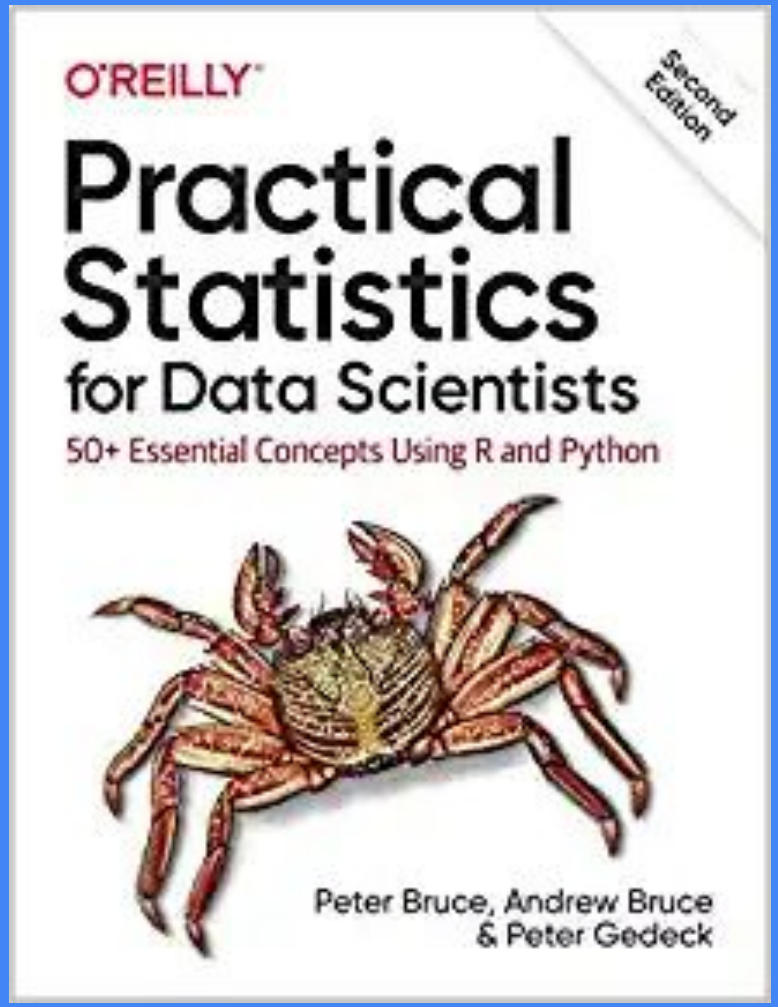


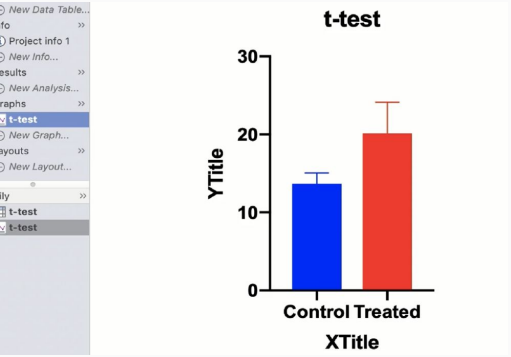
Chapter 3: t-Test, Multiple Testing, Degrees of Freedom, ANOVA

t-Test

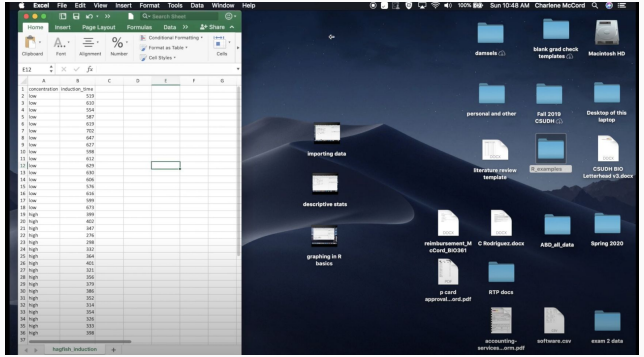
- Is a hypothesis test based on the Student's t-distribution.
- Is used in A / B testing, usually comparing the means of the two groups.
- Need to standardize the data to compare them to a Student's t-distribution of reference.
- In Data Science the t-Test is getting substituted by permutation tests where the null model can be created automatically the data points (large number of them).

How to run a 2 samples t-Test ?

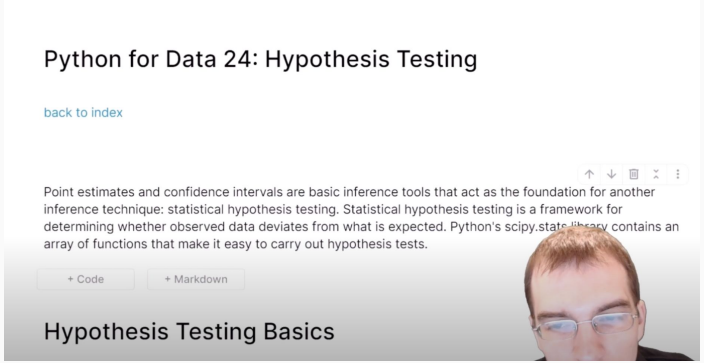
[Prism](#)



[R](#)



[Python](#)



Multiple testing

- “Torture the data long enough and it will confess” means that if we perform multiple tests on our data, we might come across with a false positive result.
- A adjustment factor needs to be applied, ex Bonferroni adjustment.
- In Data Science we attempt to escape from being fooled by chance with:
  - Validation set
  - Cross-validation
  - Resampling

Degrees of freedom

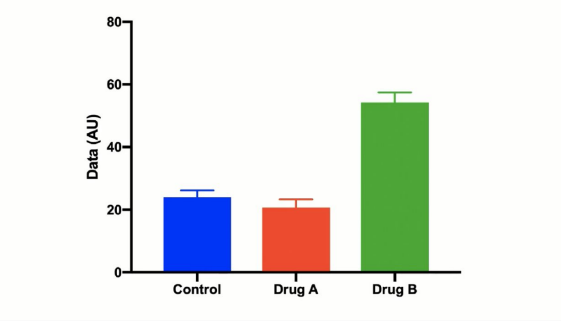
- Degrees of freedom stand for the number of values that are free to vary in the distribution of a statistic created from sample data.
- In Data Science, degrees of freedom are getting into use when factoring the categorical variables into n - 1 binary variables (avoid multicollinearity error)
- In Statistics, degrees of freedom are input in many statistical tests to standardize the test statistic to match a distribution of reference.

ANOVA

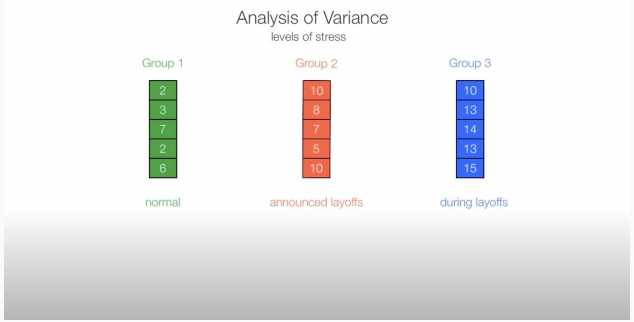
- ANOVA (Analysis of Variance) is the statistical procedure that test if the difference between two groups or more is statistically significant.
- Uses the difference of variances as the test statistic.
- ANOVA can be performed using either a permutation test or a statistical test, based on the F-statistic.

How to run an ANOVA test ?

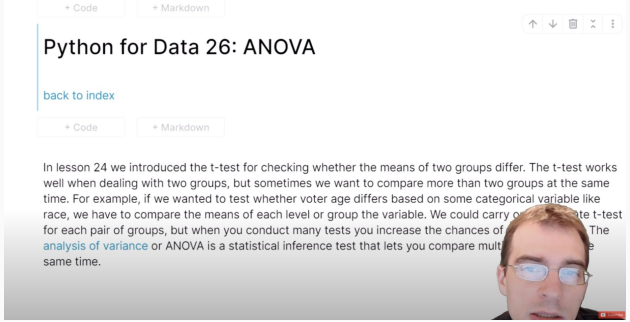
Prism



R



Python



Additional References

- https://www.graphpad.com/data-analysis-resource-center/
- https://www.youtube.com/watch?v=0Pd3dc1GcHc&t=257s
- http://www.sthda.com/

That was a summary from the book:

Practical Statistics for Data Scientists by Peter Bruce, Andrew Bruce and Peter Gedeck

Created by the members of the **#66DaysOfData** study group:

William Guesdon

Rea Kalampaliki

#66DAYSOFDATA