EDDA - Assignment 2 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

Exercise 1

Moldy bread If left alone bread will become moldy, rot or decay otherwise. To investigate the influence of temperature and humidity on this process, the time to decay was measured for 18 slices of white bread, which were placed in 3 different environments and humidified or not. The data are given in the filebread.txt, with the first column time to decay in hours, the second column the environment (cold, warm or intermediate temperature) and the third column the humidity.

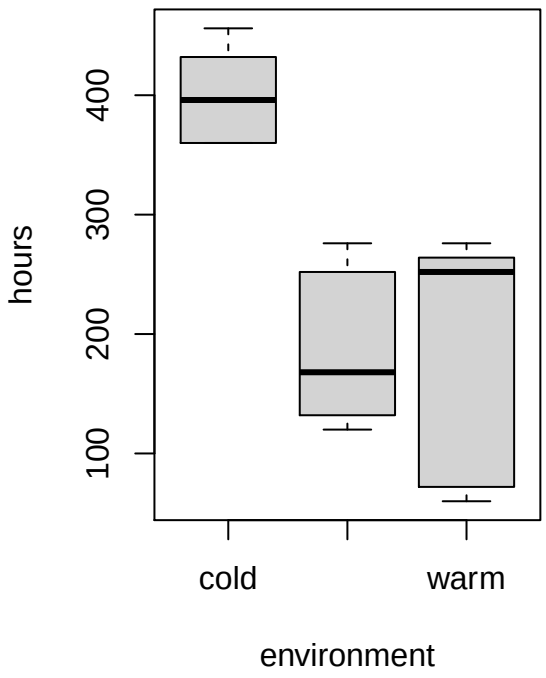
a)The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.

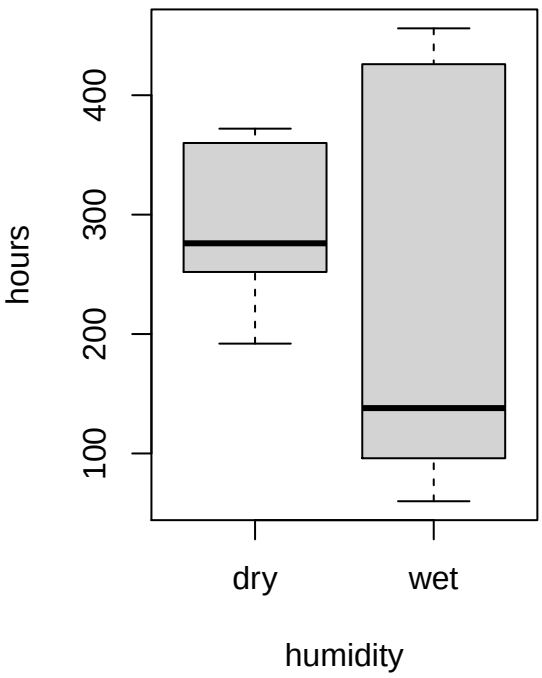
```
data <- read.table(file="data/bread.txt",header=TRUE)
I=3; J=2; N=3
rbind(rep(1:I,each=N*J),rep(1:J,N*I),sample(1:(N*I*J)))
```

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

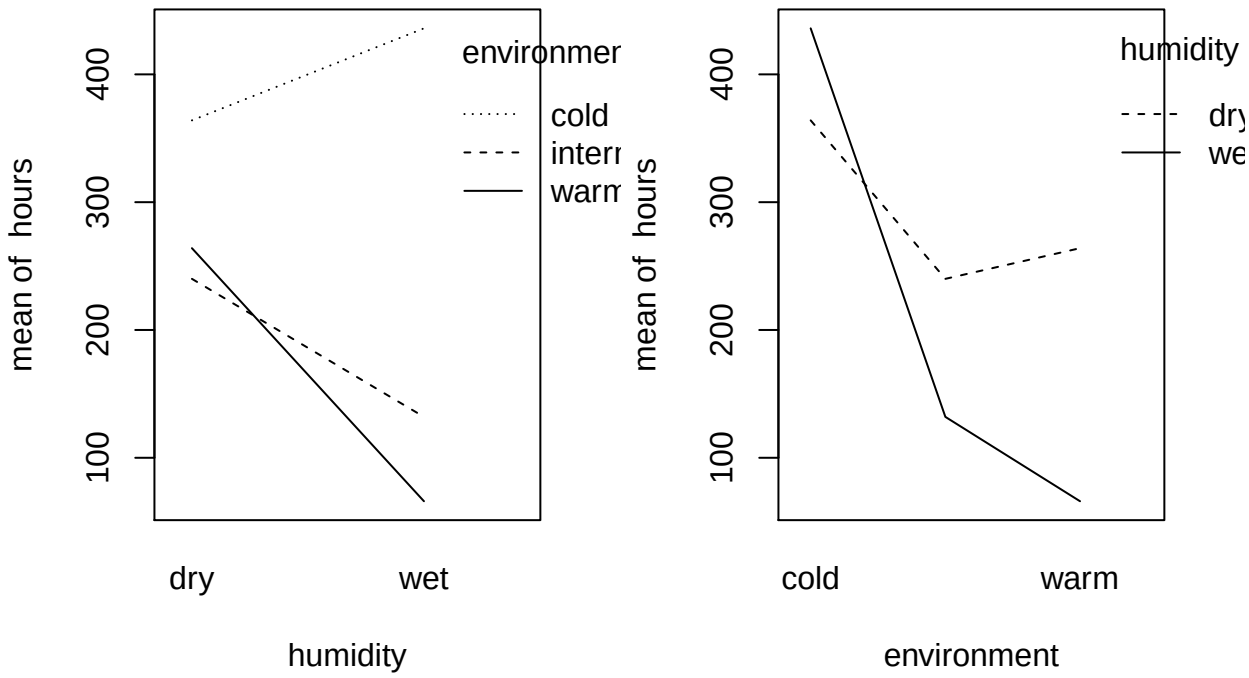
b)Make two boxplots ofhoursversus the two factors and two interaction plots (keeping the two factors fixed in turn).

```
par(mfrow=c(1,2))
attach(data)
boxplot(hours~environment)
boxplot(hours~humidity)
```






```
interaction.plot(humidity,environment,hours)
interaction.plot(environment,humidity,hours)
```



c)Perform an analysis of variance to test for effect of the factors temperature,humidity, and the interaction. Describe the interaction effect in words.

```
data$environment=as.factor(data$environment)
data$humidity=as.factor(data$humidity)
dataaov=lm(hours~humidity*environment)
anova(dataaov)
```

```
## Analysis of Variance Table
##
## Response: hours
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## humidity              1  15162   15162    32.3 0.00014 ***
## environment           2 183791   91895   195.9 2.5e-09 ***
## humidity:environment  2  52071   26036    55.5 1.8e-06 ***
## Residuals            11   5160     469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dataaov)
```

```
##
## Call:
## lm(formula = hours ~ humidity * environment)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##    -48     -6      0     12     36
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           364.0       12.5   29.11  9.3e-12 ***
## humiditywet                            72.0       17.7    4.07  0.00185 **
## environmentintermediate              -124.0       17.7   -7.01  2.2e-05 ***
## environmentwarm                      -100.0       17.7   -5.65  0.00015 ***
## humiditywet:environmentintermediate  -180.0       25.0   -7.20  1.8e-05 ***
## humiditywet:environmentwarm          -270.0       26.5  -10.18  6.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.7 on 11 degrees of freedom
## Multiple R-squared:  0.98,   Adjusted R-squared:  0.971
## F-statistic:  107 on 5 and 11 DF,  p-value: 6.04e-09
```

```
# Without interaction

data$humidity=as.factor(data$humidity)
data$environment=as.factor(data$environment)
dataaov=lm(hours~humidity+environment,data=data)
anova(dataaov)
```

```
## Analysis of Variance Table
##
## Response: hours
##             Df Sum Sq Mean Sq F value  Pr(>F)
## humidity     1  15162   15162    3.44   0.086 .
## environment  2 183791   91895   20.87 8.7e-05 ***
## Residuals   13  57231    4402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the humidity:environment column show a significant p-value which means that the there is evidence for the interaction effect between humidity and environment.

d) Which of the two factors has the greatest (numerical) influence on the decay? Is this a good question?

When we want to know which factor has the greatest influence we want to use the additive model as used above. This shows a p-value of 0.026 for humidity and a p-value of 3.7e-05 for environemnt. This means that the environment has the greatest influence. Is this a good question?

e) Check the model assumptions by using relevant diagnostic tools. Are there any outliers?

```
par(mfrow=c(1,2))
contrasts(data$humidity)=contr.sum; contrasts(data$environment)=contr.sum
dataaov2=lm(hours~humidity*environment,data=data);
summary(dataaov2)
```

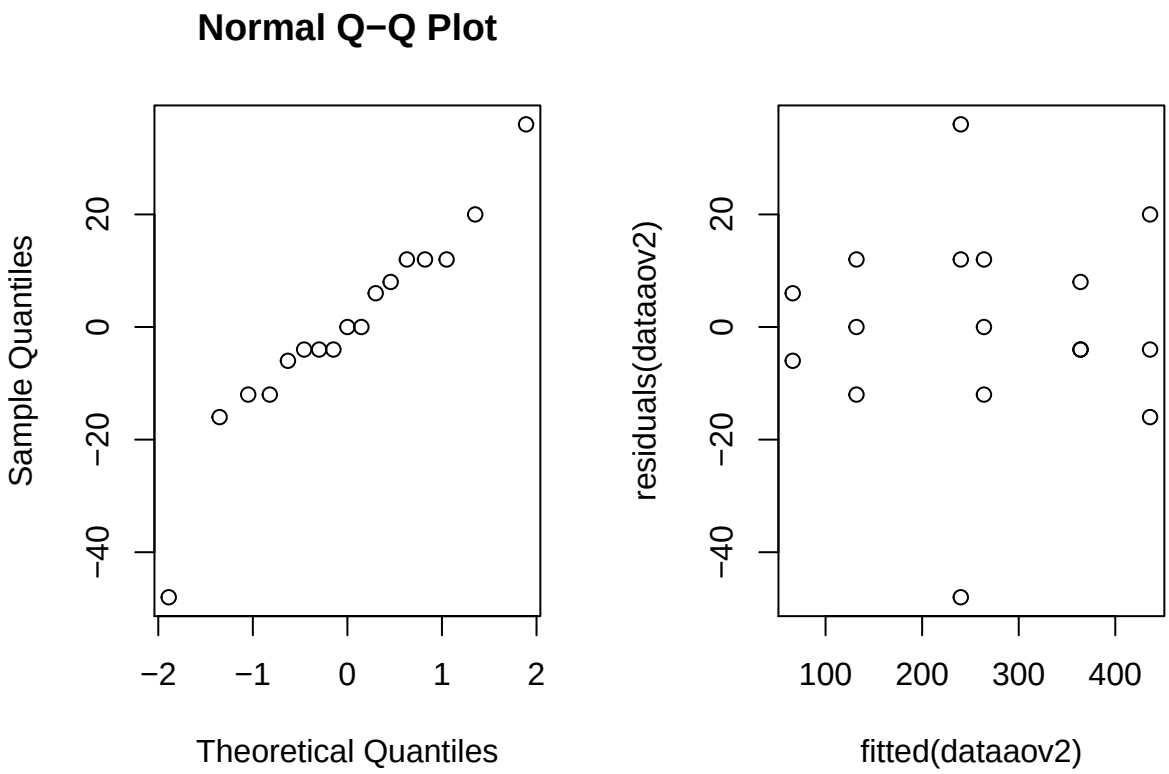
```
##
## Call:
```

```
## lm(formula = hours ~ humidity * environment, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##    -48     -6      0     12     36
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                250.33       5.31   47.11  4.8e-14 ***
## humidity1                   39.00       5.31    7.34  1.5e-05 ***
## environment1               149.67       7.37   20.31  4.5e-10 ***
## environment2               -64.33       7.37   -8.73  2.8e-06 ***
## humidity1:environment1     -75.00       7.37  -10.18  6.2e-07 ***
## humidity1:environment2      15.00       7.37    2.04    0.067 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.7 on 11 degrees of freedom
## Multiple R-squared:  0.98,   Adjusted R-squared:  0.971
## F-statistic:  107 on 5 and 11 DF,  p-value: 6.04e-09
```

```
qqnorm(residuals(dataaov2))
shapiro.test(residuals(dataaov2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(dataaov2)
## W = 0.9, p-value = 0.2
```

```
plot(fitted(dataaov2),residuals(dataaov2))
```



The qqplot shows a somewhat linear line which means that based on the qqplot we can state that the data is normally distributed. Furthermore we used a Shapiro-Wilks test to see if the test can back this assumption. The Shapiro-Wils test showed a p-value of 0.2 which means that the residual data is normally distributed. There is also looked at the spread of the residuals, which showed that there are two outliers around 240.

Exercise 4

Stochastic models for word counts are used in quantitative studies on literary styles. Statistical analysis of the counts can, for example, be used to solve controversies about true author ships. Another example is the analysis of word frequencies in relation to Jane Austen's novel Sanditon. At the time Austen died, this novel was only partly completed. Austen, however, had made a summary for the remaining part. An admirer of Austen's work finished the novel, imitating Austen's style as much as possible. The file austen.txt contains counts of different words in some of Austen's novels: chapters 1 and 3 of Sense and Sensibility(stored in the Sense column), chapters 1, 2 and 3 of Emma(column Emma), chapters 1 and 6 of Sanditon(both written by Austen herself, column Sand1) and chapters 12 and 24 of Sanditon(both written by the admirer,Sand2)

a) Discuss whether a contingency table test for independence or for homogeneity is most appropriate here.

The contingency table test for independence or for homogeneity is approriate because we look at a a count of units in different categories of two factors which are words and story

b) Using the given data set, investigate whether Austen herself was consistent in her different novels. Where are the main inconsistencies?

```
data=read.table(file="data/austen.txt",header=TRUE)
austen = data[,1:3]
```

```
z = chisq.test(austen)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  austen
## X-squared = 12, df = 10, p-value = 0.3
```

```
residuals(z)
```

```
##           Sense   Emma  Sand1
## a       -1.0300 -0.129  1.594
## an       0.4473 -0.159 -0.375
## this     0.0513  0.294 -0.504
## that     0.7482  0.287 -1.442
## with    -0.0475  0.521 -0.704
## without  1.0654 -1.588  0.893
```

She is not inconsistent as the p-value is above 0.05. The main inconsistency where the words "a", "that" and "without"

```
z = chisq.test(data)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 46, df = 15, p-value = 6e-05
```

```
residuals(z)
```

```
##           Sense      Emma  Sand1   Sand2
## a       -1.015 -0.112093  1.606 -0.0589
## an      -0.591 -1.219955 -1.067  3.7282
## this     0.139  0.390490 -0.444 -0.3267
## that     1.594  1.179849 -0.910 -3.0493
## with    -0.512  0.000192 -1.025  1.7482
## without  1.392 -1.341196  1.137 -1.0696
```

The fan is inconsistent as the p-value of the test is below 0.05. The main inconsistencies were for the words "that" and "an".