Assignment 1

PD, JO, and MD, group 70

17 febuary 2023

```
knitr::opts_chunk$set(echo = TRUE, fig.width = 10, fig.height = 4)
```

Exercise 1

```
birthweight = read.table(file="../datasets/birthweight.txt", header=FALSE)
birthweight = birthweight[2:189,]
birthweight = as.numeric(unlist(birthweight))
```

a) To check the data for normality we used a Shapiro-Wilk test. This test tests the null hypothesis that a sample from an unknown distribution is normal.

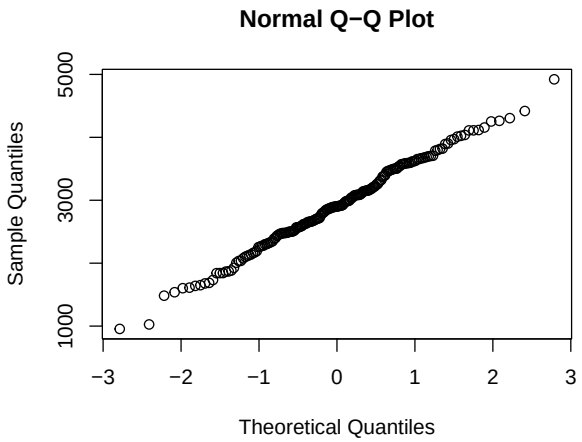
- Null hypothesis (H0): The mean weight is the same for male and female babies.
- Alternative hypothesis (Ha): The mean weight is different for male and female babies.

Our p-value of 0.90 is higher then the confidence level of 0.05 this means we cannot reject our null hypothesis. This hypothesis is complemented by the QQ-plot and histogram of the data that both approximate a normal distribution.

```
# check normality
shapiro.test(birthweight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  birthweight
## W = 0.99595, p-value = 0.8995
```

```
par(mfrow=c(1,2))
qqnorm(birthweight)
hist(birthweight)
```



Assuming normality we can construct a 96%-CI for the mean. We constructed the CI using a t-test with a confidence level of 0.96. The resulting confidence interval is [2808.08, 3018,50].

```
# 96%-CI
m = mean(birthweight)
t.test(birthweight, mu=m, conf.level = 0.96)
```

```
##
##  One Sample t-test
##
## data:  birthweight
## t = 0, df = 187, p-value = 1
## alternative hypothesis: true mean is not equal to 2913.293
## 96 percent confidence interval:
##  2808.084 3018.501
## sample estimates:
## mean of x
##  2913.293
```

The sample size needed to provide that the length of the CI is at most 100, is the sample size needed to get a CI of mu +/- 100/2. The sample size necessary is 550.

```
#sample size needed for CI length of 100
e = 100/mean(birthweight)
qnorm(0.98)^2*0.96*0.04/(e/2)^2
```

```
## [1] 549.8625
```

The bootstrap CI is very similar to the CI that we got with the t-test it is just a fraction smaller. Which could mean that the bootstrap interval is more robust than the t-test interval.

```
#bootstrap
B=1000
Tstar = numeric(B)
for(i in 1:B){
  Xstar = sample(birthweight, replace=TRUE)
```

```
  Tstar[i] = mean(Xstar)
}
Tstar2 = quantile(Tstar, 0.02)
Tstar98 = quantile(Tstar, 0.98)
sum(Tstar<Tstar2)
```

```
## [1] 20
```

```
c(2*m - Tstar98, 2*m - Tstar2)
```

```
##      98%       2%
## 2811.621 3013.394
```

b)

To verify the claim we do a t-test with H0: mu = 2800 and Ha: mu > 2800. The result of the t-test has a p-value of 0.014, which is smaller than alpha. This means we can reject H0, and Ha is accepted. The test also tells us that there is a probability of 95% that the mean is greater than 2819.20. For an appropriate sign-test we check the number of observations that are greater than 2800 and the total number of observations in the data set. With H0: mu = 2800 and Ha: mu > 2800. The result of the sign-test has a p-value of 0.033, so H0 is rejected. The results tell us that the number of observations that are greater than 2800 is greater than the number of total observations divided by two.

```
#t-test to verify mean weight is bigger than 2800
t.test(birthweight, mu=2800, alternative = "g")
```

```
##
##  One Sample t-test
##
## data:  birthweight
## t = 2.2271, df = 187, p-value = 0.01357
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829.202      Inf
## sample estimates:
## mean of x
##  2913.293
```

```
#sign test
binom.test(sum(birthweight>2800),
           length(birthweight), alternative = "g")
```

```
##
##  Exact binomial test
##
## data:  sum(birthweight > 2800) and length(birthweight)
## number of successes = 107, number of trials = 188, p-value = 0.03399
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5065781 1.0000000
## sample estimates:
## probability of success
##              0.5691489
```

c)

We can test the power of the test by simulation. By generating data and checking the fraction of the sample where H0 is rejected. The powers of the test tell us that the t-test performs better than the sign-test. This is because the sign-test is a non-parametric test, which means it makes very few assumption about the data but this can also lead to a lack of statistical power.

```
#geen idee of dit goed is
#power of t-test and sign test
B=10000; n=188; s = sd(birthweight); m = mean(birthweight)
psign = numeric(B)
pttest = numeric(B)

for(i in 1:B){
  x = rnorm(n, mean=m, sd=s)
  pttest[i] = t.test(x, mu=2800, alternative = "g")[[3]]
  psign[i] = binom.test(sum(x>2800), n, alternative ="g")[[3]]
}

sum(pttest<0.05)/B
```

```
## [1] 0.7156
```

```
sum(psign<0.05)/B
```

```
## [1] 0.5344
```

d)

The confidence interval for P(X<2600) is [0.25, 0.41], with a confidence level of 0.98.

```
n = length(birthweight)
lower_bound = 0.25
p_hat = sum(birthweight<2600)/n
margin = p_hat - lower_bound
upper_bound = lower_bound + 2*margin
z = margin / sqrt((p_hat*(1-p_hat))/n)
alpha = (1 - pnorm(z))*2
confidence_level = 1 - alpha
```

e)

To test the claim that there is a difference in the mean birth weight of male and female babies are different we can perform a prop test on the proportions of males and females that are born with a weight below 2600 gram.

- Null hypothesis (H0) : There is no difference in proportion between male and female babies born with a weight less than 2600.
- Alternative hypothesis (Ha) : There is a difference in proportion between male and female babies born with a weight less than 2600.

This test concludes with a p-value of ~0.5 and therefore we cannot reject the null hypothesis.

We can not directly reject the claim that there is a difference in the mean birth weight between male and female babies because we cannot directly measure this. However, we also have not found evidence for it in the given proportions.

```
male_weights <- c(34, 61)
female_weights <- c(28, 65)

x<-c(male_weights[1],female_weights[1])
y<-c(sum(male_weights),sum(female_weights))

prop.test(x, y)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of y
## X-squared = 0.45343, df = 1, p-value = 0.5007
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08792638  0.20156532
## sample estimates:
##    prop 1    prop 2
## 0.3578947 0.3010753
```

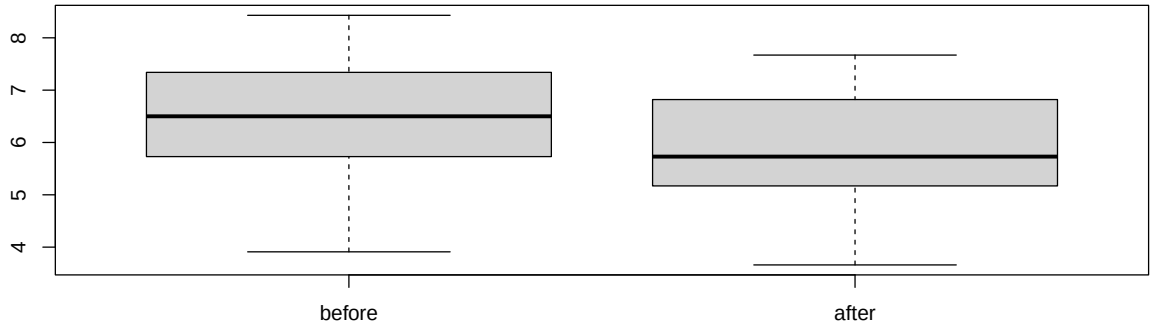
Exercise 2

```
fat = scan("../datasets/cholesterol.txt",
           what = list(before = 0, after = 0));
attach(fat);
```

a)

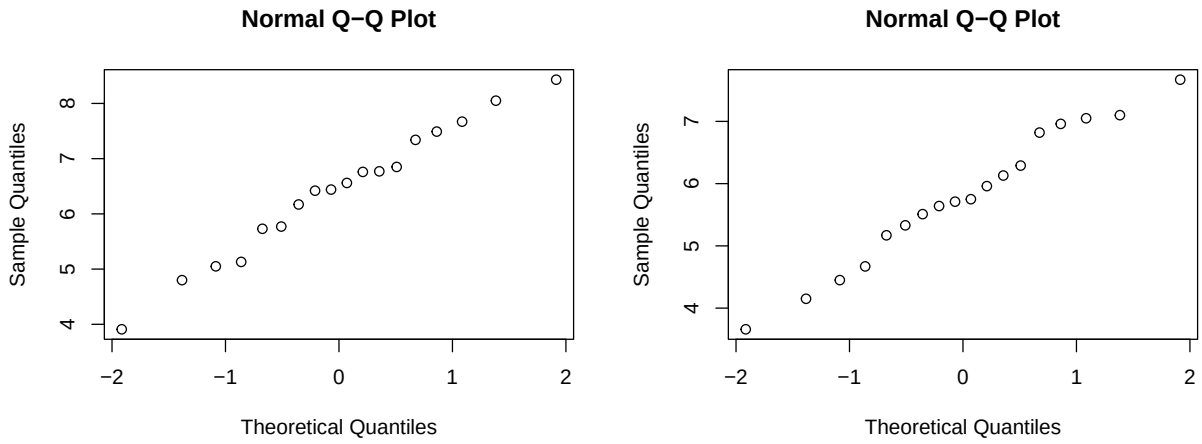
To investigate the data set we create a box plot of both columns. Judging from this we can observe a possible difference. The mean of the data from after 8 weeks appears to be lower.

```
boxplot(fat)
```



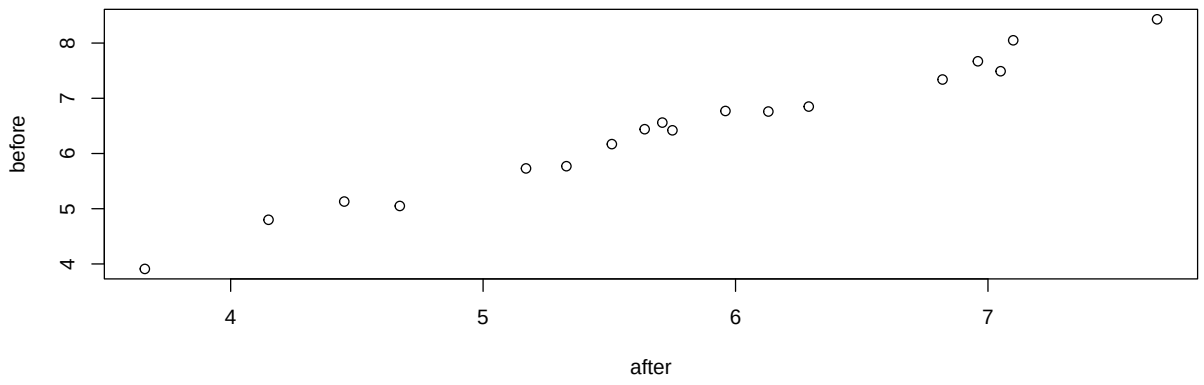
Next we plot a normal Q-Q plot to check if the data is normally distributed. this appears to be the case.

```
par(mfrow = c(1,2)); qqnorm(before); qqnorm(after)
```



To check if the before and after 8 weeks data is correlated we can plot the two data sets against each other. The plot shows a clear linear correlation between before and after. Then we can confirm this with both the Pearson's and Spearman's correlation test. Both of these give the conclusion that there indeed is a correlation and it is a strong correlation with **an r value of 0.99.**

```
plot(before~after)
```




```
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
```

```
cor.test(before, after, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  before and after
## S = 12, p-value = 9.753e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9876161
```

b)

Since the data is collected from the same patient before and after 8 weeks it is clear the data is paired. We can therefore perform a t-test to investigate whether there is a difference in the mean of the before and after data.

- Null hypothesis (H0) : mean(before) = mean(after)
- Alternative hypothesis (Ha) : mean(before) > mean(after)

Performing a t.test gives us a p-value of 1.639e-11 this is smaller than our significance level of 0.05 and we can therefore reject the null hypothesis. The before data has a mean that is greater than the mean of the after data.

```
t.test(before, after, alternative="greater", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 14.946, df = 17, p-value = 1.639e-11
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.5556906       Inf
## sample estimates:
## mean difference
##       0.6288889
```

We also perform a Wilcoxon signed rank test to check if the same conclusion can be drawn. The test gives us a p-value of 3.815e-06. with significance level of 0.05 we can reject the null hypothesis again.

```
wilcox.test(before, after, alternative = "greater", paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  before and after
## V = 171, p-value = 3.815e-06
## alternative hypothesis: true location shift is greater than 0
```

A permutation test is applicable for independent samples with any test statistic that expresses difference between the samples. This means that for the before and after data we can perform this test is the data is independent. The before and after data are independent therefore we might perform a permutation test.

c)

The estimator of theta we find by computing max(after). Now we can find the confidence interval for this estimator by the bootstrapped confidence interval method. This gives us a confidence interval of [7.67, 8.38]

```
B=1000
T1 = max(after)
Tstar=numeric(B)
c1 = after

for(i in 1:B) {
  Xstar=sample(c1,replace=TRUE)
  Tstar[i]=max(Xstar)
}
Tstar25=quantile(Tstar,0.025)
Tstar975=quantile(Tstar,0.975)

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

```
## 97.5%  2.5%
##  7.67  8.38
```

d)

To determine for which thetas we cannot reject that our estimate

- Null hypothesis (H0): p(after) = unif(3,theta) where theta in [3, 12]
- Alternative hypothesis (Ha): p(after) =/= unif(3,theta) where theta in [3, 12]

We performed the bootstrap test for the values for theta in 3 . . . 12. For the values inside the bootstrapped confidence interval we cannot reject the null hypothesis.

the Kolmogorov-Smirnoff (KS) test can be used to perform this analysis. This can be done by generating data from the uniform distribution and performing the KS test on the generated data and the after data.

```
data = after
n=length(data); t=max(data); t
```

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

```
B=1000; tstar=numeric(B)
for (i in 1:B) {
  xstar=runif(n, 3, 9)
  tstar[i]=max(xstar)
}

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

```
## [1] 0.036
```

e)

Using the sign test we are test if H0: median(after) = 6. with Ha: median(after) < 6. Performing the test gives us a p-value of 0.61111 this is not lower than our significance level of 0.05 and therefore we cannot reject the null hypothesis.

```
x = sum(after<6)
n = length(after<6)
binom.test(x, n, x/n, "l")
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 11, number of trials = 18, p-value = 0.5883
## alternative hypothesis: true probability of success is less than 0.6111111
## 95 percent confidence interval:
##  0.0000000 0.8010467
## sample estimates:
## probability of success
##              0.6111111
```

Next to the wilcox.

```
?wilcox.test(after,mu=6)
```

```
## starting httpd help server ... done
```

Exercise 3

```
diet = read.table("../datasets/diet.txt", header = TRUE)
```

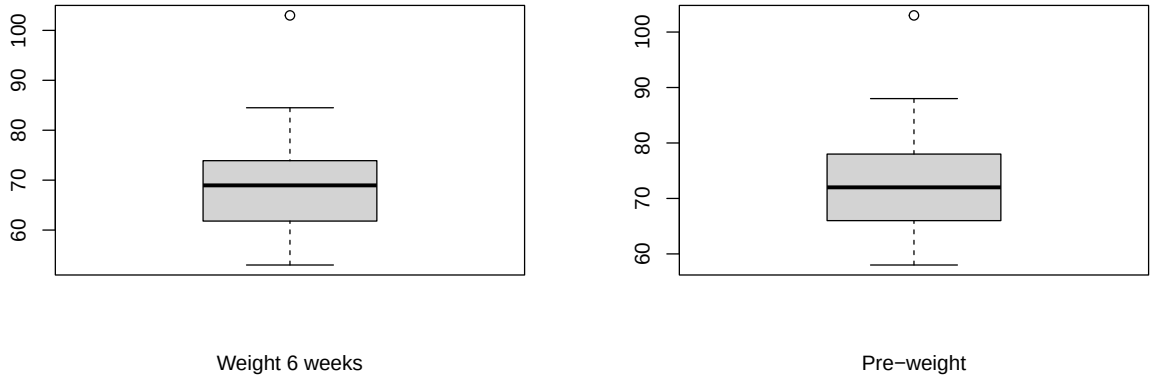
a)

We check the data by making boxplots for both pre-weight and weight after 6 weeks. From the plots we can tell that the data appears to be normally distributed, we can also tell that there is a difference between the pre-weight and the weight after 6 weeks. Lastly, we want to investigate if there is a significant difference between the mean weight before and after. We perform a T-test with significance level of 95% and hypotheses:

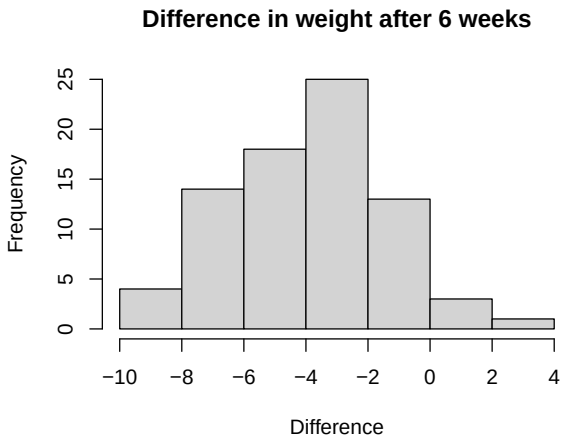
- Null hypothesis (H0): mu1 = mu2
- Alternative hypothesis (Ha): mu1 =/= mu2

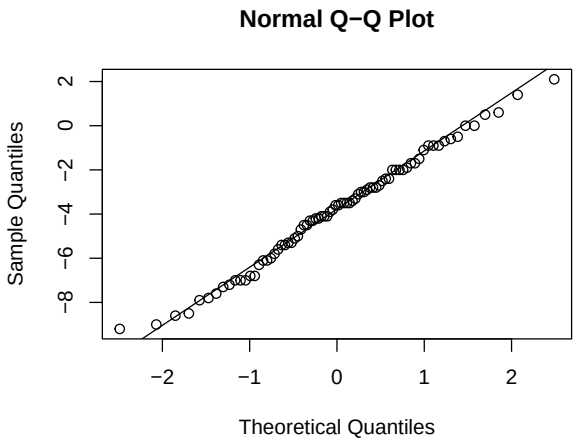
The test rejects H0, therefore we can assume that there is a difference in the mean weight before and after the diet.

```
par(mfrow=c(1,2))
boxplot(x = diet$weight6weeks,xlab= "Weight 6 weeks" )
boxplot(x = diet$preweight,  xlab= "Pre-weight")
```




```
#Check for if data is normaly distributed
hist(diet$weight6weeks - diet$preweight, main =
       "Difference in weight after 6 weeks", xlab =
       "Difference")
qqnorm(diet$weight6weeks - diet$preweight)
qqline(diet$weight6weeks - diet$preweight)
```






```
t.test(diet$preweight, diet$weight6weeks, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  diet$preweight and diet$weight6weeks
## t = 13.309, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  3.269602 4.420141
## sample estimates:
## mean difference
##        3.844872
```

b)

Our goal is to use one-way ANOVA test to compare the diets. We first draw QQ-plot to check if we can assume normality this seems to be the case so we perform ANOVA with the Hypotheses:

- Null hypothesis (H0): The means of the lost weight are equal among the three types of diets.
- Alternative hypothesis (Ha): The mean of at least one diet-type is not equal among the three types of diets.

The ANOVA

Looking at the summary of the anova table we can see clearly that the p-value of diet 3 is smaller than our significance level of 0.05. Therefor we reject the null hypothesis and assume that there is a difference between the diets. All diets have and effect this is shown in question a. Diet 3 has the biggest effect on the weight lost.

The Kruskal-Wallis is also valid because it has the same pre-conditions as the anova test the difference is that Kruskal-Wallis test can also be used when normality cannot be assumed.

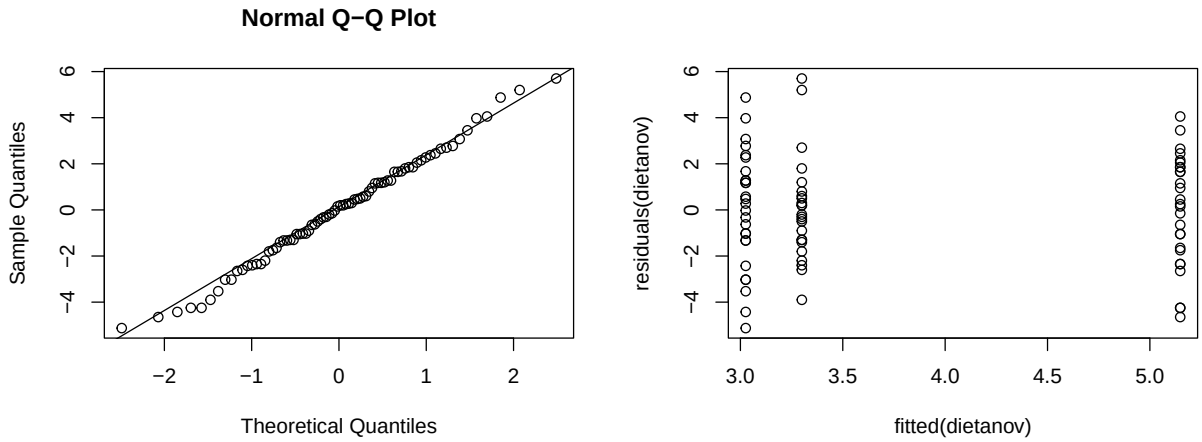
```
dietframe <- data.frame(weight=(diet$preweight-diet$weight6weeks),
                        diet=factor(diet$diet))
dietanov=lm(weight~diet ,data = dietframe)
anova(dietanov)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value   Pr(>F)
## diet       2  71.09  35.547  6.1974 0.003229 **
## Residuals 75 430.18   5.736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dietanov)
```

```
##
## Call:
## lm(formula = weight ~ diet, data = dietframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1259 -1.3815  0.1759  1.6519  5.7000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3000     0.4889   6.750 2.72e-09 ***
## diet2        -0.2741     0.6719  -0.408  0.68449
## diet3         1.8481     0.6719   2.751  0.00745 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.395 on 75 degrees of freedom
## Multiple R-squared:  0.1418, Adjusted R-squared:  0.1189
## F-statistic: 6.197 on 2 and 75 DF,  p-value: 0.003229
```

```
par(mfrow=c(1,2)); qqnorm(residuals(dietanov))
qqline(residuals(dietanov))
plot(fitted(dietanov),residuals(dietanov))
```



c)

Our goal is to use two-way ANOVA test to investigate the effects on gender and different diets on the mean lost weight. We first draw QQ-plot to check if we can assume normality this seems to be the case so we can go over to our hypothesis.

- Null hypothesis (H0): The means of the lost weight are equal among the factors.
- Alternative hypothesis (Ha): The means of the lost weight are not equal for at least one factor.

The result of the ANOVA test is that we can reject the null hypothesis. There is a difference between the means of the lost weight. There also is a small significant interaction between diet and gender with a p-value of ~0.048. Looking at the summary we can see that this is not a specific interaction between a specific diet type and gender but a general effect that gender has on all diet types.

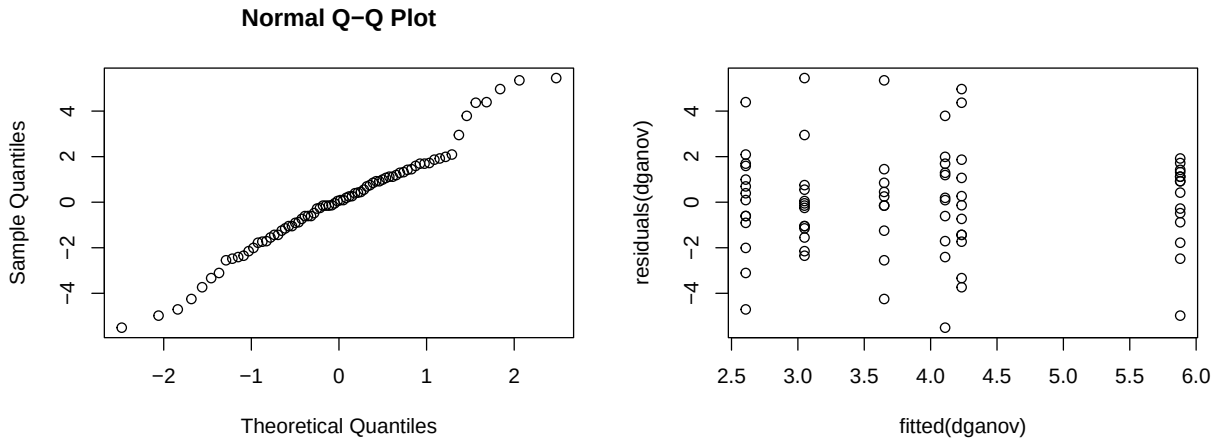
```
dgframe <- data.frame(weight=diet$preweight - diet$weight6weeks,
                      diet=factor(diet$diet), gender=factor(diet$gender))
dganov=lm(weight~diet*gender ,data = dgframe)
anova(dganov)
```

```
## Analysis of Variance Table
##
## Response: weight
##             Df Sum Sq Mean Sq F value   Pr(>F)
## diet         2  60.53 30.2635  5.6292 0.005408 **
## gender       1   0.17  0.1687  0.0314 0.859910
## diet:gender  2  33.90 16.9520  3.1532 0.048842 *
## Residuals   70 376.33  5.3761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dganov)
```

```
##
## Call:
## lm(formula = weight ~ diet * gender, data = dgframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5091 -1.2958  0.0705  1.2159  5.4500
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.0500     0.6197   4.922 5.49e-06 ***
## diet2          -0.4429     0.8764  -0.505   0.6149
## diet3           2.8300     0.8616   3.284   0.0016 **
## gender1         0.6000     0.9600   0.625   0.5340
## diet2:gender1   0.9019     1.3395   0.673   0.5030
## diet3:gender1  -2.2467     1.3145  -1.709   0.0919 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.319 on 70 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2009, Adjusted R-squared:  0.1438
## F-statistic: 3.519 on 5 and 70 DF,  p-value: 0.006775
```

```
par(mfrow=c(1,2))
qqnorm(residuals(dganov))
plot(fitted(dganov), residuals(dganov))
```



e)

Whether we want to use one-way or two-way ANOVA depends on what we want to investigate. If we want to see the effects of different diets on weight loss one-way ANOVA would be sufficient. However if we are interested in the effect of gender on the effectiveness of diet this one-way model could not work. Therefor we would have to use two-way ANOVA. Because we only want to predict the lost weight achieved by diet-type we are going to use the same linear model we used for one-way ANOVA to predict weight loss for the different diet-types.

```
summary(dietanov)
```

```
##
## Call:
## lm(formula = weight ~ diet, data = dietframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1259 -1.3815  0.1759  1.6519  5.7000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3000     0.4889   6.750 2.72e-09 ***
## diet2        -0.2741     0.6719  -0.408  0.68449
## diet3         1.8481     0.6719   2.751  0.00745 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.395 on 75 degrees of freedom
## Multiple R-squared:  0.1418, Adjusted R-squared:  0.1189
## F-statistic: 6.197 on 2 and 75 DF,  p-value: 0.003229
```

Looking at the summary for the ANOVA model we can see that the estimate average weight difference between the before and after. This would mean that the prediction for the weight loss after 6 weeks will be:

Col1	Col2
diet 1	3.3 kg
diet 2	3.1 kg
diet 3	5.1 kg

Exercise 4

a)

We loop over all the blocks and sample 2 plots for each additive.

```
#randomized plot design
plot=c(0,1,2,3)
plots = data.frame(matrix(0, ncol = 4, nrow = 24))
for(i in 1:24){
  if(i %% 4 != 1){next}

  x<-as.integer(i/4)+1
  plots[i, 4]= x
  plots[i+1, 4]= x
  plots[i+2, 4]= x
  plots[i+3, 4]= x

  for(j in 1:3){
    sample_plots = sample(plot, 2)
    plots[i+sample_plots[1],j] = 1
    plots[i+sample_plots[2],j] = 1
  }
}
colnames(plots) = c("N","P", "K", "Blocks")
plots
```

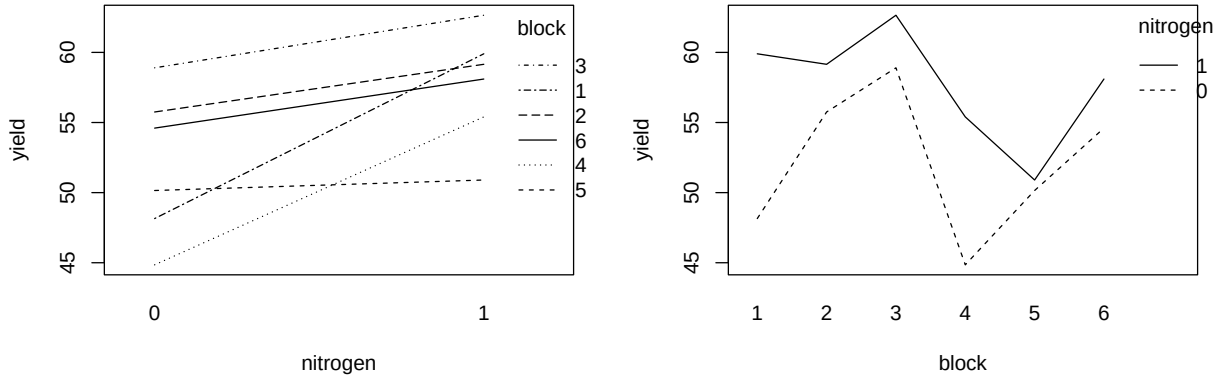
```
##    N P K Blocks
## 1  1 1 0      1
## 2  1 0 1      1
## 3  0 1 1      1
## 4  0 0 0      1
## 5  0 0 1      2
## 6  0 1 1      2
## 7  1 0 0      2
## 8  1 1 0      2
## 9  1 0 0      3
## 10 0 1 1      3
## 11 1 1 1      3
## 12 0 0 0      3
## 13 1 1 0      4
## 14 0 1 0      4
## 15 1 0 1      4
## 16 0 0 1      4
## 17 0 1 1      5
## 18 0 0 1      5
## 19 1 1 0      5
```

```
## 20 1 0 0       5
## 21 1 0 0       6
## 22 0 1 1       6
## 23 0 0 0       6
## 24 1 1 1       6
```

b)

By looking at every block you get a more precise picture of the different combinations. Because different blocks have different combinations of additives applied. This way you can determine the effect the other additives have on the yield.

```
par(mfrow=c(1,2))
interaction.plot(npk$N, npk$block, npk$yield,
                 xlab = "nitrogen", ylab = "yield", trace.label = "block")
interaction.plot(npk$block, npk$N, npk$yield,
                 xlab = "block", ylab = "yield", trace.label = "nitrogen")
```



c)

- Null hypothesis: The means of the yield are equal among the factors.
- Alternative hypothesis: The means of the yield are not equal for at least one factor.

The p-value for block is not significant therefore we cannot reject H0. The p-value for nitrogen is significant however and that means we can reject H0 and accept Ha. It is sensible to include the block factor in the model because the yield is source of variation by including it we can account for this variation and better estimate the effect nitrogen.

The Friedman test can't be used because the Friedman test is non-parametric test for comparing two groups on a dependent variable. However here we have two groups and a response variable.

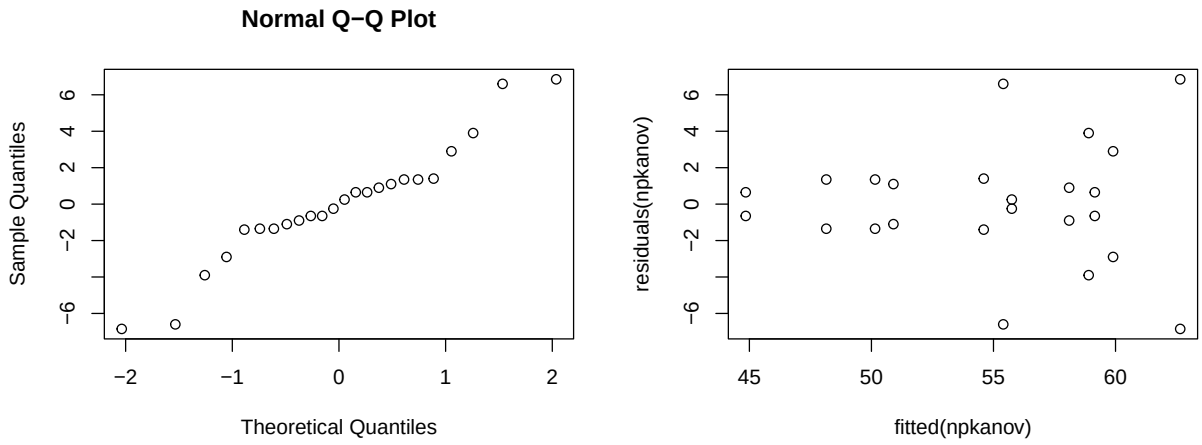
```
npk$block = as.factor(npk$block)
npk$N = as.factor(npk$N)

npkanov = lm(yield~block*N, data=npk)
anova(npkanov)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## block      5 343.29  68.659  3.3592 0.03967 *
## N          1 189.28 189.282  9.2607 0.01021 *
## block:N    5  98.52  19.704  0.9640 0.47690
## Residuals 12 245.27  20.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(npkanov)
```

```
##
## Call:
## lm(formula = yield ~ block * N, data = npk)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.85  -1.35   0.00   1.35   6.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   48.150      3.197  15.062 3.71e-09 ***
## block2         7.600      4.521   1.681   0.1186
## block3        10.750      4.521   2.378   0.0349 *
## block4        -3.300      4.521  -0.730   0.4794
## block5         2.000      4.521   0.442   0.6661
## block6         6.450      4.521   1.427   0.1792
## N1            11.750      4.521   2.599   0.0233 *
## block2:N1     -8.350      6.394  -1.306   0.2160
## block3:N1     -8.000      6.394  -1.251   0.2347
## block4:N1     -1.200      6.394  -0.188   0.8543
## block5:N1    -11.000      6.394  -1.720   0.1110
## block6:N1     -8.250      6.394  -1.290   0.2212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.521 on 12 degrees of freedom
## Multiple R-squared:  0.7201, Adjusted R-squared:  0.4636
## F-statistic: 2.807 on 11 and 12 DF,  p-value: 0.04492
```

```
par(mfrow=c(1,2))
qqnorm(residuals(npkanov))
plot(fitted(npkanov), residuals(npkanov))
```



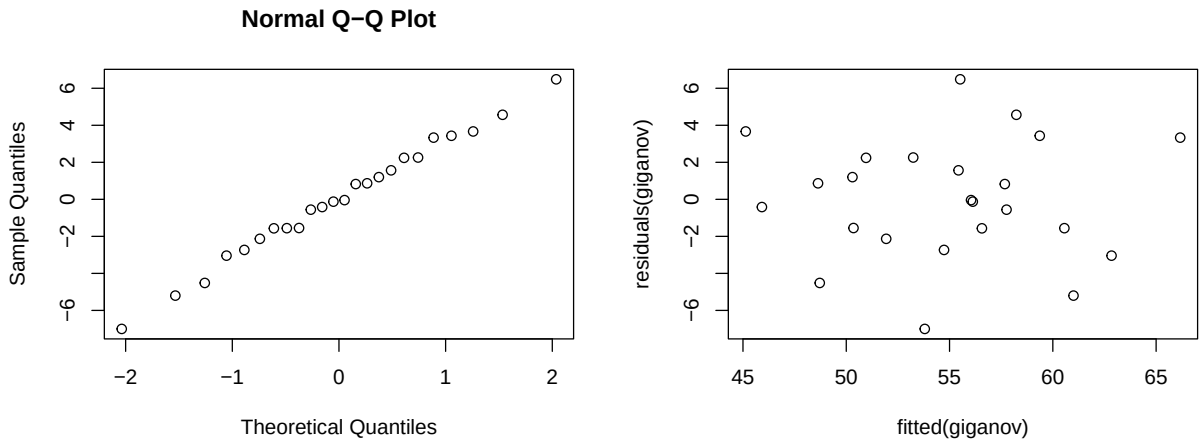
d)

We did a investigation on the significance of the variables. And removed each iteration the least significant value from the model. At the end we only had Nitrogen left and because it the only factor with significance relation we have chosen it as our favorite.

```
npk$block = as.factor(npk$block)
npk$N = as.factor(npk$N)
npk$P = as.factor(npk$P)
npk$K = as.factor(npk$K)

giganov = lm(yield~block+N+P+K, data = npk)
anova(giganov)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## block      5 343.29  68.659  4.2879 0.01272 *
## N          1 189.28 189.282 11.8210 0.00366 **
## P          1   8.40   8.402  0.5247 0.47999
## K          1  95.20  95.202  5.9455 0.02767 *
## Residuals 15 240.18  16.012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
qqnorm(residuals(giganov))
plot(fitted(giganov), residuals(giganov))
```




```
model1 = lm(yield~block+N+K, data = npk)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  4.4192 0.010172 *
## N          1 189.28 189.282 12.1829 0.003024 **
## K          1  95.20  95.202  6.1275 0.024874 *
## Residuals 16 248.59  15.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 = lm(yield~block+N, data = npk)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  3.3951 0.026173 *
## N          1 189.28 189.282  9.3598 0.007095 **
## Residuals 17 343.79  20.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e)

The output of the mixed effect analysis will generally return a smaller standard error than that of the fixed effects model because the mixed effects model also accounts for variation that is caused by the difference of the blocks. Furthermore we can use the random effect of the block to estimate the variance in yield.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```r
npklmer=lmer(yield ~ N + (1|block), data = npk)
summary(npklmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: yield ~ N + (1 | block)
##    Data: npk
##
## REML criterion at convergence: 139.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.34436 -0.61483 -0.03131  0.49213  1.70506
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  block    (Intercept) 12.11    3.480
##  Residual             20.22    4.497
## Number of obs: 24, groups:  block, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   52.067      1.924  27.056
## N1             5.617      1.836   3.059
##
## Correlation of Fixed Effects:
##    (Intr)
## N1 -0.477
```