# Part I: Linear Models

## Chapter I.6

## ANOVA – Analysis of Variance

# Topics

**To be discussed...**

- Motivation: two-sample t-test
- one-factorial ANOVA
- effect representation
- F-test
- two-factorial ANOVA

# Two sample t-tests

### ▶ I.6.1 Example

In order to test the impaired roadworthiness affected by a drug, two test series are conducted. In the experiment, the response time to an event is measured. In group I, the test persons receive a placebo whereas group II gets the treatment. We get the following measurements (in seconds):

Group I (placebo): `1 1.29 1.31 2.48 0.98 2.26 1.89 2.29 2.47 1.45 1.21 1.05 0.85 2.16 1.89 2.23 1.53`

Group II (treatment): `2.68 1.62 2.11 2.56 3.38 2.57 2.55 2.1 2.01 3 3.36 3.14 1.78 3.28 2.09 2.86 3.43 3.44 2.15`

❯ **Problem**: Does the medication have an influence on the response time?

# Two sample t-tests

> **I.6.2 Model**
>
> Let $X_1, \ldots, X_{n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$ be independent samples with $n_1, n_2 \geqslant 2$. The parameters $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma^2 > 0$ are supposed unknown.

Relevant hypotheses are:

> **I.6.3 Hypotheses (mean comparisons)**
>
> $$H_0 : \mu_1 \leqslant \mu_2 \qquad \longleftrightarrow \qquad H_1 : \mu_1 > \mu_2$$
> $$H_0 : \mu_1 \geqslant \mu_2 \qquad \longleftrightarrow \qquad H_1 : \mu_1 < \mu_2$$
> $$H_0 : \mu_1 = \mu_2 \qquad \longleftrightarrow \qquad H_1 : \mu_1 \neq \mu_2$$

# Two sample t-tests

The two-sample t-test is constructed as follows:

- The difference $\mu_1 - \mu_2$ is estimated by the difference of the sample means $\overline{\Delta} = \overline{X} - \overline{Y}$.

  Due to the independence of the samples, $\overline{X}$ and $\overline{Y}$ are independent so that

  $$\overline{\Delta} = \overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \sigma_{\Delta}^2\right)$$

  where $\sigma_{\Delta}^2 = \text{Var}\left(\overline{X} - \overline{Y}\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2$.

- The variance $\sigma^2$ is estimated by

  $$\widehat{\sigma}_{\text{pool}}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{j=1}^{n_2}(Y_j - \overline{Y})^2\right).$$

- Test statistic: $\widehat{D} = \dfrac{\overline{X} - \overline{Y}}{\widehat{\sigma}_{\text{pool}}\sqrt{1/n_1 + 1/n_2}}$.

# Two sample t-tests

## ▣ I.6.4 Remark

- ❯ Since the variance is supposed identical in Model I.6.2, both samples provide information about $\sigma^2$. Thus, its estimator $\widehat{\sigma}^2_{\text{pool}}$ is based on **both** samples!

- ❯ Notice that

$$\widehat{\sigma}^2_{\text{pool}} = \frac{n_1 - 1}{n_1 + n_2 - 2}\widehat{\sigma}^2_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\widehat{\sigma}^2_2$$

where $\widehat{\sigma}^2_1$ and $\widehat{\sigma}^2_2$ are the sample variance of the X- and the Y-sample, respectively.

- ❯ Since $\widehat{\sigma}^2_1$ and $\widehat{\sigma}^2_2$ are unbiased estimates of $\sigma^2$, $\widehat{\sigma}^2_{\text{pool}}$ is unbiased, too, that is, for all $\sigma^2 > 0$

$$\mathrm{E}\widehat{\sigma}^2_{\text{pool}} = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathrm{E}\widehat{\sigma}^2_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathrm{E}\widehat{\sigma}^2_2 = \sigma^2.$$

# Two sample t-tests

## ▶ I.6.4 Remark

- From Corollary I.2.17, we get $\overline{\Delta}$ and $\widehat{\sigma}_{\mathsf{pool}}^2$ are independent as well as

$$\frac{(n_1 + n_2 - 2)\widehat{\sigma}_{\mathsf{pool}}^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

  Therefore,

$$\widehat{D} = \frac{\overline{X} - \overline{Y}}{\widehat{\sigma}_{\mathsf{pool}}\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2) \quad (\text{if } \mu_1 = \mu_2).$$

- In Model I.6.2, $(\overline{X}, \overline{Y}, \frac{n_1 + n_2 - 2}{n_1 + n_2}\widehat{\sigma}_{\mathsf{pool}}^2)$ is the MLE of $(\mu_1, \mu_2, \sigma^2)$. This follows directly from Theorem I.4.31 and Remark I.6.7.

  Furthermore, using properties of MLEs, we have that $(\overline{X} - \overline{Y}, \frac{n_1 + n_2 - 2}{n_1 + n_2}\widehat{\sigma}_{\mathsf{pool}}^2)$ is the MLE of $(\mu_1 - \mu_2, \sigma^2)$.

- In Model I.6.2, it is assumed that the variance is unknown (and identical in both samples).

  Assuming the variances to be known, the test statistic $\widetilde{D} = \frac{\overline{X} - \overline{Y}}{\sigma_\Delta}$ defines two-sample Gaussian tests. In this case, the variances in each sample may be different.

# Two sample t-tests

## ▶ I.6.5 Procedures (Decision rules for the hypotheses)

|  |  | $\sigma^2$ known | $\sigma^2$ unknown |
|---|---|---|---|
| $H_0$ | $H_1$ |  | $H_0$ is rejected if |
| $\mu_1 \leqslant \mu_2$ | $\mu_1 > \mu_2$ | $\widetilde{D} > u_{1-\alpha}$ | $\widehat{D} > t_{1-\alpha}(n_1 + n_2 - 2)$ |
| $\mu_1 \geqslant \mu_2$ | $\mu_1 < \mu_2$ | $\widetilde{D} < -u_{1-\alpha}$ | $\widehat{D} < -t_{1-\alpha}(n_1 + n_2 - 2)$ |
| $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ | $|\widetilde{D}| > u_{1-\frac{\alpha}{2}}$ | $|\widehat{D}| > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)$ |

$u_{1-\alpha}$ and $t_{1-\alpha}(m)$ are the quantiles of the standard normal distribution and the t-distribution with $m$ degrees of freedom, respectively.

# Two sample t-tests

## ▶ I.6.6 Example

In Example I.6.1, we get the following estimates and results:

- ◉ $\hat{\mu}_1 = 1.667058824$, $\hat{\mu}_2 = 2.637368421$, $\hat{\mu}_1 - \hat{\mu}_2 = -0.970309597$

  $\hat{\sigma}_1^2 = 0.325897059$, $\quad \hat{\sigma}_2^2 = 0.360031579$

  $\hat{\sigma}_{\mathsf{pool}}^2 = 0.343968275$, $\hat{\sigma}_{\mathsf{pool}} = 0.586488086$

- ◉ Furthermore, F-test does not reject null hypothesis of equal variances

- ◉ $\hat{D} = \dfrac{1.667058824 - 2.637368421}{0.586488086 \cdot \sqrt{1/17 + 1/19}} = -4.955655903$

- ◉ Given level $\alpha = 0.01$, we have $t_{0.99}(34) = 2.441149628$ und $t_{0.995}(34) = 2.728394367$
  - ◉ null hypothesis $H_0 : \mu_1 \geqslant \mu_2$ and $H_0 : \mu_1 \neq \mu_2$ are rejected.
  - ◉ the test shows that the drug affects the impaired roadworthiness; the response time becomes larger

# Two sample t-tests as LM

> ▷ **I.6.7 Remark**

The two-sample Model I.6.2 can be written as a LM:

Let $n_1, n_2 \in \mathbb{N}$, $(\varepsilon_{ij})_{i=1,\ldots,2,j=1,\ldots,n_i} \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$, and $\mu_1, \mu_2 \in \mathbb{R}$ be parameters. Then,

$$(X_j =) \, Y_{1j} = \mu_1 + \varepsilon_{1j}, \quad j = 1, \ldots, n_1,$$
$$(Y_j =) \, Y_{2j} = \mu_2 + \varepsilon_{2j}, \quad j = 1, \ldots, n_2$$

forms a LM $\mathbf{Y} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with parameter vector $\boldsymbol{\beta} = (\mu_1, \mu_2)'$, design matrix $B = \begin{bmatrix} \mathbb{1}_{n_1} & 0 \\ 0 & \mathbb{1}_{n_2} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times 2}$ and $\mathbf{Y} = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2})'$.

In a study, three learning methods to learn vocabulary are evaluated. Method i is used in group i. The participants are randomly assigned to the groups and asked to learn 100 words in a foreign language with the corresponding method. The success is measured by a test after one week. The respective numbers of correct vocabulary are given in the following table:

| Method 1 | Method 2 | Method 3 |
|----------|----------|----------|
| 82 | 57 | 61 |
| 93 | 59 | 50 |
| 80 | 71 | 66 |
| 79 | 46 | 70 |
| 87 | 49 | 39 |
| 69 | 51 | 53 |
| 78 | 54 | 45 |
| 91 | 61 | 57 |
|  | 63 | 68 |
|  | 55 |  |

# Analysis of variance (ANOVA)

## ▶ I.6.9 Example (one factorial analysis of variance, 3 factors)

Let $n_1, n_2, n_3 \in \mathbb{N}$, $(\varepsilon_{ij})_{i=1,\ldots,3, j=1,\ldots,n_i} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$, and $\mu_1, \mu_2, \mu_3 \in \mathbb{R}$ parameters. Then,

$$Y_{1j} = \mu_1 + \varepsilon_{1j}, \quad j = 1, \ldots, n_1,$$
$$Y_{2j} = \mu_2 + \varepsilon_{2j}, \quad j = 1, \ldots, n_2,$$
$$Y_{3j} = \mu_3 + \varepsilon_{3j}, \quad j = 1, \ldots, n_3,$$

forms a LM $\mathbf{Y} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with parameter vector $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)'$, design matrix $B = \begin{bmatrix} \mathbb{1}_{n_1} & 0 & 0 \\ 0 & \mathbb{1}_{n_2} & 0 \\ 0 & 0 & \mathbb{1}_{n_3} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2+n_3) \times 3}$ and $\mathbf{Y} = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, Y_{31}, \ldots, Y_{3n_3})'$.

## ▣ I.6.10 Remark

- ❯ The dependent variable $Y$ is measured for different groups that differ w.r.t. one or more aspects; these are called **factors** (treatment)

- ❯ The values of a factor are called **factor level**.

- ❯ The aim of the analysis is to identify differences in the data w.r.t. to the factor levels.

- ❯ We consider only **fixed effects**, that is, factors and their levels are known/fixed. It is also possible to study models with random effects.

# Examples for one-factorial ANOVA

## ▣ I.6.11 Example

❶ In Example I.6.8 the dependent variable Y (**No. of correct words**) is influenced by the factor **Method**. The factor **Method** has levels **Method 1**, **Method 2**, and **Method 3**.

❷ Five marketing strategies are tested. The dependent variable is defined as the sales volume Y per week. The levels are the five strategies.

❸ In order to study the influence of a fertilizer on the growth of crop, ten different fertilizers (factor) are sown on grainfields. After five weeks, the biomass Y per square meter is measured. Alternatively, one may consider the crop per square meter.

❹ In order to treat hypertension, six drugs should be tested. After medication, the reduction (Y) of blood pressure is measured to evaluate the success of the treatment. The levels are given by the different drugs.

❺ The effect of four preparative therapies on the success of a psychotherapy should be evaluated. Thus, 20 patients are randomly selected and assigned to four groups which are treated by the preparatives.

# Analysis of variance (ANOVA)

## ▶ I.6.12 Example (one factorial analysis of variance, effect representation)

Let $n_1, n_2, n_3 \in \mathbb{N}$ with $n_1 + n_2 + n_3 \geqslant 4$, $(\varepsilon_{ij})_{i=1,\dots,3, j=1,\dots,n_i} \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$, and $\mu, \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ be parameters. Then,

$$Y_{1j} = \mu + \alpha_1 + \varepsilon_{1j}, \quad j = 1, \dots, n_1,$$
$$Y_{2j} = \mu + \alpha_2 + \varepsilon_{2j}, \quad j = 1, \dots, n_2,$$
$$Y_{3j} = \mu + \alpha_3 + \varepsilon_{3j}, \quad j = 1, \dots, n_3,$$

forms a LM $\mathbf{Y} = B_* \boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$ with parameter vector $\boldsymbol{\beta}_* = (\mu, \alpha_1, \alpha_2, \alpha_3)'$, design matrix $B_* = \begin{bmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & 0 & 0 \\ \mathbb{1}_{n_2} & 0 & \mathbb{1}_{n_2} & 0 \\ \mathbb{1}_{n_3} & 0 & 0 & \mathbb{1}_{n_3} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2+n_3) \times 4}$ and $\mathbf{Y} = (Y_{11}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{2n_2}, Y_{31}, \dots, Y_{3n_3})'$.

- ❯ $B_*$ has not full rank (rank$(B_*) = 3$)
- ❯ The model is **overparameterized**.

# Analysis of variance (ANOVA)

▶ **I.6.13 Remark (one-factorial ANOVA)**

- ▶ In the above chosen parametrization (here: $p = 3$), we call
  - ▶ $\mu = \frac{1}{n} \sum_{i=1}^{p} n_i \mu_i$ with $n = n_{\bullet} = \sum_{i=1}^{p} n_i$ (**grand mean**)
  - ▶ $\alpha_i = \mu_i - \mu$ (**effect of $i$-th factor level/treatment**)

- ▶ Notice that the parameter vector $(\mu, \alpha_1, \alpha_2, \alpha_3)'$ is not identifiable, since $B_*$ has not full rank.

  In order to ensure identifiability, one may introduce a **side conditions** $\nu' \beta_* = 0$ such that $\begin{bmatrix} B_* \\ \nu' \end{bmatrix}$ has full rank (see Theorem I.6.17). Then, the LSE for any estimable function is always the same (that is, independent of the chosen constraint)!

- ▶ Common choices for the mentioned constraint are:
  - ▶ $\nu = (0, n_1, \ldots, n_p)'$ and $\nu' \beta_* = \sum_{i=1}^{p} n_i \alpha_i = 0$
  - ▶ $\nu = e_{1, p+1}$ and $\nu' \beta_* = \mu = 0$
  - ▶ $\nu = e_{i, p+1}$ for some $i \in \{2, \ldots, p+1\}$, and $\nu' \beta_* = \alpha_i = 0$

- ▶ Thus, we have the constraint NoLM

$$Y = B_* \beta_* + \varepsilon \quad \text{with } \nu' \beta_* = 0.$$

# Analysis of variance (ANOVA)

> **I.6.14 Definition (one factorial analysis of variance with $p$ factor levels)**

Let $n_1, \ldots, n_p \in \mathbb{N}$, $n = n_\bullet = \sum_{j=1}^{p} n_j$, $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$, and $\mu_1, \ldots, \mu_p \in \mathbb{R}$ be parameters. Then, the **one-factorial analysis of variance** model (one-factorial ANOVA) is defined by the LM

$$\mathbf{Y} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with parameter vector $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_p)'$, design matrix $B = \begin{bmatrix} \mathbb{1}_{n_1} & 0 & \cdots & \cdots & 0 \\ 0 & \mathbb{1}_{n_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \mathbb{1}_{n_p} \end{bmatrix} \in \mathbb{R}^{n_\bullet \times p}$ and

$\mathbf{Y} = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{p1}, \ldots, Y_{pn_p})'$.

$(n_1, \ldots, n_p)$ is called **(experimental) design** of the ANOVA model. For $n_1 = \cdots = n_p \in \mathbb{N}$, the design/model is called **balanced** and otherwise **unbalanced**.

**I.6.15 Definition (one factorial analysis of variance with $p$ factor levels in effect representation)**

Let $n_1, \ldots, n_p \in \mathbb{N}$, $n = n_\bullet = \sum_{j=1}^{p} n_j$, $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$, and $\mu, \alpha_1, \ldots, \alpha_p \in \mathbb{R}$ be parameters. Then, the LM

$$Y = B_* \beta_* + \varepsilon$$

with parameter vector $\beta_* = (\mu, \alpha_1, \ldots, \alpha_p)'$, design matrix $B_* = \begin{bmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & 0 & \cdots & \cdots & 0 \\ \mathbb{1}_{n_2} & 0 & \mathbb{1}_{n_2} & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ \mathbb{1}_{n_p} & 0 & \cdots & \cdots & 0 & \mathbb{1}_{n_p} \end{bmatrix} \in$

$\mathbb{R}^{n_\bullet \times (p+1)}$ and $Y = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{p1}, \ldots, Y_{pn_p})'$ is called **one-factorial ANOVA in effect representation**.

**I.6.16 Remark**

The design matrix

- $B \in \mathbb{R}^{n_\bullet \times p}$ in Model I.6.14 has $\text{rank}(B) = p$.
- $B_* \in \mathbb{R}^{n_\bullet \times (p+1)}$ in Model I.6.15 has $\text{rank}(B_*) = p < p + 1$.

# Estimation with a constraint/side condition

> **I.6.17 Theorem**
>
> Consider a LM $Y = B_* \beta_* + \varepsilon$ with design matrix $B_* \in \mathbb{R}^{n \times k}$ and $\text{rank}(B_*) = r < k \leqslant n$. Furthermore, let $T \in \mathbb{R}^{(k-r) \times k}$ with $\text{rank}(T) = k - r$ such that $B = \left[ \begin{smallmatrix} B_* \\ T \end{smallmatrix} \right]$ satisfies $\text{rank}(B) = k$.
>
> Then, there is a unique vector $\widehat{\beta}_*$ that satisfies both the normal equation $B_*' B_* \widehat{\beta}_* = B_*' y$ and the side condition $T \widehat{\beta}_* = 0$.
>
> If $K\beta_*$ is an estimable function, then $K\widehat{\beta}_*$ is the unique (unbiased) LSE.

> **I.6.18 Remark**
>
> - Notice that the conditions on $T$ in Theorem I.6.17 mean that the rows of $B_*$ and $T$ are linearly independent.
>
> - The condition on $T$ mean that the function $T\beta$ is not (linearly) estimable.
>
> - Moreover, a unique LSE exists in the model $Y = B\beta_* + \varepsilon$. It is given by
>
> $$\widehat{\beta}_* = (B'B)^{-1} B' \left[ \begin{smallmatrix} y \\ 0 \end{smallmatrix} \right] = (B_*' B_* + T'T)^{-1} B_*' y$$
>
> Further details can be found in, e.g., Rencher, Schaalje (2008).

# Analysis of variance (ANOVA)

**▣ I.6.19 Remark**

Consider the model I.6.15 of one-factorial ANOVA in effect representation. Then,

$$B_* = \begin{bmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & 0 & \cdots & \cdots & 0 \\ \mathbb{1}_{n_2} & 0 & \mathbb{1}_{n_2} & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ \mathbb{1}_{n_p} & 0 & \cdots & \cdots & 0 & \mathbb{1}_{n_p} \end{bmatrix} \in \mathbb{R}^{n_\bullet \times (p+1)}$$

has $\operatorname{rank}(B_*) = p$.

Notice that the **side conditions** specified by the vector $v$ mentioned in Remark I.6.13 are linearly independent of the rows of $B_*$, that is,

- ● $v = (0, n_1, \ldots, n_p)'$ and $v'\beta_* = \sum_{i=1}^{p} n_i \alpha_i = 0$
- ● $v = e_{1,p+1}$ and $v'\beta_* = \mu = 0$
- ● $v = e_{i,p+1}$ for some $i \in \{2, \ldots, p+1\}$, and $v'\beta_* = \alpha_i = 0$

# Analysis of variance (ANOVA)

**▶ I.6.20 Remark (one-factorial ANOVA)**

  **❯** We consider the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_p \longleftrightarrow H_1 : \exists\, i, j \text{ mit } \mu_i \neq \mu_j.$$

  **❯** Alternatively, $H_0$ can be written as

$$H_0 : \alpha_1 = \cdots = \alpha_p = 0 \longleftrightarrow H_1 : \exists\, i \text{ mit } \alpha_i \neq 0.$$

**▶ I.6.21 Theorem**

The LSE in Model I.6.14 is given by $\widehat{\beta}_i = \widehat{\mu}_i = \dfrac{1}{n_i} \sum\limits_{j=1}^{n_i} Y_{ij} = \overline{Y}_{i\bullet}, \ 1 \leqslant i \leqslant p$. For the effect model, the LSE are given by

$$\widehat{\mu} = \frac{1}{n_\bullet} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij} = \overline{Y}_{\bullet\bullet}, \quad \widehat{\alpha}_i = \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}, 1 \leqslant i \leqslant p.$$

Notice that $\widehat{\alpha}_i, \ 1 \leqslant i \leqslant p$, satisfies the equation $\sum\limits_{i=1}^{n} n_i \widehat{\alpha}_i = 0$.

In a one-factorial ANOVA, we have:

$$\text{SST} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet} + \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left[ (Y_{ij} - \overline{Y}_{i\bullet})^2 + 2(Y_{ij} - \overline{Y}_{i\bullet})(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) + (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \right]$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2 + 2 \sum_{i=1}^{p} \underbrace{\left[ \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet}) \right]}_{=0} (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) + \sum_{i=1}^{p} n_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2 + \sum_{i=1}^{p} n_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 = \text{SSE} + \text{SSR}.$$

where

- $\text{SST} = \sum\limits_{i=1}^{p} \sum\limits_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2$
- $\text{SSE} = \sum\limits_{i=1}^{p} \sum\limits_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2$
- $\text{SSR} = \sum\limits_{i=1}^{p} n_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$

# Geometrical interpretation

fictitious example
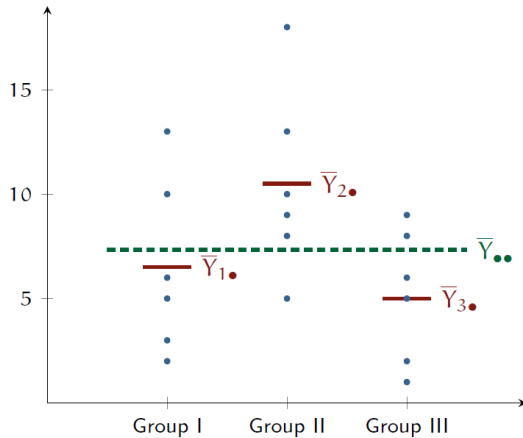
# One-factorial ANOVA: F-test

Using Theorem I.4.32, we get:

> ▶ **I.6.22 Theorem**
> Consider the NoLM of one-factorial ANOVA as given in Definition I.6.14. Let $J_0 = \mathbb{1}_{n \times n}/n$ and $J = B(B'B)^{-1}B'$. Then,
>
> ① $J, J_0$ are orthogonal projectors with $JJ_0 = J_0J = J_0$
> ② $\text{rank}(J) = p$, $\text{rank}(J_0) = 1$, $\text{rank}(J - J_0) = p - 1$
> ③ $SSR, SSE$ are independent statistics
> ④ $SST/\sigma^2 \sim \chi^2(n - 1)$
> ⑤ $SSR/\sigma^2 \sim \chi^2(p - 1)$, $SSE/\sigma^2 \sim \chi^2(n - p)$
> ⑥ $\dfrac{SSR/(p - 1)}{SSE/(n - p)} \sim F(p - 1, n - p)$

# ANOVA-table for a one-factorial ANOVA & F-test

| Source of variation | degrees of freedom | sum of squares |
|---|:---:|:---:|
| treatment | $p - 1$ | SSR |
| residual | $n - p$ | SSE |
| total | $n - 1$ | SST |

Then, a $(1 - \alpha)$-level statistical test for $H_0$ is given by the decision rule

$$\text{Reject } H_0 \text{ if } F = \frac{SSR/(p-1)}{SSE/(n-p)} > F(p-1, n-p)$$

For the data in Example I.6.8, we get the ANOVA-table ($n = 27$, $p = 3$):

| Source of variation | degrees of freedom | sum of squares | variance estimate |
|---|---|---|---|
| treatment | 2 | $SSR = 3746{,}17$ | $\widehat{\sigma}^2_{SSR} = 1873{,}08$ |
| residual | 24 | $SSE = 1826{,}50$ | $\widehat{\sigma}^2_{SSE} = 76{,}10$ |
| total | 26 | $SST = 5572{,}67$ | $\widehat{\sigma}^2_{SST} = 214{,}33$ |

For the test statistics, we get $\frac{\widehat{\sigma}^2_{SSR}}{\widehat{\sigma}^2_{SSE}} = \frac{1873{,}08}{76{,}10} = 24{,}61$. Since $F_{0{,}99}(2{,}24) = 5{,}61$ this yields

- ⊙ the null hypothesis can be rejected at the level of significance $\alpha = 1\%$.
- ⊙ a difference of the means can be shown.
- ⊙ However: **Where is the difference?**

**▣ I.6.24 Remark**

- ❯ The previous results are based on assumptions which have to be checked before applying the tests:
  - ❯ Homoscedasticity (☞ Levene's test)
  - ❯ normal distribution (☞ Shapiro-Wilks' test)

- ❯ In order to check, which treatment effects are non-zero; various methods can be used.

  If the null hypothesis in a one-factorial ANOVA model is rejected, then (a-posteriori) multiple comparisons cab be applied to test which treatment effect is non-zero. The following tests can be used:
  - ❯ Tukey (see next section)
  - ❯ Scheffé (see next section)
  - ❯ Bonferroni (see next section)
  - ❯ Newman-Keuls

# Two-factorial ANOVA

An ANOVA model can also be discussed for more than one factor.

- For two factors $A$ and $B$, the model reads (in effect representation and balanced design)

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n.$$

- factor $A$ has $a$ levels, factor $B$ has $b$ levels.
  - $\alpha_i$: effect of treatment $A$
  - $\beta_j$: effect of treatment $B$
  - $\gamma_{ij}$: interaction
- To ensure identifiability, we add the following $a + b + 2$ **side conditions**:

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = 0, j = 1, \ldots, b, \quad \sum_{j=1}^{b} \gamma_{ij} = 0, i = 1, \ldots, a. \qquad \text{(I.13)}$$

## ▸ I.6.25 Example

In order to study the influence of the factors fertilizer and soil quality on crop growth, different fertilizers $A \in \{1, \ldots, a\}$ are applied to fields with a different soil quality $B \in \{1, \ldots, b\}$. Then, the crop $Y$ is measured for combinations $(i, j) \in A \times B$ of fertilizer and soil quality.
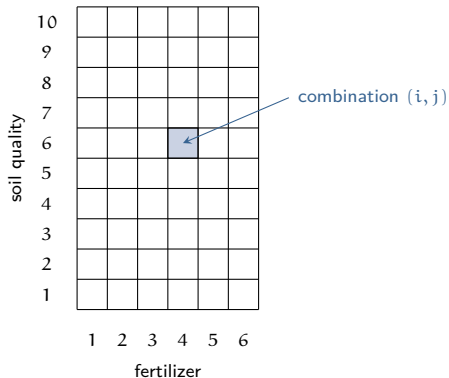
**⊠ I.6.26 Remark**

- ❯ Under the side conditions (I.13), a unique LSE exists. In fact, using Theorem I.6.17, we get similar results to those obtained for one-factorial ANOVA. One gets

$$\widehat{\mu} = \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\alpha}_i = \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\beta}_j = \overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\gamma}_{ij} = \overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}.$$

- ❯ Alternatively, the model representation

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad k = 1, \ldots, n.$$

  with a full rank design matrix B can be discussed. In this case,

$$\widehat{\mu}_{ij} = \overline{Y}_{ij\bullet} = \frac{1}{n} \sum_{k=1}^{n} Y_{ijk}, \quad i = 1, \ldots, a, j = 1, \ldots, b,$$

  defines the LSE for the parameters. All these estimators are unbiased for the respective parameter.

- ❯ The following testing problems are usually considered:
  - ❯ $H_0 : \alpha_i = 0, i \in \{1, \ldots, a\}$ (no effect of treatment A)
  - ❯ $H_0 : \beta_j = 0, j \in \{1, \ldots, b\}$ (no effect of treatment B)
  - ❯ $H_0 : \gamma_{ij} = 0, i \in \{1, \ldots, a\}, j \in \{1, \ldots, b\}$ (no interaction)

- ❯ For more information, we refer to the literature.

# Part I: Linear Models

## Chapter I.6

## Multiple Comparisons

# Topics

**To be discussed...**

- Tukey's test
- Scheffé's test
- Bonferroni correction
- Details and proofs are provided in the literature.

# Tukey's test (honest significant difference method)

Consider the test problem

$$H_0 : \alpha_1 = \cdots = \alpha_p = 0 \quad \longleftrightarrow \quad H_1 : \exists\, i, j : \alpha_i \neq \alpha_j$$

- ❯ **Tukey's range test** for multiple comparisons is based on the (studentized) statistic

$$T = \frac{\max_{i,j \in \{1,\ldots,p\}}\{|\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}|\}}{\sqrt{\text{SSE}/(n-p)}},$$

where a **balanced design is assumed**.

- ❯ Using the order statistics $\overline{Y}_{(1\bullet)} \leqslant \cdots \leqslant \overline{Y}_{(p\bullet)}$, $T$ can be written as

$$T = \frac{\overline{Y}_{(p\bullet)} - \overline{Y}_{(1\bullet)}}{\sqrt{\text{SSE}/(n-p)}}.$$

- ❯ The corresponding distribution is called **studentized range distribution** with parameter $p$ and $n - p$ (Notation: $T(p, n - p)$). Quantile tables are available in the literature.

# Tukey's test

For the test statistic T, we have

$$T = \frac{\max_{i,j \in \{1,\ldots,p\}}\{|\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}|\}}{\sqrt{SSE/(n-p)}} > T_{1-\alpha}(p, n-p)$$

$$\iff |\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}| > T_{1-\alpha}(p, n-p)\sqrt{SSE/(n-p)} \text{ for a pair } i, j \in \{1,\ldots,p\}$$

Thus, for pairs $(i, j)$ satisfying this condition, we get

$$0 \notin \left[\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet} - T_{1-\alpha}(p, n-p)\sqrt{\frac{SSE}{n-p}}, \overline{Y}_{i\bullet} - \overline{Y}_{j\bullet} + T_{1-\alpha}(p, n-p)\sqrt{\frac{SSE}{n-p}}\right]$$

☞ $(1-\alpha)$-**level confidence interval for the difference** $\mu_i - \mu_j$

Four tire types $A, B, C, D$ are evaluated w.r.t. their braking characteristic; we observe the stopping distance StopDist at a speed of 60 mile/h.

| No. | StopDist | tire | No. | StopDist | tire |
|-----|----------|------|-----|----------|------|
| 1 | 391 | A | 13 | 435 | C |
| 2 | 374 | A | 14 | 415 | C |
| 3 | 416 | A | 15 | 403 | C |
| 4 | 363 | A | 16 | 418 | C |
| 5 | 353 | A | 17 | 434 | C |
| 6 | 381 | A | 18 | 425 | C |
| 7 | 394 | B | 19 | 422 | D |
| 8 | 413 | B | 20 | 378 | D |
| 9 | 398 | B | 21 | 409 | D |
| 10 | 396 | B | 22 | 447 | D |
| 11 | 428 | B | 23 | 417 | D |
| 12 | 402 | B | 24 | 389 | D |

- left diagram: Boxplots of variable StopDist for factors

- right diagram: means for factors
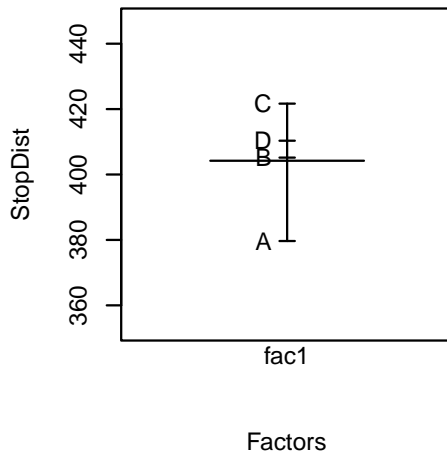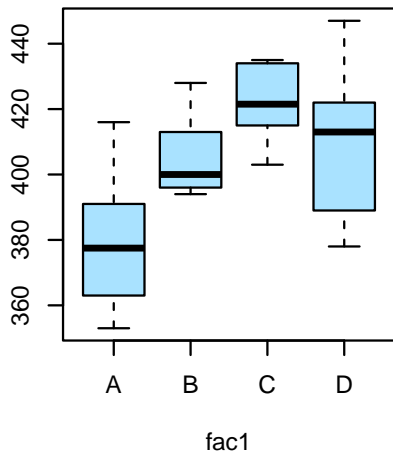
## ▶ I.6.27 Example

### ● ANOVA

```
                 Df  Sum Sq  Mean Sq  F value   Pr(>F)
     tire         3    5673   1891.0    5.328  0.00732**
     Residuals   20    7099    354.9

     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

☞ $H_0$ is rejected

### ● Multiple comparison: Tukey's test

```
              Tukey multiple comparisons of means
               95% family-wise confidence level

          Fit: aov(formula = StopDist   tire)

     $tire
              diff        lwr        upr       p adj
     B-A  25.500000  -4.9446409  55.94464  0.1213153
     C-A  42.000000  11.5553591  72.44464  0.0049515
     D-A  30.666667   0.2220258  61.11131  0.0479540
     C-B  16.500000 -13.9446409  46.94464  0.4464584
     D-B   5.166667 -25.2779742  35.61131  0.9637307
     D-C -11.333333 -41.7779742  19.11131  0.7273681
```
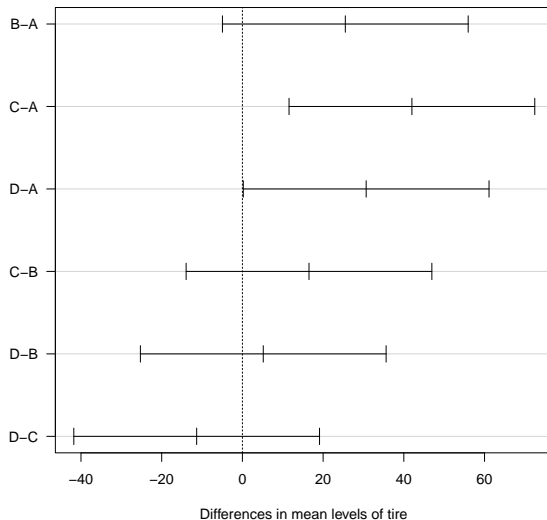
**95% family–wise confidence level**

Differences in mean levels of tire

# Scheffé's test

- As an alternative to Tukey's test, Scheffé's test can be used. It can also be used in cases of unbalanced designs. It is based on the decision rule

$$\text{Reject } H_0 \text{ if } |\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}| > S_{i,j;\alpha} \text{ for a pair } i, j \in \{1, \ldots, p\}$$

where $S_{i,j;\alpha} = \sqrt{(p-1)\widehat{\sigma}_{SSE}^2 \left( \dfrac{1}{n_i} + \dfrac{1}{n_j} \right) F_{1-\alpha}(p-1, n-p)}$.

- Scheffé's test can also be applied to a set of contrasts $\mathbf{c}_1, \ldots, \mathbf{c}_\ell$ (i.e., $H_0 : \mathbf{c}_i'\boldsymbol{\beta} = 0, i = 1, \ldots, \ell$).
- in this case, replace

  - $\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}$ by $\sum\limits_{k=1}^{p} c_k \overline{Y}_{k\bullet}$
  - $\dfrac{1}{n_i} + \dfrac{1}{n_j}$ by $\sum\limits_{k=1}^{p} \dfrac{c_k^2}{n_k}$.

- For multiple comparisons, we have $\mathbf{c}_{ij} = e_{i,p} - e_{j,p}$, $1 \leqslant i < j \leqslant p$, with standard basis vectors $e_{i,p}, e_{j,p}$.

# Scheffé's test

> **I.6.29 Example**

In Example I.6.8, we obtain with $F_{0,95}(2,24) = 3,40$:

$$i = 1, j = 2 : \qquad \sqrt{2\widehat{\sigma}_{SSE}^2 \left(\tfrac{1}{8} + \tfrac{1}{10}\right) F_{0,95}(2,24)} = 10{,}791$$

$$i = 1, j = 3 : \qquad \sqrt{2\widehat{\sigma}_{SSE}^2 \left(\tfrac{1}{8} + \tfrac{1}{9}\right) F_{0,95}(2,24)} = 11{,}054$$

$$i = 2, j = 3 : \qquad \sqrt{2\widehat{\sigma}_{SSE}^2 \left(\tfrac{1}{10} + \tfrac{1}{9}\right) F_{0,95}(2,24)} = 10{,}452$$

⊙ $|\overline{y}_{1\bullet} - \overline{y}_{2\bullet}| = 25{,}775$, $|\overline{y}_{1\bullet} - \overline{y}_{3\bullet}| = 25{,}819$, $|\overline{y}_{2\bullet} - \overline{y}_{3\bullet}| = 0{,}044$

☞ $\mu_1 \neq \mu_2$ and $\mu_1 \neq \mu_3$, respectively

☞ no significant difference between $\mu_2$ and $\mu_3$ is detectable

# Bonferroni method

- Testing problem $H_0 : \mathbf{c}_i' \boldsymbol{\beta} = 0, i = 1, \dots, \ell$
- Decision rule:

$$\text{Reject } H_0 \text{ if } \frac{\left( \sum_{k=1}^p c_{ki} \overline{Y}_{k\bullet} \right)^2 / \sum_{k=1}^p \frac{c_{ki}^2}{n_k}}{\text{SSE}/(n-p)} > F_{1-\alpha/\ell}(1, n-p) \text{ for some } i \in \{1, \dots, \ell\}.$$

- For multiple comparisons, choose $\mathbf{c}_{ij} = e_{i,p} - e_{j,p}$, $1 \leqslant i < j \leqslant p$, with standard basis vectors $e_{i,p}, e_{j,p}$.
- **Proof**: Use Bonferroni inequality for (arbitrary) sets $A_1, \dots, A_\ell$:

$$P\left( \bigcup_{i=1}^\ell A_i \right) \leqslant \sum_{i=1}^\ell P(A_i).$$