Assignment 1

Alexia Salomons, Nathan Maxwell Jones, Yauheniya Makarevich, group 71

27 February 2023

Exercise 1.

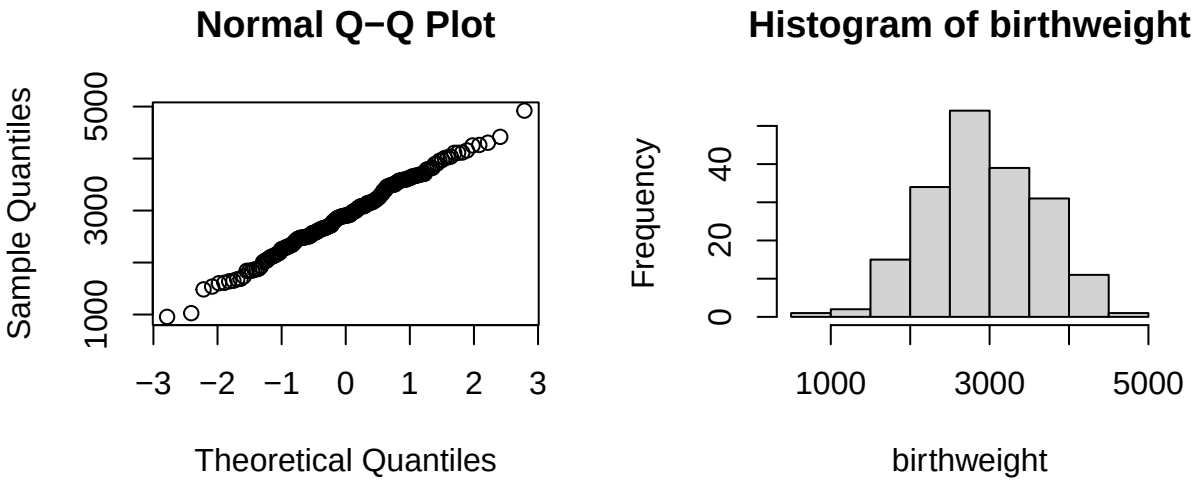
Sample mean:

```
birthweight_mean <- mean(birthweight)
birthweight_mean
```

```
## [1] 2913
```

a)

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



Given the QQ-plot and the histogram above, the data appears to be normal.

To calculate the CI-96% in R:

```
t.test(birthweight, conf.level = 0.96)
```

```
##
##  One Sample t-test
##
## data:  birthweight
## t = 57, df = 187, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
```

```
## 96 percent confidence interval:
##  2808 3019
## sample estimates:
## mean of x
##      2913
```

With a 96% CI, we get [2808-3019]. In order to decrease this range to 100, we can reverse the calculations in order to check the required sample size. CI is given by $\left[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$. Thus we can conclude that $t_{\alpha/2}\frac{s}{\sqrt{n}} \leq 50$. From the data we can determine that $s = 698$ and $t_{0.04/2} = 2.07$ (for 96% CI), and thus solve for n as follows:

```
s = sd(birthweight) # 698
t_02 = qt(1-0.02,df=188-1) # 2.07
(n = (t_02 * s / 50)^2)
```

```
## [1] 832
```

Therefore, a sample size of approximately 832 babies would be needed in order to have a 96% CI with a range of 100.

```
B <- 1000
alpha <- 0.04
T_star <- numeric(B)

for(i in 1:B) {
  X_star <- sample(birthweight, replace = TRUE)
  T_star[i] <- mean(X_star)
}

T_star_q2 <- quantile(T_star, alpha/2)
T_star_q98 <- quantile(T_star, 1 - alpha/2)

c(2*birthweight_mean - T_star_q98, 2*birthweight_mean - T_star_q2)
```

```
##  98%   2%
## 2814 3015
```

The CI in this case is very similar to the previous one, only changing the lower bound by 1.

b)

$$H_0 = \mu =< 2800$$

$$H_1 = \mu > 2800$$

```
t.test(birthweight, alternative = "greater", mu=2800)
```

```
##
##  One Sample t-test
##
## data:  birthweight
```

```
## t = 2, df = 187, p-value = 0.01
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829  Inf
## sample estimates:
## mean of x
##      2913
```

As p is smaller than 0.05, we reject the null hypothesis, supporting the claim made by the expert. The CI is infinite on the right side, since the test is one-sided.

For a sign test, we can use:

```
greater_weight <- as.integer(birthweight > 2800)
binom.test(sum(greater_weight), length(greater_weight), p=0.5, alt="g")
```

```
##
##  Exact binomial test
##
## data:  sum(greater_weight) and length(greater_weight)
## number of successes = 107, number of trials = 188, p-value = 0.03
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.507 1.000
## sample estimates:
## probability of success
##                  0.569
```

We again reject H_0, accepting H_1, that mean of the sample is bigger than 2800.

Both test confirmed the hypothesis that the mean of the sample is bigger than 2800.

c) To compute the powers of both tests, we can assume that the data we have was sampled from a normal distribution with the same variance and $\mu = 2900$. We can then compute the probability of each test correctly rejecting $H_0 : \mu \leq 2800$ as follows. To better approximate the current situation, we will choose a sample size of $n = 188$ newborn babies.

```
B=1000; n=188 ; mu = 2900; stdev = sd(birthweight);
psign=numeric(B)
pttest=numeric(B)
for(i in 1:B) {
  x=rnorm(n, mean=mu, sd=stdev)
  pttest[i]=t.test(x, alternative = "g", mu=2800)[[3]]
  psign[i]=binom.test(sum(x>2800), n, p=0.5)[[3]]
}
```

T-test power:

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

```
## [1] 0.624
```

Sign test power:

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

```
## [1] 0.349
```

We see that the power of the t-test is higher than that of the sign test. This makes sense because the sign test discards valuable information when considering only the signs, while the t-test is designed to test normal distributions, which is the case here.

d)

e)

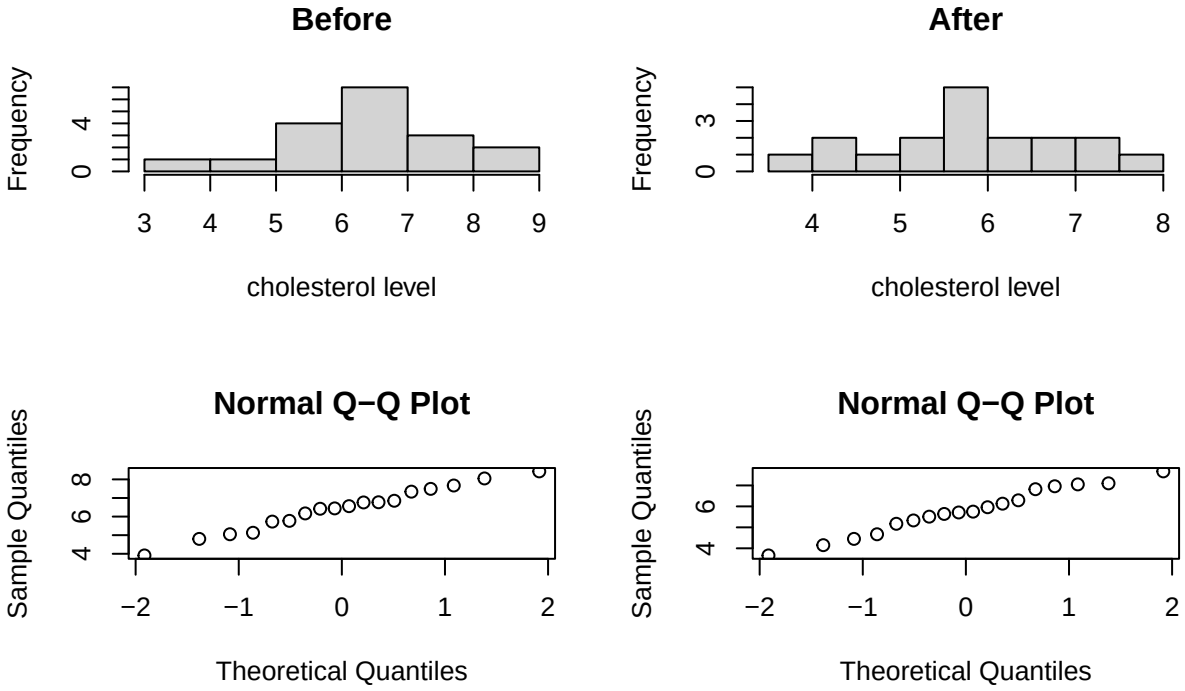
$$H_0 = P_{male} - P_{female} = 0$$
$$H_1 = P_{male} - P_{female} \neq 0$$

```r
prop.test(c(61, 65), c(95, 93))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(61, 65) out of c(95, 93)
## X-squared = 0.5, df = 1, p-value = 0.5
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2016  0.0879
## sample estimates:
## prop 1 prop 2
##  0.642  0.699
```

As the p-value is bigger than 0.05, we fail to reject the null-hypothesis, meaning that there is no true difference between the mean weight of males and females.

Exercise 2



a)

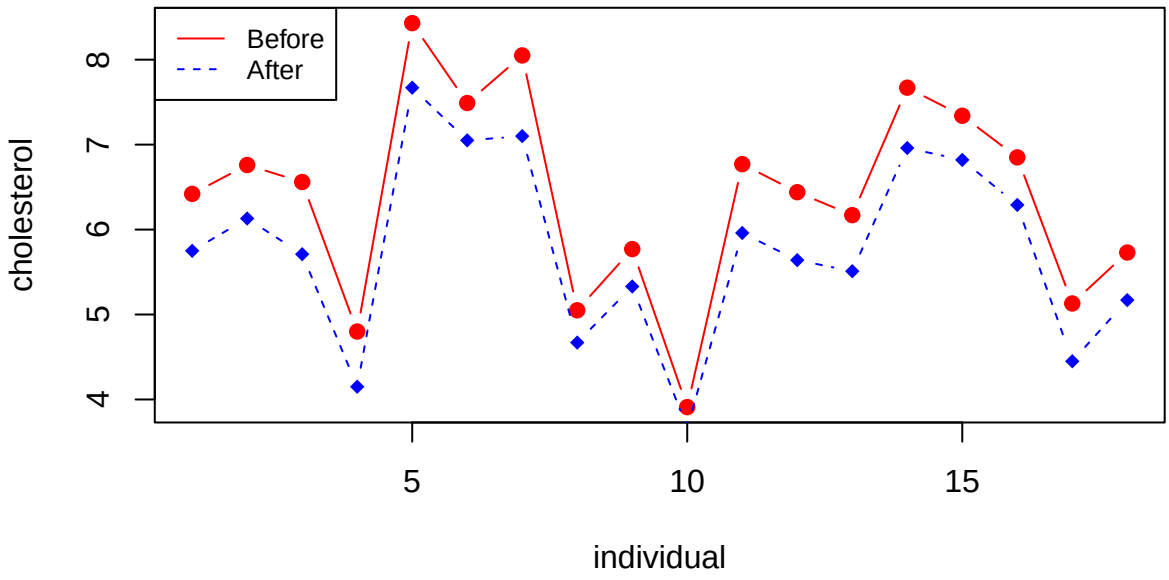
```
shapiro.test(df[, 1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df[, 1]
## W = 1, p-value = 1
```

```
shapiro.test(df[, 2])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df[, 2]
## W = 1, p-value = 0.9
```

As we can see from the plots and Shapiro test, the data is distributed normally.



We draw the data as lines to see if there any consistences in data such as data incompleteness, outliers or meaningless data points.

```
cor.test(df[, 1], df[, 2], method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## t = 29, df = 16, p-value = 2e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.975 0.997
## sample estimates:
##   cor
## 0.991
```

From the result of Pearson's test we can conclude that data before and data after are strongly correlated.

b) Data is paired since it is two different measurements of the same person and two samples are correlated.

In order to investigate has the diet affected individuals, we are going to apply tests (Paired t-test and Permutation test) that check difference between mean of different samples. Permutation test is applicable because it doesn't take normality into account and we are only testing for a difference between means, not how they relate to each other.

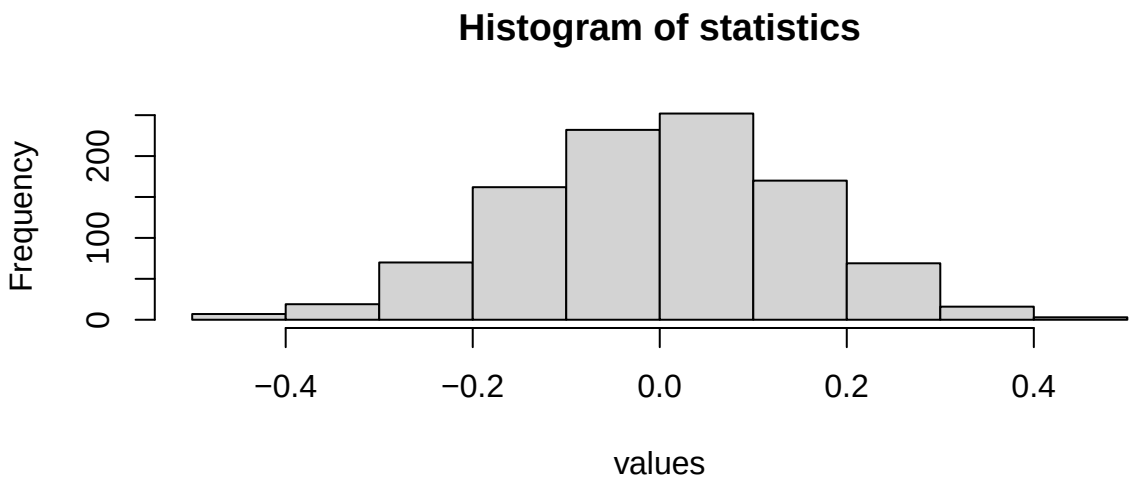
```
t.test(df[, 1], df[, 2], paired=2)
```

```
##
##  Paired t-test
##
## data:  df[, 1] and df[, 2]
```

```
## t = 15, df = 17, p-value = 3e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.540 0.718
## sample estimates:
## mean difference
##           0.629
```

Based on p-value obtained for paired T-test we reject null-hypothesis that samples have the same mean. Hence, there is a difference between these two samples.

```
diff_mean <- function(x, y) {
  return(mean(x-y))
}

stats <- diff_mean(df[, 1], df[, 2])

B <- 1000
t_star <- numeric(B)

for (i in 1:B) {
  diff_star <- t(apply(cbind(df[, 1], df[, 2]), 1, sample))
  t_star[i] <- diff_mean(diff_star[, 1], diff_star[, 2])
}

hist(t_star, main="Histogram of statistics", xlab="values")
lines(rep(stats, 2), c(0, 50), col="red", lwd=2)
```




```
# calculating p-value
pl <- sum(t_star < stats) / B
pr <- sum(t_star > stats) / B

p <- 2*min(pl, pr)
print(paste("P-value =", p))
```

```
## [1] "P-value = 0"
```

We rejecting the null hypothesis that there is no difference between two samples.

c) We have sample $x_i \in R[3, \theta]$, $i = \overline{1, 18}$, $\theta > 3$. Analytic mean for this distribution will be $\mu = \frac{3+\theta}{2}$.

We can estimate the mean and the standard distribution using our sample.

```
## [1] 5.78
```

```
## [1] 1.1
```

From this value we can obtain the estimation for the parameter θ. Knowing $\hat{\mu}$ $\hat{\theta} = 2 * \hat{\mu} - 3$.

```
## [1] "theta_hat = 8.55777777777778"
```

Then, using Central Limit Theorem, we can obtain confidence interval for parameter θ:

$$[t_{-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} + \frac{\theta + 3}{2}, t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} + \frac{\theta + 3}{2}]$$

```
alpha <- 0.05

n <- length(df[, 2])
t_alpha <- qt(1 - alpha/2, df=n)
theta_l <- theta_hat - t_alpha*s/sqrt(n)
theta_r <- theta_hat + t_alpha*s/sqrt(n)

c(theta_l, theta_r)
```

```
## [1] 8.01 9.10
```

We can improve the CI by having more individuals in the samples, as increasing in parameter n causes decreasing in standard deviation for sample of means. We can improve since we know the distribution and we know the estimation for theta.

d)

```
t <- max(df[, 2]); n <- length(df[, 2])

for (theta in 3:12) {
  B <- 1000
  t_star <- numeric(B)

  for (i in 1:B) {
    x_star <- runif(n, min = 3, max = theta)
    t_star[i] <- max(x_star)
  }

  pl <- sum(t_star < t)/B
  pr <- sum(t_star > t)/B

  p <- 2* min(pl, pr)
  print(paste("Theta =", theta, ", ", p))
}
```

```
## [1] "Theta = 3 ,  0"
## [1] "Theta = 4 ,  0"
## [1] "Theta = 5 ,  0"
## [1] "Theta = 6 ,  0"
## [1] "Theta = 7 ,  0"
## [1] "Theta = 8 ,  0.668"
## [1] "Theta = 9 ,  0.026"
## [1] "Theta = 10 ,  0"
## [1] "Theta = 11 ,  0"
## [1] "Theta = 12 ,  0"
```

Using bootstrap we have values of $\theta = 8$ and $\theta = 9$ for which our null hypothesis is not rejected (p-value > 0.05). We can apply Kolmogorov-Smirnov test to examine the sample since it examine if two samples were drawn from the one distribution. We can generate sample from uniform distribution and apply test on them.

```
ks.test(df[, 2], runif(100000, min = 3, max = 8))
```

```
## Warning in ks.test.default(df[, 2], runif(1e+05, min = 3, max = 8)): p-value
## will be approximate in the presence of ties
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  df[, 2] and runif(1e+05, min = 3, max = 8)
## D = 0.2, p-value = 0.4
## alternative hypothesis: two-sided
```

P-value $= 0.4$, so we accept null hypothesis and can confirm that the sample was drown from $U(3, \theta)$.

e) We are using sign test to test if the median of the cholesterol level is ł6.

```
less_chol <- as.integer(df[, 2] < 6)
binom.test(sum(less_chol), length(less_chol), p=0.5, alt="g")
```

```
##
##  Exact binomial test
##
## data:  sum(less_chol) and length(less_chol)
## number of successes = 11, number of trials = 18, p-value = 0.2
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.392 1.000
## sample estimates:
## probability of success
##                  0.611
```

P-value for the test equals 0.2 which rejects alternative hypothesis.

proportion tests

```
less_chol <- as.integer(df[, 2] < 4.5)
binom.test(sum(less_chol),length(less_chol),p=0.25, alt="l")
```

```
##
##  Exact binomial test
##
## data:  sum(less_chol) and length(less_chol)
## number of successes = 3, number of trials = 18, p-value = 0.3
## alternative hypothesis: true probability of success is less than 0.25
## 95 percent confidence interval:
##  0.000 0.377
## sample estimates:
## probability of success
##                  0.167
```

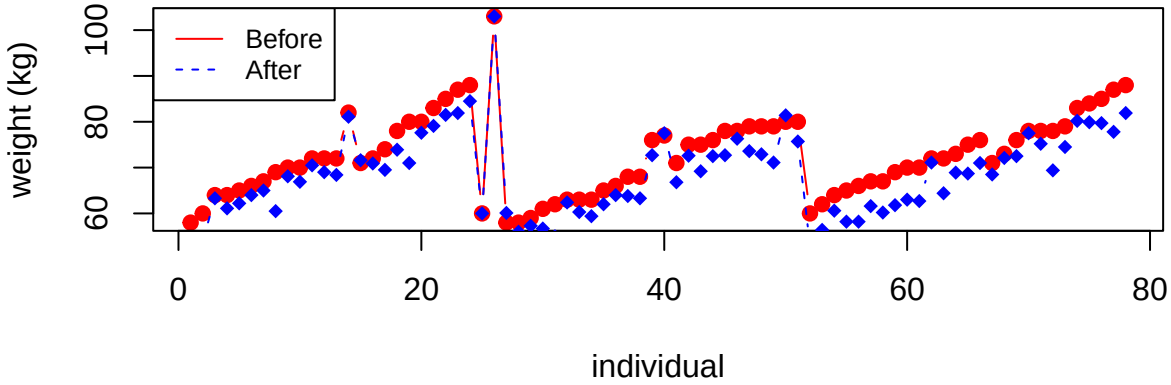
Exercise 3

We can compute and add the variable *weight.lost* as follows:

```
df <- as.data.frame(read.table("data/diet.txt", header=TRUE))
df["weight.lost"] <- df["preweight"] - df["weight6weeks"]
head(df)
```

```
##   person gender age height preweight diet weight6weeks weight.lost
## 1      1      0  22    159        58    1         54.2         3.8
## 2      2      0  46    192        60    1         54.0         6.0
## 3      3      0  55    170        64    1         63.3         0.7
## 4      4      0  33    171        64    1         61.1         2.9
## 5      5      0  50    170        65    1         62.2         2.8
## 6      6      0  50    201        66    1         64.0         2.0
```

a) To visualize the effect of diet on weight lost, we can plot the weights before and after the diets for every individual.



We see the trend that the weights after 6 weeks of diet are lower for nearly all individuals. To test whether this trend is significant, we can apply a paired t-test as follows:

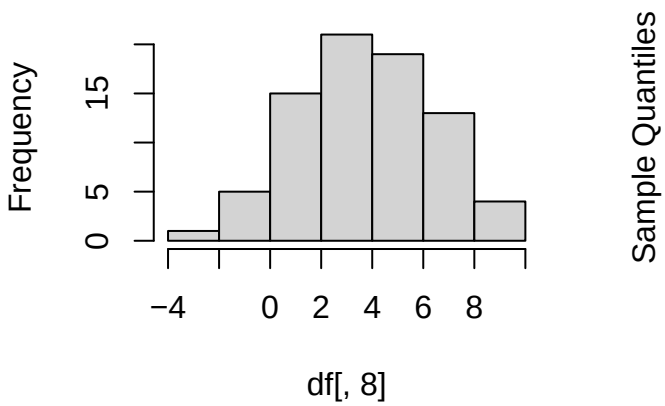
```
t.test(df[,5], df[,7], paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  df[, 5] and df[, 7]
```

```
## t = 13, df = 77, p-value <2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  3.27 4.42
## sample estimates:
## mean difference
##            3.84
```

A p-value < 2.2e-16 indicates that diet does indeed have a significant effect on weight loss, with an estimated average loss of 3.845kg over the 6 week period.

To check the assumptions of the test, we need to verify that the difference between *preweight* and *weight6weeks* (ie. *weight.lost*) follows a normal distribution.




```
shapiro.test(df[,8])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df[, 8]
## W = 1, p-value = 0.8
```

Looking at the shape of the Q-Q plot and histogram as well as the result of the result of the Shapiro-Wilk normality test, we can conclude that *weight.lost* follows a normal distribution. Thus the test assumptions are valid.

b) Format data

```
df$diet <- as.factor(df$diet)
dietaov=lm(weight.lost~diet,data=df)
anova(dietaov)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## diet       2     71    35.5     6.2 0.0032 **
## Residuals 75    430     5.7
## ---
```

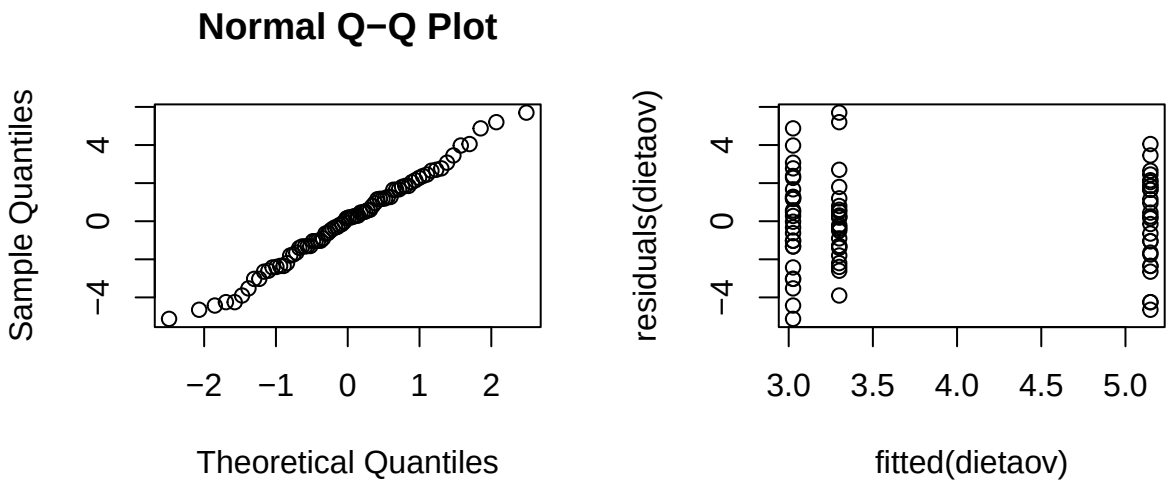
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dietaov)
```

```
##
## Call:
## lm(formula = weight.lost ~ diet, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.126 -1.381  0.176  1.652  5.700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.300      0.489    6.75  2.7e-09 ***
## diet2         -0.274      0.672   -0.41   0.6845
## diet3          1.848      0.672    2.75   0.0075 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.39 on 75 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.119
## F-statistic:  6.2 on 2 and 75 DF,  p-value: 0.00323
```

From the ANOVA table, we can see that p = 0.0032, meaning that the effect of diet on weight loss is significant. From the summary table, we can see that diet 2 is worse than diet 1, however, this difference is not significant. Furthermore, we can see that diet 3 is better than diet 1 and that this difference is significant with a p=0.0075. Therefore, diet 3 is the best diet.

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



The data seems to be relatively normal, therefore, it is appropriate to use ANOVA. Generally, the Kruskal-Wallis test is used when the data does not meet the assumptions for ANOVA, even though the data meets the assumptions here, the Kruskal-Wallis can still be used:

```
kruskal.test(df$weight.lost, df$diet)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$weight.lost and df$diet
## Kruskal-Wallis chi-squared = 10, df = 2, p-value = 0.005
```

This supports the ANOVA result as the p-value is smaller than 0.05.

c)

```
df$gender <- as.factor(df$gender)
dietgenderaov <- lm(weight.lost~gender*diet,data=df)
anova(dietgenderaov)
```

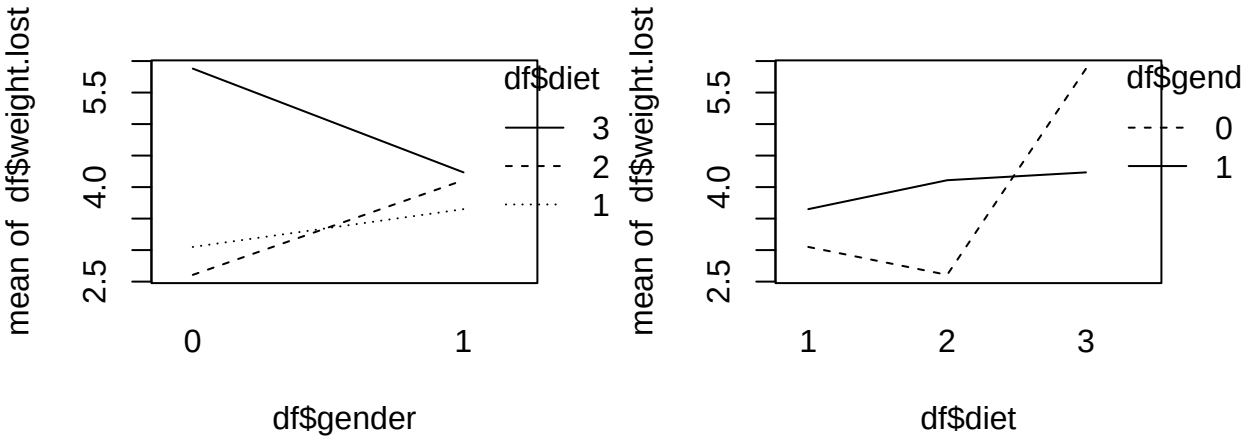
```
## Analysis of Variance Table
##
## Response: weight.lost
##             Df Sum Sq Mean Sq F value Pr(>F)
## gender       1      0    0.28    0.05 0.8206
## diet         2     60   30.21    5.62 0.0055 **
## gender:diet  2     34   16.95    3.15 0.0488 *
## Residuals   70    376    5.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this table we can see that gender does not have an effect of its own, however, it interacts with diet to affect weight loss.

```
summary(dietgenderaov)
```

```
##
## Call:
## lm(formula = weight.lost ~ gender * diet, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.51  -1.30   0.07   1.22   5.45
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.050      0.620    4.92  5.5e-06 ***
## gender1          0.600      0.960    0.62   0.5340
## diet2           -0.443      0.876   -0.51   0.6149
## diet3            2.830      0.862    3.28   0.0016 **
## gender1:diet2    0.902      1.340    0.67   0.5030
## gender1:diet3   -2.247      1.315   -1.71   0.0919 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.32 on 70 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.201,  Adjusted R-squared:  0.144
## F-statistic: 3.52 on 5 and 70 DF,  p-value: 0.00677
```



Assumption: diet depends on gender, we can see that effect of diet varies for women.

```
genderaov <- lm(weight.lost~gender,data=df)
anova(genderaov)
```

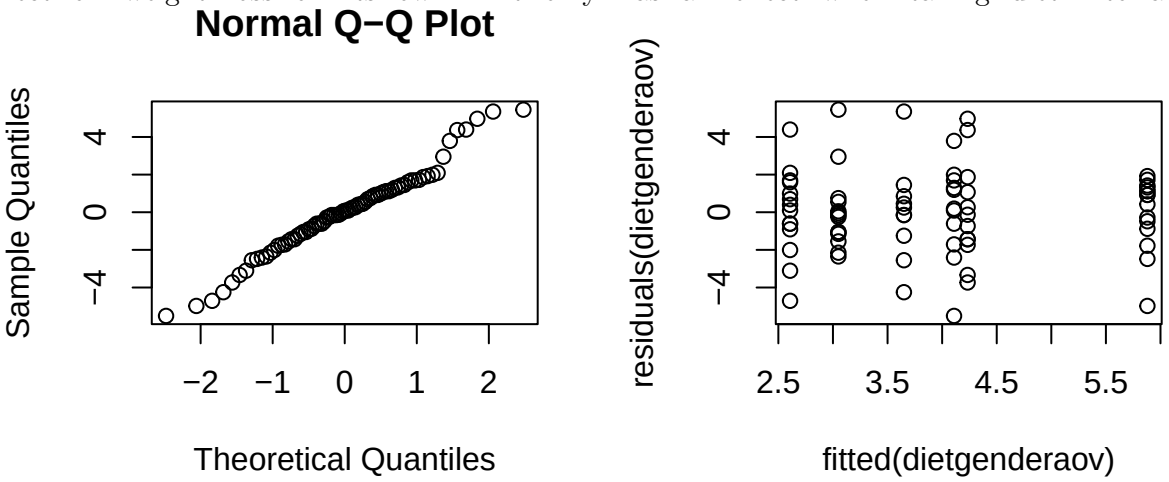
```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## gender     1      0    0.28    0.04   0.83
## Residuals 74    471    6.36
```

```
summary(genderaov)
```

```
##
## Call:
## lm(formula = weight.lost ~ gender, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.993 -1.685 -0.204  1.726  5.185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.893      0.385   10.12  1.3e-15 ***
## gender1        0.122      0.584    0.21     0.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.52 on 74 degrees of freedom
##   (2 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.000591,   Adjusted R-squared:  -0.0129
## F-statistic: 0.0438 on 1 and 74 DF,  p-value: 0.835
```

AS has also been shown in the previous ANOVA table, gender does not have an effect on weight loss on its own. It only has an effect when taking diet into account.



The residuals and QQ-plot of the gender and diet plot do not appear to be normally distributed, which violates our assumptions about normality. Therefore, one must take our previous results with caution.

d) Skipped, as instructed.

e) We prefer the model from b as c looks irrelevant for the weight loss and violates the needed assumptions.

```
c(dietaov$coefficients[1], dietaov$coefficients[1] + dietaov$coefficients[2], dietaov$coefficie
```

```
## (Intercept) (Intercept) (Intercept)
##        3.30        3.03        5.15
```

An average person would lose approximately 3 kg for both diet 1 and 2, while losing approximately 5 kg for diet 3.

Exercise 4

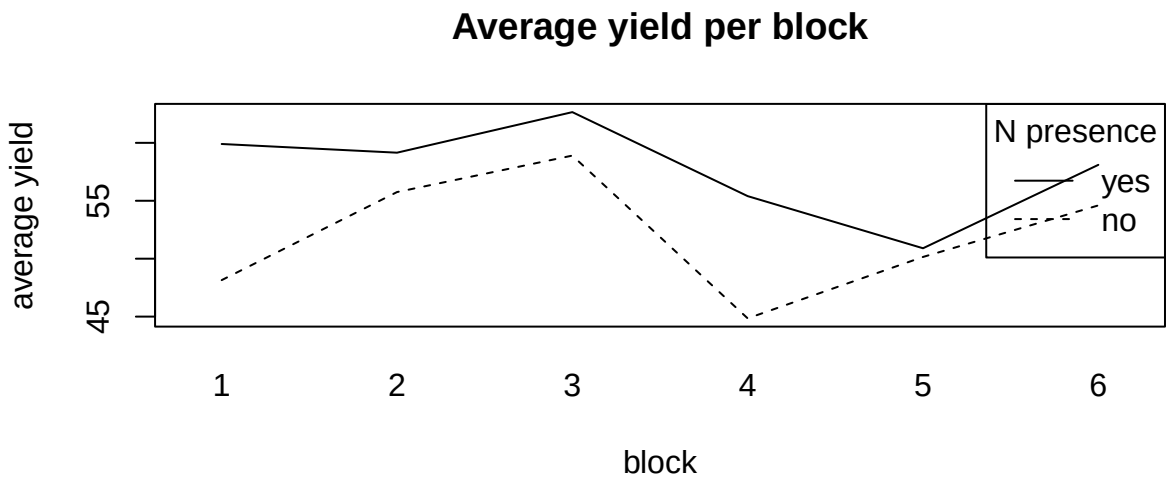
a)

```
B <- 6; P <- 4; T <- 3

process <- c()
for (i in 1:B) {
  block <- c()
  for (tr in 1:T) {
    block <- cbind(block, as.numeric(sample(1:P) > 2))
  }
  process <- rbind(process, block)
}

process <- t(process)
```

```
rownames(process) <- c("N", "P", "K")
colnames(process) <- paste0(rep(1:6, each=4), paste0(".", rep(1:4, 6)))
process
```

```
##   1.1 1.2 1.3 1.4 2.1 2.2 2.3 2.4 3.1 3.2 3.3 3.4 4.1 4.2 4.3 4.4 5.1 5.2 5.3
## N   0   1   0   1   0   1   1   0   1   0   0   1   1   0   1   0   1   1   0
## P   1   0   0   1   1   1   0   0   0   0   1   1   0   1   0   1   1   0   0
## K   1   1   0   0   1   0   0   1   0   0   1   1   0   1   1   0   0   1   1
##   5.4 6.1 6.2 6.3 6.4
## N   0   1   0   0   1
## P   1   1   0   1   0
## K   0   0   1   0   1
```

In the table rows represent every soil additive and columns represent 6 blocks, each with 4 plots (first number - block, second number - plot). As you can see, every additive appears twice in each block.



b)

We have reason to believe that *block* may affect *yield*. This could happen because of slightly different environmental conditions: sun exposure, soil composition, etc. The plot supports this idea since it appears that average yield varies depending on the block when N is both present and absent.

c)

```
n_block_lm <- lm(yield~block*N, data=npk)
anova(n_block_lm)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    3.36   0.04 *
## N          1    189   189.3    9.26   0.01 *
## block:N    5     99    19.7    0.96   0.48
## Residuals 12    245    20.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is seen from the results that N and *block* are significant, but their interaction is not. Let's check summary for every level to see how each block individually affect the *yield*.

```
summary(n_block_lm)
```

```
##
## 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)    48.15       3.20   15.06  3.7e-09 ***
## block2          7.60       4.52    1.68    0.119
## block3         10.75       4.52    2.38    0.035 *
## block4         -3.30       4.52   -0.73    0.479
## block5          2.00       4.52    0.44    0.666
## block6          6.45       4.52    1.43    0.179
## N1             11.75       4.52    2.60    0.023 *
## block2:N1      -8.35       6.39   -1.31    0.216
## block3:N1      -8.00       6.39   -1.25    0.235
## block4:N1      -1.20       6.39   -0.19    0.854
## block5:N1     -11.00       6.39   -1.72    0.111
## block6:N1      -8.25       6.39   -1.29    0.221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.52 on 12 degrees of freedom
## Multiple R-squared:  0.72,   Adjusted R-squared:  0.464
## F-statistic: 2.81 on 11 and 12 DF,  p-value: 0.0449
```

We can see that the difference between blocks is not significant except for block 3. Additionally, it seen that presence of N in the soil makes significant difference to the *yield*. Moreover, interaction between *block* and N is insignificant.

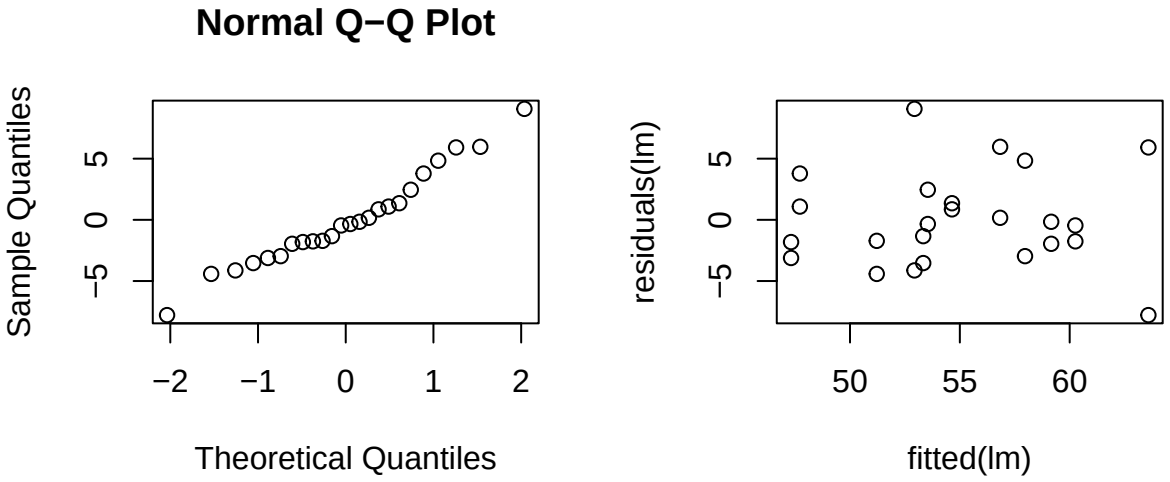
Because interaction is not significant in our case but *block* is significant by itself, we should go for the additive model.

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

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    3.40 0.0262 *
## N          1    189   189.3    9.36 0.0071 **
```

```
## Residuals 17    344    20.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the additive model we see that both *block* and *N* are still significant. Finally, we are going to check the model assumptions about normality of residuals.



The residuals and the Q-Q plot appear fairly normal.

Lastly, we cannot use Friedman test as we have two observations for each combination of soil additives.

d)

```
npklm1 <- lm(yield~P + K + block*N, data=npk)
npklm2 <- lm(yield~N + K + block*P, data=npk)
npklm3 <- lm(yield~N + P + block*K, data=npk)
npklm4 <- lm(yield~block + N + P + K, data=npk)
npklm5 <- lm(yield~N + P + K, data=npk)
```

```
print('Y ~ P + K + block*N')
```

```
## [1] "Y ~ P + K + block*N"
```

```
anova(npklm1)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## P          1      8     8.4    0.59 0.4590
## K          1     95    95.2    6.72 0.0268 *
## block      5    343    68.7    4.85 0.0164 *
## N          1    189   189.3   13.36 0.0044 **
## block:N    5     99    19.7    1.39 0.3066
## Residuals 10    142    14.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print('Y ~ N + K + block*P')
```

```
## [1] "Y ~ N + K + block*P"
```

```
anova(npklm2)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## N          1    189   189.3   11.21 0.0074 **
## K          1     95    95.2    5.64 0.0389 *
## block      5    343    68.7    4.07 0.0282 *
## P          1      8     8.4    0.50 0.4966
## block:P    5     71    14.3    0.85 0.5473
## Residuals 10    169    16.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print('Y ~ N + P + block*K')
```

```
## [1] "Y ~ N + P + block*K"
```

```
anova(npklm3)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## N          1    189   189.3   11.14 0.0075 **
## P          1      8     8.4    0.49 0.4980
## block      5    343    68.7    4.04 0.0288 *
## K          1     95    95.2    5.60 0.0395 *
## block:K    5     70    14.1    0.83 0.5583
## Residuals 10    170    17.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print('Y ~ block + N + P + K')
```

```
## [1] "Y ~ block + N + P + K"
```

```
anova(npklm4)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    4.29 0.0127 *
## N          1    189   189.3   11.82 0.0037 **
## P          1      8     8.4    0.52 0.4800
```

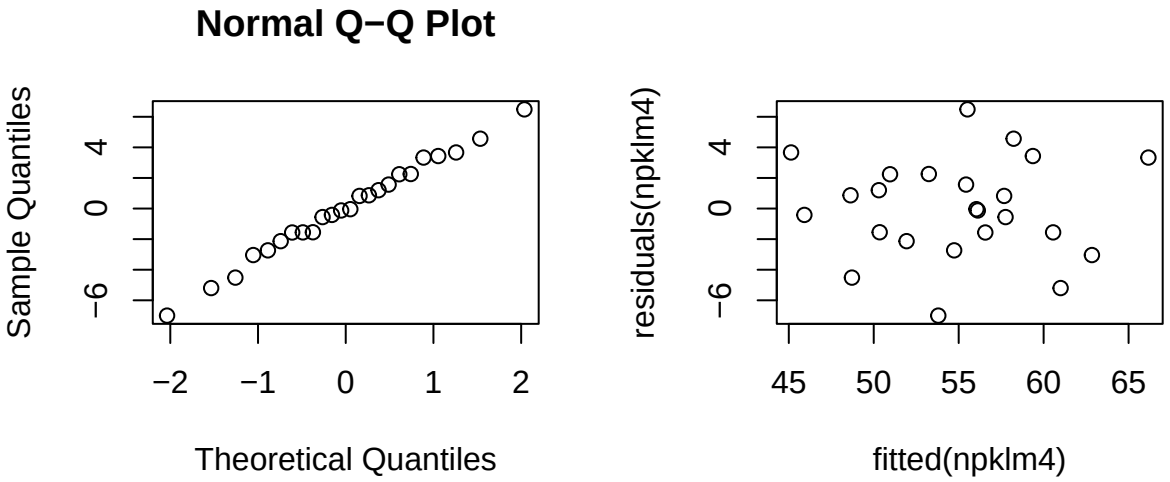
```
## K          1     95    95.2    5.95 0.0277 *
## Residuals 15    240    16.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

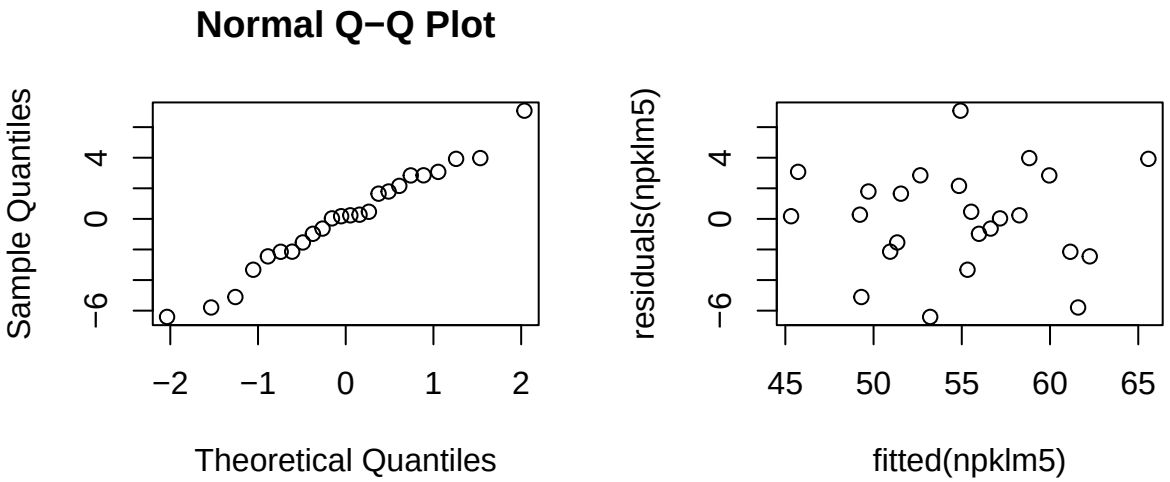
We have tested interaction models as well as additive. All the possible interactions between *block* and soil additives are insignificant. Therefore our preference goes to additive model as it shows significance of independent factors. Moreover, it is shown that P is an insignificant factor for the analysis.

```
npklm5 <- lm(yield~block + N + K, data=npk)
anova(npklm5)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5    343    68.7    4.42  0.010 *
## N          1    189   189.3   12.18  0.003 **
## K          1     95    95.2    6.13  0.025 *
## Residuals 16    249    15.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We believe that the model presented above is the best one, since all the factors are significant. But to examine further we have checked the normality for additive model with all the factors and for the additive model without P.





Since we cannot assume normality for the model without P factor, we are going to stick with the additive model which contains all the presented factors.

e)

```
require(lme4)
```

```
## Loading required package: lme4
```

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

```
npklmer <- lmer(yield~N+(1|block), REML=FALSE, data=npk)
npklmer1 <- lmer(yield~(1|block), REML=FALSE, data=npk)
anova(npklmer1, npklmer)
```

```
## Data: npk
## Models:
## npklmer1: yield ~ (1 | block)
## npklmer: yield ~ N + (1 | block)
##          npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## npklmer1    3 159 163  -76.7      153
## npklmer     4 154 158  -72.7      146   7.9  1      0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the additional analysis can be seen that N has a significant effect on the *yield*, which supports what we found in the point **(c)**.