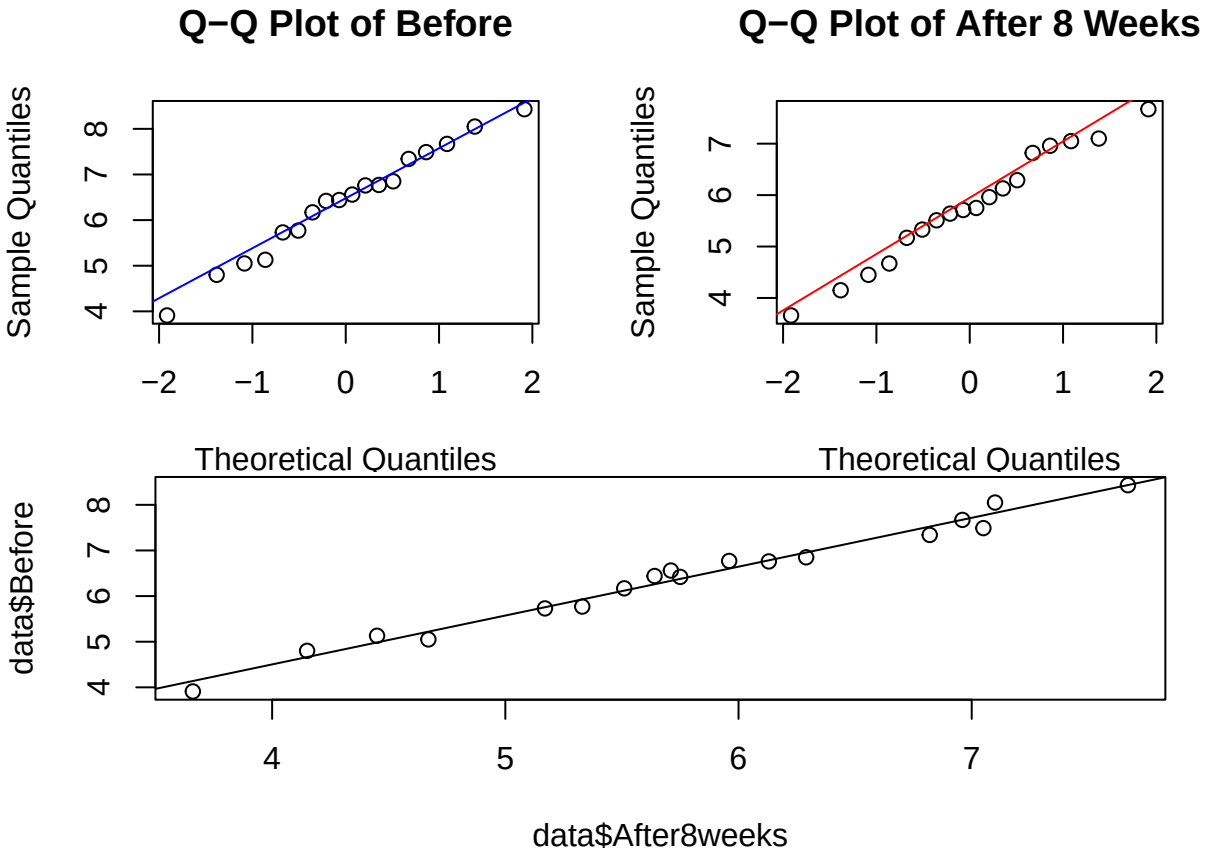
# Assignment 1 Report

2025-02-23

## Exercise 1

### a

We first plot the data to determine whether a normal distribution is followed. It is done by interpreting qqplots. The correlation between the data from before and after the diet is also tested through a correlation test (Pearson) and a scatterplot of the data against each other. These are shown below.




```
print(cor(data$Before,data$After8weeks))
```

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

```
cor.test(data$Before,data$After8weeks)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$Before and data$After8weeks
## 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
```

Due to the strong linearity of the qqplots of the data from before and after the low fat low cholesterol diet, it is assumed that the data follows a normal distribution. In terms of correlation between the data, the scatter plot suggests a strong positive trend, which is highlighted by the fitted regrsssion line. Furthermore, the Pearson correlation test, with the null hypothesis that the true correlation between the data from before and 8 weeks after adopting the diet is equal to zero, returns a p-value of 2.321e-15. Since the p-value is below 0.05, we reject the null hypothesis at a 95% confidence level, providing strong evidence of a significant correlation between the two sets of data.

**b**

In order to verify whether the diet with low fat margarine has a significant effect, various relevant tests can be applied. Identifying whether the data is paired or not allows us to choose the appropriate tests to run.Since the data from before and after 8 Weeks represent the same individual measured twice, the data is paired.Thus, a paired t-test and a Wilcoxon signed rank test are applied to the data.

```
t_paired <- t.test(data$Before, data$After8weeks, paired = TRUE)
wil_paired <- wilcox.test(data$Before, data$After8weeks, paired = TRUE)
print(t_paired)
```

```
##
##  Paired t-test
##
## data:  data$Before and data$After8weeks
## 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
```

```
print(wil_paired)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Before and data$After8weeks
## V = 171, p-value = 7.629e-06
## alternative hypothesis: true location shift is not equal to 0
```

The paired t-test and the Wilcoxon signed rank test with the null-hypothesis that the data from before and after 8 weeks of dieting are equal have respectively p-values of 3.279e-11 and 7.629e-06. Since both p-values are below 0.05, we reject the null hypothesis at a 95% confidence level, providing strong evidence that the data from before and after the dieting are not equal. Further tests providing useful information on the data are the permutation test and the Mann-Whitney test. While the permutation test does not have contrasting assumptions to the data, the Mann-Whitney test requires independent data, which makes it unsuitable for paired data.

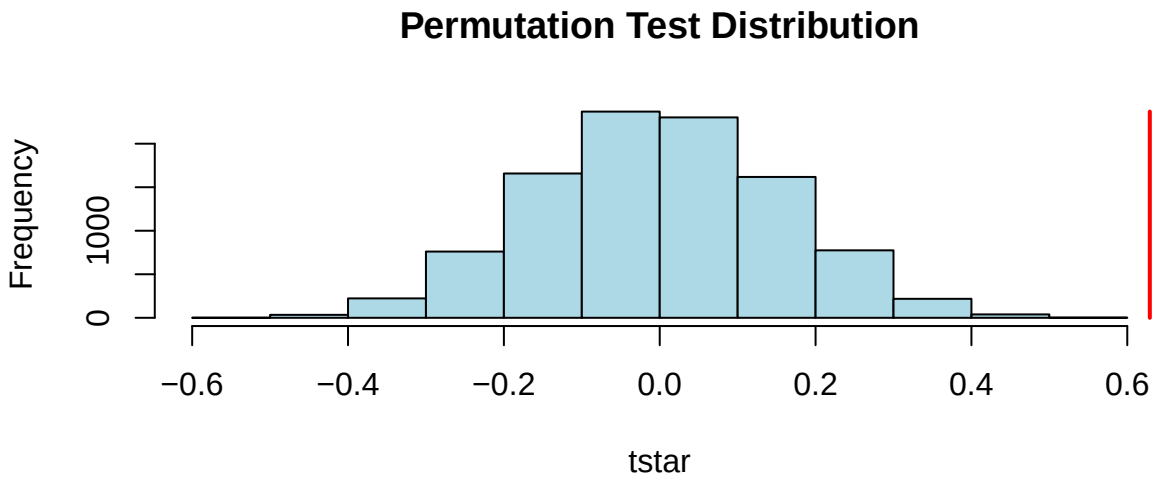
```
mystat=function(x,y) {mean(x-y)}
B=10000; tstar=numeric(B)
for(i in 1:B){
```

```
  datastar=t(apply(cbind(data$Before,data$After8weeks),1,sample))
  tstar[i]=mystat(datastar[,1],datastar[,2]) }
myt=mystat(data$Before,data$After8weeks)
print(myt)
```

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

```
pl <- sum(tstar < myt) / B
pr <- sum(tstar > myt) / B
p <- 2 * min(pl, pr)
print(p)
```

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




10000 permutations are run during which each pair (Before, After8weeks) is randomly shuffled and its mean difference is calculated. The original mean difference of the shuffled data is approximately 0.6289. A histogram representing the permuted test statistics is shown, where the red line indicates this original observed mean difference. pl counts the proportion of permuted test statistics that are less than the observed value (myt). pr counts the proportion of permuted test statistics that are greater than the observed value (myt). The permutation test p-value is 0. Since $p < 0.05$, we reject the null hypothesis that the data from before and after the diet are the same, meaning the difference between them is statistically significant.

**c**

A 97% Confidence Interval (CI) based on normality and a bootstrap 97%-CI for the mean data after 8 weeks of diet are constructed and compared to test the robustness of findings.

```
# 97% CI based on normality
mean_after <- mean(data$After8weeks)
sd_after <- sd(data$After8weeks)
n <- length(data$After8weeks)
error <- qt(0.985, df=n-1) * (sd_after/sqrt(n))
param_CI <- c(mean_after - error, mean_after + error)
print(param_CI)
```

```
## [1] 5.163850 6.393928
```

```
# Bootstrap
B <- 100000
Tstar <- numeric(B)
# Bootstrap resampling
for (i in 1:B) {
```

```
  Xstar <- sample(data$After8weeks, replace = TRUE)
  Tstar[i] <- mean(Xstar)
}
# Calculate the quantiles for the confidence interval
Tstar15 <- quantile(Tstar, 0.015)
Tstar985 <- quantile(Tstar, 0.985)
# Print the confidence interval
cat("Bootstrap 97% Confidence Interval:", Tstar15, "to", Tstar985, "\n")
```

```
## Bootstrap 97% Confidence Interval: 5.225 to 6.322222
```

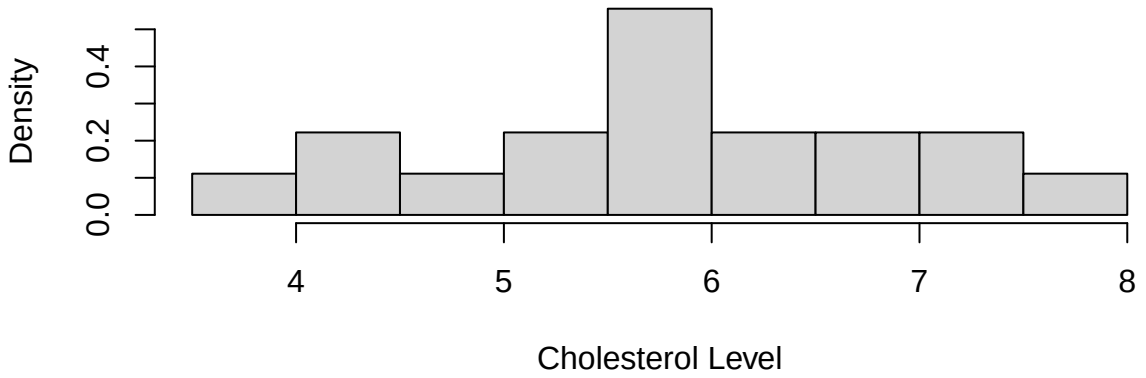
The parametric CI, which assumes normality, results in the interval [5.163, 6.393], while the bootstrap CI, which is non-parametric and does not require a normality assumption, returns the interval [5.225, 6.322]. Since both intervals do not include 0, the null hypothesis that the diet does not have a significant effect is rejected.

### d

A bootstrap test with test statistic T = max(X1, X2, ..., X18) is run to verify a hypothesis about the upper bound (theta) of the distribution of data$After8weeks. As given by the exercise description, the possible theta values, which represent the upper bounds for the cholesterol levels, range from 3 to 12.

```
X = data$After8weeks
n = length(X)
theta_values = seq(3, 12, by=0.1)
T_obs = max(X)
pvals <- numeric(length(theta_values))
bootstrap_test <- function(X, theta, B = 10000) {
  # Observed statistic
  T_obs <- max(X)
  n <- length(X)
  # Generate bootstrap replicates of T
  tstar <- numeric(B)
  for (b in 1:B) {
    xstar <- runif(n, min=3, max=theta)
    tstar[b] <- max(xstar)
  }
  pr <- mean(tstar > T_obs)
  pval <- 2 * min(pr, 1 - pr)
  return(pval)
}
for (i in seq_along(theta_values)) {
  pvals[i] = bootstrap_test(X, theta=theta_values[i], B = 10000)
}
not_rejected = theta_values[pvals >= 0.05]
not_rejected
```

```
##  [1] 7.7 7.8 7.9 8.0 8.1 8.2 8.3 8.4 8.5 8.6 8.7
```




The values ranging from 7.7 to 8.7, as printed above, are not rejected as possible true upper bounds of cholesterol levels after 8 weeks, while theta $< 7.7$ and theta $> 8.7$ are rejected. The histogram of the cholesterol level confirms that the upper bound value is approximately 8, which resides in the range [7.7, 8.7].

The Kolmogorov-Smirnov test is not considered as it assumes the data to be independent. Since in this case the data is dependent on each other, it would give misleading results.

**e**

To verify whether the median cholesterol level after 8 weeks of low fat diet is less than 6, and to check if the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25% binomial test are performed.

```
bin_test_1 <- binom.test(sum(data$After8weeks < 6), length(data$After8weeks), p = 0.5, alternative = "l
print(bin_test_1)
```

```
##
##  Exact binomial test
##
## data:  sum(data$After8weeks < 6) and length(data$After8weeks)
## 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
```

```
successes <- sum(data$After8weeks < 4.5)
n <- length(data$After8weeks)
bin_test_2 <- binom.test(successes, n, p = 0.25, alternative = "g")
print(bin_test_2)
```

```
##
##  Exact binomial test
##
## data:  successes and n
## number of successes = 3, number of trials = 18, p-value = 0.8647
## alternative hypothesis: true probability of success is greater than 0.25
## 95 percent confidence interval:
##  0.04702488 1.00000000
```

```
## sample estimates:
## probability of success
##              0.1666667
```

For the first binomial test the null hypothesis is that the probability of a cholesterol level being below 6 is at least 0.5. Since the p-value is 0.8811, which is greater than 0.05, we fail to reject the null hypothesis, meaning there is not satisfactory evidence that the median cholesterol level is below 6. For the second test the null hypothesis is that the proportion of cholesterol levels below 4.5 is $\leq 0.25$. Since the p-value is 0.8647, which is greater than 0.05, we fail to reject the null hypothesis, meaning there is not enough evidence supporting that more than 25% of cholesterol levels are below 4.5.