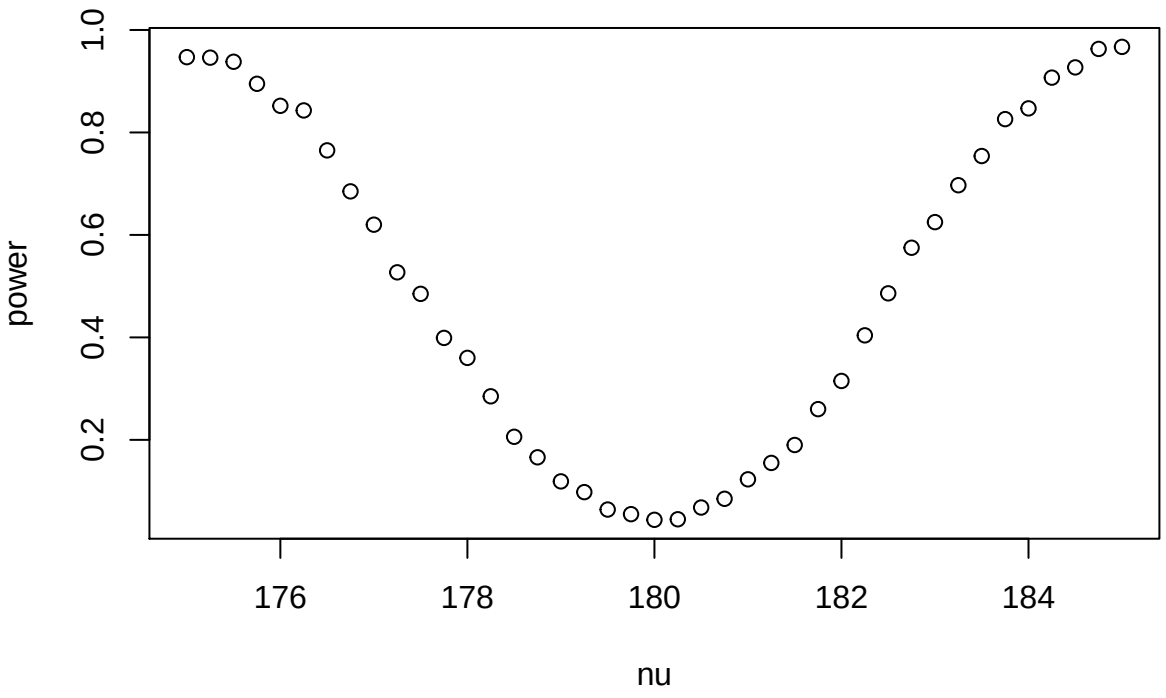
EDDA Assignment 1

Exercise 1

a)

We generate two samples of sizes 30 from a standard normal distribution and programme loops to calculate the power of the t-test for every value of nu:

```
n=m=30;mu=180;sd=5;B=1000;p=numeric(B);
nu=seq(175,185,by=0.25);
power=vector();
for(i in 1:41){
  for(b in 1:B){
    x=rnorm(n,mu,sd);
    y=rnorm(m,nu[i],sd);
    p[b]=t.test(x,y,var.equal = TRUE)[[3]];
  }
  power[i]=mean(p<0.05)
}
plot(nu,power,type="p");
```

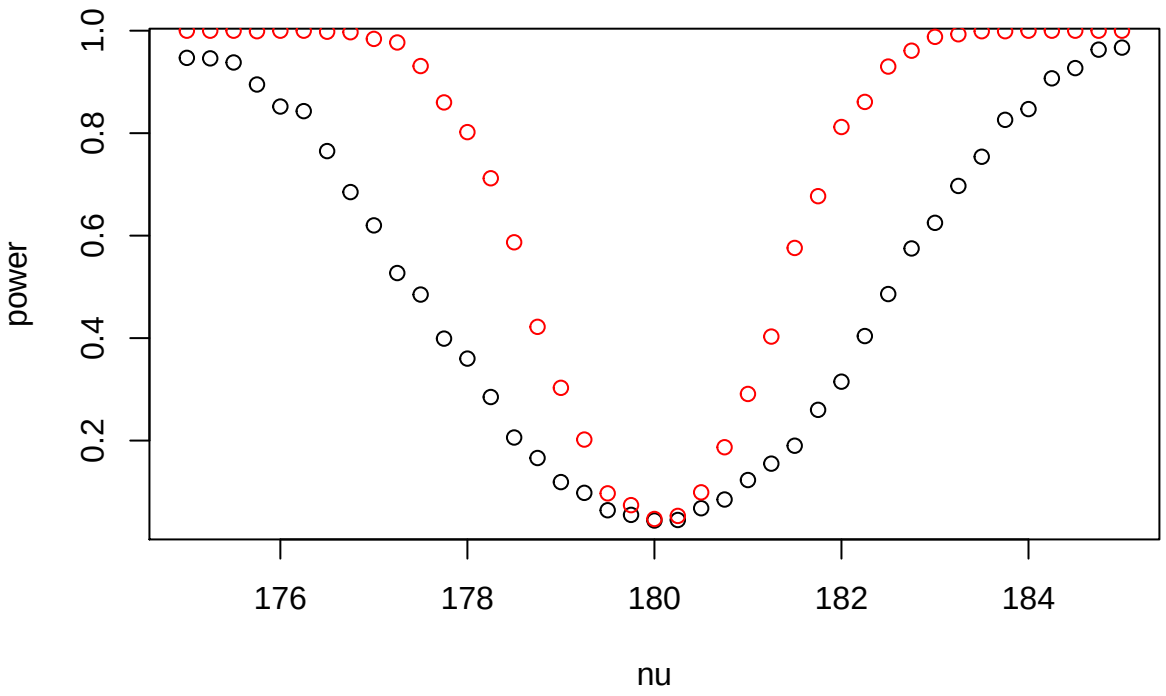


b)

Change n and m to 30:

```
p2=numeric(B);
power2=vector();
n2=m2=100;
for(i in 1:41){
  for(b in 1:B){
    x2=rnorm(n2,mu,sd);
    y2=rnorm(m2,nu[i],sd);
```

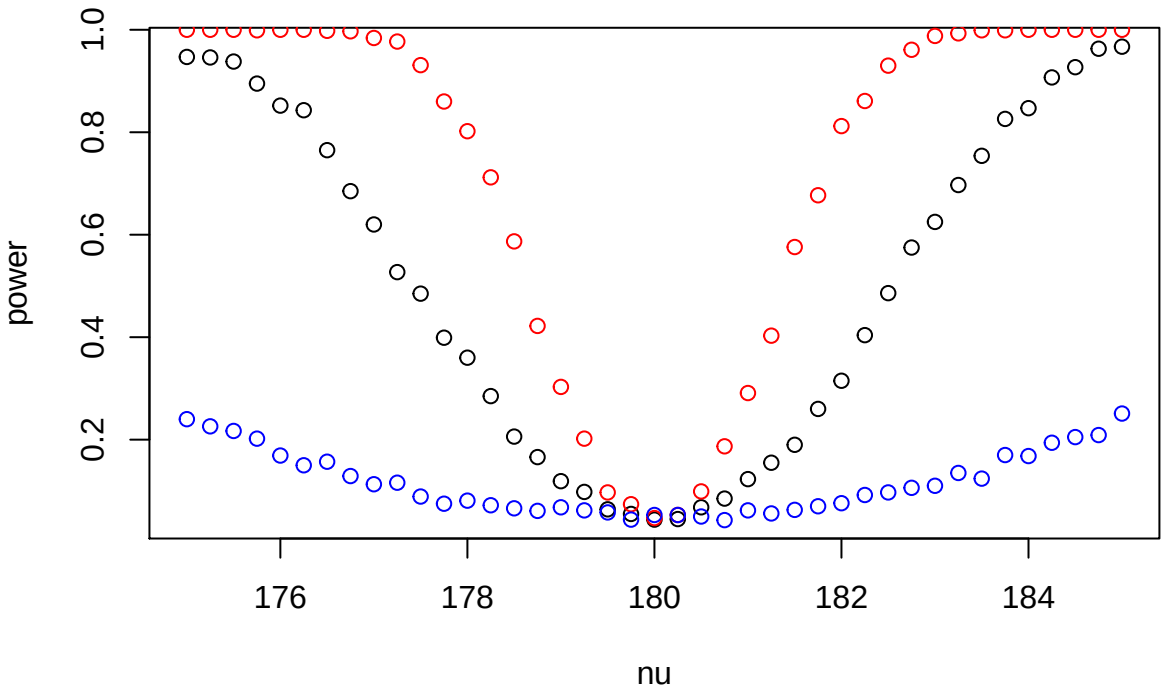
```
    p2[b]=t.test(x2,y2,var.equal = TRUE)[[3]];
  }
  power2[i]=mean(p2<0.05);
}
plot(nu,power,type="p");
lines(nu,power2,type="p",col="red");
```



c)

Change sd to 15:

```
sd3=15;p3=numeric(B);
power3=vector();
for(i in 1:41){
  for(b in 1:B){
    x3=rnorm(n,mu,sd3);
    y3=rnorm(m,nu[i],sd3);
    p3[b]=t.test(x3,y3,var.equal = TRUE)[[3]];
  }
  power3[i]=mean(p3<0.05);
}
plot(nu,power,type="p");
lines(nu,power2,type="p",col="red");
lines(nu,power3,type="p",col="blue");
```



d)

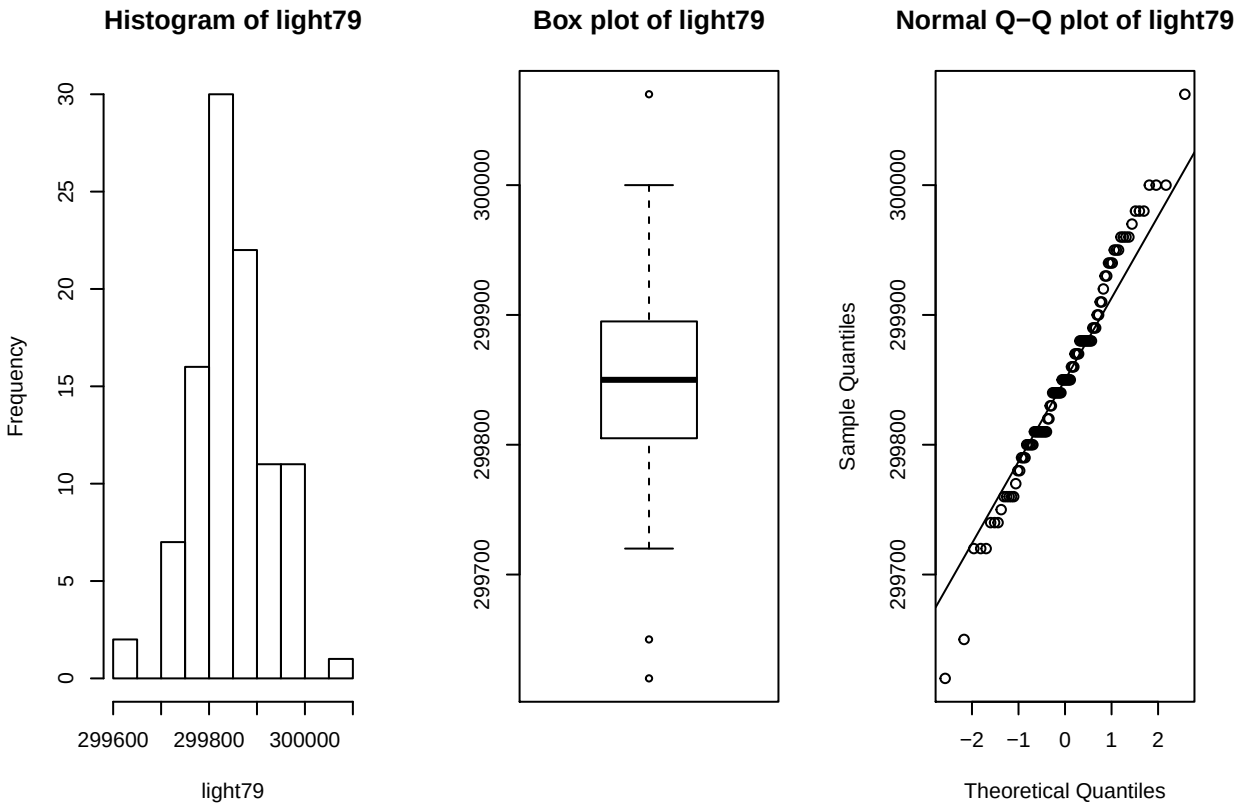
As in a), we can see that if nu gets closer to 180 which equals mu, the power of t-test gets smaller. Compare b) with a), we can find out that if we enlarge the size of samples, the nu has to be more closer to 180 to get a low power. It is more accurate. When it comes to c), we can see that with the growth of sd, the rate of the change of power curve becomes smaller, which means the probability that t-test rejects the null hypothesis becomes smaller.

Exercise 2

a)

$H_0 : \mu = \mu_0$ (population is normally distributed)

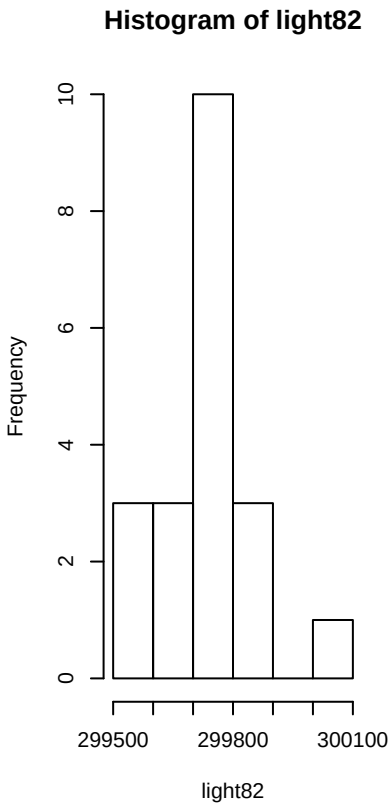
```
light79_raw <- read.delim("./light1879.txt", header = FALSE, sep = "")
light82_raw <- read.delim("./light1882.txt", header = FALSE, fill = TRUE, sep = "")
light_raw <- read.delim("./light.txt", header = FALSE, fill = TRUE, sep = "")

light79 <- unlist(na.omit(light79_raw) + 299000, use.name = FALSE)
light82 <- unlist(na.omit(light82_raw) + 299000, use.name = FALSE)
light <- unlist((7.442 / ((na.omit(light_raw)/1000) + 24.8) * 1000000), use.names = FALSE)

par(mfrow = c(1,3))
hist(light79)
boxplot(light79, main = "Box plot of light79")
qqnorm(light79, main = "Normal Q-Q plot of light79")
qqline(light79)
```

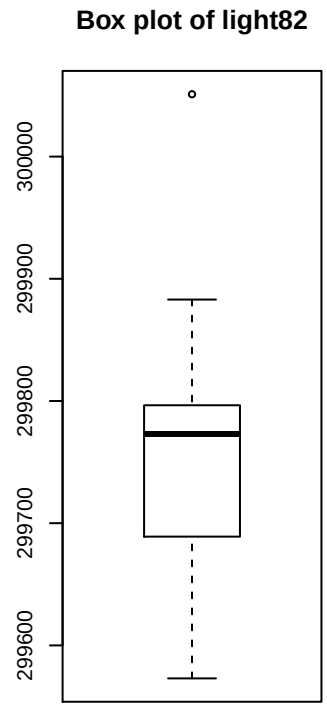


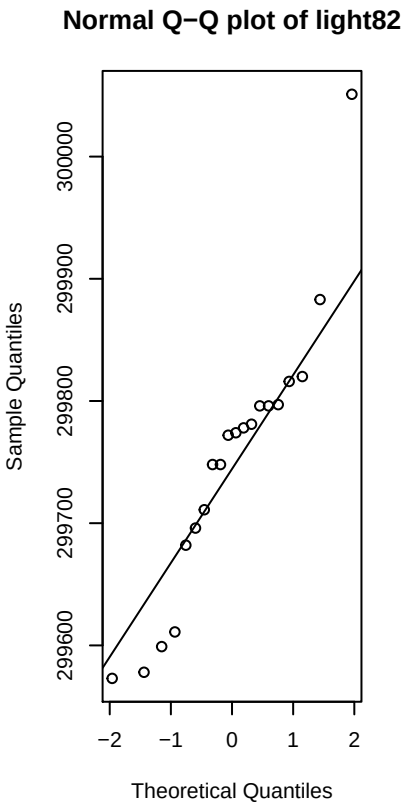

```
shapiro.test(light79)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  light79
## W = 0.98807, p-value = 0.5137
```

```
par(mfrow = c(1,3))
hist(light82)
boxplot(light82, main = "Box plot of light82")
qqnorm(light82, main = "Normal Q-Q plot of light82")
qqline(light82)
```



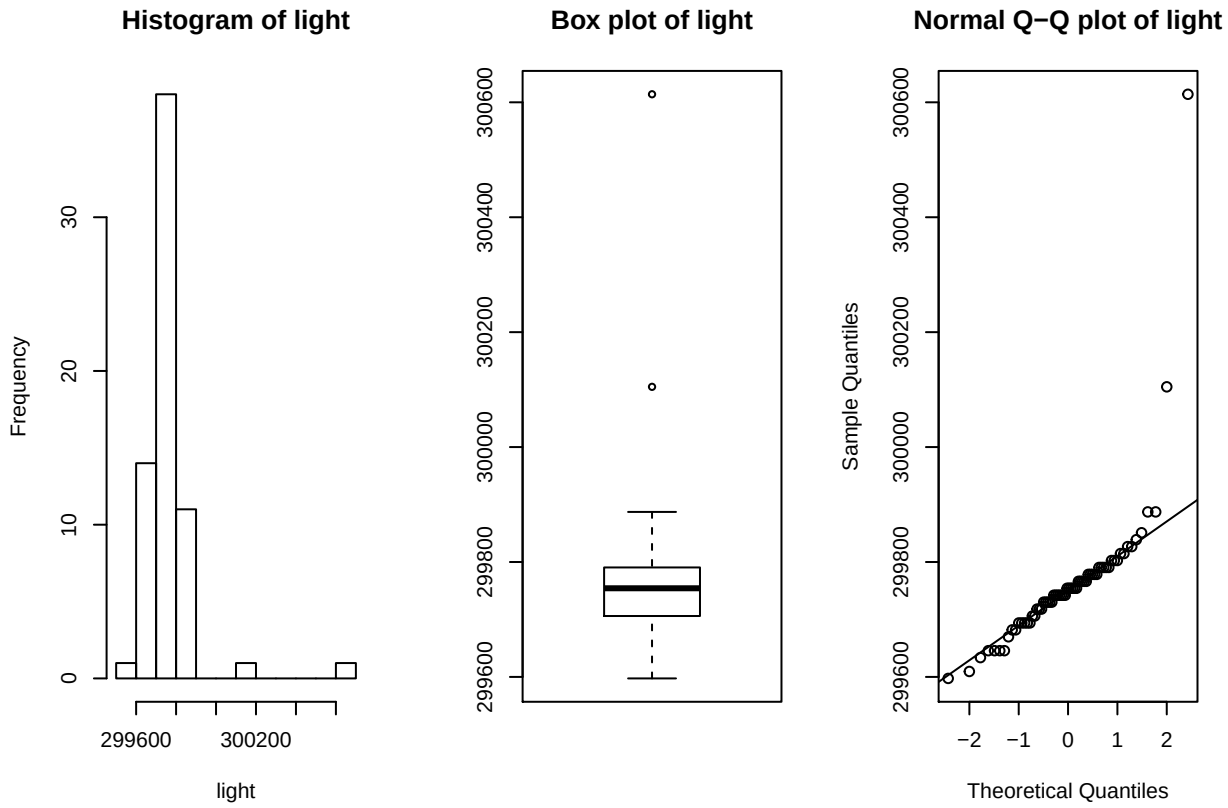





```
shapiro.test(light82)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  light82
## W = 0.91932, p-value = 0.09613
```

```
par(mfrow = c(1,3))
hist(light)
boxplot(light, main = "Box plot of light")
qqnorm(light, main = "Normal Q-Q plot of light")
qqline(light)
```




```
shapiro.test(light)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  light
## W = 0.59001, p-value = 2.724e-12
```

b)

```
T1 <- mean(light79)
T2 <- mean(light82)
T3 <- mean(light)

conf_interval <- function(light_data, T_val) {
  B <- 1000
  Tstar <- numeric(B)
  for(i in 1:B) {
    Xstar <- sample(light_data, replace = TRUE)
    Tstar[i] <- mean(Xstar)
  }
  Tstar25 <- quantile(Tstar, 0.025)
  Tstar975 <- quantile(Tstar, 0.975)
  return(c(2*T_val - Tstar975, 2*T_val - Tstar25))
}

light79_int <- conf_interval(light79, T1)
light82_int <- conf_interval(light82, T2)
```

```r
light_int <- conf_interval(light, T3)

cat("Light 79: Interval = ", light79_int, "Mean = ", T1, "\n")
```

```
## Light 79: Interval =  299836.7 299866.8 Mean =  299852.4
```

```r
cat("Light 82: Interval = ", light82_int, "Mean = ", T2, "\n")
```

```
## Light 82: Interval =  299699 299797.5 Mean =  299750.5
```

```r
cat("Light: Interval = ", light_int, "Mean = ", T3, "\n")
```

```
## Light: Interval =  299728.5 299789.2 Mean =  299763.9
```

c)

```r
current_c <- 299792 # https://www.space.com/15830-light-speed.html
# assuming normality for 79 and 82 we can use t-test, for light we need wilcoxon test
t.test(light79, mu = current_c)
```

```
##
##  One Sample t-test
##
## data:  light79
## t = 7.6445, df = 99, p-value = 1.374e-11
## alternative hypothesis: true mean is not equal to 299792
## 95 percent confidence interval:
##  299836.7 299868.1
## sample estimates:
## mean of x
##  299852.4
```

```r
t.test(light82, mu = current_c)
```

```
##
##  One Sample t-test
##
## data:  light82
## t = -1.6552, df = 19, p-value = 0.1143
## alternative hypothesis: true mean is not equal to 299792
## 95 percent confidence interval:
##  299698 299803
## sample estimates:
## mean of x
##  299750.5
```

```r
wilcox.test(light, mu = current_c)
```

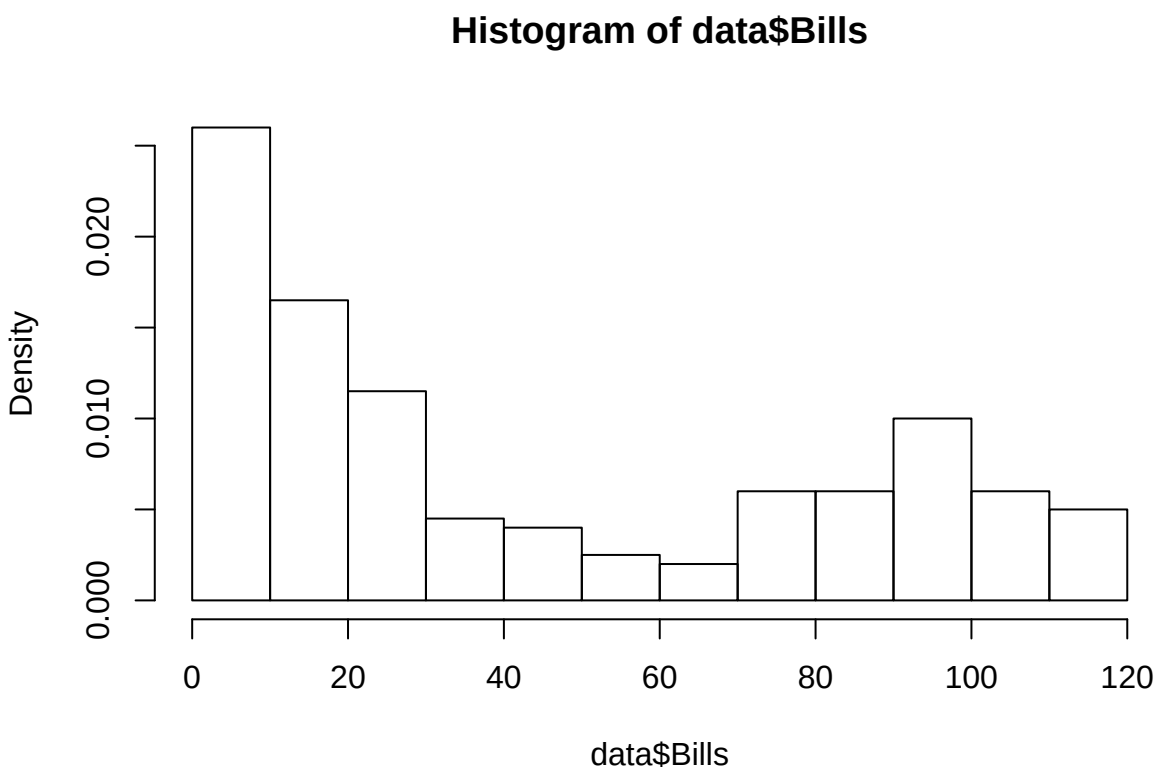
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  light
## V = 387, p-value = 4.451e-06
## alternative hypothesis: true location is not equal to 299792
```

Exercise 3

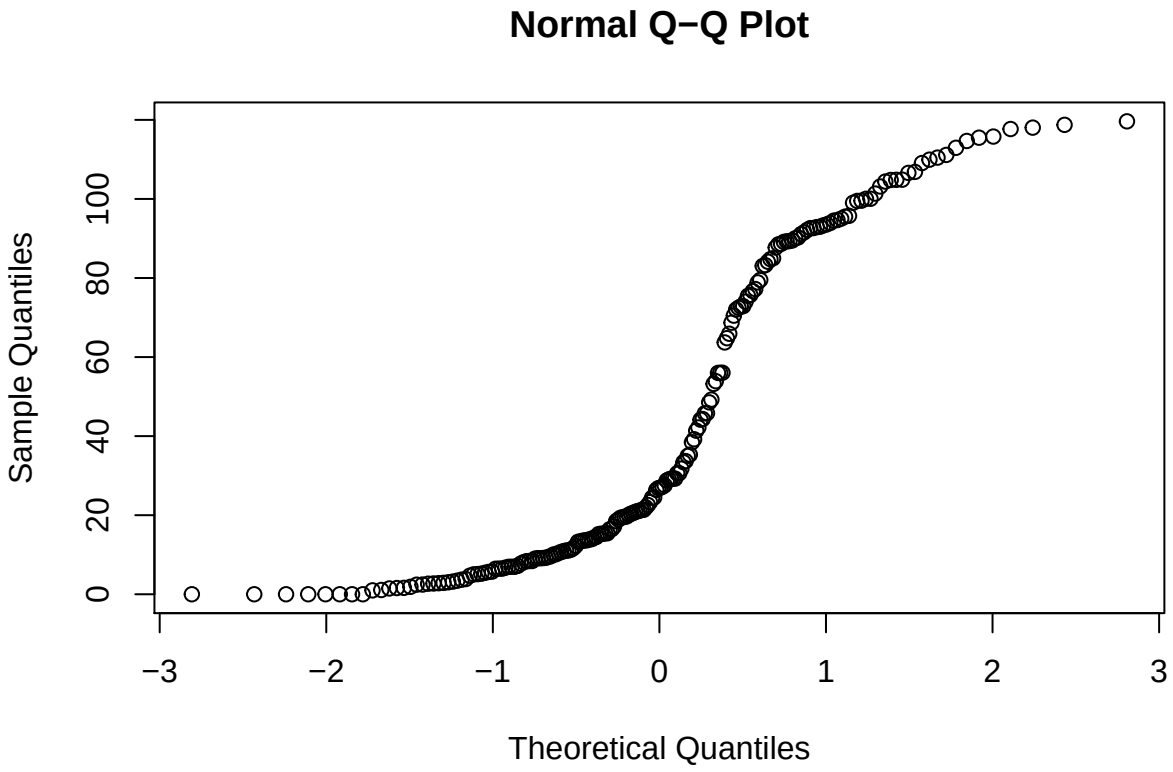
a) Make an appropriate plot of this data set. What marketing advice(s) would you give to the marketing manager? Are there any inconsistencies in the data? If so, try to fix these.

We can make a histiogram and a QQ-plot to show the data set. From the histogram and QQ-plot of the first month bills, we can see that most subscribers' bills is between 0 to 20; bills more than 70 are also of a large quantity. So my advice for the manager is to set two kinds of month bills: one is between 0 to 20, the other is more than 70. We don't think there is any inconsistency in the data. For 0, people may just don't use mobile phones.

```
data=read.table(file="telephone.txt",header=TRUE)
hist(data$Bills,prob=T);
```




```
qqnorm(data$Bills);
```



b) By using a bootstrap test with the test statistic T = median(X1,. . . ,X200), test whether the data telephone.txt stems from the exponential distribution Exp(lambda) with some lambda from [0.01, 0.1].

In order to test lamda from[0.01,0.1],let lamda=seq(0.01,0.1,by=0.01);for every lamda, use bootstrap test.

```
lamda=seq(0.01,0.1,by=0.01);
length(lamda);
```

```
## [1] 10
```

```
bills=data$Bills;
t=median(bills);
B=1000;
tstar=numeric(B);
p=numeric(10);
n=length(bills);
for(i in 1:10){
  for (j in 1:B){
    xstar=rexp(n,lamda[i]);
    tstar[j]=median(xstar);
  }
  pl=sum(tstar<t)/B;
  pr=sum(tstar>t)/B;
  p[i]=2*min(pl,pr);
}
p;
```

```
##  [1] 0.000 0.012 0.122 0.000 0.000 0.000 0.000 0.000 0.000 0.000
```

We can easily find out that p[3] is bigger than 0.05, which means when data$Bills stems from the exponential

distribution Exp(0.03);

c) Construct a 95% bootstrap confidence interval for the population median of the sample.

Confidence 1-2*alpha=0.95, according to the formula its confidence interval should be [2*T-Tstar(1-alpha),2*T=Tstar(alpha)];

```
bills=data$Bills;
B=1000;
Tstar=numeric(B);
T1=median(bills);
for(i in 1:B){
  Xstar=sample(bills,replace = TRUE);
  Tstar[i]=median(Xstar);
}
Tstar25=quantile(Tstar,0.025);
Tstar975=quantile(Tstar,0.975);
sum(Tstar<Tstar25);
```

```
## [1] 24
```

```
c(2*T1-Tstar975,2*T1-Tstar25);
```

```
##  97.5%   2.5%
## 19.450 33.555
```

d) Assuming X1 , . . . Xn ~ Exp(lambda) and using the central limit theorem for the sample mean, estimate lambda and construct again a 95% confidence interval for the population median. Comment on your findings.

Central limit theorem establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. We can set up 1,000 sets of samples, each group sampling 50 data. The mean of 1000 samples is close to the population mean. According to the exponential distribution, mean=1/lamda, so lamda is about 0.022.

```
B=1000;
s=numeric(B);
for(i in 1:B){
  sample=sample(bills,50);
  sample_mean=mean(sample);
  s[i]=sample_mean;
}
s_mean=mean(s);
s_var=var(s);
lamda=1/s_mean;
lamda
```

```
## [1] 0.02295191
```

According to the above distribution,rexp(n,lamda):

```
B=1000;
Tstar=numeric(B);
T1=median(bills);
for(i in 1:B){
  Xstar=sample(rexp(n,lamda));
```

```
  Tstar[i]=median(Xstar);
}
Tstar25=quantile(Tstar,0.025);
Tstar975=quantile(Tstar,0.975);
sum(Tstar<Tstar25);
```

```
## [1] 25
```

```
c(2*T1-Tstar975,2*T1-Tstar25);
```

```
##    97.5%     2.5%
## 17.40727 29.02297
```

e) Using an appropriate test, test the null hypothesis that the median bill is bigger or equal to 40 euro against the alternative that the median bill is smaller than 40 euro. Next, design and perform a test

to check whether the fraction of the bills less than 10 euro is at most 25%. Null hypothesis: median bill >= 40 euro. Choose sign test. The result shows p-value = 0.009698, which is smaller than 0.05, so reject H0, the median bill is smaller than 40 euro.

```
binom.test(sum(bills>40),200,al="l");
```

```
##
##  Exact binomial test
##
## data:  sum(bills > 40) and 200
## number of successes = 83, number of trials = 200, p-value = 0.009698
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.4754731
## sample estimates:
## probability of success
##                  0.415
```

Null hypothesis: the fraction of bills less than 10 euro is at most 25%; p-value=0.3983,the fraction of the bills less than 10 euro is at most 25%.

```
low=sum(bills<10);
binom.test(low,200,p=0.25,alternative="greater",conf.level = 0.9)
```

```
##
##  Exact binomial test
##
## data:  low and 200
## number of successes = 52, number of trials = 200, p-value = 0.3983
## alternative hypothesis: true probability of success is greater than 0.25
## 90 percent confidence interval:
##  0.2197061 1.0000000
## sample estimates:
## probability of success
##                   0.26
```

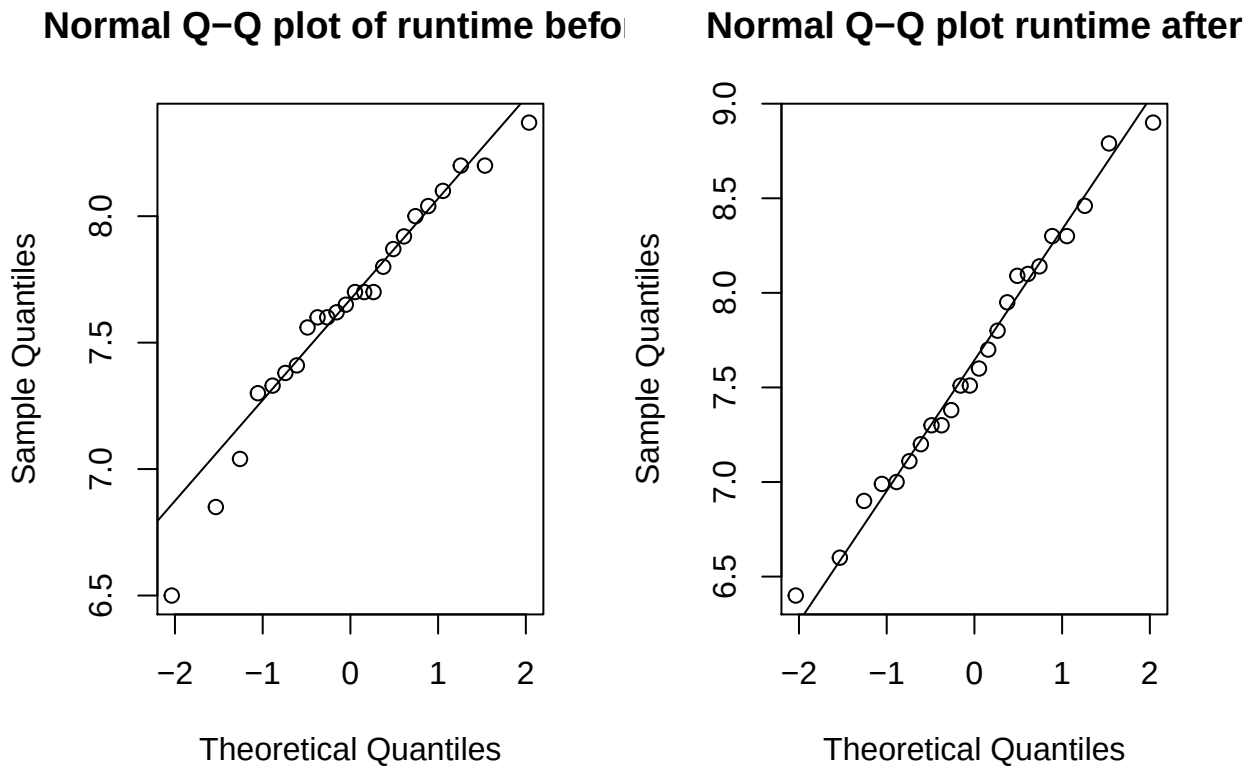
Exercise 4

a) Disregarding the type of drink, test whether the run times before drink and after are correlated.

```
run_raw <- read.table("run.txt", header = TRUE)
before <- run_raw$before
after = run_raw$after
cor.test(before, after)
```

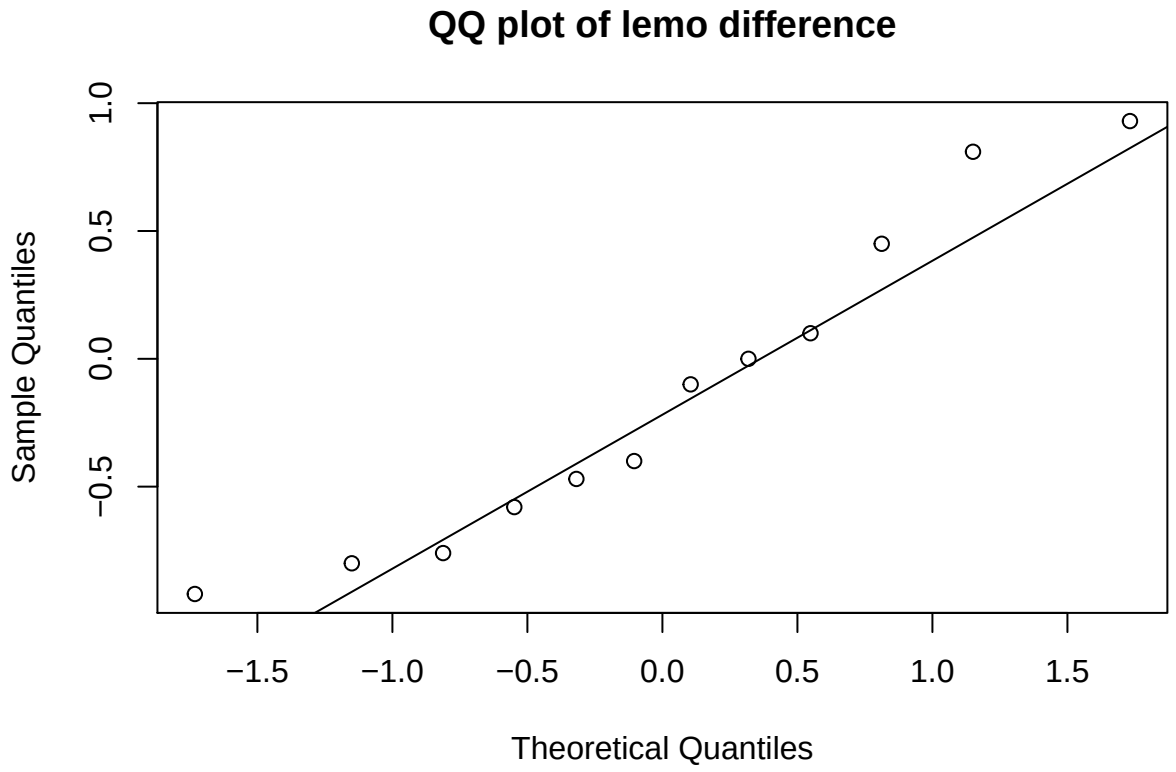
```
##
##  Pearson's product-moment correlation
##
## data:  before and after
## t = 3.8944, df = 22, p-value = 0.00078
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3171271 0.8286612
## sample estimates:
##      cor
## 0.638803
```

```
# Check normality assumption
par(mfrow = c(1,2))
qqnorm(before, main = "Normal Q-Q plot of runtime before")
qqline(before)
qqnorm(after, main = "Normal Q-Q plot runtime after")
qqline(after)
```

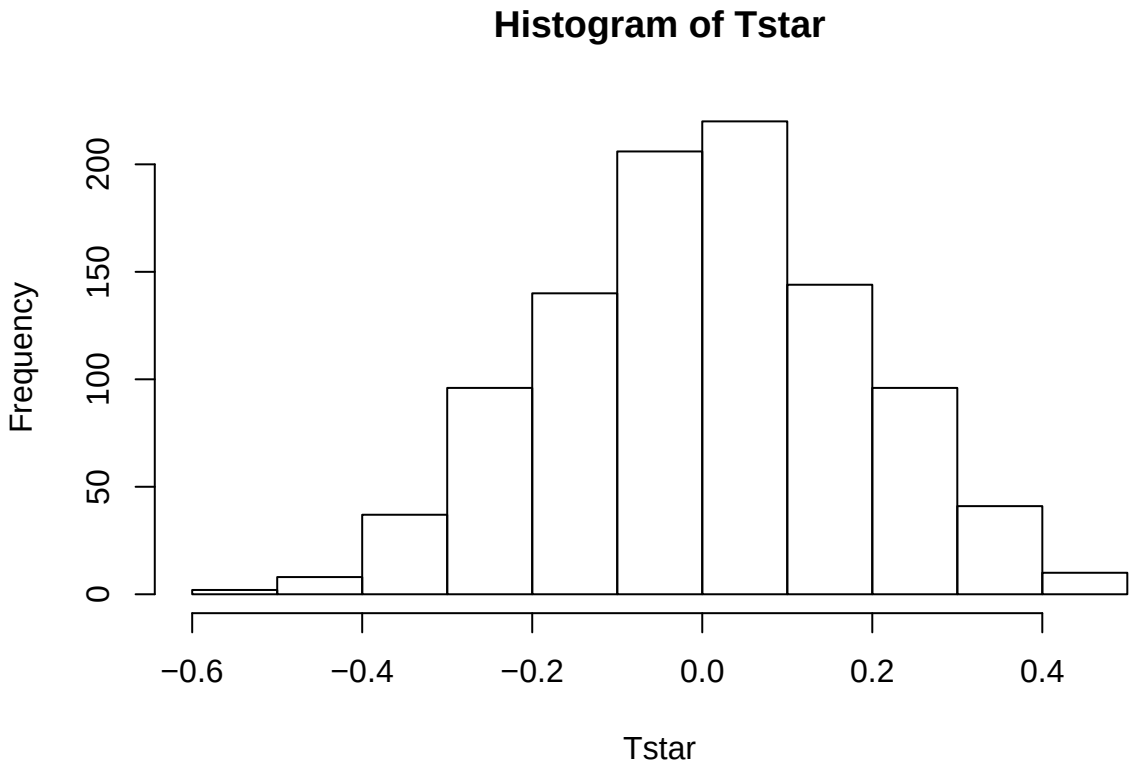


b) Test separately, for both the softdrink and the energy drink conditions, whether there is a differencein speed in the two running tasks.

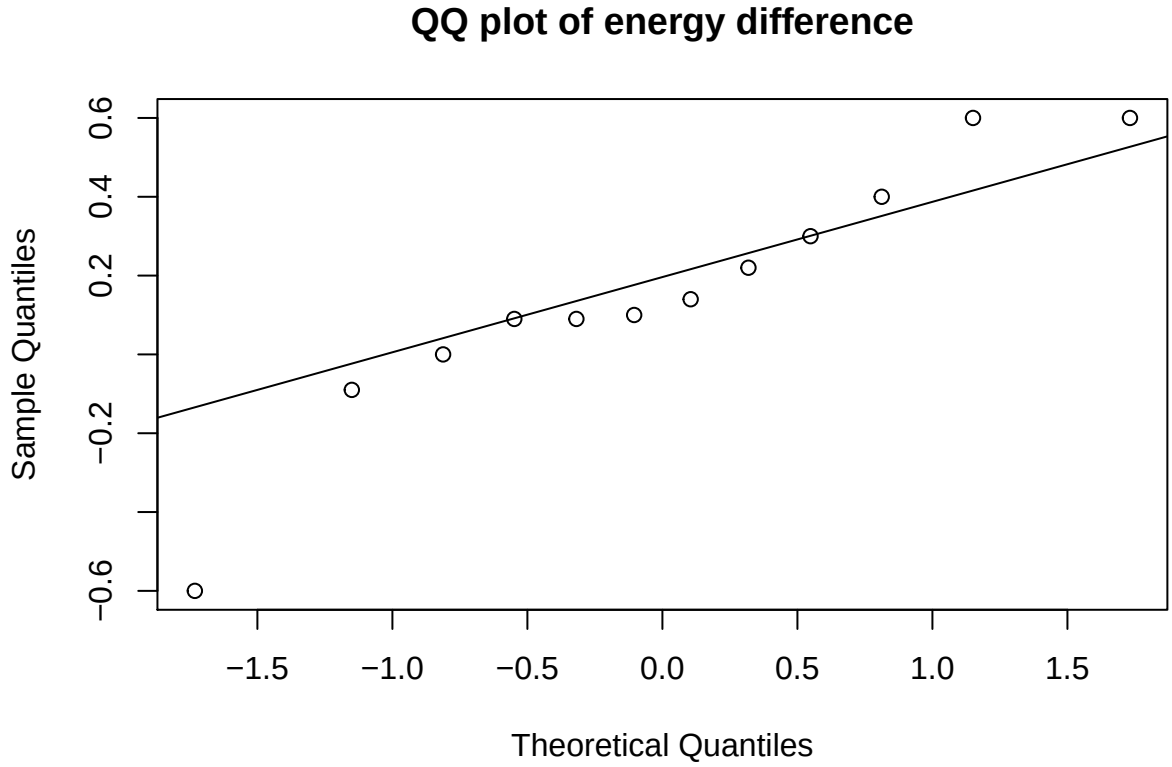
```
lemo <- split(run_raw, f = run_raw$drink)["lemo"]
energy <- split(run_raw, f = run_raw$drink)["energy"]

permutation_test <- function(x, y) {
  B <- 1000
  Tstar <- numeric(B)
  for(i in 1:B) {
    Xstar <- t(apply(cbind(x, y), 1, sample))
    Tstar[i] <- mean(Xstar[,1] - Xstar[,2])
  }
  t <- mean(x - y)
  hist(Tstar)
  p1 <- sum(Tstar < t) / B
  pr <- sum(Tstar > t) / B
  p <- 2 * min(p1, pr)
  return(p)
}

par(mfrow = c(1,1))
qqnorm(lemo$lemo$before - lemo$lemo$after, main = "QQ plot of lemo difference")
qqline(lemo$lemo$before - lemo$lemo$after)
```



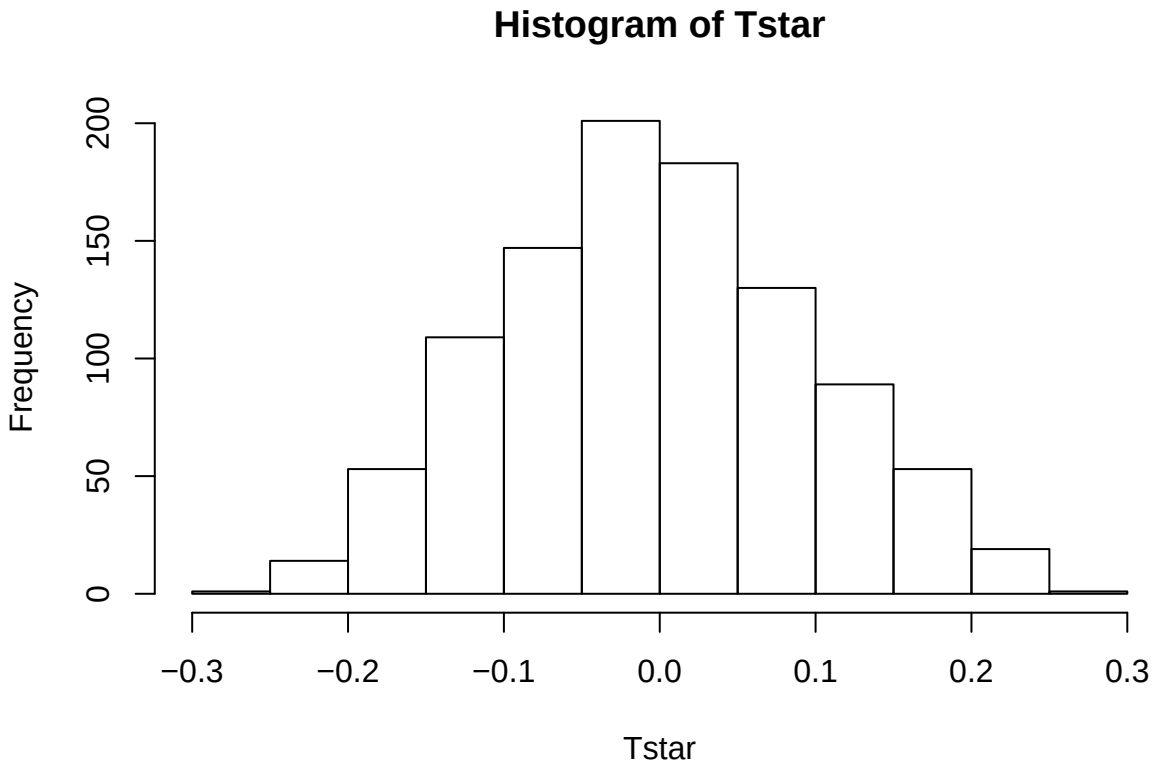

```
p_lemo <- permutation_test(lemo$lemo$before, lemo$lemo$after)
```




```
par(mfrow = c(1,1))
qqnorm(energy$energy$before - energy$energy$after, main = "QQ plot of energy difference")
qqline(energy$energy$before - energy$energy$after)
```



```r
p_energy <- permutation_test(energy$energy$before, energy$energy$after)
```



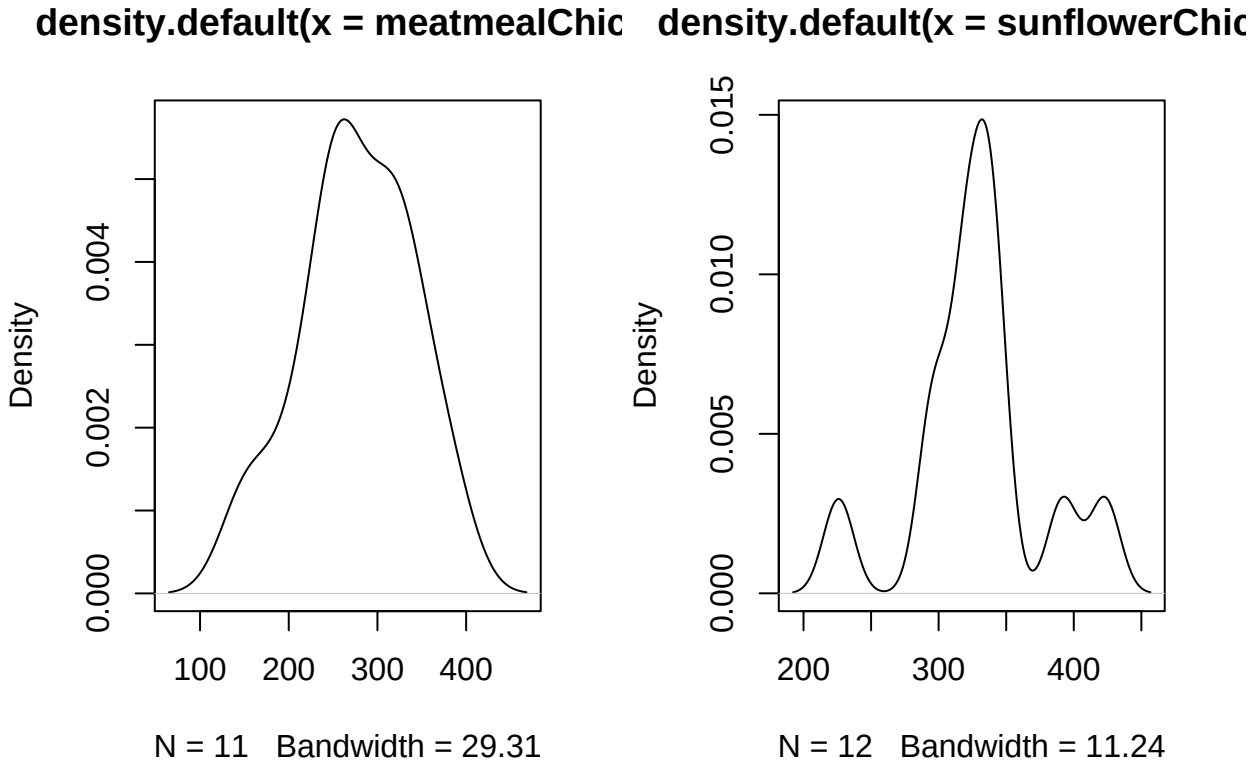
c) For each pupil compute the time difference between the two running tasks. Test whether these time differences are effected by the type of drink.

d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

Exercise 5

a) Test whether the distributions of the chicken weights for meatmeal and sunflower groups are different by performing three tests: the two samples t-test (argue whether the data are paired or not), the Mann-Whitney test and the Kolmogorov-Smirnov test. Comment on your findings.

```r
data("chickwts")
meatmealChick <- chickwts$weight[chickwts$feed == 'meatmeal']
sunflowerChick <- chickwts$weight[chickwts$feed == 'sunflower']
par(mfrow = c(1,2))
plot(density(meatmealChick))
plot(density(sunflowerChick))
```




```
data("chickwts")
par(mfrow = c(1,1))
t.test(meatmealChick, sunflowerChick, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  meatmealChick and sunflowerChick
## t = -2.1838, df = 21, p-value = 0.04047
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -101.53409   -2.48106
## sample estimates:
## mean of x mean of y
##  276.9091  328.9167
```

p-value = 0.04047 < 0.05, so we reject H0 and conclude that a significant difference does exist, which means the mean chicken weights of meatmeal and sunflower groups are different.

It requires both types of weights should be normally distributed and Variances between two groups be equal. According to density plots, the distribution of sunflower is non-normal distribution, so the two-sample t-test is invalid.

```
wilcox.test(meatmealChick, sunflowerChick)
```

```
##
##  Wilcoxon rank sum test
##
## data:  meatmealChick and sunflowerChick
## W = 36, p-value = 0.06882
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is 0.06882 > 0.05, If the p-value is larger than 0.05, we cannot conclude that a significant difference exists. The conclusion is that the chicken weights of meatmeal and sunflower groups share the same median.

We require the two data samples are independent and the samples do not affect each other. So in this case, the test is valid.

```
ks.test(meatmealChick, sunflowerChick)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  meatmealChick and sunflowerChick
## D = 0.47727, p-value = 0.1085
## alternative hypothesis: two-sided
```

The p-value is 0.1085 > 0.05, the chicken weights of meatmeal and sunflower groups should be identical and balanced in median, variability, and the shape of the distribution.

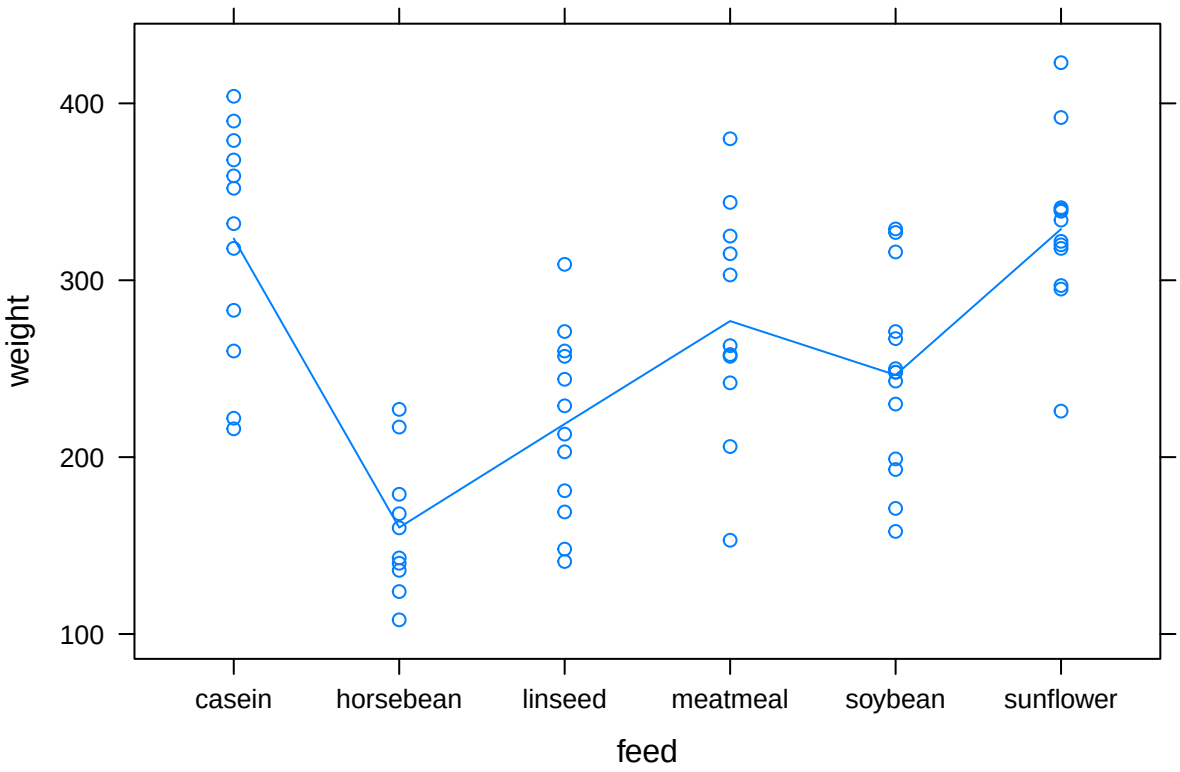
b) Conduct a one-way ANOVA to determine whether the type of feed supplement has an effect on the weight of the chicks. Give the estimated chick weights for each of the six feed supplements. What is the best feed supplement?

```
chickaov=lm(weight ~ feed, data=chickwts)
anova(chickaov)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## feed       5 231129   46226  15.365 5.936e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the test is highly significant (p=5.94e-10), therefore we conclude the alternative, that at least one feed type has a different average weight. Hence the type of feed supplement has an effect on the weight of the chicks.

```
library("lattice")
xyplot(weight ~ feed, data=chickwts, main="", type=c("p","a"))
```



The plot above explains which feed types are producing the highest average weights.

```
summary(chickaov)
```

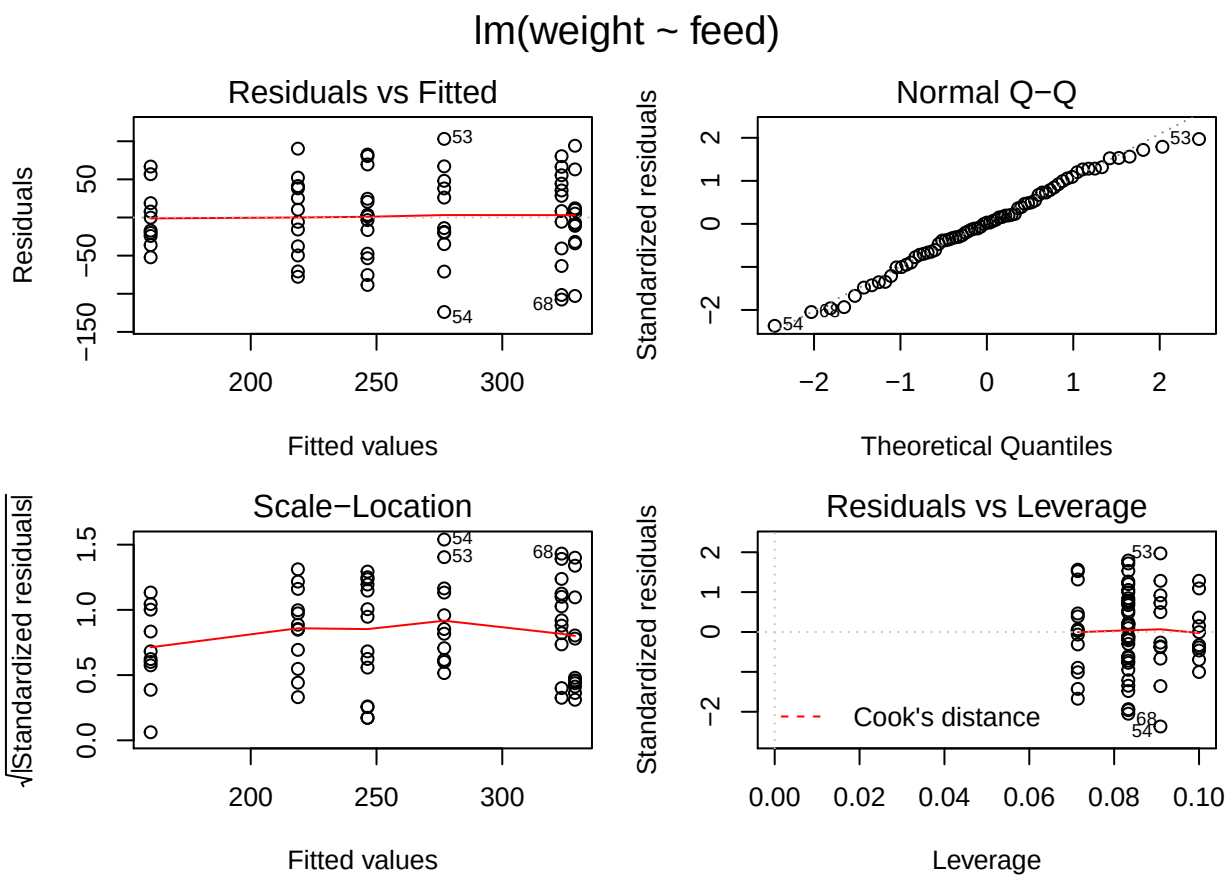
```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    323.583     15.834  20.436  < 2e-16 ***
## feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
## feedsoybean    -77.155     21.578  -3.576 0.000665 ***
## feedsunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

The estimated chick weights for each of the six supplements are: casein 323.6, horsebean: 160.2, linseed: 218.8, meatmeal: 276.9, soybean: 246.4, sunflower: 328.9.

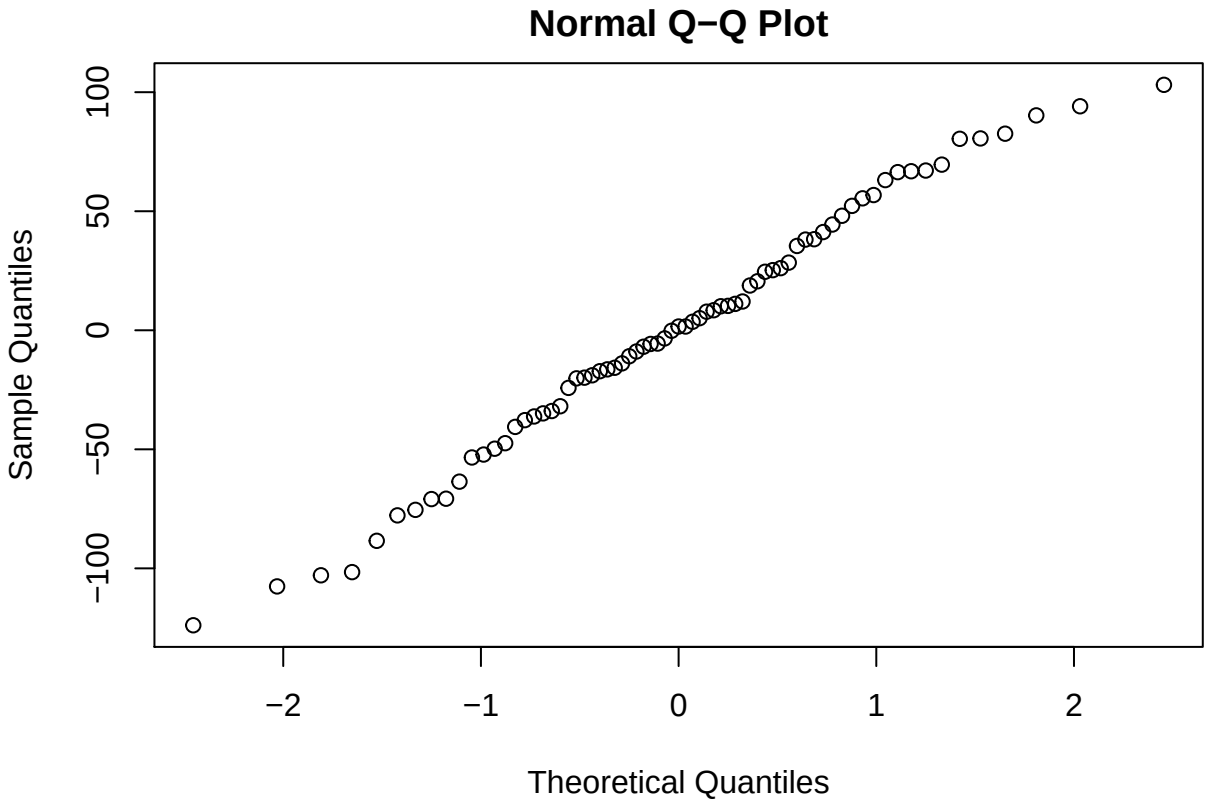
Sunflower and casein are the best feed supplements.

c) Check the ANOVA model assumptions by using relevant diagnostic tools.

```
opar <- par(mfrow = c(2, 2), oma = c(0, 0, 1.1, 0),
            mar = c(4.1, 4.1, 2.1, 1.1))
plot(chickaov)
```




```
par(mfrow=c(1,1))
qqnorm(residuals(chickaov))
```



The plots demonstrate that the assumptions of ANOVA were satisfied as the residuals versus fitted values plot shows roughly constant variance, and the QQ-Plot shows normality of the residuals.

d) Does the Kruskal-Wallis test arrive at the same conclusion about the effect of feed supplement as the test in b)? Explain possible differences between this conclusion and the conclusion from b).

```
kruskal.test(weight ~ feed, data=chickwts)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight by feed
## Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

The command kruskal.test performs the Kruskal-Wallis test and yields a p-value. The p-value for testing H0 is 5.113e-07, hence H0 is rejected. the type of feed supplement has an effect on the weight of the chicks.

The one-way ANOVA also yield a significant difference. From the qq plot above,the residuals do not seem to deviate significantly from normal, and both tests could be used here.