University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

*Statistics for Economics and Business*

*Stefano Bonnini & Valentina Mini*

# Hierarchical and Non hierarchical CA: Questions and answers

*Lecture 12 – 29nd of March 2019*

In cluster analysis, the group membership of the individual observations is determined such that:

1. The various groups are as heterogeneous as possible and the observations within a group are as homogeneous as possible

2. The various groups are as homogeneous as possible and the observations within a group are as homogeneous as possible

3. The various groups are as homogeneous as possible and the observations within a group are as heterogeneous as possible

In hierarchical cluster analysis, the agglomerative methods:

1. Start treating the entire sample as a cluster
2. Start treating each statistical units as a cluster
3. Stop the process treating each statistical units as a cluster

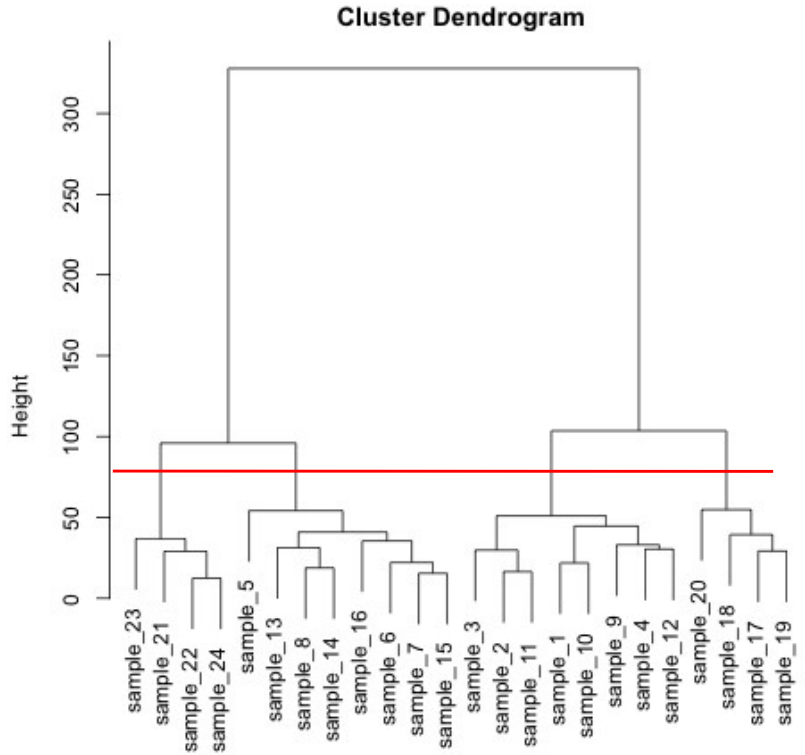
In hierarchical cluster analysis the distance is a key information, and it's computed for:

1. Numeric variables
2. Categorical variables
3. Both

We may represent the hierarchical cluster analysis using a dendogram.

Considering the following dendogram, at distance level of 70, how many clusters we individuate?

1. 2 clusters
2. 3 clusters
3. 4 clusters



Considering the Minkowski distance:

$$_m d_{iu} = \left[\sum_{j=1}^{k} |x_{ij} - x_{uj}|^m\right]^{1/m}$$

when m=2 we obtain the:

1. Chebicev distance
2. Manhattan distance
3. Euclidean distance

Euclidean distance: $_2 d_{iu} = \|\boldsymbol{x}_i - \boldsymbol{x}_u\| = \left[\sum_{j=1}^{k} (x_{ij} - x_{uj})^2\right]^{1/2}$

In hierarchical cluster analysis, the single linkage method is based on:

1. The minimum distance between two clusters
2. The maximum distance between two clusters
3. The average distance between two clusters

Among the non-hierarchical cluster, the K-means method is the most popular. It is based on:

1. The centroids calculation
2. The density calculation
3. Both

The non hierarchical cluster partitioning method (i.e. k-means):

1. Begins with an initial indication of variables choice
2. Begins with an initial definition of hierarchical structure
3. Begins with an initial solution in terms of number of clusters

Example using R:
CAanalysis=kmeans(standardized data matrix, number of clusters)

Which of the following sentences is true?

1. the k-means partitioning technique doesn't allow objects to change group membership through the cluster formation process

2. the k-means partitioning technique allows objects to change group membership through the cluster formation process

3. The hierarchical cluster technique allows objects to change group membership through the cluster formation process

University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

*Statistics for Economics and Business*

# General Questions and answers

The product between a row vector and a column vector is:

1. a matrix
2. a scalar
3. a diagonal matrix

$$\underset{(m\times n)}{\mathbf{A}} \cdot \underset{(n\times h)}{\mathbf{B}} = \underset{(m\times h)}{\mathbf{C}}$$

Thus the product between a row vector and a column vector is a scalar; the product between a column vector and a row vector is a matrix:

$$\underset{1\times n}{\mathbf{a}} \cdot \underset{n\times 1}{\mathbf{b}} = \underset{1\times 1}{c} \qquad \underset{n\times 1}{\mathbf{b}} \cdot \underset{1\times n}{\mathbf{a}} = \underset{n\times n}{\mathbf{C}}$$

Given two vectors **a** and ***b***, such that :

$$\mathbf{a} = \begin{pmatrix} 2 & 4 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

The product between a x b is:

1. 10
2. 42
3. 18

$$\underset{1\times 2}{\mathbf{a}} = \begin{pmatrix} 2 & 4 \end{pmatrix} \qquad \underset{2\times 1}{\mathbf{b}} = \begin{pmatrix} 5 \\ 2 \end{pmatrix} \qquad \Rightarrow \qquad \mathbf{a} \times \mathbf{b} = 2 \cdot 5 + 4 \cdot 2 = 18$$

Given the matrix A= $\begin{pmatrix} 3 & 6 & 4 \\ 2 & 8 & 9 \\ 2 & 5 & 1 \end{pmatrix}$

the transpose matrix A' is:

1. A' = $\begin{pmatrix} 3 & 2 & 2 \\ 6 & 8 & 5 \\ 4 & 9 & 1 \end{pmatrix}$

2. A' = $\begin{pmatrix} 2 & 2 & 3 \\ 5 & 8 & 6 \\ 1 & 9 & 4 \end{pmatrix}$

3. A' = $\begin{pmatrix} 2 & 5 & 1 \\ 2 & 8 & 9 \\ 3 & 6 & 4 \end{pmatrix}$

The **transpose** of the matrix $\boldsymbol{A}=(a_{ij})$ is the matrix $\boldsymbol{A}'=(a_{ji})$ whose rows correspond to the columns of $\boldsymbol{A}$:

$$\mathbf{A} = \begin{pmatrix} 3 & 6 & 4 \\ 2 & 8 & 9 \\ 2 & 5 & 1 \end{pmatrix} \qquad \mathbf{A}' = \begin{pmatrix} 3 & 2 & 2 \\ 6 & 8 & 5 \\ 4 & 9 & 1 \end{pmatrix}$$

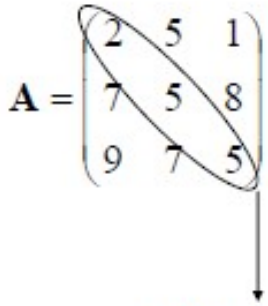
Given the matrix A = $\begin{pmatrix} 2 & 5 & 1 \\ 7 & 5 & 8 \\ 9 & 7 & 5 \end{pmatrix}$

the trace of A = tr(A) is:

1. 50
2. 12
3. 15

The **trace** of $\boldsymbol{A}=(a_{ij})$ is the sum of the elements in the main diagonal of $\boldsymbol{A}$:

$$tr(\boldsymbol{A})=\Sigma_i\, a_{ii}$$




$$tr(\mathbf{A}) = 2+5+5 = 12$$

Given two matrix **A** and **B**, such that:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 6 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 2 & 5 \\ 7 & 3 \end{pmatrix}$$

The *det*(A*B) is:

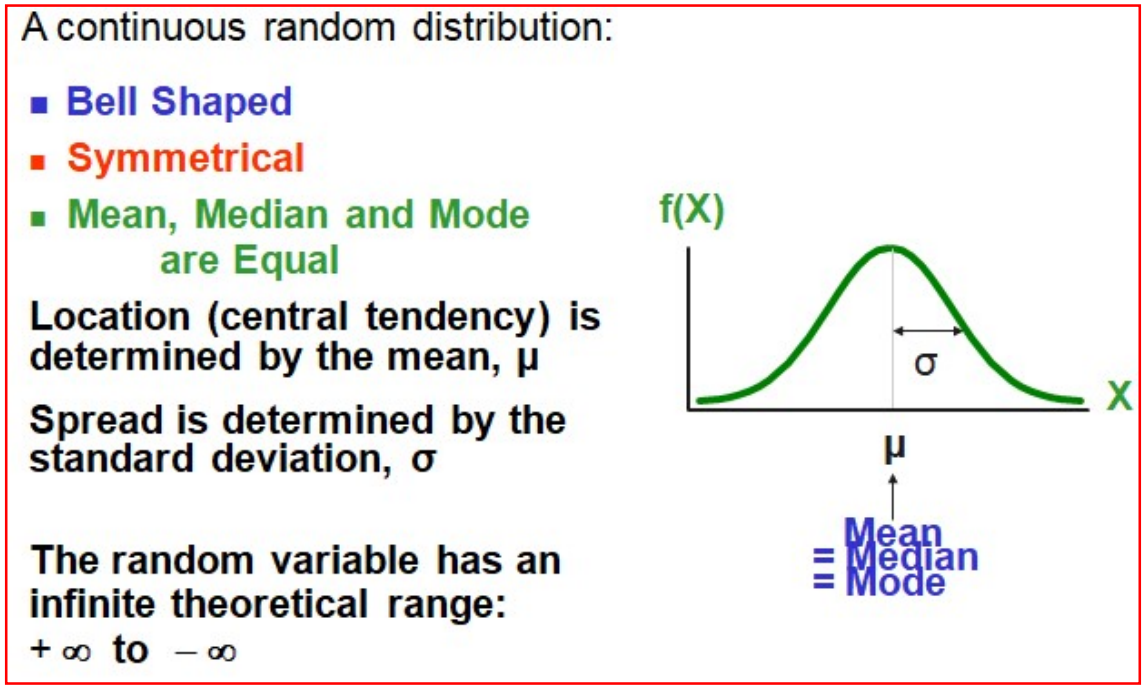
1. 20
2. 38
3. -261

$$\det(\mathbf{A}) = 2 \cdot 6 - 3 \cdot 1 = 9 \qquad \det(\mathbf{B}) = 2 \cdot 3 - 5 \cdot 7 = -29$$

$$\det(\mathbf{A}) \cdot \det(\mathbf{B}) = 9 \cdot (-29) = -261$$

A normal continuous distribution is characterized by:

1. a bell shape
2. a difference between median and mean
3. a random variable which has a finite theoretical range



Simple linear regression analysis is used to:

1. Explain the impact on the dependent variable of changes in independent (explanatory) variable X
2. Explain the independent variable using a dependent variable
3. Explain the independent variable using 2 or more dependents variables

**Y** (dependent variable) ← **X** (independent variable)

One dependent variable explained by one independent variable
We try to explain the impact on Y caused by changes in X

Considering the Simple Linear Regression analysis, the least squares method identifies the regression coefficients by finding the values that:

1. Minimize the difference between Y and X
2. Minimize the sum of the squared differences between Y and $\hat{Y}$
3. Maximize the sum of the squared difference between Y and $\hat{Y}$

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Considering the Simple Linear Regression analysis:

1. $b_0$ is the estimated mean value of Y when X is 0
2. $b_0$ is the estimated mean value of Y when X is 1
3. $b_0$ is the estimated change in Y when X changes by 1 unit

$$b_0 = \bar{Y} - b_1 x_1$$

Considering the following output

```
> cor(torta)
                 settimana     vendita     prezzo       pubb       pr_non.surge   pr_panna  vendita.panna  giorni.di.festa
settimana        1.00000000
vendita          0.03360076  1.00000000
prezzo          -0.10014845 -0.10209557  1.000000000
pubb             0.19279946  0.19514066 -0.001526334  1.000000000
prezzo_non.surge -0.33221180 -0.36502135 -0.113666725  0.052860721      1.00000000
prezzo_panna     -0.23453792 -0.05114394  0.654599388 -0.090582798     -0.01416071  1.00000000
vendita.panna     0.05384546  0.80734983 -0.111172219 -0.033649346     -0.30566582 -0.08676635    1.00000000
giorni.di.festa   0.09359796 -0.33030785 -0.215219045  0.025079631      0.32507725 -0.07861741   -0.12425313      1.00000000
```

we can say that:

1. The correlation between “pubb” and “vendita” is negative
2. The correlation between “prezzo_panna” and “vendita” is negative
3. The correlation between “pubb” and “pubb” doesn’t exist

Among the following, which is not an assumptions of linear regression model ?

1. Linearity assumption (between y and x)
2. Independence assumption
3. Discrete distribution assumption

Assumptions of the model:

- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

Considering the following output,

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x           1 105.748 105.748  113.23 1.823e-07 ***
Residuals 12  11.207   0.934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

the coefficient of determination is:

1. 0.934
2. 0.105
3. 0.904

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as $r^2$

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } sum \text{ of squares}}{total \text{ sum of squares}}$$

note: $0 \le r^2 \le 1$

*in our example: 105.748/(105.748+11.207)*

Considering the following output,

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated linear regression model is:

1. Y = 0.5262- 0.1569*X
2. Y = 1.6699 + 0.9645*X
3. Y = 0.9645+ 1.6699*X

Considering the following output,

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9664 on 12 degrees of freedom
```

The T-statistic value is:

1. 10.641
2. 1.6699
3. 1.833

Considering the following output,

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1751 -0.4982  0.1616  0.6278  2.3758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12803    0.17372   0.737 0.000331 ***
LIKE_AROMA   0.42853    0.05601   7.652 1.63e-13 ***
LIKE_SWEET   0.19714    0.05441   3.623 0.000331 ***
LIKE_TASTE   0.24836    0.05409   4.591 5.99e-06 ***
---
:

Residual standard error: 1.042 on 382 degrees of freedom
Multiple R-squared:  0.5919,    Adjusted R-squared:  0.5887
F-statistic: 184.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

The performed analysis concerns:

1. Simple linear regression
2. Multiple linear regression
3. Hierarchical Cluster

Considering the following output,

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1751 -0.4982  0.1616  0.6278  2.3758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12803    0.17372   0.737 0.000331 ***
LIKE_AROMA   0.42853    0.05601   7.652 1.63e-13 ***
LIKE_SWEET   0.19714    0.05441   3.623 0.000331 ***
LIKE_TASTE   0.24836    0.05409   4.591 5.99e-06 ***
---
:

Residual standard error: 1.042 on 382 degrees of freedom
Multiple R-squared:  0.5919,    Adjusted R-squared:  0.5887
F-statistic: 184.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

which percentage of Y variability is explained by the model?

1. Around 10%
2. Around 59%
3. Around 62%

Considering the following output,

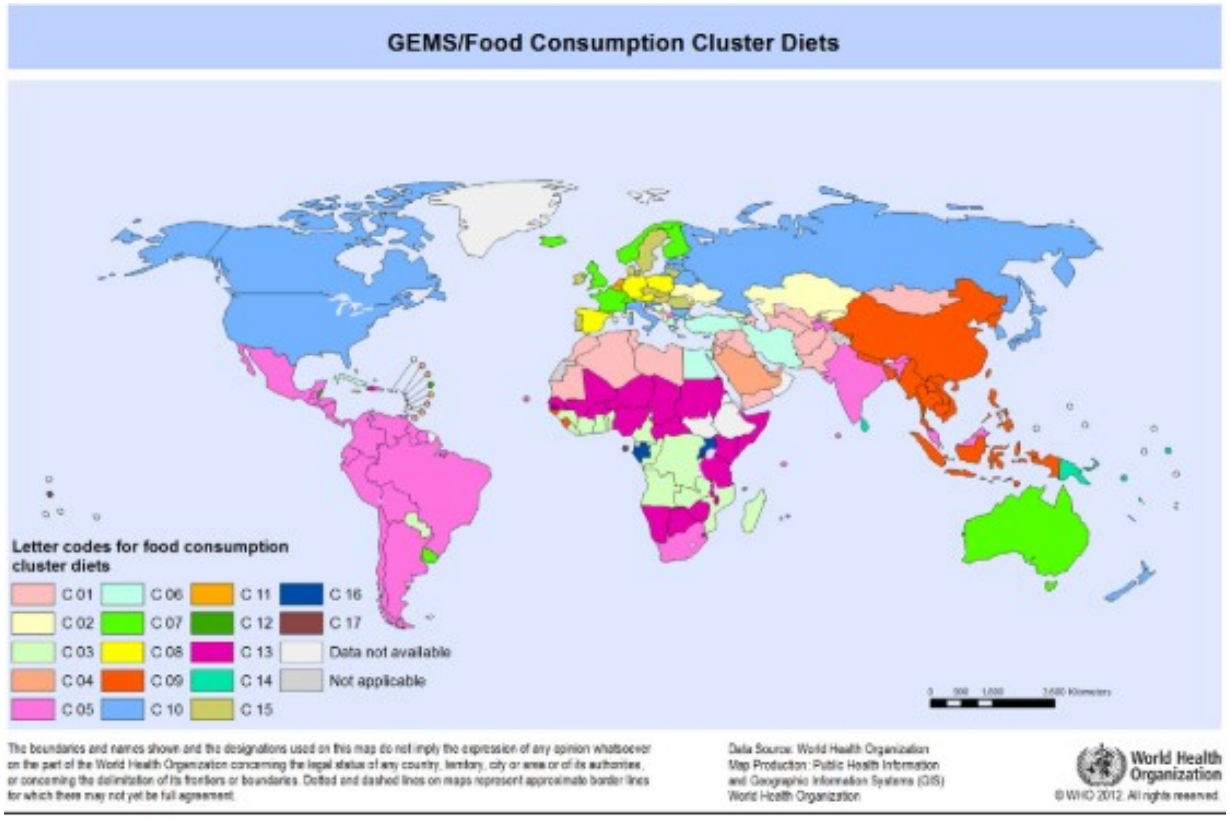
```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1751 -0.4982  0.1616  0.6278  2.3758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12803    0.17372   0.737 0.000331 ***
LIKE_AROMA   0.42853    0.05601   7.652 1.63e-13 ***
LIKE_SWEET   0.19714    0.05441   3.623 0.000331 ***
LIKE_TASTE   0.24836    0.05409   4.591 5.99e-06 ***
---
:

Residual standard error: 1.042 on 382 degrees of freedom
Multiple R-squared:  0.5919,    Adjusted R-squared:  0.5887
F-statistic: 184.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

we reject the null hypothesis that all the **regression** coefficients are equal to zero: thus, we can say that there are sufficient evidence about the predictive capability of our model…at which level of confidence level?

1. 90%
2. 95%
3. 99.9%

# Lab using R

# Profiling countries
# by their food consumption



**Research questions**:

Can we profile countries categories observing their food consumption? What results we may achieve?

**Database**:

eating

**Method**:

Cluster Analysis

*Please, perform the analysis and write-up a small dissertation on the topic and results*

# Example

*in a word document set up all the following sections*

- Title
- Contents
- Introduction (contextualizing the issue)
- The dataset
- The research questions and the variables of interest
- The statistical method applied
- Discussion of the main results and comments
- Final conclusions
- *Appendix = script and commands' description*
- References