Three shades of ANOVA

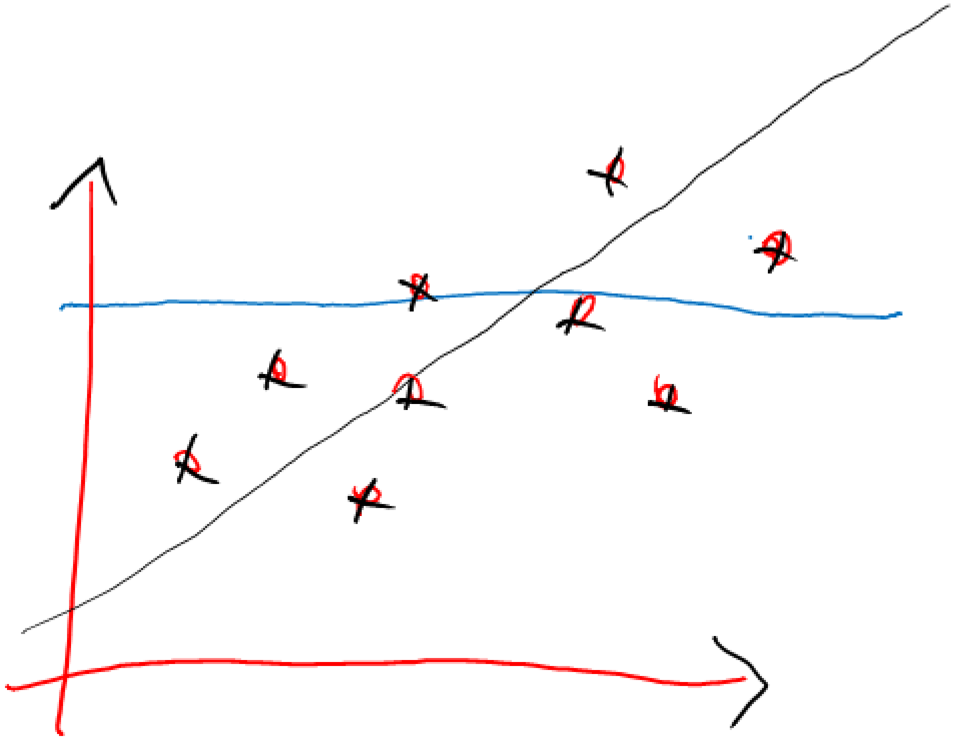
Chapter I.

Agoston Torok @ ELTE

Many *t*-tests vs. ANOVA

- What is the difference?
  - The way how the effects are calculated from the variance explained

- Sum of squares (*SS)*
  - See figure
  - Shows how well the model describes the data (i.e. how ‚far’ are the datapoints from the average)



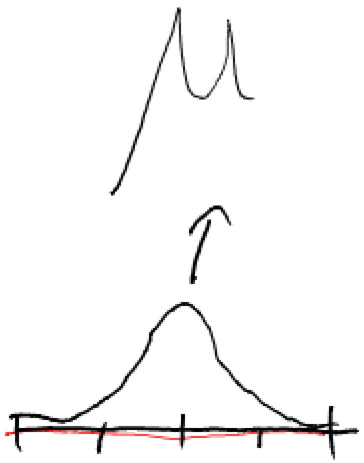
What is inside the ANOVA model?

Six potential subsamples of the whole sample

Data

A	B
1	1
1	2
1	3
2	1
2	2
2	3

Without knowing the Factors (here A and B) our best guess is the average in a normal distributed population



Design Matrix

A		B			A*B					
A1	A2	B1	B2	B3	A1 B1	A1 B2	A1 B3	A2 B1	A2 B2	A2 B3
1	0	1	0	0	1	0	0	0	0	0
1	0	0	1	0	0	1	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0
0	1	1	0	0	0	0	0	1	0	0
0	1	0	1	0	0	0	0	0	1	0
0	1	0	0	1	0	0	0	0	0	1

Main effects

Interaction

+e

Residual Error

Calculation of metrics

- Let’s take a simple experiment we want to decide whether reading Harry Potter books in Hungarian, English or Swahili makes people more happy. Eight participants were recruited and took part in the experiment, after reading all HP books they had to answer how happy they are on 7 point Likert scale. The results are as follows:

Factor level	Results		
Hungarian	3	4	3
English	2	3	4
Swahili	7	6	

NOTE:
We use colours here on purpose to be able to follow later which parameter/caluculation belongs where

Baseline model

Let's use the term model from now more frequently

- Our best guess without knowing on which language they read the book is the average of the whole sample

$$\mu = \frac{\sum \text{observations}}{\text{N}_{\text{observations}}} = (3 + 4 + 3 + 2 + 3 + 4 + 7 + 6) / 8 = 4$$

- The residual variance of the model can be described by its SS

$$\text{SS}_{\text{total}} = \sum_N^i (\mu - \text{y})^2 = (3\text{-}4)^2 + (4\text{-}4)^2 + (3\text{-}4)^2 + (2\text{-}4)^2 + (3\text{-}4)^2 + (4\text{-}4)^2 + (7\text{-}4)^2 + (6\text{-}4)^2 = \ldots$$

$$1 + 0 + 1 + 4 + 1 + 0 + 9 + 4 = \underline{20}$$

$$\text{df} = \text{N}_{\text{observations}} - 1 = \underline{7}$$

- The last component of the calculation is the Mean of the Squares

$$\text{MS} = \frac{SS_{total}}{df} = 20 / 7 = 2.86$$

Factorial model concepts

Remember: in the ANOVA the null hypothesis is that there is no effect of the factors
We are building the ANOVA model on the baseline model, that is we include the baseline model as the **intercept** of our factorial ANOVA model:

$$\hat{y}_{ij} = \mu + \hat{A}_i + \epsilon_{ij}$$

Here *i* is the group (can be Hungarian, English or Swahili) and *j* is the ID of the participant (can be 1 to 8). The hat on the *y* and *A* means that those are estimates, the μ has been calculated previously (4) and ϵ is the error that remains in the sample.

Factorial model

- Let's calculate the *A* values. Note: because we included the μ in our model the A is going to be a relative value

$$\hat{A}_i = \text{Mean}_i - \mu$$

- $\hat{A}_{Hungarian} = \text{Mean}_{Hungarian} - \mu = \frac{3+4+3}{3} - 4 = \underline{-0.67}$
- $\hat{A}_{English} = \text{Mean}_{English} - \mu = \frac{2+3+4}{3} - 4 = \underline{-1}$
- $\hat{A}_{Swahili} = \text{Mean}_{Swahili} - \mu = \frac{7+6}{2} - 4 = \underline{2.5}$

The SS between groups is the following

$$SS_{between} = \sum \widehat{A}_i^2 * n_i = -0.67^{\,2} * 3 + -1^{\,2} * 3 + 2.5^{\,2} * 2 = 1.35 + 3 + 12.5 \underline{= 16.85}$$

$$SS_{within} = \sum SS_i = \underline{3.15}$$

One step away from the *F* value

- We need the Mean of the Squares

$$MS_{Factors} = \frac{SS_{between}}{df_{factor\ levels}} = \frac{16.85}{2} = 8.425$$

$$MS_{Error} = \frac{SS_{within}}{df_{residual\ error}} = \frac{3.15}{5} = 0.63$$

We have two *df*-s that needs to be calculated. $df_{factor\ levels}$ is the number of factor levels -1, here 3-1 = 2. $df_{residual\ error}$ is the number of participants minus the number of factor levels, here 8 – 3 = 5. So finally:

$$F = \frac{MS_{Factors}}{MS_{Error}} = \underline{13.37}$$

In the next chapter

- What is an unbalanced design?
- Why is an unbalanced design problematic?
- Type 1,2,3 ANOVAs
- Outlook to Linear regression