exercise2

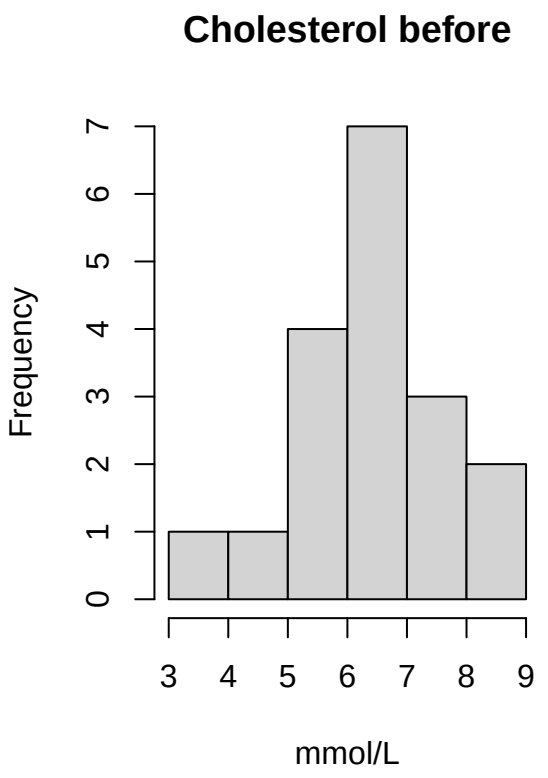
2023-02-20

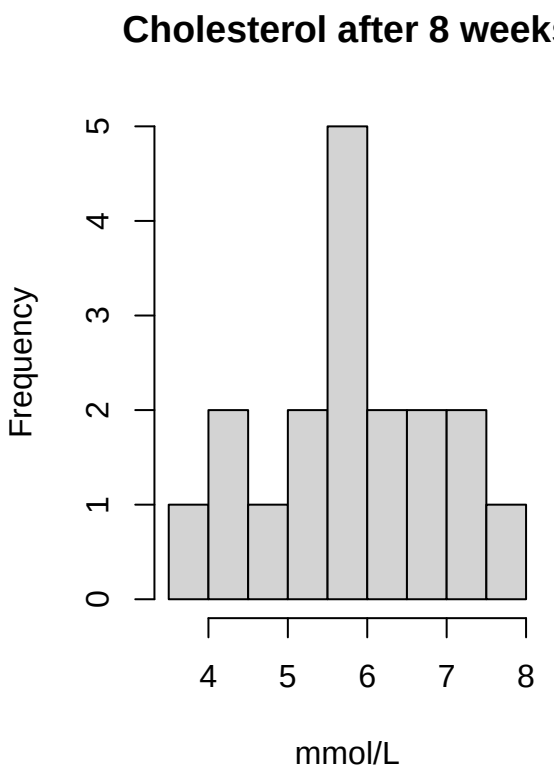
R Markdown

Exercise 2:

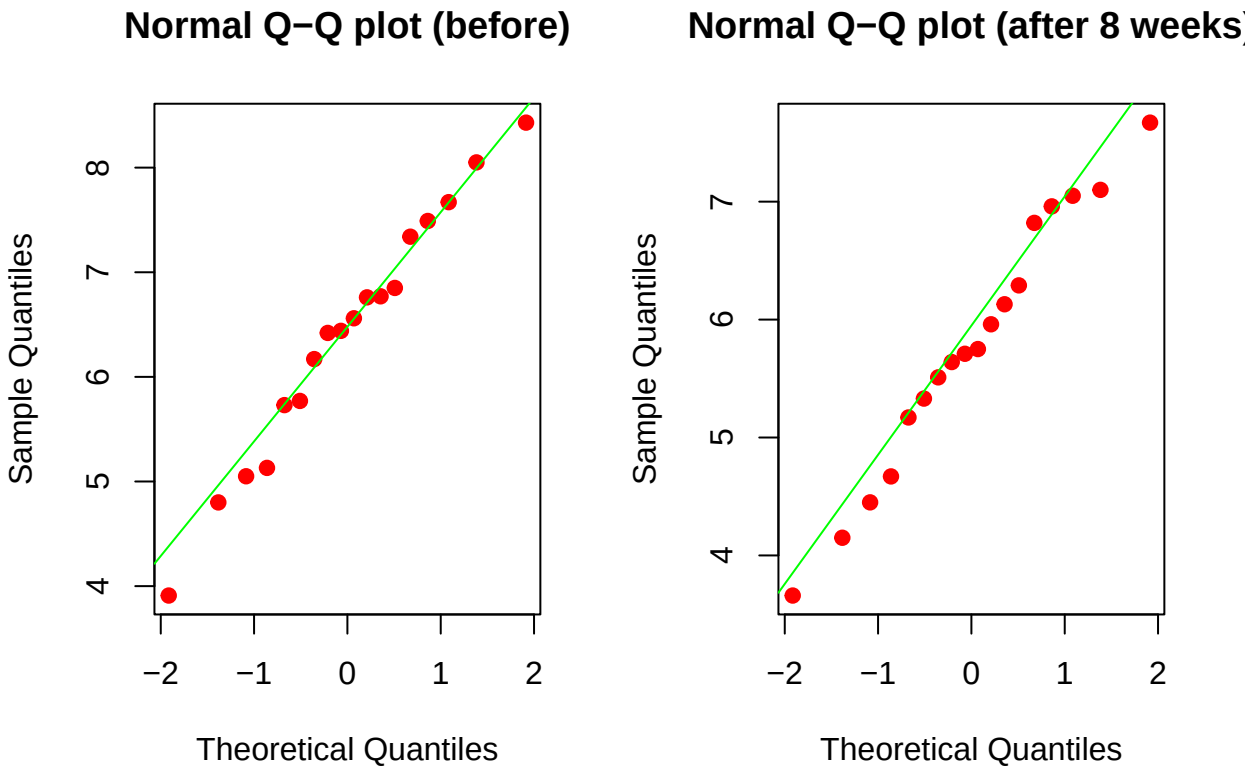
a) Histogram plots for cholesterol samples before low fat diet and after 8 weeks of low fat diet imply that data is normally distributed, since both histograms show symmetrical "bell" shape distribution. Normality is also implied by QQ plots in which a straight diagonal line show that theoretical quantiles of normal distribution match with sample quantiles (true for both samples). However, it should be taken in to the account, when assuming normality, that in both samples we only have 18 observations. Correlation between samples of cholesterol levels before, and after 8 week diet was computed to be 0.991. Since cholesterol levels were measured for the same sample of people before and after low fat diet, it can be expected that data will be highly correlated.

```
df = read.csv('Data/cholesterol.txt', header = TRUE, sep = "")
par(mfrow=c(1,2))
hist(df$Before, main = 'Cholesterol before', xlab = 'mmol/L')
hist(df$After8weeks, main = 'Cholesterol after 8 weeks', xlab = 'mmol/L')
```

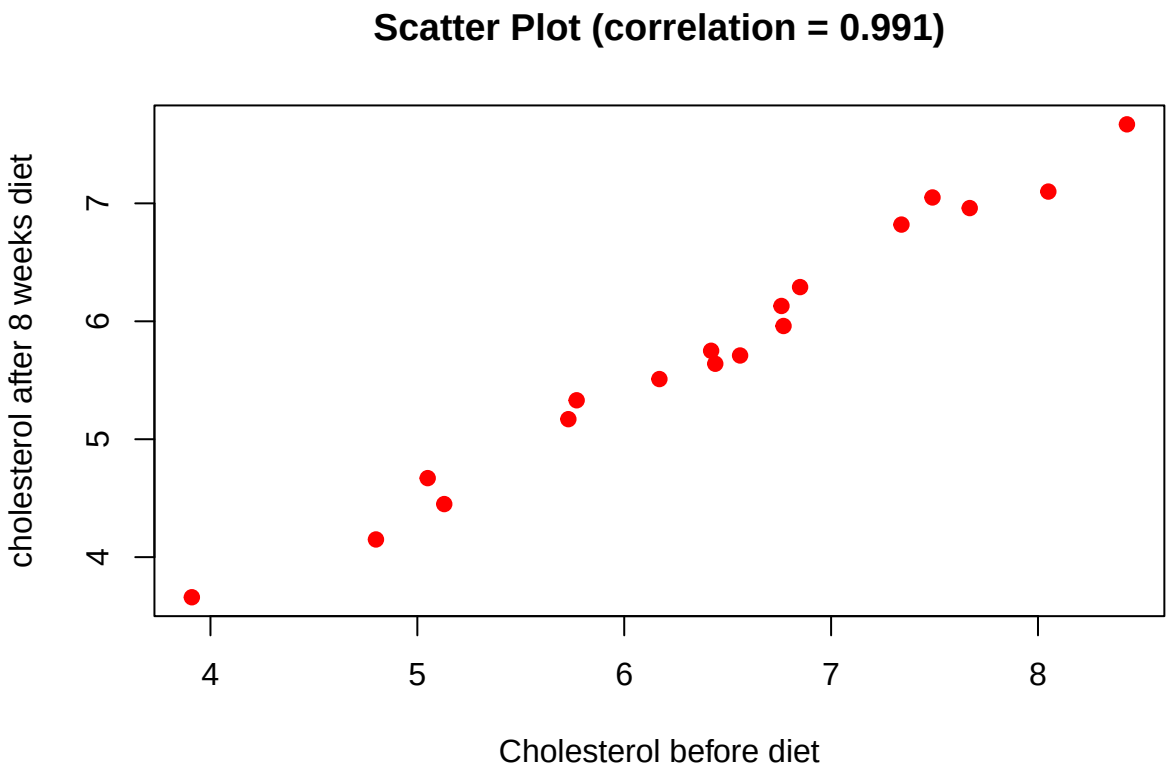




```
par(mfrow=c(1,2))
qqnorm(df$Before, col='red', pch=19, main="Normal Q-Q plot (before)")
qqline(df$Before, col='green')
qqnorm(df$After8weeks, col='red', pch=19, main="Normal Q-Q plot (after 8 weeks)")
qqline(df$After8weeks, col='green')
```




```
plot(df$Before, df$After8weeks, xlab = "Cholesterol before diet",
     ylab = "cholesterol after 8 weeks diet",
     main = paste("Scatter Plot (correlation = ", round(cor(df$Before,df$After8weeks), 3),
                  collapes=")", sep=""), col='red',pch = 19)
```



b) To verify that low fat diet is effective in lowering cholesterol levels, paired t-test and paired Wilcoxon signed rank test were constructed, where $H_0 : \mu_{before} \leq \mu_{after8weeks}$ and $H_1 : \mu_{before} > \mu_{after8weeks}$. T-test provided us with p-value equal to 0.000 which allowed us to reject H_0, therefore we can conclude that low fat diet is indeed effective in lowering cholesterol levels. The Wilcoxon signed rank test having the same hypothesis resulted in p-value also equal to 0.000, which also allows us to confirm alternative hypothesis that $\mu_{before} > \mu_{after8weeks}$ is true. Our motives for choosing t-test and Wilcoxon signed rank test come from our data properties. The data set cholesterol features two-paired samples, in which experimental units (18 people) have two numerical outcomes (cholesterol levels (mmol/L)) - before treatment (diet) and after it. Also, it must be mentioned that both samples imply to be normally distributed (see Q-Q plots above). Therefore, two-paired nature of data and normality allows us to conduct paired t-test, and symmetry of data allows us computing Wilcoxon signed rank test.

Permutation test can also be applied in this case since we have a setting of two normally distributed paired samples.

```
ttest = t.test(df$Before, df$After8weeks, alt='g', paired=TRUE)
print(paste("p-value of two-paired t-test: ",round(ttest$p.value,3)))
```

```
## [1] "p-value of two-paired t-test:  0"
```

```
wilcox_test = wilcox.test(df$Before, df$After8weeks, alt='g',paired = TRUE)
print(paste("p-value of two-paired Wilcoxon signed rank test: ",round(wilcox_test$p.value,3)))
```

```
## [1] "p-value of two-paired Wilcoxon signed rank test:  0"
```

c) Assuming that $X_1, ..., X_{18} \sim Unif[3, \theta]$, where $X_1, ..., X_{18}$ is random variable from column *after8weeks*, we applied central limit theorem by drawing 18 samples with replacement from column *after8weeks* and calculating max cholesterol level in drawn sample, this step is repeated 1000 times to collect a set of maximum values. By computing mean for aforementioned maximum values set we estimate that $\hat{\theta} = 7.43$. Our computed 95% confidence interval - $[6.96, 7.67]$.

```
sample_maxs <- c()
n = 1000
for (i in 1:n){
  sample_maxs[i] = max(sample(df$After8weeks, nrow(df), replace=TRUE))
}
estimated_upper_limit = mean(sample_maxs)
cat("Estimated Theta:", estimated_upper_limit, "\n")
```

```
## Estimated Theta: 7.46
```

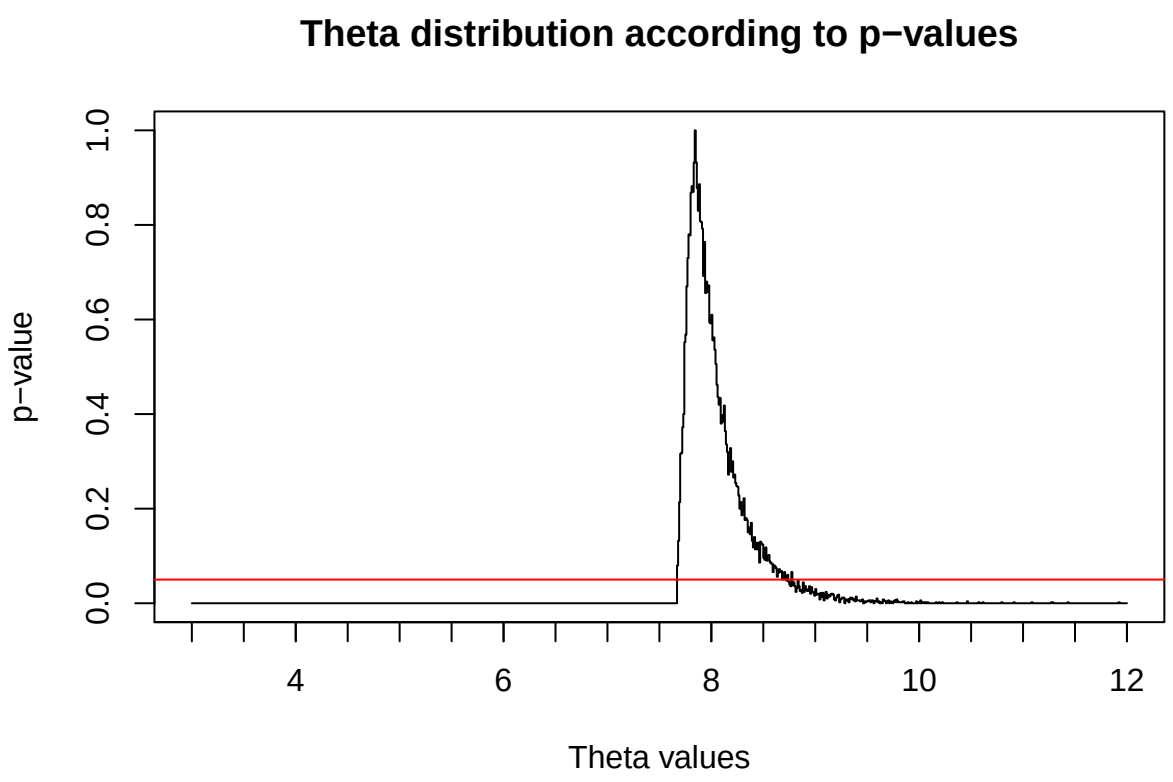
```
cat("Confidence interval:", quantile(sample_maxs, probs=c(0.025,0.975 )))
```

```
## Confidence interval: 6.96 7.67
```

d) Here, we construct bootstrap test with test statistic $T = max(X_1, ..., X_{18})$ and null hypothesis $H_0 : X_1, ..., X_{18} \sim Unif[3, \theta]$. in each bootstrap test iteration we take 18 samples from $\sim unif[3, \theta]$ distribution and we compute test statistic. This step is repeated $B = 1000$ times to obtain sample T^*, from which we then estimate p-value. At first we set $\theta = 3$ and we increment it by 0.01 after each bootstrap test, this is done until condition $\theta \leq 12$ is being satisfied. At the end we are left with a two sets of 901 θ and p-value values. Below, we plot theta and corresponding p-values, the red horizontal line marks 0.05 p-value, therefore, we can see that H_0 is not rejected when $\theta \in [7.68, 8.78]$.

Kolmogorov-Smirnov test can also be applied in this case. Let F_x denote *after8weeks* sample distribution and F_{X^*} denote distribution of $X^* \sim unif[3, \theta]$. Then with Kolmogorov-Smirnov test we would test whether we can reject $H_0 : F_x = F_{X^*}$. We would have to repeat this test for every θ in interval [3, 12] to find its values that fail to reject H_0.

```
theta = 3.00; t=max(df$After8weeks); counter = 1;B=1000;
tstar=numeric(B); p_values = c(); thetas = c();
while (theta <= 12) {
  for (i in 1:B){
    xstar = runif(n=nrow(df), min=3, max=theta)
    tstar[i]=max(xstar)
    }
  p_left=sum(tstar<t)/B; p_right=sum(tstar>t)/B;
  p_values[counter]= 2*min(p_left,p_right)
  thetas[counter] = theta
  counter = counter + 1
  theta = theta + 0.01 #increment theta by  0.01
}
plot(x=thetas, y = p_values, type = "S", xlab = "Theta values", ylab="p-value", main = "Theta distributi
axis(1,at=seq(0,12,0.5),labels=NA)
abline(a=0.05,b=0, col='red')
```




```
df_theta = data.frame(thetas,p_values)
df_theta = dplyr::filter(df_theta, p_values > 0.05)
interval <- c(min(df_theta$theta), max(df_theta$theta))

print(interval)
```

```
## [1] 7.68 8.78
```

e) To test whether median cholesterol level after 8 weeks of low fat diet is less than 6, we chose to conduct sign test, with $H_0 : m_x \leq 6$ and $H_1 : m_x > 6$. The obtained p-value of 0.9 fails to reject null hypothesis, therefore, we confirm that the median cholesterol levels after 8 weeks of low fat diet is less than 6.

To check whether the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%, we also chose sign test, with $H_0 : m_x \leq 4.5$ and $H_1 : m_x > 4.5$. Since sign test for median in R is done with binom.test(), we can specify the hypothesized probability of success of 0.25. The p-value of this sign test was compute to be equal to 0, therefore, we can reject null hypothesis. Thus, we confirm that the alternative hypothesis is true that the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%.

```
binom.test(sum(df$After8weeks > 6), nrow(df), alt='g')
```

```
##
##  Exact binomial test
##
```

```
## data:  sum(df$After8weeks > 6) and nrow(df)
## number of successes = 7, number of trials = 18, p-value = 0.9
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.199 1.000
## sample estimates:
## probability of success
##                  0.389
```

```
binom.test(sum(df$After8weeks > 4.5), nrow(df), alt='g', p=0.25)
```

```
##
##  Exact binomial test
##
## data:  sum(df$After8weeks > 4.5) and nrow(df)
## number of successes = 15, number of trials = 18, p-value = 3e-07
## alternative hypothesis: true probability of success is greater than 0.25
## 95 percent confidence interval:
##  0.623 1.000
## sample estimates:
## probability of success
##                  0.833
```