Advanced Statistics
Part one: Parametrical Statistics

M. Pérez-Casany

Dept. of Statistics and Operations Research and DAMA-UPC
Technical University of Catalunya

First Semester: Third Session

3. Linear Models: analysis of variance

Assume that one is interested in quantifying the influence of

- 2 CPU schedulling algorithms,
- 3 CPU types and
- 3 types of workloads in a given metric

Those variables are called *factors* and each one takes an small number of cathegories also known as *levels*.

The objective of the ANOVA is to quantify the impact of each factor in the response variable, and to determine if there exists significative differences between the levels of a given factor.

3. Linear Models: Analysis of variance

The simplest model is the **One factor ANOVA.**

Example: We want to compare the *Execution Time* (Y) based on a different *workload types* (X)

Factor has a levels, and from i-thm level one has n_i observations

Level	Observations	
1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	$\overline{y}_{1.}$
2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	$\overline{y}_{2.}$
$\vdots$		
a	$y_{a1}, y_{a2}, \cdots, y_{an_a}$	$\overline{y}_{a.}$

Where $\overline{y}_{i.} = 1/n_i \cdot \sum_{j=1}^{n_i} y_{ij}$,
$\overline{y}_{..} = 1/N \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij} = 1/N \sum_{i=1}^{a} n_i \overline{y}_{i.}$ and $N = \sum_{i=1}^{a} n_i$.

3. Linear Models: analysis of variance

If $n_i = n \;\; \forall i \in \{1, \cdots, a\}$ the experiment is said to be **balanced.**

The appropiate model is:

$$y_{ij} = \mu + \tau_i + e_{ij}, \;\; i = 1, \cdots, a \;\, j = 1, \cdots, n_i$$

where

- $\mu_i = \mu + \tau_i$ is the expected response under level i, and
- e_{ij}, the errors, are indep. and Normal$(0, \sigma^2)$ distributed.

3. Linear models: analysis of variance

Goal: To see if there exists significative differences between the levels of the Factor.

That is to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \;\; vs \;\; H_1 : \exists (i,j) \; \mu_i \neq \mu_j,$$

at a given significance level α. Which is equivalen to:

$$H_0 : \tau_i = 0 \, \forall i \;\; vs \;\; H_1 : \exists i \quad \tau_i \neq 0,$$

Observation: The decision rule *not to reject* H_0 *when the null hyp. of all the 2 by 2 comparisons has not been rejected* is not appropiate. Since, in that case,

$$P(\text{reject } H_0 | H_0 \text{ is true}) = 1 - (1-\alpha)^{a(a-1)/2} \geq \alpha$$

The total variability in the data must be partitioned in two parts, the one caused by the different levels of the factor, and the one due to error.

Denoting by $\overline{y}_{..}$ the total sample mean,

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{..})^2=\sum_{i=1}^{a}n_i\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2$$

Which is usuallly denoted by:

$$SS_T=SS_A+SS_E$$

These sums of squares have, respectively $N-1$, $a-1$ and $N-a$ degrees of freedom

3. Linear models: Analysis of variance

The **one-way ANOVA table** is equal to:

	SE	d. f.	MSE	F
Factor A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$F_0 = \frac{SS_A/(a-1)}{SS_E/(N-a)}$
Error	SS_E	$N-a$	$\frac{SS_E}{N-a}$	
Total	SS_T	$N-1$	$\frac{SS_T}{N-1}$	

Given that SS_A and SS_E are indep., and that under H_0 $(a-1)SS_A/\sigma^2 \sim \chi^2_{a-1}$ we have the following decision rule:

$$\text{reject } H_0 \ \text{ when } F_0 \geq F_{\alpha,a-1,N-a}$$

Important: This rule generalizes the comparation of two means under normality and homocedasticity.

3. Linear models: Analysis of variance

Parameter estimation: applying MLE method, we want to minimize:

$$f(\mu, \tau_1, \cdots, \tau_a) = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - (\hat{\mu} + \hat{\tau}_i))^2$$

Differenciating one has that:

$$\frac{\partial f}{\partial \mu} = (-2)\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$$

$$\frac{\partial f}{\partial \tau_i} = (-2)\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0 \;\; \forall i = 1..a$$

$a+1$ parameters, a equations because $\sum_{i=1}^{a} \frac{\partial f}{\partial \tau_i} = \frac{\partial f}{\partial \mu}$

3. Linear models: Analysis of variance

To solve the system, some parameter restrictions (constraints) must be assumed.

The most common are:

- Corner point restrictions, usually $\tau_1 = 0$ or $\tau_a = 0$.
- Add up to zero $\sum_{i=1}^{a} \tau_i = 0$

IMPORTANT:

- Parameter estimations will change depending on the constraints. Its interpretation also changes.
- Predicted values ($\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i$) will be the same indep. on the constraints.
- In the one way anova $\hat{\mu}_i - \hat{\mu}_j$ is also uniquely estimated regardless of the constraint.

3. Linear models: Analysis of variance

In matrix form, the LM for the one way anova may be written as:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{ij} \\ \vdots \\ Y_{an_a} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & & \\ 1 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_a \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{ij} \\ \vdots \\ e_{an_a} \end{pmatrix}$$

Observation: We have what is called a **collinearity** problem.

A corner point restriction is equivalent to supress one of the a last columns and solve the problem. Parameters may also be estimated assuming an add to zero constraint.

3. Linear models: Analysis of variance

Assuming $\sum_{i=1}^{a} \tau_i = 0$,

$$\hat{\mu} = \overline{y}_{..}, \quad \hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{..}$$

Parameter interpretation:

- μ is interpreted as the overall mean, the mean if one doesn't know from which level the observation comes from.
- τ_i is the deviation from the overall mean if one knows that the observtion belongs to level i.

Predicted value for level i, $\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}$

Predicted difference of two levels $\hat{\mu}_i - \hat{\mu}_j = \hat{\tau}_i - \hat{\tau}_j = \overline{y}_{i.} - \overline{y}_{j.}$

3. Linear models: Analysis of variance

Assuming $\tau_1 = 0$,

$$\hat{\mu} = \overline{y}_{1.}, \;\; \hat{\tau}_1 = 0, \;\; \hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{1.}$$

Parameter interpretation:

- μ is interpreted as the mean response at the first level.
- τ_i is the deviation with respect to the first level obtained at the i-thm level.

Predicted value for level i, $\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}$

Predicted difference of two levels $\hat{\mu}_i - \hat{\mu}_j = \hat{\tau}_i - \hat{\tau}_j = \overline{y}_{i.} - \overline{y}_{j.}$

3. Linear models: Analysis of variance

Estimation of the variance parameter:

It is verified that

$$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{i.})^2}{\sigma^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i} e_{ij}^2}{\sigma^2} \sim \chi^2_{N-a}$$

Applyint the moment estimation method, one has that:

$$E\left(\frac{SSE}{\sigma^2}\right) = N - a \quad \Longrightarrow \hat{\sigma} = \sqrt{\frac{SSE}{N-a}} = \sqrt{MSE}$$

known as **square root of the mean square error**

3. Linear models: Analysis of variance

In the case where H_0 is rejected, we can go further to determine where does the true differences really exist.

There are methods that allow:

1) to perform two by two comparaisons

$$H_0 : \mu_i = \mu_j \;\; vs \;\; H_1 : \mu_i \neq \mu_j$$

The most important ones are: Newman-Kéuls, Duncan Tukey and LSD,

2) two compare

$$H_0 : \sum_{i=1}^{a} c_i \, \mu_i = 0 \;\; vs \;\; H_1 : \sum_{i=1}^{a} c_i \, \mu_i \neq 0$$

where $\sum_{i=1}^{a} c_i = 0$, which is known as a **contrast**.

3. Linear models: Analysis of variance

Visual diagnosis test of the model assumptions

To visualy check the model assumptions:

- Perform a **normal quantile-quantile plot** for the residuals, to check normality. It should be linear.
- Perform an scatter plot of **residuals versus predicted**, to check that tere is no trend in the residuals or their spread.
- Perform the scatter plot of **predicted versus le covariate levels**. Again no trends in the spread must be observed.

Comment: If the relative magnitude of errors is smaller than the response by an order of magnitude or more, the trends may be ignored.

3. Linear models: Analysis of variance

The **Two factor ANOVA with crossed Factors.**

Example: We want to compare the *Execution Time* (Y) based on a different *workload types* (X_1) and b different *CPU* (X_2).

Factors A and B have a and b levels respect. and for each combination (i, j) one has n_{ij} observations.

Balanced case $n_{ij} = n \; \forall (i, j)$

A|B	1	2	$\cdots$	b	
1	$y_{111} \cdots y_{11n}$	$y_{121} \cdots y_{12n}$	$\cdots$	$y_{1b1} \cdots y_{1bn}$	$\overline{y}_{1..}$
2	$y_{211} \cdots y_{21n}$	$y_{221} \cdots y_{22n}$	$\cdots$	$y_{2b1} \cdots y_{2bn}$	$\overline{y}_{2..}$
$\vdots$					
a	$y_{a11} \cdots y_{a1n}$	$y_{a21} \cdots y_{a2n}$	$\cdots$	$y_{ab1} \cdots y_{abn}$	$\overline{y}_{a..}$
	$\overline{y}_{.1.}$	$\overline{y}_{.2.}$		$\overline{y}_{.b.}$	$\overline{y}_{...}$

3. Linear models: Analysis of variance

The **two-way ANOVA**:

Additive model

$$y_{ijk} = \mu + \tau_i + \beta_j + e_{ijk}$$

assumptions:

- $e_{ijk} \sim \mathrm{N}(0, \sigma^2)$,
- indep. of errors
- $\sum_i \tau_i = 0 \;\; \sum_j \beta_j = 0$.

Observation 1: It is possible to have $n = 1$.

Observation 2: In the case where the factor effects may be multiplicative, it is necessary to consider $\log(Y)$ as a response r.v.

$$\mu_{ijk} = \mu \cdot \tau_i \cdot \beta_j \iff \log(\mu_{ijk}) = \log(\mu) + \log(\tau_i) + \log(\beta_j)$$

3. Linear models: analysis of variance

Goal: To see if there exists significative differences between the levels of each one of the factors.

That is to test, for a fixed value of α,

$$H_0^1 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \;\; vs \;\; H_1^1 : \exists i \; \tau_i \neq 0,$$

and

$$H_0^2 : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \;\; vs \;\; H_1^2 : \exists i \; \beta_i \neq 0,$$

3. Linear models: Analysis of variance

The **two-way ANOVA table** without interaction is equal to:

	SE	d. f.	MSE	F
Factor A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$F_0^1 = \frac{SS_A/(a-1)}{SS_E/(N-(a+b-1))}$
Factor B	SS_B	$b-1$	$\frac{SS_B}{b-1}$	$F_0^2 = \frac{SS_B/(b-1)}{SS_E/(N-(a+b-1))}$
Error	SS_E	$N-a-b+1$	$\frac{SS_E}{N-a-b+1}$	
Total	SS_T	$N-1$	$\frac{SS_T}{N-1}$	

Decision rules:

$$\text{reject } H_0^1 \text{ if } F_0^1 \geq F_{\alpha,a-1,N-a}$$

$$\text{reject } H_0^2 \text{ if } F_0^2 \geq F_{\alpha,b-1,N-a}$$

3. Linear models: Analysis of variance

In the case where one thinks that the effect of a level of one factor depends on the level of the other factor with which it is combined, one has to consider the model with **interaction** term. That is:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\gamma)_{ij} + e_{ijk}$$

with assumptions:

- $e_{ijk} \sim \mathrm{N}(0, \sigma^2)$,
- indep. of errors
- $\sum_i \tau_i = 0,\ \sum_j \beta_j = 0, \forall\, i,\ \sum_j \gamma_{ij} = 0, \forall\, j,\ \sum_i \gamma_{ij} = 0,$

Graphically, interaction is observed in the lak of parallelism in the poligonal plots corresponding to the mean cells.

3. Linear models: Analysis of variance

In interaction model we also want to test if:

$$H_0^3 : \gamma_{ij} = 0, \forall (i,j) \;\; vs \;\; H_1^3 : \exists (i,j) \; \gamma_{ij} \neq 0,$$

If the interaction is significative, it means that the two factors do not act independently.

Observation: The interaction model requires that $n \geq 2$, otherwise we do not have degrees of freedom for the error term.

3. Linear models: Analysis of variance

The **two-way ANOVA table** with interaction is equal to:

	SE	d. f.	MSE	F
Factor A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$F_0^1 = \frac{SS_A/(a-1)}{SS_E/(ab(n-1)))}$
Factor B	SS_B	$b-1$	$\frac{SS_B}{b-1}$	$F_0^2 = \frac{SS_B/(b-1)}{SS_E/(ab(n-1)))}$
AB	SS_{AB}	$(a-1)(b-1)$	$\frac{SS_{AB}}{(a-1)(b-1)}$	$F_0^3 = \frac{SS_{AB}/(a-1)(b-1)}{SS_E/(ab(n-1))}$
Error	SS_E	$ab(n-1)$	$\frac{SS_E}{ab(n-1)}$	
Total	SS_T	$N-1$	$\frac{SS_T}{N-1}$	

Similar decision rules

3. Linear models: Analysis of variance

If the interaction is statistically different from zero, the same tools used to compare the levels of a factor may be used, once the level of the other factor has been fixed.

This means that, for instance, the two by two mean comparaisons for Factor A may be perform when j takes a fixed value, and the other way around.

3. Linear models: Analysis of variance

Two way ANOVA with nested factors

Factor B is *nested* in Factor A, when the levels of B change by changing the level of A.

EXAMPLE: We want to compare several computers in terms of *execution times of several benchmarks* (Y). One has a total number of six different computers (FACTOR A), and for each computer 4 different benchmarks (FACTOR B) are tested.

If the 4 benchmarks are the same for all the computers, the two factors are crossed. Otherwise, FACTOR B is nested in FACTOR A.

Observation 1: It exists the corresponding ANOVA table for nested factors.

Observation 2: Interaction has no sense with nested factors.

3. Linear models: Analysis of variance

In general,

The amount of variability in the data explained by the assumed model is computed as:

$$R^2 = \frac{SS_{model}}{SS_T}\%$$

where SS_{model} contains the sums of squares of all the terms in the total variability descomposition, with the exception of the error sum of squares.

As larger is R^2 better the model explains the data.

There also exists the *adjunted*-R^2 that penalize models with more parameters.

$$R^2_{adj} = 1 - \frac{(1 - R^2)\,(N-1)}{N-p-1}$$

3. Linear models: Analysis of variance

Other type of sums of squares in ANOVA

There exists several types of sums of squares.

- Type I sums of squares: they add the total sum of squares. Its value depend on the order in which the factors are introduced. Also called *sequential* sums of squares.
- Type II Computed for main effects in all different possible permutations of them, assuming no interaction.
- Type III Test one main effect after the other and the interaction being included.
- Sums of squares arising out of mutually orthogonal sets of functions.