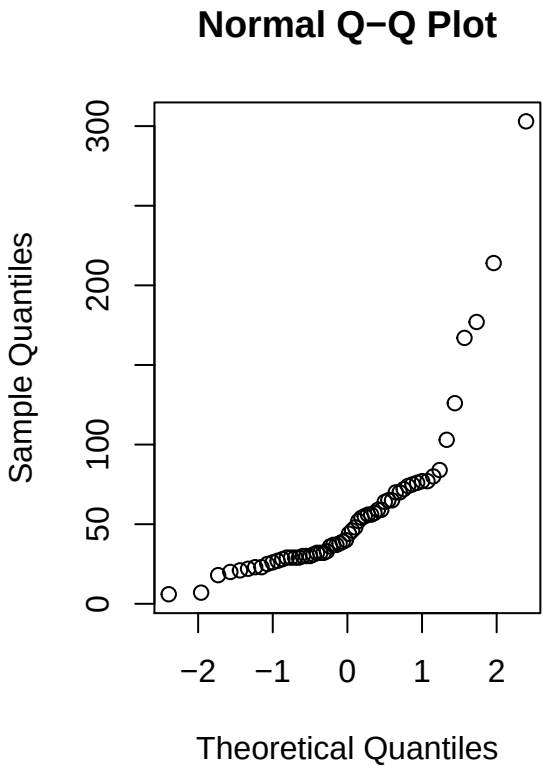
```
## Warning in wilcox.test.default(klm, mu = 30): cannot compute exact p-value
## with zeroes
```

```
## [1] 0.004790985
```

2.

```
klm = scan("klm.txt")
par(mfrow=c(1,2))
hist(klm)
qqnorm(klm)
```





As it can be seen in the plots, in this case we cannot make the symmetric assumption, and therefore we have to perform a sign Binomial test.

```
binom.test(sum(klm>70),sum(!!klm),p=0.1)
```

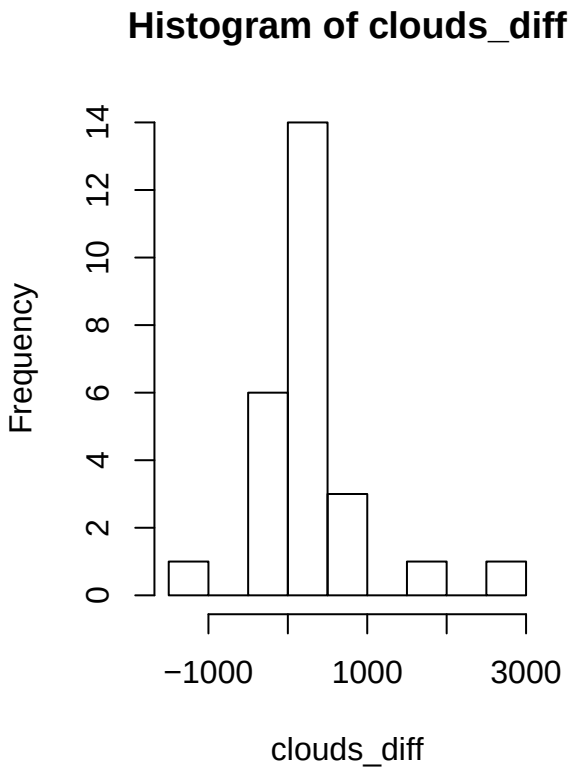
```
##
##  Exact binomial test
##
## data:  sum(klm > 70) and sum(!(!klm))
## number of successes = 14, number of trials = 60, p-value =
## 0.002028
## alternative hypothesis: true probability of success is not equal to 0.1
## 95 percent confidence interval:
##  0.1338373 0.3603828
## sample estimates:
## probability of success
##              0.2333333
```

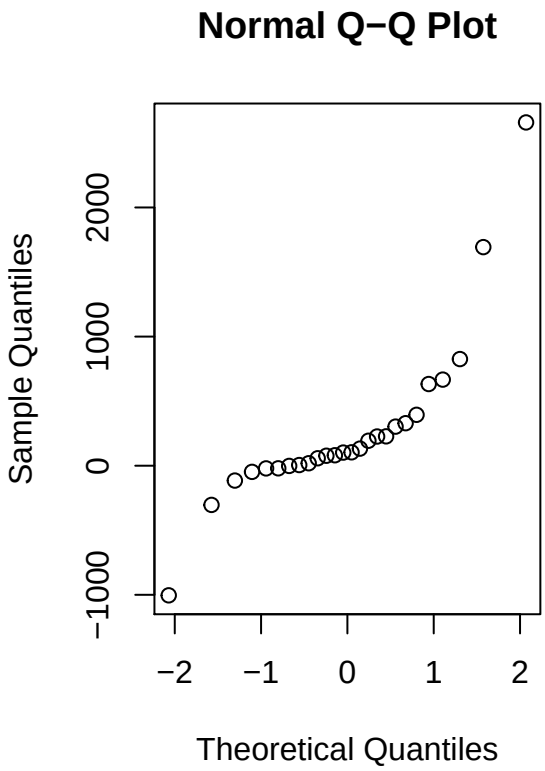
The test returns a confidence interval between 13.38% and 36.04%, and therefore KLM criterion (at most 10% of parts take more than 70 days) is not met.

Exercise 4

```
par(mfrow=c(1,2))
clouds=read.table("clouds.txt",header=TRUE)

clouds_diff = clouds[,1]-clouds[,2]
hist(clouds_diff)
qqnorm(clouds_diff)
```





Two-paired test

Data is clearly paired as it arises from the same individual (cloud) at different points in time.

```
t.test(clouds[,1],clouds[,2],paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  clouds[, 1] and clouds[, 2]
## t = 2.1204, df = 25, p-value = 0.04407
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##    7.961957 546.883428
## sample estimates:
## mean of the differences
##                277.4227
```

According to a paired t-test, silver nitrate does have an effect, giving a confidence interval of difference of (8, 547). As it is a confidence interval with positive values, silver nitrate would result in more rain.
Although this is the result we expected, a Two-Paired test assumes that the sample comes from a normal distribution, which we cannot as it can be seen in the previous plots.

Mann-Whitney test

```
wilcox.test(clouds[,1],clouds[,2])
```

```
## Warning in wilcox.test.default(clouds[, 1], clouds[, 2]): cannot compute
## exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clouds[, 1] and clouds[, 2]
## W = 473, p-value = 0.01383
## alternative hypothesis: true location shift is not equal to 0
```

In order to perform a Mann-Whitney test, we assume that the first sample stems for population F and the second sample from population G, and we test if F=G. As we can make this assumption, the test is appropiate.
As the p-value is low, we can reject the hypothesis that seeded and unseeded clouds come from the same population. This means that there is certainly a difference between clouds with silver nitrate and without it.
The underlying distribution of seeded clouds is shifted to the right from that of unseeded. (i.e. seeded clouds have bigger values than unseeded)

Kolmogorov-Smirnov test

```
ks.test(clouds[,1],clouds[,2])
```

```
## Warning in ks.test(clouds[, 1], clouds[, 2]): cannot compute exact p-value
## with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  clouds[, 1] and clouds[, 2]
## D = 0.42308, p-value = 0.01905
## alternative hypothesis: two-sided
```

Again, the assumption for Kolmogorov is similar to Mann-Whitney's and therefore we can perform the test. The p-value is low again and therefore we can conclude that the samples come from different populations, being the mean of seeded clouds higher than those unseeded.

2.

```
sqrt_clouds_1 = sqrt(clouds[,1])
sqrt_clouds_2 = sqrt(clouds[,2])
t.test(sqrt_clouds_1,sqrt_clouds_2,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  sqrt_clouds_1 and sqrt_clouds_2
## t = 2.682, df = 25, p-value = 0.01278
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.656326 12.617061
## sample estimates:
## mean of the differences
##                7.136693
```

```
wilcox.test(sqrt_clouds_1,sqrt_clouds_2)
```

```
## Warning in wilcox.test.default(sqrt_clouds_1, sqrt_clouds_2): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sqrt_clouds_1 and sqrt_clouds_2
## W = 473, p-value = 0.01383
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(sqrt_clouds_1,sqrt_clouds_2)
```

```
## Warning in ks.test(sqrt_clouds_1, sqrt_clouds_2): cannot compute exact p-
## value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  sqrt_clouds_1 and sqrt_clouds_2
## D = 0.42308, p-value = 0.01905
## alternative hypothesis: two-sided
```

For the Two-Paired test we get an smaller confidence interval, but the hypothesis is still rejected, with an even smaller p-value. For Mann-Whitney and Kolmogorov we get the same exact values, for p-values, W (in Mann-Whitney) and D (in Kolmogorov)

3.

```
sqrt_clouds_1 = sqrt(sqrt(clouds[,1]))
sqrt_clouds_2 = sqrt(sqrt(clouds[,2]))
t.test(sqrt_clouds_1,sqrt_clouds_2,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  sqrt_clouds_1 and sqrt_clouds_2
## t = 2.8413, df = 25, p-value = 0.008811
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2673422 1.6759523
## sample estimates:
## mean of the differences
##               0.9716472
```

```
wilcox.test(sqrt_clouds_1,sqrt_clouds_2)
```

```
## Warning in wilcox.test.default(sqrt_clouds_1, sqrt_clouds_2): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sqrt_clouds_1 and sqrt_clouds_2
## W = 473, p-value = 0.01383
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(sqrt_clouds_1,sqrt_clouds_2)
```

```
## Warning in ks.test(sqrt_clouds_1, sqrt_clouds_2): cannot compute exact p-
## value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  sqrt_clouds_1 and sqrt_clouds_2
## D = 0.42308, p-value = 0.01905
## alternative hypothesis: two-sided
```

Again, we get a smallet confidence interval for Two-Paired test with an even smaller p-value, while Mann-Whitney and Kolmogorov maintain their same exact values.