Introduction to Statistics 02402/02323

Lecture 11: One-way Analysis of Variance, ANOVA

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

Analysis of Variance - ANOVA

"ANalysis Of VAriance" (ANOVA) was introduced by R.A. Fisher about 100 years ago as a systematic way to analyze groups and has since been fundamental to the development of statistics.

- Today: A single classification criterion (one-way ANOVA)
- *In course 02402 - Next week: Two classification criteria (two-way ANOVA)*
- Classification criterion = factor
- The first factor is typically called *treatment*, the second factor *block*

Overview

One-Way Analysis of Variance - Example

Group A	Group B	Group C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

Is there a difference (in mean) between groups A, B, and C?

Analysis of variance (ANOVA) can be used for the analysis, provided the observations in each group can be assumed to be normally distributed.

Example in Python

- Go to today's Python notebook in VS Code
    - "Example: Intro to ANOVA"

Overview

One-Way Analysis of Variance - Model

- The model can be written as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,,$$

where it is assumed that ε_{ij} are i.i.d. with

$$\varepsilon_{ij} \sim N(0, \sigma^2)\,.$$

- μ is the overall mean
- α_i indicates the effect of group (treatment) i
- Y_{ij} is measurement j in group i (j ranges from 1 to n_i)

One-Way Analysis of Variance - Hypothesis Test

- We will now compare (more than two) means $(\mu + \alpha_i)$ in the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)\,.$$

- The null hypothesis is given by:

$$H_0: \quad \alpha_i = 0 \quad \text{for all } i.$$

- The alternative hypothesis is given by:

$$H_1: \quad \alpha_i \neq 0 \quad \text{for at least one } i.$$

Overview

One-Way ANOVA - Decomposition and ANOVA Table

- With the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

  the total variation in data can be decomposed:

$$SST = SS(Tr) + SSE\,.$$

- 'One-way' implies that there is only one factor in the experiment (with k levels).
- The method is called analysis of variance because testing is done by comparing variances.

Formulas for Sum of Squares

- Total variation:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

Formulas for Sum of Squares

- Total variation:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- Variation within groups (Residual variation left after the model):

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Formulas for Sum of Squares

- Total variation:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

- Variation within groups (Residual variation left after the model):

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

- Variation between groups (Variation explained by the model):

$$SS(Tr) = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

Kahoot!
(x4)

One-Way ANOVA - Parameter Estimates

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = \bar{y}_i - \bar{y}$
- $\hat{\sigma}^2 = MSE = \frac{SSE}{n-k}$

Visual Studio Code

(μ, α_i and σ^2)

ANOVA Table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$
Total	$n-1$	SST	

n: Number of observations

k: Number of groups

Visual Studio Code

(ANOVA table)

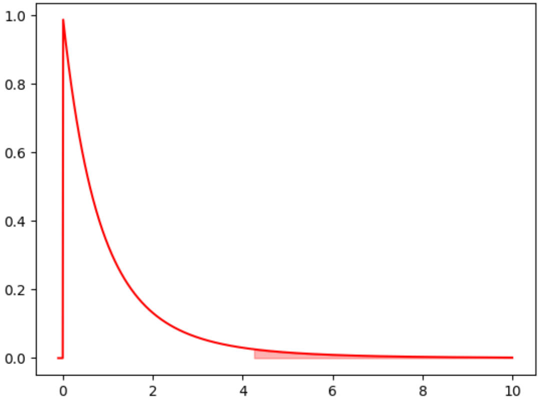
Overview

One-Way ANOVA - F-Test

- We have (Theorem 8.2)
$$SST = SS(Tr) + SSE$$
- From this, the test statistic can be derived:
$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{MS(Tr)}{MSE} = \frac{\text{"between group variation"}}{\text{"within group variation"}},$$
where
  - k is the number of groups,
  - n is the number of observations.
- Choose a significance level α and compute the test statistic F.
- Compare the test statistic with the $(1-\alpha)$ quantile in the F distribution:
$$F \sim F(k-1, n-k) \text{ (Theorem 8.6)}$$

F Distribution and F-Test

Plot of F-distribution with $n = 12$ and $k = 3$:



The 5% *most extreme* values are above 4.26.

Analysis of variance table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
treatment	$k-1$	$SS(Tr)$	$MS(Tr)=\frac{SS(Tr)}{k-1}$	$F_{\text{obs}}=\frac{MS(Tr)}{MSE}$	$P(F>F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE=\frac{SSE}{n-k}$		
Total	$n-1$	SST			

Visual Studio Code

(F-test)

Overview

Variability and relation with the t-test for two samples (Theorem 8.4)

The residual sum of squares, SSE, divided by $n-k$, also called residual mean square, $MSE = SSE/(n-k)$, is the average within-group variability:

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{n-k},$$

$$s_i^2 = \frac{1}{n_i-1}\sum_{i=1}^{n_i}(y_{ij}-\bar{y}_i)^2.$$

ONLY when $k=2$: (cf. Method 3.52)

$$MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2},$$

$$F_{\text{obs}} = t_{\text{obs}}^2,$$

where t_{obs} is the pooled t-test statistic from Methods 3.52 and 3.53.

Overview

Why is Post-hoc Comparison Necessary?

- ANOVA tests the overall null hypothesis that all group means are equal $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.
- If ANOVA shows a significant result (i.e., $p - value < \alpha$), it only indicates that at least one group mean differs, but it doesn't specify which groups are different from each other.
- Post-hoc comparisons are performed to pinpoint which specific group means are different.

Post hoc confidence interval – Method 8.9

- A single *planned* comparison of the difference between treatment i and j is found by:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{\frac{SSE}{n-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

  where $t_{1-\alpha/2}$ is from the t-distribution with $n-k$ degrees of freedom.

- Note the fewer degrees of freedom, since more parameters are estimated in calculating $MSE = SSE/(n-k) = s_p^2$ (the pooled variance estimate)

- If all $M = k(k-1)/2$ combinations of pairwise confidence intervals are calculated, use the formula M times, each time with $\alpha_{\text{Bonferroni}} = \alpha/M$.

Post hoc pairwise hypothesis test – Method 8.10

- For a single *planned* hypothesis test

$$H_0:\ \mu_i = \mu_j, \qquad H_1:\ \mu_i \neq \mu_j, \quad i \neq j$$

a t-test with $n-k$ degrees of freedom can be used with test statistic

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{\frac{SSE}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}.$$

- If all $M = k(k-1)/2$ combinations of pairwise tests are done, then the α level can be adjusted to control the type I error rate using the Bonferroni approach:

$$\alpha_{\text{Bonferroni}} = \frac{\alpha}{M}.$$

Overview

Model Control

Our model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)\,.$$

What is our assumption(s)?

Can we do anything to check if the assumptions are valid?

Visual Studio Code

(Model control)

Overview

A Worked Example – From the Book

Introduction to Statistics

Agendas eNotes Course Material Podcast Forum Quiz Admin

Dokumentegenskaber...

8.2.5 A complete worked through example: plastic types for lamps

Example 8.17 Plastic types for lamps

On a lamp two plastic screens are to be mounted. It is essential that these plastic screens have a good impact strength. Therefore an experiment is carried out for 5 different types of plastic. 6 samples in each plastic type are tested. The strengths of these items are determined. The following measurement data was found (strength in kJ/m^2):

Type of plastic				
I	II	III	IV	V
44.6	52.8	53.1	51.5	48.2
50.5	58.3	50.0	53.7	40.8
46.3	55.4	54.4	50.5	44.5
48.5	57.4	55.3	54.4	43.9
45.2	58.1	50.6	47.5	45.9
52.3	54.6	53.4	47.8	42.5

Overview