Assignment 2

Gijs Smeets, Daan Wijnhorst, Moos Middelkoop group 35

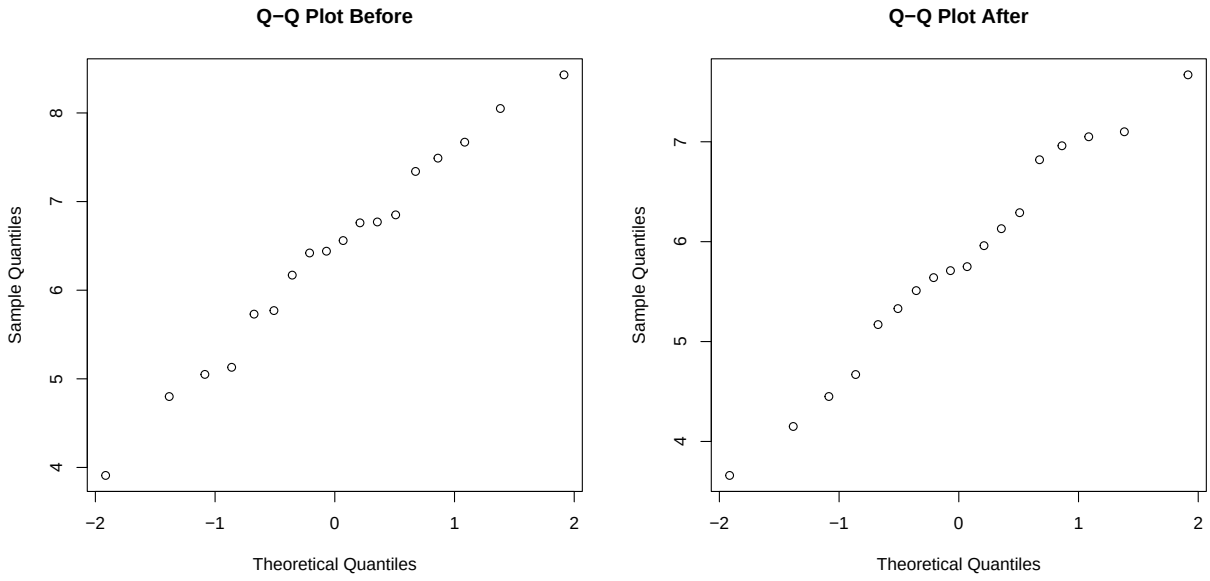
February 2023

Exercise 2

A

```
data = read.table('../data/cholesterol.txt',header=TRUE)
before = data$Before
after = data$After8weeks
```

To check for normality, we look at the Q-Q plots



Here we see that the sampled data is approximately normal

```
summary(data)
```

```
##      Before       After8weeks
##  Min.   :3.910   Min.   :3.660
##  1st Qu.:5.740   1st Qu.:5.210
##  Median :6.500   Median :5.730
##  Mean   :6.408   Mean   :5.779
##  3rd Qu.:7.218   3rd Qu.:6.688
##  Max.   :8.430   Max.   :7.670
```

```
shapiro.test(before)
```

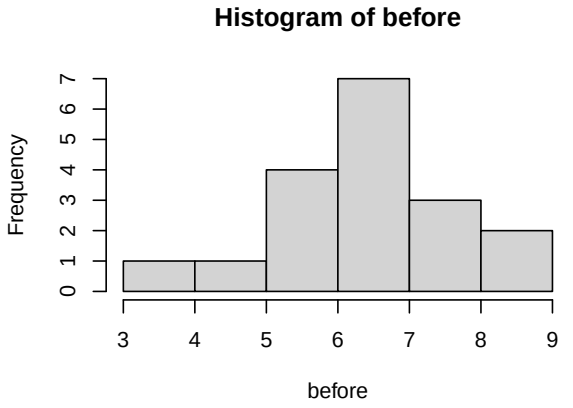
```
##
##  Shapiro-Wilk normality test
##
## data:  before
## W = 0.9819, p-value = 0.9675
```

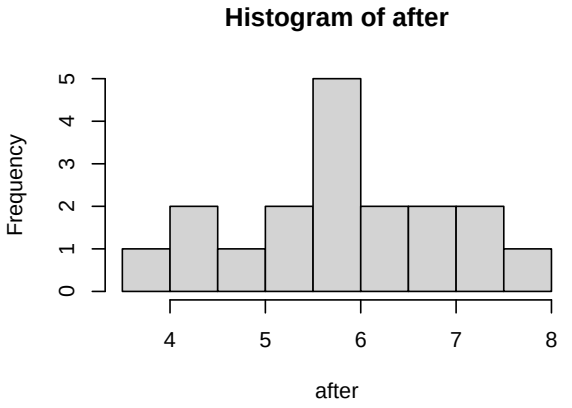
```
shapiro.test(after)
```

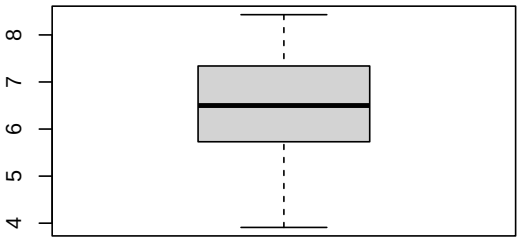
```
##
##  Shapiro-Wilk normality test
##
## data:  after
## W = 0.97733, p-value = 0.9183
```

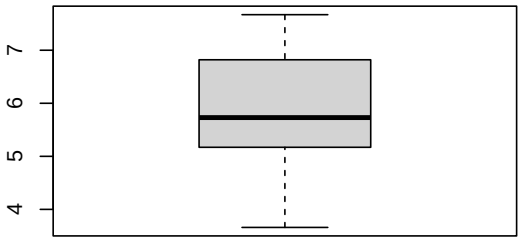
To complement the visual check, we use the Shapiro-Wilk test. In this test, the H0 is that the data is normal. H1 is that the data is not normal.

As the p-value from the Shapiro-Wilk test is high for both values we do not have enough evidence to reject the null-hypothesis. If we combine this result with the visual check we can safely assume the data to be normally distributed.

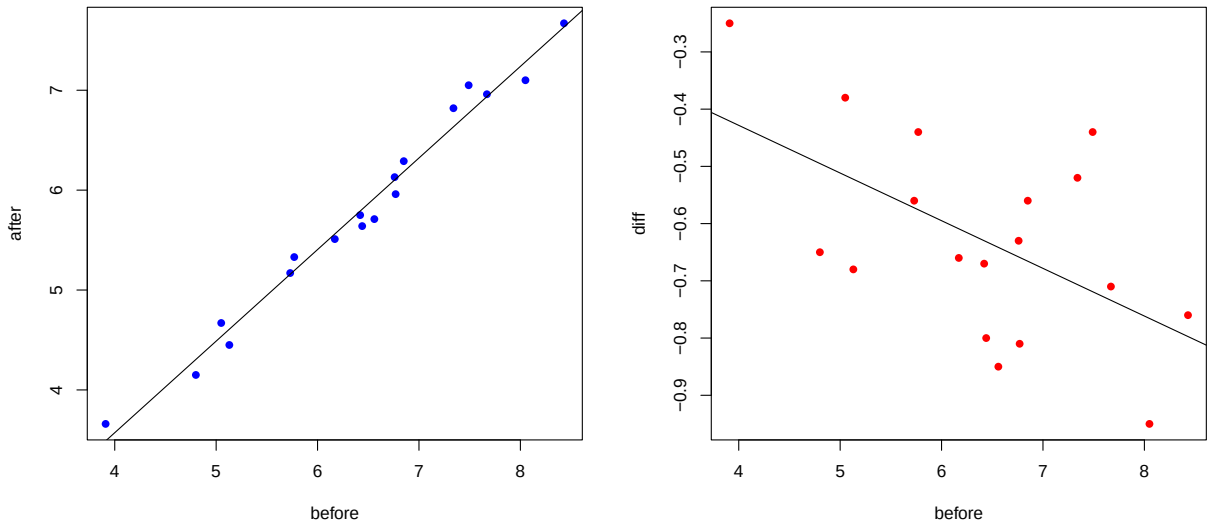









```
##
## Call:
## lm(formula = after ~ before)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20843 -0.15401  0.01823  0.12825  0.27904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.09513    0.20283  -0.469    0.645
## before       0.91670    0.03115  29.428 2.32e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.153 on 16 degrees of freedom
## Multiple R-squared:  0.9819, Adjusted R-squared:  0.9807
## F-statistic:   866 on 1 and 16 DF,  p-value: 2.321e-15
```




```
##
## Call:
## lm(formula = diff ~ before)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20843 -0.15401  0.01823  0.12825  0.27904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.09513    0.20283  -0.469   0.6454
## before      -0.08330    0.03115  -2.674   0.0166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.153 on 16 degrees of freedom
## Multiple R-squared:  0.3089, Adjusted R-squared:  0.2657
## F-statistic: 7.151 on 1 and 16 DF,  p-value: 0.01663
```

adj R^2 = 0.98 explained, correlated positively. From the plots it becomes clear that people who have higher cholesterol levels in the before data have a higher absolute decrease of cholesterol after the experiment. It becomes easier to see if we remove before bias and only look at the differences (see the right red plot)

B

Pearson Correlation test As we have concluded the data to be normally distributed in A, we can conduct a Paired-Sampled T-test. In this test, H0: the mean difference between the values of X and Y are 0 H1: this is not the case, the difference is not 0.

```r
#T-test
t.test(before, after, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 14.946, df = 17, p-value = 3.279e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.5401131 0.7176646
## sample estimates:
## mean difference
##       0.6288889
```

As the p-value is below 0.05 we reject H0 of the T-test and conclude the mean paired differences are not equal to zero.

Another relevant test is testing for correlation.

We use the Pearson's correlation test as we have considered the data to be normally distributed in subsection A.

```r
#pearson corr. test
par(mfrow=c(1,2))
cor.test(before,after)
```

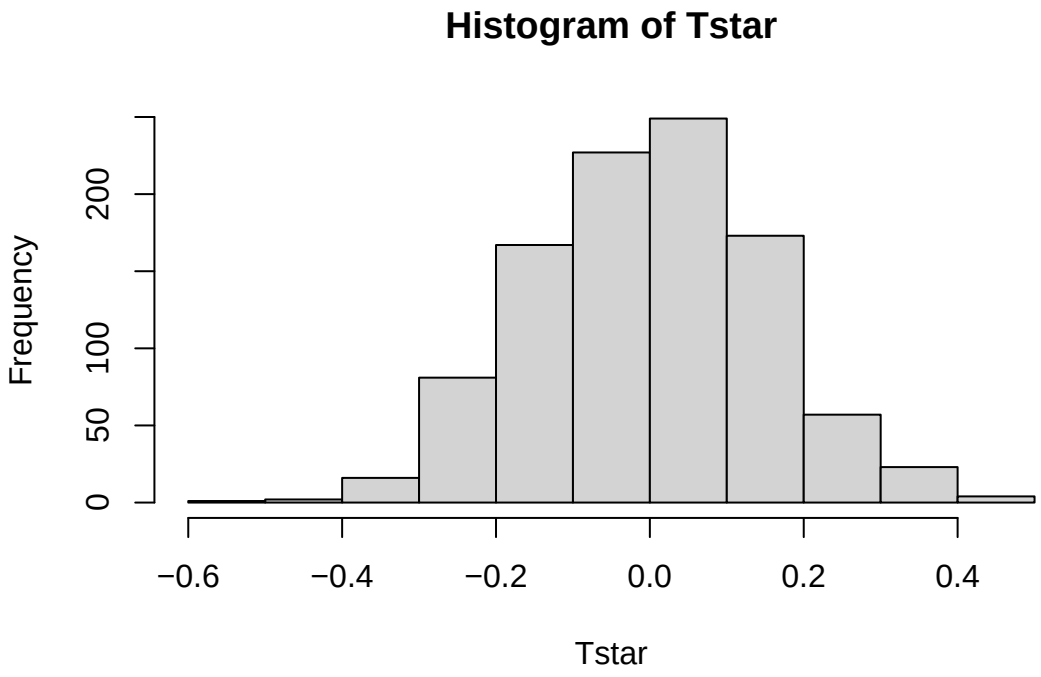
```
##
##  Pearson's product-moment correlation
##
## data:  before and after
## t = 29.428, df = 16, p-value = 2.321e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9751289 0.9966788
## sample estimates:
##       cor
## 0.9908885
```

Pearson's returns the correlation value of 0.9908885 herefore we conclude there exists a strong (positive) correlation between the two variables.

Combining the results of the two tests, we conclude that the low fat margarine diet has an effect.

Is a permutation test applicable? As we are dealing with numerical outcomes, two conditions per experimental unit and we are interested in possible differences between two outcomes per unit, we can devise a permutation test.

```r
meansamples = function(x,y) {mean(x-y)}
B=1000; Tstar = numeric(B)
for (i in 1:B){
  dietstar = t(apply(cbind(before,after),1,sample))
  Tstar[i] = meansamples(dietstar[,1],dietstar[,2])
}
myt = meansamples(before,after);
hist(Tstar)
```




```r
pl = sum(Tstar<myt)/B
pr = sum(Tstar>myt)/B
p = 2*min(pl,pr)
p
```

```
## [1] 0
```

As P = 0 we conclude that there is indeed a significant difference between the before and after data.

C

As we can assume from the assignment that the sample has an underlying uniform distribution, we can construct the point estimate as followed:

```r
# function that computes variance of uniformly distr. variable
unif_var = function(a,b){
  return ((1/12)*((b-a)^2))
}

# E(X) = (a+b)/2 = mean(after)
# So point estimate of b is 2*E(X)-3
theta_est = 2*mean(after)-3
```

Now we have the point estimate, we can construct the confidence interval for level 1-alpha. As we know sigma we can compute the CI as follows:

```r
#build 95%-CI for theta_est
n = length(after)
ci_theta = c(theta_est - qnorm(0.975)*(sqrt(unif_var(3,theta_est))/sqrt(n)), theta_est + qnorm
ci_theta
```

```
## [1] 7.816600 9.298956
```

This returns us the interval 7.816600, 9.298956. we can try to improve the CI via choosing a different estimating statistic?

D

We conduct the bootstrap test (to test H0: samples are sampled from a uniform distribution between 3 and theta) as followed:

```r
t = max(after); t
```

```
## [1] 7.67
```

```r
vP = rep(0,9)
vT = rep(0,9)
for (theta in 3:12){
  B = 1000; tstar=numeric(B)

  for (i in 1:B){
    xstar=runif(n,3,theta)
    tstar[i] = max(xstar)
  }

  pl = sum(tstar<t)/B; pr = sum(tstar>t)/B
  p = 2*min(pl,pr); p

  vT[theta-2] = theta
```

```
  vP[theta-2] = p

}
res = rbind(vT,vP); res
```

```
##    [,1] [,2] [,3] [,4] [,5]  [,6] [,7] [,8] [,9] [,10]
## vT    3    4    5    6    7 8.000 9.00   10   11    12
## vP    0    0    0    0    0 0.584 0.03    0    0     0
```

As seen in the P-values vP we do not have enough evidence to reject H0 regarding the sampled distribution belonging to a uniform distribution for Theta = 6 As B is very large, the estimation error is considered to be very small.

Kolmogorov-Smirnov test can be applied as this test tests for H0 that the two underlying distributions are the same. The uniform distributions with the range of different thetas can be distribution 1 and the sampled data can be distribution 2.

E

As we are given a small sample size, we conduct a sign test. H0: population median m = m0

```
## EXERCISE 2E
# sign test for the median with binomial distr.
s = sum(after<6); s
```

```
## [1] 11
```

```
binom.test(s,n,p=0.5,alt="g")
```

```
##
##  Exact binomial test
##
## data:  s and n
## number of successes = 11, number of trials = 18, p-value = 0.2403
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.3921553 1.0000000
## sample estimates:
## probability of success
##              0.6111111
```

As this results in a p-value of 0.2403 we do not have enough evidence to reject the null-hypothesis.

Next, we check whether the fraction of cholesterol levels lower than 4.5 in after is at most 0.25 (H0).

```r
s2 = (sum(after<4.5));
binom.test(s2,n,p=0.25,alt="l")
```

```
##
##  Exact binomial test
##
## data:  s2 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
```

As the returned p-value = 0.3057, we do not have enough evidence to reject H0.