ANOVA: introduction

The Analysis of Variance (ANOVA) is used to analyse the influence of a discrete variable having multiple factors on a continuous independent variables. We are here just concerned with the One-Way ANOVA, which can be thought of as a generalization of the t-test.

The model is:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where the parameter μ is the overall mean across all factors, α_i and β_j are main effects, within and between factors respectively and eventually ϵ_{ij} is the residual part of the variance that connot be explained by other parameters, that is some kind of error term, obeying a Normal distribution.

'iris' dataset

iris: dataset of 150 observations x 5 variables.

Sepal.Length: continuous variable
Sepal.Width: continuous variable
Petal.Length: continuous variables
Petal.Width: continuous variable
Species.: discrete variables, setosa, virginica, versicolor

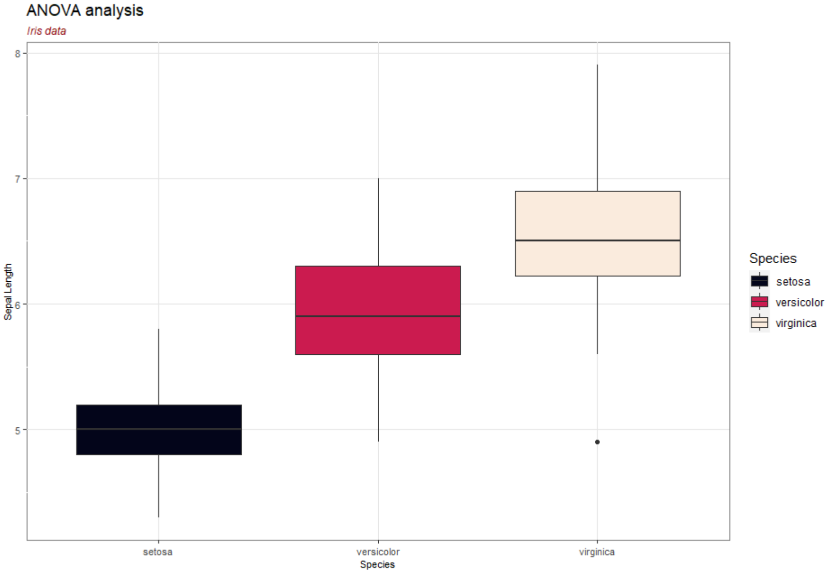
```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Summary

In order to perform a One-Way ANOVA, we call the function 'aov()' on the continuous variable, linked to the discrete variable with means of the operator ' ', like when fitting a linear model. We then call the function 'summary()' on this object to access the summary. The main interpretations are given in the next slides.

```
> # load dataset
> data(iris)
>
> # perform ANOVA
> anova_result = aov(Sepal.Length ~ Species, data=iris)
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
     1
```

ANOVA Boxplots



Post-hoc test and plot

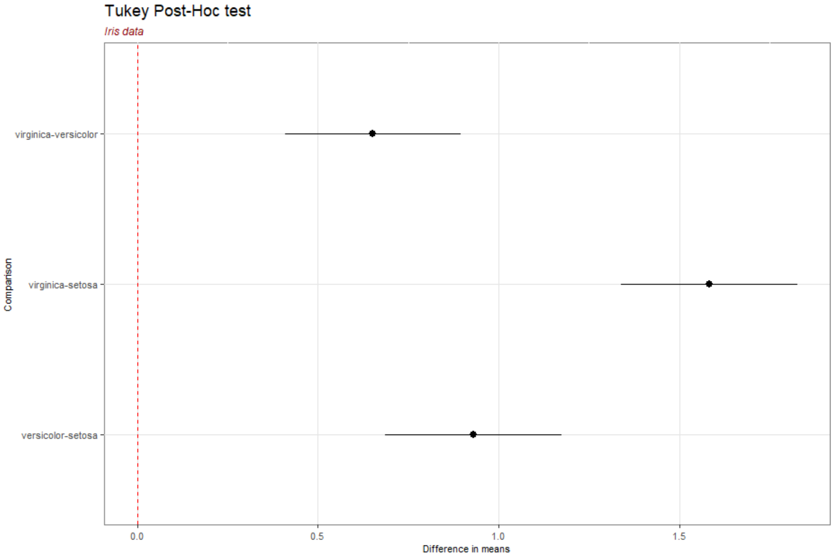
Post-Hoc tests, here we only consider the Tukey test, are used after the initial analysis in order to have an idea as to which factor(s) differs significantly from the other and thus cannot be assumed to belong to the same general population.

```
> TukeyHSD(anova_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sepal.Length ~ Species, data = iris)

$Species
                      diff       lwr       upr p adj
versicolor-setosa    0.930 0.6862273 1.1737727     0
virginica-setosa     1.582 1.3382273 1.8257727     0
virginica-versicolor 0.652 0.4082273 0.8957727     0
```

Tukey Post-Hoc test plot



Main observations

- Null Hypothesis: The mean sepal lengths of the three iris species are equal.

- F-Statistic and P-Value: The F-statistic is 119.26, and the p-value is less than $2e - 16$.

- Significant Result: The p-value is much smaller than 0.05, so we reject the null hypothesis that all petal length are about equal.

- Post-Hoc Analysis: Significant differences exist among the species, and a post-hoc test can pinpoint which species differ.

References

Bijma, F., Jonker M., Van der Vaart, A. (2016), An Introduction to Mathematical Statistics. Amsterdam University Press. ISBN 978 94 6298 5100

The R Project for Statistical Computing: https://www.r-project.org/