Training Fiche Template

Title	Correspondence Analysis, AC
Keywords (meta tags)	AC, qualitative variables, explained inertia, eigenvalues
Language	English
Objectives / Goals / Learnig outcomes	**The aim of this module is to introduce and explain the Principal Component Analysis technique.** **At the end of this module you will be able to:** **- Know the logic of AC** **- Know the requirements** **- Conduct an AC** **- Conduct an AC in R with the FactoMineR** package
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 presented the multidimensional analysis technique called Correspondence Analysis, AC. Correspondence Analysis is a form of multidimensional scaling, which essentially builds a kind of spatial model that shows the associations between a set of categorical variables. If the set includes only two variables, the method is usually called Simple Correspondence Analysis (SCA). If the analysis involves more than two variables, then it is usually called Multiple Correspondence Analysis (MCA). In this module we will deal with the analysis of simple correspondences, the objective of this analysis is to reduce the dimensionality of the phenomenon under investigation while preserving the information

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.

	contained by it. The technique is applicable to phenomena measured with qualitative variables. The last part of the module will be dedicated to the application of AC with the R software.
Contents arranged in 3 levels	**1. INTRODUCTION** Correspondence analysis, AC, is a multidimensional analysis technique that is capable of translating almost any type of table consisting of numerical data into graphical form. The object of the AC are the contingency matrices, whose elements indicate the number of times the characteristics of two different quantities have been detected together. The main goal of the AC is to analyze the relationships between two variabiland qualitative observed on a collective of statistical units. This is done through the identification of an "optimal" space, i.e. of a reduced dimension that represents the synthesis of the structural information contained in the original data. The purpose of the analysis is to bring to light the interweaving of links, or correspondences, that exist between the data under examination. **2. REQUIREMENTS FOR MATCHING ANALYSIS** In order to conduct correspondence analysis it is important to analyze the variables to be used to have clear some of their characteristics. Specifically, the variables must have the following requirements: - *The **variables** must be **Qualitative**:* Qualitative variables are variables that are not represented by numbers, but by modalities, for example: gender, level of education, marital status, etc. These modalities, also called categories, must be exhaustive and mutually exclusive. Mutually exclusive means that the variable modalities must not contain the same type of information. For example, for the variable "hair color" you can not enter the modes "dark hair" and "brown hair", as dark hair also means brown hair and vice versa. Estaustive means that the modalities of a variable must take into account all possibilities. For example, for the variable "level of education" the modalities "diploma", "bachelor's degree", "second-level degree" are inserted. These three modalities do not take into account all possible liville of education.

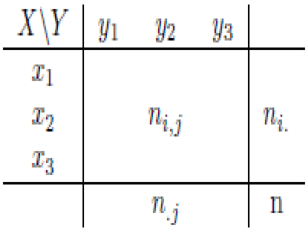
- *The variables must be interdependent*:
  Before performing the analysis of the correspondences it is necessary to verify the degree of interdependence between the two variables considered, as if they were to be independent it may not make sense to conduct the analysis of the matches.
  To do this, perform the chi-square test:
  H_0: the two variables are independent
  H_1: the two variables are not independent
  To interpret the results of the test we observe the p-value:
  p-value < 0.05: the null hypothesis is rejected and consequently the variables are considered with a certain degree of dependence.

3. How to conduct AC

After verifying the CA requirements, you can move on to the actual analysis.

3.1) Contingency tables

In correspondence analysis we work with contingency tables, which contain the joint frequencies of the modes of the two qualitative variables X and Y. These matrices are always made up of never negative integers that are counts , i.e. simple records of what has occurred. In addition, both categorical variables play a symmetric role in which all elements have the same nature.



$X \backslash Y$	y_1	y_2	y_3	
x_1				
x_2		$n_{i,j}$		$n_{i.}$
x_3				
		$n_{.j}$		n

X, Y are the qualitative variables.

x_1, x_2 , x_3 : are the modes of the variable of X

y_1, y_2 , y_3 : are the modes of the variable of Y

$n_{i,j}$: are the absolute joint frequencies, i.e. the frequencies of the pairs, for example $n_{1,1}$: $X = x_1; Y = y_1$

$n_{i\cdot}$: are the row marginals: $n_{i.} = \sum_{j=1}^{C} n_{i,j}$

$n_{\cdot j}$: are the column marginals: $n_{.j} = \sum_{i=1}^{R} n_{i,j}$

These are the sum for the fixed row (or column) of the joint frequencies on the modes of Y (for the columns on the modes of X).

n = is the sample number, which can be obtained by adding the marginals of row or column: $n = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{i,j} \quad \forall\, i,j$

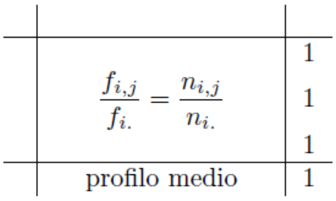
You can switch from absolute frequencies to relative frequencies by dividing each absolute frequency by n: $f_{i,j} = \frac{n_{i,j}}{n}$

3.2) Row Profile Matrix and Column Profile Matrix

The row profiles matrix is obtained by dividing the absolute frequencies (or relative frequencies) by the respective row marginals. Therefore:

$$\frac{n_{i,j}}{n_i} = \frac{f_{i,j}}{f_{i.}} \qquad \forall\, i,j$$

The contingency table will be:



$\frac{f_{i,j}}{f_{i.}} = \frac{n_{i,j}}{n_{i.}}$	1
	1
	1
profilo medio	1

On the marginals of row we have all 1 and this represents the sum of the row profiles.

On the marginals of the column there are the average profiles that are obtained by adding the relative frequencies per column; or by averaging the elements of the row profile array, per column. This is a weighted average, where the masses are represented by the row marginals $f_{i.}$.

Working with frequencies loses a dimension, so the row space is represented by a space C-1 dimensions, i.e.

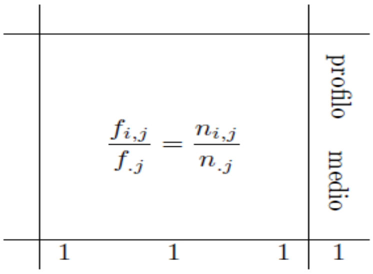
Can be construct a **diagonal matrix of row marginals** $\boldsymbol{D_R}$, that has row profiles on the major diagonal. The diagonal matrix of row marginals is a matrix **R· R,** which has dimensions equal to the rows and on the major diagonal contains the row marginals of the relative frequency table. A diagonal matrix is a matrix whose generic element on the major diagonal is the marginal of row, at above or below it, there are all zeros. It is always a symmetrical and square matrix. With the diagonal matrix of row margins one can construct the **array of row profiles**: it is obtained by dividing the relative frequencies by the row marginals $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$. The dimensions of **F** are R· C, while $\boldsymbol{D_R}$ it has dimension R· R, since the division between matrices cannot be done, one calculates the inverse of $\boldsymbol{D_R}$ and multiplies by **F**, thus solving the dimensionality problem: $\boldsymbol{D_R^{-1}} \cdot \boldsymbol{F}$.

The same goes for the columns, with some small differences.

The column profiles matrix is constructed by dividing the absolute frequencies by the relative column margins:

$$\frac{n_{i,j}}{n_{.j}} = \frac{f_{i,j}}{f_{.j}} \qquad \forall\, i,j$$

The contingency table you get will be:



In this case on the marginals of the column you will have all 1 and on the marginals of row you have the average column profile. In this case the masses are represented by the column marginals $f_{.j}$. Obviously, even in column space you work at less than one dimension, so the column space is R-1.

Can be construct a diagonal matrix of marginals column $\boldsymbol{D_C}$**,** that has column profiles on the major diagonal. The diagonal matrix of column marginals is a matrix **C·C,** that has dimensions equal to the columns and on the major diagonal contains the column marginals of the relative frequency table. A diagonal matrix is a matrix whose generic element on the major diagonal is the marginal of column, above or below it, there are all zeros. It is always a symmetrical and square matrix. With the diagonal matrix of column marginals one can construct the **matrix of column profiles**: it is obtained by dividing the relative frequencies by the column marginals $\frac{\boldsymbol{F}}{\boldsymbol{D_R}}$**.** The dimensions of **F** are R· C, while $\boldsymbol{D_C}$ having dimension C·C, since the division between matrices cannot be done, one calculates the inverse of $\boldsymbol{D_C}$ and post-multiplies to **F**, thus solving the dimensionality problem: $\boldsymbol{F} \cdot \boldsymbol{D_C^{-1}}$.

3.3) Dinstances

In the correspondence analysis it is necessary to understand what distance there is between the values, this in order to understand if the modalities are far or close to each other and therefore if they resemble each other or not. You can do this by observing the frequencies: the lower they are, the closer they are and vice versa. There are various methods for calculating distance: **Euclidean distance** and **chi-square distance**.

The **Euclidean distance** is the simplest and rewards the highest distances at the expense of the lower ones. It is calculated by making the difference of the relative frequencies by raising them to the square.

For row profiles:

$$d_{(i,i')} = \sqrt{\sum_{j=1}^{C} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right)^2}$$

For column profiles:

$$d_{(j,j')} = \sqrt{\sum_{i=1}^{R} \left(\frac{f_{i,j}}{f_{.j}} - \frac{f_{i,j'}}{f_{.j'}} \right)^2}$$

The **chi-square dinstance** rewards the lower distances because the frequencies with low number are reweighted with respect to the rows, inserting in the formula the inverse of the column marginal (respect to the columns, inserting in the formula the inverse of the marginal of row). The disadvantage of chi-square distance is that the reciprocal of column (or row) marginals can tend to zero and therefore a single response can contribute excessively to the calculation of the distance.

3.4) Rows Space and Columns Space

In **rows space** the two components are:

- Row profile: $\mathbf{D_R^{-1} \cdot F}$
- Metric: $\mathbf{D_C^{-1}}$

Let's start with the formula:

$$\boldsymbol{\Psi_{n\times 1} = X_{n\times p} \cdot u_{p\times 1}}$$

By making appropriate substitutions:

$$\boldsymbol{\Psi = D_R^{-1} \cdot F \ \cdot D_C^{-1} \cdot u}$$

The objective of correspondence analysis is the set of unit axes that allow to maximize the distances between the projections of the row profiles. We must, therefore, look for those vectors that maximize projections. Since vectors $\boldsymbol{u}$ can be infinite, the unit norm constraint is added.

$$\boldsymbol{u^T \cdot \ \mathbf{D_C^{-1}} \cdot u = 1}$$

Maximization problem: Maximize the explained inertia (explained variation), which corresponds to the variability for quantitative variables.

$$\begin{cases} \text{MAX: } \left\{\hat{\psi}^T D_R \hat{\psi}\right\} \\ v^T D_C^{-1} v = 1 \end{cases}$$

To solve the constrained maximization problem, use the method of Lagrange multipliers:

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.

$$\mathcal{L}(v,\lambda) = (\hat{\psi}^T D_R \hat{\psi}) - \lambda(v^T D_C^{-1} v - 1)$$

$\boldsymbol{\lambda}$= Lagrange multiplier, which is a scalar;

u= vector of weights we are looking for

By making the necessary replacements, we will have:

$$\mathcal{L}(v,\lambda) = (D_R^{-1} F D_C^{-1} v)^T D_R (D_R^{-1} F D_C^{-1} v) - \lambda(v^T D_C^{-1} v - 1)$$

We perform the transposition operations, substitute a $\boldsymbol{D_R \cdot D_R^{-1}}$ for the identity matrix ***I*** and $[(-\boldsymbol{\lambda}) \cdot (-1)]$ replace it with $\boldsymbol{\lambda}$**.** We can then remove the transpose from the diagonal matrices $\mathbf{D_C^{-1}}$ and $\mathbf{D_R^{-1}}$, since the transpose of a diagonal matrix does not change. Get:

$$\mathcal{L}(v,\lambda) = v^T D_C^{-1} F^T D_R^{-1} F D_C^{-1} v - \lambda v^T D_C^{-1} v + \lambda$$

We calculate the partial derivatives, deriving the Lagrangian respect to $\boldsymbol{u}$ and put them equal to 0:

$$\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow \frac{\partial \mathcal{L}}{\partial v} = 2F^T D_R^{-1} F D_C^{-1} v - 2\lambda v = 0$$

Multiply the equation by $\mathbf{D_C^{-1}}$**:**

$$F^T D_R^{-1} F D_C^{-1} v = \lambda v$$

If we replace the transpose of row profiles and the matrix of column profiles with ***S***, we can write the characteristic equation as:

$$Sv = \lambda v$$

Maximizing the explained inertia of row profiles is equivalent to decomposing this matrix into eigenvalues and eigenvectors of the same. The first eigenvalue is associated with the first eigenvector that explaining the maximum inertia. The eigenvectors that are extracted subsequently, will be extracted orthogonally placing the orthogonality constraint

$$\boldsymbol{u}_1^T \cdot \boldsymbol{D}_C^{-1} \cdot \boldsymbol{u}_2 = \boldsymbol{0}$$

We use the orthogonality constraint to be able to choose the second component that will explain the inertia that is not explained by the first component. Obviously, the first extracted component explains the maximum inertia, that is the maximum elongation of the points cloud.

In **the columns space** two components are:

- Column profile: $\mathbf{F} \cdot \mathbf{D}_C^{-1}$
- Metric: $\mathbf{D}_R^{-1}$

Let's start with the formula:

$$\boldsymbol{\varphi}_{p\times1} = \left(\boldsymbol{X}_{n\times p}^T\right)_{p\times n} \cdot \boldsymbol{v}_{n\times1}$$

We replace and get

$$\boldsymbol{\varphi} = \mathbf{D}_C^{-1}\boldsymbol{F}^T\mathbf{D}_R^{-1}\boldsymbol{v}$$

The maximization problem to be solved with Lagrange multipliers is:

$$\begin{cases} \text{MAX: } \{\hat{\varphi}^T D_C \hat{\varphi}\} \\ \nu^T D_R^{-1}\nu = 1 \end{cases}$$

Proceeding as in the space of the rows, finally we will get:

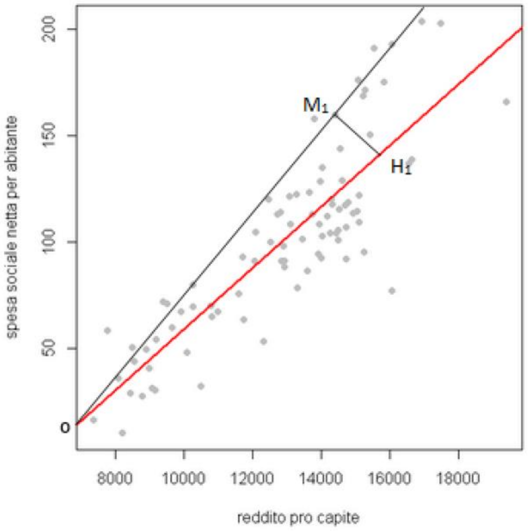
$$F D_C^{-1} F^T D_R^{-1} \nu = \mu\nu$$

Substituting the matrix of column profiles and the transposed metric of row profiles with $\boldsymbol{S}^{*}$ we obtain the characteristic equation:

$$S^{*}\nu = \mu\nu$$

Geometrically maximizing the explained inertia, i.e. making the lost information as small as possible and the observed information as large as possible, will be: make the distance M_1H_1as small as possible and the distance OH_1 as large as possible.

Figura 1.3: Diagramma di dispersione



We must therefore find the straight line f (in red) interpolating the points of vector space, so the distance between all points of the space and points projected orthogonally on the straight line f is the minimum possible.

Eigenvalues in rows space correspond to eigenvectors in column space, so the eigenvalues of **S** correspond to those of $\boldsymbol{S}^{*}$. Eigenvectors are equal to each other except for one constant. So when we have to maximize we don't need to decompose into eigenvalues and eigenvectors **S** and $\boldsymbol{S}^{*}$, just do it with one. The amount of inertia explained is equal whether we calculate **S** or $\boldsymbol{S}^{*}$, the relation between the two spaces is represented by the **transition formulas**:

$$\boldsymbol{S} \to \boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F} \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon} \equiv \boldsymbol{S}^* \to \boldsymbol{\upsilon} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Rows space:

$$\hat{\psi} = \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon}$$

With:

$$\boldsymbol{\upsilon} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

By applying the appropriate substitutions:

$$\frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F}' \boldsymbol{D}_R^{-1} \boldsymbol{\upsilon} \to \frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F}' \hat{\psi}$$

Get:

$$\sqrt{\lambda}\hat{\psi} = \boldsymbol{D}_C^{-1} \boldsymbol{F}' \hat{\psi} \to \hat{\psi} = \frac{1}{\sqrt{\lambda}} \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\psi} \to \sqrt{\lambda}\hat{\psi} = \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\psi}$$

For the space of the rows, therefore:

$$\sqrt{\lambda}\hat{\psi} = \boldsymbol{D}_C^{-1} \boldsymbol{F} \hat{\psi} \equiv \hat{\psi} = \sqrt{\lambda}\hat{\psi}$$

Column space:

$$\hat{\psi} = \boldsymbol{D}_R^{-1} \boldsymbol{\nu}$$

Where:

$$\boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \boldsymbol{F} \boldsymbol{D}_C^{-1} \boldsymbol{\upsilon}$$

By applying the appropriate substitutions:

$$\frac{1}{\sqrt{\lambda}}\boldsymbol{D}_R^{-1}\boldsymbol{F}\boldsymbol{D}_C^{-1}\boldsymbol{v} \rightarrow \frac{1}{\sqrt{\lambda}}\boldsymbol{D}_R^{-1}\boldsymbol{F}\hat{\psi}$$

Get:

$$\frac{1}{\sqrt{\lambda}}\boldsymbol{D}_R^{-1}\boldsymbol{F}\hat{\psi} \rightarrow \sqrt{\lambda}\hat{\psi} \rightarrow \boldsymbol{D}_R^{-1}\boldsymbol{F}\hat{\psi}$$

For column space:

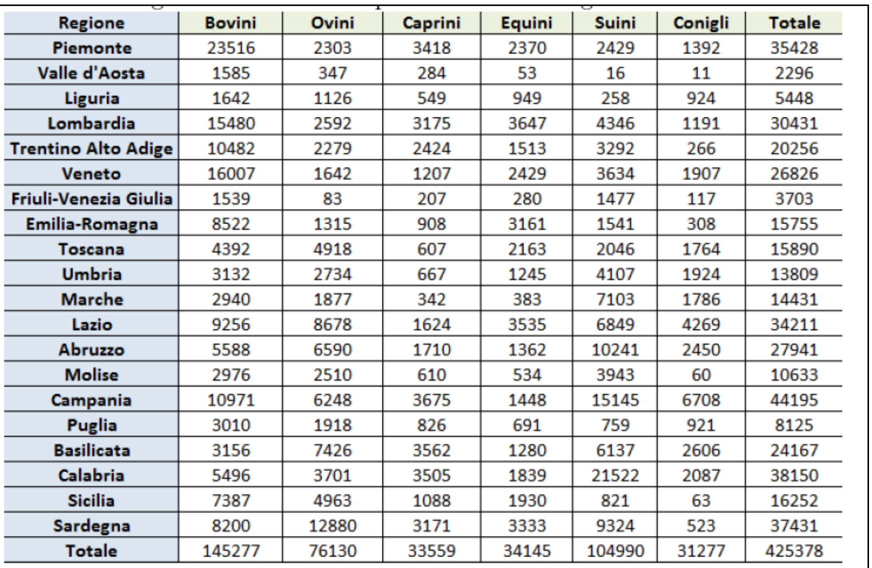
$$\sqrt{\lambda}\hat{\psi} = \boldsymbol{D}_R^{-1}\boldsymbol{F}\hat{\psi} \equiv \hat{\psi} = \sqrt{\lambda}\hat{\psi}$$

4) Example with R software

Verify a possible relationship between the distributions of livestock and the different Italian regions. The data refer to the year 2011, collected by the banks available on the Istat website.

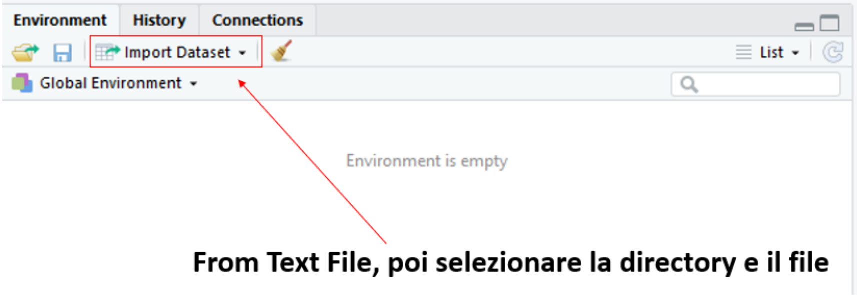
Hypothesis: the various regions, depending on the territorial characteristics and the needs of the population, choose to raise some heads of cattle rather than others.

Dataset:



Regione	Bovini	Ovini	Caprini	Equini	Suini	Conigli	Totale
Piemonte	23516	2303	3418	2370	2429	1392	35428
Valle d'Aosta	1585	347	284	53	16	11	2296
Liguria	1642	1126	549	949	258	924	5448
Lombardia	15480	2592	3175	3647	4346	1191	30431
Trentino Alto Adige	10482	2279	2424	1513	3292	266	20256
Veneto	16007	1642	1207	2429	3634	1907	26826
Friuli-Venezia Giulia	1539	83	207	280	1477	117	3703
Emilia-Romagna	8522	1315	908	3161	1541	308	15755
Toscana	4392	4918	607	2163	2046	1764	15890
Umbria	3132	2734	667	1245	4107	1924	13809
Marche	2940	1877	342	383	7103	1786	14431
Lazio	9256	8678	1624	3535	6849	4269	34211
Abruzzo	5588	6590	1710	1362	10241	2450	27941
Molise	2976	2510	610	534	3943	60	10633
Campania	10971	6248	3675	1448	15145	6708	44195
Puglia	3010	1918	826	691	759	921	8125
Basilicata	3156	7426	3562	1280	6137	2606	24167
Calabria	5496	3701	3505	1839	21522	2087	38150
Sicilia	7387	4963	1088	1930	821	63	16252
Sardegna	8200	12880	3171	3333	9324	523	37431
Totale	145277	76130	33559	34145	104990	31277	425378

We import the dataset:

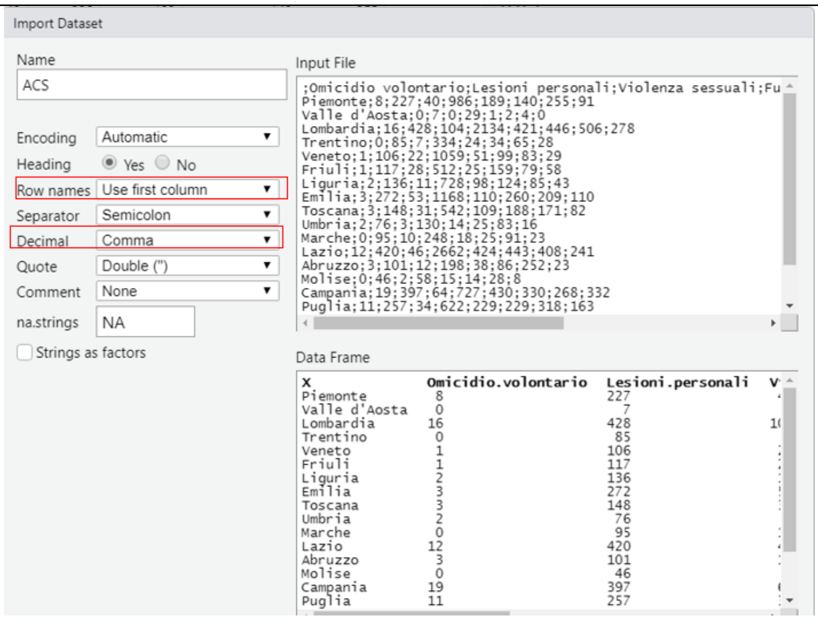


Environment is empty

From Text File, poi selezionare la directory e il file

In the ***row names*** field, select the wording: "***use first column***" in order to have the labels of both individuals and variables on the graphs.

In the ***decimal*** field we select "***comma***".



Import Dataset

Name: ACS

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Use first column

Separator: Semicolon

Decimal: Comma

Quote: Double (")

Comment: None

na.strings: NA

□ Strings as factors

Input File

```
;Omicidio volontario;Lesioni personali;Violenza sessuali;Fu
Piemonte;8;227;40;986;189;140;255;91
Valle d'Aosta;0;7;0;29;1;2;4;0
Lombardia;16;428;104;2134;421;446;506;278
Trentino;0;85;7;334;24;34;65;28
Veneto;1;106;22;1059;51;99;83;29
Friuli;1;117;28;512;25;159;79;58
Liguria;2;136;11;728;98;124;85;43
Emilia;3;272;53;1168;110;260;209;110
Toscana;3;148;31;542;109;188;171;82
Umbria;2;76;3;130;14;25;83;16
Marche;0;95;10;248;18;25;91;23
Lazio;12;420;46;2662;424;443;408;241
Abruzzo;3;101;12;198;38;86;252;23
Molise;0;46;2;58;15;14;28;8
Campania;19;397;64;727;430;330;268;332
Puglia;11;257;34;622;229;229;318;163
```

Data Frame

```
X               Omicidio.volontario   Lesioni.personali   V
Piemonte         8                    227
Valle d'Aosta    0                      7
Lombardia       16                    428
Trentino         0                     85
Veneto           1                    106
Friuli           1                    117
Liguria          2                    136
Emilia           3                    272
Toscana          3                    148
Umbria           2                     76
Marche           0                     95
Lazio           12                    420
Abruzzo          3                    101
Molise           0                     46
Campania        19                    397
Puglia          11                    257
```

With the command:

X<-as.matrix(nome_del_dataset)

We attribute to **X**, as an object, the dataset used in the analysis.

Before being able to perform the AC it is necessary to establish the degree of interdependence between the two characters considered, this is because in the event that they are independent it may not make sense to continue AC. To verify this we perform th chi-square test.

The command is:

chiquadro<-chisq.test(X)

```
        Pearson's Chi-squared test

data:  X
X-squared = 126691.2, df = 95, p-value < 2.2e-16
```

It can be observed that the ***p-value*** is lower than the most commonly used significance level i.e. 0.05. We can therefore reject the null hypothesis of statistical independence between the two variables and we can continue with the analysis.

Now we want to create a matrix of relative frequencies **F**.

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.

We calculate the sample number, with the command:

```
n<-sum(X)
```

and then dividing the starting matrix (therefore all the joint frequencies) by the sample number we obtain the matrix **F**. Command:

```
F<-X/n
```

The next step is to get the **row and column profile** tables. In order to do this, first of all, it is necessary to calculate the marginals of row and column. Respectively the commands are:

```
sumrow<-apply(F,1,sum)
sumcol<-apply(F,2,sum)
```

Then we calculate the diagonal matrix of the marginals of row and its inverse with the commands:

```
Dr<-diag(sumrow)
Dr_inv<-solve(Dr)
```

Now we can calculate row profiles. In matrix terms we premultiply the inverse of the diagonal matrix of the marginal row to the matrix of relative frequencies. The command to use is:

```
Pr<-Dr_inv%*%F
```

The same thing for column profiles, remembering that in this case the inverse of the column matrix must be post-multiplied to the matrix of relative frequencies.

```
Dc<-diag(sumcol)
Dc_inv<-solve(Dc)
Pc<-F%*%Dc_inv
```

Now we can calculate the distances between the points. As already mentioned, there are two types of distance: **Euclidean** and **Chi-square**.

Euclidean distance **row profiles**:

```
d_euc_r<-dist(rbind(Pr[1,],Pr[2,]))
```

Euclidean distance **column profiles**:

```
d_euc_c<-dist(rbind(Pr[,1],Pr[,2]))
```

Distance of chi-square **row profiles**:

```
d_r<-pr[1,]-pr[2,]
d<-d_r^2/sumcol
d_chi_r<-sqrt(sum(d))
```

Chi-square distance **column profiles:**

```
dc<-Pr[,1]-Pr[,2]
dc<-dc^2/sumrow
d_chi_c<-sqrt(sum(dc))
```

The characteristic equation of the **row profile** matrix:

```
S<-t(Pr)%*%Pc
```

Since the matrix S is not symmetric, it is necessary to diagonalize it to obtain **S_tilde:**

```
A<-t(F)%*%Dr_inv%*%F #simmetria

Dc_12<-diag(sumcol^(-1/2))

S_tilde<-Dc_12%*%A%*%Dc_12
```

Now we have to maximize the inertia explained by decomposing the matrix into eigenvalues and eigenvectors:

```
AC<-eigen(S_tilde)

lambda<-as.matrix(AC$values)

lambda<-lambda[-1,]

w<-AC$vectors

u<-Dc^(1/2)%*%w

u<-u[,-1]
```

The characteristic equation of the column **profile** matrix **:**

```
S_star<-F%*%Dc_inv%*%t(F)%*%Dr_inv
```

To move from ***u*** to ***v***, we use transition formulas (since the amount of inertia explained is equal in both row and column space).

```
sq_lambda<-diag((sqrt(lambda))^(-1))

v<-F%*%Dc_inv%*%u%*%sq_lambda
```

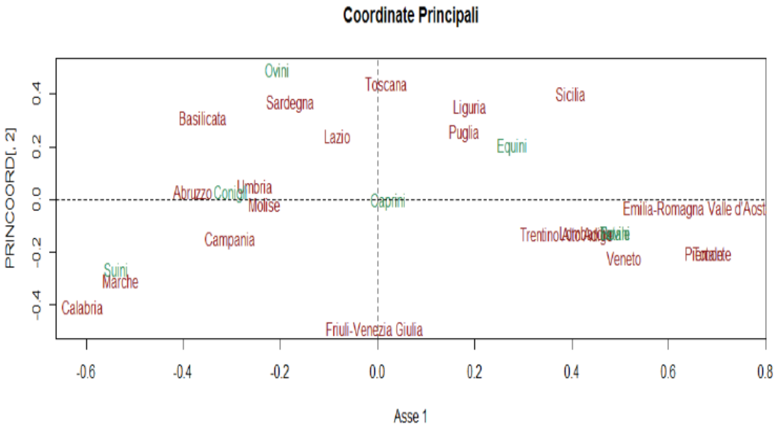
We calculate factors and coordinates, first row space and then columns:

```
fp_r<-Dc_inv%*%u

fp_c<-Dr_inv%*%v

PHI_coord<-Dc_inv%*%t(F)%*%fp_c

PSI_coord<-Dr_inv%*%F%*%fp_r
```

We display the graph of the main coordinates:

```
PRINCOORD<-rbind(PSI_coord,PHI_coord)

rows<-row.names(X);columns<-colnames(X)

plot(PRINCOORD[,1],PRINCOORD[,2],type="n",main="Main Coordinates",xlab="Axis1",ylab="Axis2")+
text(PRINCOORD[1:20,1],PRINCOORD[1:20,2],labels=rows,col="springgreen4")

text(PRINCOORD[21:29,1],PRINCOORD[21:29,2],labels=columns,col="violetred")

abline(h=0,v=0,lty=2,lwd=1.5)
```

We obtain:



Looking at this graph we can say, for example, that in regions such as Abruzzo, Molise, Umbria rabbits are mainly bred.

We choose the components:

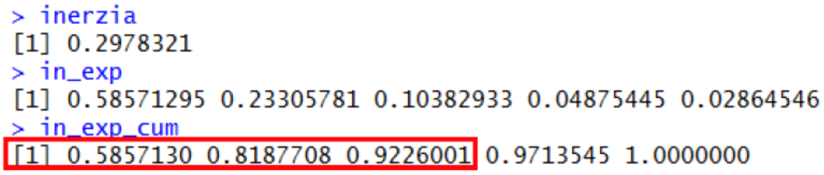
inertia<-sum(diag(S))-1

sum(lambda)

in_exp<-lambda/inertia

in_exp_<-cumsum(in_exp)

We visualize the results obtained:



```
> inerzia
[1] 0.2978321
> in_exp
[1] 0.58571295 0.23305781 0.10382933 0.04875445 0.02864546
> in_exp_cum
[1] 0.5857130 0.8187708 0.9226001 0.9713545 1.0000000
```

The first dimension alone explains 58.57% of the variability, and the first three together explain 92.26% of the overall variability of the data.

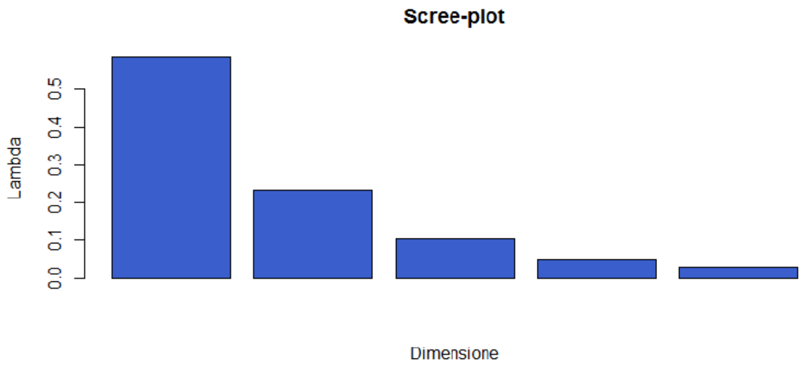
The results obtained can be displayed graphically with the **scree-plot of the inertia explained**:

screeplot<-barplot(in_exp,main="Scree-plot inertia", xlab="Size", ylab="Lambda", col="lightblue")

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.

Figura 1.10: Scree-plot dell'inerzia spiegata



For the quality of the representation:

- to evaluate how much a mode influences or participates in the factorial axis we calculate **the absolute contributions**, **CA,** both for rows and columns:

ca_r<-Dr%*%fp_c^2

ca_c<-DC%*%fp_r^2

- To evaluate the quality of the representation we calculate the **relative contributions, CR.** These give a better measure of the representation of the points on the axes and is given by the cosine of the angle formed by the projection vector of the point and the relative vector i (or j) at *the* point *i* (or *j*) in its original space:

G<-matrix(sumcol,20,9,byrow=T)

di<-(Pr-G)^2%*%Dc_inv

d_ig<-apply(di,1,sum)

cos2r<-PSI_coord^2/d_ig

H<-matrix(sumrow,20,9)

dj<-Dr_inv%*%(Pc-H)^2

d_jh<-apply(dj,2,sum)

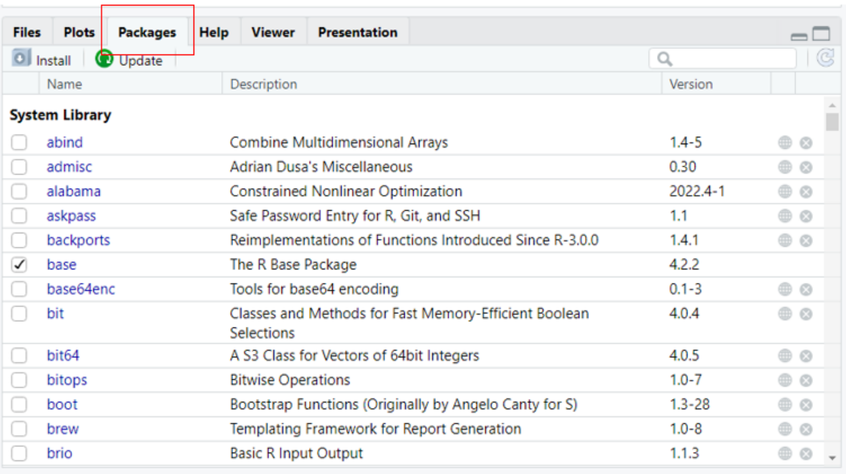
COS2C<-PHI_coord^2/d_jh

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

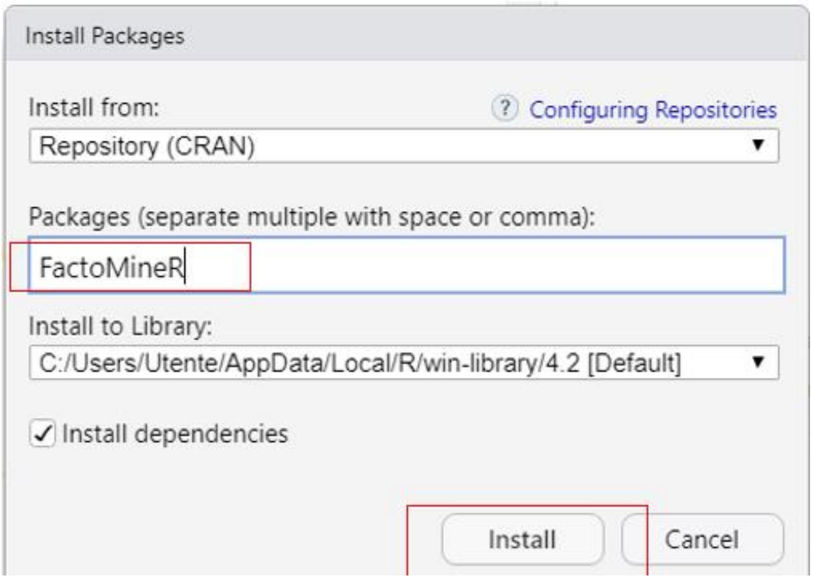
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

R for the analysis of correspondences provides a package called **FactoMineR**, which adds information on individuals and variables and allows you to create a joint two-dimensional graph of individuals and variables.

On R to be able to use this package you must first download it:



	Name	Description	Version
	System Library		
□	abind	Combine Multidimensional Arrays	1.4-5
□	admisc	Adrian Dusa's Miscellaneous	0.30
□	alabama	Constrained Nonlinear Optimization	2022.4-1
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
☑	base	The R Base Package	4.2.2
□	base64enc	Tools for base64 encoding	0.1-3
□	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
□	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
□	bitops	Bitwise Operations	1.0-7
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	brew	Templating Framework for Report Generation	1.0-8
□	brio	Basic R Input Output	1.1.3



Install from: Repository (CRAN)

Packages (separate multiple with space or comma): FactoMineR

Install to Library: C:/Users/Utente/AppData/Local/R/win-library/4.2 [Default]

☑ Install dependencies

After installing it you need to call it with the command

library(FactoMineR)

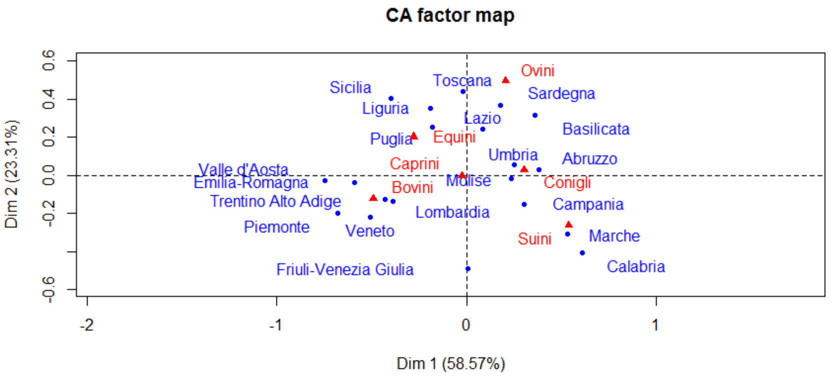
Let's move on to the creation of the two-dimensional graph Individuals and variables:

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.

CA(X, ncp = 5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL)

Graphically we will have:



Interpretation of results:

We can say that the initial hypothesis is confirmed. In particular, the regions most dedicated to sheep farming seem to be Tuscany, Sardinia and Basilicata, and this can be explained by the fact that these regions are mountain and transhumance areas. Horses are mostly bred in Puglia, Liguria and Sicily because these animals have always been used for work in the countryside. Cattle are present in Trentino Alto-Adige, Veneto, Piedmont, Lombardy and Emilia-Romagna; In fact, these regions have a tradition of more developed breeding for food use. Rabbits appear mainly in Umbria, Abruzzo and Molise. Instead, pigs seem to be more reared in the Marche, Campania and Molise; These regions also have a tradition of more developed breeding for food use. Goats, on the other hand, are placed in the middle of the axes, probably because there aren't regions that prefer their breeding.

Self-assessment (multiple choice queries and answers)

1. What do transition formulas do?

   A) Switch between spaces
   B) Move from the representation of absolute contributions to that of related contributions
   C) Switch from the matrix of frequencies relative to those of the profiles

	2. Why do the chi-square test before implementing AC? A) To verify whether the variables are quantitative B) To assess whether the variables are qualitative C) To analyze the existence of interdependence between the two variables 3. What is the goal of Correspondence Analysis? A) Maximize the explained variability B) Maximize the explained inertia C) Minimize explained inertia
Resources (videos, reference link)	
Related material	
Related PPT	
Bibliography	van der Heijden, P. G. M. & de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis, Psychometrika, 50, pp. 429-447. Le, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18. Mineo, A. M. (2003). Una Guida all'utilizzo dell'Ambiente Statistico R, http://cran.r-project.org/doc/contrib/Mineo-dispensaR.pdf.
Provided by	[Unisalento]