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

```
1 iris.head(6)
2
3 sepal_length sepal_width petal_length petal_width species
4 0 5.1 3.5 1.4 0.2 setosa
5 1 4.9 3.0 1.4 0.2 setosa
6 2 4.7 3.2 1.3 0.2 setosa
7 3 4.6 3.1 1.5 0.2 setosa
8 4 5.0 3.6 1.4 0.2 setosa
9 5 5.4 3.9 1.7 0.4 setosa
```

Summary

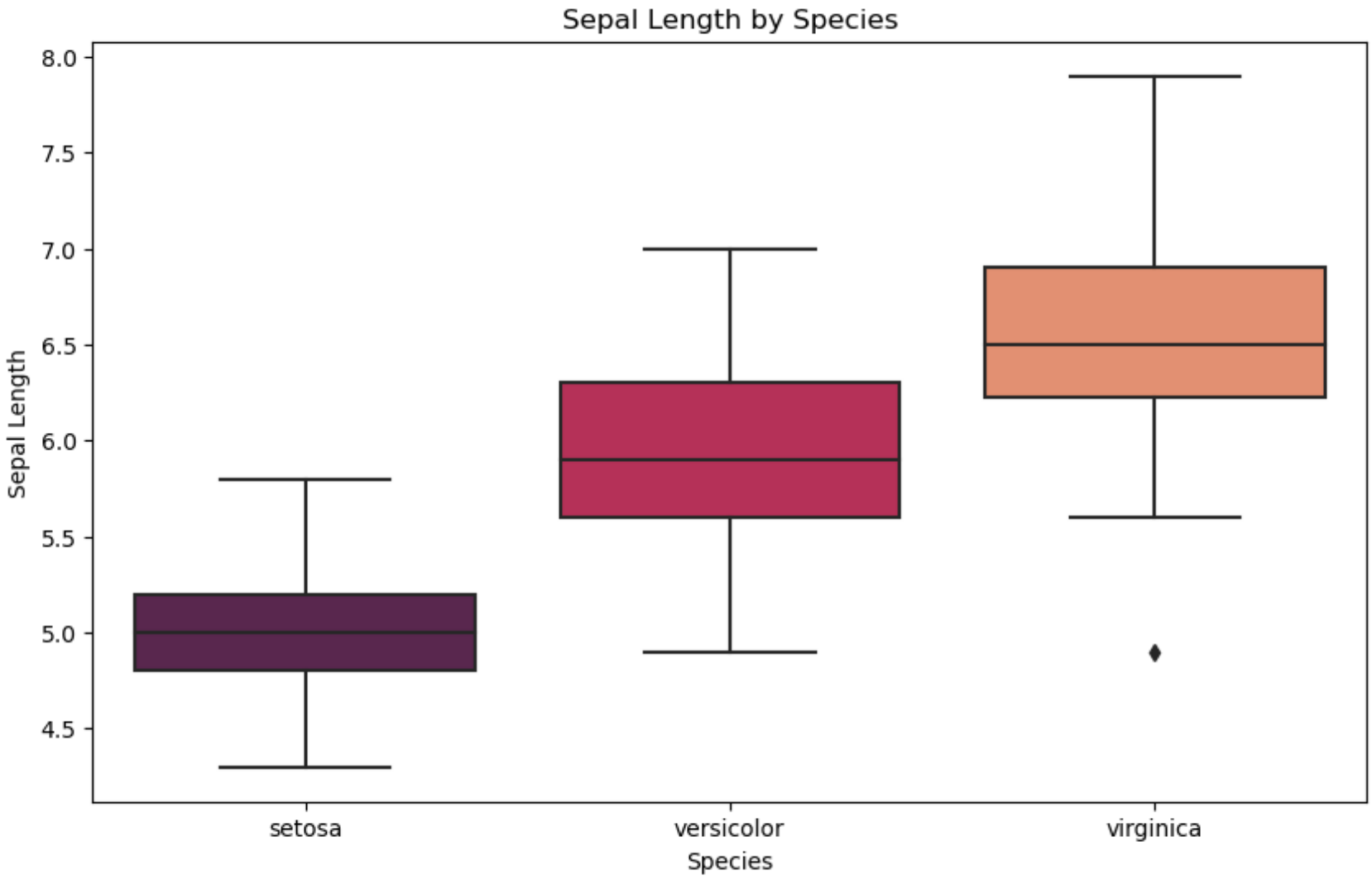
In order to perform a One-Way ANOVA, we use $'sm.stats.anova_lm()'$ from the library statsmodels. There are other ways to perform a One-Way ANOVA. The main interpretations are given in the next slides.

```
# Load the iris dataset
iris = sns.load_dataset("iris")

# Perform one-way ANOVA
model = ols('sepal_length ~ C(species)', data=iris).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table

sum_sq df F PR(>F)
C(species) 63.212133 2.0 119.264502 1.669669e-31
Residual 38.956200 147.0 NaN NaN
```

ANOVA Boxplots



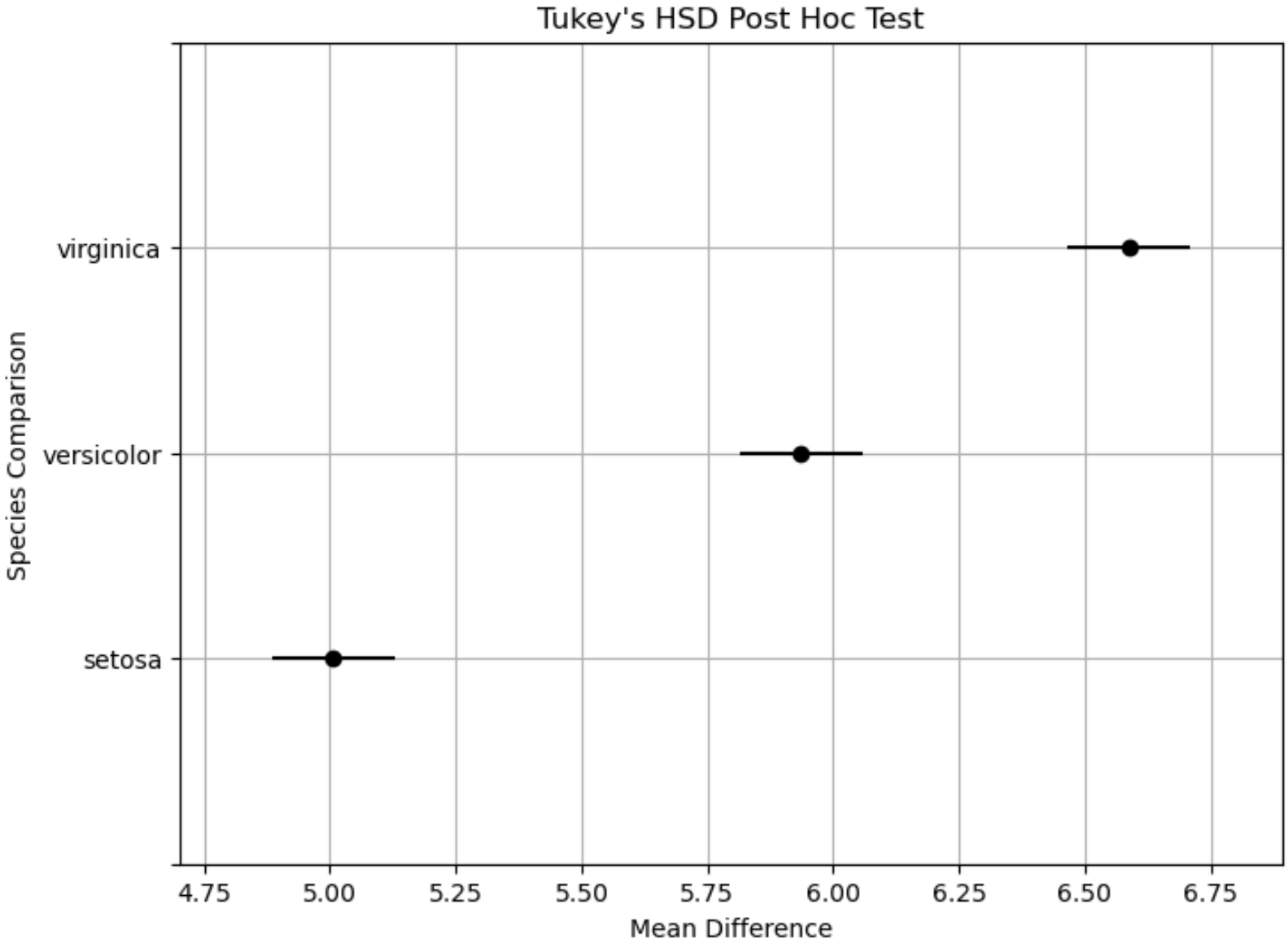
Post-hoc test and plot

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.

```
# Post-Hoc test using Tukey's HSD
tukey = pairwise_tukeyhsd(endog=iris['sepal_length'], groups=iris['species'],
     alpha=0.05)
print(tukey)

   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1     group2   meandiff p-adj lower  upper  reject
-----------------------------------------------------
    setosa versicolor     0.93   0.0 0.6862 1.1738   True
    setosa  virginica    1.582   0.0 1.3382 1.8258   True
versicolor  virginica    0.652   0.0 0.4082 0.8958   True
-----------------------------------------------------
```

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

Python:
https://www.python.org/