# Exercise Anova and Tukey test on Rivers data

*Emil Brohus*

*25 February 2016*

I found the way of assigning factor levels as shown during the lecture with the g vector of 1, 2 and 3s to be a bit laborious. However I could not get the aov and Tukey to work if not all the data was in one vector.

```
data = read.table(url('https://raw.githubusercontent.com/EmBro/MDA/master/rivers.csv'), dec = ',', sep

show(data)
```

```
##    Varde  Ribe Skjern
## 1  166.5 155.7  145.1
## 2  163.2 141.8  143.6
## 3  146.1 140.9  139.8
## 4  142.4 129.1  131.4
## 5  159.5 132.7  133.2
## 6  151.3 140.6  138.1
## 7  136.6 137.0  133.6
## 8  148.6 124.8  128.2
## 9  151.9 154.6  144.1
## 10 155.4 135.4  137.9
```

I tried to rearrange in many ways, but found out in the end that by using the 'stack' function, it did all the work for me.

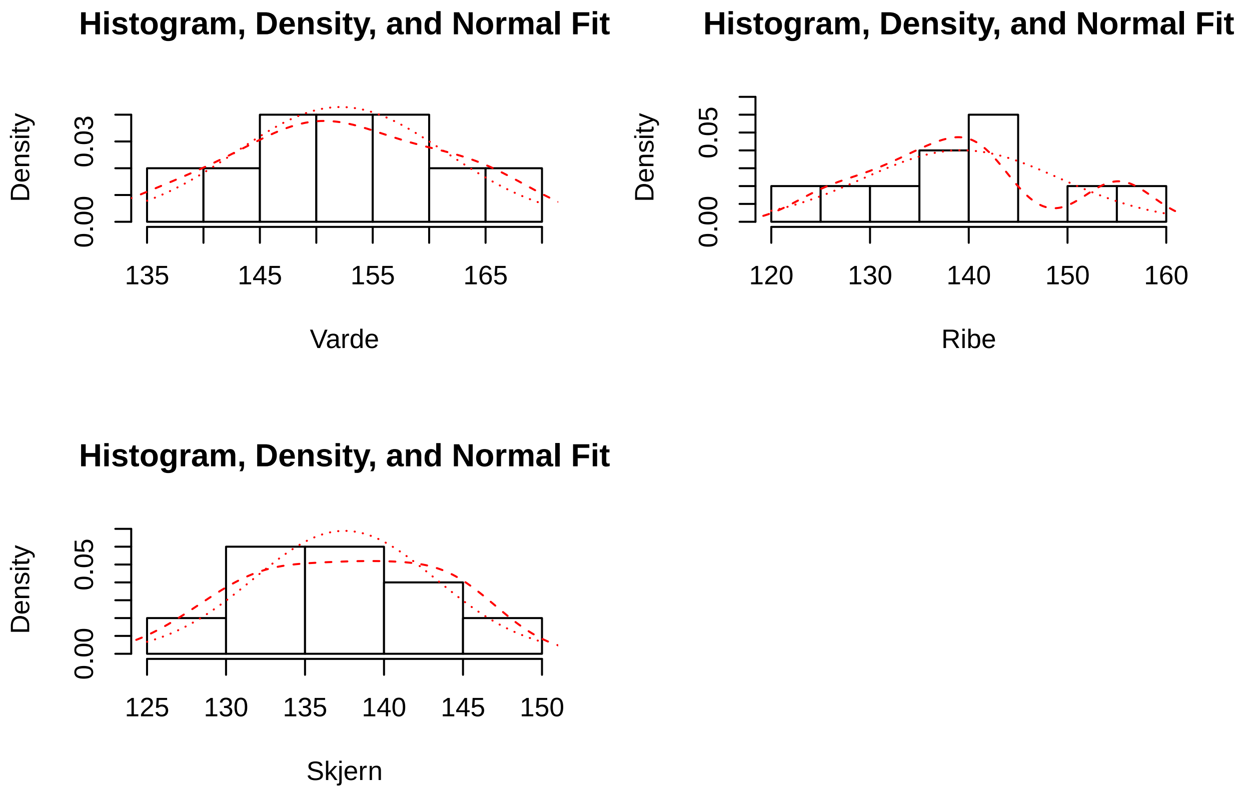
```
data1 = stack(data)
show(data1)
```

```
##    values    ind
## 1   166.5  Varde
## 2   163.2  Varde
## 3   146.1  Varde
## 4   142.4  Varde
## 5   159.5  Varde
## 6   151.3  Varde
## 7   136.6  Varde
## 8   148.6  Varde
## 9   151.9  Varde
## 10  155.4  Varde
## 11  155.7   Ribe
## 12  141.8   Ribe
## 13  140.9   Ribe
## 14  129.1   Ribe
## 15  132.7   Ribe
## 16  140.6   Ribe
## 17  137.0   Ribe
## 18  124.8   Ribe
## 19  154.6   Ribe
## 20  135.4   Ribe
## 21  145.1 Skjern
## 22  143.6 Skjern
## 23  139.8 Skjern
## 24  131.4 Skjern
## 25  133.2 Skjern
## 26  138.1 Skjern
## 27  133.6 Skjern
## 28  128.2 Skjern
## 29  144.1 Skjern
## 30  137.9 Skjern
```

Data is now ready to be analyzed with boxplot, Anova and Tukey test

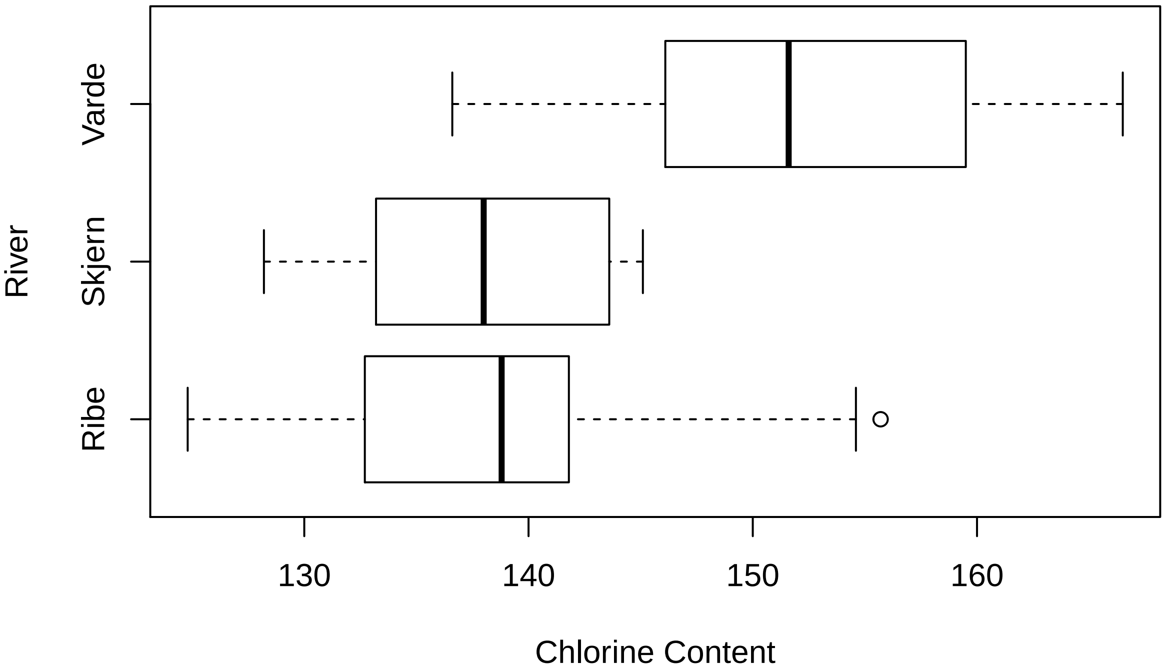
But before, and just for the fun of it, I have decided to see normality with histograms. I found this need little package that does exactly that

```
library(psych)
multi.hist(data, freq = F, dcol = 'red')
```



It looks fairly normal distributed, however Ribe a little odd. I proceed with boxplot, Anova and tukey test

```
boxplot(data1$values ~ data1$ind, horizontal = T,
            xlab = 'Chlorine Content', ylab = 'River')
```



The boxplot gives us an indication about the differences, however we cannot conclude anything certain. It also reveals a slight outlier in the sample set of Ribe River. We proceed with the anova test.

```
r = aov(data1$values ~ data1$ind)
summary(r)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## data1$ind    2   1280   639.8   8.755 0.00117 **
## Residuals   27   1973    73.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test tells that within a 95% Confidence interval that there is a difference in the level of Chlorine in the rivers. However the anova test doesn't tell which one. Therefore the Tukey test

```
Tukey = TukeyHSD(r)
show(Tukey)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$values ~ data1$ind)
##
## $`data1$ind`
##                diff        lwr       upr     p adj
## Skjern-Ribe   -1.76 -11.238861  7.718861 0.8902165
## Varde-Ribe    12.89   3.411139 22.368861 0.0062140
## Varde-Skjern 14.65   5.171139 24.128861 0.0019298
```

The result of the Tukey test allows us to conclude, that there is a difference in the mean level of chlorine in the rivers Varde-Skjern and Varde-Ribe. Since everything is different, there is a difference en mean chlorin content in rivers Skjern-Ribe, however in this sample it is not significant.