ONE FACTOR ANOVA

By Alessandro Maurizio Idrovo Gavilanes

Introduction

One-factor ANOVA is a statistical method for testing the null hypothesis that three or more population means are equal, against the alternative hypothesis that at least one of the populations differs from the others in terms of its expected value. This contrast is fundamental in the analysis of experimental results, in which it is of interest to compare the results of K 'treatments' or 'factors' with respect to the dependent variable or variable of interest.

Analysis of variance assumptions

ANOVA requires the following assumptions to be met:

- The populations follow a normal distribution.
- The K samples on which the treatments are applied are independent.
- The populations all have equal variance (homoscedasticity).

One-factor ANOVA

When you have *k* different levels corresponding to the *k* population means, which are also the treatments, an ANOVA attempts to answer the question: Are the k population means all the same, or is at least one mean different from the others?

Then if there are k independent random samples of size n_i (i=1, 2, ……, *k*), coming from k different populations (concerning *k* treatments, *k* groups, *k* production methods, etc.) they can be ordered as follows:

$$\begin{matrix} T_1 & T_2 & \cdots & T_k \\ y_{11} & y_{21} & \cdots & y_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1n_1} & y_{2n_2} & \cdots & y_{kn_k} \end{matrix}$$

The sample means of the groups are denoted as $\bar{y}1$, $\bar{y}2$, ………, $\bar{y}K$ and $\bar{y}$ is the mean of all $N = \sum_{i=1}^{k} = n_i$ observations.

- **Variability**

The responses generated in an experimental situation always show some variability. In an analysis of variance, the total variability of the responses is divided into two parts: one due to **between-treatment variability** (measured by the sum of the squares of the treatments, SC_{trat}) and another part due to **within-treatment variability** or error (measured by the sum of the squares of the error, SC_{error}):

Total variability= between-treatment variability + within-treatment variability

$$SC_{total} = SC_{trat} + SC_{error}$$

- **Between-treatment variability**

$$SC_{trat} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

It measures the discrepancy between the groups and the global mean, so that if there are no differences between them (the null hypothesis is not rejected) we will obtain small variabilities. If, on the other hand, the null hypothesis is rejected, the variability between groups can be expected to be large.

- **Within-treatment variability or error variability**

$$SC_{error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

It measures the intrinsic variability of the observations, i.e., if the experiment is well designed and no variation factors other than the one studied are included, it should be purely random error produced as a result of the biological variability of the experimental material.

ANOVA table

The analysis of variance contrast is based on the comparison of the variability between and the variability within. Thus, the null hypothesis is rejected whenever the "**between**" variability is large with respect to the "**within**" variability.

The complete information is summarized in the following table, known as the ANOVA table, which compiles all the information necessary to perform the corresponding test.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

Example

A study is done on the effectiveness of three brands of spray to kill flies. To do this, each product is applied to a group of 100 flies, and the number of dead flies is counted, expressed in percentages. Six replicates are made and the results obtained are shown below (Gutiérrez Pulido & De la Vara Salazar, 2008).

Atomizer brand	Number of replicas					
	1	2	3	4	5	6
1	72	65	67	75	62	73
2	55	59	68	70	53	50
3	64	74	61	58	51	69

Assumptions verification (Minitab software)

Verification of the normality assumption

1. **Set the hypotheses.**

Ho= The residues follow a normal distribution

H1= The residues do not follow a normal distribution

2. **Set the significance level.**

α= 0.05

3. **Set the test statistic.**

Anderson- Darling.

4. **Set the decision.**

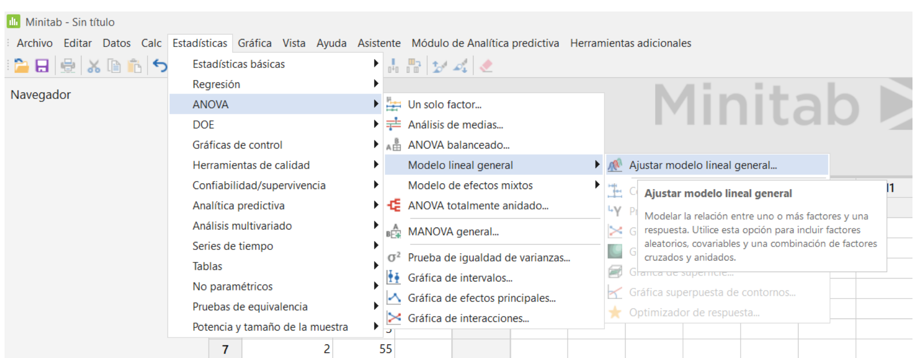
The null hypothesis is rejected if p-value < α.

5. **Resolution**

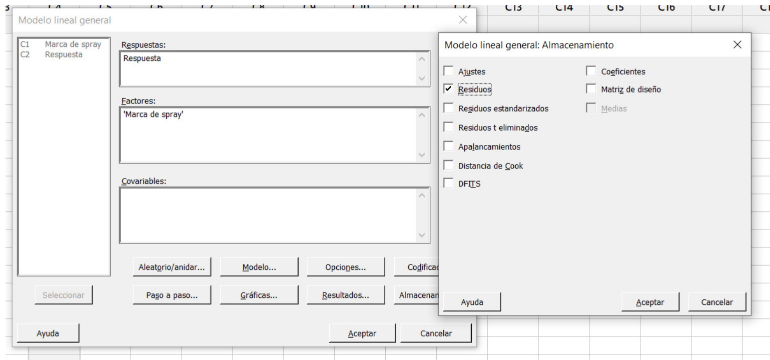
1) The data is placed in Minitab, entering them in two columns, one corresponding to the treatments and the other corresponding to the responses.

	C1 Atomizer Brand	C2 Response
1	1	72
2	1	65
3	1	67
4	1	75
5	1	62
6	1	73
7	2	55
8	2	59
9	2	68
10	2	70
11	2	53
12	2	50
13	3	64
14	3	74
15	3	61
16	3	58
17	3	51
18	3	69

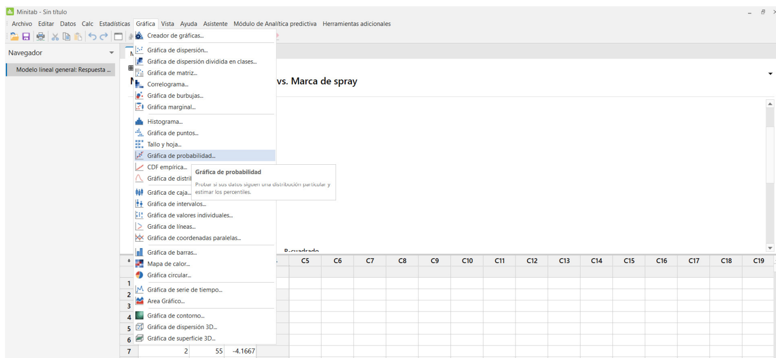
2) Select the ANOVA option to fit general linear model.



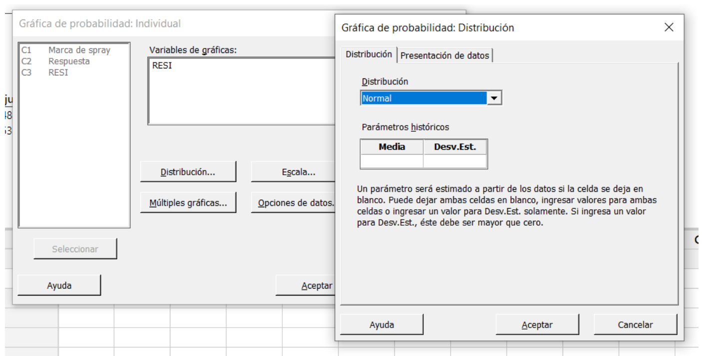
3) Select the responses and treatments in the open window in addition to selecting the waste in storage option.



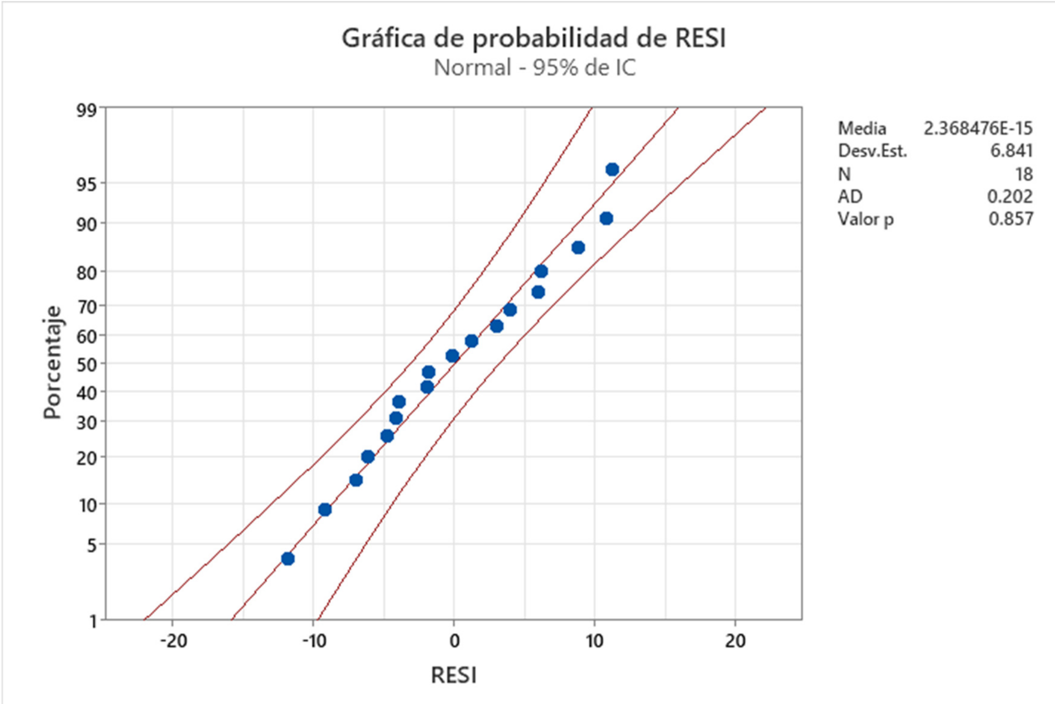
4) Once the residuals have been calculated, we go to the graph section and select probability graph.



5) Select residuals as the response variable and a normal distribution.



6) Results according to Minitab



Since the p-value is greater than the significance level, the null hypothesis is not rejected, therefore, we cannot reject with a 95% confidence level that the residuals follow a normal distribution **(the assumption of normality is met)**.

Verification of the homoscedasticity assumption (Minitab software)

1. **Set the hypotheses.**

Ho= The variances are equal.

H1= At least one of the variances is different.

2. **Set the significance level.**

α= 0.05

3. **Set the test statistic.**

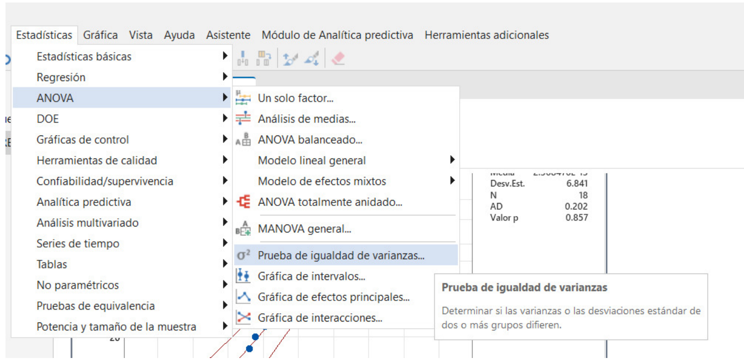
Bartlett.

4. **Set the decision.**

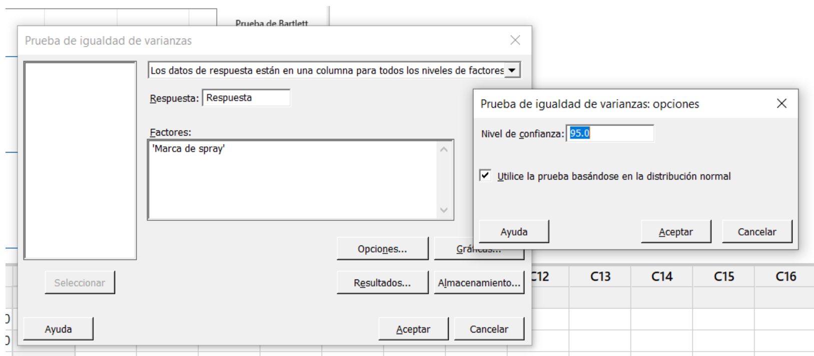
The null hypothesis is rejected if p-value < α.

5. **Resolution**

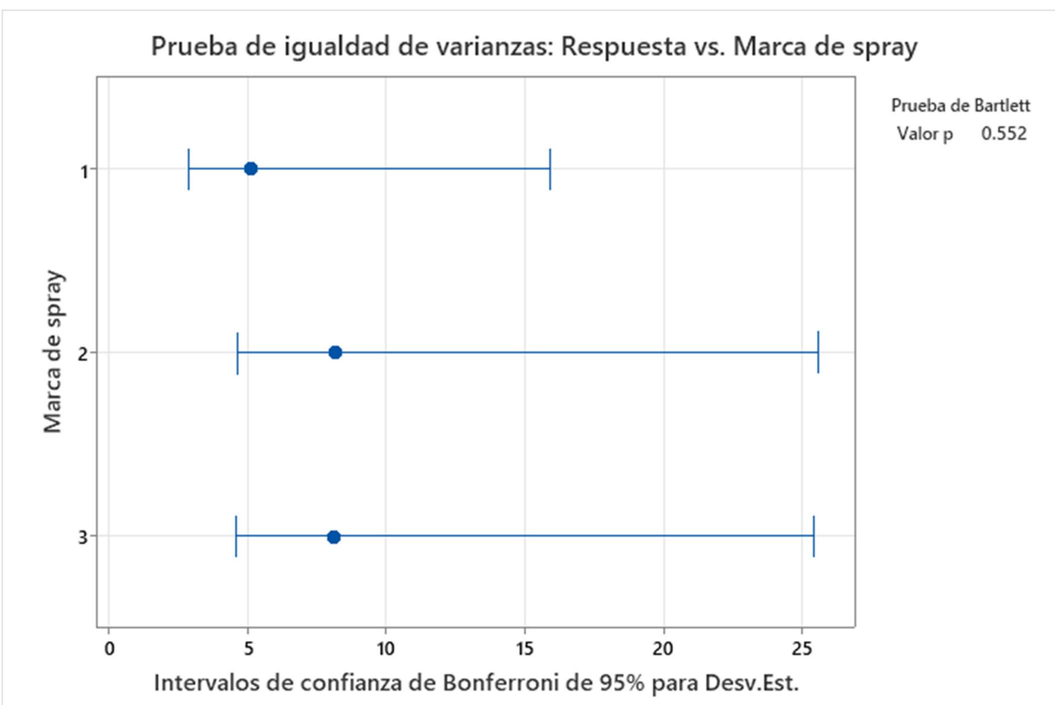
1) Select ANOVA test of equality of variances.



2) In the open window we select the response and corresponding treatments, in addition to indicating that the test is based on a normal distribution.



3) Results according to Minitab.



Since the p-value is greater than the significance level, the null hypothesis is not rejected, therefore, we cannot reject with a 95% confidence level that the variances are equal **(the assumption of homoscedasticity is met)**.

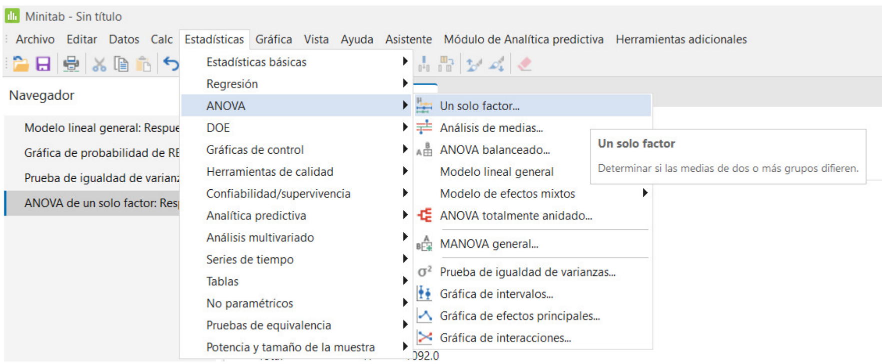
Verification of the independence assumption

Since the populations come from different treatments, we assume that the populations are independent **(the assumption of independence is met)**.

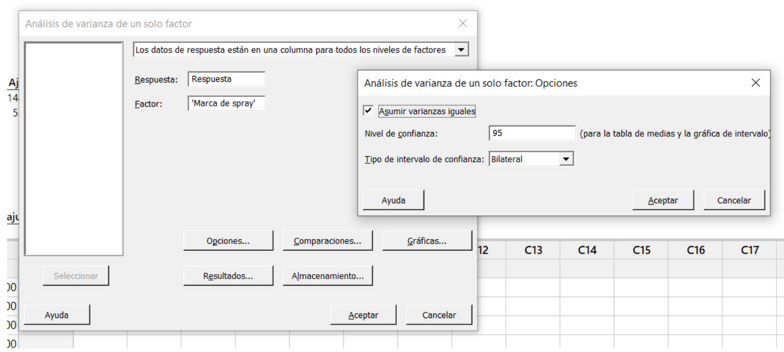
ANOVA resolution (Minitab)

Once the assumptions have been verified, the ANOVA table is created.

1) Select one-factor ANOVA.



2) Select the treatments and response in the corresponding boxes, in addition to checking the box to assume equal variances.



3) Results according to Minitab.

Método

Hipótesis nula	Todas las medias son iguales
Hipótesis alterna	No todas las medias son iguales
Nivel de significancia	α = 0.05

Se presupuso igualdad de varianzas para el análisis.

Información del factor

Factor	Niveles	Valores
Marca de spray	3	1; 2; 3

Análisis de Varianza

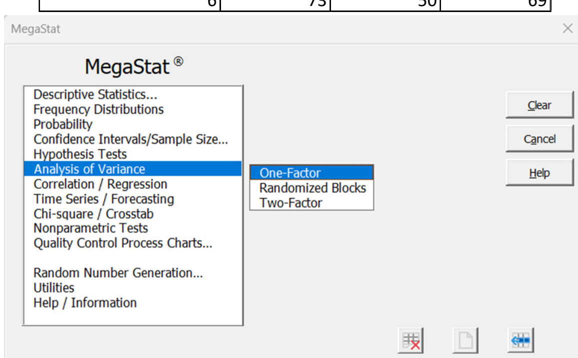
Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Marca de spray	2	296.3	148.17	2.79	0.093
Error	15	795.7	53.04		
Total	17	1092.0			

Since the p-value is greater than the significance level, the null hypothesis is not rejected, therefore, **we cannot reject with a 95% confidence level that the population means are equal.**

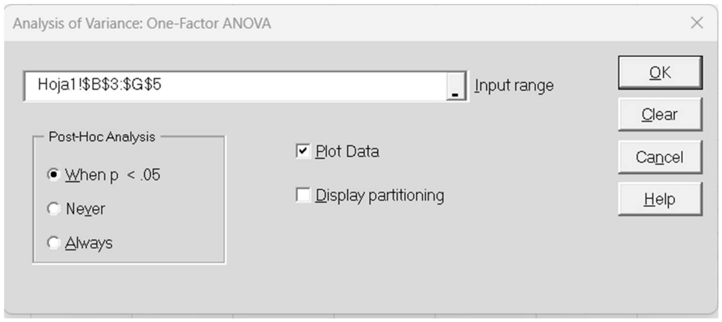
ANOVA resolution (MegaStat)

1) Open MegaStat in Excel and select analysis of variance one-factor.

Number of replicas	Atomizer brand		
	1	2	3
1	72	55	64
2	65	59	74
3	67	68	61
4	75	70	58
5	62	53	51
6	73	50	69



2) Select the data of treatments and number of replicas in the input range, in addition to checking the box of plot data.



3) Results according to MegaStat.

One factor ANOVA

Mean	n	Std. Dev	
69,0	6	5,10	A
59,2	6	8,18	B
62,8	6	8,13	C
63,7	18	8,01	Total

ANOVA table

Source	SS	df	MS	F	p-value
Treatment	296,33	2	148,167	2,79	,0931
Error	795,67	15	53,044		
Total	1.092,00	17			

Since the p-value is greater than the significance level, the null hypothesis is not rejected, therefore, **we cannot reject with a 95% confidence level that the population means are equal.**

Conclusion

One-factor ANOVA is a very useful statistical method for testing the null hypothesis that three or more population means are equal, against the alternative hypothesis that at least one of the populations differs from the others in terms of its expected value.

Bibliography and recommended links

Gutiérrez Pulido, H., & De la Vara Salazar, R. (2008). Análisis y Diseño de experimentos. Ciudad de México: McGraw-Hill Interamericana.

Milton, J. (2014). Estadística para Biología y Ciencias de la Salud. Ciudad de México: McGraw-Hill Interamericana.

http://www.ub.edu/aplica_infor/spss/cap4-7.htm

https://www.youtube.com/watch?v=efSGqgvZCtw&ab_channel=LEARN%26APPLY%3ALeanandSixSigma