DataScience
in Human & Social Science for Women Empowerment

https://datascience-project.eu/

Training Fiche Template

Title	Generalized linear models: ANOVA
Keywords (meta tags)	Multivariate analysis, between and within variability, hypothesis testing, linear models
Language	English
Objectives / Goals / Learning outcomes	**The aim of this module is to present the basic concepts of the one-factor and two-factors Analysis of Variance (ANOVA), which can be understood as a basic linear model** **At the end of this module you will be able to:** - **How ANOVA can be useful to test if there are differences between the mean value of a continuous variable across different levels of one or several categorical variables.** - **Understand and identify the conditions required to apply these techniques.** - **Conduct one-way and multiple Analysis of Variance and interpret the results obtained.**
Training course:	
Data Science Literacy	
Data Visualisation and Visual Analytics Module	X
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	
Description	In this training module you will be introduced to the use of basic linear modeling to understand how mean differences can be attributed or not to the effect of categorical variables. The analysis presented here is the basis of linear regression, which also considers the effect of continuous variables. The techniques described in this training module limit themselves to the case of categorical (qualitative) variables. On this regard, you can approach the contents of this module as an introduction to General Linear Modeling (GLM) that

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	uses only categorical factors to explain variability in a (continuous) variable of interest. The procedure presented here bases on decomposing the total variability measured in the sample into different sources: some are residual (or unexplained by the factors considered) while some are coming from a systematic part that can be attributed to the different categories of the categorical factors.
Contents arranged in 3 levels	**1. INTRODUCTION** The GLM techniques presented here in the form of Analysis Of Variance (ANOVA) allow for responding to potentially interesting questions. Some examples: a. Are male and female workers in a region making the same mean annual wage? b. Do the students of a course following different teaching methods getting the same mean grade? c. Is the mean weekly consumption of certain medicine different across age groups and/or gender? One-factor ANOVA is fine for questions 1 and 2, while question 3 requires of two-factor ANOVA. Our goal is to test for the effect of an independent variable (*factor*) classified into k several categories (*levels*) on a numerical dependent variable (*response variable*), and it bases on decomposing total sample variability. We can approach this problem as a statistical hypothesis test of a null hypothesis (H0; our default) versus and alternative (H1; an alternative worldview). The test is formulated in terms of the population means of the response variable across the levels of our factor(s). $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ H_1: At least two different μ_i The assumptions required to conduct the ANOVA test are: - Normal populations: the distribution of the response variable on each and every level should be normal - Equality of variances: the variances of the response variable across levels must be the same

- Independent simples: the sample data on each level of the factor is nor correlated with the other sample data (collected from the other levels)

2. ONE FACTOR ANOVA

2.1. The Procedure

The ANOVA procedure with one factor bases on the following equation:

$$X_{\mathrm{ir}} = \mu + \alpha_i + \mathrm{u}_{\mathrm{ir}}$$

where X_{ir} is the value of our response variable for individual r at category (level) i. We assume that this value is the sum of three effects:

- A grand mean value (μ), common to all the individuals and levels
- A shift (α_i) that captures the mean influence of belonging to level i
- A residual (u_{ir}) , which accounts for random, uncontrolled variations. This residual is assumed to distribute normally with zero mean

The ANOVA test is equivalent to test if the α_i terms are identical across the k levels. If not, there will be significant differences in the means.

We take sample data on X and decompose its variability (dispersion around the sample means) into two parts:

a. The within group (SSW) accounts for the internal variability.
b. The between variability (SSB) accounts for the differences between each group sample mean and the grand mean.

The total variability (SST) is just the sum of SSW+SSB. If SSB is much larger than SSW, it indicates than there are significant differences in the group means. So, there will be significan differences in the means across the levels of the factor.

In order to compare the relative weight of SSB and SSW on the total variability, we scale them dividing by the number of degrees of freedom, producing the values MSB and MSW respectively.

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$$d = \frac{MSB}{MSW} = \frac{\frac{1}{k-1}\chi^2_{k-1}}{\frac{1}{n-k}\chi^2_{n-k}} \sim F^{k-1}_{n-k}$$

If the assumptions required hold, the statistic (d) computed as MSB⁄MSW distributes as a F-model. This statistic allows for making a decision about the test: the higher its value, the larger (relatively) is the between part when compared with the within variability.

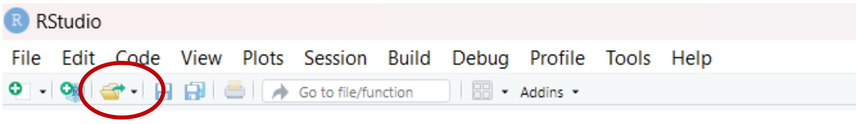
But, how can we know if d is high or not? By calculating the p-value associated to this test: we compute the p-value (the probability at the right tail of the relevant F-distribution) and if this p-value is low we reject the null (i.e., there are significant differences in the mean across levels)

2.2. An example

As an illustrative example, suppose want to test if the design of the packages on which a specific brand of milk is sold, has any influence on the sales. With this objective, we take a sample of 12 stores with similar characteristics and, settign the same price for the milk, we randomly assign one type of packaging (1, 2 or 3). Then we get the sample data of our response varaible "Sales", which measures how many thousands milk bottles were sold in one month, as depicted below:

	Sales	Package
1	2.2	1
2	2.5	1
3	2.4	1
4	2.6	1
5	3.1	2
6	2.8	2
7	3.2	2
8	3.3	2
9	2.5	3
10	2.8	3
11	3.2	3
12	2.5	3

Our sample data shown above is contained in a R file, which we can open by going here (we are calling this data file "Milk"):



We want to test if there are statistically significant differences in the mean sales, depending on the designg of the package. We are applying ANOVA with R, which requires installing specific packages:

```
#install and load the relevant packages
install.packages("car")
install.packages("dplyr")
library(car)
library(dplyr)
```

In order to apply ANOVA, we first need to make sure that the assumptions required actually hold, so we run the following pieces of code:

```
# test normality (by group)
Milk %>%
  group_by(Package) %>%
  summarise(statistic = shapiro.test(Sales)$statistic,
            p.value = shapiro.test(Sales)$p.value)
```

These lines first indicated the dataset that is considered ("Milk"), then group the data by the levels of the factor ("Package") and finally runs a Spahiro normality test on our response variable ("Sales") across groups:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

```
  Package statistic p.value
    <dbl>     <dbl>   <dbl>
1       1     0.971   0.850
2       2     0.927   0.577
3       3     0.854   0.241
```

The high p-values of this normality test for all the levels allow us to work under the required assumption of normality. Additionally, we also assume to have equal variances, which leads us to run a Barlett test of homogenous variances as shown below:

```
# test for homogeneous variances (by group)
bartlett.test(Milk$Sales, Milk$Package)
```

The p-value displayed below suggests that this assumption is hihgly realistic:

```
        Bartlett test of homogeneity of variances

data:  Milk$Sales and Milk$Package
Bartlett's K-squared = 1.2076, df = 2, p-value = 0.5467
```

Given that the necessary assumption seem to hold, we conduct the ANOVA methodology by running the following code lines:

```
# run the ANOVA
anova(lm(Sales ~ Package, Milk))
```

Which produces the following output:

```
Analysis of Variance Table

Response: Sales
          Df  Sum Sq Mean Sq F value Pr(>F)
Package    1 0.21125 0.21125  1.6794 0.2241
Residuals 10 1.25792 0.12579
>
```

The results of the ANOVA test indicates that the differnt designs of the packages seem not to impact on the mean sales: the part of variability explained by the different levels of the factor “Package” (between variability) is not significantly larger than the residual part (within variations). As a consecuence, the p-value associated to this test is high and telss us that there are not reasons to reject the null hypothesis of equal mean sales across designs.

3. Two-factor ANOVA

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.1 The procedure

The ideas explained for the one-factor ANOVA case can be extended to accommodate problems on which more than one factor can be affecting my response variable. Now, the ANOVA test is now extended to account for a second factor plus a possible interaction as:

$$X_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_{ijr}$$

Where X_{ijr} is the value of our response variable for individual r at category (level) i of factor α and level j of factor β. We assume that these values depart from the grand mean (μ), as the sum of four effects:

a. A shift (α_i) that captures the mean influence of belonging to level i of factor α
b. A second shift (β_j) that captures the mean influence of belonging to level j of factor β
c. An interaction term between these two factors $(\alpha\beta)_{ij}$
d. A residual u_{ir}, which accounts for random, uncontrolled variations. This residual is assumed to distribute normally with zero mean

Now the comparisons between the different parts of the variability are more complex. Each source of variation is compared (conveniently scaled by the number of degrees of freedom) with the residual variance. The intuition is the same as in one-factor ANOVA, but there are three different tests, as summarized in the table below:

SOURCE OF VARIATION	SUM OF SQUARES	d.f.	MEAN OF SQUARES	F
Factor α	SSα	k-1	MSα	MSα/MSR
Factor β	SSβ	h-1	MSβ	MSβ/MSR
Interaction (αβ)	SSαβ	(k-1)(h-1)	MSαβ	MSαβ/MSR
Residual	SSR	n-hk	MSR	
Total	SST	n-1		

Co-funded by the Erasmus+ Programme of the European Union

3.2. An example

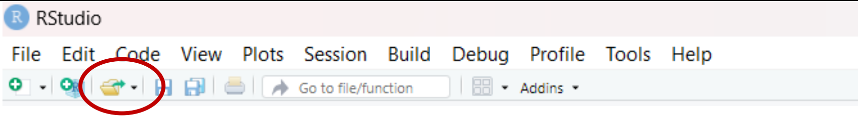
We are going to illustrate empirically of the two-factor ANOVA works, assuming that we have the following problem: A health centre wants to analyze the potential influence of age and sex on the use of a medicine. A sample survey is conducted for this purpose and users were grouped by age into four categories (children, teenagers, adults, seniors) and gender. A sample of 24 individuals was drawn, independently selecting 3 individuals by gender and age group. The response variable is the monthly consumption of this medicine (in €), and we have the followng dataset:

	consumption	sex	age
1	3.0	Male	Child
2	4.0	Male	Child
3	2.8	Male	Child
4	3.2	Female	Child
5	3.0	Female	Child
6	4.1	Female	Child
7	1.8	Male	Teenager
8	1.0	Male	Teenager
9	1.5	Male	Teenager
10	2.1	Female	Teenager
11	1.2	Female	Teenager
12	1.7	Female	Teenager
13	2.5	Male	Adult
14	2.8	Male	Adult
15	3.0	Male	Adult
16	3.0	Female	Adult
17	4.0	Female	Adult
18	2.9	Female	Adult
19	5.0	Male	Senior
20	5.2	Male	Senior
21	6.0	Male	Senior
22	4.9	Female	Senior
23	5.1	Female	Senior
24	6.2	Female	Senior

Again, the sample data shown above (contained in a R file called “medicine), can be loaded in Rstudio by going here:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Now, we are applying a two-factor (age and gender) ANOVA with R, which requires installing and loading specific packages:

```
#install and load the relevant packages
install.packages("car")
install.packages("dplyr")
library(car)
library(dplyr)
```

In order to apply ANOVA, we first test if the assumptions required actually hold, by running normality and equal-variances test. Normality tests (across all the age groups and the two genders) are conducted by running:

```
# we test normality by group first
Medicine %>%
  group_by(age,sex) %>%
  summarise(statistic = shapiro.test(consumption)$statistic,
            p.value = shapiro.test(consumption)$p.value)
```

We first indicate the dataset that is considered ("Medicine"), then group the data by the levels of the tow factors considered in our anlaysis ("age" and "sex") and finally runs a Spahiro normality test on variable "consumption" across all the groups:

```
  age      sex     statistic p.value
  <fct>    <fct>       <dbl>   <dbl>
1 Child    Male        0.871   0.298
2 Child    Female      0.881   0.328
3 Teenager Male        0.980   0.726
4 Teenager Female      0.996   0.878
5 Adult    Male        0.987   0.780
6 Adult    Female      0.818   0.157
7 Senior   Male        0.893   0.363
8 Senior   Female      0.862   0.274
```

Note that now, when referring to the levels of the two factors, we need to consider all pairs of possible categories between them. Agian we find high p-values for this normality test in all the cases, which allow us to work under the required assumption of normality. Moreover, homogenous variances are required as well, and in this case this assumption is tested by conducting a Levene test as:

	`` #testing for equal variances leveneTest(consumption ~ age*sex, data=Medicine, center="mean") ``` The p-value found indicates that we do not have empirical evidence in the sample against this assumption either: ``` Levene's Test for Homogeneity of Variance (center = "mean") Df F value Pr(>F) group 7 0.9575 0.4926 ``` Since the assumptions required to conduct a two-factor ANOVA process seem to hold, we do it by running the following code lines: ``` # two factor ANOVA analysis anova(lm(consumption ~ age*sex, Medicine)) ``` The output of the analysis comes in the form of the following multiple ANOVA table: ``` Analysis of Variance Table Response: consumption Df Sum Sq Mean Sq F value Pr(>F) age 3 45.250 15.0833 51.8625 1.827e-08 *** sex 1 0.327 0.3267 1.1232 0.305 age:sex 3 0.223 0.0744 0.2560 0.856 Residuals 16 4.653 0.2908 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ``` The results of this two-factor ANOVA provides very useful information that allows to give a data-based response to our research question. The tests conducted indicates that the mean values of the consumption of the medicine are significantly different across the four levels of the factor “age” (note that is the only case when we have a low p-value, which leds to reject the null hypothesis of equal means). However, neither we find significant differences in the mean consumption by gender or across the interactions between age-group and gender.
Self-assessment (multiple choice queries and answers)	In one-factor ANOVA, the residuals: a) Are assumed to be correlated b) Are assumed to be normal c) We do not need any assumptions on the residuals The null hypothesis in a one-factor ANOVA states that: a) All the means are the same across levels b) There are only two means that are the same

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	c) All the means are different The two-factor ANOVA statistic to test for the significance of factor $\boldsymbol{\alpha}$ has a distribution: a) Chi-square b) Student's t c) Snedecor's F
Resources (videos, reference link)	
Related material	
Related PPT	
Bibliography	NEWBOLD, P. et al. (2008): Statistics for Management and Economics, (6th edition) Ed. Prentice Hall. Chapter 17, pp. 635-661.
Provided by	[Uniovi]