Experimental Design and Data Analysis - Assignment 1

Group 11 - Björn van der Haas, Deividas Aksomaitis, Nur Basak Ozer
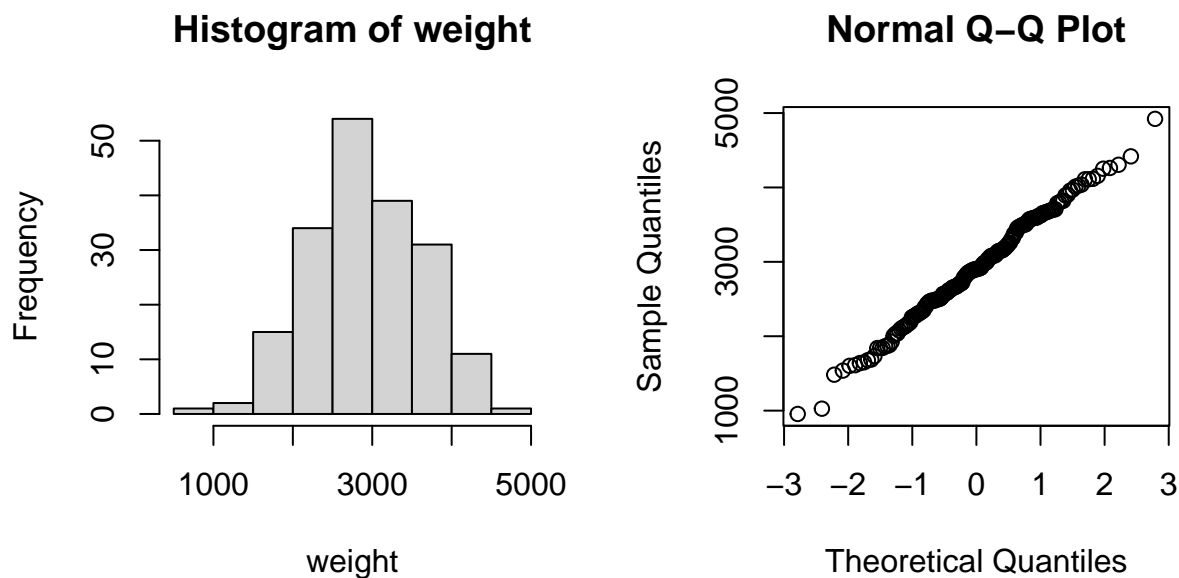
2023-02-27

Exercise 1: Birthweights

Section a

```
par(mfrow=c(1,2))
data <- read.table(file="birthweight.txt",header=TRUE)
weight <- data$birthweight

# Check for normality of the data
hist(weight); qqnorm(weight)
```



```
round(shapiro.test(weight)$p.value, 3)
```

```
## [1] 0.9
```

The histogram somewhat resembles bell shape and the qq plot is somewhat straight - these are the signs of normal distribution. After performing Shapiro-Wilk test on weights, we get p=0.9, which is larger than 0.05. Therefore, we cannot reject H_0, which states that normal distribution is assumed. However we must consider the fact that Shapiro-Wilk test is not reliable when not rejecting.

```
# Confidence interval
m = mean(weight); sd = sd(weight); len = length(weight); error = qnorm(0.98)*sd/sqrt(len)
```

```
round(c(m-error, m+error))
```

```
## [1] 2809 3018
```

Bounded 96%-CI for μ is (2809, 3018).

```
# Sample size for 96%-CI is at most 100.
round((qnorm(0.98)*sd/50)^2)
```

```
## [1] 821
```

The sample size needed to provide that the length of the 96%-CI is at most 100 is 821.

```
# Bootstrap 96%-CI for mean
B=999
Tstar=numeric(B)

for(i in 1:B) {
    Xstar=sample(weight,replace=TRUE)
    Tstar[i]=mean(Xstar) }

Tstar2=quantile(Tstar,0.02)
Tstar98=quantile(Tstar,0.98)
sum(Tstar<Tstar2)
```

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

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

```
##  98%   2%
## 2804 3022
```

Bootstrap 96%-CI for μ is: (2808, 3026). It will be slightly different every time the calculation is performed because of sampling. Although the methods to calculate bounded CI and bootstrap CI differ, the intervals calculated here are quite similar.

Section b

To determine whether the mean birthweight is bigger than 2800 grams we can use one-sample t-test, because the data has a normal distribution and we are testing the mean of a single population. Our test hypotheses would be H_0: μ=2800 and H_1: μ>2800. Moreover, t-value for 187 degrees of freedom, one-tailed is 1.653.

```
# t-test
ttest <- t.test(weight, mu=2800, alternative="greater")
round(ttest$statistic, 3)
```

```
##     t
## 2.227
```

The calculated t-value is 2.227, which is greater than 1.653, which means that we reject the null hypothesis.

```
round(ttest$conf.int)
```

```
## [1] 2829  Inf
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval in the R-output represents the range of values for the true population mean birthweight with 95% confidence. The interval is from 2829 to infinity, which means that we can be 95% confident that the true mean birthweight is at least 2829 grams.

We can also perform a one-sample sign test. The data does not have to be normal and the test will be performed to test the median. Our test hypotheses would be H_0: m=2800 and H_1: m>2800.

```r
# sign test
median(weight)
```

```
## [1] 2901.5
```

```r
round(binom.test(sum(weight > 2800), len, p=0.5, alternative="greater")$p.value, 3)
```

```
## [1] 0.034
```

The test gives a p-value of 0.034 which is lower than 0.05 significance level, therefore we reject the null hypothesis.

Wilcoxon signed rank test for one sample is another test we can perform. The test hypotheses would be H_0 : m=2800 and H_1 : m>2800.

```r
# Wilcoxon signed rank test for one sample
round(wilcox.test(weight, mu=2800, alternative="greater")$p.value, 3)
```

```
## [1] 0.015
```

We get a p-value of 0.015 which is lower than 0.05 significance level, therefore we reject the null hypothesis.

Section c

We know that power of a test is $P(\text{rejecting } H_0 \mid H_0 \text{ is false}) = 1 - P_{H1}(\text{not rejecting } H_0)$. To compute the powers of the t-test and sign test we would need to specify an alternative hypothesis first. In our case, it would be H_0: μ=2800 and H_1: μ>2800.

```r
B <- 1000; n <- 100
psign <- numeric(B); pttest <- numeric(B)

for(i in 1:B) {
  x=rnorm(n, mean=2850, sd=200)
  pttest[i] <- t.test(x, mu=2800, alternative="greater")$p.value
  psign[i] <- binom.test(sum(x > 2800), n, p=0.5, alternative="greater")$p.value }

sum(psign<0.05)/B
```

```
## [1] 0.589
```

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

```
## [1] 0.807
```

To generate data, we choose alternative mean μ as 2850, standard deviation as 200 and sample size as 100. We generate this data 1000 times and each time we perform both t-test and sign test on it. In the end, we check how many times the tests correctly rejected the null hypothesis. In our case, sign test got it 626 times (P=0.626), while t-test got it 800 times (P=0.8). The t-test has a higher power, therefore it is better. The higher the power (closer to 1), the better the test performance is. If the data is normal, t-test should always be used instead of sign test.

Section d

```r
proportion = round(sum(weight < 2600) / 188, 2); margin = proportion - 0.25;
z = round(margin / sqrt( (proportion * (1 - proportion)) / 188 ), 2)
```

$$\hat{p} = 0.33$$

3

$$margin = 0.08$$

$$\hat{p}_r = 0.33 + 0.08 = 0.41$$

$$z = 2.33$$

The confidence interval is [0.25, 0.41]. Since the z-score is 2.33, the area is 0.9901 and the confidence level is 98%.

Section e

To test whether the mean weight is different for male and female babies, we can use two-sample t-test. Our test hypotheses would be H_0: $\mu=0$ and H_1: $\mu \neq 0$.

```
gender <- rep(NA, len); gender[weight < 2600] <- c(rep("male", 34), rep("female", 28))
gender[weight >= 2600] <- c(rep("male", 61), rep("female", 65))
round(t.test(weight ~ gender, var.equal = TRUE)$p.value, 3)
```

```
## [1] 0.183
```

After performing the test, we get a p-value of 0.183, which is higher than our significance level of 0.05. Therefore we do not reject the null hypothesis. This means that the mean weight is not different for male and female babies.
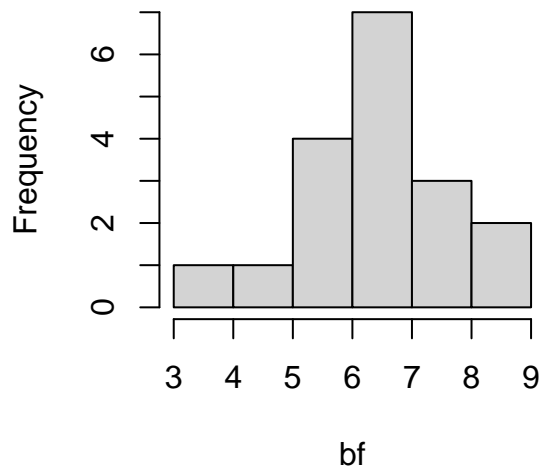
Exercise 2: Cholesterol

```
par(mfrow=c(1,2))
ch <- read.table(file="cholesterol.txt",header=TRUE)
bf <- ch$Before; a8 <- ch$After8weeks
```
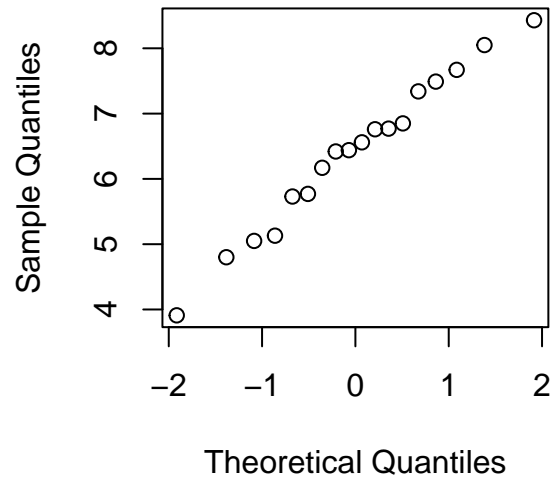
Section a

Here we make the most relevant plots of the data in this case. We will use these to comment on normality and sample size.

```
# For the column 'Before'
par(mfrow=c(1,2))
hist(bf); qqnorm(bf)
```
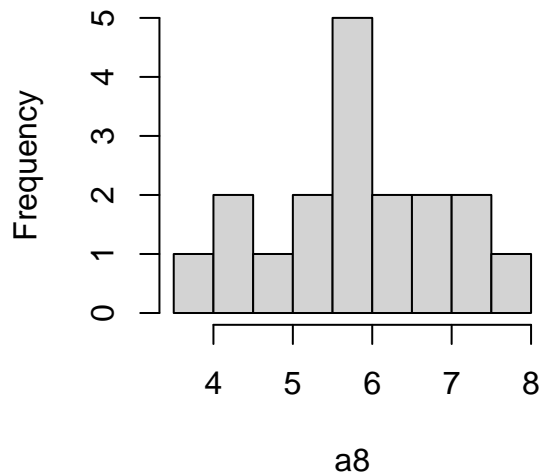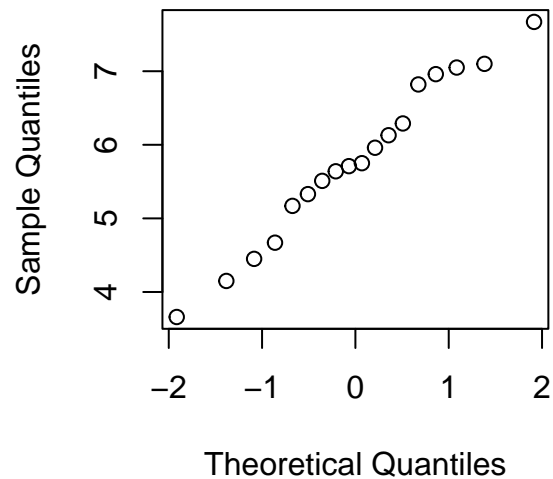
Histogram of bf

bf

Normal Q–Q Plot


```
# For the column 'Before'
par(mfrow=c(1,2))
hist(a8); qqnorm(a8)
```

Histogram of a8

a8

Normal Q–Q Plot

The histograms would resemble a bell curve if the data comes from a normal distribution. Hence we deduce that the samples appear to be normally distributed. We also know that if the points on a qq-plot are approximately on a straight line, then the data can be assumed to be sampled from a normal distribution. Judging by this fact, the samples indeed look normally distributed.

```
boxplot(ch)
```
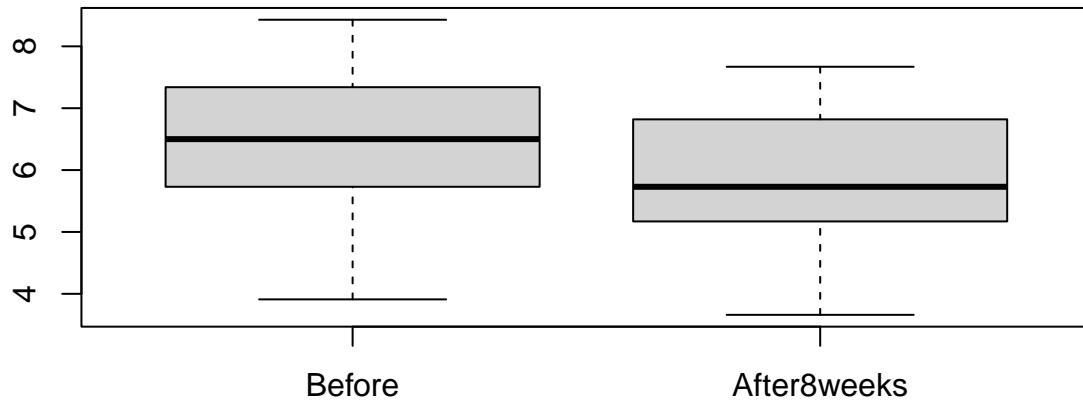

It looks like the sample size is relatively small. Moreover, there are not many outliers outside the box, hence the data seems consistent.

We can also corroborate the normality of data by performing the Shapiro-Wilk test. Here we test the correctness of hypothesis H_0 which states that P is a normal distribution:

```
# For the column 'Before'
round(shapiro.test(bf)$p.value, 3)
```

```
## [1] 0.967
```

```
# For the column 'After8weeks'
round(shapiro.test(a8)$p.value, 3)
```

```
## [1] 0.918
```

The p-value is larger than 0.05 in both cases, which means there is no strong evidence against the normality of the samples. Hence we don't reject H_0 for both cases. However we must consider the fact that Shapiro-Wilk test is not reliable when not rejecting.

Now we investigate whether the columns Before and After8weeks are correlated:

```
round(cor(ch),3)
```

```
##            Before After8weeks
## Before      1.000       0.991
## After8weeks  0.991       1.000
```

Based on these results, we observe that the variables are correlated, exhibiting a linear relationship.

Section b

First and foremost, we must address whether the samples are paired or not. We know that this is an experiment with exactly 2 numerical outcomes per experimental unit (or simply, every row consists of 2 measurements that come from the same individual). Additionally, we assume that the experiments performed

on the same unit are under different conditions, i.e. each row is reserved for the cholesterol levels recorded before and after the experiment for the same individual. Hence we conclude that **the data samples are paired.** Furthermore, we assumed that the data comes from a normal distribution. Hence we can apply paired sample t-test to investigate the effect of margarine (in other words, the difference between samples recorded before and after the diet respectively). Recall that we are testing hypothesis H_0, which states that the difference between the paired samples is 0:

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

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

The p-value we obtained here is significantly smaller than 0.05. This means that we reject H_0 and deduce that the margarine was effective in reducing cholesterol levels.

As an alternative, we can also use the Wilcox signed test here. This is because the sample data comes from a symmetric population (Normal distribution is known to be symmetric). Thus, to test the previous hypothesis, we apply the wilcox.test function and set the "paired" argument as TRUE:

```
wilcox.test(bf,a8,paired=TRUE)
```

```
##
##   Wilcoxon signed rank exact test
##
## data:  bf and a8
## V = 171, p-value = 7.629e-06
## alternative hypothesis: true location shift is not equal to 0
```

Again, the p-value is significantly smaller than 0.05. It supports our conclusion that margarine was effective in reducing cholesterol levels. Hence we reject H_0.

In addition, we can apply the permutation test in this case. This is because the data consists of two paired samples. Moreover, due to the lack of normality assumption, the permutation test can be performed over both normal and non-normal data. Lastly, we use a test statistic that expresses the difference between the X and Y within pairs, such as the difference of the means of the paired samples.

Section c

For this section of the exercise, we assume that the column 'After8weeks' is uniformly distributed with min parameter 3 and max parameter θ where $\theta > 3$. We want to estimate this θ, so we need to find the point estimate $\hat{\Theta}$ using the definition of the CLT. With using this point estimate $\hat{\Theta}$, the mean of the distribution would now be $(3+\hat{\Theta})/2$. Let us equate this mean to the mean of the column 'After8weeks', say m. Then the point estimate $\hat{\Theta}$ would be 2m -3. We also know that the variance of the distribution would be $(\theta-3)/12$. Here we can use our point estimate to replace the distribution's variance:

```
m <- mean(a8); t_est <- 2*(m) - 3; est_var <- (t_est-3)*(t_est-3)/12
```

Now we need to inspect whether we have chosen a good estimator by constructing the CI as our next step. We know that the $(1 - \alpha)$-confidence interval for θ would be $\hat{\Theta} \pm z\alpha/2*(\sigma/\sqrt{n})$. To find the value of $z\alpha/2$, we

will use the command qnorm(0.975). Here, 0.975 is the value of the upper quantile, which we obtained from the following calculations:

$$1 - 0.95 = 0.05$$

$$0.05/2 = 0.025$$

$$1 - 0.025 = 0.975$$

```r
up_qt <- qnorm(0.975)
```

Now we can calculate the margin of error using the formula mentioned above. We must also calculate the sample variance as well:

```r
sample_size <- length(a8); marg_err <- up_qt*sqrt(est_var/sample_size)
marg_err
```

```
## [1] 0.7411778
```

Using the value we found for the margin of the error, we can now determine the lower and upper bounds of the confidence interval:

```r
lb <- t_est - marg_err; ub <- t_est + marg_err
```

```r
round(lb,3)
```

```
## [1] 7.817
```

```r
round(ub,3)
```

```
## [1] 9.299
```

Based on this result, we obtain a 95% CI of [7.817, 9.299].

- **Can we improve this CI?** We can do a few things to improve the confidence interval we found above. We know that it is essentially the probability that our random interval contains the true value. Hence, if we take a larger sample size of n, the confidence interval will be narrower, which means we will obtain more certainty at the same confidence level. Moreover, if we get a smaller σ (or s) value, the confidence interval will also be narrower, and we will obtain less uncertainty at the same confidence level. Lastly, if we take a larger α, the confidence interval will again be narrower, and we will have increased accuracy. However because 1-α will be smaller, we would have a lower confidence level.

Section d

In this section, we will perform a bootstrap test. Here we wish to test hypothesis H_0 that the sample 'After8weeks' comes from a uniform distribution with the min parameter set to 3 and the max parameter set to θ, where θ belongs to the interval [3,12], by applying the following procedure:

```r
sample_size <- length(a8); t_stat <- max(a8)
t_stat
```

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

We obtain our test statistics which is the max value of the sample. We will use this value when calculating the p-values.

```r
B <- 1000; thetas <- seq(3, 12, by = 0.1); p_values <- rep(0,length(thetas))
```

We set the value of B as 1000, which is the number of times we will repeat the bootstrap test. Here we also define the interval for θ as an R sequence as we intend to calculate the p-values for each possible θ value. We then proceed with the bootstrap test, where we perform our calculations inside a nested loop. Inside the outer loop, we define the vector for T* values, which will be stored during the execution of the inner loop.

Moreover, the length of the vector will be the same as the number of trials (B). Inside the inner loop, we simulate a uniform distribution for each θ value from our defined interval. Moreover, we get the max of the result and store it in our t-array before calculating the p-value of each θ value.

```r
for (t in 1:length(thetas)){

  tstar <- rep(0,B)

  for(i in 1:B){

    xstar <- runif(sample_size,min=3,max=thetas[t])
    tstar[i] <- max(xstar)}

  p <- sum(tstar>=t_stat)/B
  p_values[t] <- p

}
```

After performing the test, we need to calculate the p-values which determine the acceptance and rejection region. More precisely, we ensure that T^* values reside in the rejection region and more probable T^* belongs to the acceptance region. Here, we are interested in the p-values that do not cause the rejection of H_0. Hence we must calculate the probabilities where T^* is greater than or equal to our statistics.

```r
plot(thetas, p_values, type = "l", main = "p-values for each theta value",
     xlab = "theta values", ylab = "p-values", xlim = c(3,12))
abline(h = 0.05, lty = 2, lwd = 2, col = "red")
```

p–values for each theta value
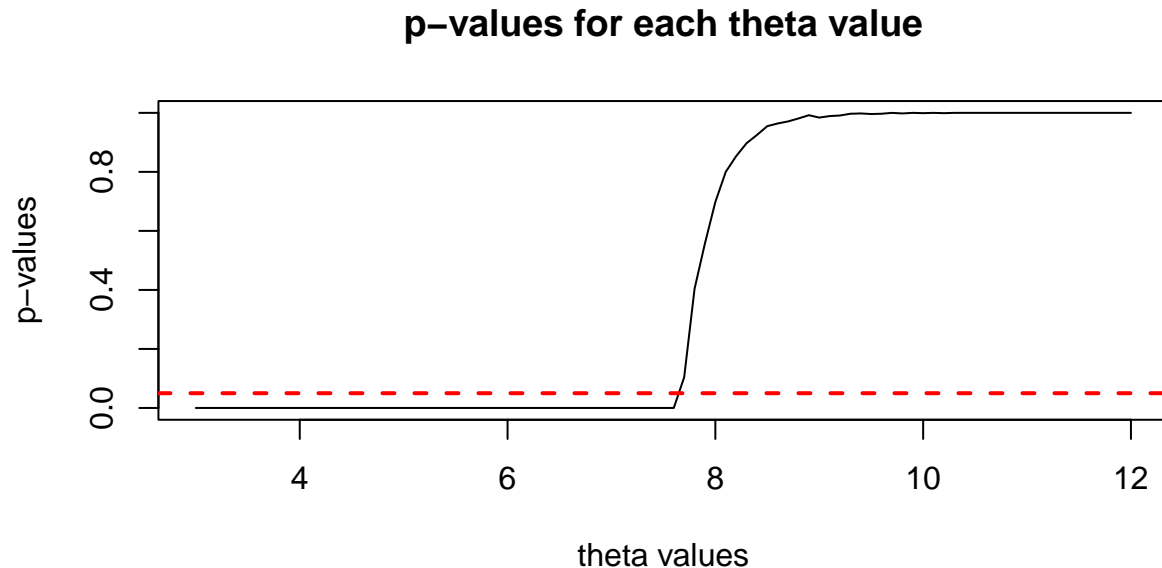
Judging by the plot above, we can observe the θ values that yield p-values that are larger than 0.05. This means that when these values are substituted into the above-mentioned distribution, the hypothesis H_0 would not be rejected.

On another note, we cannot apply the Kolmogorov-Smirnov test as it assumes that the test distribution is fully specified, including the parameters. Since the hypothesis is based on a uniform distribution with an unknown max parameter, this assumption is violated, thus making the test inapplicable for this situation.

9

Section e

Here, we would like to confirm the fact that the median m of the cholesterol levels after 8 weeks of a low-fat diet is less than 6. Based on this requirement, we determine our hypotheses as H_0: m=6 and H_1: m < 6. To test this hypothesis, we can again perform a Wilcox signed rank test using the parameter alt="l":

```
round(wilcox.test(a8,mu=6,alt="l",exact=FALSE)$p.value, 3)
```

```
## [1] 0.223
```

Since the p-value we obtained is larger than 0.05, we cannot reject H_0, which means there is no strong evidence that suggests the median m is less than 6 after 8 weeks.

Moreover, we are requested to apply a test that investigates whether the fraction of the cholesterol levels after 8 weeks of a low-fat diet less than 4.5 (less_45) is at most 25%. For that, we can perform a sign test here where the tested hypotheses would be H_0: less_45=0.25 and H_1: less_45 < 0.25.

```
num_of_succ <- sum(a8 < 4.5)
num_of_succ
```

```
## [1] 3
```

Here, we found that the number of successes (i.e. cholesterol levels after 8 weeks of a low-fat diet less than 4.5) to be 3. We will set the parameter t to this number and proceed with the sign test:

```
round(binom.test(num_of_succ,sample_size,p=0.25,alt="l")$p.value, 3)
```

```
## [1] 0.306
```

Again, we obtained a p-value that is larger than 0.05. Thus we cannot reject H_0 and conclude that the fraction less_45 is at most 25%.

Exercise 3: Diet

Note that we must check every respective variable is properly classified as a factor with is.factor(x), with a TRUE outcome, and is.numeric(x), with a FALSE outcome.

```
# Reading data
datadi <- read.table(file="diet.txt", header=TRUE)

# Adding variable weight.lost
datadi$weight.lost <- c(datadi$preweight - datadi$weight6weeks);
datadi$diet <- as.factor(datadi$diet)
```
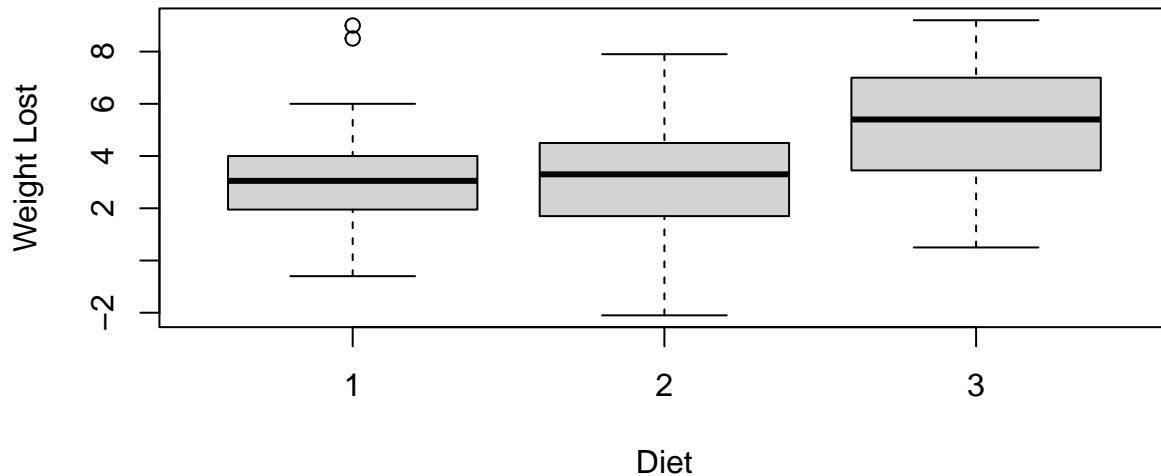
Section a

Informative graphical summary for the effect of diet on weight loss Boxplot (as opposed to e.g. scatterplot) makes the most sense as diet is a factor.

```
plot(datadi$diet, datadi$weight.lost, xlab = "Diet", ylab = "Weight Lost",
     main = "Effect of diet on weight lost")
```

Effect of diet on weight lost

Since each measure comes from the same individual we are dealing with paired samples.

```
# Paired t-test for preweight and weight6weeks
t.test(datadi$preweight, datadi$weight6weeks, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  datadi$preweight and datadi$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
```

$p < 0.05$, which means there is a significant effect of diet on weight.
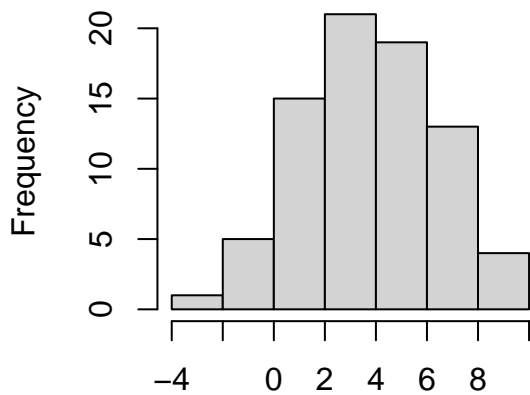
This test requires the assumption of normality on the differences. Hence we check this with Q-Q plot, histogram and Shapiro-Wilk test.

```
par(mfrow=c(1,2))
qqnorm(datadi$preweight-datadi$weight6weeks, main = "Normal Q-Q Plot Difference")
hist(datadi$preweight - datadi$weight6weeks, main = "Histogram of Difference")
```

Normal Q−Q Plot Difference
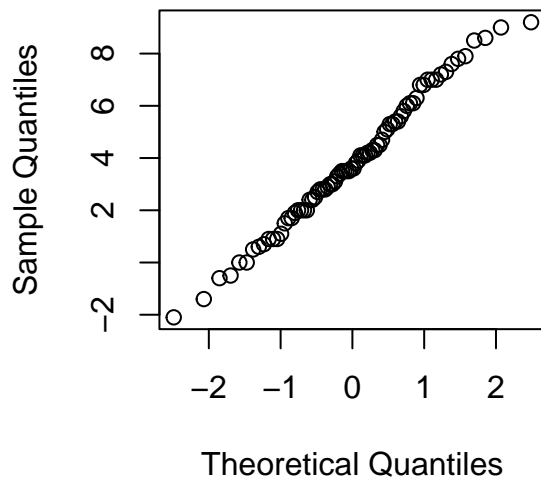
Histogram of Difference

Q-Q plot looks like a fairly straight line. The histogram also looks normal.

```
round(shapiro.test(datadi$preweight-datadi$weight6weeks)$p.value, 3)
```

```
## [1] 0.802
```

Shapiro-Wilk test does not reject normality (inconclusive as it is not reliable when not rejecting). Considering Shapiro-Wilk does not reject, Q-Q plot looks fairly straight and histogram looks normal, we can assume normality.

Section b

```
modeldi <- lm(weight.lost ~ diet, data = datadi); anova(modeldi)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##           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
```

$p < 0.05$ (in fact $P < 0.01$) so we reject H_0, concluding that diet has a significant effect on weight lost

```
summary(modeldi)
```

```
##
## Call:
## lm(formula = weight.lost ~ diet, data = datadi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1259 -1.3815  0.1759  1.6519  5.7000
##
```

12

```
## 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
```

Based on this summary table, diet 3 is the highest, mu + a3 = 3.3 + 1.8481 = 5.15. It is followed by diet 1 with 3.3 and then diet 2 with 3.3 - 0.2741 = 3.0259.

```
# Diagnostics for ANOVA
par(mfrow=c(1,2))
qqnorm(residuals(modeldi)); plot(fitted(modeldi), residuals(modeldi))
```
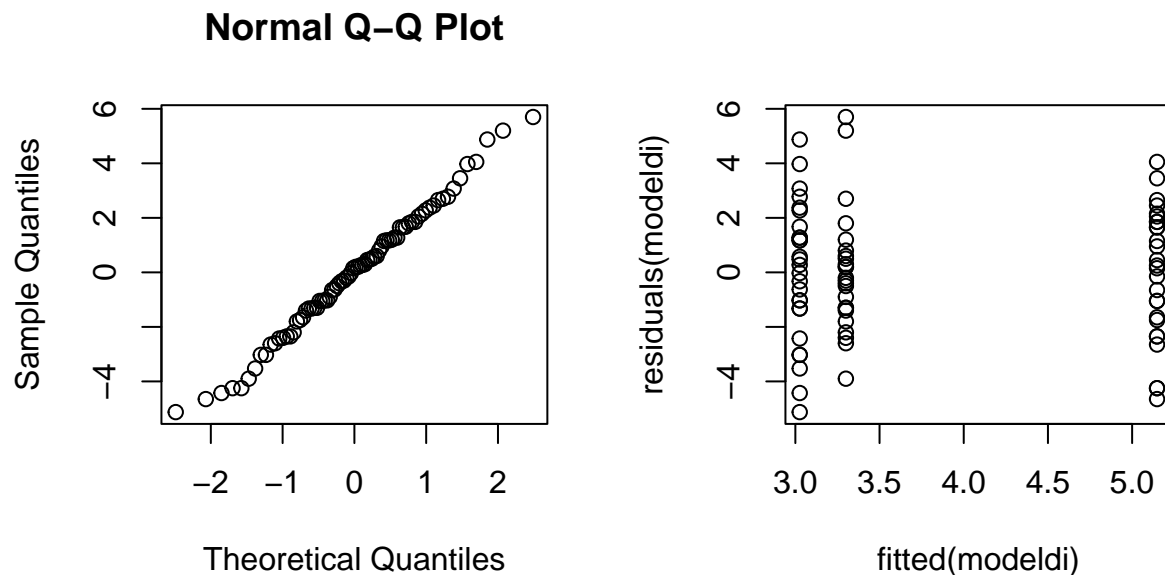

Q-Q plot indicates the normality of residuals (straight line). We also observe no clear relationship in Fitted vs Residuals plot, which is the desired outcome. As we can safely assume normality for this ANOVA, the Kruskal-Wallis test would be superfluous.

Section c

As there are two missing values these ids have to be removed now that we are looking at gender.

```
datadi2 <- datadi[complete.cases(datadi), ]; datadi2$gender <- as.factor(datadi2$gender)
```

```
# Two-Way ANOVA
modeldi2 = lm(weight.lost ~ diet * gender, data = datadi2); anova(modeldi2)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
```

```
##              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
```

$p < 0.05$, thus we reject H0 and can conclude that there is a significant interaction effect.

```
# Diagnostics
par(mfrow=c(1,2))
qqnorm(residuals(modeldi2)); plot(fitted(modeldi2), residuals(modeldi2))
```
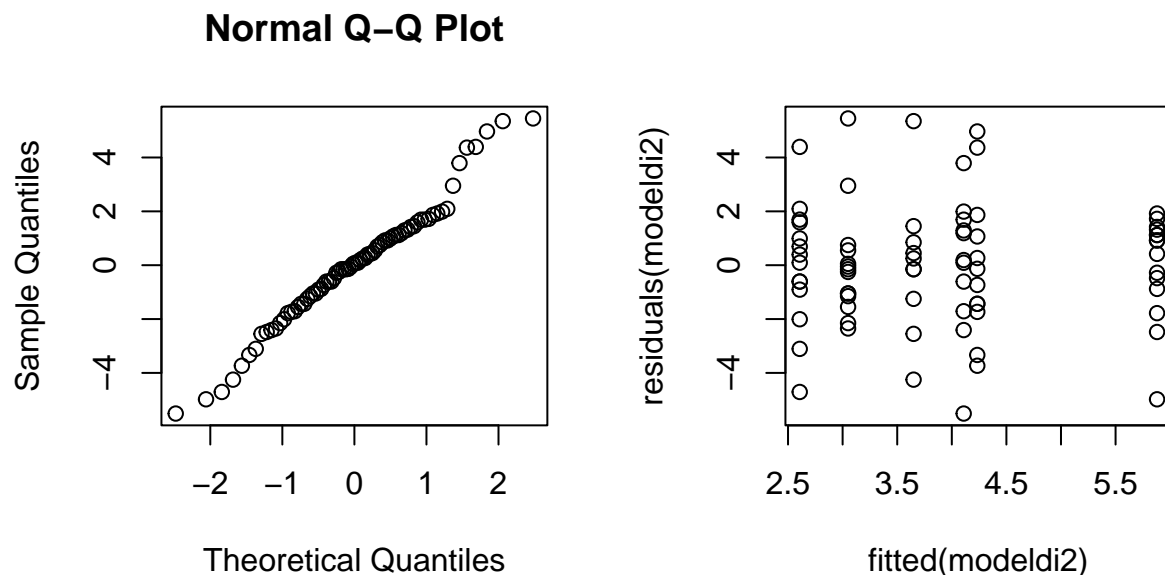
Normal Q–Q Plot

Q-Q plot is somewhat straight but does skew, so normality is doubtful. Furthermore, we do not observe a relation in the Fitted vs Residuals plot, as desired.

Section e

As normality is slightly in doubt, the argument can be made that the approach in c) is better. However, normality is not entirely rejected and thus the approach in d) would be better as we know that gender interacts significantly with diet and should thus be considered for weight loss predictions.

```
dietfem <- data.frame(diet = unique(datadi2$diet), gender = "0")
predict(modeldi2, dietfem, type = "response")
```

```
##        1        2        3
## 3.050000 2.607143 5.880000
```

```
dietmale <- data.frame(diet = unique(datadi2$diet), gender = "1")
predict(modeldi2, dietmale, type = "response")
```

```
##        1        2        3
## 3.650000 4.109091 4.233333
```

The predictions are quite different between male and female. For female, diet #3 clearly has the highest

impact on weight loss, while diet #1 and diet #2 are similar. For male, the differences are less stark, but also indicate diet #3 is the better diet for weight loss followed by diet #2 and then diet #1.

Exercise 4: Yield of peas

Note that we must check every respective variable is properly classified as a factor with is.factor(x), with a TRUE outcome, and is.numeric(x), with a FALSE outcome.

Section a

```
library(MASS)

n_blocks <- 6; n_plots <- 4

scheme <- data.frame(block = rep(1:n_blocks, each = n_plots),
                     N = rep(0, n_blocks*n_plots),
                     P = rep(0, n_blocks*n_plots),
                     K = rep(0, n_blocks*n_plots))

for (i in 1:n_blocks) {

    j <- (i-1)*n_plots+1
    k <- n_plots*i

    scheme[sample(j:k, 2), "N"] <- 1
    scheme[sample(j:k, 2), "P"] <- 1
    scheme[sample(j:k, 2), "K"] <- 1}
```

Section b

```
avg_treated_no <- numeric(n_blocks); avg_treated_yes <- numeric(n_blocks)

for (i in 1:n_blocks) {

    j <- (i-1)*n_plots+1; k <- n_plots*i

    nitro <- npk[j:k, "N"]; yield <- npk[j:k, "yield"]

    avg_treated_no[i] <- sum(yield[nitro == 0]) / 2
    avg_treated_yes[i] <- sum(yield[nitro == 1]) / 2 }

plot(1:n_blocks, avg_treated_yes, type="l", col="blue",
     xlab="Block number", ylab="Average yield", ylim=c(0, 70))
points(1:n_blocks, avg_treated_no, type="l", col="red")
legend("bottomright", legend=c("Treated", "Untreated"), col=c("blue", "red"), lty=1)
```
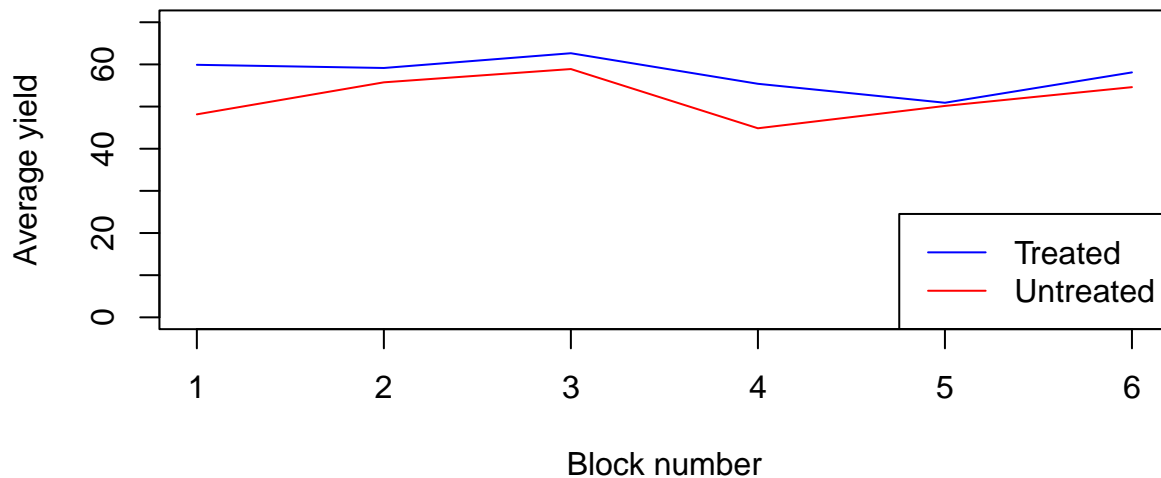
The plot shows us that the average yields for treated soil are a bit higher than for untreated soil. The averages for each block (except for the first one) and treatment tend to have a similar change. For example, block 3 has higher averages for both treatments than block 2. This could mean that the soils are the same in each block, but differ from other blocks.

We do not know whether the soil is all the same or different. The block factor itself is not of interest, but it is there to ensure that the treatments are assigned randomly to each soil within a block as well as reduce the variation and get more precise results.

Section c

```
datac <- npk; npk$block <- as.factor(npk$block); npk$N <- as.factor(npk$N)
dfc <- npk[c("block", "N", "yield")]
```

```
# Two-Way ANOVA
modelc <- lm(yield ~ block * N, data = dfc); anova(modelc)
```

```
## 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
```
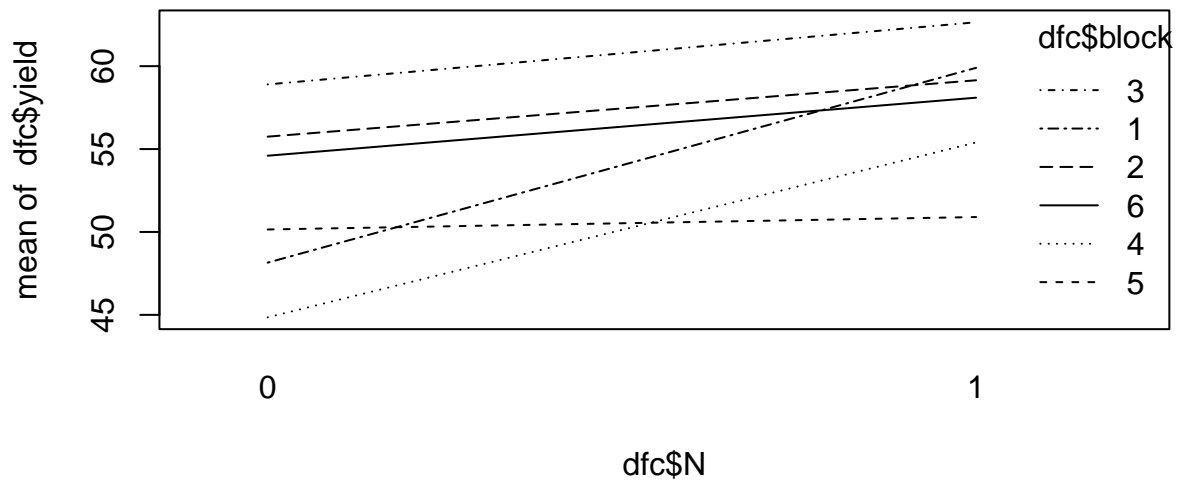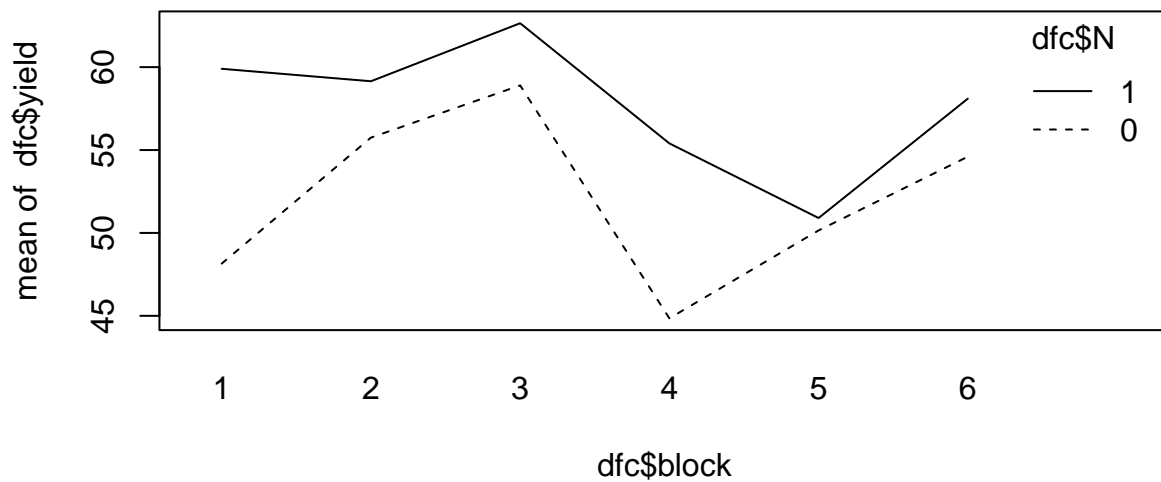
$p > 0.05$, which means there is no significant evidence of interaction effect.

```
interaction.plot(dfc$N,dfc$block,dfc$yield)
```

```
interaction.plot(dfc$block,dfc$N,dfc$yield)
```


Interaction plot for block also display parallel lines, indicating no interaction.

As no significant interaction effect was found, we try an additive model:

```
modeladd <- lm(yield ~ block + N, data = dfc); anova(modeladd)

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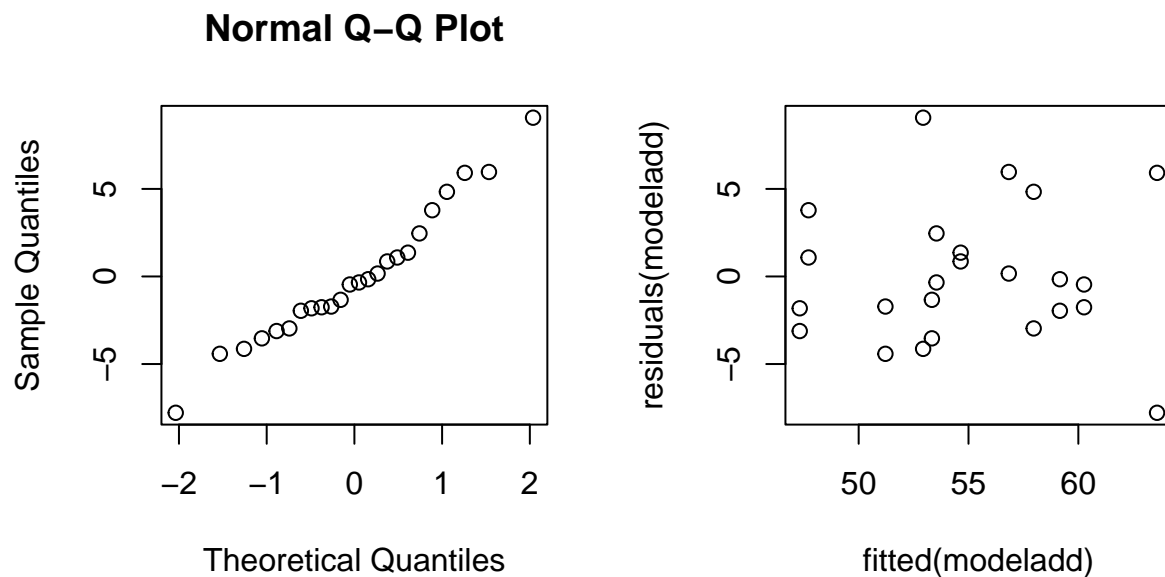
```
## block       5 343.29  68.659  3.3951 0.026173 *
## N           1 189.28 189.282  9.3598 0.007095 **
## Residuals  17 343.79  20.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In both cases $p < 0.05$, so both factors have a main effect in the additive model.

```
# Diagnostics:
par(mfrow=c(1,2))
qqnorm(residuals(modeladd)); plot(fitted(modeladd),residuals(modeladd))
```

Normal Q–Q Plot

There is a slight curve in the Q-Q plot but it is more or less a straight line so it is likely normal. There is no clear relation in Fitted vs Residuals plot, as desired.

- **Was it sensible to include factor block into this model?** By adding block, we have a randomized block design. As the plot can influence the treatment factor (N), but is not directly of interest, this does make sense. By doing so we assume it has a significant effect, which it indeed has.
- **Can we also apply the Friedman test for this situation?** We do have randomized block design, but in this dataset, the treatments are not completely randomized; instead, each block contains two plots that receive and two plots that do not receive each soil additive. Therefore, a Friedman test would not be appropriate for this dataset.

Section d

```
pairwiseP <- lm(yield ~ P*block + N + K, data = datac)
pairwiseK <- lm(yield ~ K*block + P + N, data = datac)
pairwiseN <- lm(yield ~ N*block + K + P, data = datac)

anova(pairwiseP); anova(pairwiseK); anova(pairwiseN)

## Analysis of Variance Table
##
## Response: yield
```

```
##            Df Sum Sq Mean Sq F value   Pr(>F)
## P          1    8.40   8.402  0.4978 0.496588
## block      5 343.29  68.659  4.0678 0.028234 *
## N          1 189.28 189.282 11.2143 0.007381 **
## K          1   95.20  95.202  5.6404 0.038947 *
## P:block    5   71.40  14.280  0.8460 0.547341
## Residuals 10 168.79  16.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: yield
##            Df Sum Sq Mean Sq F value   Pr(>F)
## K          1   95.20  95.202  5.6028 0.039477 *
## block      5 343.29  68.659  4.0407 0.028799 *
## P          1    8.40   8.402  0.4945 0.497989
## N          1 189.28 189.282 11.1397 0.007521 **
## K:block    5   70.27  14.054  0.8271 0.558263
## Residuals 10 169.92  16.992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: yield
##            Df Sum Sq Mean Sq F value   Pr(>F)
## N          1 189.28 189.282 13.3611 0.004423 **
## block      5 343.29  68.659  4.8465 0.016439 *
## K          1   95.20  95.202  6.7201 0.026843 *
## P          1    8.40   8.402  0.5931 0.459045
## N:block    5   98.52  19.704  1.3908 0.306583
## Residuals 10 141.67  14.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No interaction effect for either of the three, thus we do an additive model:

```
modeladd_d <- lm(yield ~ P + K + N + block, data = datac); anova(modeladd_d)
```

```
## Analysis of Variance Table
##
## Response: yield
##            Df Sum Sq Mean Sq F value  Pr(>F)
## P          1    8.40   8.402  0.5247 0.47999
## K          1   95.20  95.202  5.9455 0.02767 *
## N          1 189.28 189.282 11.8210 0.00366 **
## block      5 343.29  68.659  4.2879 0.01272 *
## Residuals 15 240.19  16.012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p < 0.05$ for N and K, showing a main effect. $p > 0.05$ for P, so we can conclude that there is no significant effect.

We conclude that the best model is the additive ANOVA model. It provides an overall indication of the effects of each factor on yield, which the pairwise models can not do. Furthermore, the additive ANOVA

allows us to compare the effects of each factor while controlling for the others, giving us a more accurate estimation which factors influence yield. The pairwise models are worse as they only focus on the interaction between two factors at a time, thus they cannot compare the effects of each factor while controlling for others, and are therefore not as comprehensive.

Section e

```
library(lme4)
npkmer <- lmer(yield ~ N+P+K+(1|block),data=npk,REML=FALSE)
pkmer <- lmer(yield ~ P+K+(1|block),data=npk,REML=FALSE)
anova(npkmer,pkmer)
```

```
## Data: npk
## Models:
## pkmer: yield ~ P + K + (1 | block)
## npkmer: yield ~ N + P + K + (1 | block)
##        npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## pkmer     5 159.49 165.38 -74.745   149.49
## npkmer    6 151.03 158.10 -69.514   139.03 10.46  1    0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarly to Section c, we can conclude that N has a significant effect on yield $p < 0.05$ (in fact $p < 0.01$).