Analysis of Variance

Cathy Maugis-Rabusseau

INSA Toulouse / IMT
GMM 116
cathy.maugis@insa-toulouse.fr

2021-2022

Outline

1. Introduction

Context - Notation

- ANOVA = analysis of variance
- Aim: Explain a **quantitative variable** Y using **qualitative** explanatory variables called **factors**
- The modalities of a factor = **levels** (sub-groups in the sample)

Context - Notation

- Here we will not address the issue of **experimental design**, just this vocabulary:

Definition

1. A **block** of an experimental design = group of observations associated to a combination of controlled factors
2. An experimental design is called **full** if there is at least one observation in each block
3. An experimental design is called **repeated** if there are several observations per block
4. An experimental design is called **balanced** if there is the same number of observations per block

Outline

2 One-way ANOVA

Outline

2. **One-way ANOVA**
   - Context and Example
   - Regular model
   - Singular model
   - Predictions, residuals and variance
   - Confidence interval
   - Test: effect of the factor?

Context

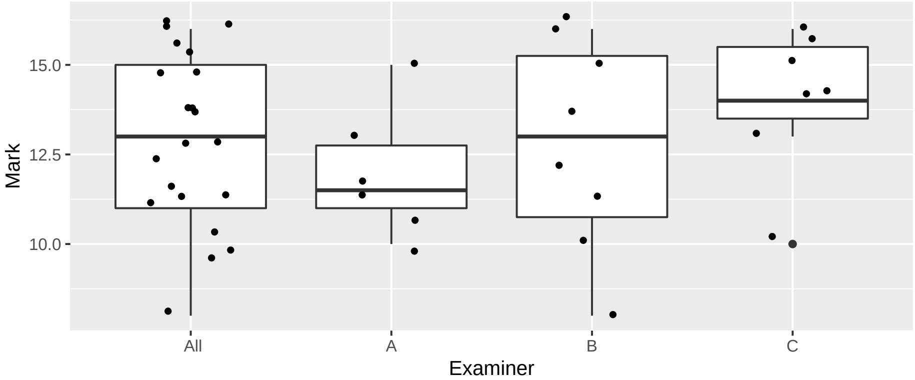
- Data: One quantitative response variable Y and **one** factor having I levels
- Notation:
  - Y_{ij} = value for individual j in group i (level of the factor)
  - Group i has n_i individuals
  - $Y_{i.}$ is the mean value for group i: $Y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$
  - $n = \sum_{i=1}^{I} n_i$ is the total number of individuals
- Question: potential effect of the factor on the response Y ? $\Leftrightarrow$ Difference of the average response per group

Example

- We are interested in the marks obtained by students in an oral examination.
- Is there a potential effect of the examiner on the mark obtained?

Examiner (i)	A	B	C
Mark Y_{ij}	10, 11, 11	8, 10, 11, 12	10, 13, 14, 14
	12, 13, 15	14, 15, 16, 16	15, 16, 16
Number n_i	6	8	7
Average $Y_{i.}$	12	12.75	14

Example



Outline

2 **One-way ANOVA**

- Context and Example
- Regular model
- Singular model
- Predictions, residuals and variance
- Confidence interval
- Test: effect of the factor?

Regular model

- **Regular model**:

$$\begin{cases} Y_{ij} = m_i + \varepsilon_{ij}, \ \forall i = 1, \cdots I, \ \forall j = 1, \cdots, n_i \\ \\ \varepsilon_{ij} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

$$Y = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1_{n_1} & 0_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & 0_{n_2} & \cdots & 0_{n_2} \\ 0_{n_3} & 0_{n_3} & 1_{n_3} & \cdots & 0_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{n_I} & 0_{n_I} & 0_{n_I} & \cdots & 1_{n_I} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_I \end{pmatrix} + \varepsilon$$

with $\varepsilon \sim \mathcal{N}_n \left(0_n, \sigma^2 I_n\right)$.

- Unknown parameters: $\theta = (m_1, \ldots, m_I)'$ $[k = I]$ and σ^2.

Estimation of θ

- $X'X = diag(n_1, \ldots, n_I)$ is invertible $\Rightarrow$ regular model
- $\hat{\theta} = (X'X)^{-1}X'Y$ thus $\widehat{m}_i = Y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$

```
anReg<-lm(Marks~Exam -1)
summary(anReg)
```

```
Call:
lm(formula = Marks ~ Exam - 1)

Residuals:
   Min     1Q Median     3Q    Max 
 -4.75  -1.00   0.00   2.00   3.25 

Coefficients:
      Estimate Std. Error t value Pr(>|t|)    
ExamA  12.0000     0.9789   12.26 3.58e-10 ***
ExamB  12.7500     0.8478   15.04 1.23e-11 ***
ExamC  14.0000     0.9063   15.45 7.88e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.398 on 18 degrees of freedom
Multiple R-squared:  0.9716,    Adjusted R-squared:  0.9668 
F-statistic:   205 on 3 and 18 DF,  p-value: 4.226e-14
```

Estimation of θ

```python
import statsmodels.api as sm
from statsmodels.formula.api import ols
anRegpy = ols('Marks ~ Exam-1', data=Datapy).fit();
anRegpy.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Marks   R-squared:                       0.115
Model:                            OLS   Adj. R-squared:                  0.017
Method:                 Least Squares   F-statistic:                     1.170
Date:                Mar, 12 oct 2021   Prob (F-statistic):              0.333
Time:                        23:09:09   Log-Likelihood:                -46.546
No. Observations:                  21   AIC:                             99.09
Df Residuals:                      18   BIC:                             102.2
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Exam[A]       12.0000      0.979     12.258      0.000       9.943      14.057
Exam[B]       12.7500      0.848     15.039      0.000      10.969      14.531
Exam[C]       14.0000      0.906     15.447      0.000      12.096      15.904
==============================================================================
Omnibus:                        0.750   Durbin-Watson:                   1.388
Prob(Omnibus):                  0.687   Jarque-Bera (JB):                0.773
Skew:                          -0.356   Prob(JB):                        0.679
Kurtosis:                       2.386   Cond. No.                         1.15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Outline

2 One-way ANOVA

Singular model

- For interpretation reasons, we may be interested in an other parametrization

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \forall i = 1, \cdots I,\ \forall j = 1, \cdots, n_i$$

where

- $\mu =$ average effect
- $\alpha_i = m_i - \mu =$ differential effect of group i.
- But this model is over-parameterized [I+1 parameters]

  $\Rightarrow$ one constraint is required to have an identifiable model

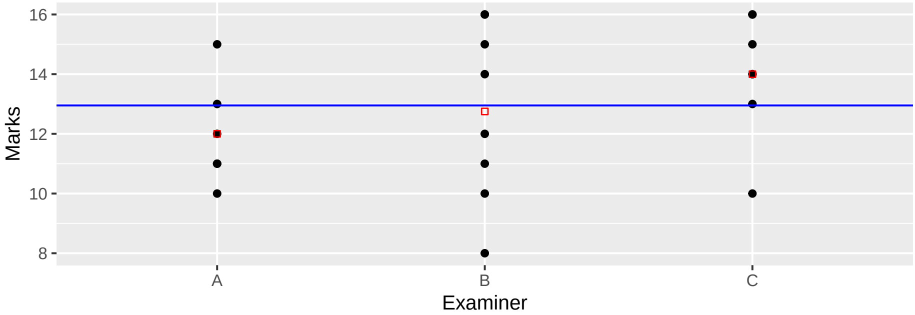
  - Orthogonal constraint : $\sum_{i=1}^{I} n_i \alpha_i = 0$
  - By default in R: $\alpha_1 = 0$

Estimation of θ - Orthogonal constraints

- Model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \theta = (\mu, \alpha_1, \ldots, \alpha_I)'$$

- The orthogonal constraint $\sum_{i=1}^{I} n_i \alpha_i = 0$
- Estimators: $\begin{cases} \widehat{\mu} = Y_{..} \\ \widehat{\alpha}_i = Y_{i.} - Y_{..} \end{cases}$

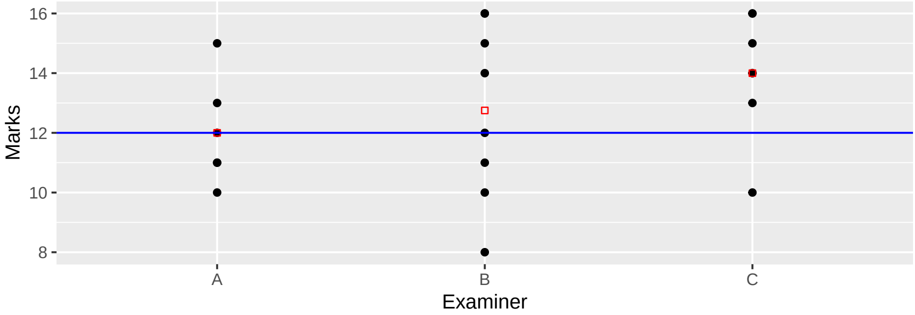


Estimation of θ - By default in R

- Model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \theta = (\mu, \alpha_1, \ldots, \alpha_I)'$$

- The constraint by default in R: $\alpha_1 = 0$
- Estimators: $\begin{cases} \widehat{\mu} = Y_{1.} \\ \widehat{\alpha}_i = Y_{i.} - Y_{1.} \end{cases}$



Example

```
anSing <- lm(Notes~Exam,data=Data)
summary(anSing)
```

```

Call:
lm(formula = Notes ~ Exam, data = Data)

Residuals:
   Min     1Q Median     3Q    Max 
 -4.75  -1.00   0.00   2.00   3.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.0000     0.9789  12.258 3.58e-10 ***
ExamB         0.7500     1.2950   0.579    0.570    
ExamC         2.0000     1.3341   1.499    0.151    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.398 on 18 degrees of freedom
Multiple R-squared:  0.115, Adjusted R-squared:  0.01669 
F-statistic:  1.17 on 2 and 18 DF,  p-value: 0.333
```

Example

```
anSingpy = ols('Marks ~ Exam', data=Datapy).fit()
anSingpy.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Marks   R-squared:                       0.115
Model:                            OLS   Adj. R-squared:                  0.017
Method:                 Least Squares   F-statistic:                     1.170
Date:                Mar, 12 oct 2021   Prob (F-statistic):              0.333
Time:                        23:09:11   Log-Likelihood:                -46.546
No. Observations:                  21   AIC:                             99.09
Df Residuals:                      18   BIC:                             102.2
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     12.0000      0.979     12.258      0.000       9.943      14.057
Exam[T.B]      0.7500      1.295      0.579      0.570      -1.971       3.471
Exam[T.C]      2.0000      1.334      1.499      0.151      -0.803       4.803
==============================================================================
Omnibus:                        0.750   Durbin-Watson:                   1.388
Prob(Omnibus):                  0.687   Jarque-Bera (JB):                0.773
Skew:                          -0.356   Prob(JB):                        0.679
Kurtosis:                       2.386   Cond. No.                         4.00
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
"""
```

Outline

Predictions, residuals and variance

- Predicted values: $\widehat{Y} = X\widehat{\theta}$

$$\Leftrightarrow \forall i,\ \forall j,\quad \widehat{Y}_{ij} = \widehat{m}_i = \widehat{\mu} + \widehat{\alpha}_i = Y_{i.}$$

- Residuals: $\widehat{\varepsilon} = Y - \widehat{Y}$

$$\Leftrightarrow \forall i,\ \forall j,\quad \widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{Y}_{ij} = Y_{ij} - Y_{i.}$$

- Estimator of the variance σ^2:

$$\begin{aligned}
\widehat{\sigma^2} &= \frac{\|Y - \widehat{Y}\|^2}{n - I} = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{Y}_{ij})^2 \\
&= \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{\varepsilon}_{ij})^2 = \frac{SSR}{n-I}
\end{aligned}$$

Properties

Proposition

- The mean of residuals per block is null: $\forall i = 1, \cdots, I,\ \frac{1}{n_i} \sum_{j=1}^{n_i} \widehat{\varepsilon}_{ij} = 0$.
- The mean of residuals is null: $\frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \widehat{\varepsilon}_{ij} = 0$.
- $\frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \widehat{Y_{ij}} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}$.
- $cov(\widehat{\varepsilon}, \widehat{Y}) = 0$.
- $var(Y) = var(\widehat{Y}) + var(\widehat{\varepsilon})$.

Proof in exercise

Decomposition of the variance

- **Between-group variance** :

$$var(\widehat{Y}) = \sum_{i=1}^{I} \frac{n_i}{n} (Y_{i.} - Y_{..})^2$$

- **Within-group variance** (or residual variance):

$$var(\widehat{\varepsilon}) = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2 = \frac{1}{n} \sum_{i=1}^{I} n_i var_i(Y)$$

- The equality $var(Y) = var(\widehat{Y}) + var(\widehat{\varepsilon})$:

$$\text{Total Variance} = \text{Between Variance} + \text{Within Variance}.$$

Coefficient of determination R^2

The coefficient of determination R^2 is the ratio of the between-group variance on the total variance:

$$R^2 = \frac{var(\widehat{Y})}{var(Y)} = 1 - \frac{var(\widehat{\varepsilon})}{var(Y)}.$$

It is a measure of connection between a quantitative variable and a qualitative variable.

Remarks:

1. $R^2 = 1 \leftrightarrow \widehat{\varepsilon} = 0_n \leftrightarrow \forall j = 1, \cdots, n_i, Y_{ij} = Y_{i.}$
2. $R^2 = 0 \leftrightarrow var(\widehat{Y}) = 0 \leftrightarrow \forall i = 1, \cdots, I, Y_{i.} = Y_{..}$

Outline

2 One-way ANOVA

- Context and Example
- Regular model
- Singular model
- Predictions, residuals and variance
- Confidence interval
- Test: effect of the factor?

Confidence interval for m_i

- m_i is estimated by $\widehat{m_i} = Y_{i.} \sim \mathcal{N}(m_i, \sigma^2/n_i)$

since Y_{ij} i.i.d $\mathcal{N}(m_i, \sigma^2)$ $(j = 1, \ldots, n_i)$

- By Cochran's theorem, $\frac{(n-I)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-I)$ and $\widehat{m_i} \perp\!\!\!\perp \widehat{\sigma}^2$

- We deduce that

$$\sqrt{n_i}\,\frac{\widehat{m_i} - m_i}{\widehat{\sigma}} \sim \mathcal{T}(n-1).$$

- Let $t_{n-I,1-\alpha/2}$ be the $(1-\alpha/2)$ quantile of $\mathcal{T}(n-1)$. Then,

$$\mathbb{P}\left(m_i \in \left[\widehat{m_i} \pm t_{n-I,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2}{n_i}}\,\right]\right) = 1 - \alpha.$$

Example

- With R:

```
anReg<-lm(Marks~Exam -1)
confint(anReg)
```

```
          2.5 %   97.5 %
ExamA  9.943313 14.05669
ExamB 10.968857 14.53114
ExamC 12.095878 15.90412
```

- With python:

```
anRegpy.conf_int(alpha=0.05)
```

```
                0          1
Exam[A]   9.943313  14.056687
Exam[B]  10.968857  14.531143
Exam[C]  12.095878  15.904122
```

Exercise

- Build the following confidence intervals:

```
anSing<-lm(Marks~Exam)
confint(anSing)
```

```
                 2.5 %    97.5 %
(Intercept)  9.9433129 14.056687
ExamB       -1.9707414  3.470741
ExamC       -0.8027921  4.802792
```

```
anSingpy.conf_int(alpha=0.05)
```

```
                  0          1
Intercept   9.943313  14.056687
Exam[T.B] -1.970741   3.470741
Exam[T.C] -0.802792   4.802792
```

Indications: determine the law of $\hat{\mu} = Y_{1.}$ *and* $\hat{\alpha}_i = Y_{i.} - Y_{1.}$

Outline

2. **One-way ANOVA**
   - Context and Example
   - Regular model
   - Singular model
   - Predictions, residuals and variance
   - Confidence interval
   - Test: effect of the factor?

Test: effect of the factor?

- Testing procedure:

$$\mathcal{H}_0 : m_1 = m_2 = \cdots = m_I = m \Longleftrightarrow \forall i = 1, \cdots, I,\ \alpha_i = 0$$

against

$$\mathcal{H}_1 : \exists (i, i') \text{ such that } m_i \neq m_{i'}.$$

- Fisher's test of the sub-model:

$$(M_0) : Y_{ij} = m + \varepsilon_{ij} \text{ with } \widehat{m} = Y_{..} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}$$

$$(M_1) : Y_{ij} = m_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Test: effect of the factor?

- Fisher's statistics:

$$F = \frac{\sum_{i=1}^{I} n_i (Y_{i.} - Y_{..})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2/(n-I)} = \frac{SSE/(I-1)}{SSR/(n-I)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-I),$$

- We reject $\mathcal{H}_0$ if $F > f_{1-\alpha, I-1, n-I}$.

Example

- With R:

```r
anmequal<-lm(Marks~1)
anova(anmequal,anReg)
```

```
Analysis of Variance Table

Model 1: Marks ~ 1
Model 2: Marks ~ Exam - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 116.95
2     18 103.50  2    13.452 1.1698  0.333
```

- With Python:

```python
from statsmodels.stats.anova import anova_lm
anmequalpy = ols('Marks ~1', data=Datapy).fit();
anova_lm(anmequalpy,anRegpy)
```

```
   df_resid         ssr  df_diff    ss_diff         F    Pr(>F)
0      20.0  116.952381      0.0        NaN       NaN       NaN
1      18.0  103.500000      2.0  13.452381  1.169772  0.332952
```

Summary table of one-way anova

	df	Sum of squares	Average of squares	F	$f_{1-\alpha}$
Factor	$I-1$	$SSE = \sum_{i=1}^{I} n_i(Y_{i.} - Y_{..})^2$	$\frac{SSE}{I-1} = MSE$	$\frac{MSE}{\widehat{\sigma}^2}$	$f_{1-\alpha,I-1,n-I}$
Residual	$n-I$	$SSR = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})^2$	$\frac{SSR}{n-I} = \widehat{\sigma}^2$		
Total	$n-1$	$SST = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - Y_{..})^2$			

Outline

3 Two-way ANOVA

Outline

3 Two-way ANOVA

Example (Husson et Pagès, 2013)

In a study of factors influencing wheat yield, three varieties of wheat (L, N and NF) and two nitrogen inputs were compared (normal supply = dose 1, intensive supply = dose 2).
Three repetitions for each couple (variety, dose) were performed and the yield (in q / ha) was measured.
We are interested in the differences that could exist from one variety to another, and in the possible interactions of varieties with nitrogen inputs.

```
summary(Ble)
```

```
 Dose  Variety     Yield
 1:9   L :6    Min.   :55.65
 2:9   N :6    1st Qu.:62.82
       NF:6    Median :65.50
               Mean   :66.58
               3rd Qu.:69.75
               Max.   :79.83
```

Notation

- Two factors (qualitative explanatory variables):
  - First factor = factor A *[Dose]* with I *[=2]* levels
  - Second factor = factor B *[Variety]* with J *[=3]* levels
- Y = quantitative response variable *[wheat yield]*
  - Y_{ijk} = measure of the k-th individual for level i of factor A and level j of factor B
  - n_{ij} = nb of obs. for level i of factor A and level j of factor B
  - $Y_{ij.}$ = mean of observations for block (i, j)
- Notation:

$$Y_{i..} = \frac{1}{n_{i+}} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} Y_{ijk} \text{ with } n_{i+} = \sum_{j=1}^{J} n_{ij}$$

$$Y_{.j.} = \frac{1}{n_{+j}} \sum_{i=1}^{I} \sum_{k=1}^{n_{ij}} Y_{ijk} \text{ with } n_{+j} = \sum_{i=1}^{I} n_{ij}$$

$$Y_{...} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} Y_{ijk} \text{ with } n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j}$$

Example

Variety	L ($j=1$)	N ($j=2$)	NF ($j=3$)
Dose 1 ($i=1$) $(Y_{1,j,k})$ $n_{1j}=3$	70.35 63.59 79.83 $Y_{11.}=71.26$	62.56 58.89 55.65 $Y_{12.}=59.03$	69.45 64.84 66.12 $Y_{13.}=66.80$
Dose 2 ($i=2$) $(Y_{2,j,k})$ $n_{2j}=3$	74.97 69.12 77.18 $Y_{21.}=73.76$	58.78 64.39 60.83 $Y_{22.}=61.33$	69.85 64.89 67.15 $Y_{23.}=67.30$

- We are interested in the effect of **Variety** and **Dose** on **wheat yield** with a possible interaction between the two factors.

Outline

3. **Two-way ANOVA**
   - Introduction
   - Models
   - Estimation of θ
   - Predicted values, residuals and variance
   - Decomposition of the variability
   - Interaction plot
   - Testing procedures
   - Summary table of two-way anova

Regular model vs Singular model

- **Regular** model:

$$Y_{ijk} = m_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

but all the effects are included in m_{ij}

- **Singular** model (with interaction):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

distinction of effects ... but constraints are required for parameter estimation.

$1 + I + J + IJ$ parameters and IJ ddl thus $1 + I + J$ constraints

Parameters of the singular model

The IJ parameters m_{ij} are thus decomposed into:

- $\mu =$ centering parameter (intercept),
- α_i, $I-1$ parameters (main effect of factor A),
- β_j, $J-1$ parameters (main effect of factor B),
- γ_{ij}, $(I-1)(J-1)$ parameters (interaction effects of factors).

Orthogonal design

Proposition

In the two-way anova framework with interaction, there exist some constraints such that the model is orthogonal if and only if

$$n_{ij} = \frac{n_{i+}n_{+j}}{n}.$$

In this case, the constraints (called Type I) are

$$\sum_{i=1}^{I} n_{i+}\alpha_i = 0;\ \sum_{j=1}^{J} n_{+j}\beta_j = 0;\ \forall i,\ \sum_{j=1}^{J} n_{ij}\gamma_{ij} = 0;\ \forall j,\ \sum_{i=1}^{I} n_{ij}\gamma_{ij} = 0.$$

Other constraints

- In practice, the following constraints (Type III) may be used:

$$\sum_i \alpha_i = 0, \sum_j \beta_j = 0, \forall i, \sum_j \gamma_{ij} = 0 \text{ et } \forall j, \sum_j \gamma_{ij} = 0$$

  With these constraints, the model is orthogonal only if n_{ij} are constant.

- With R, the constraints by default are

$$\alpha_1 = \beta_1 = 0,\ \gamma_{1j} = 0 \forall j,\ \gamma_{i1} = 0, \forall i$$

In the sequel, the model is assumed to be orthogonal.

Additive two-way ANOVA

- The additive two-way ANOVA model = two-way ANOVA model without interaction

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

- Exercise : Determine the constraints such that this model is orthogonal.

Outline

3 Two-way ANOVA

Estimation - Regular model

- Model:

$$\left\{\begin{array}{l} Y_{ijk} = m_i + \varepsilon_{ijk} \\ \varepsilon_{ijk} \text{ i.i.d } \mathcal{N}(0,\sigma^2) \end{array}\right. \Leftrightarrow Y = X\theta + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

- $\theta = (m_1, \dots, m_I)'$ is estimated by $\hat{\theta} = (X'X)^{-1}X'Y$

Proposition

$$\widehat{m}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} = Y_{ij.} \sim \mathcal{N}\left(m_{ij}, \frac{\sigma^2}{n_{ij}}\right).$$

Estimation - Singular model

- Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$
- Orthogonal constraints: If $n_{ij} = \frac{n_{i+}n_{+j}}{n}$, the orthogonal constraints (type I) are

$\sum_{i=1}^{I} n_{i+}\alpha_i = 0$; $\sum_{j=1}^{J} n_{+j}\beta_j = 0$; $\forall i$, $\sum_{j=1}^{J} n_{ij}\gamma_{ij} = 0$; $\forall j$, $\sum_{i=1}^{I} n_{ij}\gamma_{ij} = 0$.

Proposition

Under the orthogonal constraints,

$$\begin{cases} \widehat{\mu} = Y_{...} \\ \widehat{\alpha}_i = Y_{i..} - Y_{...} \\ \widehat{\beta}_j = Y_{.j.} - Y_{...} \\ \widehat{\gamma}_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...} \end{cases}$$

Estimation - Singular model

- Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$
- By default in R: $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0,\ \forall i,\ \forall j$

Proposition

$$\begin{cases} \widehat{\mu} = Y_{11.} \\ \widehat{\alpha}_i = Y_{i1.} - Y_{11.} \\ \widehat{\beta}_j = Y_{1j.} - Y_{11.} \\ \widehat{\gamma}_{ij} = Y_{ij.} - Y_{i1.} - Y_{1j.} + Y_{11.} \end{cases}$$

Example

```r
anov2 = lm(Yield~Dose*Variety,data=Ble)
summary(anov2)
```

```
Call:
lm(formula = Yield ~ Dose * Variety, data = Ble)

Residuals:
   Min     1Q Median     3Q    Max 
-7.667 -2.296 -0.325  2.623  8.573 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       71.257      2.536  28.101 2.55e-12 ***
Dose2              2.500      3.586   0.697  0.49899    
VarietyN         -12.223      3.586  -3.409  0.00519 ** 
VarietyNF         -4.453      3.586  -1.242  0.23801    
Dose2:VarietyN    -0.200      5.071  -0.039  0.96919    
Dose2:VarietyNF   -2.007      5.071  -0.396  0.69928    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.392 on 12 degrees of freedom
Multiple R-squared:  0.6725,	Adjusted R-squared:  0.536 
F-statistic: 4.928 on 5 and 12 DF,  p-value: 0.01105
```

Example

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
Blepy=r.Ble
Blepy['Dose']=Blepy['Dose'].astype(str)
Blepy['Variety']=Blepy['Variety'].astype(str)
anov2Singpy = ols('Yield ~ Dose * Variety', data=Blepy).fit();
anov2Singpy.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Yield   R-squared:                       0.672
Model:                            OLS   Adj. R-squared:                  0.536
Method:                 Least Squares   F-statistic:                     4.928
Date:                Mar, 12 oct 2021   Prob (F-statistic):             0.0111
Time:                        23:09:13   Log-Likelihood:                -48.528
No. Observations:                  18   AIC:                             109.1
Df Residuals:                      12   BIC:                             114.4
Df Model:                           5
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                  71.2567      2.536     28.101      0.000      65.732      76.781
Dose[T.2]                   2.5000      3.586      0.697      0.499      -5.313      10.313
Variety[T.N]              -12.2233      3.586     -3.409      0.005     -20.037      -4.410
Variety[T.NF]              -4.4533      3.586     -1.242      0.238     -12.267       3.360
Dose[T.2]:Variety[T.N]     -0.2000      5.071     -0.039      0.969     -11.250      10.850
Dose[T.2]:Variety[T.NF]    -2.0067      5.071     -0.396      0.699     -13.056       9.043
==============================================================================
Omnibus:                        1.202   Durbin-Watson:                   2.764
Prob(Omnibus):                  0.548   Jarque-Bera (JB):                0.199
```

Outline

3 Two-way ANOVA

Predicted values, residuals and variance

- Predicted values:

$$\widehat{Y}_{ijk} = \widehat{m}_{ij} = Y_{ij.} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij}$$

- Residuals:

$$\widehat{\varepsilon}_{ijk} = Y_{ijk} - Y_{ij.}$$

- Estimator of the variance σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-IJ}\sum_{ijk}(\widehat{\varepsilon}_{ijk})^2 = \frac{1}{n-IJ}\sum_{ijk}(Y_{ijk} - Y_{ij.})^2 = \frac{SSR}{n-IJ}$$

and

$$\frac{(n-IJ)\ \widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-IJ)$$

Outline

3 Two-way ANOVA

Decomposition of the variability

- Decomposition of SST:

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - Y_{...})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(Y_{ij.} - Y_{...})^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} var_{ij}(Y)}_{\text{SSR}}$$

with $var_{ij}(Y) = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}}(Y_{ijk} - Y_{ij.})^2$.

- Under the orthogonal constraints, $SSE = SSA + SSB + SSI$

$$SSA = \sum_{i=1}^{I} n_{i+}(Y_{i..} - Y_{...})^2 = \sum_{i=1}^{I} n_{i+}(\widehat{\alpha}_i)^2$$

$$SSB = \sum_{j=1}^{J} n_{+j}(Y_{.j.} - Y_{...})^2 = \sum_{j=1}^{J} n_{+j}(\widehat{\beta}_j)^2$$

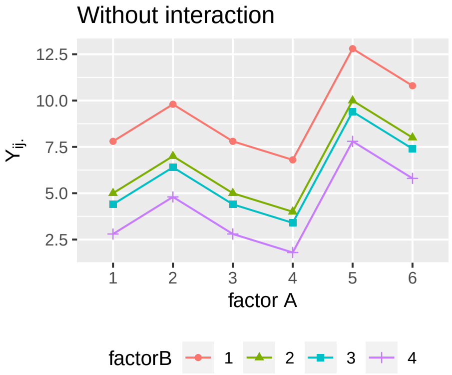
$$SSI = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\widehat{\gamma}_{ij})^2$$

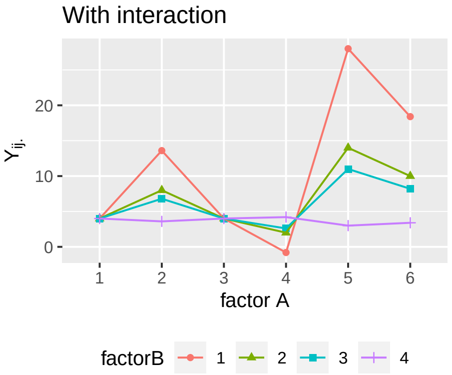
Outline

3 Two-way ANOVA

Interaction plot

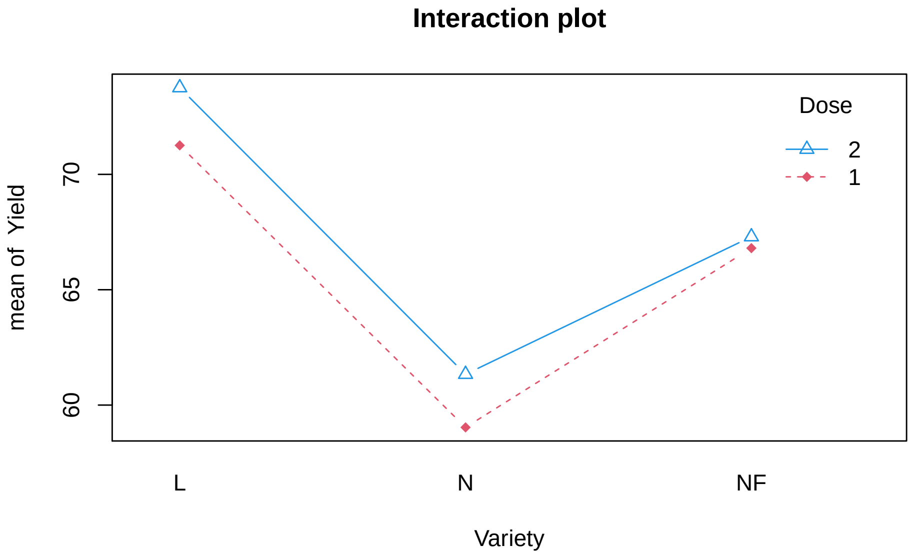
- Graphical control of the possible interaction





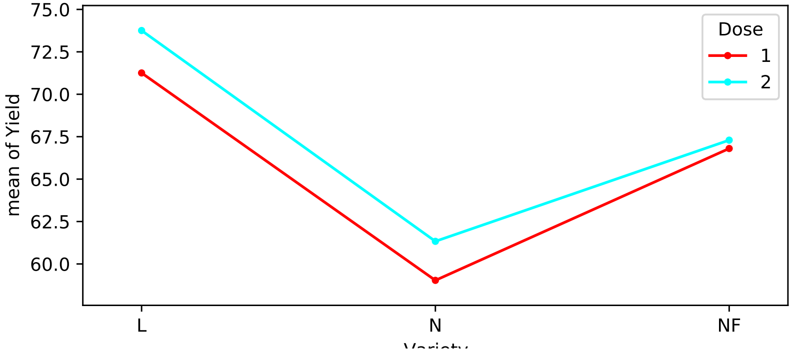
Example

```r
attach(Ble)
interaction.plot(Variety,Dose,Yield,col=c(2,4),pch=c(18,24),main="Interaction plot",type="b")
```



Example

```
from statsmodels.graphics.factorplots import interaction_plot
from matplotlib import pyplot as plt
interaction_plot(Blepy['Variety'],Blepy['Dose'],Blepy['Yield']);
plt.show()
```



Outline

3. **Two-way ANOVA**

Testing procedures

Testing procedures

- Warning: If there are interactions between the two factors, then the principal effect of each factor which constitute this interaction must be integrated into the model.
- In order to simplify the model, it is necessary to
  - test if there is an interaction effect firstly,
  - if we retain the model without interaction, we can then test the effect of each factor

Non-interaction testing

- Hypothesis:

$$\mathcal{H}_I : \gamma_{ij} = 0,\ \forall i = 1, \cdots, I,\ \forall j = 1, \cdots, J$$

- Fisher's test of sub-model:
  - $[M_1]$ $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ (model with interaction)
  - $[M_0]$ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ (additive model)
- Test statistics:

$$F = \frac{SSI/(I-1)(J-1)}{SSR/(n-IJ)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}((I-1)(J-1), n-IJ).$$

Example

- With R:

```
anov2add = lm(Yield ~Variety + Dose, data=Ble)
anova(anov2add,anov2)
```

```
Analysis of Variance Table

Model 1: Yield ~ Variety + Dose
Model 2: Yield ~ Dose * Variety
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     14 235.14
2     12 231.47  2    3.6654 0.095   0.91
```

- With python:

```
from statsmodels.stats.anova import anova_lm
anov2addpy = ols('Yield~Dose + Variety', data=Blepy).fit()
anovaResults = anova_lm(anov2addpy,anov2Singpy)
print(anovaResults)
```

```
   df_resid         ssr  df_diff   ss_diff        F    Pr(>F)
0      14.0  235.137778      0.0       NaN      NaN       NaN
1      12.0  231.472400      2.0  3.665378  0.09501  0.910041
```

Test for the effect of factor A

- Hypothesis: $\mathcal{H}_A : \alpha_i = 0, \ \forall i = 1, \cdots, I$
- Fisher's testing of sub-model:
  - $[M_1]$ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ (additive model)
  - $[M_0]$ $Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$ (one-way ANOVA)
- Test statistics:

$$F = \frac{SSA/(I-1)}{SSRAB/(n-(I+J-1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-(I+J-1)),$$

where $SSRAB =$ residual sum of squares for the additive model.

Example

- With R:

```
anovA = lm(Yield ~Variety,data=Ble)
anova(anovA,anov2add)
```

```
Analysis of Variance Table

Model 1: Yield ~ Variety
Model 2: Yield ~ Variety + Dose
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     15 249.15
2     14 235.14  1     14.01 0.8341 0.3765
```

- With python:

```
anovApy = ols('Yield~Variety', data=Blepy).fit()
print(anova_lm(anovApy,anov2addpy))
```

```
   df_resid         ssr  df_diff    ss_diff         F    Pr(>F)
0      15.0  249.147467      0.0        NaN       NaN       NaN
1      14.0  235.137778      1.0  14.009689  0.834131  0.376541
```

Test for the effect of factor B

- Hypothesis: $\mathcal{H}_B : \beta_j = 0, \ \forall j = 1, \cdots, J$
- Fisher's testing of sub-model:
  - $[M_1]$ $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ (additive model)
  - $[M_0]$ $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$ (one-way ANOVA)
- Test statistics:

$$F = \frac{SSB/(J-1)}{SSRAB/(n-(I+J-1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(J-1, n-(I+J-1)),$$

where $SSRAB$ = residual sum of squares for the additive model.

Example

- With R:

```r
anovB = lm(Yield~Dose,data=Ble)
anova(anovB,anov2add)
```

```
Analysis of Variance Table

Model 1: Yield ~ Dose
Model 2: Yield ~ Variety + Dose
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     16 692.72
2     14 235.14  2    457.58 13.622 0.0005192 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- With python:

```python
anovBpy = ols('Yield~Dose', data=Blepy).fit()
print(anova_lm(anovBpy,anov2addpy))
```

```
   df_resid         ssr  df_diff     ss_diff          F    Pr(>F)
0      16.0  692.719511      0.0         NaN        NaN       NaN
1      14.0  235.137778      2.0  457.581733  13.622108  0.000519
```

Outline

3 Two-way ANOVA

Summary table

- Two-way Anova with interaction + Orthogonal design
- Decomposition of the variance

$$SST = SSE + SSR = SSA + SSB + SSI + SSR.$$

Source	df	Sum of squares	Average of squares	F	$f_{1-\alpha}$
Factor A	$I-1$	SSA	$MSA = \frac{SSA}{I-1}$	$MSA/\widehat{\sigma}^2$	$f_{1-\alpha,I-1,n-IJ}$
Factor B	$J-1$	SSB	$MSB = \frac{SSB}{J-1}$	$MSB/\widehat{\sigma}^2$	$f_{1-\alpha,J-1,n-IJ}$
Interaction	$(I-1)(J-1)$	SSI	$MSI = \frac{SSI}{(I-1)(J-1)}$	$MSI/\widehat{\sigma}^2$	$f_{1-\alpha,(I-1)(J-1),n-IJ}$
Residual	$n-IJ$	SSR	$\widehat{\sigma}^2 = \frac{SSR}{n-IJ}$		
Total	$n-1$	SST			

Outline

Conclusion

Summary

- Know how to write an ANOVA model with one and two factors (individually and matrix), regular and singular
- Know how to distinguish a regular model from a singular model
- Know how to estimate the parameters of the ANOVA model in the regular case and in the singular case (by adapting to the chosen constraint (s))
- Know how to construct a confidence interval for a parameter of the ANOVA model
- Know how to build a procedure to test the effect of a factor, the interaction effect between factors, ... and know how to organize these tests
- Know how to interpret an interaction plot
- Know how to handle SSA, SSB, SSI, SSE, SSR in the case of an orthogonal design.