Introduction to Mathematical Statistics 02403

Lecture 10: 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)
- Tomorrow: Two classification criteria (two-way ANOVA)
- Classification criterion = factor
- The first factor is typically called *treatment*, the second factor *block*

Overview

F-distribution

If $Q_1 \sim \chi^2(n_1)$ and $Q_2 \sim \chi^2(n_2)$, and Q_1 and Q_2 independent then

$$F = \frac{Q_1/n_1}{Q_2/n_2} \tag{1}$$

an F-distribution with n_1 and n_2 degrees of freedom.

The F-distribution as a sample distribution

Let $Y_{1,1}, \ldots, Y_{1,n_1}$ be iid. $N(\mu_1, \sigma^2)$ and let $Y_{2,1}, \ldots, Y_{2,n_2}$ være i.i.d. $N(\mu_2, \sigma)$ then

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \tag{2}$$

where S_1^2 and S_2^2 are the sample variance for Y_1 hhv. Y_2.

The pooled two-sample t-test statistic

Calculation of the pooled variance test statistics (Method 3.63 and 3.64)

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0: \ \delta = \delta_0$$

the pooled two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

With $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$.

Pooled variance set up

Assume that $Y_{1,i} \sim N(\mu_1, \sigma)$ and $Y_{2,j} \sim N(\mu_2, \sigma)$. Then the pooled two-sample statistic seen as a random variable (Theorem 3.54, Example 2.85 og Exercise 2.16):

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}} \tag{3}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

Pooled variance set up

Assume that $Y_{1,i} \sim N(\mu_1, \sigma)$ and $Y_{2,j} \sim N(\mu_2, \sigma)$.

We want to insestigate the hypothesis

$$H_0: \quad \mu_1 = \mu_2$$

Then under the assumptions and the null hypothesis, the test statistics

$$T_{obs} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{1/n_1 + 1/n_2}}$$

follow a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

And hence

$$T_{obs}^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{S_p^2(1/n_1 + 1/n_2)} \sim F(1, n_1 + n_2 - 2).$$

The aim of today

Assume that $Y_{i,j} \sim N(\mu_i, \sigma^2)$, and iid. with $i \in \{1,...,k\}$ and $j \in \{1,...,n_i\}$
We want to investigate the hypothesis

$$\mu_1 = \mu_2 = ... = \mu_k$$

The general idea is to generalize the two-sample t-test, the result is

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{S_p^2} \sim F(k-1, n-k)$$

where S_p^2 is the best estimator for the σ^2.

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

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad \text{and iid.}$$

such that $E[Y_{ij}] = \mu_i$. The null-hypothesis is $H_0 : \mu_i = \mu_j$ for all (i, j).
The model can also be formulated as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

The design matrix $\boldsymbol{X}$ can be parametrized in different ways, a simple form is

$$\boldsymbol{X}_1 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & & \ddots & \vdots \\ \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \dots & \mathbf{1}_{n_K} \end{bmatrix},$$

in this case $\beta_i = \mu_i$, and the null-hypothesis is $\beta_i = \beta_j$.

Example

One way ANOVA (the model):

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{ i.i.d. } N(0, \sigma^2), \quad i = 1,2,3, \quad j = 1,2.$$

An expanded view of this model is:

$$\begin{array}{lllll} y_{11} = \mu_1 & & & + \varepsilon_{11} \\ y_{12} = \mu_1 & & & + \varepsilon_{12} \\ y_{21} = & \mu_2 & & + \varepsilon_{21} \\ y_{22} = & \mu_2 & & + \varepsilon_{22} \\ y_{31} = & & \mu_3 & + \varepsilon_{31} \\ y_{32} = & & \mu_3 & + \varepsilon_{32} \end{array}$$

The exact same in matrix notation:

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}}_{\varepsilon}$$

One-Way Analysis of Variance - Parametrization

While the previous parametrization is simple it is not so common and a more common parametrization is

$$\boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & & \ddots & \vdots \\ \mathbf{1}_{n_K} & \mathbf{0}_{n_K} & \cdots & \mathbf{1}_{n_K} \end{bmatrix},$$

and the null hypothesis translate to

$$H_0: \quad \beta_i = 0 \quad \text{for } i > 1.$$

One-Way Analysis of Variance - Parametrization

Chapter 8 use the parametrization

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

with the addition constraint $\sum_i n_i \alpha_i = 0$, which correspond to the parametrization

$$\boldsymbol{X}_3 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \cdots & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & & \ddots & \vdots \\ \mathbf{1}_{n_{K-1}} & \mathbf{0}_{n_{K-1}} & \cdots & & \mathbf{1}_{n_{K-1}} \\ \mathbf{1}_{n_K} & -\frac{n_1}{n_K}\mathbf{1}_{n_K} & \cdots & \cdots & -\frac{n_{K-1}}{n_K}\mathbf{1}_{n_K} \end{bmatrix},$$

and the null hypothesis translate to

$$H_0: \quad \beta_i = 0 \quad \text{for } i > 1.$$

Null hypothesis

In all cases (parametrizations) the null-hypothesis correspond to the design matrix

$$\boldsymbol{X}_0 = \mathbf{1}$$

and the partioning of variation is

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}-\boldsymbol{H}_0)\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$$

with

- $Tr(\boldsymbol{H}_0) = 1$, $Tr(\boldsymbol{H}) = k$, $Tr(\boldsymbol{I}) = \sum_{i=1}^{k} n_i = n$

and under the null-hypothesis

$$F_{obs} = \frac{\boldsymbol{Y}^T(\boldsymbol{H}-\boldsymbol{H}_0)\boldsymbol{Y}/(k-1)}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}/(n-k)} \sim F(k-1, n-k).$$

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:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H}_0) \boldsymbol{Y}$$

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

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

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

$$SS(Tr) = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 = \boldsymbol{Y}^T (\boldsymbol{H} - \boldsymbol{H}_0) \boldsymbol{Y}$$

One-Way ANOVA - Parameter Estimates

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

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

or

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_3^T \boldsymbol{X}_3)^{-1} \boldsymbol{X}_3^T \boldsymbol{Y}$$

with

$$\hat{\boldsymbol{\beta}} = [\hat{\mu}, \hat{\alpha}_1, ..., \hat{\alpha}_{k-1}].$$

with $\hat{\alpha}_k = -\sum_{i=1}^{k-1} \frac{n_i}{n_k} \hat{\alpha}_i$.

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:

$$F \sim F(k-1, n-k) \text{ (Theorem 8.6)}$$

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			

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 validation

Our model:

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

Check the ususal assumptions

- Identically distributed (check within group variation, e.g. box-plots).
- Distribution (qq-plot of residuals (usually any kind go, but see Lecture 9)).
- If data is a time series, check serial correlation.

Overview

1. F-distribution
2. Introduction
3. Model and Hypotheses
4. Computation: Variance Decomposition and ANOVA Table
5. Hypothesis Testing (F-test)
6. Variability and relation with the t-test for two samples
7. Post hoc comparisons
8. Model