MEM-264 Applied Statistics

Department of Mathematics and Applied Mathematics, University of Crete

Costas Smaragdakis (kesmarag@uoc.gr)

16th lecture - 02-04-2021

ANOVA test

Analytis of Variance (ANOVA) is a statistical method for comparing samples based on their means and the variance of the data.

ANOVA

- Checks if the means obtained by three or more groups are significantly different from each other!
- For two groups gives the same results as the t-test (see Lecture 15).

Contains the calculation of two basic elements

- Variation within each group
- Variation between groups

In terms of a test hypothesis problem

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{vs} \quad H_1 : \mu_\ell \neq \mu_m \text{ for at least one pair of indices} (\ell, m).$$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu \text{ vs } \quad H_1 : \mu_\ell \neq \mu_m \text{ for at least one pair of indices} (\ell, m)$$

The general form of the one-way-ANOVA is described by the following model

parameter

if Ho is true

$\mu - \mu_i \to 0$

Random errors

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \sum_{i=1}^{I} \alpha_i = 0, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

group order parameters

$$\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$$

recoverable parameters

$\mu, \alpha_1, \ldots, \alpha_{I-1}$

$\underline{I\text{-parameters}}$

- GROUP i=1 : $\{Y_{11}, \ldots, Y_{1,n_1}\}$ i.i.d sample η_1, σ^2
- GROUP i=2 : $\{Y_{21}, \ldots, Y_{2,n_2}\}$ i.i.d sample η_2, σ^2

$$Y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2)$$

$\vdots$

- GROUP i=I : $\{Y_{I1}, \ldots, Y_{I,n_I}\}$ i.i.d sample η_{n_I}, σ^2
- **The ANOVA model assumes that all groups have the same variance.**

Sample mean for the i-th GROUP

GROUP = POPULATION

$$\bar{Y}_{i,n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

, i fixed.

$\mathbb{E}\{\bar{Y}_{i,n_i}\} = \mu_i$

Sample variance for the i-th GROUP

$$S^2_{i,n_i} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} e^2_{ij}, \quad e_{ij} = Y_{ij} - \underline{\bar{Y}_{i,n_i}}$$

$\mathbb{E}\{S^2_{i,n_i}\} = \sigma_i^2 = \sigma^2$

Pooled sample variance

$$S^2_p = \frac{\sum_{i=1}^{I}(n_i - 1)S^2_{i,n_i}}{\sum_{i=1}^{I}(n_i - 1)} = \frac{\sum_{i=1}^{I}(n_i - 1)\, S^2_{i,n_i}}{\underbrace{\sum n_i}_{N} - I}$$

N = the total number of random variabl

Overall sample mean

$N = \sum_{i=1}^{I} n_i$

$$\bar{Y} = \frac{1}{\sum_{i=1}^{I} n_i} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}$$

under H_0

Sample variance between groups

$\frac{1}{I-1} \sum \left(\sqrt{n_i}\, \bar{Y}_{i,n_i} - \sqrt{n_i}\, \bar{Y} \right)^2$

"

$$S_b^2 = \frac{1}{I-1} \sum_{i=1}^{I} n_i (\bar{Y}_{i,n_i} - \bar{Y})^2$$

$\bar{Y}_{i,n_i} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n_i}\right)$

$\sqrt{n_i}\, \bar{Y}_{i,n_i} \sim \mathcal{N}(\sqrt{n_i}\mu, \sigma^2)$

$\sqrt{n_1}\, \bar{Y}_{1,n_1}, \sqrt{n_2}\, \bar{Y}_{2,n_2}, \sqrt{n_3}\, \bar{Y}_{3,n_3}$

Under the null hypothesis

$$Y_{ij} = \mu + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

A single parameter. μ

- Must be $S_p^2 \approx S_b^2$

$\sigma^2 \uparrow \quad \sigma^2 \uparrow$

F statistic

$$F = \frac{S_b^2}{S_p^2}$$

- If H_0 is true then $F \approx 1$
- IF H_0 is false then $F > 1$