Session 3:
Profiling

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information System track

Prof. Lidia Montero

lidia.montero@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Information is stored in tables

Rows of table

Represent individuals or instances, Also called sample, example, record, … they can be repeated, at least theoretically, forming the population under study.

Thing to be classified, associated, or clustered

Characterized by a predetermined set of attributes

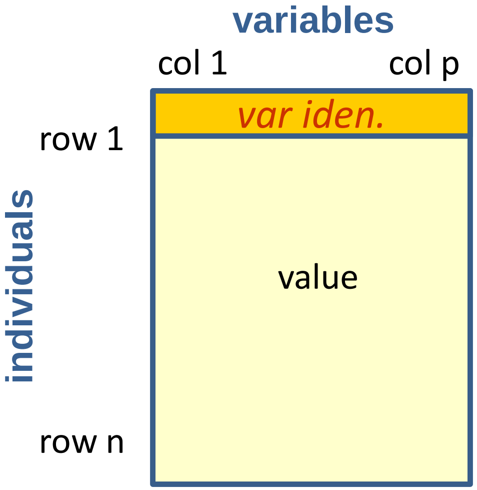
Columns of a table

Each instance is described by a predefined set of features, its variables or "attributes". A variable is a measure of individuals which can take different values (according a probabilistic function).

Possible attribute types ("levels of measurement"): binary, nominal, ordinal, interval, ratio, textual, …

Restriction: Same variables measured in all individuals and in the same order, but different formats are possible (fixed, csv, …), forming a Table.

First rows usually contain the dictionary of variables (var. labels)



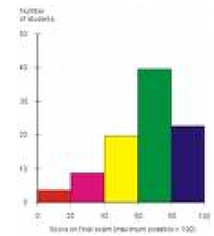
Table

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

FIB
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

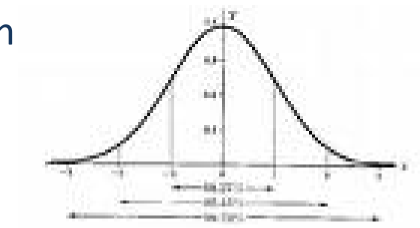
Coding of variables

- **Categorical**
  - Binary (yes/no variable, boolean attribute. Binomial distribution)
  - Nominal (marital status, region, …. Multinomial distribution)
  - Ordinal (size of clothe's size, social class, …., Ordered multinomial)
  - Pareto chart



Prob. distribution

- **Continuous or numeric**
  - Count data ("number of words of a sentence", "number of unem country", …. Poisson distribution)
  - Interval ("temperature,… Laplace-Gauss distribution)
  - Ratio data (age, speed, … Laplace Gauss distribution)



But distinctions are often blurred. Ordinal data can be treated as continuous, …

- **Multiresponse**

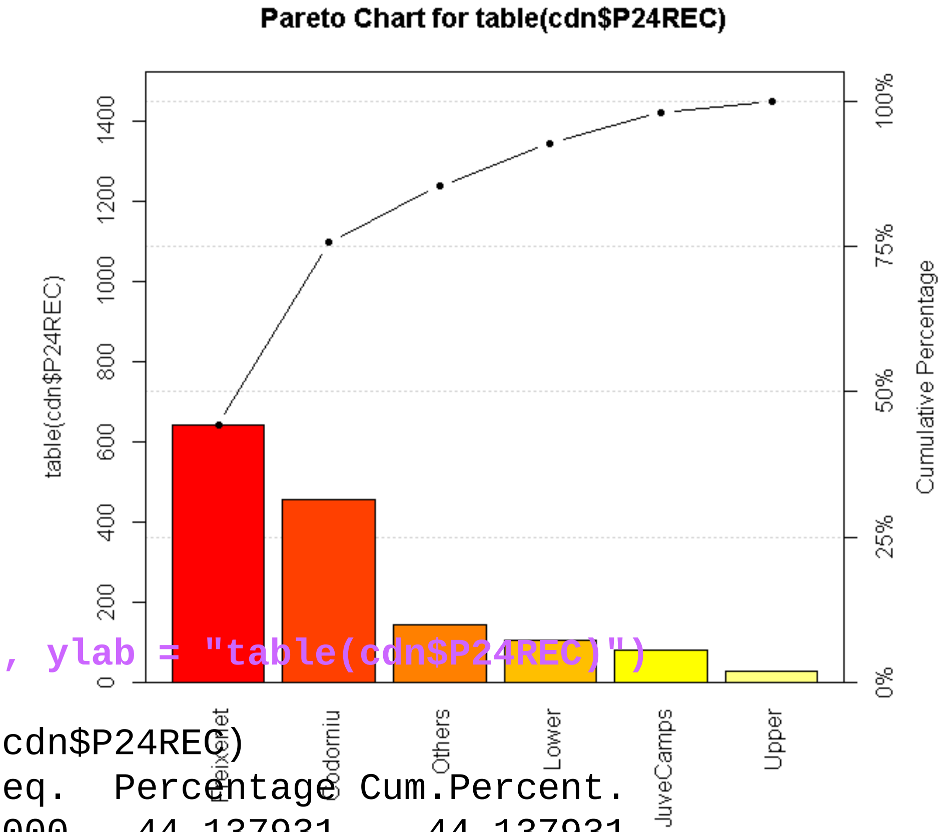
  When several responses are possible.

  i.e. Visited countries?. Each potential visited country becomes a binary variable

- **Textual**

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Pareto chart: a classical Quality Control chart




```
> library(qcc)
Loading required package: MASS
Package 'qcc', version 2.2

> pareto.chart(table(cdn$P24REC), ylab = "table(cdn$P24REC)")
```

```
Pareto chart analysis for table(cdn$P24REC)
             Frequency   Cum.Freq.  Percentage Cum.Percent.
  Freixenet 640.000000  640.000000   44.137931    44.137931
  Codorniu  457.000000 1097.000000   31.517241    75.655172
  Others    142.000000 1239.000000    9.793103    85.448276
  Lower     105.000000 1344.000000    7.241379    92.689655
  JuveCamps  79.000000 1423.000000    5.448276    98.137931
  Upper      27.000000 1450.000000    1.862069   100.000000
```

Role of variables

- **Response**

  Variables that we want to study, by building a model, finding associations, … (number of products bought, passing or failing a course, income, …)

  It can be either continuous or categorical

- **Explanatory**

  Variables which serve to explain the behaviour of the response variables (all the variables present in the data matrix except the response)

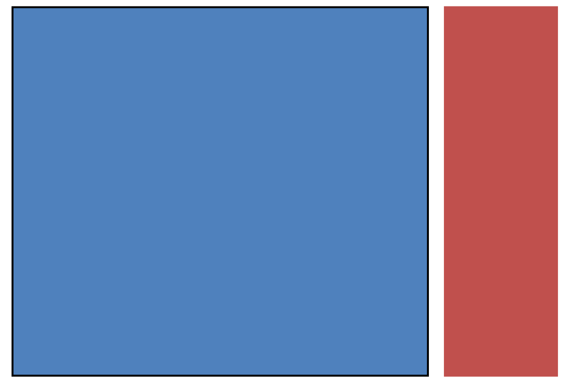
  They can be either continuous or categorical

Types or data matrix

With or without response(s) variable

i.e. transactions data

Inputs Output(s)



Data to explore, to describe, to find associations (i.e. itemsets), …

Idem, but we want to **find a model to predict the response**

Paradigm

Any stored data from any process always contain information about the generating phenomenon (**statistical regularity**).

Goal: **To reveal the information** (model, patterns, associations, trends, clusters, … hidden in the data

Data are routinely stored (and most will never be analyzed)

Data is a treasure for organizations (be aware of the data quality)

Any transactional process con be enhanced by analysis of its collected data

How? *Selecting and reporting what is interesting*

SQL queries are NOT ENOUGH. How many A products sold last month?.

Profiling. What is the profile of A buyers? *Automatic detection of significant deviations*

Automatic profiling of groups of individuals

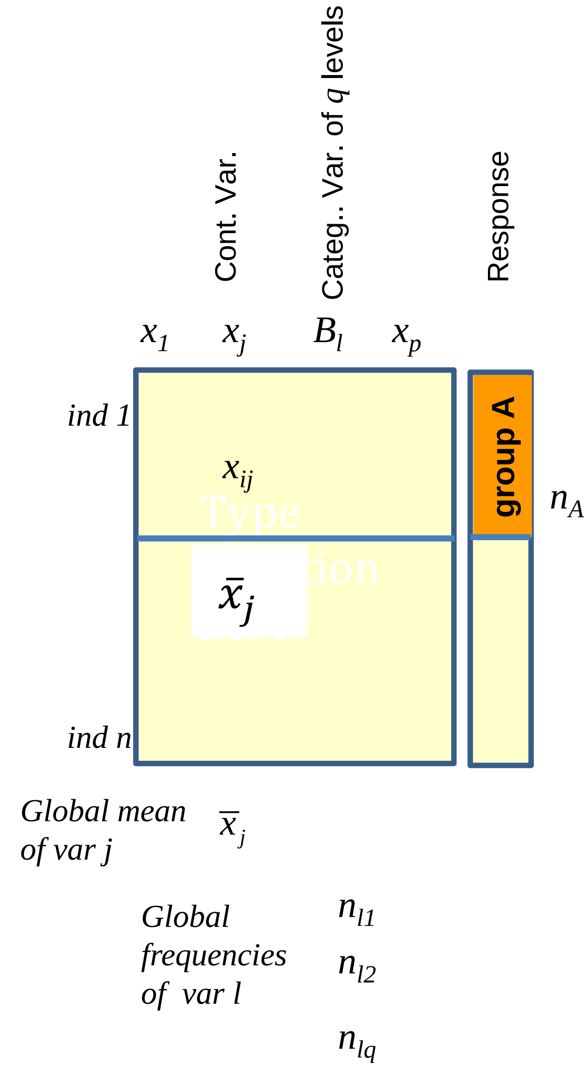
We have a group of individuals defined by a level of a categorical variable (target).

Problem: For every group of individuals detect which other groups of individuals (identified by the levels of the explanatory variables) or what continues variables, deviate significantly from what were expected.

- We take as response variable the variable identifying the groups that we want to find their profile.
- The explanatory variables are either categorical or continuous.

Tool: Hypothesis test

- For each group to profile, rank the modalities of the categorical explanatory variables according their p-value (ascending). Likewise, rank the continuous variables according their p-value
- Select the most significant by a threshold (0.05, 0.01, ..) defined a priori. (what matters is the ordering, actual significance depends on the number of individuals)



R function available in FactoMineR

We will use FactoMineR Package (cran R)

You can also consult (and download this R function from) http://factominer.free.fr/ where a large documentation is provided, with theoretical background, examples, tutorials and so on.

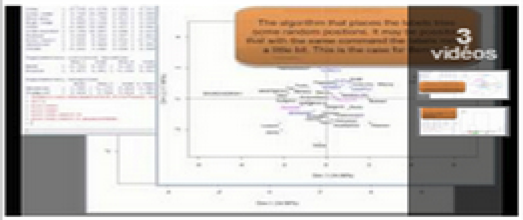
The functions of this package corresponding to this sessions are:

- **Catdes:** description of the categories of a categorical variable by quantitative variables, categorical variables and categories
- **Condes:** description of a quantitative variable by quantitative and categorical variables

Home | Classical Methods | Advanced Methods | Interface | Facto's best | Excel | F.A.Q. | Docs | Contact

FACTOMINER

News bulletin



Exploratory multivariate analysis with R and FactoMineR

Videos on the use of FactoMineR (for PCA, multiple factor analysis, clustering, etc.)

The version 1.24 of FactoMineR has a new graphical module that place the labels in an "optimal" way, that allows to select some elements to draw, etc.

Four reviews on the book Exploratory Multivariate Analysis by Example using R are available in this site. To see the complete review done by Gary Evans (for Journal of Statistical Software)

A new useR group to ask questions on FactoMineR and on Exploratory Multivariate Data Analysis has been created. Join this group to have news about FactoMineR and to ask questions

missMDA: a new package to handle missing values in PCA, MCA or MFA with FactoMineR

English Version

Version française

Top Menu

- Home
- Classical Methods
- Advanced Methods
- Interface
- Facto's best
- FactoMineR and Excel
- F.A.Q.
- Documents
- Contact

Useful Links

- Agrocampus Rennes Applied Maths Department
- R Project
- CRAN

Elements of a Hypothesis test

Tool to validate a hypothesis

Hypothesis: Group k of individuals is different from the population

To test a hypothesis we need:

Null hypothesis H_0: denial of the hypothesis (like an evil advocate). Group k is equal to the population

Alternative hypothesis H_1: the hypothesis we want to validate

Test statistic: its depends on the problem.

Reference distribution: Distribution of the *test statistic* if the H_0 is true.

Significance threshold: Risk that we are ready to incur to reject H_0 when it is true (significance depend on the number of individuals, thus classical statistical thresholds need to be adapted).

In our case:

$$H_0: \quad \text{Individuals of group k are taken at random}$$
$$H_1: \quad \text{Individuals of group k are NOT taken at random}$$

General rationale of any test

Test

We want to see if problematic posed by the user

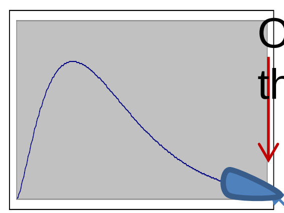
Null hypothesis **H_0**: conservative hypothesis.

Alternative hypothesis **H_1**: hypothesis in which the user is interested in

Data *data are observed, generally a sample*

Test statistic: a statistic is a function of the sample (=observed data) and thus
a variate or random variable

Reference distribution: Distribution of the *test statistic* under H_0 (that is, if H_0 is true).



Observed value, as computed on the data

Significance threshold: Risk of rejecting H_0 although H_0 being true

≤ 0.05 H_0 is rejected

> 0.05 H_0 is not rejected

This area= *p*-value

12

Elements of a Hypothesis Test

The test statistic depends on the level of measurement of variables

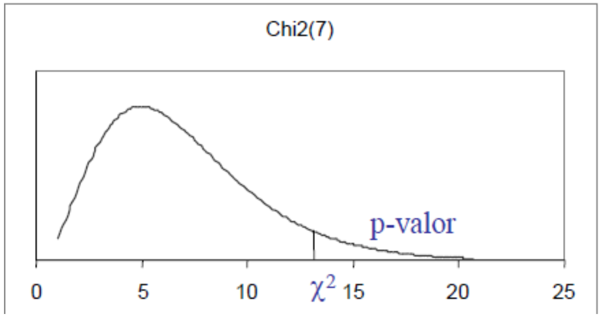
Null hypothesis

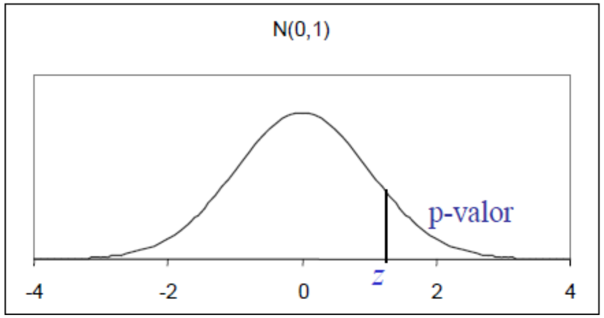
- For continuous variables, the *mean* in group should be equal to the global mean
- For categorical variables: the *proportion* in group *k* should be equal to the global proportion

Compute per each test, *the p-value* of the test statistic

p-value: probability to find a *test statistic* equal or farthest that the actual test statistic.
It is the probability that if the H_0 is true, of actually finding the observed test statistic

.





Finally rank the *p-values* in ascending order, and select the significant characteristics from those *p-values* lower than the chosen significance threshold

Profiling a categorical target from a continuous variable

groups	*means*	*counts*
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
p	$\bar{x}_p$	n_p
Global	$\bar{x}$	n

Ronald Fisher 1890, 1962

$$H_0\colon \mu_1 = \cdots = \mu_p = \mu$$

Null Hypothesis: All group means are equal to the global mean

In R:

- Assuming normal distribution on X: oneway.test(X~A).
- Without normality assumption (non –parametric test): Kruskal-Wallis test
  kruskal.test(X~A)

Profiling target categorical variables from continuous variables

groups	means	counts
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
p	$\bar{x}_p$	n_p
Global	$\bar{x}$	n

$$H_0 : \mu_k = \mu \quad k = 1, \ldots, p$$

Test statistic: Difference between the mean in group *k* and the global mean

Are there any groups with a significant different mean?

$$t = \frac{\bar{x}_k - \bar{x}}{\sqrt{(1 - \frac{n_k}{n}) \frac{s^2}{n_k}}} \sim t_{n-1}$$

Student's *t*

William Gosset "Student", English, 1876, 1937

Rank the continuous variables by p.value (ascending)

Function to compute p-values for profiling a categorical target from continuous variables

To Rank variables and groups according to pvalues:

```
p.xk <- function(vec,fac)  {
        nk <- as.vector(table(fac));
        n <- sum(nk);
        xk <- tapply(vec,fac,mean);
        txk <- (xk-mean(vec))/(sd(vec)*sqrt((n-
nk)/(n*nk)));
        pxk <- pt(txk,n-1,lower.tail=F)}
```

Rank the continuous variables by p.value (ascending)

FactoMineR solution:

- catdes(data.frame,num.var): sections
  - Link between the cluster variable and the quantitative variables
  - Description of each cluster by quantitative variables

Profiling categories from categorical variables

$$
\begin{array}{cc}
 & 1\ldots \quad j \quad \ldots q \\
\begin{array}{c} 1 \\ k \\ p \end{array} & \boxed{\begin{array}{ccc} & \vdots & \\ \cdots & n_{kj} & \cdots \\ & \vdots & \end{array}} \; n_k \\
 & n_j
\end{array}
$$

– **Global Relationship between each category** of the target variable and other categorical variables: a chi-square-test is performed

– **Relationship between each category** of the variable target and each category of another categorical variable: comparison of two proportions, taking into account an hypergeometric model

–Descriptive tools: contingency tables (numeric) and mosaic plot (graphical)

Rank the levels of the categorical explanatory variables/ the categories by p-value (ascending)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

FIB UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

(Global) Characterization of a target categorical variable by the other categorical variables

Test

Null hypothesis — **H_0**: conservative hypothesis. Both variables are independent

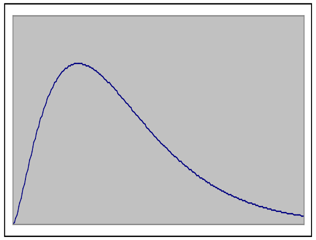
Alternative hypothesis — **H_1**: Both variables are not independent

Test statistic:

$$\chi^2_{obs} = \sum_i \sum_j \frac{\left(n_{ij} - {n_{i.}n_{.j}}/{n}\right)^2}{{n_{i.}n_{.j}}/{n}} = \sum_i \sum_j \frac{\left(n_{ij} - np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}}$$

Reference distribution: Distribution of the *test statistic* under H_0 (that is, if H_0 is true).

Chi-2 distribution, with the convenient degrees of freedom



Significance threshold: Risk of rejecting H_0 although H_0 being true

(significance depends on the number of individuals) P-value

Characterization of a target categorical variable by the levels of other categorical variables

$$
\begin{array}{c|ccccc|c}
 & 1\ldots & & j & & \ldots q & \\
\hline
1 & & & \vdots & & & \\
k & \cdots & & n_{kj} & & \cdots & n_k \\
p & & & \vdots & & & \\
\hline
 & & & n_j & & &
\end{array}
$$

$$H_0 : p_{j \cdot k} = p_j \quad k = 1, \ldots, p;\ j = 1, \ldots, q$$

Assumption of normality of proportions:

$$\frac{n_{kj}}{n_k} \sim N\left(p_j = \frac{n_j}{n}, \left(1 - \frac{n_k}{n}\right) \frac{p_j (1 - p_j)}{n_k} \right)$$

Test statistic: Difference between proportion of modality *j* in group *k* and proportion of modality *j* in whole data

$$z = \frac{\frac{n_{kj}}{n_k} - \frac{n_j}{n}}{\sqrt{\left(1 - \frac{n_k}{n}\right)\left(\frac{p_j (1 - p_j)}{n_k}\right)}} \sim N(0,1)$$

Rank the levels of the categorical explanatory variables by p.value (ascending)

R function to compute the table of p-values of a categorical variable

```
p.zkj <- function(res,expl){
   taula <- table(res,expl)
   n <- sum(taula);
   pk <- apply(taula,1,sum)/n;
   pj <- apply(taula,2,sum)/n;
   pf <- taula/(n*pk);
   pjm <- matrix(data=pj,nrow=nrow(pf),ncol=ncol(pf), byrow=T);
   dpf <- pf - pjm;
   dvt <- sqrt(((1-pk)/(n*pk))%*%t(pj*(1-pj)));
   zkj <- dpf/dvt;
   pzkj <- pnorm(zkj,lower.tail=F);
list(rowpf=pf,vtest=zkj,pval=pzkj)}
```

FactoMineR solution:

- catdes(data.frame,num.var)
  - Link between the cluster variable and the categorical variables (chi-square test)
  - Description of each cluster by categories

Example: SwissLabor data in AER library

Usage

data("SwissLabor")

Format

A data frame containing 872 observations on 7 variables.

```
levels(SwissLabor$participation)<-
       paste("Parti.",sep="",levels(SwissLabor$participation))
levels(SwissLabor$foreign)<-
       paste("Foreign.",sep="",levels(SwissLabor$foreign))
```

participation	Factor. Did the individual participate in the labor force?
income	Logarithm of nonlabor income.
age	Age in decades (years divided by 10).
education	Years of formal education.
youngkids	Number of young children (under 7 years of age).
oldkids	Number of older children (over 7 years of age).
foreign	Factor. Is the individual a foreigner (i.e., not Swiss)?

Profiling a categorical target by the categories of the other categorical variables

In SwissLabor dataset in library(AER): Participation –Yes (target) vs Foreign

H_0: The category “foreign=NO” is neither infra nor supra represented

H_1: The category “foreign=NO” is infra (versus supra) represented

```
>table(SwissLabor$foreign,SwissLabor$participation)
               Target-no Target-yes
Foreign-no     402     254
Foreign-yes    69      147
➢prop.table(table(SwissLabor$foreign,SwissLabor$participation),1)
               Target-no Target-yes
Foreign-no     0.6128049 0.3871951
Foreign-yes    0.3194444 0.6805556
>
prop.table(table(SwissLabor$foreign,SwissLabor$participation),2)
               Target-no Target-yes
Foreign-no     0.8535032 0.6334165
Foreign-yes    0.1464968 0.3665835
```

```
round(prop.table
(table(SwissLabor$foreign)),dig=2)
Foreign.no Foreign.yes
    0.75        0.25
```

Characterization of a categorical variable by the categories of the other categorical variables

In SwissLabor dataset in library(AER): Participation –Yes (target) vs Foreign

H_0: The category "foreign=NO" is neither infra nor supra represented

H_1: The category "foreign=NO" is infra (versus supra) represented

68% of class Foreign-Yes belongs to Category Target-Yes

36.7% of Category Target-Yes belongs to class Foreign-Yes

```
$category$`Target-yes`
                        Cla/Mod  Mod/Cla   Global      p.value   v.test
foreign=Foreign-yes 68.05556 36.65835 24.77064 5.591005e-14  7.517321
foreign=Foreign-no  38.71951 63.34165 75.22936 5.591005e-14 -7.517321
```

```
prop.table(table(SwissLabor$foreign))
Foreign-no Foreign-yes
0.7522936  0.2477064
```

Foreign women represent 25% of the sample, but 36.7% in the target class Target-Yes

FIB UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Profiling a quantitative target from quantitative or categorical variables

– **Description by quantitative variables (condes) : global association**

Correlation (Pearson)-> cor(*data.frame*)

– **Description by categorical variables and categories**

ANOVA: test F (global association) and t-tests

Relationship between a quantitative target and the other quantitative variables

H$_0$) no relationship (correlation is null ρ=0)
H$_1$) relationship (correlation non null $\rho \neq 0$)

Statistics:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson linear correlation

```
> condes(SwissLabor,2) #Numeric target income
$quanti
            correlation      p.value
education    0.3273458 3.166132e-23
oldkids      0.1391036 3.758541e-05
```


28

Global Relationship between the quantitative target "income" and the categorical variables

income (foreign group *k*)= mean + effect of foreign group *k* + error

$$Y_{ki} = \mu + \alpha_k + \varepsilon_{ki}$$

H_0 (no category effect): $\alpha_1 = \ldots = \alpha_k = \ldots = \alpha_K = 0$
H_1: There are at least two "factor" levels *k* and *k'* such as: $\alpha_k \neq \alpha_{k'}$

```
> condes(SwissLabor,2) #Numeric target income
…

$quali
                      R2      p.value
foreign       0.04389655 4.170824e-10
participation 0.02989118 2.794460e-07
```


29

Relationship between the quantitative variable "income" and the levels of categorical variables

H_0 The coefficient of category *k* is null $\alpha_k = 0$

H_1: The coefficient of category *k* is non-null $\alpha_k \neq 0$

```
> condes(SwissLabor,2) #Numeric target income

$category
                 Estimate      p.value
Foreign.no    0.10004281 4.170824e-10
Parti.no      0.07150532 2.794460e-07
Parti.yes    -0.07150532 2.794460e-07
Foreign.yes  -0.10004281 4.170824e-10
```

Example: SwissLabor data in AER library

Usage

data("SwissLabor")

Format

A data frame containing 872 observations on 7 variables.

participation	Factor. Did the individual participate in the labor force?
income	Logarithm of nonlabor income.
age	Age in decades (years divided by 10).
education	Years of formal education.
youngkids	Number of young children (under 7 years of age).
oldkids	Number of older children (over 7 years of age).
foreign	Factor. Is the individual a foreigner (i.e., not Swiss)?

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

FIB UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Example: SwissLabor data in AER library

```
> condes(SwissLabor,2) #Numeric target income
$quanti
          correlation      p.value
education   0.3273458 3.166132e-23
oldkids     0.1391036 3.758541e-05

$quali
                      R2      p.value
foreign       0.04389655 4.170824e-10
participation 0.02989118 2.794460e-07

$category
                Estimate      p.value
Foreign.no    0.10004281 4.170824e-10
Parti.no      0.07150532 2.794460e-07
Parti.yes    -0.07150532 2.794460e-07
Foreign.yes  -0.10004281 4.170824e-10
```

Global association

Global association

Profiling on categories