

Preprocessing

Data Mining course
Master in Information Technologies
Enginyeria Informàtica



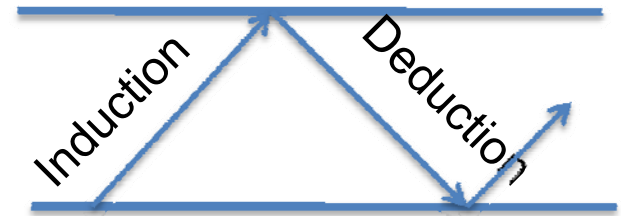
UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Why we need data?

Learning is an iteration process between the real world facts and the hypothesized world of theories

theories, hypothesis

observed data, facts



DM concerns the inductive phase of learning

Data = Fit + Noise

Data represented in flat files



Rows of a data file

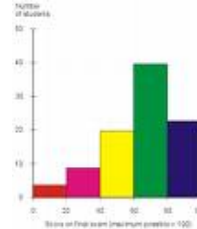
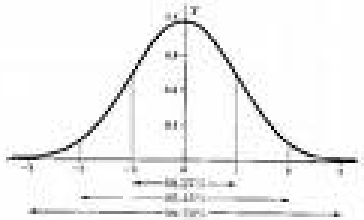
- Represent individuals or instances
 - In DM they can be from 60 to millions, ...

Also called sample, example, record, ...

- Unit that 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
- We use all available data. Sampling betrays the spirit of DM, where few individuals may keep the most precious information and only can be justified for alleviating the computational cost

Columns of the data file

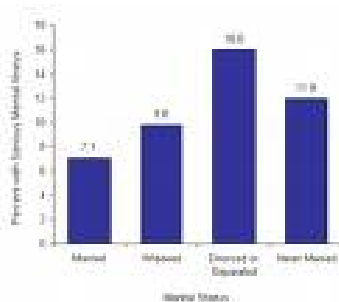
- Each instance is described by a predefined set of features, its variables or “attributes”
- A variable is a measure of individuals which can take different values (according a probabilistic function)



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

Nominal variables

- Values are distinct categories represented by symbols
 - Values themselves serve only as labels or names
- Example: attribute “marital status”
 - Values: “single”, “married”, “divorced” and “widow”
- No relation is implied among nominal values (no ordering or distance measure)
- Only percentages and tables can be calculated
- Bar plots (among many other plots) for graphical representation



Special case: Binary data
(Yes/No) (Boolean attribute)

Ordinal variables

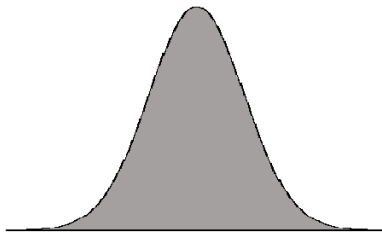
- Impose order on values, but: no distance between values defined
- Example:
attribute “size” in clothe’s size
 - Values: “XXL” > “XL” > “L” > “M” > “S”Or social status
 - Values: “upper class” > “middle/high” > “middle” > “middle/lower” > “lower class”
- Arithmetic calculations not possible. Distinction between nominal and ordinal not always clear
- Tables and percentages and bar plots, but emphasizing the order of values

Interval variables

- Interval quantities are not only ordered but measured in predefined units
- Example : “temperature” expressed in degrees Fahrenheit or Celsius.
 - Zero point is not defined
- Example: “year”
- Difference of two values makes sense, but not the quotient.
- Usually modeled by the Laplace-Gauss distribution
- Means and standard deviations. Dot plots

Ratio variables

- Ratio quantities are ones for which the measurement scheme defines a zero point
- Example: attribute “distance”
 - Distance between an object and itself is zero
- Examples: “age”, “speed”, ...
- Ratio quantities are treated as real numbers. Usually modeled by the Laplace-Gauss distribution.



Karl Gauss, 1777-1855
great German mathematician



Pierre Simon, marquis de Laplace 1749-1827
great French mathematician

- Means and standard deviations, ... dot plots, ...
- But: is there an “inherently” defined zero point? In practice difficult to distinguish from interval data.

Count (or discrete) data

- Very often a variable is a result of a count
- Examples: attribute “number of words of a sentence”, “number of students of a course”, “number of products bought”, “number of bugs per program”, “number of unemployed per country”, ...
- Usually modeled by the Poisson distribution.
- Means and standard deviations, but percentages as well.

Attribute types used in practice

- **Categorical** (factor in R and the categories are called levels)
 - Binary
 - Nominal
 - Ordinal
- **Continuous or numeric** (variate in R)
 - Count data
 - Interval
 - Ratio data
- But distinctions are often blurred. Ordinal data can be treated as continuous, ...
- DM methods may only admit one type of variables (Knn continuous, Naïve Bayes categorical, ...)

Understanding the data

- Data origin: Secondary source ..., what?. Representative sample?)
- Data quality: Outliers, missing data, errors (produce a report on it)
- Size of the data set:
 - Number of records (individuals) Sampling (balanced)
 - Number of fields (variables). *Rule of thumb: for each variable, 10 or more instances*
- Variables have meaning
 - Normally expressed in the metadata file
- **Role of variables**
 - **Response** , output or target. They are the variables we want to study, model, predict, ...
 - **Explanatory**, input or predictors. They are the variables used to predict the former.

Two types of data files

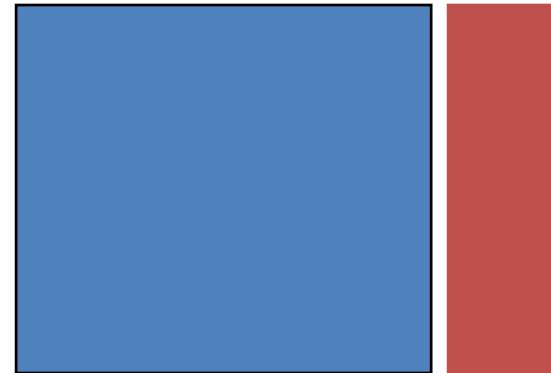
With or without response(s) variable

i.e. transactions data



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

Inputs Output(s)



Idem, but data to **find a model to predict the response**

Preprocessing Steps

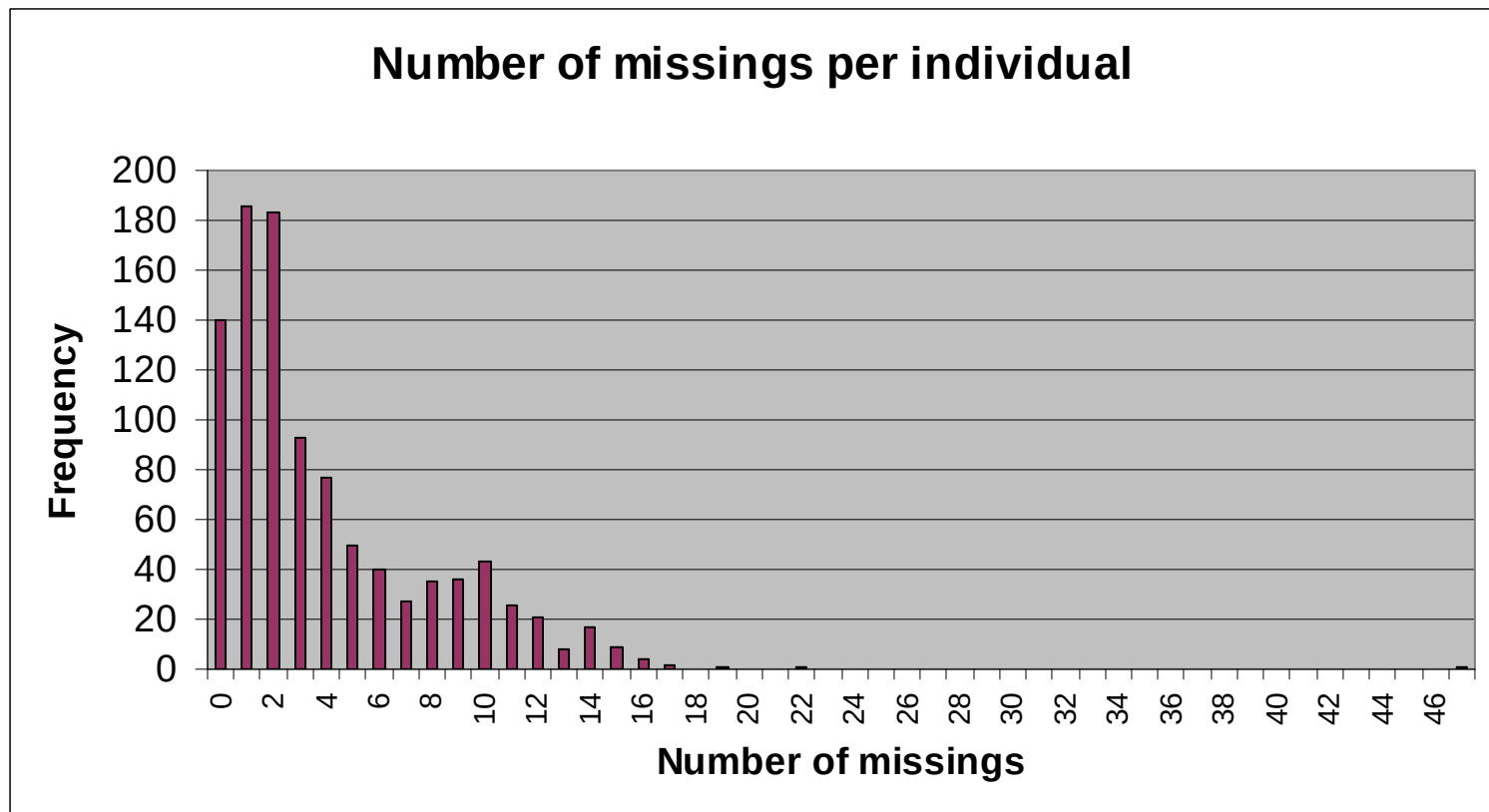
- Data Cleaning
 - Errors
 - Outliers
 - Missing values
- Preparing the data for the analysis
 - Feature selection
 - Feature extraction
 - Transformations
 - Recoding
 - Quantifying
 - Normalizing

Outliers and Errors

- Errors
 - Detect them and correct them
- Outliers are values thought to be out of range.
They can bias the results
Approaches (difficult):
 - do nothing
 - treat it as missing data

Missing Values

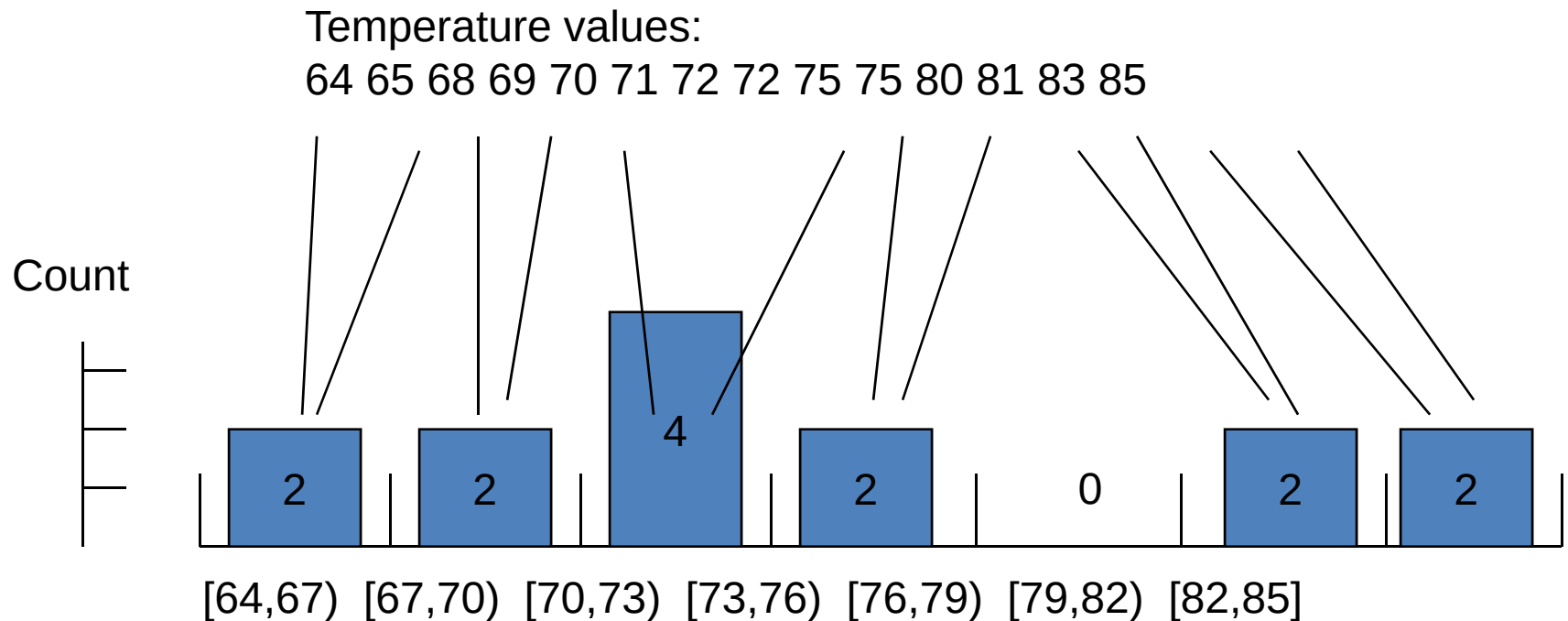
- Number of missing values are an indicator of the data quality
- Missing data can appear in several forms:
 - <empty field> “0” “.” “99999” “NA” ...
- Standardize missing value code(s)
- Exploring the missingness pattern
 - Create a new variable counting the number of missing per individual. Describe this variable
 - Count the number of missing per variable and rank them accordingly, the greater the less reliable is the variable.
 - Create a new variable per variable indicating the (missing/non missing) and compare both groups of cases.
- Dealing with missing values:
 - ignore records with missing values
 - If categorical, treat missing value as a separate value.
 - If continuous, fill in with mean or median values, 1nn, EM algorithm, DA algorithm, ...



DESCRIPTION OF : number of -- no answer --
 DESCRIPTION BY CATEGORIES
 OF CONTINUOUS VARIABLE : number of -- no answer --
 ON 1000.0 ACTIVE CASES MEAN = 4.054
 STD. DEV. = 4.190

TEST PROB. VALUE	MEAN	STD. DEV.	CATEGORIES	VARIABLE LABEL	WEIGHT
7.71 0.000	6.32	5.77	65 years and over	age of the respondent ?	169.00
6.86 0.000	6.32	5.55	doesn't know	as a method of child care,day care centers are...	139.00
5.93 0.000	5.49	5.27	unbreakable union	opinion about marriage	231.00
5.86 0.000	7.00	4.50	doesn't know	opinion about justice...	65.00
5.46 0.000	6.46	4.29	less than 2000	size of town (number of inhabitants)	83.00
5.28 0.000	5.95	5.82	never	do you occasionally invite friends for lunch or dinner ?	120.00
5.28 0.000	5.50	4.19	none	what is the highest level of education attained ?	189.00
5.17 0.000	6.54	6.96	doesn't know	does society need a change	71.00
5.08 0.000	5.11	4.92	owner	do you rent or own your lodging ?	290.00
5.03 0.000	4.64	4.19	yes	the family is the only place where we feel good...	561.00
5.01 0.000	6.92	7.27	none	do you think that the ideal number of children is ?	51.00
5.01 0.000	7.90	8.54	doesn't know	as a method of child care,housewives are...	29.00
4.91 0.000	4.96	4.80	woman more than 38 y	age and sex of interviewer ?	338.00
4.76 0.000	4.66	4.38	very much	do you like the view from your window ?	516.00
4.67 0.000	6.26	6.25	9 p.m. or earlier	generally, at what time do you go to bed ?	73.00
4.20 0.000	4.70	4.77	no answer	does your work entail risks to your health ?	424.00
4.02 0.000	4.26	4.29	no	do you share in the protection of the environment ?	874.00
3.91 0.000	4.65	4.74	no answer	does your work life sometimes clash with your personal life	433.00
3.87 0.000	6.87	4.14	farmer	respondent occupation ?	32.00
3.72 0.000	4.75	5.05	preferably the wife	who does housework and childcare ?	336.00
3.53 0.000	6.76	5.03	can't say	for society to change, what is required ?	29.00
3.22 0.001	5.79	4.76	can't say	is the change in the standard of living last years	57.00
3.18 0.001	6.86	4.57	doesn't know	is money spent on home ...	22.00
3.01 0.001	4.80	5.19	no answer	for society to change, what is required ?	223.00
3.00 0.001	4.26	4.32	no	do you own a dish-washer ?	789.00
2.96 0.002	5.10	4.57	not very/not at all	was the respondent interested ?	124.00
2.67 0.004	4.60	4.82	important	the preservation of environment is...	298.00
2.56 0.005	4.61	4.17	sometimes	do you occasionally invite friends for lunch or dinner ?	274.00
2.55 0.005	4.64	4.31	the same	change in respondent standard of living/last years	250.00
2.53 0.006	4.75	4.49	50 - 64 years	age of the respondent ?	188.00
2.45 0.007	6.08	4.22	refused	opinion about justice...	25.00

Transforming a continuous variable to an ordinal with Equal-Width



Equal Width, bins Low $\leq$ value $<$ High

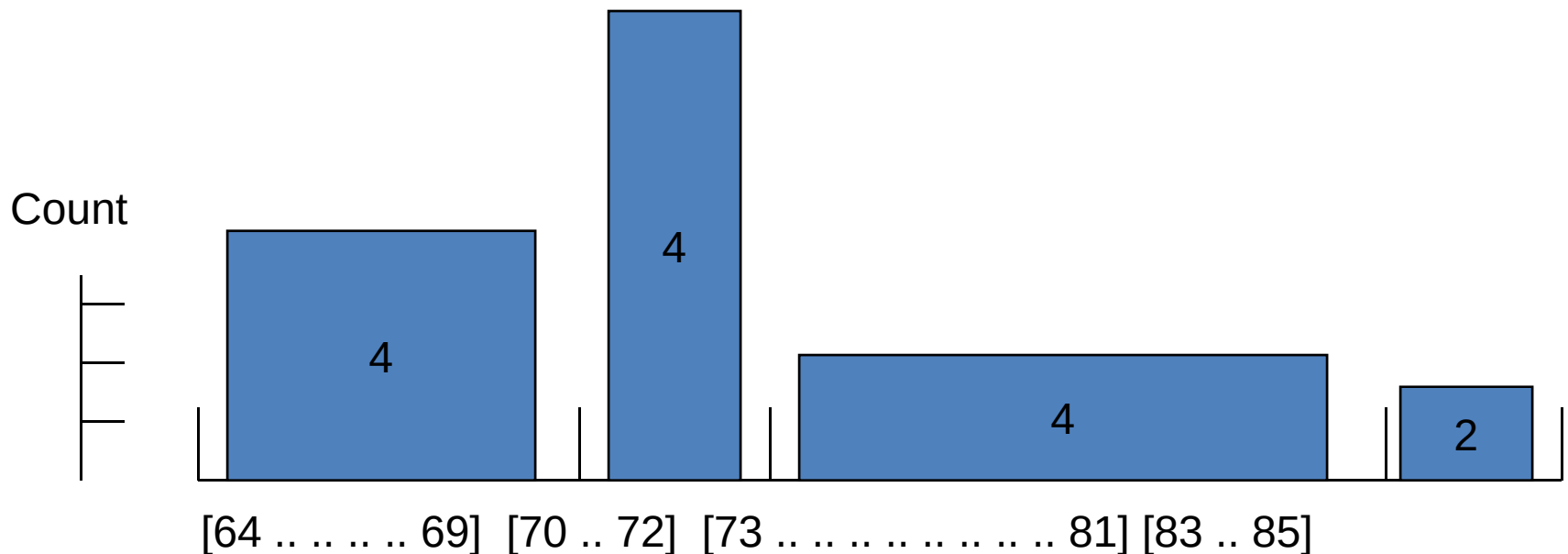
Equal-Width may produce clumping



Equal-Count

Temperature values:

64 65 68 69 70 71 72 72 75 75 80 81 83 85



Equal Count = 4, except for the last bin

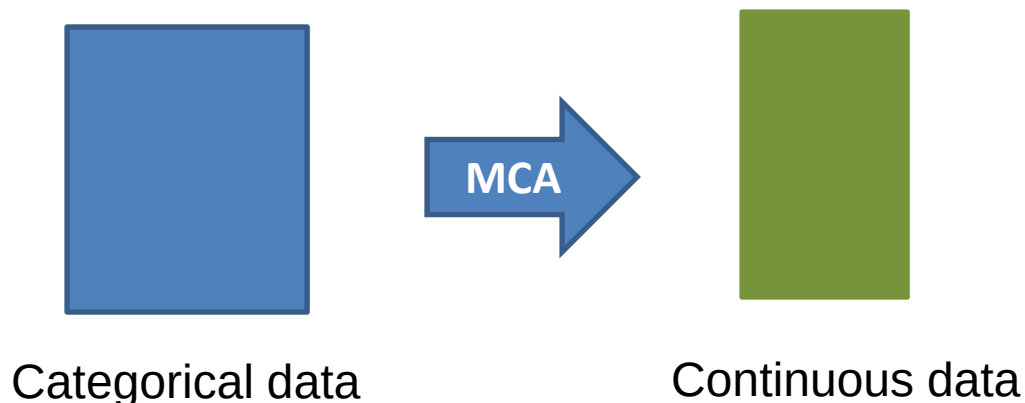
Equal count implies equal area per bin but not equal height

Discretization depending on the response

- Run a Decision tree for each predictor
- May be used for feature selection

Quantification

- Passing from a categorical variable to a numeric equivalent
- If binary assign a 0/1 code
- If ordinal assign the middle class value (problematic for extreme bins)
- If nominal
 - Quantification depending of the response: Use a decision tree
 - Multiple quantification (several nominal vars.): Use Multiple Correspondence Analysis



Feature selection

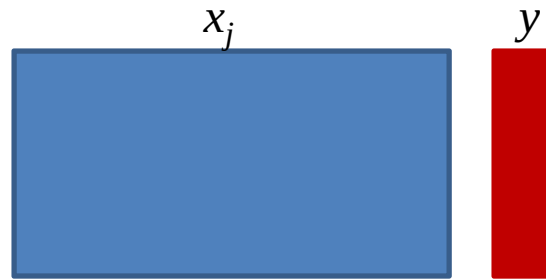
Goal: Discard non interesting variables

- It improves Classification. Most learning algorithms look for non-linear combinations of fields -- can easily find many spurious combinations given small # of records and large # of fields
- Try to avoid the curse of dimensionality
- Rank variables according their predictive power of the response variable
- Pre or post model selection



Feature Selection

- Let be a data file with y response and input vector $\mathbf{x} = (x_1, x_2, \dots, x_p)$ with large p .

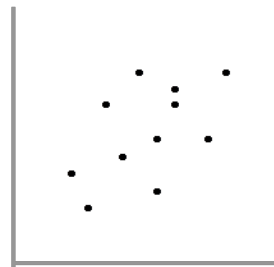


- For a given response, rank the explanatory variables x_j according to their dependence with the response y
- But, how can we establish the dependence between two variables (x_j and y)

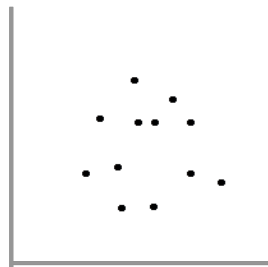
Dependence between continuous variables

Correlation coefficient

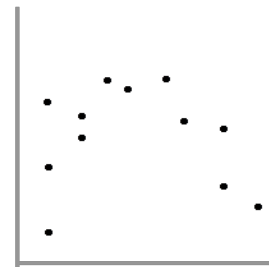
$$r_{x_j y} = \frac{\sum_i (x_{ij} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_{ij} - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{s_{x_j y}}{s_x s_y}$$



correlación 0,6



correlación 0



correlación 0
(relación cuadrática)

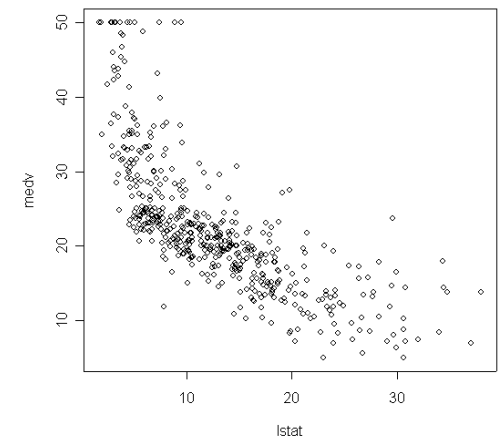


correlación -0,9

Asymptotical threshold for non significant correlation
(with 1000 individuals a correlation greater than 0.06 is already significant)

$$\frac{1.96}{\sqrt{n-1}}$$

But, it only detects linear dependencies.
Transform to ordinal to detect non linearities.



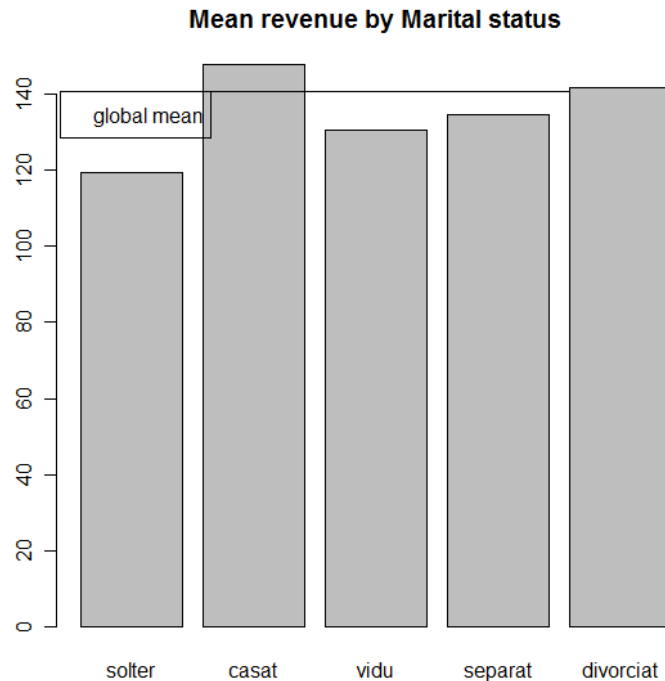
Dependence between two categorical variables

- Set the cross table

Estat_civil							Row %
Vivenda	solter	casat	vidu	separat	divorciat	Total	
lloguer	174	723	11	50	15	973	21.9%
escriptura	167	1839	50	38	12	2106	47.4%
contr_privat	26	212	3	4	1	246	5.5%
ignora_cont	1	18	0	0	1	20	0.4%
pares	507	238	0	30	7	782	17.6%
altres viv	98	208	3	8	2	319	7.2%
Total	973	3238	67	130	38	4446	
Columns %	21.9%	72.8%	1.5%	2.9%	0.9%		

If Marital Status and Housing were independent, how many single do you expect would live with their parents? and married?

Dependence between one continuous variables and one categorical



If Revenue and Marital Status were independent, what would you expect be the mean revenue of singles? and married?

How to formalize the dependence?

- **Tool: HYPOTHESIS TEST**

- Try to confirm the hypothesis that there is no relation between x_j and y

H_0 : There is no relation between the y and x_j

H_1 : There is a relation

- The type of hypothesis test depends on the level of measurement of variables
- Finally rank the variables x_j according how strongly they have refuted the null hypothesis (lower p-value). The important is the ordering, Significance depend on the number of individuals, thus classical statistical thresholds need to be adapted.

Test for the dependence between one categorical variable and one continuous.

The Fisher's F

levels	means	counts
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
q	$\bar{x}_q$	n_q

$$H_0 : \mu_1 = \dots = \mu_q = \mu$$

$\bar{x}, s^2$ global mean and global variance

$$F = \frac{S_B^2 / (q-1)}{S_W^2 / (n-q)} \sim F_{q-1, n-q}$$

$$S_W^2 = \sum_{k=1}^q \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2$$

$$S_B^2 = \sum_{k=1}^q n_k (\bar{x}_k - \bar{x})^2$$

Ranking of variables by p.values (ascending)

R function: `oneway.test(continuous~categorical, data, subset, na.action, var.equal = FALSE)`

Test for the dependence between two categorical variables: The χ^2

	1... j ...q	
1	$\vdots$	
k	$\cdots n_{kj} \cdots$	n_k
p	$\vdots$	
	n_j	n

$$H_0 : n_{kj} = np_k p_j \quad \forall kj$$

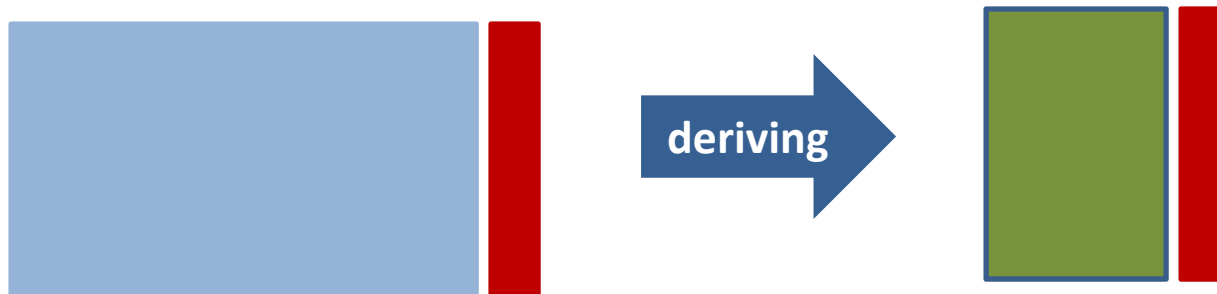
$$X^2 = \sum_{k=1}^p \sum_{j=1}^q \frac{(n_{kj} - \frac{n_k n_j}{n})^2}{\frac{n_k n_j}{n}} \sim \chi^2_{(p-1)(q-1)}$$

Rank the variables according to the p-values (ascending)

R function: `chisq.test(response, explanatory)`

Feature extraction

- Better to have a fair modeling method and good variables, than to have the best modeling method and poor variables.
- Models never are true, but we just need to be useful
- Implies to work with (or to be a) domain expert to define new sounded derived variables (i.e. ratio height/width to classify manuscript characters from pixelated images)
- Reduction of dimensionality (PCA, MCA, CCA, PLS, ..)



Data Base description

- Data are routinely stored (and most will never be analyzed)
- But stored data from any process always contain information about the generating phenomenon (statistical regularity).

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

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

- How? Selecting and reporting what is interesting
- What does it mean?
 - **SQL queries NOT ENOUGH.** How many A products sold last month?.
 - **Profiling.** What is the profile of A buyers? *Automatic detection of significant deviations*

First insight into data

- Look at meta-data
- Summary of data to inspect for anomalies and errors, missing data, outliers.
- First report data quality.
- Preprocess of data
- Plan for verification - verify the results after each procedure

Automatic profiling of groups of individuals

- For every group of individuals, identified by a level of the response variable, detect which other groups of individuals (identified by the levels of the explanatory variables) or what continuous variables, **deviate** significantly from what were expected.
(take the variable identifying the groups that we want to find their profile as response, and the variables serving to define the profile as explanatory)
- Tool: Hypothesis test
For each group to profile, rank the modalities of the categorical explicative variables according to their p-value (ascending).
Likewise, rank the continuous variables according to their p-value and select the most significant.
What matters is the ordering, actual significance depends on the number of individuals

Dependence of one level of a categorical variable (group of individuals) with a continuous variable

$$H_0 : \mu_k = \mu \quad k = 1, \dots, q$$

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

Student's t

Rank the continuous variables by p.value (ascending)

Dependence of one level of a categorical variable (group of individuals) with another level of a categorical explanatory variable

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

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

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

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