Solution to Series 1

1. An experiment was conducted to study the effect of 3 types of diets on the coagulation time of blood. In the first part we will load and plot the data. In the second part we carry out an analysis of variance step by step and in the last part we will check the model assumptions.

a) Loading the data and exploratory plots

1. The results of the experiments are saved in the file `blood.csv`, which you can find on the course webpage. Open it with a spreadsheet reader (like Excel) and look at the data. Then import them in R using the function `read.csv` or with the importing tool of RStudio (see upper right panel).

```
blood <- read.csv(file='blood.csv')
```

2. How many patients are there? how many patients were assigned to each diet type?

```
## number of patients:
nrow(blood)
```

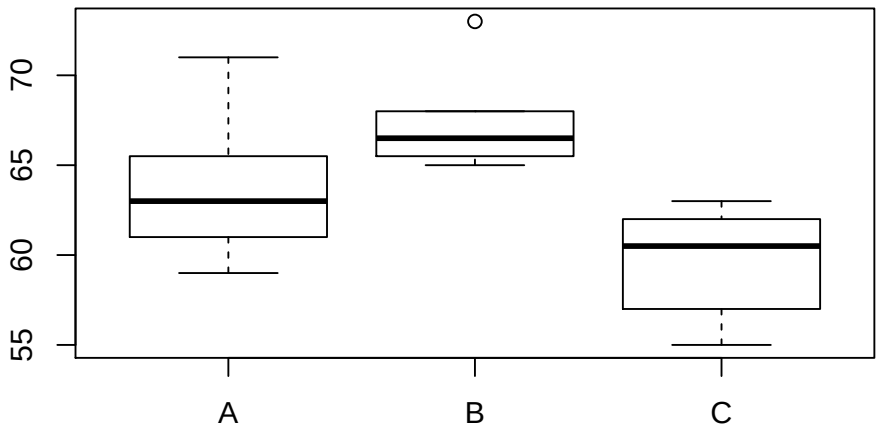
```
[1] 24
```

```
## for each group:
table(blood$diet)
```

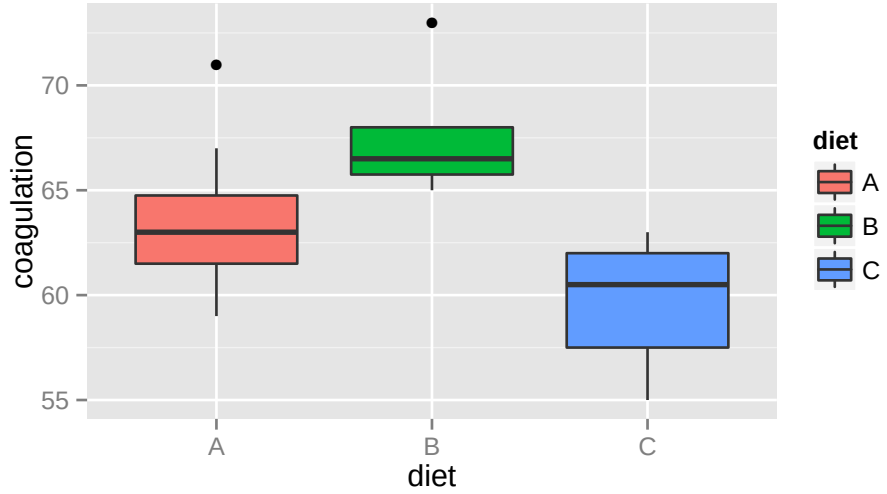
```
A B C
8 8 8
```

3. *Boxplot*

```
boxplot(coagulation~diet, data=blood)
```

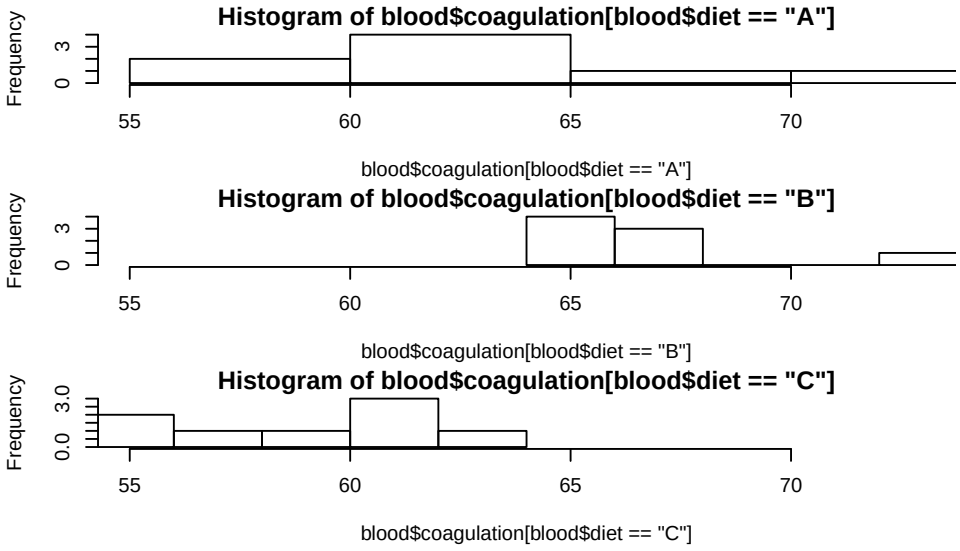



```
require(ggplot2)
g <- ggplot(data=blood, aes(x=diet, y=coagulation))
g <- g + geom_boxplot(aes(fill=diet))
g
```

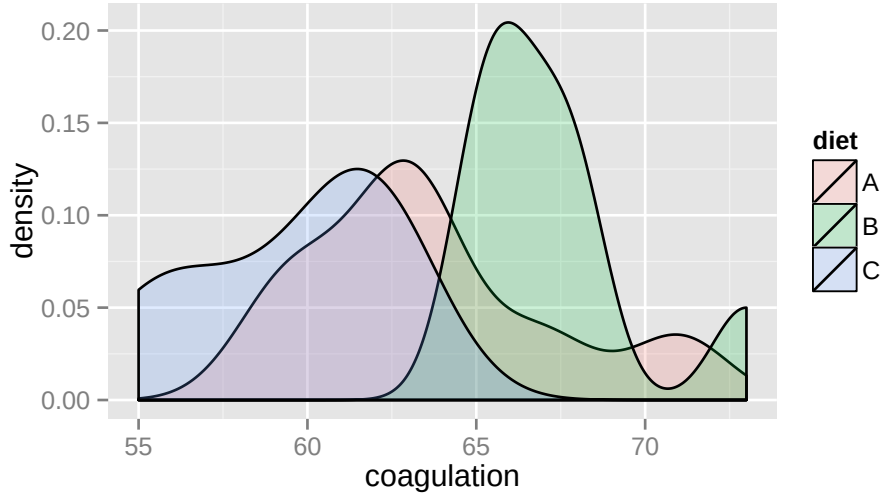


4. *Density plot*

```
interval <- range(blood$coagulation) ## calculate min and max
par(mfrow=c(3,1)) # split the window in 3 rows and 1 column
## set xlim the same for all plots for better comparison:
hist(blood$coagulation[blood$diet=='A'], xlim=interval)
hist(blood$coagulation[blood$diet=='B'], xlim=interval)
hist(blood$coagulation[blood$diet=='C'], xlim=interval)
par(mfrow=c(1,1)) # reset default value
```




```
g <- ggplot(data=blood, aes(x=coagulation, fill=diet))
g <- g + geom_density(alpha=.2) ## alpha sets the transparency
g
```



5. From these plots, do you think that the type of diet followed has a significant effect on the coagulation of blood?
   Yes it seems that the diet B results in clearly bigger coagulation. Furthermore the variance within group is smaller than the variance between group, thus the effect is expected to be significant.

b) We will now compute individually all the elements needed and perform a one-way ANOVA.

1. Compute the overall mean.

```
mtot <- mean(blood$coagulation)
mtot
```

```
[1] 63.5
```

2. Compute the mean for each group separately.

```
ma <- mean(blood$coagulation[blood$diet=='A'])
mb <- mean(blood$coagulation[blood$diet=='B'])
mc <- mean(blood$coagulation[blood$diet=='C'])
cat(ma,mb,mc)
```

```
63.625 67.25 59.625
```

3. Write down the linear model we are interested in and the estimate of the effects $\hat{A}_i$ if we take the convention that $\sum_i \hat{A}_i = 0$.
   $Y_{ij} = \mu + A_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ $\hat{A}_1 = 0.125$, $\hat{A}_2 = 3.75$ and $\hat{A}_3 = -3.875$
4. Compute the variance *within* group (MS_{res}).
   Hint: Compute first the variance of each group separately and then combine them appropriately (is it a balanced design?).

```
va <- var(blood$coagulation[blood$diet=='A'])
vb <- var(blood$coagulation[blood$diet=='B'])
vc <- var(blood$coagulation[blood$diet=='C'])
MS_res <- mean(c(va,vb,vc))
MS_res
```

```
[1] 10.15476
```

5. Compute the variance *between* groups.
   Hint: Compute first the SS_{treat}, using the appropriate formula and then the MS_{treat} with the correct *df*.

```
## the factor 8 is because there are 8 persons by group
SS_treat <- 8*sum( (c(ma, mb, mc)- mtot)^2 )
MS_treat <- SS_treat / 2 # 2 is the df
MS_treat
```

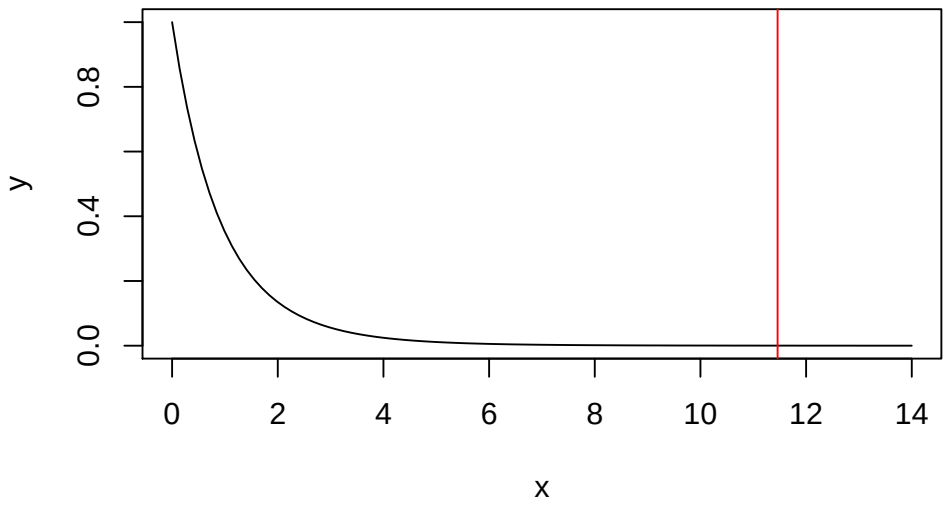
```
[1] 116.375
```

6. Compute the F-statistics. Can you make any conclusion from its value?
   Optional: Plot the probability distribution function of the F distribution with the correct degrees of freedom and look where the computed F-statistics falls (use the function `df` to compute the pdf of the F distribution).

```
F_stat <- MS_treat/MS_res
F_stat
```

```
[1] 11.46014
```

```
x <- seq(0,14,length.out = 100)
y <- df(x, df1=2, df2=21)
plot(x,y, type='l')
abline(v=F_stat, col='red')
## we see that the F-statistics is far, far in the tail...
```



7. What is the probability to obtain such a F-statistics if the null hypothesis was true? Hint: the function `pf` allows you compute the distribution function of the F distribution for any quantile. Use the correct degrees of freedom!

```
pvalue <- 1 - pf(F_stat, df1 = 2, df2= 21)
pvalue
```

```
[1] 0.0004318361
```

8. Is the effect of diet significant?
   Yes it is significant at a 5% level (also at 1%)
9. In practice you don't have to calculate everything by hand like that! Repeat the analysis using the function `lm` to fit the model and `Anova` from the package `car` to calculate the ANOVA table. You could also use `anova` (without capital A) from R base, but it gives inconsistent results when dealing with unbalanced designs and so it is not advised to use it.
   The results should be the same as you obtained before, of course!

```
require(car)
mod <- lm(coagulation~diet, data=blood)
Anova(mod)
```

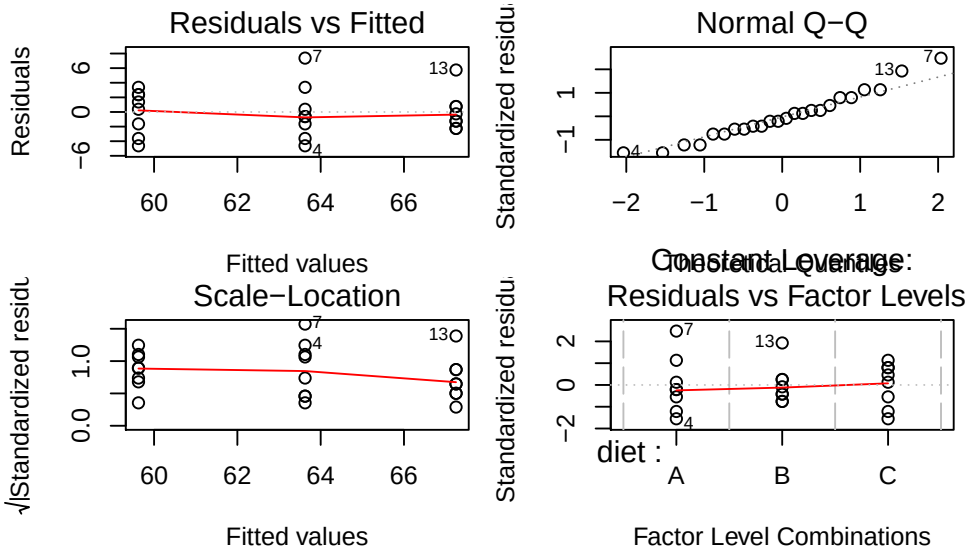
```
Anova Table (Type II tests)

Response: coagulation
          Sum Sq Df F value    Pr(>F)
diet      232.75  2   11.46 0.0004318 ***
Residuals 213.25 21
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) We will now check the model assumptions.

1. What are the model assumptions to test?
   The linear model holds: expectaction of residuals should be 0.
   The errors are normally distributed with mean 0.
   The variance is the same within each group.

2. Perform diagnostic plots. Hint: `par(mfrow=c(2,2))` splits your screen in 4 parts.

```
par(mfrow=c(2,2))
plot(mod)
```



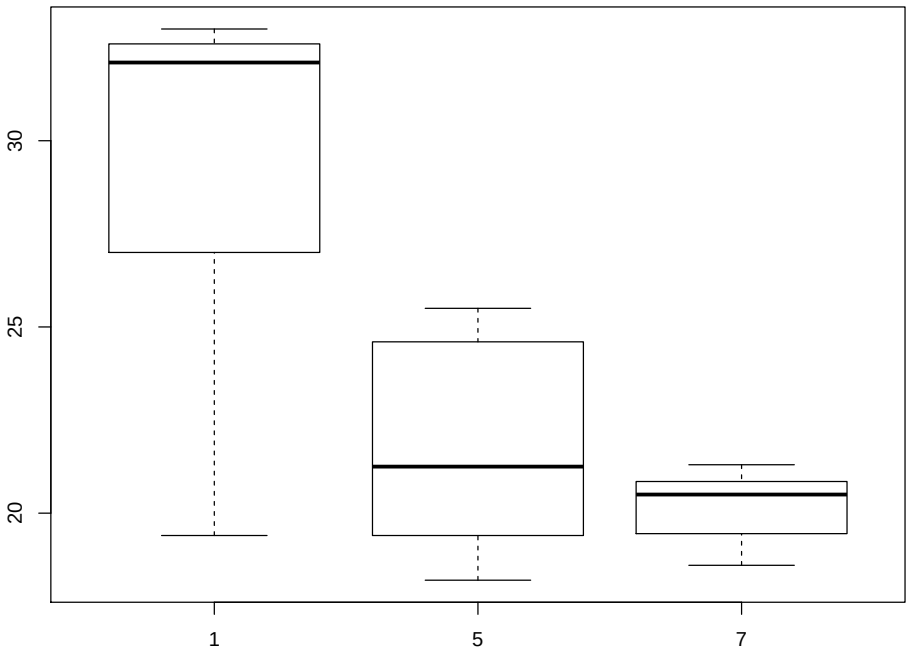
3. Interpret the plots. There is no apparent problem with the data, no outliers or particular skewness. The observation 7 could be seen as an outlier, but it's probably not (remember we deal here with a small sample size and variations are expected to occur naturally).

2. Read in the data:

```
N2 <- c(19.4,32.6,27,32.1,33,18.2,24.6,25.5,19.4,21.7,20.8,20.7,
        21,20.5,18.8,18.6,20.1,21.3)
strain <- c(1,1,1,1,1,5,5,5,5,5,5,7,7,7,7,7,7,7)
r.data <- data.frame(cbind(N2,strain))
r.data$strain <- as.factor(r.data$strain)
```

a) Plot the data.

```
plot(r.data$strain,r.data$N2)
```



The variance between strains looks larger then the variance within strains. This could be an indicator for a significant difference of nitrogen contents for different Rhizobium strains.

b) Carry out an analysis of variance.

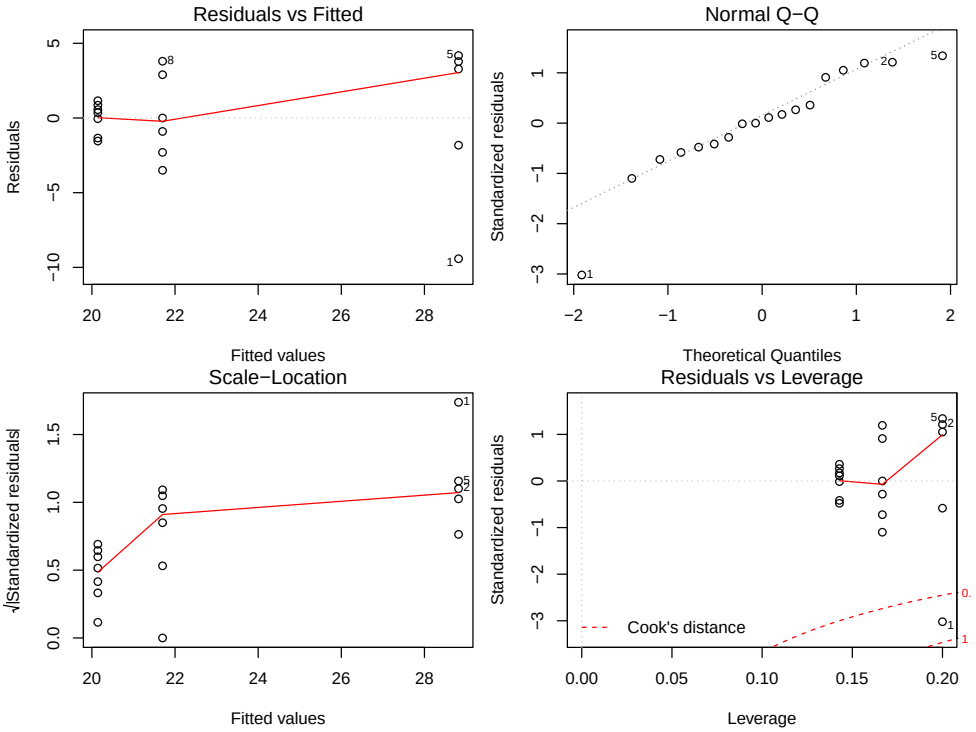
```
fit.n2 <- aov(r.data$N2 ~ r.data$strain)
summary(fit.n2)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
r.data$strain  2  236.6  118.28   9.723 0.00196 **
Residuals     15  182.5   12.16
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-value equals 9.72. By looking at the P-value we see that there are significant differences in nitrogen contents for different strains of Rhizobium.

c) Check the model assumptions.

```
par(mfrow=c(2,2))
plot(fit.n2)
```

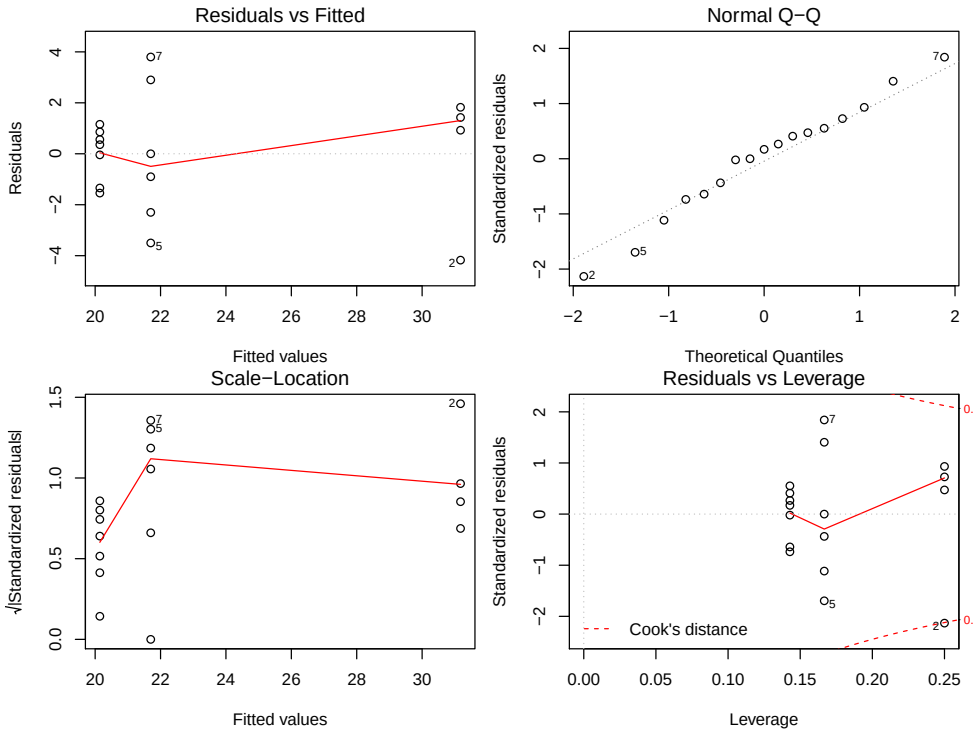


From the diagnostic plots we see that there exists an outlier. On the basis of the plots, observation number 1 can be clearly identified as an outlier. After removing the outlier we repeat the analysis.

```
rr.data <- r.data[-1,]
fit.n2mod <- aov(rr.data$N2~rr.data$strain)
summary(fit.n2mod)
```

```
               Df Sum Sq Mean Sq F value   Pr(>F)
rr.data$strain  2  333.2  166.60    32.6 5.39e-06 ***
Residuals      14   71.5    5.11
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(fit.n2mod)
```



We see that now the model assumptions are fulfilled.