Analysis Of Variance

In the experimental research **the simultaneous comparison between the means of more than two groups** is frequent. They are composed of subjects that are undergone to different treatments or of data that have been collected in different conditions.

In order to highlight all the possible important differences among the means, **it is not correct to use Student's t-test to repeat the analysis as many times as the possible comparisons are** in couples among every single group.

With the Student's t-test, **only a part of the data is used** and therefore the power is lower.

In addition, the α probability that has been chosen in order to accept the null hypothesis, or rather **the probability to commit a type I error** (that means to refuse the null hypothesis when it is true), **is valid only for every single comparison**.

If there are many comparisons, the whole probability that at least one of them is significant because of the chance, is greater.

If the H_0 null hypothesis is valid, the probability that none of the comparisons is, by chance, significant is:

$$(1-\alpha)^k$$

where k is the number of the comparisons that have been done.

If, for example, we make 10 comparisons among the means of the groups drawn by chance from the same population and, for each of them α is equal to 0.05, the probability that none of the comparisons is significant decreases from 0.95 to around **0.60**.

As consequence, the whole probability that at least one results to be significant, because of random fluctuations, moves from 0.05 to **0.40**.

In different words, if we do k comparisons with the Student's t-test, each of them with α probability, the whole α' probability to commit at least one type I error (that the test refuses the null hypothesis even though it is true) becomes:

$$\alpha' = 1 - (1 - \alpha)^k$$

The analysis of variance, with an apparently words paradox, allows the comparison between two or more means.
It allows the simultaneous comparison among them, **keeping unchanged the α whole set probability**.

It represents, in addition, the core of the industrial statistics, as, for example, of the Capability or of the GAGE R&R.

The H_0 null hypothesis and the H_A alternative hypothesis have a more general formula, compared to the comparison between two means:

$$\boldsymbol{H_0: \ \mu_1 = \mu_2 = \cdots = \mu_p}$$

$$\boldsymbol{H_A: \exists\, i, j \mid \mu_i \neq \mu_j \quad (i, j = 1, \ldots, p)}$$

t of Beyond the mathematical notation, the alternative hypothesis is that **at least one** of the means is different from the others.
The developed methodology to check the statistical significance of the diffences among the arithmetical means of a series of the groups, is called analysis of variance and it is marked as **ANOVA**. It uses **the F(p,q) distribution where p and q are needed parameters to specify it**.

Example.

We have an establishment with 3 parallel production lines (A, B, C) that produces silicon wafers.
It wants to be checked that the mean resistivity of each of them is the same.

This is a scheme of hypothesis:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_{TOT}$$

$$H_A : \exists \mu_j \neq \mu_{TOT}$$

In the analysis of variance, the root or the cause of the data variations is called **experimental factor** or **treatment**; it can have:

- ⋆ more quantitative **levels**, as long as it is possible to discretize them;
- ⋆ more qualitative **levels** (or **methods**).

Every unit or observation inside the experimental observations' levels are **replication o repetition**.

In order to calculate the mean and the variance, every group must be made of at least two replications.

The easiest analysis of variance model, that can be seen as an extension of the Student's t-test with many independent samples, is called **one criteria classification (one way)**, or **one factor**.

For the statistics analysis, in this model **it is not required that the various groups have the same number** (n_j) **of observations or replications**.

The observations' presentation methodology, already encoded, implies that experimental data, that have been collected, are orderly carried according to the underlying table.

	TREATMENTS' MODALITY OR LEVELS				
	T_1	T_2	T_3	$\ldots$	T_p
EXPERIMENTAL UNITIES OR REPLICATIONS	x_{11}	x_{12}	x_{13}	$\ldots$	x_{1p}
	x_{21}	x_{22}	x_{23}	$\ldots$	x_{2p}
	x_{31}	x_{32}	x_{33}	$\ldots$	x_{3p}
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$	$\ldots$	$x_{n_p p}$
Treatments' means	$\overline{x}_{.1}$	$\overline{x}_{.2}$	$\overline{x}_{.3}$	$\ldots$	$\overline{x}_{.p}$
Overall mean	$\overline{\overline{x}}$				

The **single observation** is carried with two indexes. The first one regards the position occupied inside the group. The second one regards the treatment or the group it belongs to: x_{ij}.

The **mean of every group or single treatment** is overlined, with a point instead of the first index and with the group index: $\overline{x}_{.j}$. Sometimes the point as first index is omitted: $\overline{x}_j$.

The **overall mean** of all the data is written with a double line and without indexes: $\overline{\overline{x}}$.

Starting from these three quantities, deviances (**Sum of Squares**) and variances (**Mean of Squares**), useful for the analysis, are estimated.

The **total deviance** or SS_T (Total Sum of Squares) is the sum, over all observations, of the squared differences of each observation from the overall mean.

$$SS_T = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{ij} - \overline{\overline{X}})^2$$

The SS_T defines the total variability of the sample.

The **deviance between treatments** or SS_B (Between Sum of Squares) is the sum of squared residuals of every group's mean from the overall mean, multiplied the number of data of the correspondent group.

$$SS_B = \sum_{j=1}^{p} n_j (\overline{X}_{.j} - \overline{\overline{X}})^2$$

The SS_B defines the variability among groups independently of the dispersion of the single observations inside every group.

The **deviance within treatments** or SS_W (Within Sum of Squares), also called **error**, defines the sum of squared residuals of every value of the mean of its group.

$$SS_W = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{.j})^2$$

The SS_W defines the variability within every single group independently of the position of the group around the sample's overall mean.

It is possible to demonstrate that:

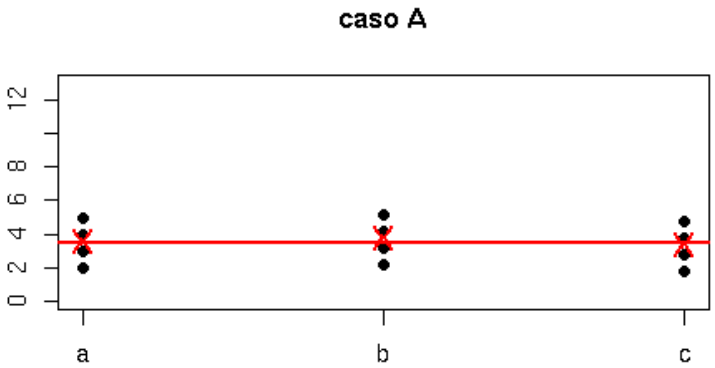
$$SS_T = SS_B + SS_W$$

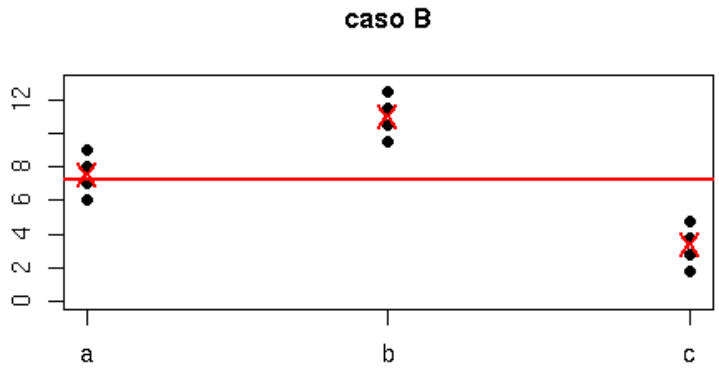
This equation shows that a sample's total variance, whose sample units are divided in groups, can be partitioned in two components:

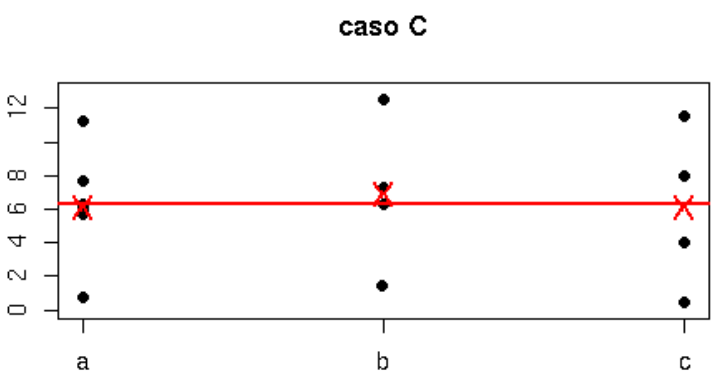
- ⋆ a component **"between"**, given by the variability among the groups;
- ⋆ a component **"within"**, given by the variability within the groups.

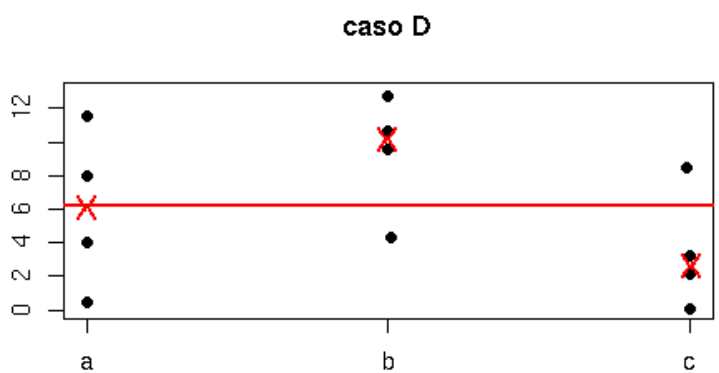
It is possible to make a comparison with the Pythagorean theorem.

Some examples:









- ⋆ **CASE A:** SS_W and SS_B low
- ⋆ **CASE B:** SS_W low but SS_B high
- ⋆ **CASE C:** SS_W high but SS_B low
- ⋆ **CASE D:** SS_W and SS_B high

In other words, given a sample with a certain total variability, its variability components "within" and "between" cannot change independently because their sum is bound to the total variability.

In the experimental practice, the reseacher is interested in finding a criteria that allows him to choose between the two underexposed situations can be considered true.

(A) The means, calculated within the subgroups, can be considered similar among them. As consequence, the treatments to which the sample units have been undergone **did not** have any effects on the answer variable.

(B) The means, calculated within the subgroups, cannot be considered similar among them. As consequence, the treatments **have** had some effects on the answer variable.

In statistical words, the points (A) and (B) can be rewritten like follows:

(A) All the sample units, independently of the subgroups, can be considered as they come from populations with the same mean. In other words, it is accepted that the factor does not have any effects on the dependent variable. In formal words, this is expressed as the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k.$$

(B) **Not** all the sample units can be considered as they come from the same population or from populations with the same mean. In formal words, this is expressed as the alternative hypothesis:

H_A : It exists at least one μ_i different from the others.

A possible way to distinguish the case (A) and the case (B), or rather to choose if the null hypothesis H_0 has to be accepted or refused, consists in comparing the $\boldsymbol{SS_B}$ and $\boldsymbol{SS_W}$ between them, taking into consideration the fact that $SS_T = SS_B + SS_W$.

If the sample means calculated within the subgroups are similar among them, then the quantity $\boldsymbol{SS_B}$ tends to be really small.

As it is true that $SS_T = SS_B + SS_W$, then the deviance within the subgroups $\boldsymbol{SS_W}$ explains the great majority of the total deviance $\boldsymbol{SS_T}$ and therefore $\boldsymbol{SS_W}$ can be considered high compared to $\boldsymbol{SS_B}$.

If, on the contrary, the sample means calculated within the subgroups are different among them, then the quantity $\boldsymbol{SS_B}$ tends to be high.

As it is true that $SS_T = SS_B + SS_W$, then the variance within the subgroups $\boldsymbol{SS_B}$ explain the great majority of the total deviance $\boldsymbol{SS_T}$ and therefore $\boldsymbol{SS_B}$ can be considered high compared to $\boldsymbol{SS_W}$.

A possible way to distinguish between the cases (A) and (B) consists in comparing, through their ratio, the two quantities SS_B and SS_W.

(A) If the ratio is **small** then the means, calculated within the subgroups, can be considered **similar** among them.

(B) If the ratio is **high** then the means, calculated within the subgroups, can be considered **different** among them.

Unluckily, the measurements **SS** grow when the number of observations increases as they represent sum of squares' (measurements).

A possible way to avoid this problem is to consider, instead of the sum of squares (**SS**), the mean of squares (**MS**, Mean of Squares).

These measurements **MS** are obtained dividing the measurements **SS** for the correspondent degrees of freedom (**df**).

The degrees of freedom are estimated by the number of independent sums required by the calculation of the relative deviances.

- ⋆ For the **sum of squares total**, SS_T, where the sum is extended to all the n data, the df_T are $n - 1$.

- ⋆ For the **sum of squares treatments**, SS_B, where the sum is extended to p groups, the df_B are $p - 1$.

- For all the sum of squares within the groups or error, $\boldsymbol{SS_W}$, the sum is extended to all the data within each group. For this reason $\boldsymbol{df_W}$ are:

$$df_W = \sum_{j=1}^{p} (n_j - 1) = n - p$$

where $\boldsymbol{n_j}$ is the size of jth group, or rather the sum of the df calculated within each group.

The following sum is valid also for the degrees of freedom:

$$df_T = df_B + df_W.$$

The **MS** measurements are calculated as:

$$MS_B = \frac{SS_B}{df_B}$$

$$MS_W = MS_E = \frac{SS_W}{df_W}$$

The measurement $\boldsymbol{MS_T}$ is usually not calculated because in the analysis of variance it is not important.

From the definition of **MS** it comes that these measurements are (estimations of) **variances**.

In particular:

- $\boldsymbol{MS_B}$ is a correct estimator of the **variance "between"** (σ_B^2); or rather:

$$E\left[MS_B\right] = \sigma_B^2.$$

- $\boldsymbol{MS_W}$ is a correct estimator the the **variance "within"** (σ_W^2); or rather:

$$E\left[MS_W\right] = \sigma_W^2.$$

Similarly to what we said before as regards the **SS**, the ratio between MS_B and MS_W, called **F**

$$F = \frac{MS_B}{MS_W}$$

can be considered as an index to evaluate if it is possible to accept

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_p$$

or if H_0 it has to be refused in favor of the alternative hypothesis H_A:

$$H_A : \exists\, i, j \mid \mu_i \neq \mu_j \quad (i, j = 1, \ldots, p)$$

- ⋆ If **F** is small, or rather MS_B is small compared to MS_W, then we accept H_0.
- ⋆ If **F** is high, or rather MS_B is high compared to MS_W, then we refuse H_0 in favor of H_A.

The **F** value results to be always positive because it is the ratio between two measurements that are always positive. In other words, the **F** value is always included between zero and infinite.

If H_0 is true, or rather if the means of all the groups are equal, **it is possible to demonstrate that both MS_B (estimator of σ_B^2) and MS_W (estimator of σ_W^2) are two independent estimations of the variance σ_ε^2** (we will see this definition later).

With the H_0 hypothesis

$$\frac{MS_B}{\sigma_B^2} = \frac{MS_B}{\sigma_\varepsilon^2} \sim \chi^2_{df_B} \qquad \text{and} \qquad \frac{MS_W}{\sigma_W^2} = \frac{MS_W}{\sigma_\varepsilon^2} \sim \chi^2_{df_W}$$

and therefore the ratio

$$F = \frac{\frac{MS_B}{\sigma_\varepsilon^2}}{\frac{MS_W}{\sigma_\varepsilon^2}} = \frac{MS_B}{MS_W}$$

tends to be distributed as a **F of Fisher-Snedecor** with $\boldsymbol{df_B = (p-1)}$ and $\boldsymbol{df_W = \sum_{j=1}^{p} (n_j - 1) = (n-p)}$ degrees of freedom: $\boldsymbol{F_{(p-1),(n-p)}}$.

$$F = \frac{MS_B}{MS_W} \sim F_{df_B, df_W}$$

If $\boldsymbol{H_0}$ is true, then the ratio **F** has to be "close" to one, or "small", as it is composed by two estimations of the same quantity.

As consequence, if we choose a probability level $\boldsymbol{\alpha}$ we calculate $(1-\alpha)$-*th* quantile of the F distribution: $\boldsymbol{F_{(p-1),(n-p);(1-\alpha)}}$. So that:

We accept $\boldsymbol{H_0}$ if $F \leq F_{(p-1),(n-p);(1-\alpha)}$

We refuse $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_A}$ if $F > F_{(p-1),(n-p);(1-\alpha)}$

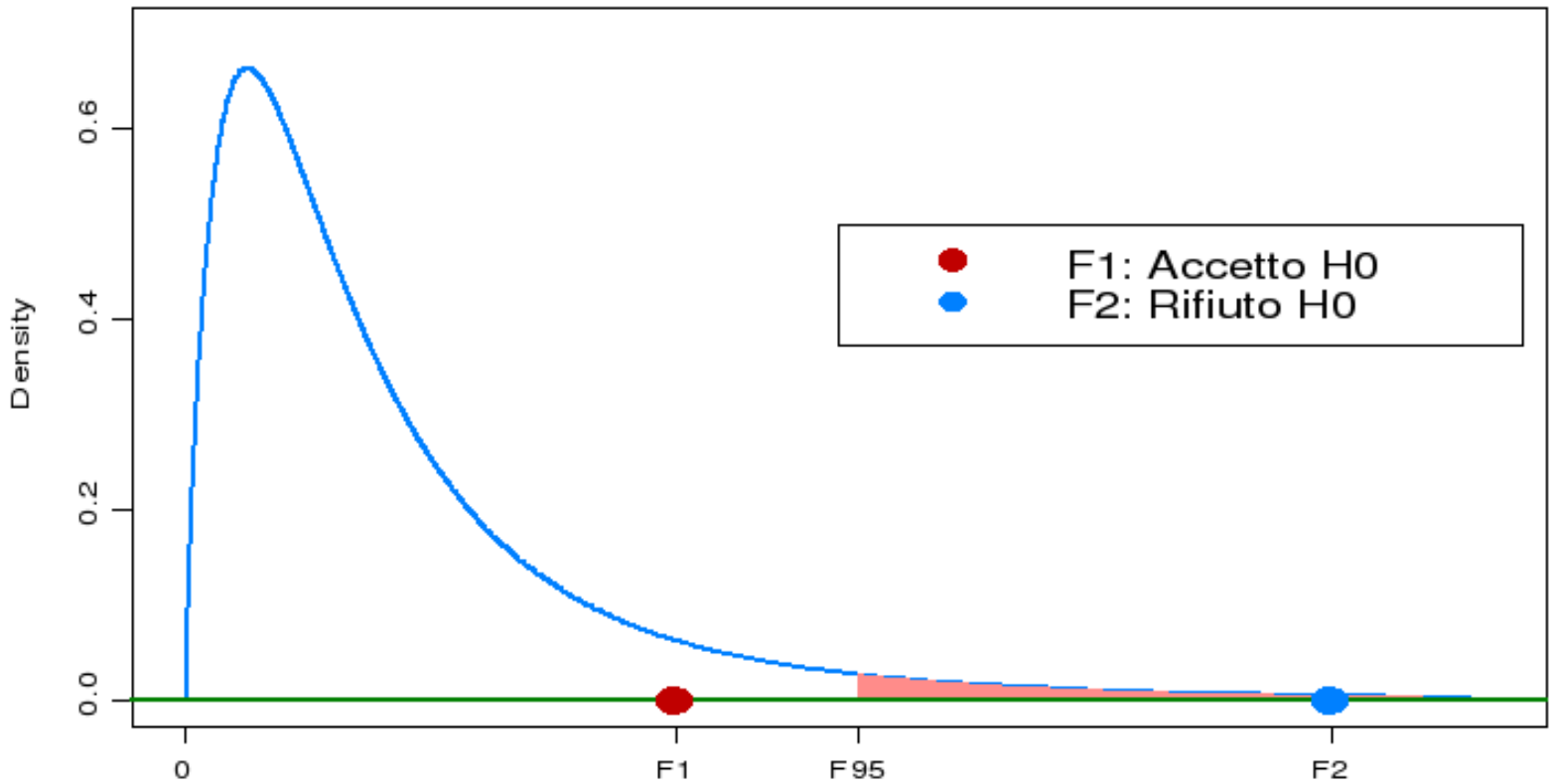
The quantile $F_{(p-1),(n-p);(1-\alpha)}$ leaves at its left a probability area equal to $\boldsymbol{1-\alpha}$, and similarly:

We accept $\boldsymbol{H_0}$ if $P(F_{(p-1),(n-p)} > F) \geq \alpha$

We refuse $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_A}$ if $P(F_{(p-1),(n-p)} > F) < \alpha$

In general $\boldsymbol{\alpha} = 0.05 = 5\%$.

In graphic terms:



The majority of the statistics packages report the result of the analysis of variance like follows:

	SS	df	MS	F	$P(F)$
$Factor$	SS_B	df_B	$MS_B = \frac{SS_B}{df_B}$	$F = \frac{MS_B}{MS_W}$	$P(F_{df_B, df_W} > F)$
$Error$	SS_W	df_W	$MS_W = \frac{SS_W}{df_W}$		
$Total$	SS_T	df_T			

In conclusion, the value of real interest for the result of the analysis of variance is the **P(F)**. This value gives the criteria of choice between the null hypothesis and the alternative hypothesis according to a α value prior set.

The α values commonly used are, as in all the tests, 0.05 and 0.01.

Esempio.

Twelve catalytic reactions are done using three different catalysts (**levels**): A, B and C.

Four reactions are assigned to each catalyst (**replications**).

For each reactions, it is measured the percentage of catalysed material.

	Catalysts		
	A	**B**	**C**
Percentage of catalysed material	99.2	99.5	99.3
	99.4	99.4	99.4
	99.3	99.2	99.5
	99.1	99.1	99.3
Mean of treatments	**99.250**	**99.300**	**99.375**
Total mean	**99.308**		

The calculated mean within the three subgroups do not seem to be especially different among them.
The analysis of variance table (ANOVA Table) results to be:

	SS	**df**	**MS**	**F**	**P(F)**
Catalysts	0.032	2	0.016	0.802	0.478
Residuals	0.177	9	0.020		
Total	0.209	11			

As $P(F) = P(F_{2,9} > 0.802) = 0.478$ we can accept the null hypothesis for the confidence level $\mathbf{1 - \alpha}$, with $\boldsymbol{\alpha}$ equal both to 0.05 and 0.01.

Or rather, the means, calculated within the subgroups (according to the catalysts) are not very different among them. As consequence, it has no influence on the variable answer (percentage of catalysed material) to use the catalysts A, B or C .

The analysis of variance is based on the additive effects of the factors that have been taken into consideration. In the easiest model, that only considers one factor at two or more levels, every single observation $\boldsymbol{X_{ij}}$ can be written as

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

as it is determined:

- ⋆ by **the overall mean μ** that defines the size of the experiment;
- ⋆ by **the factor α_j of the treatment** jth;
- ⋆ by a **random factor ε_{ij}** called **residual** or **experimental error**.

It is important to remember that **error** is not synonym for mistake, but **it shows the effect of a variety of unknown phenomenon, not valuated or not checked** in the experiment.

In this model, the **effect α of the treatment** is measured as:

$$\alpha_j = \mu_j - \mu$$

where $\boldsymbol{\mu_j}$ is the mean of the treatment jth and $\boldsymbol{\mu}$ is the overall mean, and where $\varepsilon_{ij} \in \varepsilon \sim N(0, \sigma_\varepsilon)$.

In other words, the residuals of the model are distributed with null mean and constant variance, independently of the j group they belong to and of the i observation taken into consideration.

The normality condition of the residual of the model is the validity condition of the model itself.

In descriptive terms, a similar formulation of the analysis of variance model implies that every observation is the result:

- ⋆ of the population overall mean (μ),
- ⋆ more an effect that comes from the treatment (α_j),
- ⋆ more a random effect independent of the treatment (ε_{ij}).

If the third condition is not checked, another systematic effect (not random) arises in the model beyond the treatment taken into consideration.

In other words, the difference among the means, calculated within the subgroups, is not explicable by that specific model, or rather, that model is not suitable to explain the data taken into consideration.

Going from the theoretical enunciation to the experimental data, it is possible to write that every single data $\boldsymbol{X_{ij}}$ of a specific treatment is given by

$$X_{ij} = \overline{\overline{X}} + a_j + e_{ij},$$

with

$$a_j = \overline{X}_j - \overline{\overline{X}},$$

where:

- $\overline{\overline{X}}$ is the overall mean;
- a_j is the effect of the treatment j;
- e_{ij} indicates the unknown factors.

It is clear that:

- ⋆ The overall mean $\overline{\overline{X}}$, calculated on the sample, represents an estimation of the population mean μ;
- ⋆ the effect of the treatment a_j represents an estimation of the jth treatment α_j;
- ⋆ the e_{ij} values, in order for the model to be considered valid, have to be considered as a random sample of the population of the ε;

As consequence, the analysis of variance is valid if, and only if, the empirical residuals can be considered as a sample from a normal population.

Graphic tools, as the normal probability plot or statistical tests such as the Anderson Darling test to check the normality, applied to the residuals series, are compulsory for the model validation.

The assumptions of validity of the analysis of variance depend on the error e_{ij} that:

- ⋆ must be independent among them;
- ⋆ must be normally distributed.

Moreover, the variances of the various groups must be homogeneous.

The reasons why the validation of the conditions at the basis of the model are made after the estimation of the model itself, are practice and not methodological.

The advantage of this approach will be clear when we will show models of analysis of variance with more than one factor.

When the null hypothesis is refused, a further research to **highlight which of the relative means of the various treatments are significantly different** is carried on.

In order to do that **the multiple comparison method** is used; these comparisons can be:

- ⋆ simples, or rather conducted on couple of single treatments;
- ⋆ complexes, or rather among groups of treatments.

The **multiple comparison analysis** can be seen as an in-depth analysis of the ANOVA because it allows the factorization of the variance among the treatments. This method has some advantages:

- ⋆ it uses all the data;
- ⋆ it uses the residual variance to evaluate the error;
- ⋆ does not decrease the value α for each of the possible comparison.

A great number of multiple comparison method uses the **contrasts**.

The contrasts are linear combination of the factors' mean:

$$C = \sum_{j=1}^{p} c_j \overline{x}_j.$$

The sum of the contrasts has to be equal to zero:

$$\sum_{j=1}^{p} c_j = 0.$$

In a described system with p factors, $p-1$ independent contrasts are possible, each of them with 1 df.

The deviance of each contrast is given by:

$$SS_C = \frac{(\sum_{j=1}^{p} c_j \overline{x}_j)^2}{n \sum_{j=1}^{p} c_j^2}$$

with just one degree of freedom.

As we have just said, $df_C = 1$ then $MS_C = SS_C$.

The significance of a contrast is obtained dividing its MS_C by MS_W:

$$F = \frac{MS_C}{MS_W} \sim F_{1,n-p}.$$

The **orthogonal contrasts** are one of the most important kind of contrast. Two contrasts are defined as orthogonal when the sum of their coefficients, respectively c_j and d_j, is equal to zero:

$$\sum_{j=1}^{p} c_j d_j = 0;$$

where d_j e c_j are the orthogonal among themselves coefficients.

In other words, two contrasts are orthogonal if they are equal to zero:

- ⋆ the sum of the coefficients by row;
- ⋆ the sum of the coefficients by column.

Two orthogonal contrasts are also independent.

Example.

It is carried an experiment to test two different treatments A_1 e A_2. It is added a checking group, C to the two treatments.

The F test results to be significant. It is highlighted a difference among the three treatments. In this case the contrasts' coefficients analysis is carried in order to evaluate which are the significant differences.

The most logical factorization of the problem is the following one:

- ⋆ comparison among the checking group C and the two treatments $(A_1 + A_2)$;
- ⋆ comparison between the two treatments A_1 and A_2.

- In the first comparison, C vs $A_1 + A_2$, it wants to be checked the null hypothesis

$$H_0 : \mu_c - \frac{\mu_{A1} + \mu_{A2}}{2} = 0$$

from it, it is possible to obtain the coefficients $(+1 \ -\frac{1}{2} \ -\frac{1}{2})$ or, that is the same, the coefficients $(+2 \ -1 \ -1)$

- In the second comparison, A_1 vs A_2, the null hypothesis is checked

$$H_0 : \mu_{A1} - \mu_{A2} = 0$$

and from it it is possible to obtain the coefficients $(0 \ +1 \ -1)$

The table that follows shows the comparison coefficients in the whole form

	C	$A1$	$A2$	**Sum per line**
C against $A_1 + A_2$	$+2$	-1	-1	0
A_1 against A_2	0	$+1$	-1	0
Products per column	0	-1	$+1$	0
Total column	0	-1	$+1$	0

The comparisons, formulated in this way, respect all the orthogonality conditions.

- ⋆ The comparison of C against $A_1 + A_2$ allows us to verify if the check has different effects compared to the two treatments, linked in only one group;
- ⋆ The comparison of A_1 against A_2 allows us to highlight if the two treatments lead to different effects among them.

The analysis of variance can be also applied to an experiment with only one factor (treatment) and two levels. For this case, the Student's t method has been already presented.

In reality, t test and F test are two methods that are only apparently different to carry the same analysis: **the t test can be seen as a special case of the analysis of variance**, applied to only two groups or, to better say, **the analysis of variance is the extention of the Student's t with bilateral alternative hypothesis to more groups and with more factors**.

In the case of one factor with two groups, between the t and F distributions, it exists a precise mathematical relation:

$$F_{(1,\nu)} = t^2_{(\nu)},$$

that can be written as:

$$t_{(\nu)} = \sqrt{F_{(1,\nu)}};$$

where ν is the number of the degrees of freedom.

The F value with 1 and ν degrees of freedom is equal to the square of t with ν degrees of freedom.

The two distributions of the critical values for the same probability α are equivalent.

In the case of the experiment with two factors, the collected data can be represented in a two-dimensions table:

	Factor A			
Factor B	A_1	A_2	A_3	Means
B_1	$x_{111}, x_{211}, x_{311}, x_{411}$ $\overline{X}_{11}$	$x_{112}, x_{212}, x_{312}, x_{412}$ $\overline{X}_{12}$	$x_{113}, x_{213}, x_{313}, x_{413}$ $\overline{X}_{13}$	$\overline{\overline{X}}_{1.}$
B_2	$x_{121}, x_{221}, x_{321}, x_{421}$ $\overline{X}_{21}$	$x_{122}, x_{222}, x_{322}, x_{422}$ $\overline{X}_{22}$	$x_{123}, x_{223}, x_{323}, x_{423}$ $\overline{X}_{23}$	$\overline{\overline{X}}_{2.}$
Means	$\overline{\overline{X}}_{.1}$	$\overline{\overline{X}}_{.2}$	$\overline{\overline{X}}_{.3}$	$\overline{\overline{\overline{X}}}$

In this case there are three null hypothesis that have to be checked against as much alternative hypothesis.

The first hypothesis regards the effect of the factor (A) on the answer variable:

$$H_0 : \mu_{.1} = \mu_{.2} = \cdots = \mu_{.p}$$

$$H_A : \exists i, j \mid \mu_{.i} \neq \mu_{.j} \quad (i, j = 1, \ldots, p)$$

where p is the number of levels of the factor (A).

Or rather, the factor (A), indipendentently of the factor (B), has no effect on the answer variable.

The second hypothesis regards the effect of the factor (B) on the answer variable:

$$H_0 : \mu_{1.} = \mu_{2.} = \cdots = \mu_{q.}$$

$$H_A : \exists i, j \mid \mu_{i.} \neq \mu_{j.} \quad (i, j = 1, \ldots, q)$$

where q is the number of levels of the factor (B).

Or rather, the factor (B), indipendently of the factor (A), has no effect on the answer variable.

The first two hypothesis can be considered as two independent analysis of variance with a single factor.

The third hypothesis regards the combined effect of the treatments A and B on the answer variable. The combined effect is defined **interaction**.

Example. In a catalytic reaction there are:

- ⋆ a positive effect (or rather an **increase** of the percentage of catalytised material) going from a temperature of $140°C$ to a temperature of $180°C$ using the catalyst $\boldsymbol{C_1}$;

- ⋆ a negative effect (or rather a **decrease** of the percentage of catalysed material) going from a temperature of $140°C$ to a temperature of $180°C$ using the catalyst $\boldsymbol{C_2}$;

As the algebra that regards the interactions is relatively more complex than the algebra of principal effects, we will avoid to delve into this topic.

It is enough to remember that, in a two or more factors analysis of variance, it exists also a component that regards the interaction.

The **total deviance** or $\boldsymbol{SS_{Total}}$ (Total Sum of Squares) defines the residual sum of squares of every value of the overall mean.

$$SS_{Total} = \sum_{k=1}^{q} \sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ijk} - \overline{\overline{\overline{X}}})^2$$

where:

- X_{ijk} is the ith observation that regards the jth level of the first factor and the kth level of the second factor;
- n is the size (considered constant) of the corresponding group of observations at the jth level of the first factor and at the kth level of the second factor;
- p is the number of levels for the first factor;
- q is the number of levels for the second factor.

The $\boldsymbol{SS_{Total}}$ defines the total variability of the sample.

The **deviance between treatments** or $SS_{Between}$ (Between Sum of Squares) defines the residuals sum of the squares of every group mean from the overall mean, multiplied the number of data of the corresponding group. It exists three components "between": one for the factor A, one for the factor B and one for the interaction.
The calculation of the component "between" is really similar to the calculation of the one-factor analysis of variance.

$$SS_A = qn \sum_{j=1}^{p} (\overline{\overline{X}}_{j.} - \overline{\overline{\overline{X}}})^2 \qquad \text{For the factor (A).}$$

$$SS_B = pn \sum_{k=1}^{q} (\overline{\overline{X}}_{.k} - \overline{\overline{\overline{X}}})^2 \qquad \text{For the factor (B).}$$

$$SS_{AB} = n \sum_{j=1}^{p} \sum_{k=1}^{q} (\overline{X}_{jk} - \overline{\overline{X}}_{j.} - \overline{\overline{X}}_{.k} + \overline{\overline{\overline{X}}})^2 \qquad \text{For the interaction (AB).}$$

The $SS_{Between}$ defines the variability among groups independently of the dispersion of the single observations within each group.

The **deviance within treatments** or SS_{Within} (Within Sum of Squares), also called **error**, defines the residuals sum of squares of every value of the mean of its group.

$$SS_{Within} = \sum_{k=1}^{q} \sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ijk} - \overline{X}_{jk})^2$$

The SS_{Within} defines the variability within single groups independently of the position of the group compared to/around the central mean of the sample or compared to/around the mean calculated for a single factor.

As the remaining part of the variability quota does not represent only the sum of the squares within the groups but also the errors sum of squars, in the future it will be preferable to use $\boldsymbol{SS_E}$ instead of $\boldsymbol{SS_{Within}}$ or $\boldsymbol{SS_W}$.

Similarly, the corresponding degrees of freedom will be indicated with df_E and the ratio between the sum of squares and the degrees of freedom with MS_E.

It is possible to demonstrate, also in this case, that

$$SS_{Total} = SS_{Between} + SS_E$$

where

$$SS_{Between} = SS_A + SS_B + SS_{AB}$$

with

- SS_A: sum of squares attributable to the factor (A);
- SS_B: sum of squares attributable to the factor (B);
- SS_{AB}: sum of squares attributable to the interaction between (A) and (B).

Starting from the sum of squares **SS**, it is possible to calculate the means of squares **MS** dividing by the appropriate degrees of freedom.

The calculation of the degrees of freedom still results to be quite easy.

- ⋆ The degrees of freedom of each factor are equal to the number of levels of that factor minus one.
- ⋆ The degrees of freedom of the interaction between the two factors are given by the product of the degrees of freedom of each factor.
- ⋆ The degrees of freedom of the residuals are equal to the product of the number of groups and the number of observations within each group minus one. In other words, they are equal to the number of observations minus one and minus the sum of the degrees of freedom of each factor and the interaction.

As in the univariate case, it is valid the following relation:

$$df_{Total} = df_A + df_B + df_{AB} + df_E.$$

Example. It is supposed to carry out an experiment with two factors, (A) and (B), where:

- ⋆ the factor (A) has two levels;
- ⋆ the factor (B) has three levels;
- ⋆ each combinations of factors' levels has four replications.

It follows that:

- ⋆ the total number of observations N is equal to $2 \cdot 3 \cdot 4 = 24$ from which $\boldsymbol{df_T = 23}$;
- ⋆ the factor (A) has two levels, from which $\boldsymbol{df_A = 1}$;
- ⋆ the factor (B) has three levels, from which $\boldsymbol{df_B = 2}$;
- ⋆ the interaction (AB) has degrees of freedom $\boldsymbol{df_{AB} = df_A \cdot df_B = 2}$;
- ⋆ the residual (error) has degrees of freedom equal to 6 (groups) multiplied by 3 (number of observations within each group minus one), that is $\boldsymbol{df_E = 6 \cdot 3 = 18}$.

Starting from the **SS** it is possible to calculate the means of squares MS:

$$MS_A = \frac{SS_A}{df_A}$$

$$MS_B = \frac{SS_B}{df_B}$$

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$$

$$MS_E = \frac{SS_E}{df_E}$$

The $\boldsymbol{MS_A}$, $\boldsymbol{MS_B}$ and $\boldsymbol{MS_{AB}}$, if the null hypothesis are true, or rather if the factors (A) and (B) do not have any effects on the answer variable, they are all independent estimations of the variance of σ_ε^2 from which the sample has been extracted.

$\boldsymbol{MS_E}$ is an estimation of σ_ε^2.

As consequence, if all the hypothesis H_0 are valid:

$$\frac{MS_A}{\sigma_A^2} = \frac{MS_A}{\sigma_\varepsilon^2} \sim \chi^2_{df_A} \qquad \frac{MS_B}{\sigma_B^2} = \frac{MS_B}{\sigma_\varepsilon^2} \sim \chi^2_{df_B}$$

$$\frac{MS_{AB}}{\sigma_{AB}^2} = \frac{MS_{AB}}{\sigma_\varepsilon^2} \sim \chi^2_{df_{AB}} \qquad \boxed{\frac{MS_E}{\sigma_E^2} = \frac{MS_E}{\sigma_\varepsilon^2} \sim \chi^2_{df_E}}$$

and, as consequence, the ratios

$$F_A = \frac{MS_A}{MS_E} \sim F_{df_A, df_E} \qquad F_B = \frac{MS_B}{MS_E} \sim F_{df_B, df_E}$$

$$F_{AB} = \frac{MS_{AB}}{MS_E} \sim F_{df_{AB}, df_E}$$

tend to be distributed as a **F of Fisher-Snedecor** with appropriate degrees of freedom.

It is then possible to build a table of the analysis of variance like follows:

	SS	df	MS	F	P(F)
A	SS_A	df_A	$MS_A = \frac{SS_A}{df_A}$	$\frac{MS_A}{MS_E}$	$P(F_{df_A,df_E} > F)$
B	SS_B	df_B	$MS_B = \frac{SS_B}{df_B}$	$\frac{MS_B}{MS_E}$	$P(F_{df_B,df_E} > F)$
AB	SS_{AB}	df_{AB}	$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MS_E}$	$P(F_{df_{AB},df_E} > F)$
Error	SS_E	df_E	$MS_E = \frac{SS_E}{df_E}$		
Total	SS_T	df_T			

The rules to read this table are identical to those of one-factor case.

The different concept is that, in a multivariate case, there are a lot of null hypothesis that have to be checked.

Example.
An experiment with three factors (A, B, C) requires the calculation of eight MS, or rather: A, B, C, AB, AC, BC, ABC, E.

Assuming, for example, that the interaction ABC is not significant, it implies that this interaction can be removed from the analysis.

But, the equation

$$SS_T = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} + SS_E$$

must be always true.

If we leave ABC, the equation will not be true.

In practice, the equation is defined like follows:

$$SS_T = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{E1},$$

where:

$$SS_{E1} = SS_E + SS_{ABC}.$$

The degrees of freedom have to be modified too, so that:

$$df_{E1} = df_E + df_{ABC}.$$

If the **SS** of the error and the degrees of freedom change, it will then change also the **MS** of the error:

$$MS_{E1} = \frac{SS_{E1}}{df_{E1}}.$$

It follows that all the ratios F and the relative probabilities are altered.

In conclusion, in a multivariate ANOVA analysis, it is necessary to carry on with the estimation of the more **parsimonious** model for interactions that follows, removing one terms at a time from the analysis.

The choice of the term that has to be removed first is done according to two choice criterions:

- ⋆ the degree of the term: the terms with an higher degree are removed first;
- ⋆ with the same degree, the term with smaller F value or $P(F)$ higher.

This has to stay within hierarchical models.

The mathematical model formulation of the ANOVA with two or more factors result to be an easy enlargement of the univariate model, every single observation $\boldsymbol{X_{ijk}}$ can be written as

$$X_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

as it is determined:

- ⋆ by the **overall mean** $\boldsymbol{\mu}$ that defines the dimension of the experiment;
- ⋆ by the **parameter** $\boldsymbol{\alpha_j}$ that regards the jth **level** of the factor (A);
- ⋆ by the **parameter** $\boldsymbol{\beta_k}$ that regards the kth **level** of the factor (B);
- ⋆ by the **parameter** $\boldsymbol{\gamma_{jk}}$ that regards **the interactions on the levels** jth and kth of the factors (A) and (B);
- ⋆ by a **random term** $\boldsymbol{\varepsilon_{ijk}}$ called **residual** or **experimental error**.

In this case too:

$$\varepsilon_{ijk} \in \varepsilon \sim N(0, \sigma_\varepsilon) \qquad \forall \qquad (i, j, k)$$

In other words, the error term of the model is distributed with null mean and constant variance.

The normality condition of the error term of the model is one of the validity conditions of the model itself, together with the independence of the residuals values.

In conclusion, **the residuals analysis, in the multivariate case, is identical to the univariate case.**

Basically, there are 4 types of experimental plans:

1. **balanced** where n_{ij} is equal to each combination ij, or rather, the number of data in every cell is equal;

2. **proportional** to the number of replication of the marginals of the size table. In the two-factors case:

$$n_{ij} = \frac{n_i n_j}{N}$$

where:

- n_{ij} is the number of data at the crossing between the line i and the column j;
- n_i is the total number of the observations within the line i;
- n_j is the total number of observations within the line j;
- N is the total number of the observations.

Example. An experimental plan is composed by 4 level of the factor A and 3 levels of the factor B, as shown in the table:

	a_1	a_2	a_3	a_4	**Total**
b_1	3	6	9	6	24
b_2	4	8	12	8	32
b_3	2	4	6	4	16
Total	9	18	27	18	72

3. **not balanced** where the number of data in each cell is different and not proportional;

4. **with empty cells**.

In an analysis of variance a **fix effect** is a factor whose levels cover all the possible modalities of the phenomenon that the factor expresses.

Example. In a company there are two identical measuring devices with the same duty. An experiment has to be carried out in order to understand if there is a significant difference in the way the two devices measure.

As only two devices exist, a factor with two level expresses the wholeness, or rather the whole population, of the devices.

In this case the factor (effect) is defined as **fix** and it is applied to a standard ANOVA (shown as **ANOVA I**).

Example.

An oven, used for firing a certain material, has a temperature range included between $120°C$ and $180°C$.

Two temperature are set up $130°C$ and $170°C$.

The two temperatures represent two level **fix**: low and high temperature.

In this case they are fixed effects too.

Example.

A certain phase of a manufactoring process requires a measurement by an operator.

The operators can be several and it is not possible to check whose operator will do the measurement.

It is then carried out an experiment, using two operators, in order to test if an operator effect exists in the manufactoring process.

The two operators are not the only two operators but a random sample of all the possible operators.

In this case it is ANOVA with **random effects**.

A random effects ANOVA (usually shown as **ANOVA II**) follows the same methods of a fixed effects ANOVA analysis.

Anyway, there are some difference as regards the formulation of the hypothesis:

$$H_0 : \ \sigma_B^2 = 0$$

$$H_A : \ \sigma_B^2 > 0$$

where σ_B^2 is the answer variance due to the difference among treatments.

This happens because it has no sense to test the single effect hypothesis as the chosen levels are **samples randomly extracted from the population**.

In the ANOVA II, if H_0 is refused, the multiple comparisons that have no meaning, are not done.
The difference between two specific levels is not of interest because these are representative of a wider population.

It follows that it is useful to proceed with a quantitative valuation of the variance by the calculation of the **components of the variance**.

- MS_B is a correct estimator of $\sigma_\varepsilon^2 + n\sigma_B^2$

  or rather $E(MS_B) = \sigma_\varepsilon^2 + n\sigma_B^2$

- MS_E is a correct estimator of the variance "within" σ_ε^2

  or rather $E(MS_E) = \sigma_\varepsilon^2$

In order to estimate the element of variance "between", it is used the following formula

$$\hat{\sigma}_B^2 = \frac{MS_B - MS_E}{n}.$$

If the null hypothesis is true, or rather that it does not exist variability due to different treatments that have been randomly chosen, then:

$$\sigma_B^2 = 0$$

$$E(MS_B) = \sigma_\varepsilon^2$$

And then, similarly to the fixed effects case:

$$F = \frac{\frac{MS_B}{\sigma_\varepsilon^2}}{\frac{MS_E}{\sigma_\varepsilon^2}} = \frac{MS_B}{MS_E} \sim F_{p-1,n-p}$$

If the model is not balanced, n has to be corrected:

$$\hat{n} = \overline{n} - \frac{\sum (n_i - \overline{n})^2}{(k-1)N},$$

where

- $\overline{n}$ is the mean number of data in the levels;
- n_i is the number of data within the level i;
- k is the number of levels;
- N is the total number of data.

In the case of a multifactorial ANOVA, the calculation of F ratio is modified, but the basic logic does not change.

In conclusion, a third type of analysis exists (it will be shown as **ANOVA III**), given be the **mix of fixed and random factors**.

The hypothesis for the calculation of the ratios for the F test are equal to that we have already seen for ANOVA I and ANOVA II, even though the ratios themselves change compared to these two.

The analysis of variance, as it has been presented till now, assumes that, during the experiment, every level of each factor cuts across with all the levels of the other factors.

If the number of times that a level of a factor meets all the levels of the others factors is constant for all the levels combinations, then the experiment is called **crossed balanced**.

If, instead, the same number is not equal for all the combinations, then the experiment is called **crossed not balanced**.

Another type of experiment is defined as **nested experiment**.

Example.

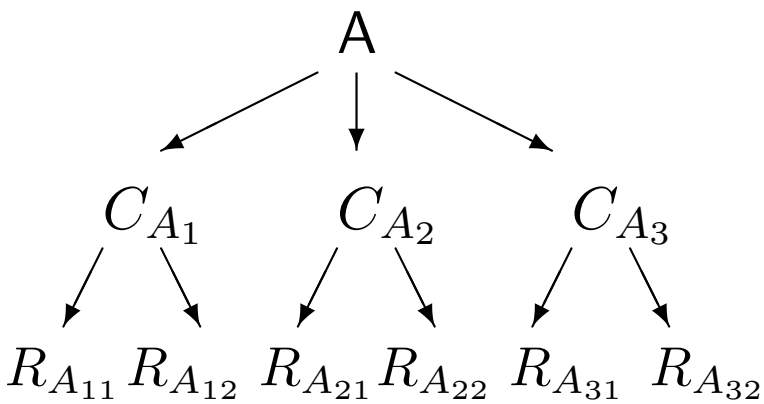
The same pharmaceutical product is produced by two companies: A e B.

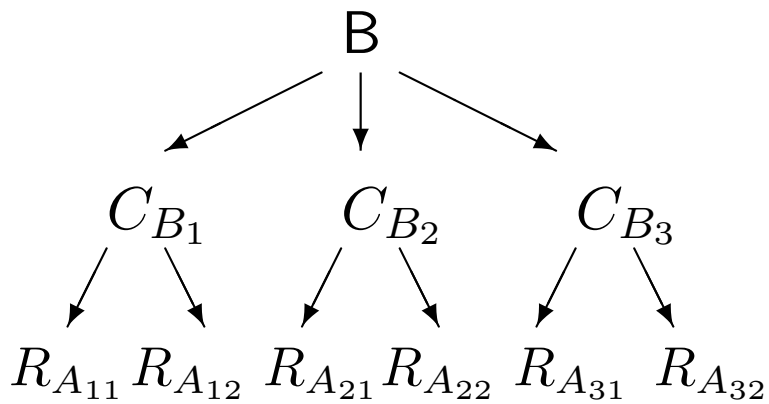
The company A produces three lots of product with different concentrations C_{A1}, C_{A2} e C_{A3}.

The company B produces three lots of product with different concentrations C_{B1}, C_{B2} e C_{B3}.

For each company, two sample revelations for each concentration have to be taken into consideration.

The situation can be summed up in the following diagram:





For the analysis of this type of experiment it is necessary to use a **hierarchical or nested classification** for factor, commonly known as **nested anova**.

The attention goes on the difference between the product lots within (nested) the company.

In other words, it wants to be checked if there is a statistically significant difference between two (or more) lots, once that the company effect has been removed.

A similar case corresponds to a statistical model like:

$$X_{ijk} = \mu + \alpha_k + \beta_{j(k)} + \varepsilon_{i(jk)}$$

where:

- μ is the overall mean;
- α_k is the kth effect of the company factor (α);
- $\beta_{j(k)}$ is the jth effect of the lot effect (β) nested in the kth level of the company factor (α);
- $\varepsilon_{i(jk)}$ is the ith error nested in the jth level of the lot factor (β) which is nested in the kth level of the company factor (α).

Two important notions are implicit:

- **the higher level affects the lower level**;
- **the lower level has importance if it is analysed within the higher**.

The calculation of the **SS** and the **MS** is made similarly to the formulations used in the case of crossed experiments.

In a similar case, with fixed effects, it is possible to build a table of the analysis of variance like:

	SS	df	MS	F	P(F)
A	SS_A	df_A	$MS_A = \frac{SS_A}{df_A}$	$\frac{MS_A}{MS_E}$	$P(F_{df_A,df_E} > F)$
B	SS_B	df_B	$MS_B = \frac{SS_B}{df_B}$	$\frac{MS_B}{MS_E}$	$P(F_{df_B,df_E} > F)$
AB	SS_{AB}	df_{AB}	$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MS_E}$	$P(F_{df_{AB},df_E} > F)$
Error	SS_E	df_E	$MS_E = \frac{SS_E}{df_E}$		
Total	SS_T	df_T			

In the nested experiments, the effects can be fixed or radom, which want different hypothesis formulations.

As consequence, also the analysis of variance of nested experiments can be divided in:

- ⋆ NESTED ANOVA I, where all the factors are fixed;
- ⋆ NESTED ANOVA II, where all the factors are random;
- ⋆ NESTED ANOVA III, where some factors are fixed and other random, usually that of lower levels.

The choice of the type of ANOVA and of the type of hypothesis to test depends on the formulation of the problem.

The calculation of the F ratios vary according to the type of NESTED ANOVA that has been considered.

The factorization of the variability of a system using an ANOVA is an algebric relation and, as consequence, it requires that some assumptions are satisfied.
The fundamental assumption requires that the observation are adequately represented by a linear model like

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

and that the errors (ε_{ij}) are independent and normally distributed with $E[\varepsilon] = 0$ and variance (σ^2) usually not known.

If these assumptions are satisfied, then the analysis of variance to test H_0 is realiable.

In order to investigate on the possible violations and to validate the model, it is useful to do an **analysis of residuals**.

It is defined as **residual of the observation i of the factor j**:

$$e_{ij} = x_{ij} - \hat{x}_{ij},$$

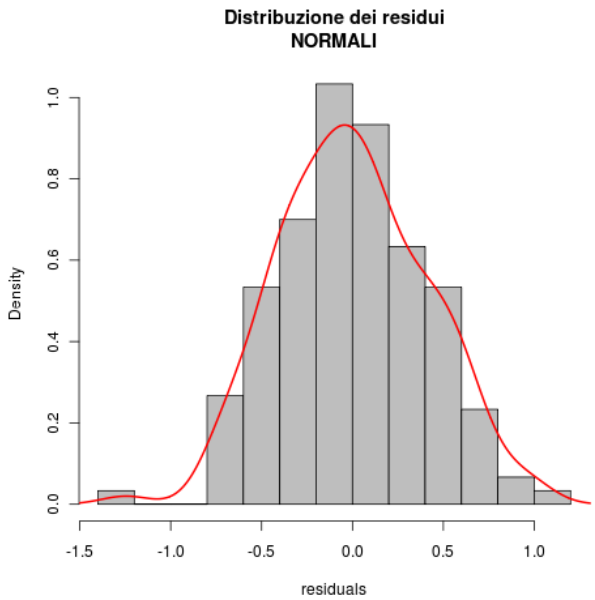
where:

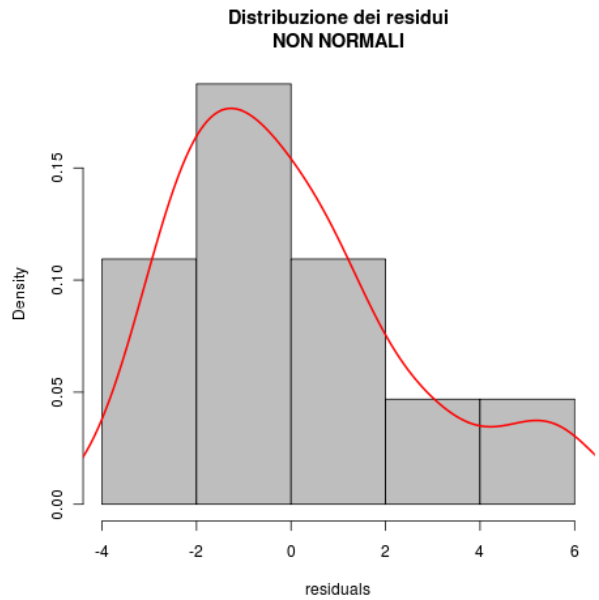
- x_{ij} is the observation i for the level j;
- $\hat{x}_{ij}$ is the estimation of the observation x_{ij} calculated from the model (in the easiest case, the mean $\overline{x}_j$).

If the model adequately describes the data, the residuals:

- do not have to be an evident pattern, or rather they have to be randomly distributed;
- do not have to be linked to any variable;
- must have a constant variability (homoscedasticity).

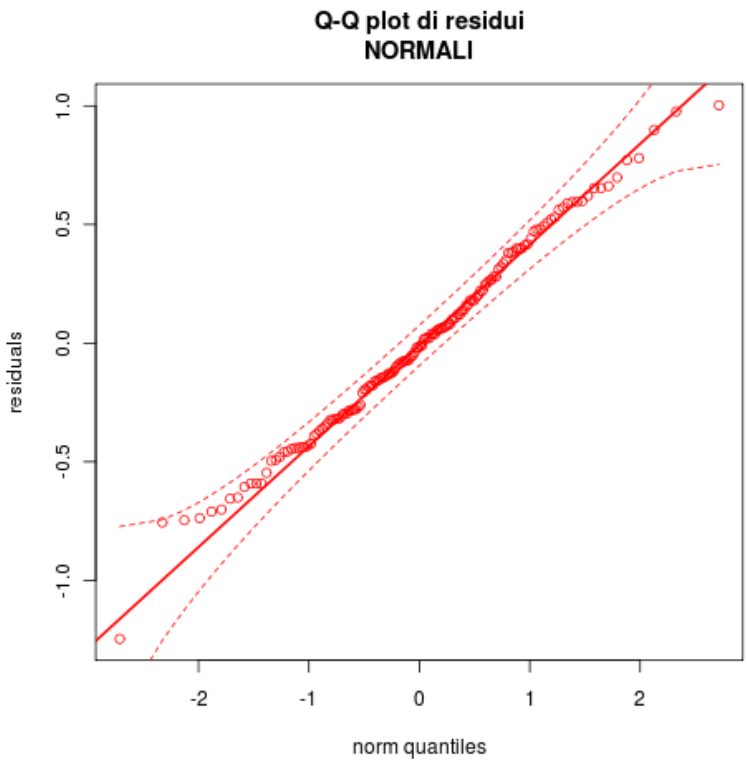
The easiest and most evident way to check the normality of the residuals, is to draw an histogram of the residuals and to check if it has a bell-shaped distribution around zero.

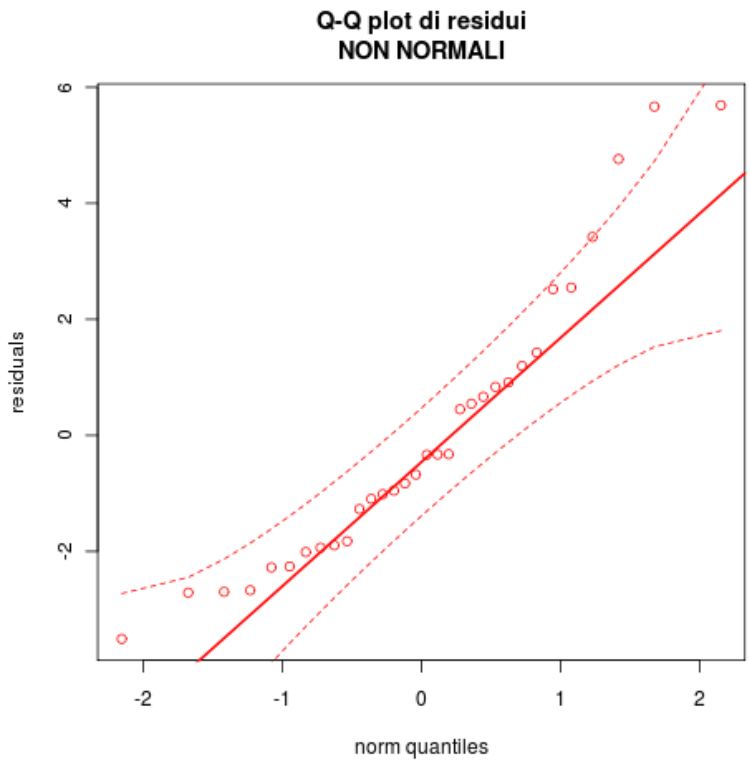




This kind of graphical approach is immediate, but it has to be confirmed by other tests because it is not very precise.

It can be useful to build a *normal probability plot* of the residuals

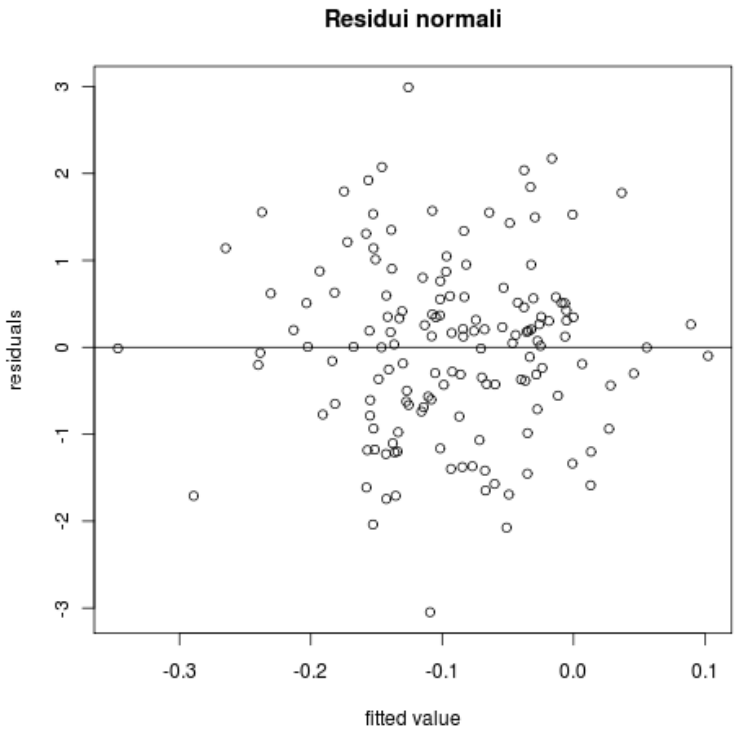


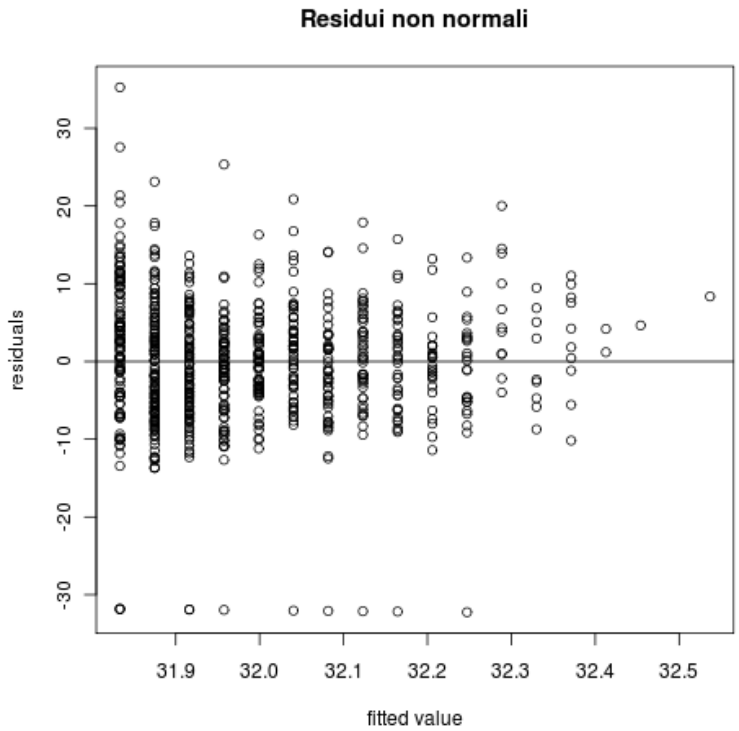


Anderson-Darling test can be used to check the normality hypothesis of the residuals.

A way to check if the formulation of the model is adequate, is that of drawing a graphic of the residuals against the estimated value of the model ($\hat{y}_{ji}$ vs e_{ji}).

This plot does not have to show any identifiable pattern.





Another graphic tool is the plot of the residuals against the explicative variables. The graphic does not have to show any identifiable functional trends.

If the residuals show a functional relation, it means that the linear model is not able to thoroughly explain the data.

It is useful to check that the residuals are **independent** among them. If there is autocorrelation in the residuals it means that there is dependence. The presence of **autocorrelation** can be caused by:

- ⋆ a wrong specification of the functional link of the model;
- ⋆ the fact that one or more explicative variables linked to the time, have not been included in the model.