Assignment 1

Shutong Cai, group 20

26 February 2023

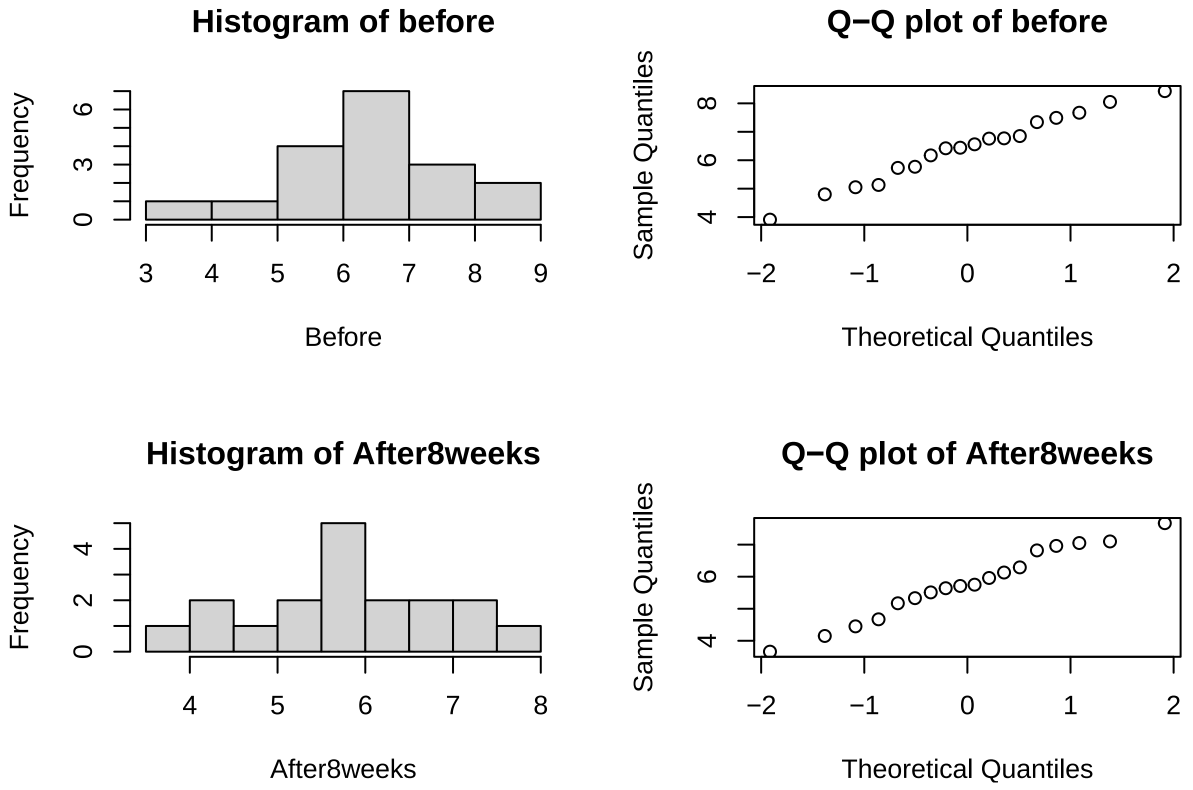
Exercise 2

Before conducting the experiment, we shall first import the data set and attach the data to the *R* search path so that all the objects can be easily accessed by giving their names.

```
cholesterol = read.table("cholesterol.txt", header = TRUE)
attach(cholesterol) # Attach the date set to the R search path
```

a) Normality check: Two kinds relevant plots are made to check the data normality, which are the histogram and Q-Q plot.

```
par(mfrow=c(2,2)) # two plots in one row next to each other
hist(Before, main = "Histogram of before")
qqnorm(Before, main = "Q-Q plot of before")
hist(After8weeks, main = "Histogram of After8weeks")
qqnorm(After8weeks, main = "Q-Q plot of After8weeks")
```

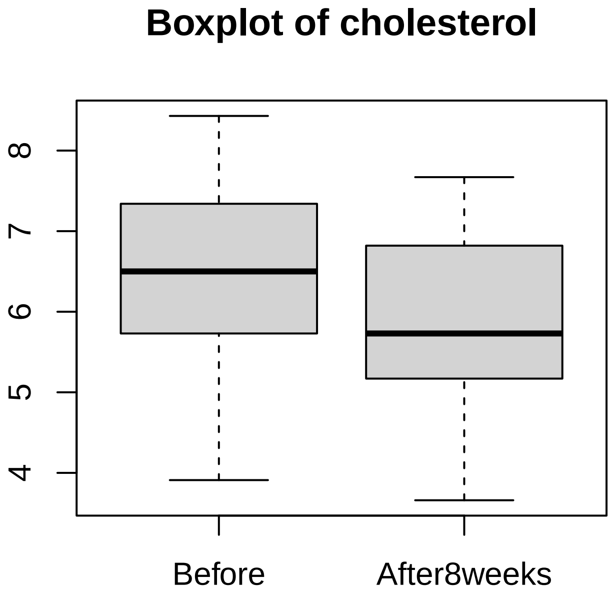


We can conclude that data of *Before* and *After8weeks* is normally distributed. The two histograms are bell-shaped the majority of the data is concentrated in the middle and the rest is distributed at two tails.

Also, these two histograms are approximately symmetrical, although the histogram of *Before* is skewed to the right slightly. As for the Q-Q plots, the points are approximately on a straight line. Thus the data of *Before* and *After8weeks* can be assumed to follow a normal distribution.

Consistency check: Box-plots are usually used to check whether there are any inconsistencies in the data.

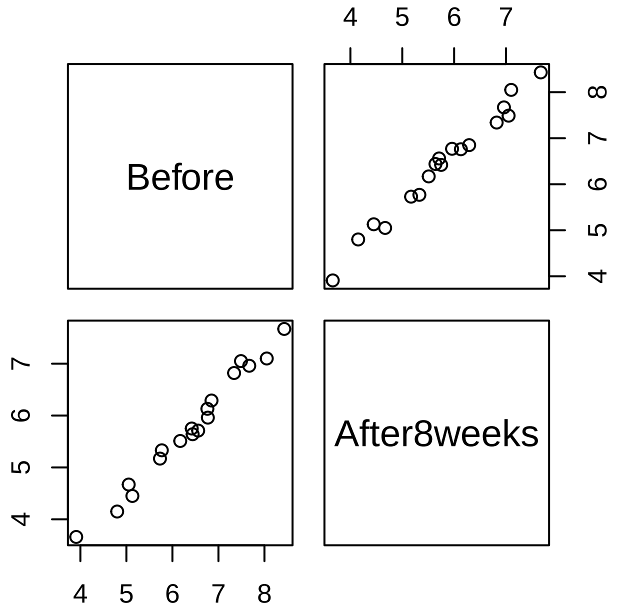
```
boxplot(Before,After8weeks,names = c("Before","After8weeks"), main = "Boxplot of cholesterol")
```



Thus we can draw the conclusion that there is no inconsistency in the data, because there is no outline shown in the box plot.

Correlation check: To check the correlation between the column *Before* and *After8weeks*, we use the scatter plot of these two variables and compute the correlation values between them.

```
pairs(cholesterol) # scatter plot of Before and After8weeks
```



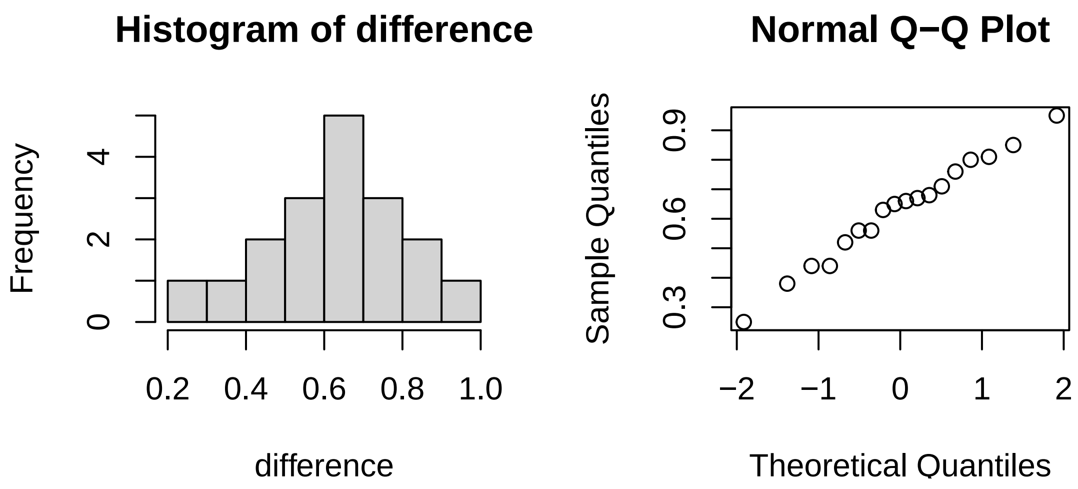
```
cor(Before, After8weeks) # calculate correlation
```

```
## [1] 0.9908885
```

From the scatter plot and the correlation value $\rho =$ `0.99`, we can conclude that these two columns *Before* and *After8weeks* are correlated, because the points of the scatter plot of these variables are approximately on a straight line and the correlation value ρ is very close to `1` which proves to be linear relation.

b) We use t-test and permutation test to verify whether the diet has an effect.
T-test: Before the t-test, we shall ensure that the data we test is normally distributed. Additionally, the two variables are matching data based on the cholesterol level, thus the data is paired. So we can use the paired t-test, first, check the normality of data, then execute the t-test, and finally conclude by p-value.

```
difference = Before-After8weeks
par(mfrow=c(1,2)) # two plots in one row next to each other
hist(difference)
qqnorm(difference)
```



```
shapiro.test(difference) # check the data normaliy of differences of before and after8weeks
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference
## W = 0.98501, p-value = 0.9869
```

```
t.test(Before - After8weeks) # execute paired t-test
```

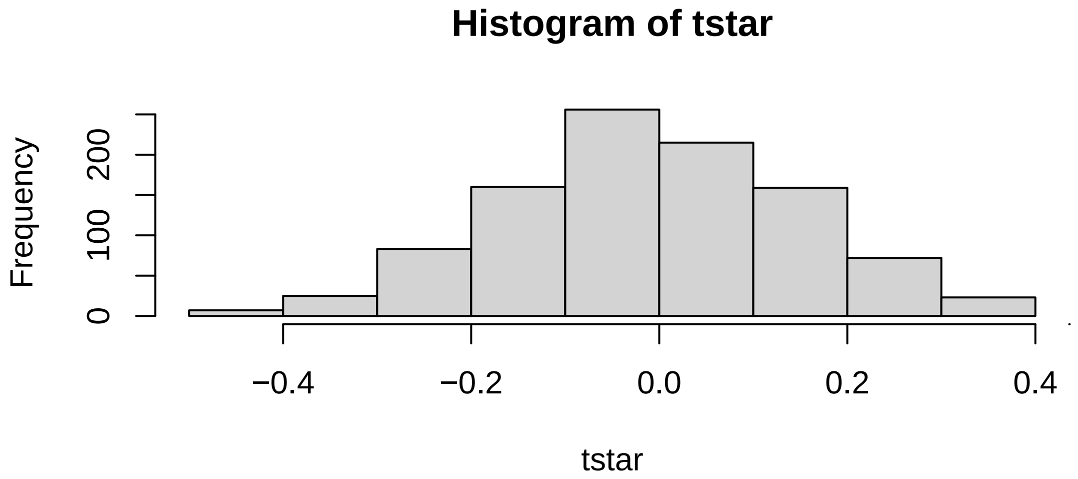
```
##
##  One Sample t-test
##
## data:  Before - After8weeks
## t = 14.946, df = 17, p-value = 3.279e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5401131 0.7176646
## sample estimates:
## mean of x
## 0.6288889
```

According to plots of *difference* and `p-value` of the Shapiro-Wilk test is `0.987 > 0.05`, the *difference* follows the normal distribution. So we can conclude that the diet does have an effect because the `p-value` of this paired sample test is `3.28e-11 < 0.05`, which means H_0(Difference is equal to 0) is rejected.
Permutation test: Because what we care about is whether the difference of *Before* and *After8weeks* is 0, permutation is applicable here. In this permutation test, we use the `mean(x-y)`as the test statistic.

```
mystat = function(x,y){ mean(x-y)} # set test statistic
B = 1000 # the number of T-star
tstar = numeric(B)
for(i in 1: B){
  cholesterolstar = t(apply(cbind(Before, After8weeks), 1, sample)) # generate x-star and y-star
  tstar[i] = mystat(cholesterolstar[,1], cholesterolstar[,2])} # generate T-star}
myt = mystat(Before, After8weeks) # compute original test statisc t
print(myt)
```

```
## [1] 0.6288889
```

```
# plot histogram of T-star with original test statistic t
par(mfrow = c(1,1))
hist(tstar)
lines(rep(myt,2),c(0,30), col = "red", lwd = 2)
axis(1,myt,expression(paste("t")))
```



```
pl = sum(tstar<myt)/B
pr = sum(tstar>myt)/B
p = 2*min(pl,pr) # calculate p-value
```

The p-value `p = 0 < 0.05`, so we reject H_0, and conclude that there are differences between these two variables, that is to say, the diet has an effect on the cholesterol level.

c) Because of the central limit theorem(CLT), the mean of $\bar{X}_i$ follows the normal distribution, which means we can estimate the mean of the population $\hat{\mu}$ equals to the mean of $\bar{X}_i$, hence we can get the $\hat{\theta}$ by using the property of uniform distribution.

```
B  = 1000; a = 3; alpha = 0.05; n = 18 # set parameters
theta = max(After8weeks)
# calculate estimate theta
x_bar = numeric(B)
for(i in 1:B){
  x = runif(n, min = a, max = theta)  # generate a bunch of samples
  x_bar[i] = mean(x)
}
mu = mean(x_bar)
estitheta = 2*mu - a # use (a+b)/2 of expectaion formula in uniform distribution
# calculate 95% confidence interval
s = (estitheta - a)^2/12 # use the variance formula in uniform distribution
```

```r
t_quantile = qt(1-alpha/2, df = n-1)
E = t_quantile*s/sqrt(n) # calculate margin error
cll = estitheta - E # calculate the left side CI
clr = estitheta + E # calculate the right side CI
```

Here we have $\hat{\theta} =$ 7.67, 95% confidence interval(CI) of [6.77,8.57]. **In order to improve the CI,**

d) First we shall apply the bootstrap test to verify the H_0, then apply the Kolmogorov-Smirnov test to check whether X follows the uniform distribution.
Bootstrap test:

```r
B = 1000; n =18; # set parameters
T = max(After8weeks) # calculate original test statistic T
Tstar = numeric(B)
for(i in 1:B){
  Theta = runif(1, min =3, max = 12) # generate a Theta randomly
  Xstar = runif(n, min = a, max = Theta) # generate the X-star
  Tstar[i] = max(Xstar) # calculate T-star
}
pl = sum(Tstar<T)/B
pr = sum(Tstar>T)/B
p = 2*min(pl,pr) # calculate p value
```

Here we get the p-value p = 0.862 > 0.05, so we accept H_0.
Kolmogorov-Smirnov test: In this case, two independent samples X_i is the generated data under H_0, and Y_i is the data follows the uniform distribution.

```r
# generate data set under H0
Theta = runif(1, min =3, max = 12)
Sampleunderh0 = runif(n, min = a, max = Theta)
# ks.test
ks.test(Sampleunderh0,"punif", min = a, max = Theta)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  Sampleunderh0
## D = 0.12998, p-value = 0.8837
## alternative hypothesis: two-sided
```

We get p-value = 0.648 > 0.05, so we accept H_0. Both the bootstrap test and the Kolmogorov-Smirnov test can be applied in this situation of checking whether X_i follows the uniform distribution with specif interval.

e) We first use the sign test to verify whether the median cholesterol level after 8 weeks of low fat diet is less than 6.

```r
sumoflessthan6 = sum(After8weeks < 6)
binom.test(sumoflessthan6, 18, p=0.5, alt = "l")
```

```
##
##  Exact binomial test
##
## data:  sumoflessthan6 and 18
## number of successes = 11, number of trials = 18, p-value = 0.8811
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
```

```
##  0.0000000 0.8010467
## sample estimates:
## probability of success
##              0.6111111
```

We get `p-value = 0.88 > 0.05`, so we accept the H_0: the median value of the data *After8weeks* is more than 6. Then to check whether the fraction of cholesterol levels after 8 weeks of low fat less than `4.5` is at most `25%`. First we shall set H_0: the fraction `p > 25%`, then we use sign test to verify the 25% quantile.

```r
sumoflessthan4_5 = sum(After8weeks < 4.5)
binom.test(sumoflessthan4_5, n, p = 0.25, alt = "l")
```

```
##
##  Exact binomial test
##
## data:  sumoflessthan4_5 and n
## number of successes = 3, number of trials = 18, p-value = 0.3057
## alternative hypothesis: true probability of success is less than 0.25
## 95 percent confidence interval:
##  0.0000000 0.3766792
## sample estimates:
## probability of success
##              0.1666667
```

We get `p-value = 0.31 > 0.05`, so we accept H_0, that is to say the fraction pf cholesterol level after 8 weeks of lo far diet less than `4.5` is more than 25%.