EDDA assignment 2

Mick IJzer, Tirza IJpma, Maud van den Berg

March 10, 2020

Exercise 1

a)
To randomize 18 slices of loaf into 6 different combinations (3 temperatures, 2 humidities) the following code in R can be used:

```
temp=3; hum=2; N=3
rbind(rep(1:temp,each=N*hum),rep(1:hum,N*temp),sample(1:(N*temp*hum)))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    1    1    1    1    1    1    2    2    2     2     2     2     3     3
## [2,]    1    2    1    2    1    2    1    2    1     2     1     2     1     2
## [3,]   14    2   11   12    7   18    6    1   13    10    16    15     5     3
##      [,15] [,16] [,17] [,18]
## [1,]     3     3     3     3
## [2,]     1     2     1     2
## [3,]     8     4     9    17
```

To interpret this table: each column can be seen as a different unit(where the id of the unit is the value in the third row), the first row can be seen as which temperature the unit has to be measured in and the second row can be seen as the humidity group of the unit.

b)

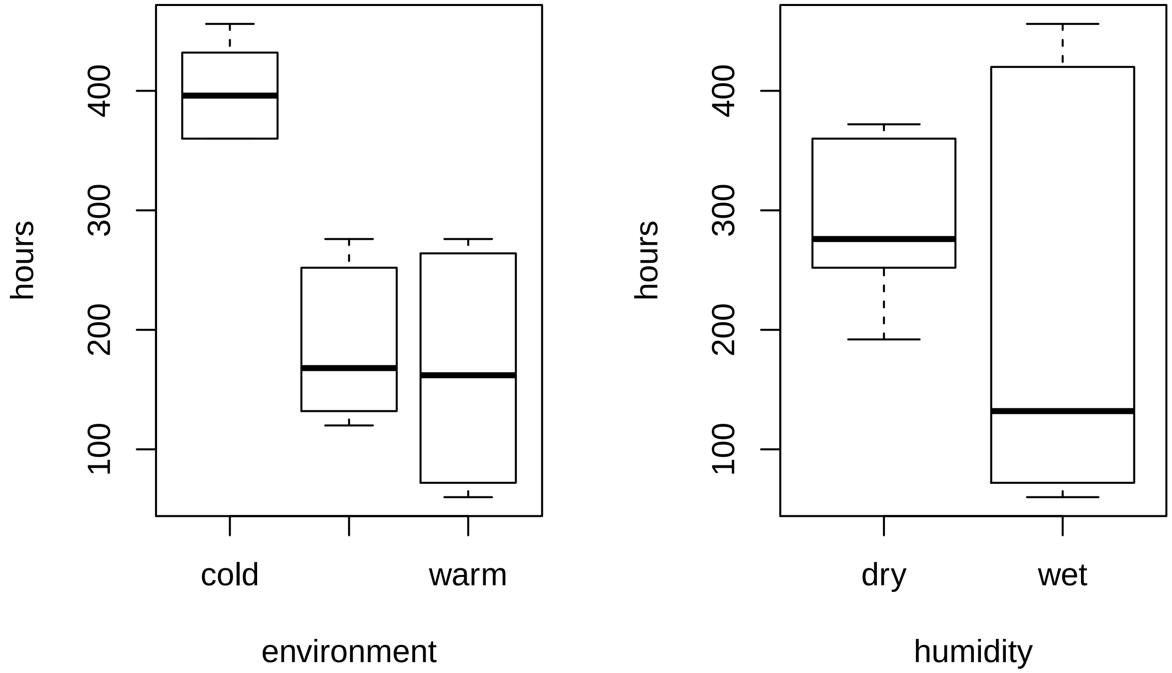


Figure 1:Time to decay against environment and humidity factors

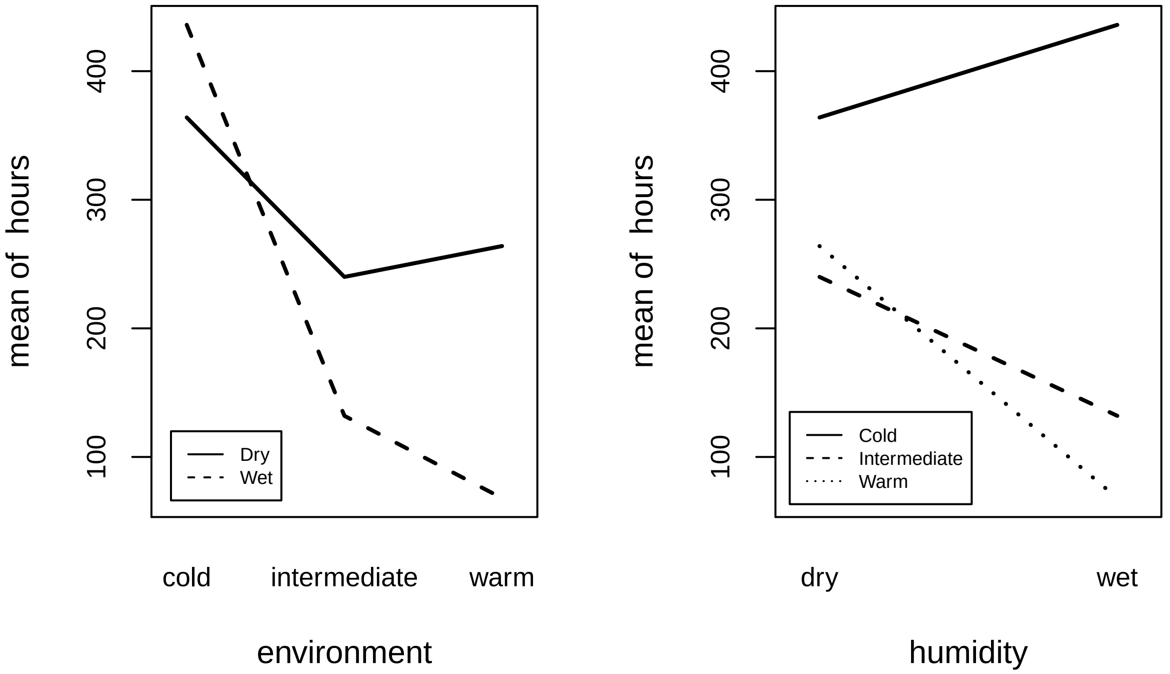


Figure 2: Interaction graph of time to decay against two fixed factors environment and humidity

c)
After performing a one way ANOVA test, the following results were obtained: the factor environment has a significant main effect on the time to decay, F = 233.685, p < .05. Also the type of humidity has a significant main effect on the time to decay, F = 62.296, p < .05. There is a significant effect for the interaction of the factors on the time to decay as well, F = 64.796, p < .05. To measure these results, the assumptions of the ANOVA test are considered to be satisfied. If this is actually the case, is examined in question e.

When inspecting the graphical representation of the interaction effect, the right interaction graph shows the differences clearly. It can be seen that for both the intermediate and warm environment the mean of hours decreases when going from a dry to a wet humidity. For a cold environment this interaction has the opposite direction, namely an increase in mean of hours when the humidity changes from dry to wet. (gem en sd toevoegen?)

d)
When interpreting the results of the ANOVA we see that environment probably(?) has the most influence on the hours it takes the bread to decay. This can be known through the bigger F value of the ANOVA as you can see in question c. Which is explained by a higher value for the explained sum of squares for environment (201904 vs. 26912 for humidity type). This isn't a good question, because there is also a significant effect for the interaction, so these results are based on the situation where both factors are taken into account. In what way the factors have an influence on the time to decay has to be explored in a different experiment.

e)
To check check the normality and the assumption of equal variances of the ANOVA, a QQ-plot of the residuals is computed. In addition also the fitted values are plotted:

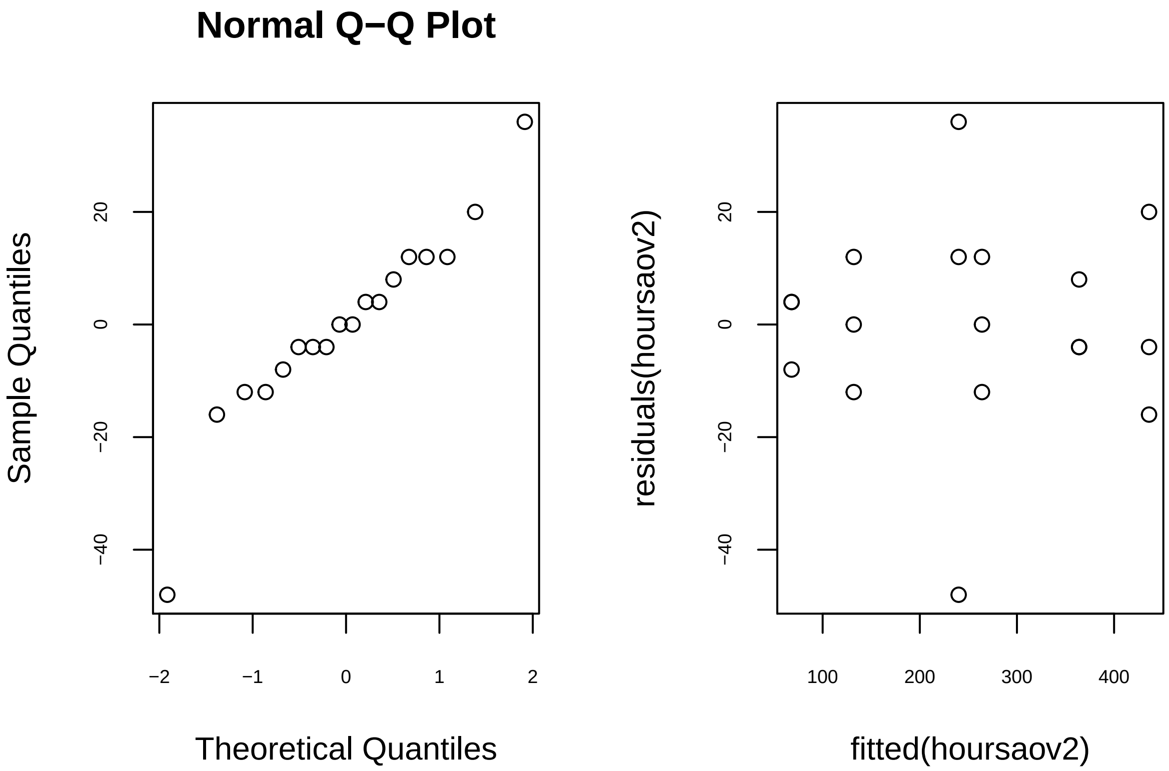


Figure 3: QQ−plot of residuals of anova and fitted residuals

The residuals, the data corrected for the different populationmeans, seem to be normally distributed. Performing the Shapiro-Wilk test confirms this (W = 0.9296, p-value = 0.1911). Also in the fitted data it can be seen that the spread in the residuals doesn't change systematically when the number of hours increases. In the knowledge that the 18 bread units are independently measured, it can be stated that the assumptions for the ANOVA are satisfied. Are there any outliers? heh? je moet juist geen patroon zien toch?

Exercise 2

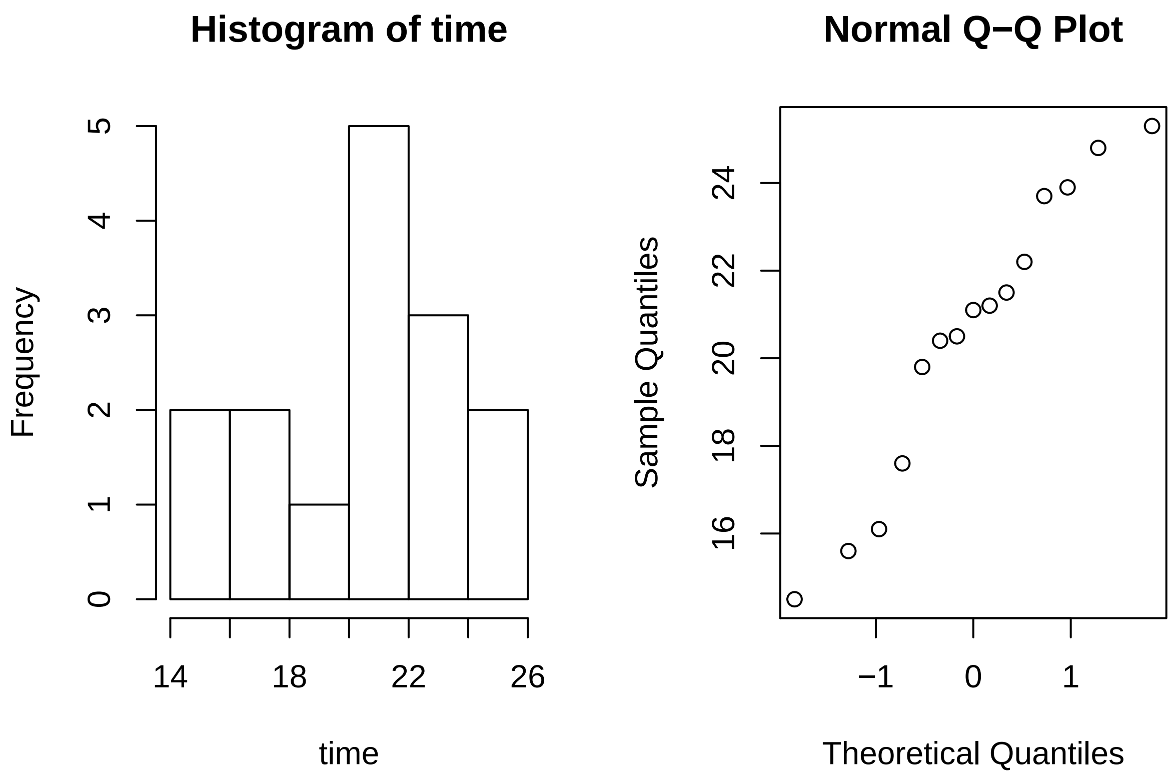
a)

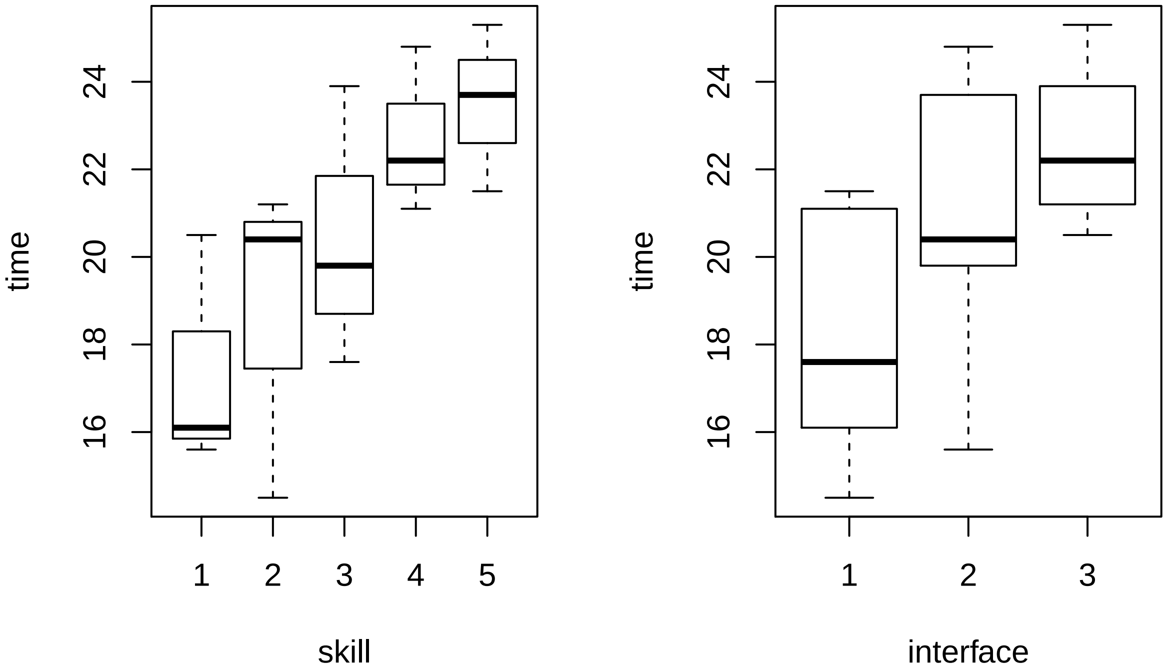
```
rbind(c(replicate(5, sample(1:3))),rep(1:5,each=3), rep(1:15))
```

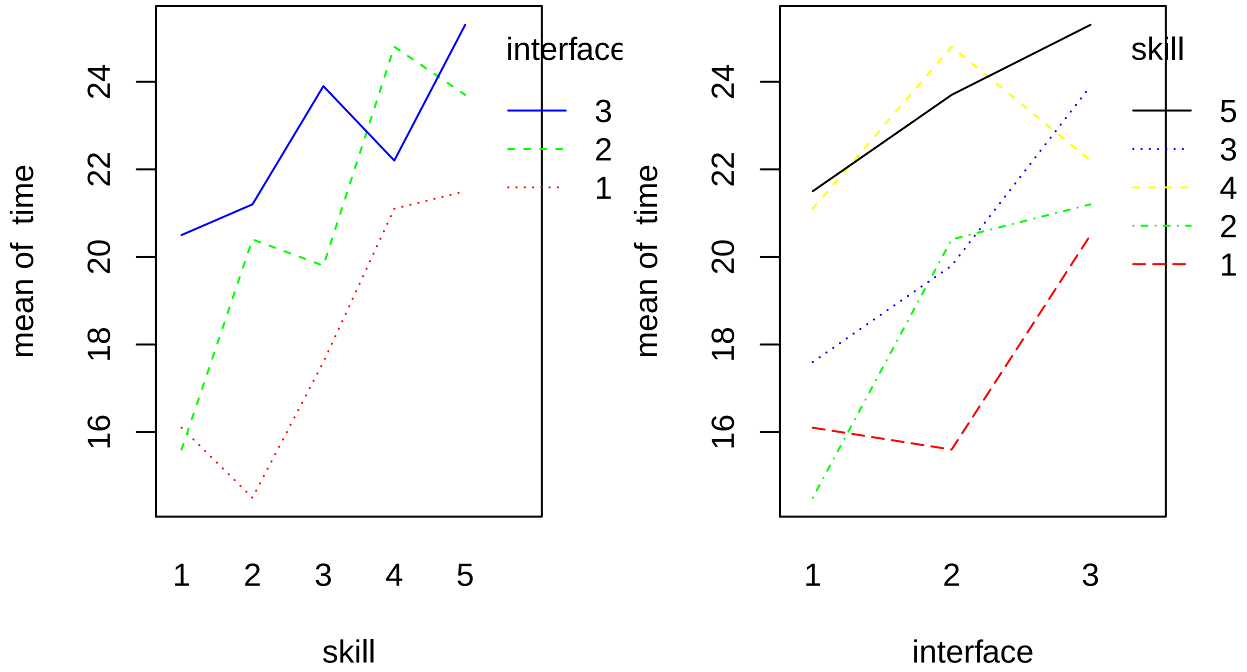
```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    1    3    2    3    1    2    3    2    1     1     3     2     3     2
## [2,]    1    1    1    2    2    2    3    3    3     4     4     4     5     5
## [3,]    1    2    3    4    5    6    7    8    9    10    11    12    13    14
##      [,15]
## [1,]     1
## [2,]     5
## [3,]    15
```

A randomizied block design was used to distribute the 15 students over the 3 possible interfaces. Because both the student ID and the skill level are fixed, only the interface had to be evenly distributed over the students. The code above shows that step. There are three students in each skill-level category, and since there are 3 possible interfaces, every one of those three students should be assigned a different interface. In the output the randomized assignment of students to each interface is shown. The first row indicates the interface, the second row shows the skill level of the students. And the third row represents the ids of the students. Each column corresponds to a student.

b)




time
16
18
20
22
24
1
2
3
4
5
skill
time
16
18
20
22
24
1
2
3
interface



First the distribution of the dependent variable is shown in a histogram and qq-plot. The data looks normally distributed. Next two boxplots are created to visualize the effect that the indpendent variables skill and interface have on the time required to finish the task. Overall the students with a lower skill level (better skills) seem to be faster than students with a high skill rank. Furthermore, the use of interface 1 seems to result in lower times. Lastly two interaction plots are created. These show that there are no obvious interaction effects. However, the previous statements are made based on the visualization of the data. No test have been used to underline these statements. Besides, any possible interaction effects cannot be reliably tested, because there is only one observation for each of factors.

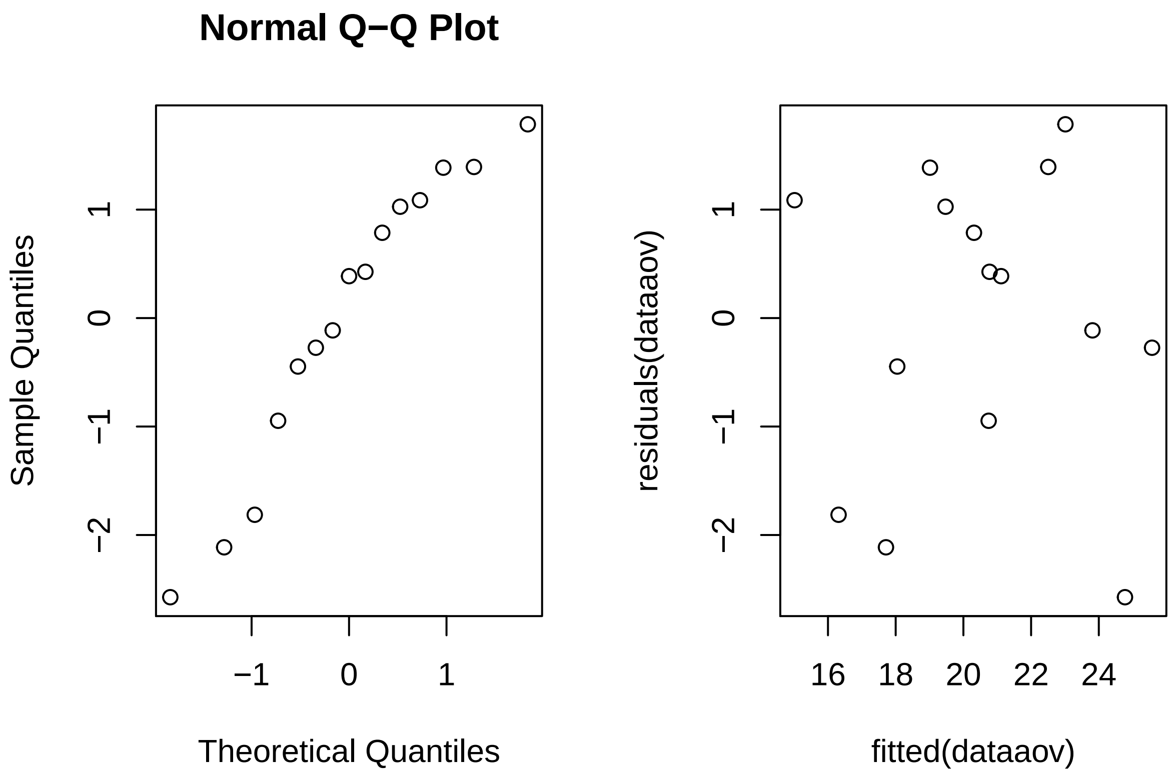
c)

```
anova(dataaov)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value  Pr(>F)
## interface  2 50.465 25.2327  7.8237 0.01310 *
## skill      4 80.051 20.0127  6.2052 0.01421 *
## Residuals  8 25.801  3.2252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To test whether interface has a main effect on the time it takes to complete the task an anova was carried out with both interface and skill as independent variables. No interaction effect was taken into account. Results show that both interface, F(2,8)=7.82, p = 0.013, and skill, F(4,8)=6.21, p = 0.014, have a significant main effect. Therefore the search time is not equal for all of the interfaces. To compute the estimated time it takes for a user with skill level 3 who uses interface 2 the summary of the anova was used. This results in an estimated time of (20.5467 + 0.3133 + -0.1133) 20.7467.

d)

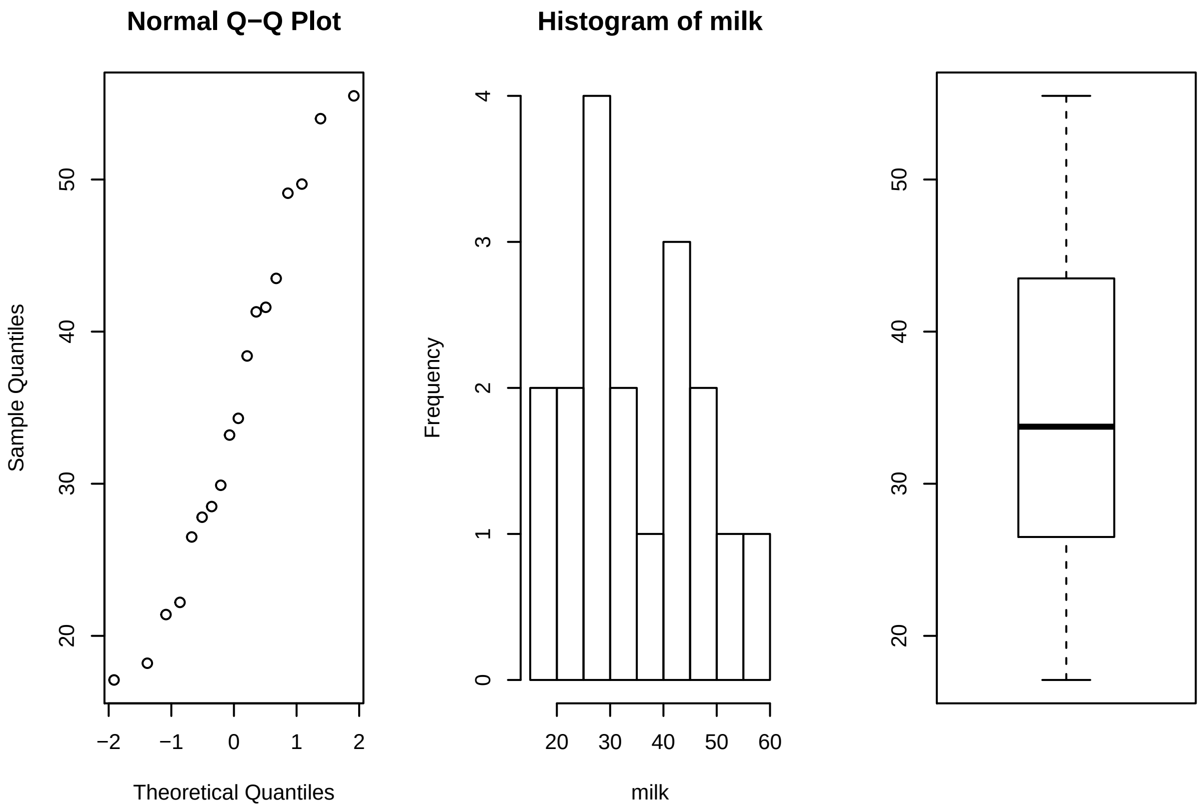


The model assumptions were checked by looking at both the normality of the residuals and a plot of the fitted values against the residuals. The residuals seem to be normally distributed. This was confirmed by a Shapiro-Wilk normality test, W=0.93, p = 0.282. The plot of the fitted values against the residuals shows that there is no systematic error.

e)
A non-parametric Friedman test was carried out to test whether there is an effect of interface. Results show that interface has a significant influence on the time it takes to complete the task, Xˆ2(2)=6.4, p = 0.041.

f)
When carrying out an analysis of variance with only the interface as independent variable, the results indicate that there is no significant effect of interface, F(2,12)=2.86, p = 0.096 on the time it takes to complete the task.It could make sense to carry out this test, however in the current context it is wrong. This is because earlier tests,as well as the visualization of the data indicate a significant effect of skill. Therefore, it is unwise to remove this factor from the analysis. For this test to be valid, tests and visualization should have indicated that there is no effect of skill on the dependent variable. If there is no effect of skill it would not matter if it was used as a factor in the analysis. Since this is not the case, it is unwise to carry out and interpret the results of this analysis.

Exercise 3



```
##
##  Shapiro-Wilk normality test
##
## data:  milk
## W = 0.95421, p-value = 0.4949
```

a)

```
cowaov=lm(milk~treatment+order+per+id,data=cow)
anova(cowaov)
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## treatment  1    0.27    0.27   0.1085  0.751470
## order      1   53.52   53.52  21.5986  0.002349 **
## per        1   25.39   25.39  10.2462  0.015046 *
## id         7 2413.96  344.85 139.1810 5.632e-07 ***
## Residuals  7   17.34    2.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, it is investigated whether the data the outcome y (milk) is normally distributed. According to the QQ-plot, histogram, boxplot it is assumed that the data is normally dsitributed. In addition, a Shapiro-Wilk-test is carried out, which also confirmed a normally distributed data. To test wheter the type of feedingstuff had an effect on the milk production, an ordinary fixed effects model is carried out. From this ANOVA it is concluded that there is a difference in milk production between the two treatments of 1.02 liter milk. However, this result is not significant (p = 0.517), which indicates that there is no effect of type feedingstuff on the milk production.

b)

```
cowaov1 = lmer(milk~treatment+order+per+(1|id),data=cow,REML=FALSE)
cowaov2 = lmer(milk~order+per+(1|id),data=cow,REML=FALSE)
anova(cowaov2 , cowaov1)
```

```
## Data: cow
## Models:
## cowaov2: milk ~ order + per + (1 | id)
## cowaov1: milk ~ treatment + order + per + (1 | id)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## cowaov2  5 117.89 122.34 -53.946   107.89
## cowaov1  6 119.31 124.65 -53.656   107.31 0.5807      1      0.446
```

A mixed effects model is carried out, to investigate if the type of feedingstuff has an effect on the milk production. Now, the effect of cow (i.e. id) is an 'random effect'. To get a Chisquare-value, another ANOVA is carried out, with this model and a model with the factor treatment left out. This resulted in [give chisquare] So, also in this case there is no effect of type feedingstuff on the milk production.

c)

The t-test resulted in P-value = 0.828. This shows that treatment does not have a significant effect on the milk production. This conclusion is compatible with the conclusion from a) because the factor order doesn't matter.

Exercise 4

a)

```
nauseadf = data.frame(naus=c(rep('no', each=180), rep('yes', each=124)),
           medicin=c(rep('chlor', each=100), rep('pent100', each=32), rep('pent150', each=48),
                     rep('chlor', each=52), rep('pent100', 35), rep('pent150', 37)))
attach(nauseadf)
xtabs(~medicin+naus)
```

```
##          naus
## medicin    no yes
##   chlor   100  52
##   pent100  32  35
##   pent150  48  37
```

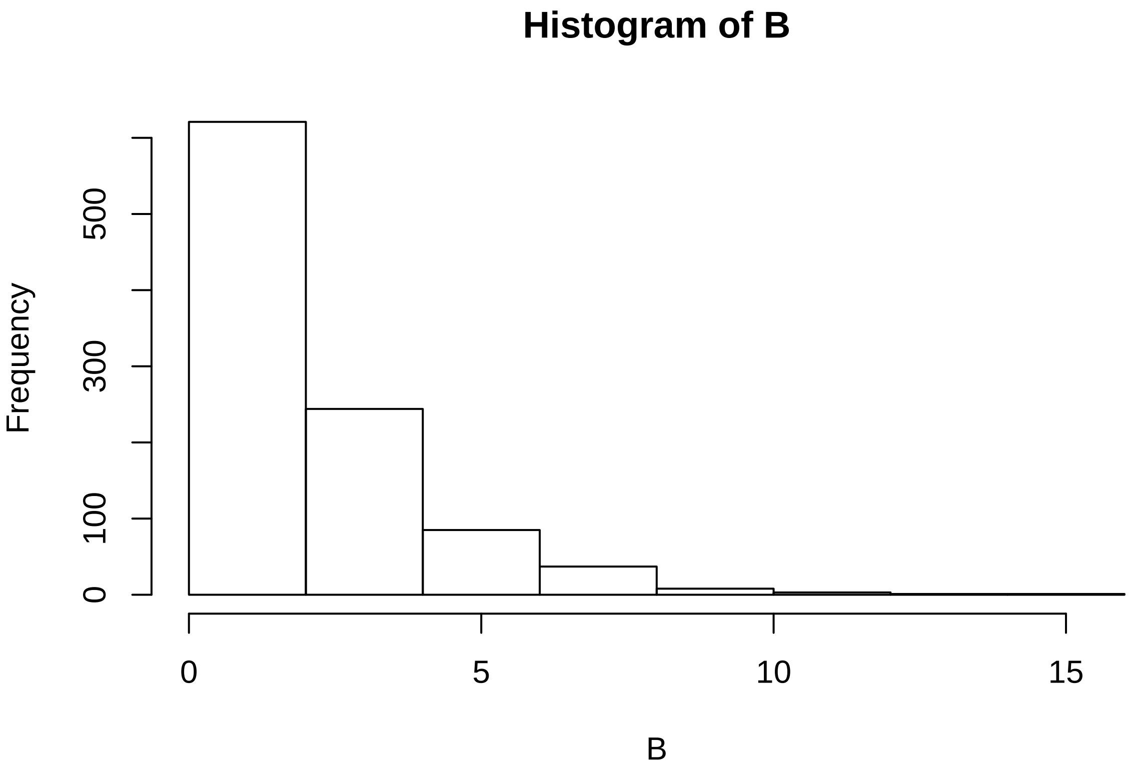
The dataframe is created by first creating a column with the number of patients with nausea (180) and without (124). Next the column for medicin is created. This is done by repeating the type of medicin a specified number of times. The number of repititions is equal to the number of patients that took a certain medicin and did not have nausea.

b)

```
B = numeric(1000)
for (i in 1:length(B)){
  nauseadf2 = transform(nauseadf, medicin=sample(medicin))
  B[i] = chisq.test(xtabs(~nauseadf2$medicin+nauseadf2$naus))[[1]]
}
chi_contingency = chisq.test(xtabs(~medicin+naus))[[1]]
p_permutation = mean(B>chi_contingency)
```

To perform a permutation test, the labels were shuffled 1.000 times. After every shuffle the chi value was computed and saved. If the null-hypothesis is true, then the chi value of the original dataset would be a probable outcome. Therefore the chi value from the original dataset was compared to the chi-values of the 1.000 permutations. The original chi value was 6.62. Only 3.6% of the permuted chi values were larger than 6.62. Therefore the null-hypothesis can be rejected, and the conclusion is that medicin and nausea are not independent.

c)



When performing a chi-square test for contingency tables, the outcome is similar. Medicin and reported nausea seem to be related, Xˆ2(2)=6.62, p = 0.036. This makes sense, because the resulting chi-values of the permutations is more or less equal to the true chi-distribution. This is illustrated by the histogram. With the permutation test, chi-values are computed under the assumption that the null-hypothesis is true. This results in the actual chi-distribution. Therefore, the p-values from permutation and contingency tables are very similar.

Exercise 5

a)

b)

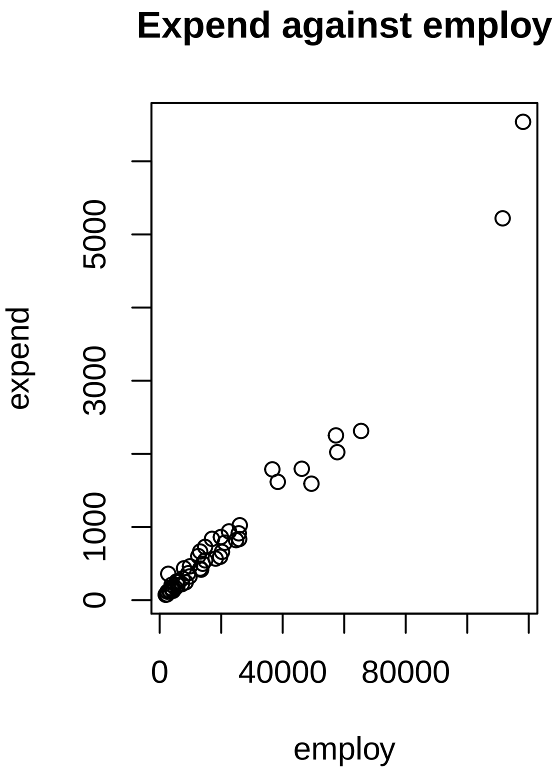
```r
summary(lm(expend~employ+lawyers,data=crimetable))
```

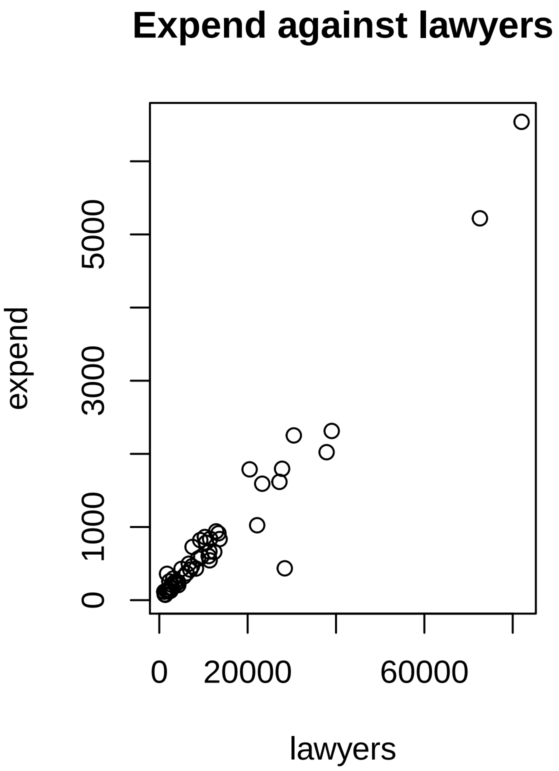
```
##
## Call:
## lm(formula = expend ~ employ + lawyers, data = crimetable)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -599.47  -94.43   36.01   91.98  936.55
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.107e+02  4.257e+01  -2.600  0.01236 *
## employ       2.971e-02  5.114e-03   5.810 4.89e-07 ***
## lawyers      2.686e-02  7.757e-03   3.463  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 232.6 on 48 degrees of freedom
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9616
## F-statistic: 627.7 on 2 and 48 DF,  p-value: < 2.2e-16
```

Both step-up and step-down methods are carried out, to find the best model. The step-up method resulted in the significant explanatory variables 'employ' and 'lawyers'. (waardes geven? F-statistic: 627.7 on 2 and 48 DF, p-value: < 2.2e-16, R-squared: 0.9632). The resulting model of the step-up method is: expend = -1.107e+02 + 2.971e-02*employ - 2.686e-02*lawyers + error.

The step-down method resulted in the same significant explanatory variables, so the two methods yield the same model.

c)


Expend against employ
expend
0
1000
3000
5000
employ
0
40000
80000


Expend against lawyers
expend
0
1000
3000
5000
lawyers
0
20000
60000