O.01 Indústria 4.0 y Gestión de Datos

Karina Gibert;

Department of Statistics and Operations Research
Knowledge Engineering and Machine Learning group
at
Intelligent Data Science and Artificial Intelligence Specific Research Center

University Institute of Research on Science and Technology of Sustainability

Universitat Politècnica de Catalunya-BarcelonaTech

karina.gibert@upc.edu
https://www.eio.upc.edu/en/homepages/karina

ÍNDICE

- Datos y Metadatos
- Primeros passos de preprocessing: La lectura de los datos
- Preliminares de R
- Actividad práctica: Reading and declaring factors
- Anàlisis descriptivo (automatic reporting)
- Preprocessing
  - Metodología
  - Selección de la matriz de trabajo
  - Datos Faltantes
  - Datos aberrantes
  - Transformaciones
  - Variables nuevas
- https://www-eio.upc.edu/~karina/datamining>MasterIndus

  Username: MasterIndus

Data, Metadata

Basic structure for analysis

The data matrix

	Weight	Height	Sex	Eyes
John	85	1.85	M	azul
	.	.	.	.
	.	.	.	.

Point cloud (video)

Rows: Individuals (study units) (i1….in)

Columns: Variables (characteristics of individuals) (X1..Xk)

Cells: Value of variables for individuals (xik)

https://www.youtube.com/watch?v=4pnQd6jnCWk

Type of variables

- Numerical: Quantitative, measure
  - continuous (real quantity): *...ght*
  - discrete (natural quantity): *..., ...es size*
- Categoric: Qualitative, adjective
  (evenctually codified)
  - Ordinal (ordering over modalities): *S...ioeconomic status*
  - Binary (two modalities): *...wear glasses*
  - Nominal (unordered modalitites): *Hair color*
- Date: Special formats, only some softwa...
- Other variables
  (no standar... rarely used in standard data mining applications)
  - Fuz... ...ables
  - ...ariables
  - ...riables
  - ...ariables
  - D...tional variables
  - Interval variables/Ratio variables (means, standard dev, dotplots)
  - Textual data

Categorization
Discretization

Mean/StDev
Histograms

RECategorization

Percentages
Tables
BarPlots

Loss information

Better avoid

From Data to Decisional Knowledge

DATA <> **INFORMATION**

prepostdif.DAT - WordPad

Home View

Cut Copy Paste

Courier New 11

Find Replace Select all

Picture Paint drawing Date and time Insert object

Clipboard Font Paragraph Insert Editing

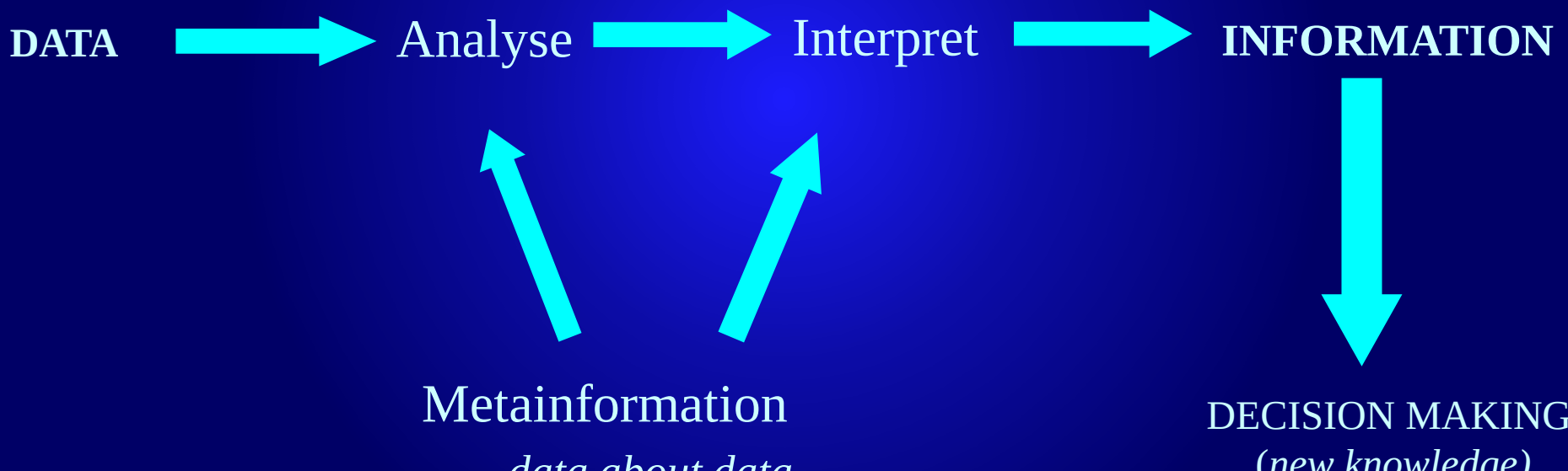
```
((0    5     5     300   300   0     0     35    35    0     2     2
       0     0     0     0     1     1     ?     ?     ?     500
       500   0     -25   -4    21    0     0     0     50    50    0
       36    0     -36))
((6    6     0     300   300   0     0     0     0     3     3     0
       0     38    38    0     0     0     ?     ?     ?     500
       500   0     -25   -25   0     0     0     0     50    50    0
       24    30    6))
((5    5     0     300   78    -222  36    40    4     5     5     0
       0     0     0     0     0     0     ?     ?     ?     500
       200   -300  1.72  3.24  1.52  0     6     6     50    18    -
32     21    42    21))
((6    6     0     300   33    -267  0     35    35    4     4     0
       41    47    6     1     3     2     ?     ?     ?     500   80
       -420  -25   -8.75 16.25 0     5     5     50    26    -24   39
       60    21))
((7    6     -1    82    52    -30   40    44    4     2     4     2
       38    53    15    0     6     6     ?     ?     ?     340
       183   -157  15.09 8.31  -6.78 2     5     3     43    28    -
15     39    39    0))
((0    5     5     300   100   -200  0     30    30    0     3     3
       0     54    54    0     6     6     ?     ?     ?     500
       210   -290  -25   5     30    0     4     4     50    20    -
30     30    0     -30))
((0    0     0     300   300   0     0     0     0     0     0     0
       0     0     0     0     0     0     ?     ?     ?     500
       500   0     -25   -25   0     0     0     0     50    50    0
       0     0     0))
((6    5     -1    60    120   60    11    15    4     4     4     0
       55    53    -2    10    6     -4    ?     ?     ?     300
       220   -80   -0.11 2.49  2.6   6     6     6     0     7     9     2
```

100%

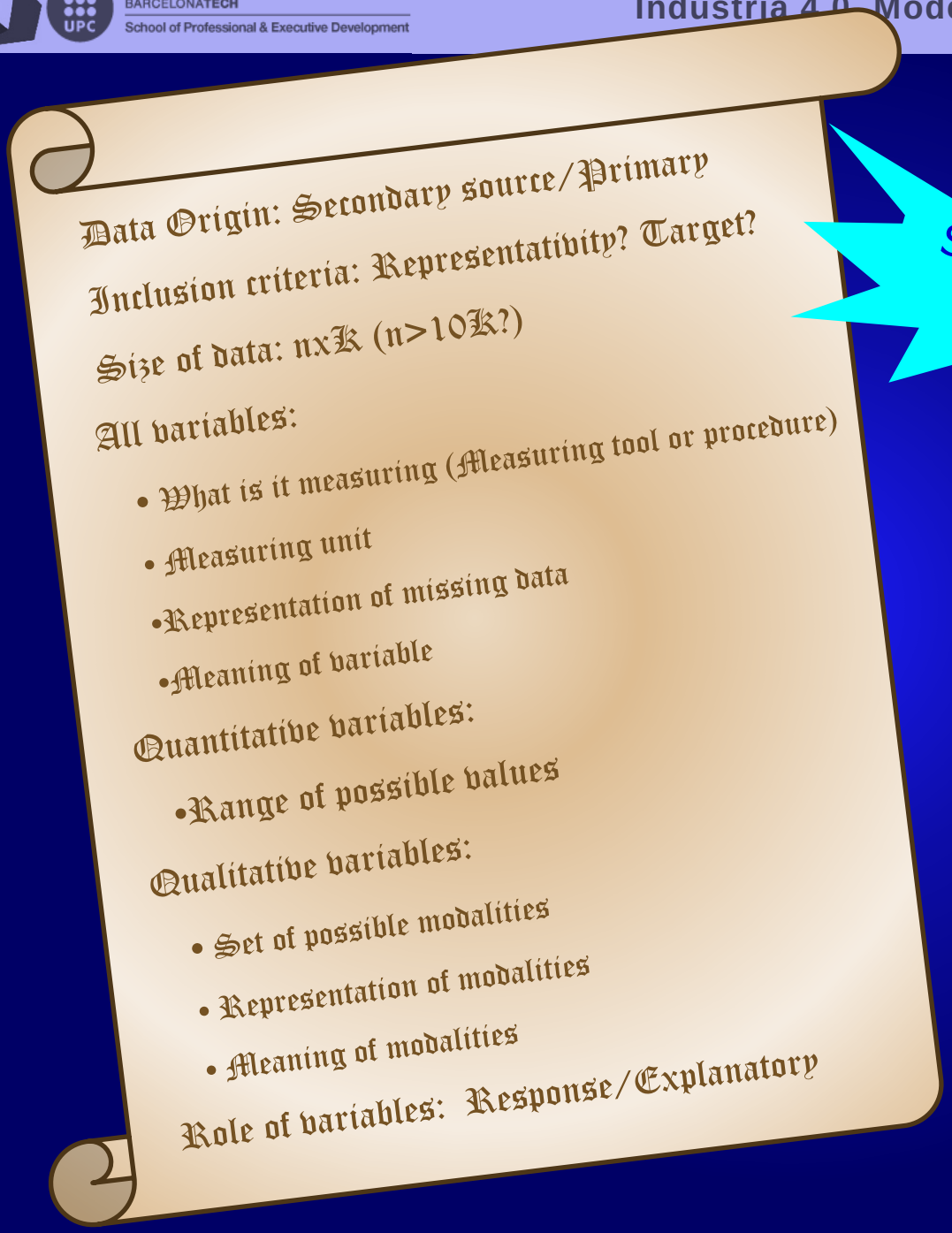
Start 7.... ori... altr... 1D... pre... ES 11:25 14/03/2014

From Data to Decisional Knowledge

DATA <> **INFORMATION**



Metadata



Still an open problem

Software do not support

- External project documentation manually managed
- Relational Data Base for very complex Data Matrices *[Gibert, MMR 92]*

Gibert, K., & Marti-Recober, M. (1992). A System for Production and Analysis of Statistical Reports. In Computational Statistics (pp. 363-368). Physica-Verlag HD.

Metadata File

url: *www.xxx.ssss.www*

Inclusion criteria: *People in [18,65] years, no hard attacks, no smoking, no cholesterol, married, with sons or daughters….*

n*: nro of rows*

K*: nro of columns*

Variable	Modalities	meaning	Type	Measuring unit	Missing code	Measuring procedure	Range	Role
Age		Age of marriage	Num	years	"*"		[1,105]	Explanatory
Sex		Gender	Quali		Unknown			Explanatory
	M	Male						
	H	Female						
FeC		Level of Iron in blood	Num	μg/dl	NA	Biochemical analysis on blood sample measuring transferrine……	[30, 200]	Explanatoyr
Anemy		The person has anemy diagnosis	Boolean		Unknown	Levels of Fec<xxx and …….		Response

First insight to Data

- Look at Metadata
- Determine rows and columns to be kept for the analysis
- Basic descriptive analysis of remanining variables
  - Inspect anomalies, errors, missing data, outliers
- First report about data quality
- Preprocessing
- Verify after each processing step
- Final descriptive analysis *(report data improvements)*

Reading Data and Variables Declaration

Reading and declaring data

- Verify that software got all rows and columns
  - Care with Spanish and English .csv files
- Verify that software understands variable types properly
  - Care with qualitative variables codified by integers
  - Care with numerical interpreted as textual variables
  - Metadata helps
- Ensure proper ordering in ordinal variables
- Use short modality labels

Reading and declaring data

Practical activity

R package preliminars

Integer
Double
Character
Logical (boolean)
Date
Vector and Matrix (numerical)
Factors (qualitative vectors)
Data Frames
Functions
Lists

Default creation with the assignment: <- or ->

Special values:

- NA: for missing data
- Inf: Infinity (division by 0...)
- NaN: error in the function results

Reading and declaring data

- Verify that software got all rows and columns
  - Care with Spanish and English .csv files
- Verify that software understands variable types properly
  - Care with qualitative variables codified by integers
  - Care with numerical interpreted as textual variables
  - Metadata helps
- Ensure proper ordering in ordinal variables
- Use short modality labels

Basic descriptive statistics and Automatic reporting

Descriptive analysis

0. Basic Principles in descriptive analysis
1. Graphical Descriptive tools for numerical variables
2. Numerical Descriptive tools for numerical variables
3. Numerical Descriptive tools for qualitative variables
4. Graphical Descriptive tools for qualitative variables
5. Descriptive Analysis of Temporal variables
6. Introduction to Space
7. Normality and Exponentiality Evaluation

Descriptive analysis

Compact and Informative view of the variable structure

DATA= FIT+ ERROR

General Pattern **Deviations**

Structural Component Random Component

Tools

1. Graphical

   Visualize variable's distribution

The human eye-brain system remains the best pattern recognition device [Goebel 1999][Ware 167]

1. Numerical

   Quantify what is observed in he graph

Goebel, M., Gruenwald, L.: A survey of data mining and knowledge discovery software tools. SIGKDD Explorations Newsletter 1, 20–33 (1999)
Ware, C.: Information visualization: perception for design. Business & Economics, Morgan Kaufmann, San Francisco, CA, 2 edn. (2004)

Graphical tools

1. Performing the graph

   Mechanical (software)

2. Reading the graph

   Technical (statistitian or data miner)

3. Interpretation

   Conceptual (domain expert)

   Contextualization

Descriptive analysis

Practical activity2

Find data *(min 20 vars, min 6 num, 6 quali)*

(use file opendatalinks.doc to inspire)

Describe automatically

Markdown file

Graphical tools for numerical variables

1. Histogram
2. Boxplot

3. Others (dotplot, stem and leaf plot….)

Histogram

Visualitzation of frequencies distribution table

Intervalo	Número de Observaciones	Observaciones Acumuladas	Frec Relativas	Frec Acumuladas
45–65	1	1	$1/17 = .06$	0.06
65-75	5	6	0.29	0.35
75-85	5	11	0.29	0.64
85-95	1	12	0.06	0.70
95-105	3	15	0.17	0.87
[illegible]115	0	15	0	0.87
[illegible]	1	16	0.06	0.93
[illegible]35	0	16	0	[illegible]
35-145	1	17	0.06	[illegible]

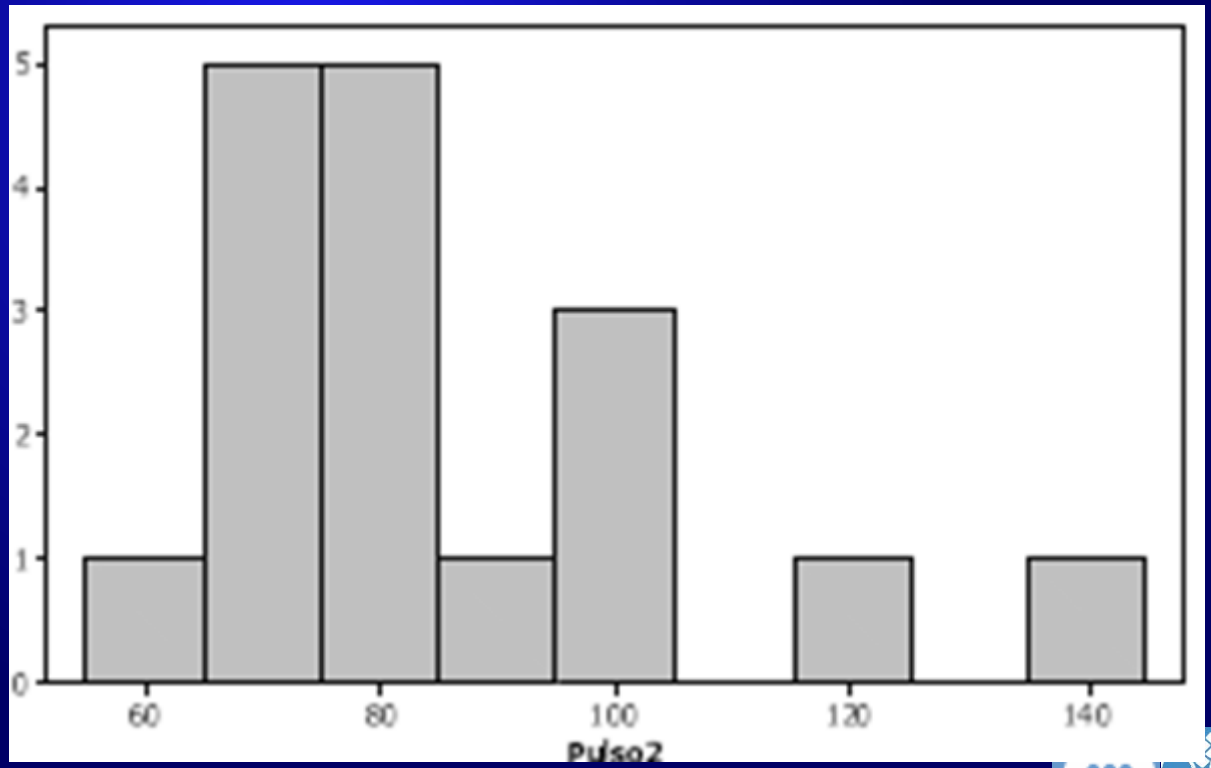
Frequency Classes?

Bars' AREAS PROPORTIONAL to frequencies

Heuristics: $\begin{cases} 6\ log_{10}(n) & ,si\ n < 100 \\ 1{,}2\sqrt{n} & ,si\ n \geq 100 \end{cases}$ $3{,}49\ s\ n^{-\frac{1}{3}}$ $2\ d_i\ n^{-\frac{1}{3}}$

READING HISTOGRAMS

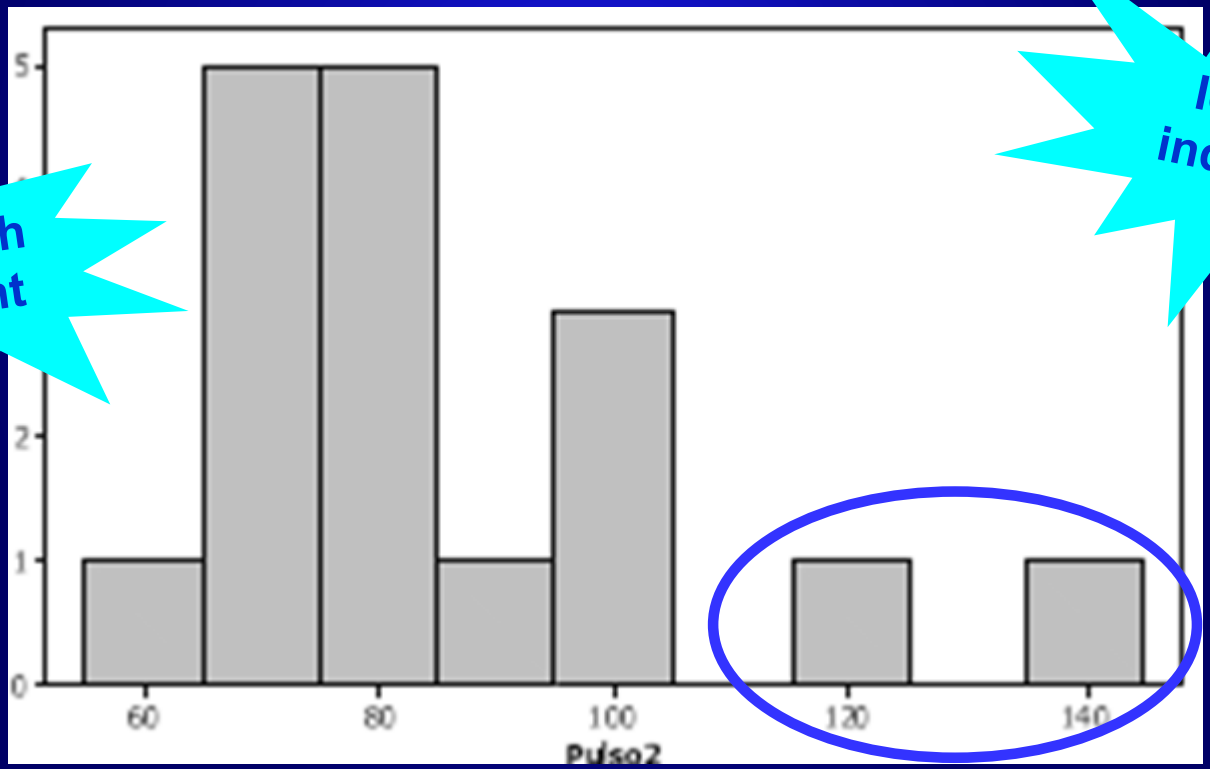
1. Range of variable (max-min)
2. Central trend
3. Dispersion
4. Simmetry
5. Anomalies



READING HISTOGRAMS

Anomalies

Outliers: Observations anormaly far from rest



CARE with Treatment

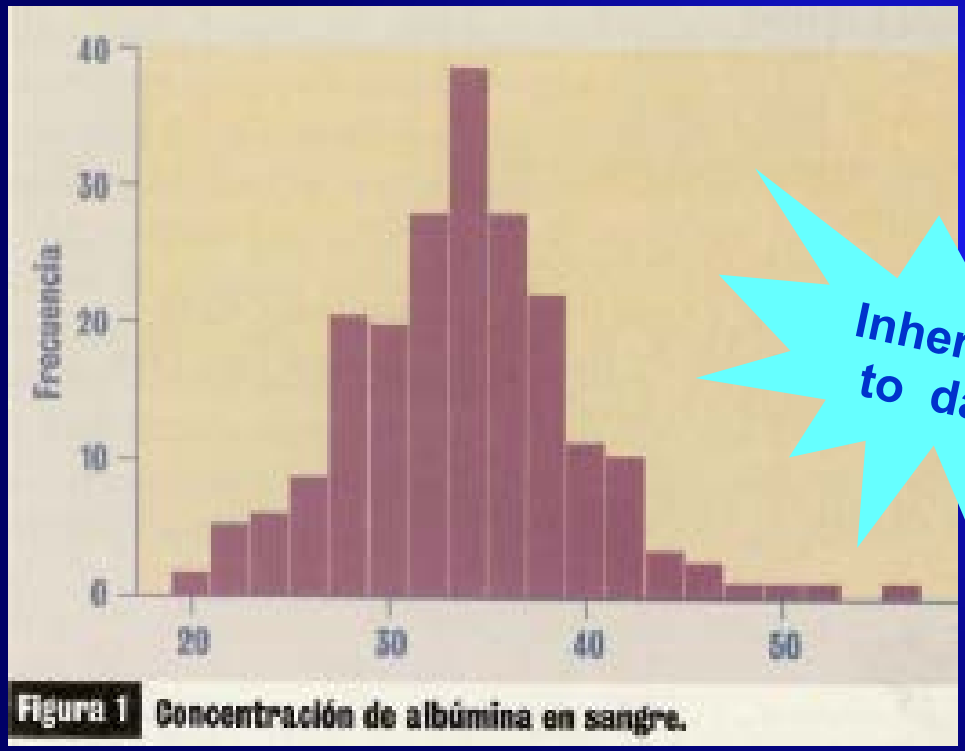
Identify individual

READING HISTOGRAMS

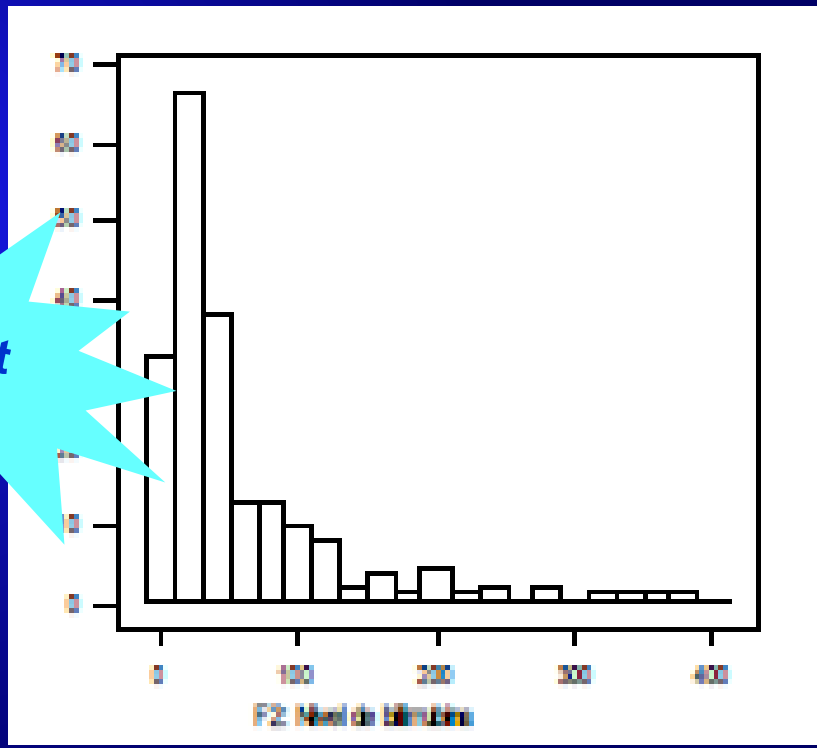
Main Patterns

[Gibert, JANO1996]

Albumine



Bilirrubine



Inherent to data

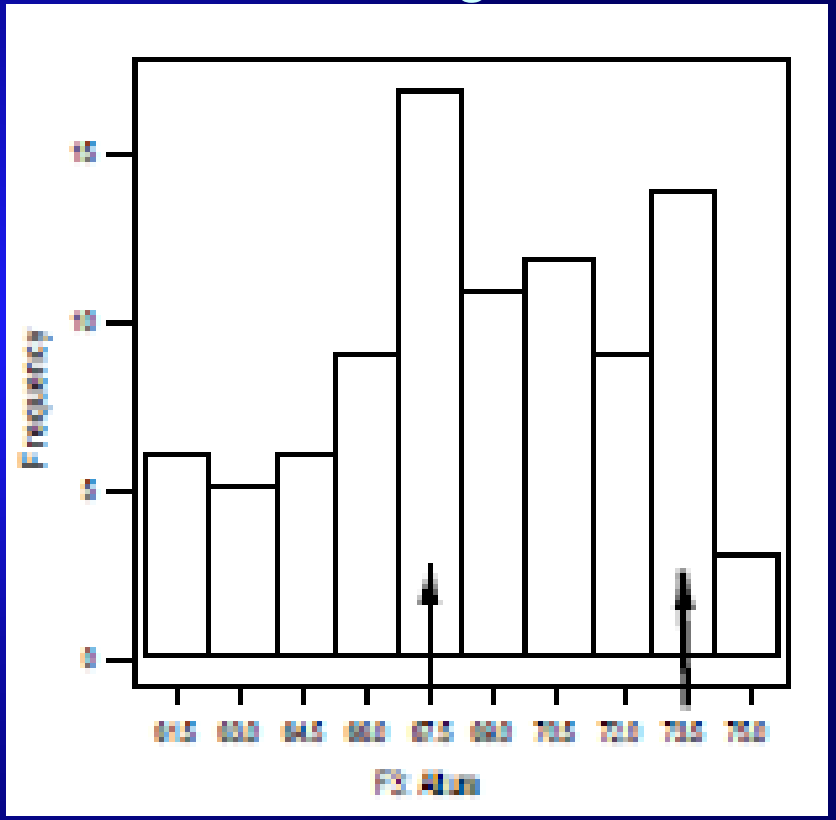
Symmetric

Asymmetric

READING HISTOGRAMS

Main Patterns

Height



Multimodality

Several central trends!!

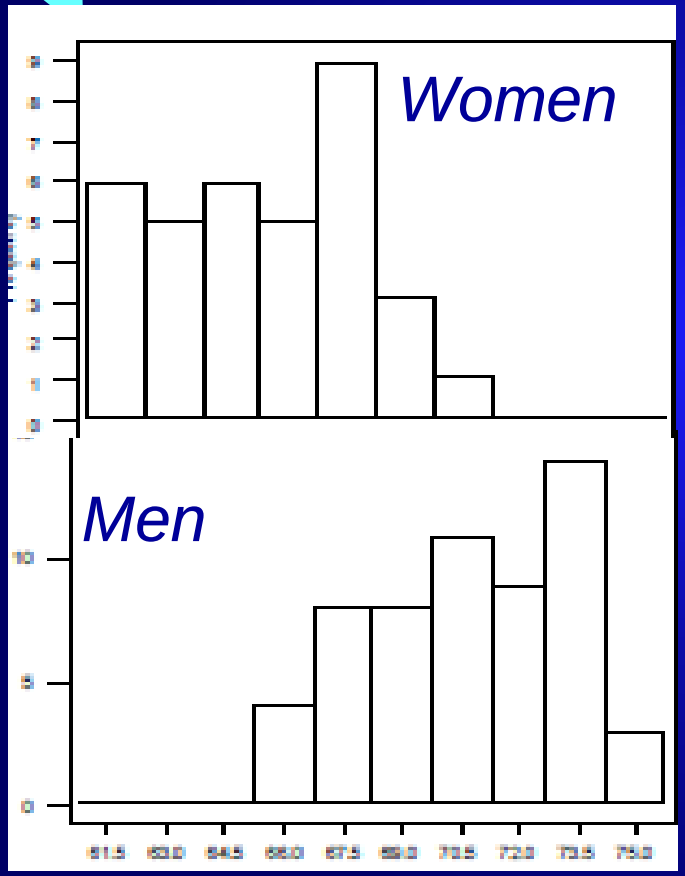
Find discriminant factor

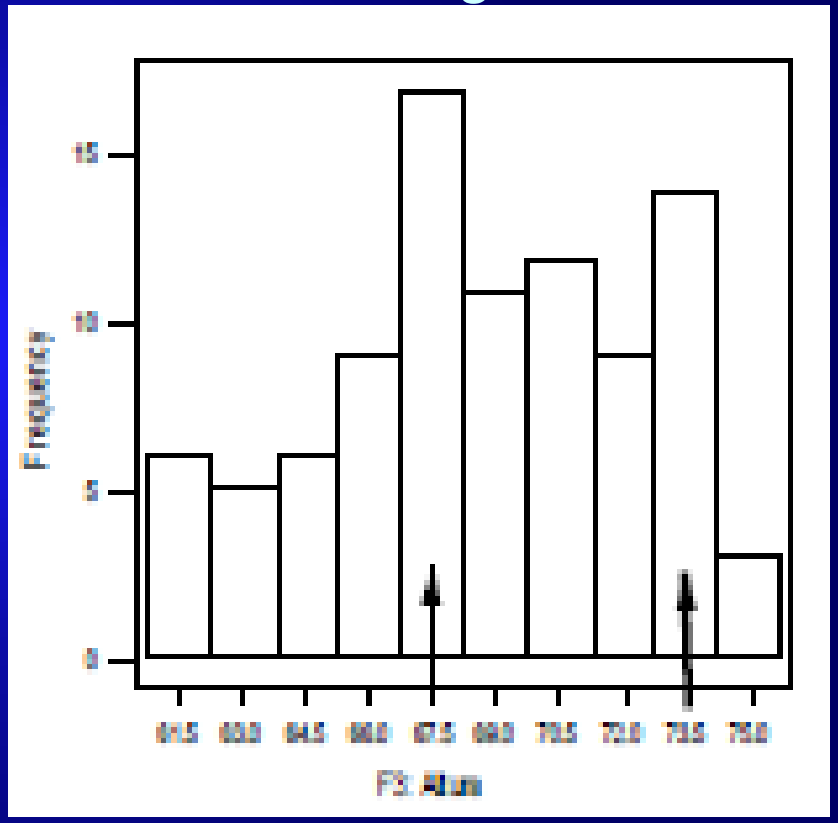
READING HISTOGRAMS

Find discriminant factor

Main Patterns

Height

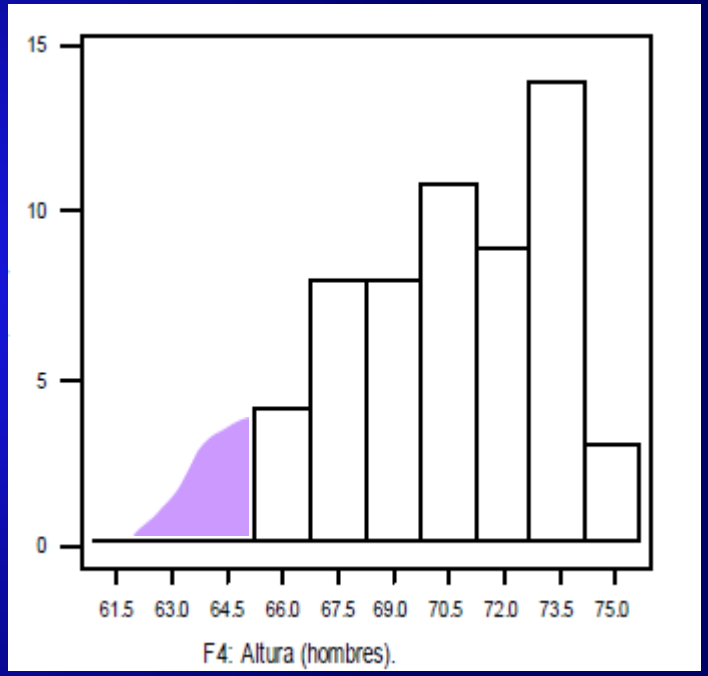




READING HISTOGRAMS

Main Patterns

Height of Men



Scarped

Part of distribution trunked!

(only adult men)

READING HISTOGRAMS

Main Patterns

Dentat

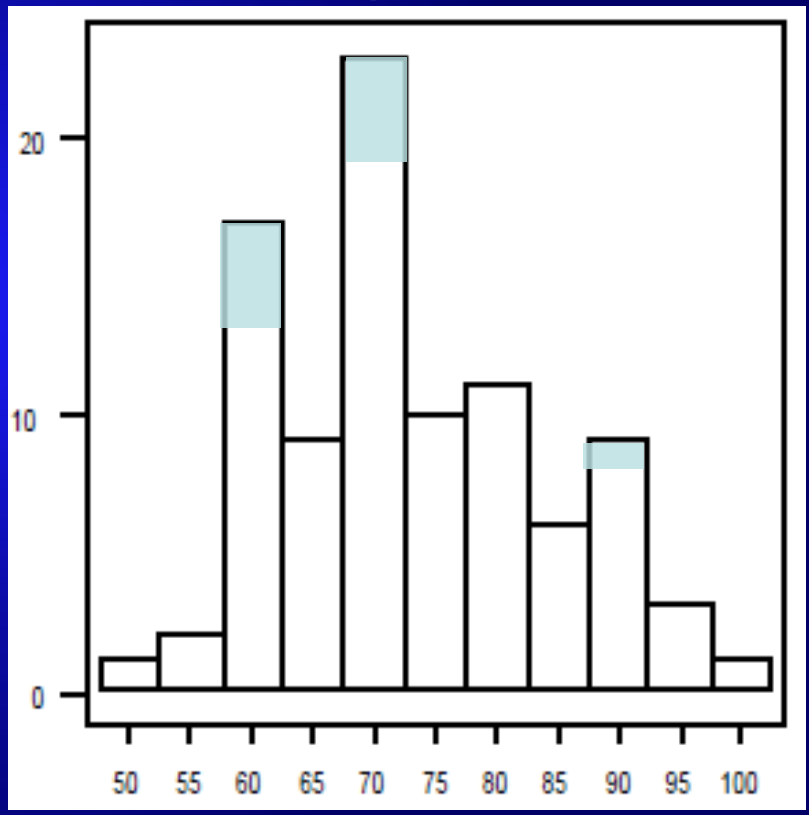
Measurement approximations!

Count one minute

Count 10 sec x 6

Count 25 sec x 4

Pulse per minute



Tools

1. Graphical

   Visualitze variable's distribution

2. Numerical

   Quantify what is observed in he graph

Numerical tools

Quantify and synthetize characteristics of a distribution

1. According to the information provided
    1. Central trend statistics
    2. Variability statistics
2. According to the stability
    1. Classic
    2. Robust

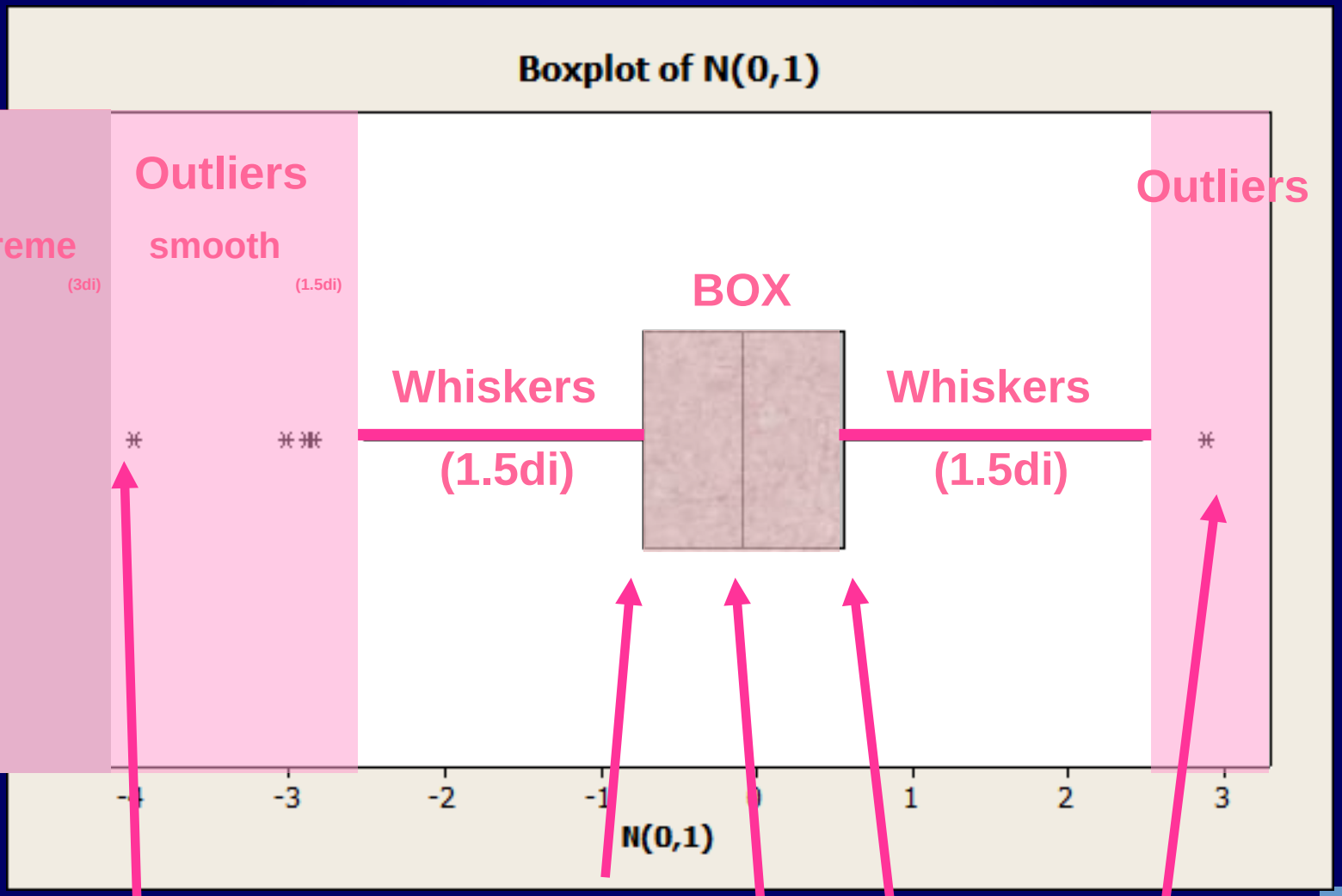
Numerical tools

for numerical variables

	Robusto	Clásico
Posición	Mediana Cuartiles Percentiles Moda	Media
Dispersión	Distancia entre cuartiles	S Desviación estándar S^2 Varianza Coef. variación Amplitud

Boxplot *[Tukey 1956]*

Symbolic representation of 5-Number Summary

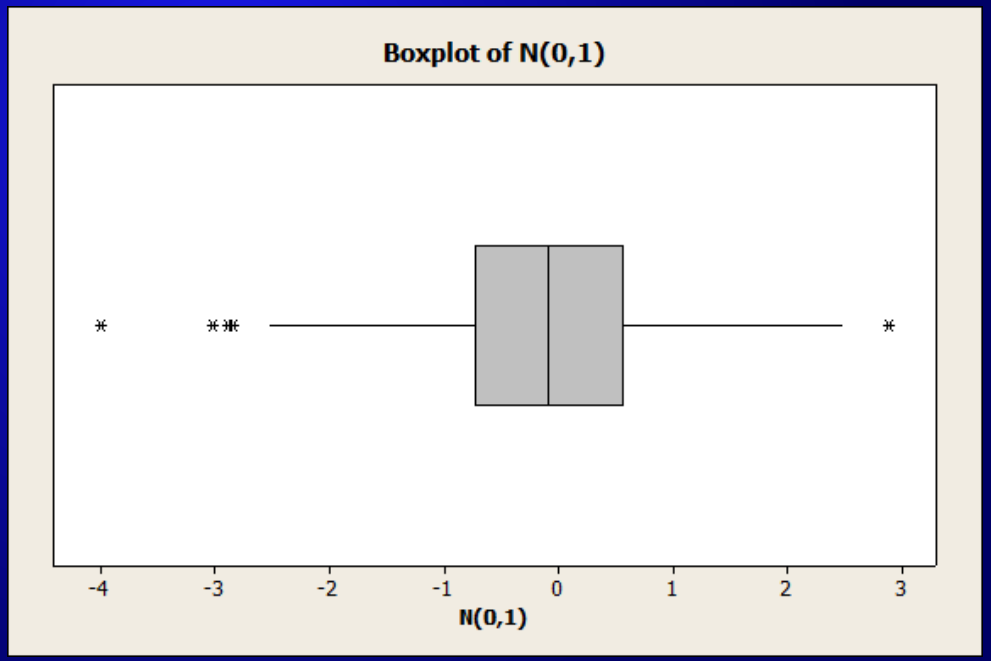


READING BOXPLOTS

1. Range of variable (max-min)
2. Central trend
3. Dispersion
4. Simmetry
5. Anomalies

READING BOXPLOTS

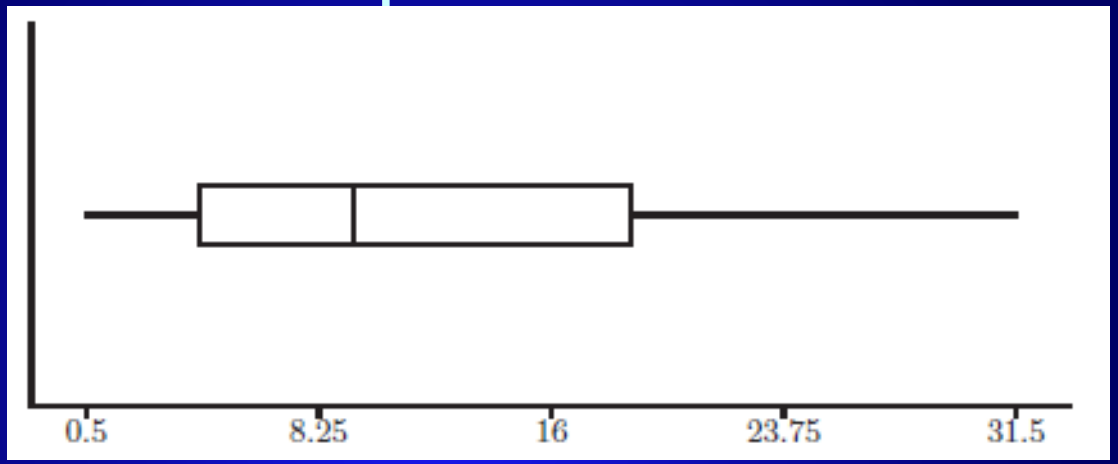
1. Range of variable (max-min)
2. Central trend
3. Dispersion
4. Simmetry
5. Anomalies



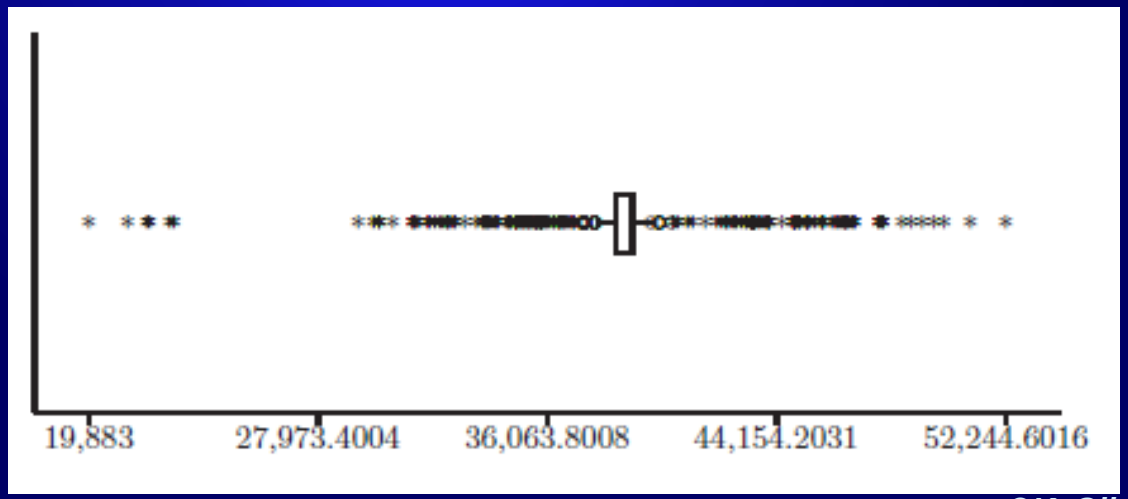
READING BOXPLOTS

Dispersion

Ammonium



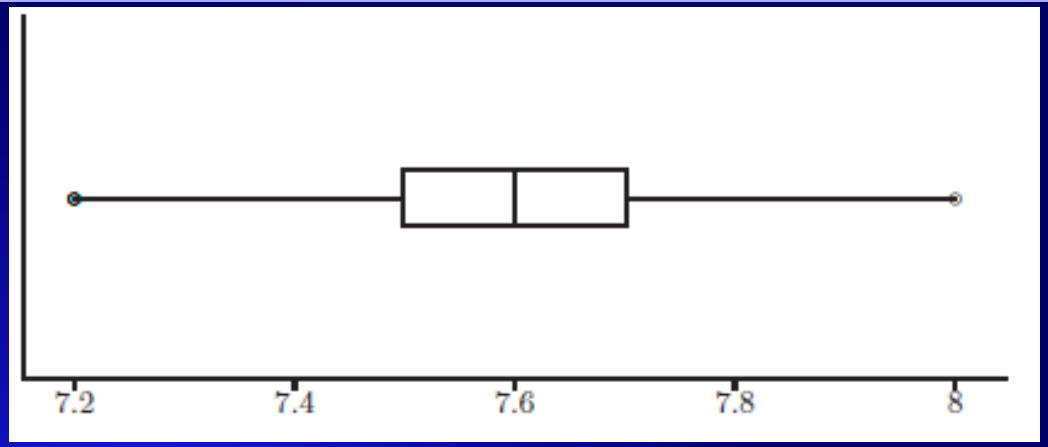
QB-B



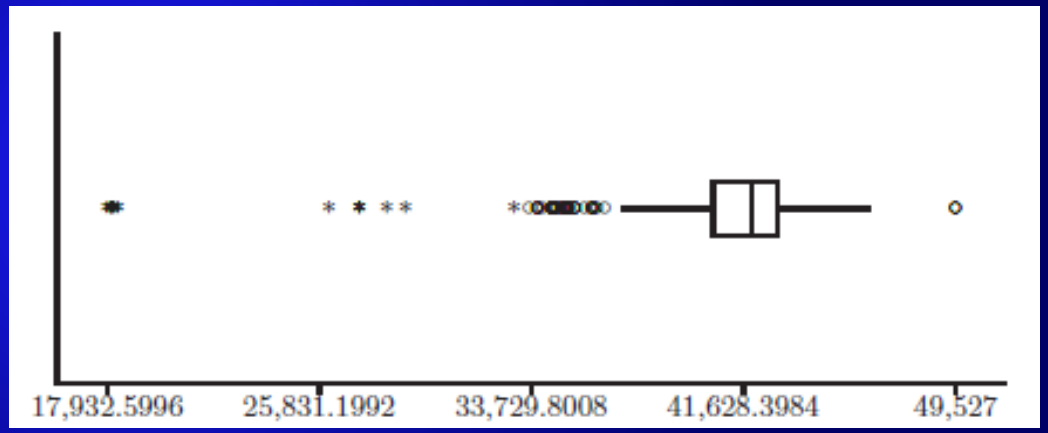
READING BOXPLOTS

Simmetry

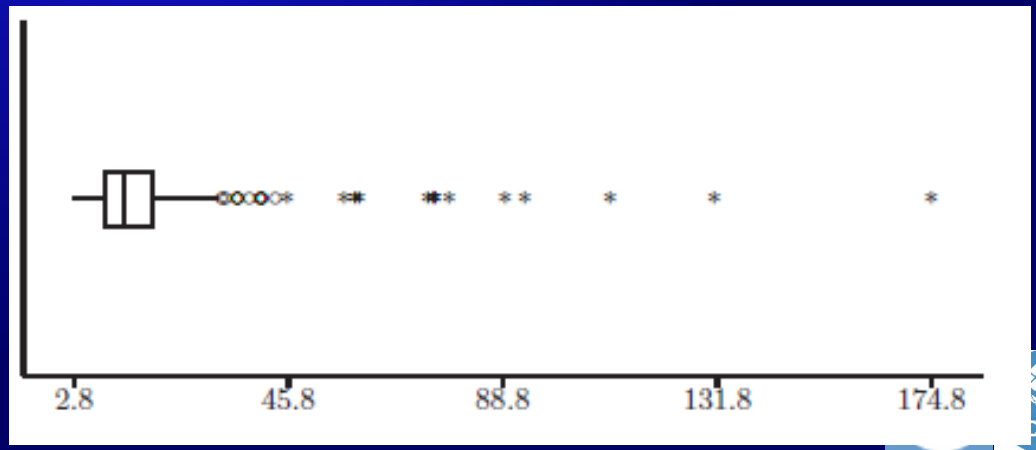
PH



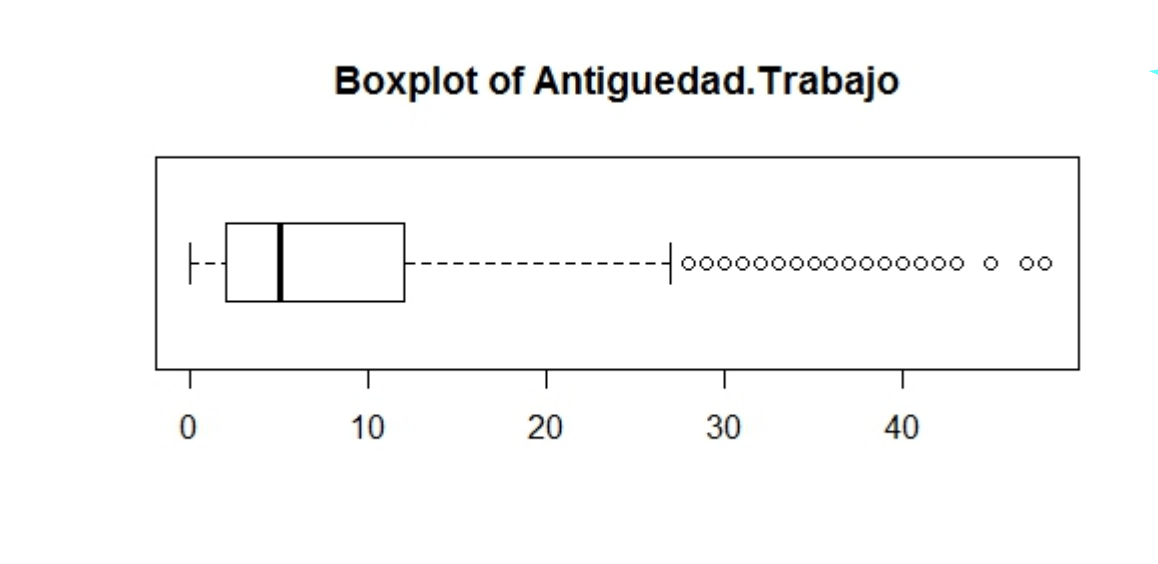
QR-G



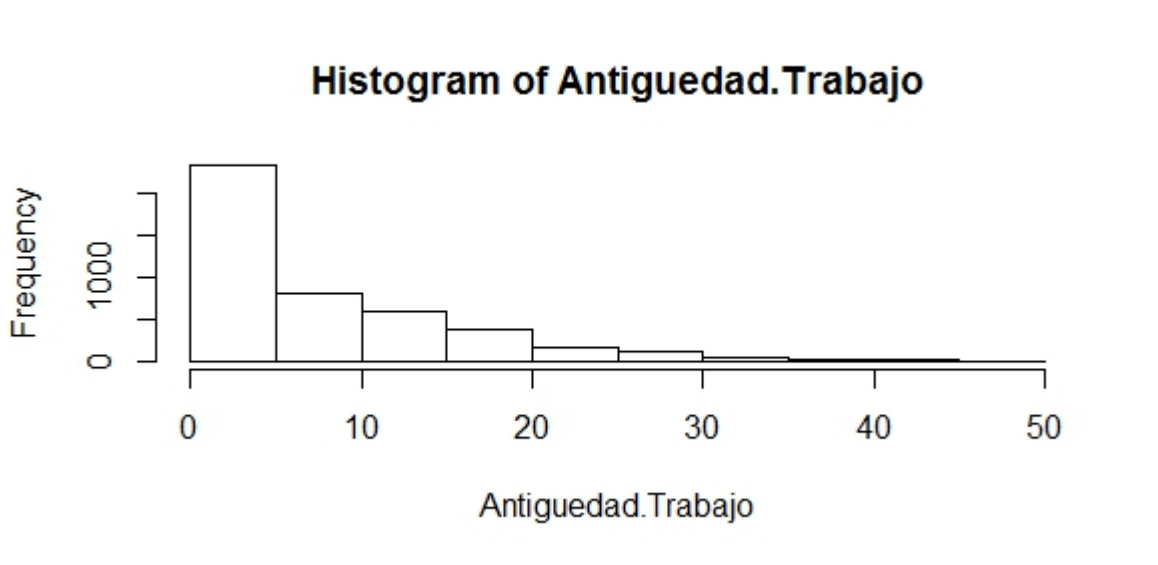
SS-S



READING BOXPLOTS



CARE With *



Symmetry

if Mean <> Median then

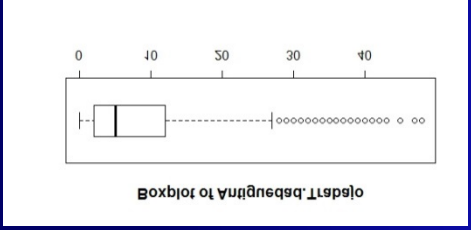
if Me-Q1 < >Q3-Me then assymmetry

else outliers

else symmetry without outliers

Synthesis

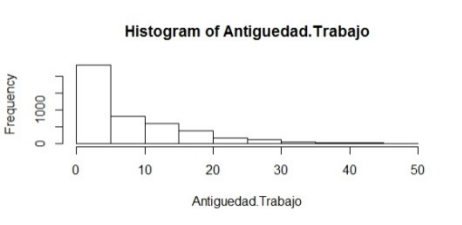
1. Descriptive analysis of numerical variable
   1. Central trend and variability (classical/robust)
   2. Graphical and Numerical tools



5-Number-Summary:

<min, Q1, Me, Q3, max>

+



mean, q-stdev, variation coefficient

   3. Characterize the variable

Central trend, variability, symmetry, n-modality….

Frequencies Distribution Table

X: qualitative variable

s: number of modalities (categories, values)

$c_1 \ldots c_s$: values

Accumulated columns only for ordinal variables

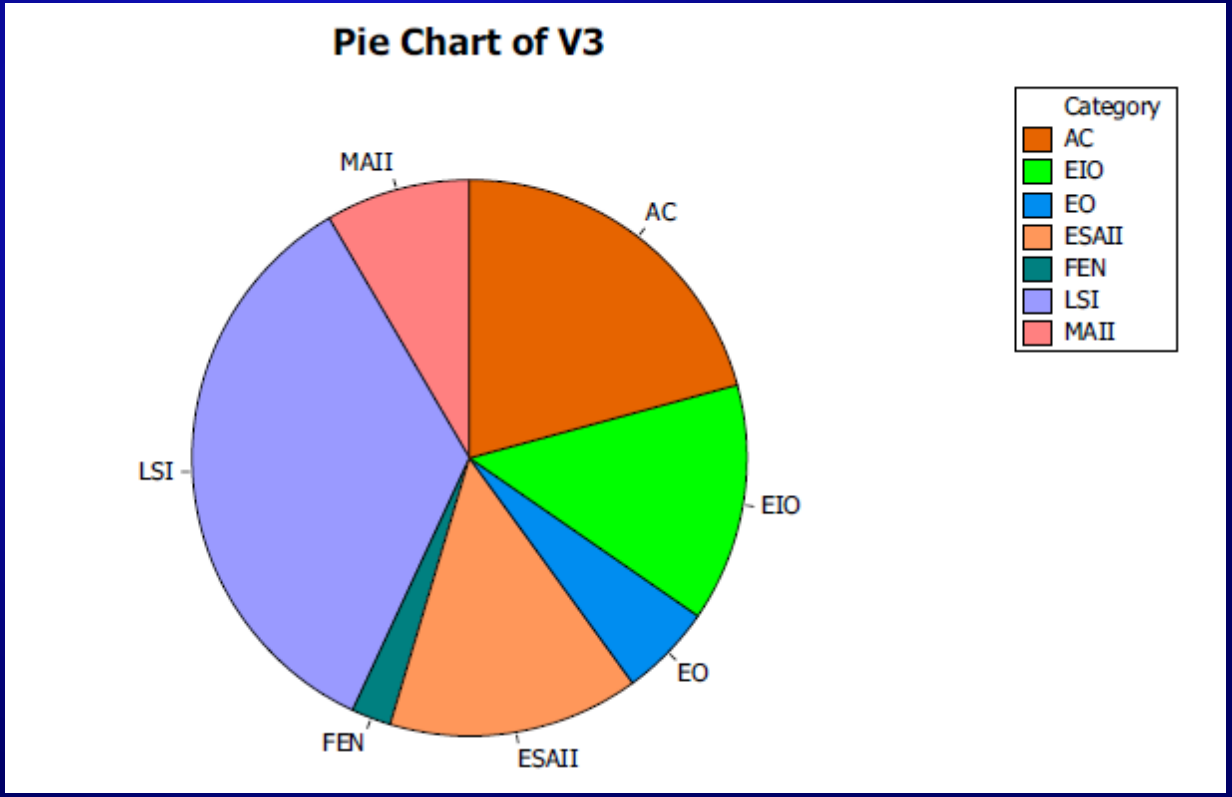
X	n_i	$\sum_{n_j}^{n_i}$	f_i	$\sum_{j=1}^{i} f_j$
c_1	n_1	n_1	$\frac{n_1}{n}$	$\frac{n_1}{n}$
c_2	n_2	$n_1 + n_2$	$\frac{n_2}{n}$	$\frac{n_1}{n} + \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_{s-1}	n_{s-1}	$n_1 + \ldots + n_{s-1}$	$\frac{n_{s-1}}{n}$	$\frac{n_1}{n} + \ldots + \frac{n_{s-1}}{n}$
c_s	n_s			1

Impact of missing in percentages

Graphical tools for qualitative variables

Pie chart

X	n_i
AC	27
EIO	18
$ESAII$	19
FEN	3
LSI	45
$MAII$	11
OE	7
	130



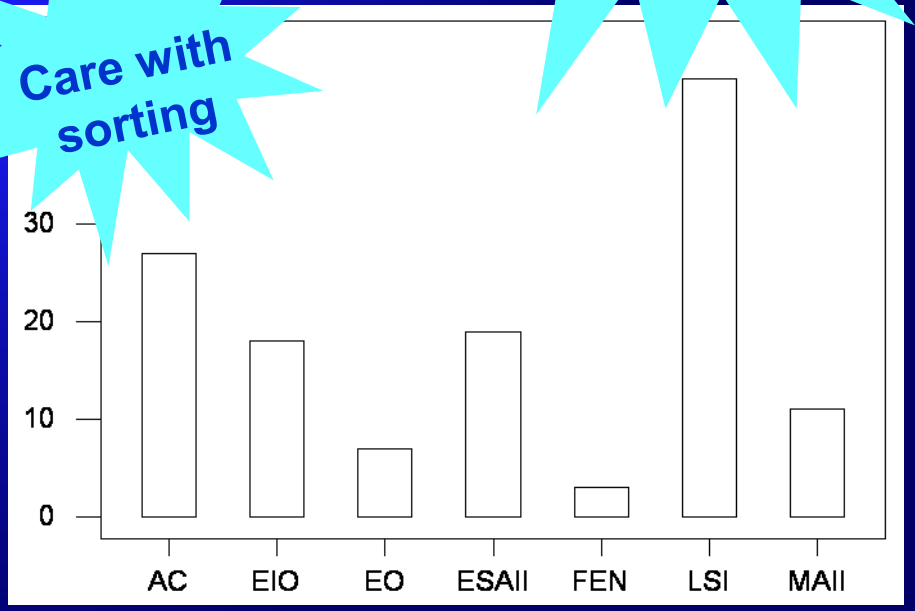
Graphical tools for qualitative variables

Bar chart (Bar Plot)

X	n_i
AC	27
EIO	18
$ESAII$	19
FEN	3
LSI	45
$MAII$	11
OE	7
	130

Care with sorting

Preferred for ordinal variables



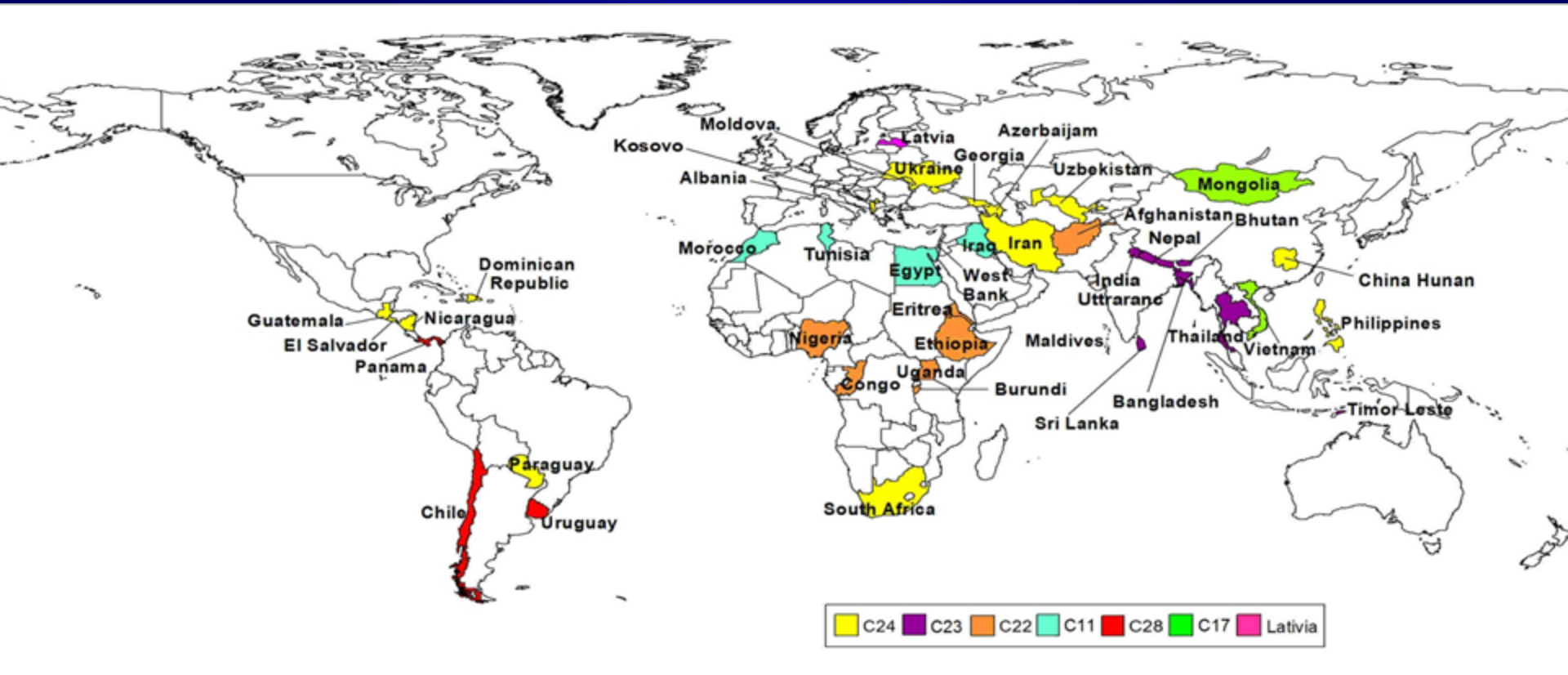
READING PIES and BARPLOTS

1. Uniform distribution?

2. Central trend (Mode)

3. Dominant modalities?

Representación de mapas

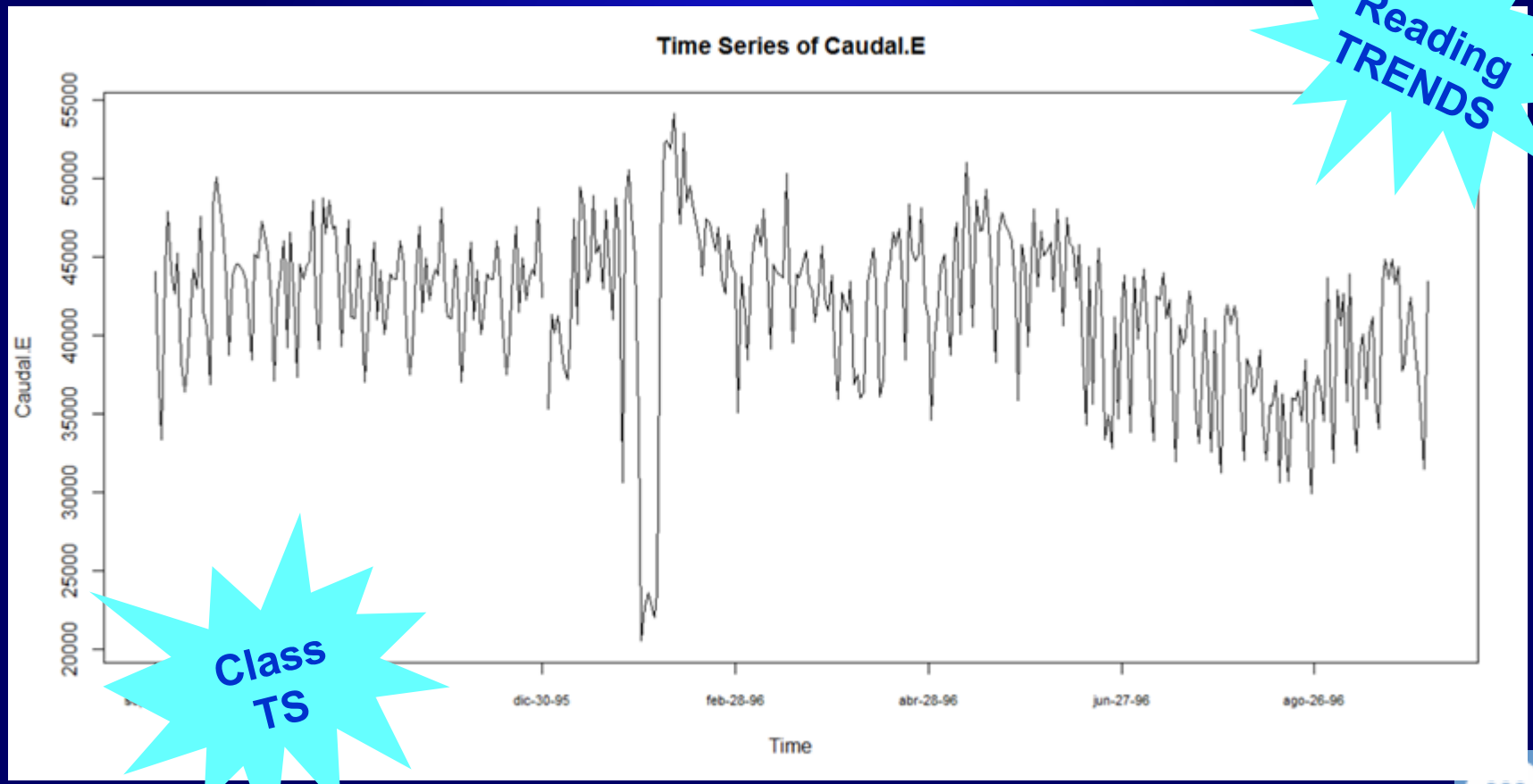


https://support.office.com/es-es/article/crear-un-gr%C3%A1fico-de-mapa-f2cfed55-d622-42cd-8ec9-ec8a358b593b

Cronological data

Observations are sequentially sorted in the dataset

Time series plot



Reading TRENDS

Class TS

Cronological data

Observations are sequentially sorted in the dataset

Evenctually the Date is Available

w.dma.ulpgc.es/profesores/personal/stat/cursoR4ULPGC/6h-Fechas.html

Cronological data

Date Objects in R

Symbo	Meaning	Example
%d **%D**	day as a number (0-31) Date format	01-31
%a **%A**	abbreviated weekday unabbreviated weekday	Mon Monday
%m	month (00-12)	00-12
%b **%B**	abbreviated month unabbreviated month	Jan January
%y **%Y**	2-digit year 4-digit year	07 2007

31/12/2014 : %d/%m/%Y

31-Dic-07: %d-%b-%y

Cronological data

Date Objects in R

Symbo	Meaning	Example
%c	Date and time	
%C	Century	
%H	Hours (00-23)	15
%I	Hours (1-12)	3
%j	Day of the year (0-365)	250
%M	minute (00-59)	January
%S	Second as integer(0-61)	07

23:12:59 = %H:%M:%S

11 12 59 = %I %M %S

Cronological data

Observations are sequentially sorted in the dataset

Evenctually the Date is Available

Class Date

To consider time

Class POSIXCT

Preprocessing

First insight to Data

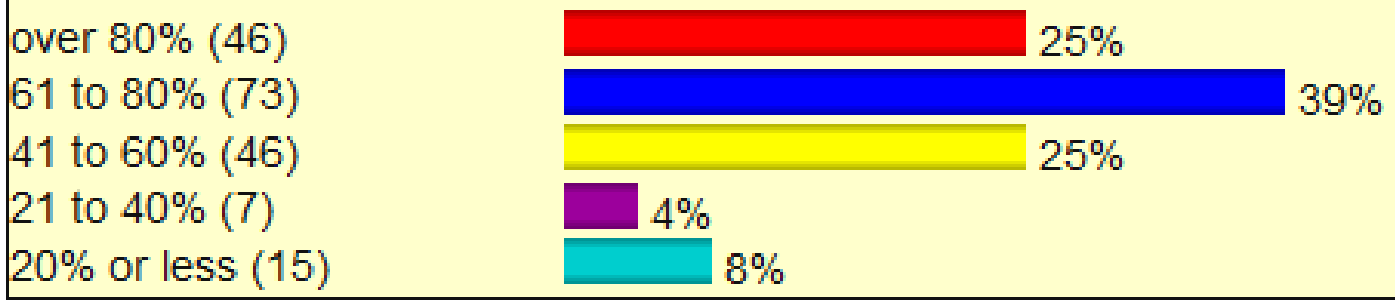
- Look at Metadata
- Determine rows and columns to be kept for the analysis
- Basic descriptive analysis of remanining variables
  - Inspect anomalies, errors, missing data, outliers
- First report about data quality
- Preprocessing
- Verify after each processing step
- Final descriptive analysis *(report data improvements)*

Impact of Preprocessing in real Data Mining projects

Data preparation part in data mining projects

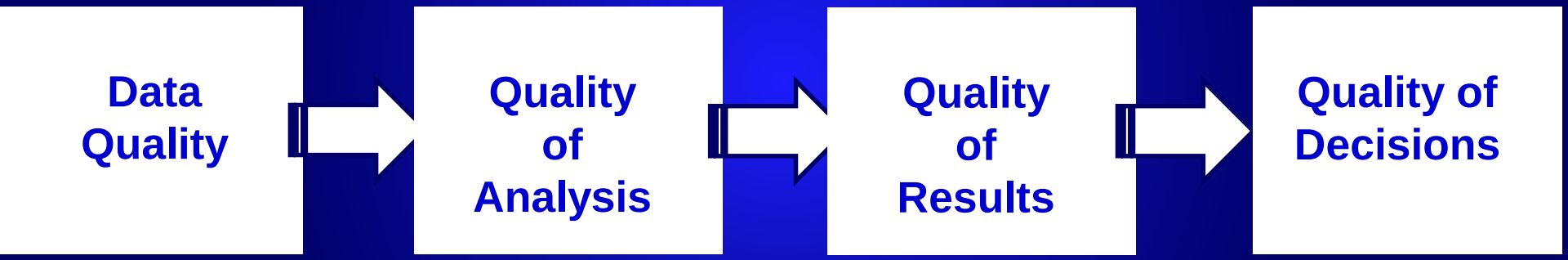
Poll

What % of time in your data mining project(s) is spent on data cleaning and preparation [187 votes total]



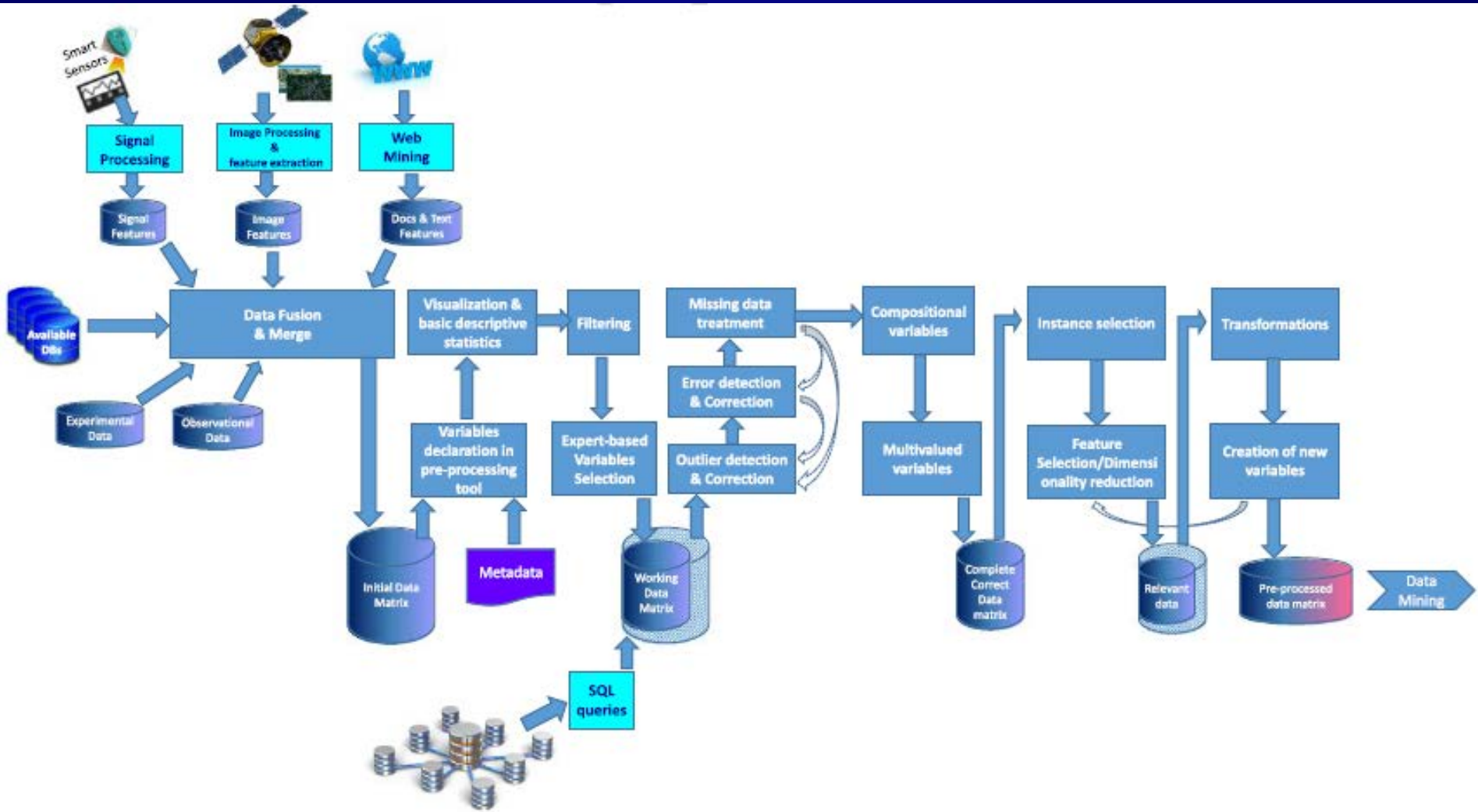
http://www.kdnuggets.com/polls/2003/data_preparation.htm

Preprocessing



Methodology

[Gibert Aicomm2016]



Gibert, K., M. Sànchez-Marrè, J. Izquierdo (2016) A Survey on Pre-processing Techniques in the Context of Environmental Data Mining. Artificial Intelligence in Communications, 29(6): 627-663, IOSPress DOI: 10.3233/AIC-160710

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Objects Selection

Inclusion/Exclusion criteria
Filtering

Select from a data base or data warehouse or from real individuals (costs are different)

- Experimental data (experimental design)
- Observational data (sample theory)

REPORT IN FINAL DOC

Define the target population

Determines scope of conclusions

Goals' oriented Variables Selection

- Often expert-guided
  (highly related with goal of analysis)

  - Be maximalists
    - Eliminate irrelevant or redundant information is less risky than detect lack of relevant things to be added in a second wave

    - Technically, to complete a final submatrix is highly costly (in both time and resources)

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Missing data

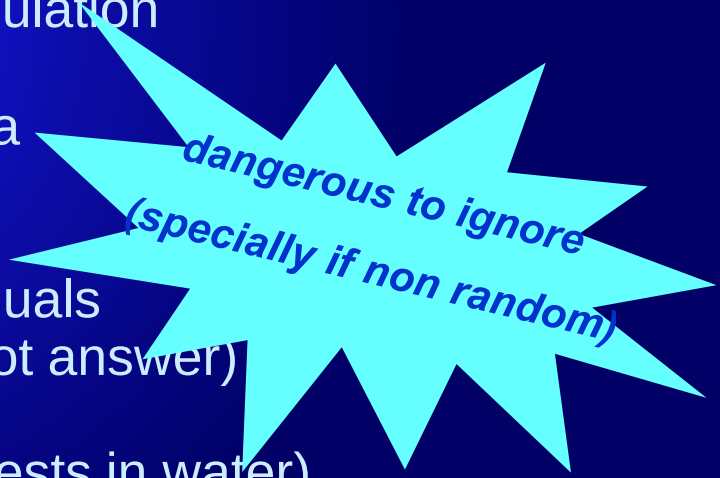
(empty cells in data matrix)

- Types and diagnosis
- Little's test
- A simple descriptive alternative
- Some methods
  - Knn
  - The MIMMI method
  - MICE
  - Interpolation (for time series)

Missing data

(empty cells in data matrix)

- Randon missing
  - non problematic
  - casual
  - follow same distribution as present data
  - inputation is easy: mean, 0
- Non random missing: absence is informative
  - come from some particular part of population
  - probably correspond to special values
  - difficult to induce from the present data
  - inputation is much difficult
  - very critical
  - very dangerous to ignore those individuals
  - asking religion in israel (muslims do not answer)
  - Asking age to a lady over 45
  - Frequency of observations (microbio tests in water)
- Non applicable value (non-random, structural,sistematico)
  - salary of a non-working person
  - number of pregnancies of a man



Missing data

Diagnoses

Little's MCAR test

H_0: Missings are completely at random (MCAR)
H_1: Missings are not random

$$d^2 = \sum_{j=1}^{J} n_j (\overline{X}_j - \overline{X}_j^*)^T \frac{1}{\hat{\sum}_j} (\overline{X}_j - \overline{X}_j^*) \sim \chi^2_{\sum r_j - K}$$

j=1:J missing patterns (subsets of missing variables in a case)

- n_j *cases in missing pattern j*
- $\overline{X}_j$ *maximum likelihood estimates of the grand means*
- $\overline{X}_j^*$ *means local to cases in missing pattern*
- $\hat{\sum}_j$ *maximum likelihood estimate of the covariance matrix*
- *rj number of complete variables for pattern j*
- *K total number of variables*

Searches significant differences in means conditioned to a certain subset of missing variables (pattern j)

- Dades UNICEF 2004 http://www.unicef.org/spanish/infobycountry/index.html missing patterns

	Països	Zona	PobTotal	INBxcap	EsperVida	%NounatsBaixPes	UsInstalSaneja	VIH	%coneixPresHomes	$ConeixPresDones	AlfabetHomes	AlfabetDones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

- Missing pattern 1: (ConeixPresHomes)
- Missing pattern 2: (ConeixPresHomes+DespesaSalut/Defensa)
- Missing pattern 3: (ConeixPresHomes+ConeixPresDones)
- Missing pattern 4: (ConeixPresHomes+ConeixPresDones+DespSal/Def)
- Missing pattern 5: (ConeixPresHomes+AlfabetsDones+DespSal/Def)
- Missing pattern 6: (VIH+ConeixPresHomes+ DespSal/Def)
- Missing pattern 7: (VIH+ConeixPresHomes+ ConeixPresHomes+ConeixPresDones+DespSal/Def)
- Missing pattern 8: (ConeixPresHomes+AlfabetHomes+AlfabetDones+DespSal/Def)

n1=4, x*1=1
n2=2, x*2=2
n3=9, x*3=2
n4=1, x*4=3
n5=1, x*5=3
n6=1, x*6=3
n7=2, x*7=4
n8=1, x*8=4

	Països	Zona	PobTotal	INBxcap	EsperVida	%NounatsBaixPes	UsInstalSaneja	VIH	%coneixPresHomes	$ConeixPresDones	AlfabetHomes	AlfabetDones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

n1=4, x*1=1
n2=2, x*2=2
n3=9, x*3=2
n4=1, x*4=3
n5=1, x*5=3
n6=1, x*6=3
n7=2, x*7=4
n8=1, x*8=4

Missing pattern 9: (VIH+CPH+ CPD+AlfaH+AlfaD+DespSal/Def)
Missing pattern 10: (INB+VIH+CPH+ CPD+AlfaH+AlfaD+DespSal/Def))
Missing pattern 1: (EV+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def)
Missing pattern 12: (INB+ EV+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def)
Missing pattern 13: (INB+ EV+NBp+IS+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def
Missing pattern 14: (Pob+INB+ EV+NBp+IS+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+

n9=3, x*9=6
n10=1, x*10=7
n11=3, x*11=9
n12=3, x*12=10
n13=1, x*13=12
n14=1, x*14=13

	Països	Zona	PobTotal	INBxcap	EsperVida	%NounatsBaixPes	UsInstalSaneja	VIH	%coneixPresHomes	$ConeixPresDones	AlfabetHomes	AlfabetDones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

- n1=4, x*1=1
- n2=2, x*2=2
- n3=9, x*3=2
- n4=1, x*4=3
- n5=1, x*5=3
- n6=1, x*6=3
- n7=2, x*7=4
- n8=1, x*8=4

- n9=3, x*9=6
- n10=1, x*10=7
- n11=3, x*11=
- n14=1

15 patterns

	Països	Zona	PobTotal	INBxcap	Esper Vida	%NounatsBaixPes	UsInstalSaneja	VIH	%coneixPresHomes	$ConeixPresDones	Alfabet Homes	Alfabet Dones	Mortalitat	Natalitat	Economia	despesaSalutDefensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

n3=9, x*3=2, r3= 13-2=11

$$\overline{X}_3 = [\overline{X}_{Pob}, \overline{X}_{INB}, \overline{X}_{EV}, \overline{X}_{\%NbP}, \overline{X}_{AlfaH}, \overline{X}_{AlfaD}, \overline{X}_M, \overline{X}_N]$$

$$\bar{X}_3^* = [\bar{X}^*_{Pob}, \bar{X}^*_{INB}, \bar{X}^*_{EV}, \bar{X}^*_{\%NbP}, \bar{X}^*_{AlphaH}, \bar{X}^*_{AlphaD}, \bar{X}^*_M, \bar{X}^*_N]$$

$\hat{\sum}_3$ = Variances and covariances of full variables

$\overline{X}_{Pob} = 38120$

$\overline{X}^*_{Pob} = 172062$

$$d^2 = \sum_{j=1}^{J} n_j (\overline{X}_j - \overline{X}^*_j)^T \frac{1}{\hat{\sum}_j} (\overline{X}_j - \overline{X}^*_j) \sim \chi^2_{\sum r_j - K}$$

	Països		PobTotal	INBxcap	EsperVida	%NounatsBaixPes	UsInstalSaneja	VIH	%coneixPresHomes	$ConeixPresDones	AlfabetHomes	AlfabetDones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

The Little test in R

- LittleMCAR {BaylorEdPsych}
- **USAGE:** LittleMCAR(x)

  x: dataframe, matrix less than 50 variables

- **Returns:**

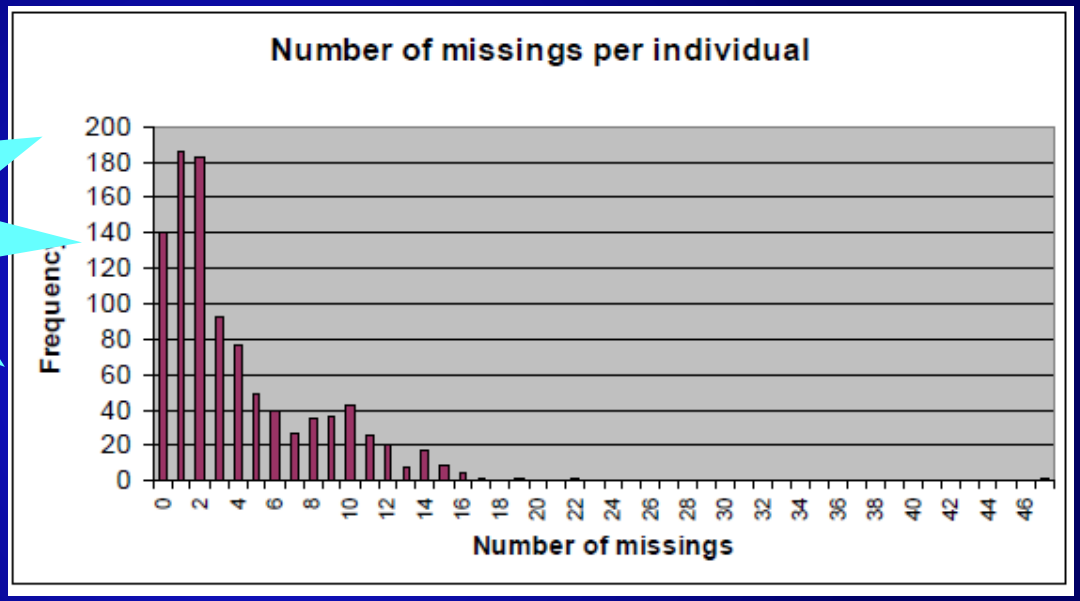
chi.square	Chi-square value
df	Degrees of freedom used for chi-square
missing.patterns	Number of missing data patterns
amount.missing	Amount and percent of mssing data
data	The data, organized my missing data patterns

Missing data

(Simple alternative)

- Build new variable counting number of missings per individuals.
- Describe this variable

Reliability of individual



- Count nr of missing per variable and rank variables
- *Provides reliability*

- Create indicator of missing/non-missing per variable
- *and compare both groups of cases*

Missing data

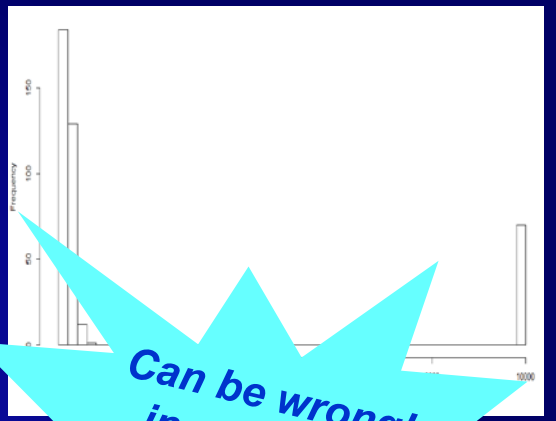
(empty cells in data matrix)

- Representation:
- *,?, “ “, depending on software
  - numerical variables: sometimes codified (0, 99999, -1…)
  - categorical variables: special modality (Ns/Nc, …)

- Standardize missing representation

- Causes of missing data:
  - *voluntary hidden (religion in israel) (always non-random)*
  - *data non-provided*
  - *data non-achieveable*
    - *technical limitations (example anemometers IKE hurrican)*
    - *accessibility (no privileges, sensitive information)*
  - *data lost*
  - *data forced to missing (as a result of correction)*

- *Identification:*
  - *Numerical indicators (stdev…)*



Can be wrongly incorporated in analysis

Missing data treatment

- Depends on analysis goals!!!!!
- keep it as a missing: only eventually
  - Can signifficantly reduce the treated observations
- Inputing: Substituting by a useful value (open problem, difficult)
  - Qualitative variable: Substitute by “Unkown<varName>”
  - Stardard way, expert knowledge required
    - use 0
    - use global mean
    - use conditional mean for local groups
    - inputation models (complex)
    - Nearest neighbor (R)
    - Intelligent inputation
    - MIMMI
  - non parametric aproach (montecarlo methods, multiple inputation)
    - special software required
    - technical hypothesis about variable distributions required
    - Final models integration required
      - *Example: French survey, global incomes of household*

- Essential to treat missing data BEFORE producing derivated variables

Missing data treatment

- Missing values frequent in real data
- Imputation before analysis CRITICAL
- Most statistical packages:
  - simple inputation by global mean
  - listwise deletion (dangerous)
- Specific softwares:
  - dedicated to sophisticated inputation methods
  - highly time consuming
  - non-exportable complete data matrices
- Find a trade-of between precision and simplicity

Knn method

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1569,0	84585,0	09/07/2003 0:00	7	7	6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7	7	7	7	7	7	7	7	7
1645,0	84990,0	21/07/2003 0:00	7		6	6	2	6	6	6	3
1666,0	91980,0	09/03/2004 0:00	7	7	7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7	7	7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7	7	6	7	6	6	6	6	7
1858,0	76281,0	26/09/2002 0:00	7		5	7	7	6	5	5	7
1900,0	84368,0	01/07/2003 0:00	6	6	4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4	7	4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7	7	7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7	5	3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7	7	6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1	1	1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5	5	1	1	1	1	6	5	1
2267,0	86796,0	01/10/2003 0:00	7		7	7	7	7	6	6	7
2524,0	76436,0	02/10/2002 0:00	7	7	6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7	7	7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7	7	6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7	7	7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7	7	7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7	7	7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6	6	4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5	5	2	5	2	1	5	5	4

Knn method

Original uncomplete data

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de roc
1569,0	84585,0	09/07/2003 0:00	7	7	6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7	7	7	7	7	7	7	7	7
1645,0	84990,0	21/07/2003 0:00	7		6	6	2	6	6	6	3
1666,0	91980,0	09/03/2004 0:00	7	7	7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7	7	7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7	7	6	7	6	6	6	6	7
1858,0	76281,0	26/09/2002 0:00	7		5	7	7	6	5	5	7
1900,0	84368,0	01/07/2003 0:00	6	6	4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4	7	4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7	7	7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7	5	3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7	7	6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1	1	1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5	5	1	1	1	1	6	5	1
2267,0	86796,0	01/10/2003 0:00	7		7	7	7	7	6	6	7
2524,0	76436,0	02/10/2002 0:00	7	7	6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7	7	7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7	7	6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7	7	7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7	7	7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7	7	7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6	6	4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5	5	2	5	2	1	5	5	4

SPLIT

Cures Full

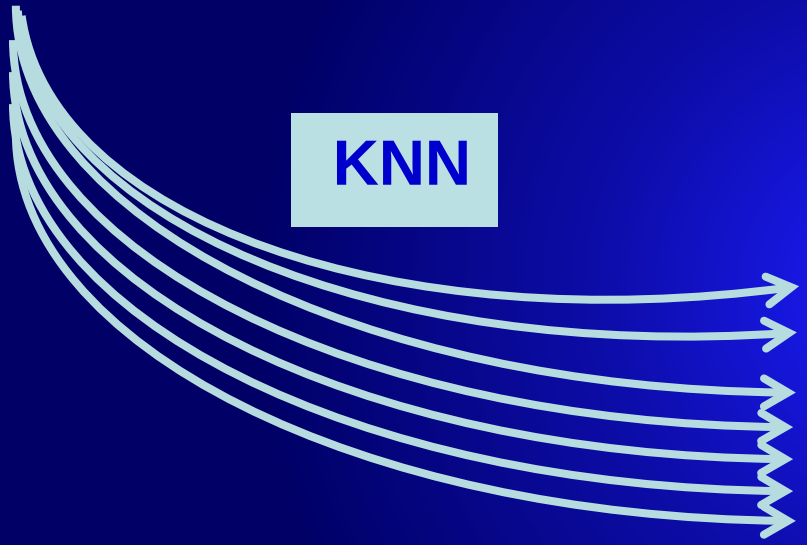
Cures Missing

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de roc
1569,0	84585,0	09/07/2003 0:00	7	7	6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7	7	7	7	7	7	7	7	7
1666,0	91980,0	09/03/2004 0:00	7	7	7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7	7	7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7	7	6	7	6	6	6	6	7
1900,0	84368,0	01/07/2003 0:00	6	6	4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4	7	4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7	7	7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7	5	3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7	7	6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1	1	1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5	5	1	1	1	1	6	5	1
2524,0	76436,0	02/10/2002 0:00	7	7	6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7	7	7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7	7	6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7	7	7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7	7	7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7	7	7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6	6	4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5	5	2	5	2	1	5	5	4

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de roc
1645,0	84990,0	21/07/2003 0:00	7		6	6	2	6	6	6	3
1858,0	76281,0	26/09/2002 0:00	7		5	7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7		7	7	7	7	6	6	7

Knn method

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1645,0	84990,0	21/07/2003 0:00	7			6	2	6	6	6	3
1858,0	76281,0	26/09/2002 0:00	7			7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7			7	7	7	6	6	7



C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de roc
1569,0	84585,0	09/07/2003 0:00	7		6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7		7	7	7	7	7	7	7
1666,0	91980,0	09/03/2004 0:00	7		7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7		7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7		6	7	6	6	6	6	7
1900,0	84368,0	01/07/2003 0:00	6		4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4		4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7		7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7		3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7		6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1		1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5		1	1	1	1	6	5	1
2524,0	76436,0	02/10/2002 0:00	7		6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7		7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7		6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7		7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7		7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7		7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6		4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5		2	5	2	1	5	5	4

Knn method

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1645,0	84990,0	21/07/2003 0:00	7			6	2	6	6	6	3
1858,0	76281,0	26/09/2002 0:00	7			7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7	7		7	7	7	6	6	7

KNN

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de roc
1569,0	84585,0	09/07/2003 0:00	7		6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7		7	7	7	7	7	7	7
1666,0	91980,0	09/03/2004 0:00	7		7	7	7	7	6	5	7
1694,0	83561,0	03/06/2003 0:00	7		7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7		6	7	6	6	6	6	7
1900,0	84368,0	01/07/2003 0:00	6		4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4		4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7		7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7		3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7		6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1		1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5		1	1	1	1	6	5	1
2524,0	76436,0	02/10/2002 0:00	7		6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7		7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7		6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7		7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7		7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7		7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6		4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5		2	5	2	1	5	5	4

Mixed Intelligent-Multivariate Missing Imputation *[Gibert 2013]*

- Select a small number of quasi-full relevant variables

  (*with small ratio of missing values*)
- Use ***intelligent inputation*** on that reduced data matrix

  (expert-based inputation, vertical-horizontal)
- Multivariate hierarchical clustering using the inputed variables
- Determine a partition of the data *(2nd or 3rd best cut)*
- Compute class local means for ALL variables
- Inpute remaining missing data with comp[illegible] loca[illegible]

Trade-off
Accuracy/required time

Gibert, K (2013) Mixed Intelligent-Multivariate Missing Imputation, International Journal of Computer Mathematics 91(1):85-96

MIMMI method on WHO-AIMS database

Selection of 16 core variables

WHO-AIMSname	Meaning	KLASSname
polplanr	Presence of policy or plan	polplanr
legisl	Presence of legislation	Legisl
d1f5i5rec	Affordability of anitpsychotic medicine	d1f5i5rec(antipsych)
d1f5i6rec	Affordability of anitdepressant medicine	d1f5i6rec(antidepr)
D2F1I2	Oragnization of services	D2f1i2(orgServices)
cbusrate	Community based inpatient units per 100,000 population	cbusrate
mhrate	mental hospitals per 100 000 population	mhrate
outpfrate	outpatient facilities per 100 000 population	outpfrate
daytrfrate	day treatment facilities per 100 000	daytrfrate
D4F1I11	psychiatrists per 100 000	d4f1i11(psychi)
D4F1I12	other doctors per 100 000	D4F1I12(doctors)
D4F1I13	nurses per 100 000	D4F1I13(nurses)
D4F1I14	psychologists per 100 000	D4F1I14(psycho)
D4F1I15	Social workers per 100 000	D4F1I15(socWorK)
d3f1i3	availability of treatment and assessment manuals	d3f1i3(manuals)
d5f2i51	formal collaborative relationship with department of primary care	d5f2i51(relprimcare)
D6F1I1	formally defined min data set	D6F1I1(mindataset)

MIMMI method

- Selection of 16 core variables:
  - Characteristic information of the whole 6 domains
  - Related with decisional variables (composite indicators)
  - Low tax of missing data

KLASSname	nMissing
outpfrate	1
D4F1I12(Other)	1
D4F1I13(nurses)	1
D4F1I14(psycho)	2
D4F1I15(socWorK)	1
d3f1i3(manuals)	1
d5f2i51(relprimcare)	1

- total of 8 missing cells to be [illegible]

Complex process

time consuming

applicable in real projects with experts

- Intelligent inputation of 8 missing values

Intelligent inputation of 8 missing values

Country	Missing variable and imputed value
South Africa	outpfrate = 2.0
China	D4F1I12(Other) = 1.20
India	D4F1I13(nurses) = 0.15
China	D4F1I14(psycho) = 0.16
Paraguay	D4F1I14(psycho) = 1.4
Nepal	D4F1I15(socWorK)= 0.15
Moldova	d3f1i3(manuals) = B
Azerbaijan	d5f2i51(relprimcare) = N

Intelligent inputation of 8 missing values

Country	Missing variable and imputed value
South Africa	outpfrate = 2.0
China	D4F1I12(Other) = 1.20
	D4F1I13(nurses) = 0.1
	D4F1I14(psycho) = 0.16
Paraguay	D4F1I14(psycho) = 1.4
Nepal	D4F1I15(socWorK)= 0.15
Moldova	d3f1i3(manuals) = B
Azerbaijan	d5f2i51(relprimcare) = N

Vertical Imputation

Rate of "other medical doctors" in China
Very institutional system.
Higher than other Asian countries.
Between Vietnam and Thailand.

Two Strategies

Intelligent inputation of 8 missing values

Country	Missing variable and imputed value
[illegible]ca	outpfrate = 2.0
	D4F1I12(Other) = 1.20
India	D4F1I13(nurses) = 0.15
China	D4F1I14(psycho) = 0.16
Paraguay	D4F1I14(psycho) = 1.4
Nepal	D4F1I15(socWorK)= 0.15
Moldova	d3f1i3(manuals) = B
Azerbaijan	d5f2i51(relprimcare) = N

Horizontal Imputation

India-Uttaranchal (nurses rate) High where hospitals are. India do not has mental hospital

Choose a low value

Two Strategies

Inputation:

Complete the 42x16 data matrix

Clustering the full matrix

Hierarchical

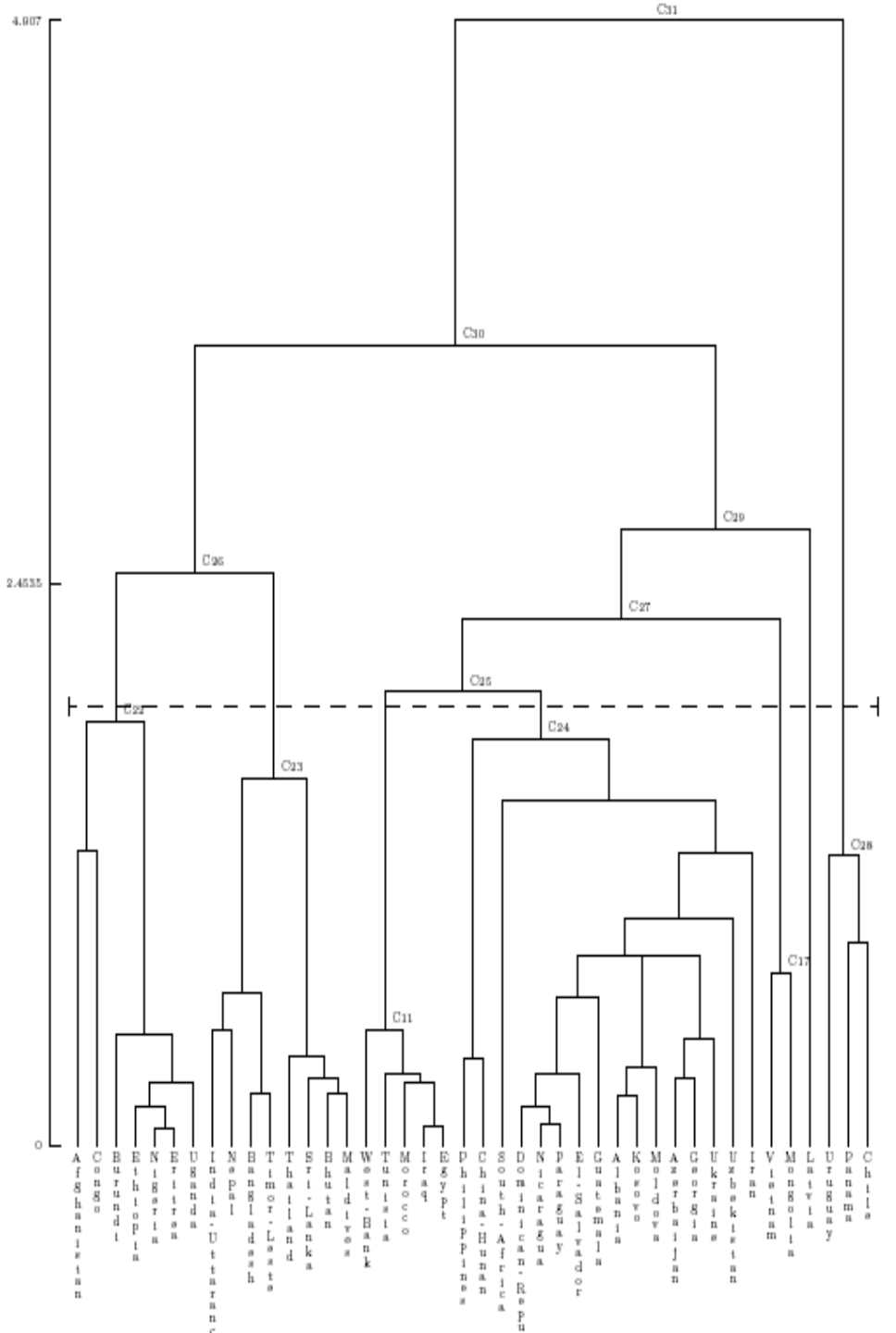
Ward criterion

Gibert's mixed metrics

[Gibert 96]

Determine the classes (7)

Find partition

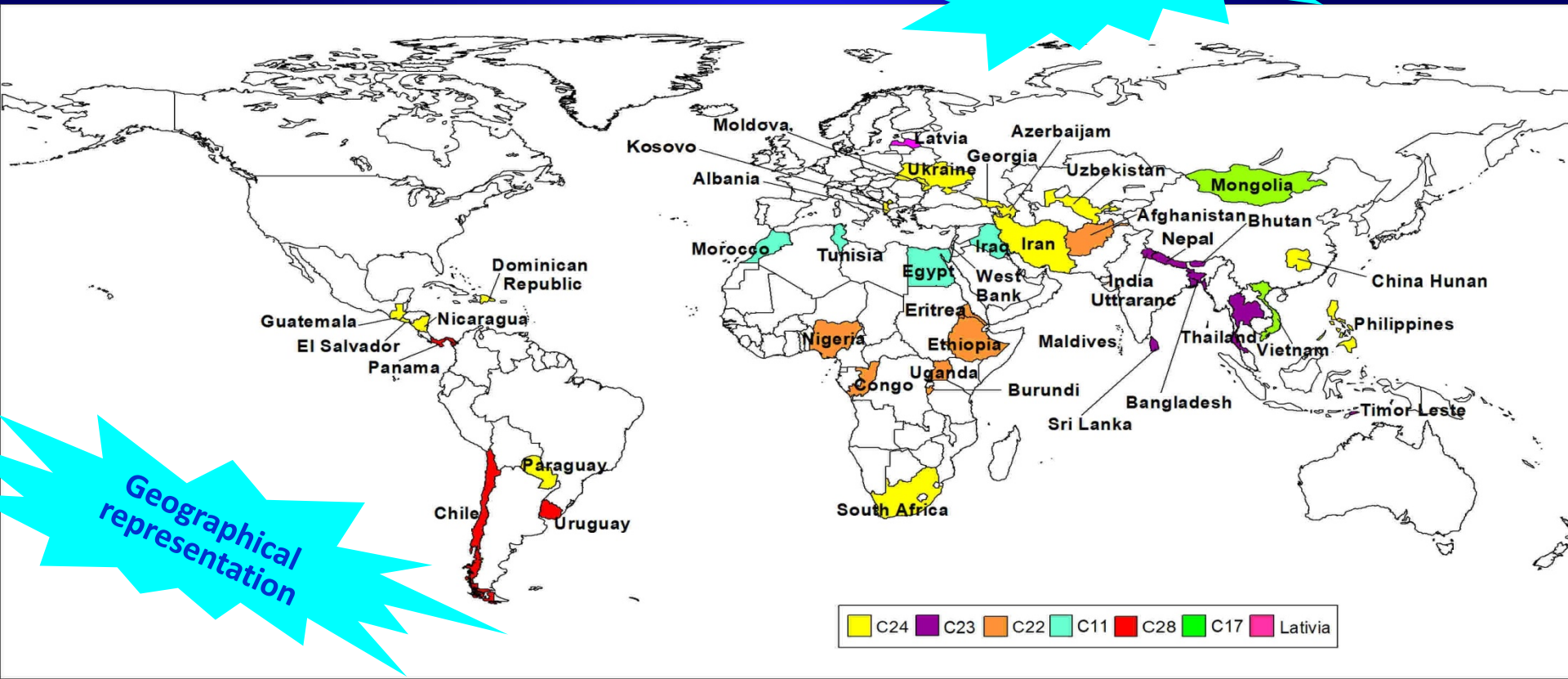


Gibert, K., and Cortés, U. (1997). "Weighing quantitative and qualitative variables in clustering methods." Mathware and Soft Computing, 4(3), 251-266.

Seven classes recommended

- **C22:** Afghanistan, Burundi, Congo, Eritrea, Ethiopia, Nigeria, Uganda
- **C24:** Albania, Azerbaijan, China-Hunan, Dominican-Repu, El-Salvador, Georgia, Guatemala, Iran, Kosovo, Moldova, Nicaragua, Paraguay, Philippines, South-Africa, Ukraine, Uzbekistan
- **C23:** Bangladeh, Bhutan, India-Uttaranc, Maldives, Nepal, Sri-Lanka,
- **C28:** Chile, Panama, Uruguay
- **C11:** Egypt, Iraq, Morocco, Tunisia, West-Bank

Burundi



Geographical representation

Local class means of numerical variables among the 256 variables

	CLASSE	C22	C24	C23	C28	C11	Latvia	C17
VARIABLE	$N = 42$	$n_c = 7$	$n_c = 16$	$n_c = 8$	$n_c = 3$	$n_c = 5$	$n_c = 1$	$n_c = 2$
totprofmh	$\bar{X}$	1.28	13.5507	5.1017	21.15	4.53	47.23	8.775
	S	0.9711	10.0276	6.0099	7.5041	2.6071		7.3468
	N*	0	2	2	0	1	0	0
treatpre	$\bar{X}$	192.22	1219.4614	59.7	1037.8201	547.0175	3490.75	1251.71
	S	121.1716	1447.8447	39.4424	519.8082	550.17	?	?
	N*	3	3	6	1	1	0	1
lundpararectrail	$\bar{X}$	1.09	0.9	0.49	1.2	1.1567	0.19	0.76
	S	0.7916	0.5622	?	0.0566	0.8864	?	?
	N*	3	7	7	1	2	0	1
comcarewor	$\bar{X}$	0.0314	0.0856	0.0197	0.0269	0.624	0.1991	0.1313
	S	0.0083	0.0196	?	0.0067			
	N*	5	10	7	1	4	0	1
usmhexperca	$\bar{X}$	0.2646	0.4961	0.2466	2.7995	0.4102	10.172	0.256
	S	0.5312	0.5059	0.2915	2.5926	0.2307		
	N*	0	1	1	0	1	0	1
d1f5i2exmhos	$\bar{X}$	0.7783	0.7954	0.7463	0.4963	0.5768	0.804	0.636
	S	0.2019	0.2121	0.1582	0.2051	0.1106		?
	N*	1	1	4	0	1	0	1

MIMMI *Method*

- Complex process
- Viable with an initial kernel of variables with few missing values

- Horizontal inputation:
  - use the value of other variables of the same
  - individual as predictors of the missing value.
  - *inputing 0 in the income of 4th person if the household has only 1,2 or 3 persons*
- Vertical inputation:
  - use the value of the same variable in other similar individuals
  - *use the mean of the salary of 4rt persons over 18 years old if the household hasmore than 4per*

MICE method

[vanBuuren2010]

multiple imputation by chained equations

Multiple imputation (MI):

- Replace missing values with *plausible* substitutes
  - Distribution-based
    maximum-likelihood based Markov-chain Monte Carlo (MCMC)
  - Inject *the right amount of randomness* to reflect uncertainty
- Repeat *m* > 1 times to procude m imputed datasets
- Analyse datasets individually, but identically
- Combine the models, get confidence intervals using Rubin's rules (`micombine`)

Context:
Multiple regression
In general

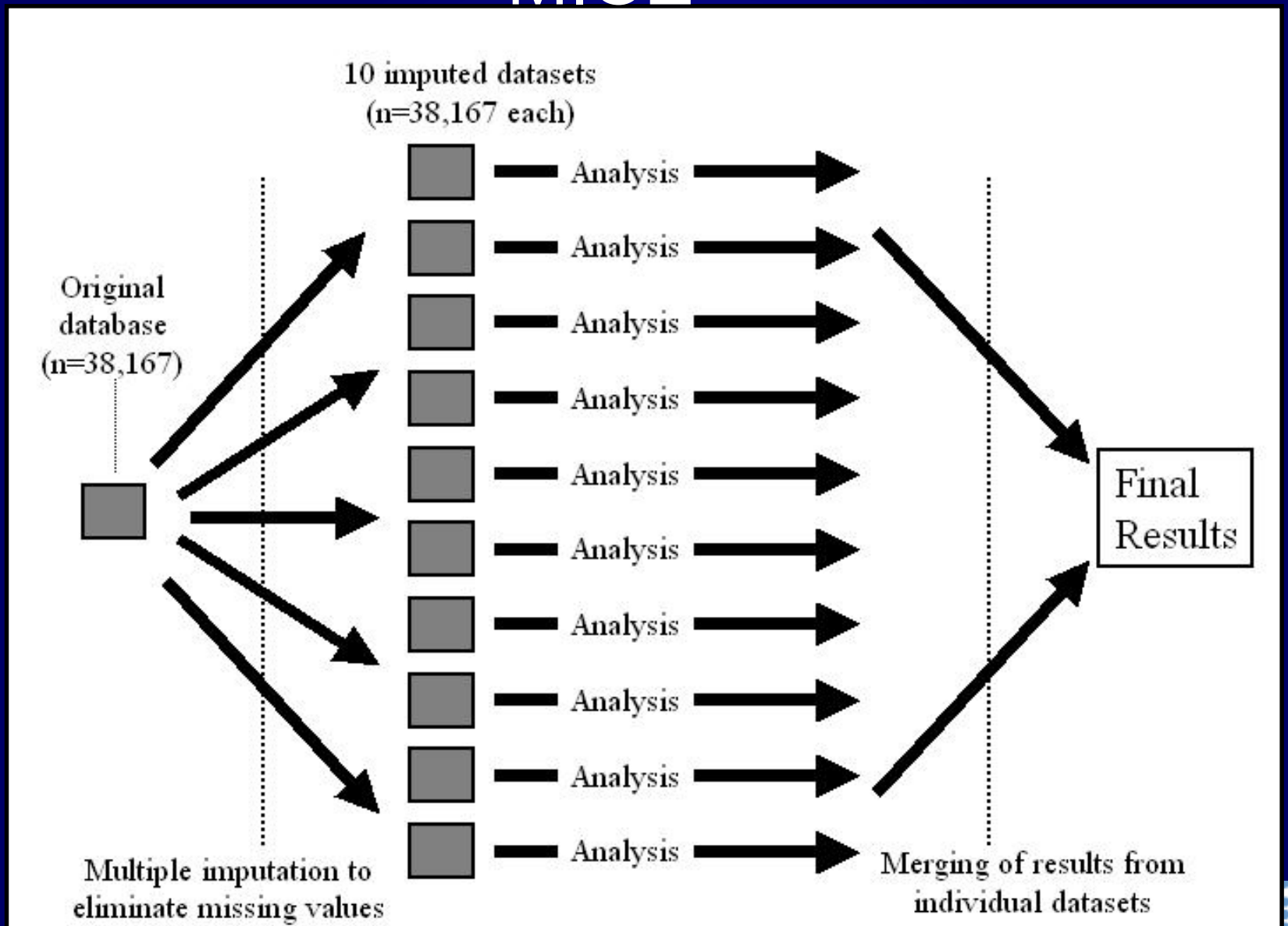
The MICE approach has three components:

- Univariate – implemented in `uvis`
- Multivariate – implemented in `ice`
- Multiple – implemented in `mice`

- `ice` = imputation by chained equations

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.

MICE


10 imputed datasets
(n=38,167 each)
Original
database
(n=38,167)
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Final
Results
Multiple imputation to
eliminate missing values
Merging of results from
individual datasets

MICE

The overall *estimate of your parameter* (Q-bar) is its mean across the m imputations

$$\bar{Q} = m^{-1} \sum \hat{Q}^{(\ell)}$$

The *within-imputation variance* (U-bar) of the Q parameter is the mean of the variances across the m imputations

$$\bar{U} = m^{-1} \sum U^{(\ell)}$$

The *between-imputation variance* (B) of the Q parameter is standard deviation of Q across the m imputations

$$B = (m-1)^{-1} \sum (\hat{Q}^{(\ell)} - \bar{Q})^2$$

The *total variance* of Q is a function of U-bar and B. This total variance is used to calculate the standard error used for test statistics

$$T = (1 + m^{-1})B + \bar{U}$$

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu$$

The *degrees of freedom* (v) are adjusted for the amount of information lost to missing data

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2$$

MICE

- MICE method is very flexible – but demands thought when creating the imputation model
- Strongly recommend mastering the `eq()`, `passive()` and `substitute()` options
- Can deal with interactions using `passive()`
- Choice of *m* is important
  - may need to be (much) larger than 5
  - See Royston (2004, SJ 4:227-41) for discussion
- available in MICE Rpackage

Easy for MAP-REDUCE strategies

WARNING reproducibility

WARNING final model consensus

Royston, P. (2004). Multiple imputation of missing values. *Stata journal*, *4*(3), 227-41.

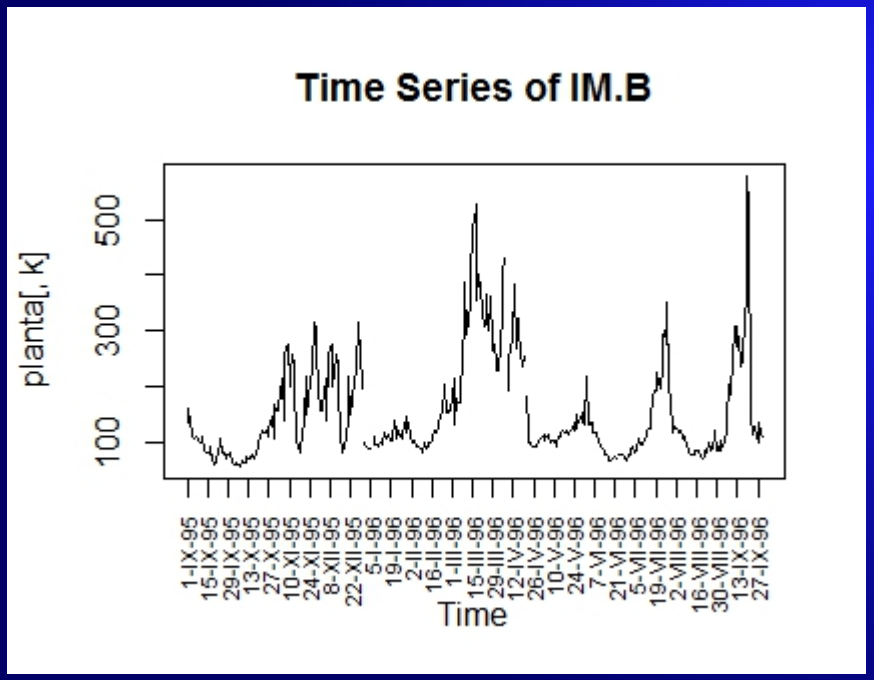
Interpolation

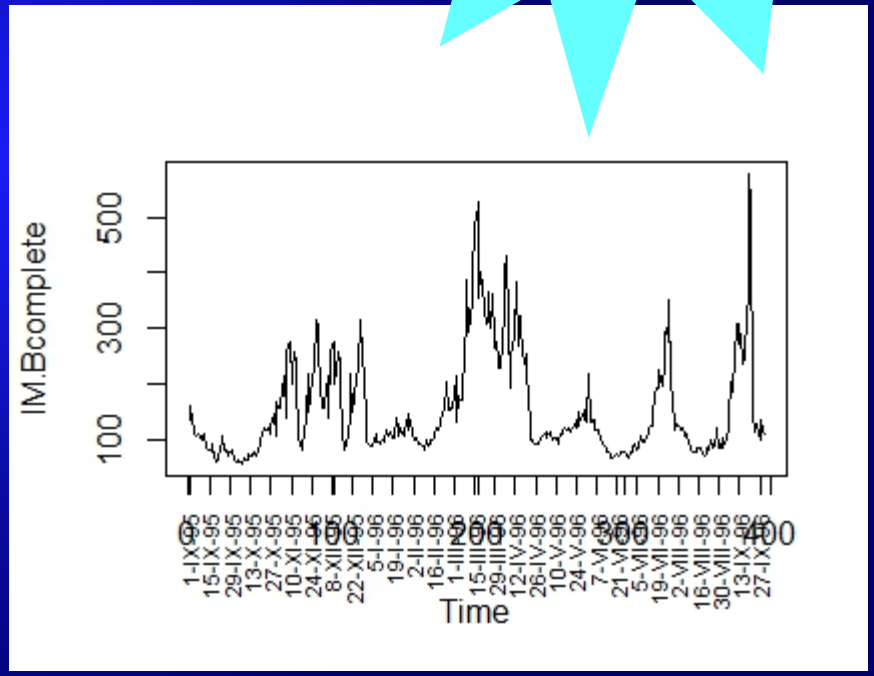
Usefull for time-series

- Linear assumption between observed points

(assume monotonic behaviour between observations)

ALTERNATIVE
Assume constant between Measurements (slow dynamics)





IM.B: Mass index of mixed liquor (na.approx {zoo})

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Outlier

- Rare observation (presumed out of range)
- Multivariate vs univariate outlier

BIAS

- Types of outliers:
  - Mistake (Transcription Error or Measurement Error)
    - A person 560 years old
    - FIRST VERIFY *If possible correct*.
      - » *If not, substitute by missing*
  - Informative point
    - A single informative point of a missing part of the population
    - *Complete the sample*
      - *when impossible, restrict scope of analysis*
  - Extreme value of the population
    - Very old person, 99 years old
    - *Keep*

Care with suppressions

  - Value of another population
    - One swedish in the middle of cannibal tribu, measuring height
    - *Treat apart. CLEARLY REPORT ABOUT IT*
  - Missing code
    - *Substitute by missing or inpute*

The danger of suppressions

- *In 1985 British scientists reported a hole in the ozone layer of the earth's atmosphere over the South Pole. This is disturbing, since ozone protects us from cancer-causing ultraviolet radiation. The British report was at first disredarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software used o analyze the data had automatically syppressed these values as erroneous outliers. Readings dating back to 1979 were reanalized and showed a large and growing hole in the ozone layer that is unexplained and possibily dangerous Computers analyzing large volumnes of data are often programmed to suppress outliers as protection againts errors in the data. As the example of the hole in the ozone layer illustrated, suppressing an outliers without inv estigating it can kleep valuable information out of the sight*

Moore, McCabe, Introduction to the practice of Statistics, 5th Edition, Freeman

From the paper of John Gleick in New York Times, July 1985

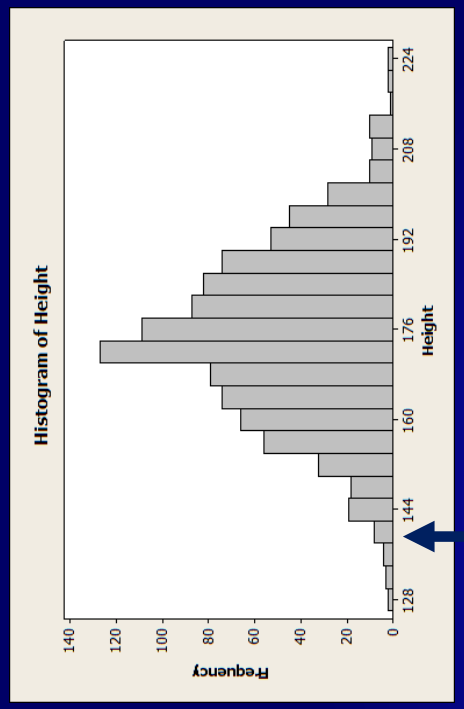
http://www.nytimes.com/1986/07/29/science/hole-in-ozone-over-south-pole-worries-scientists.html?pagewanted=all

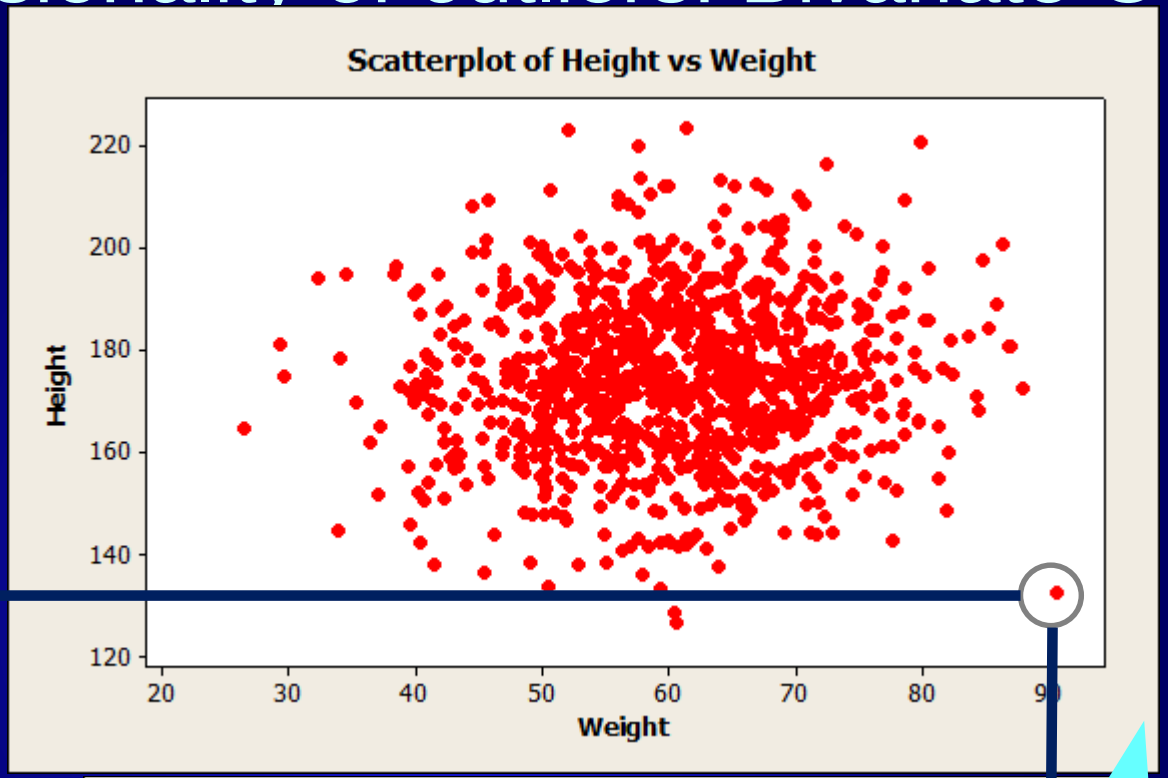
Outlier detection

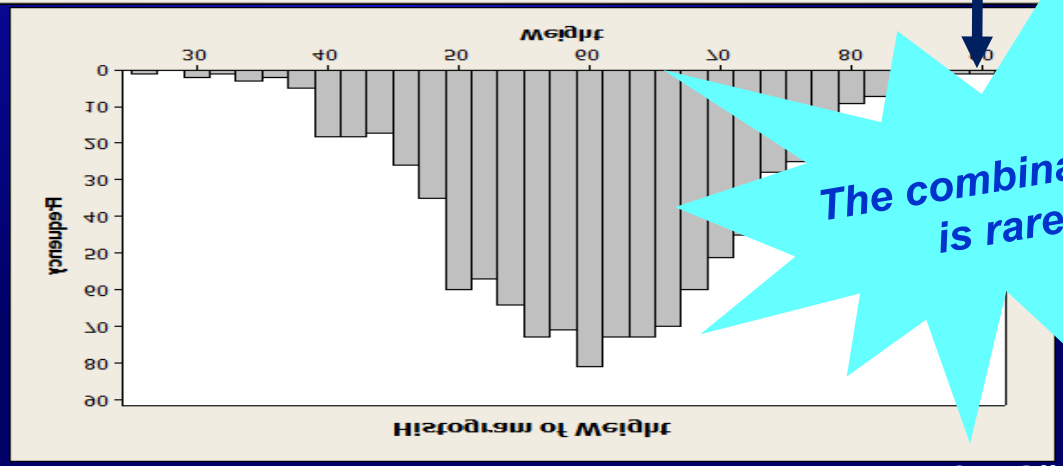
- Specific statistical Tests:
  - Depend on the software
  - Usually for specific distributions

- Graphical representation of the distribution of data
  - Univariate (histogram or boxplot)
  - Bivariate (plots)
  - Clustering for multivariate outliers (singleto[illegible])

Dimensionality of outliers: Bivariate Outlier



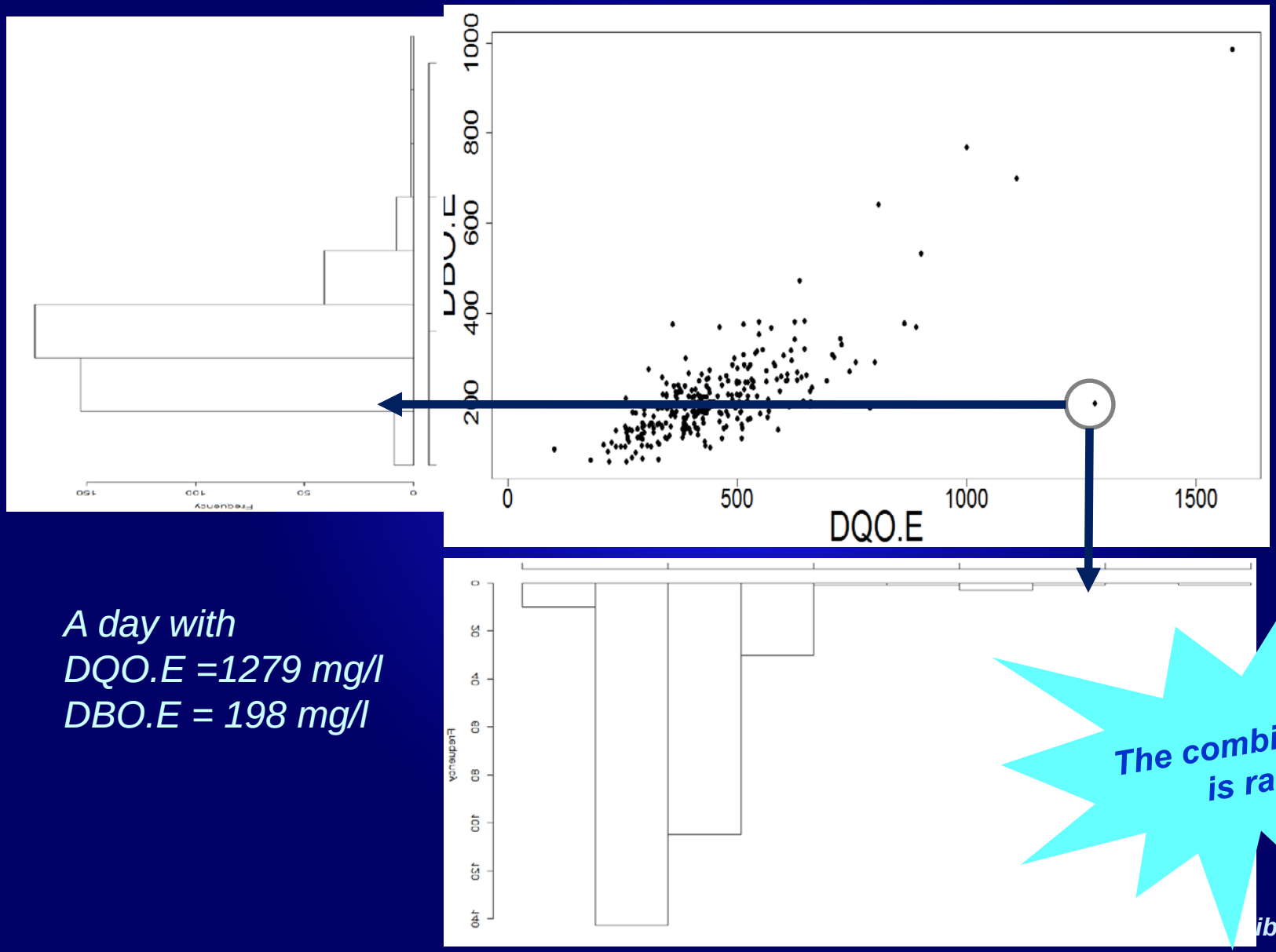




A person with 90Kg and 1,32 m

The combination is rare

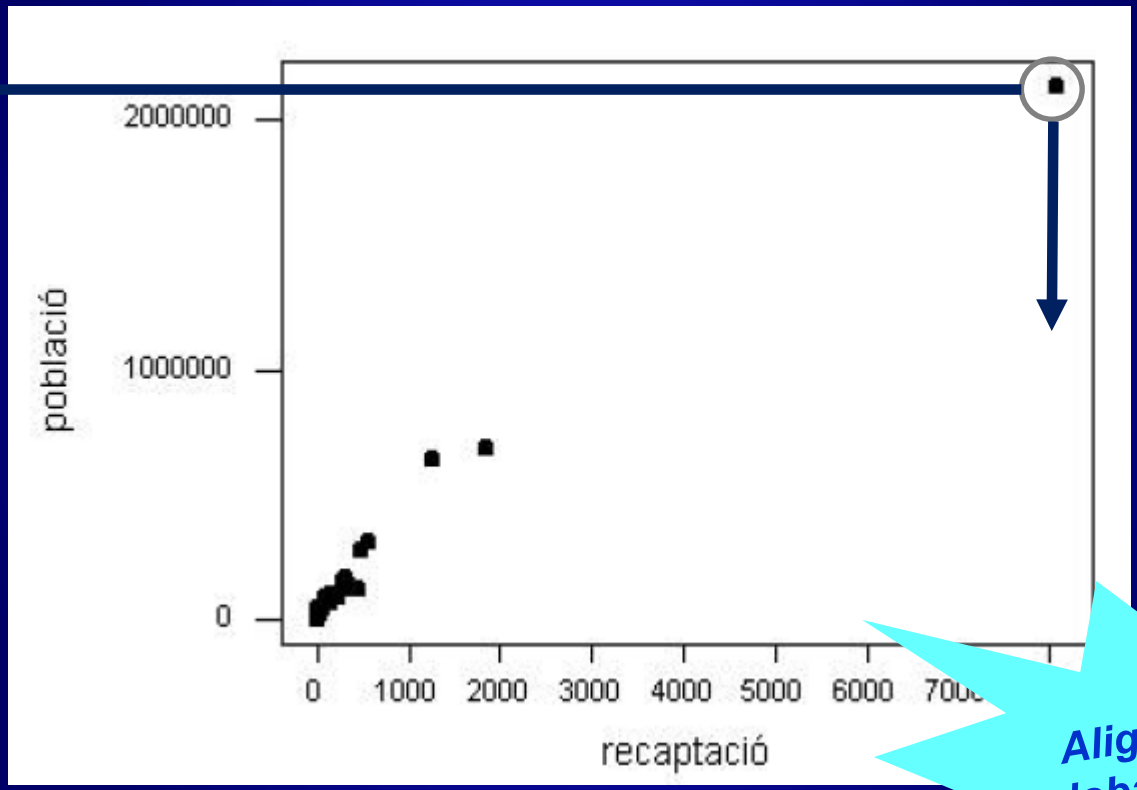
Dimensionality of outliers: Bivariate Outlier



A day with
DQO.E =1279 mg/l
DBO.E = 198 mg/l

The combination is rare

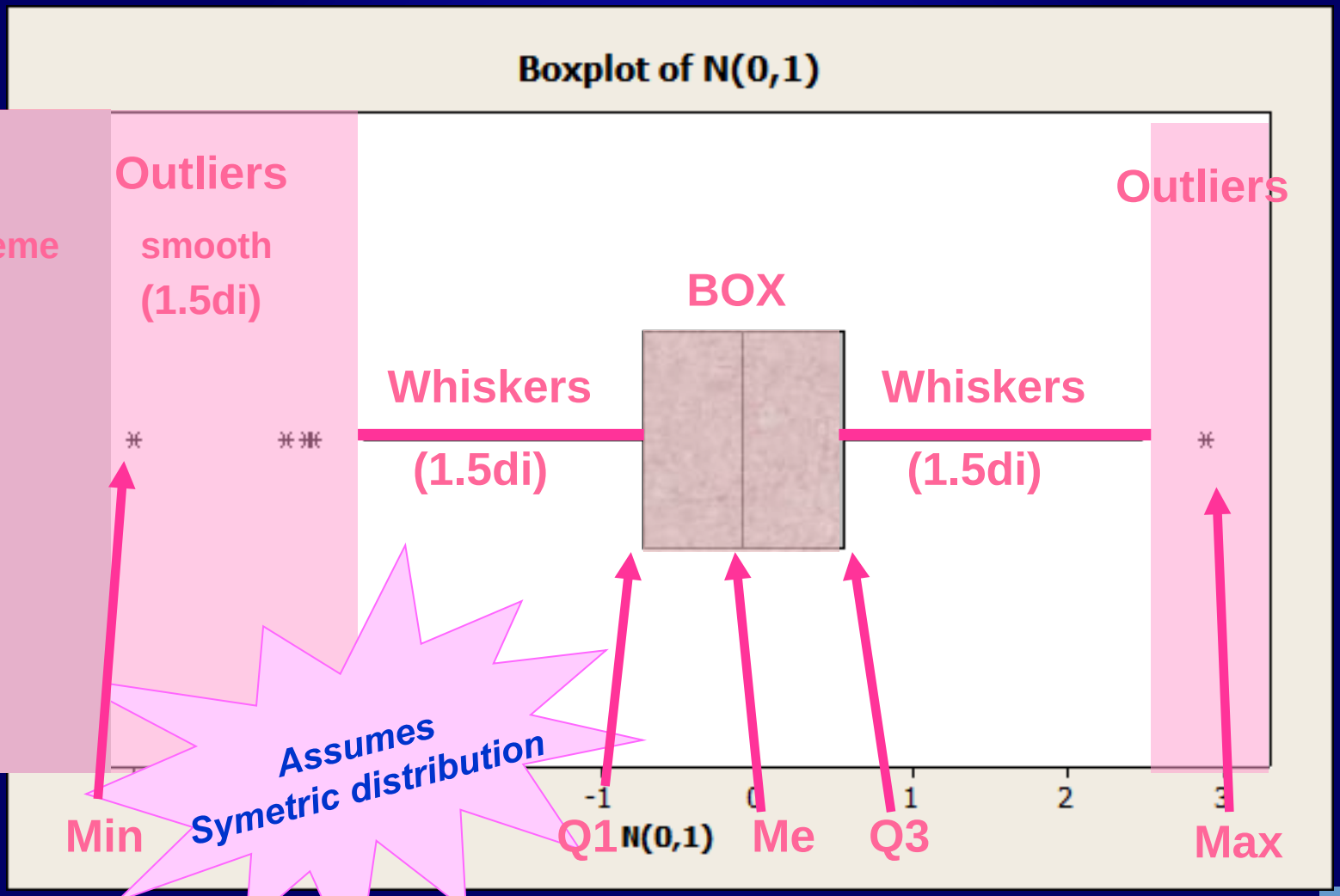
Dimensionality of outliers: Univariate Outlier



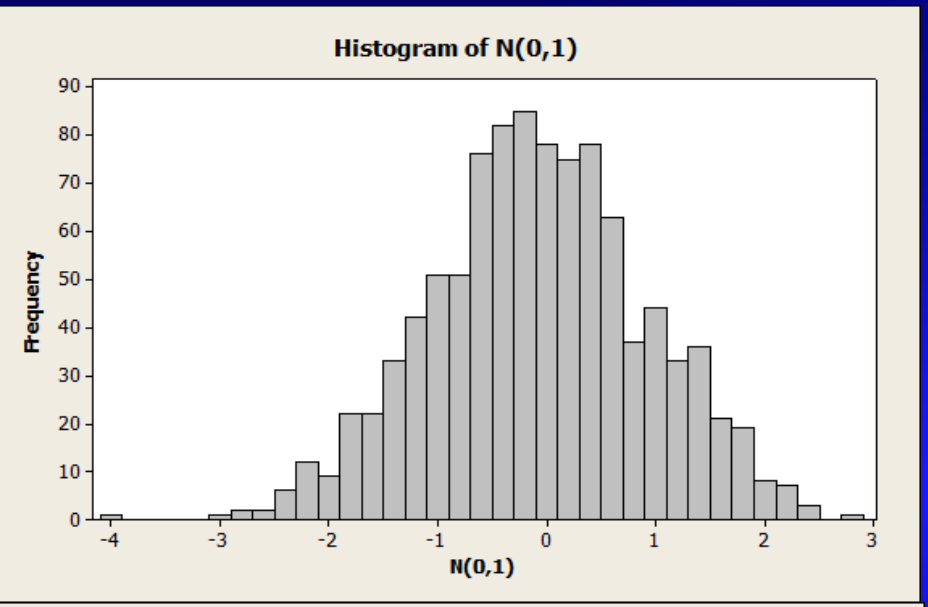
Aligned with global model!!!!

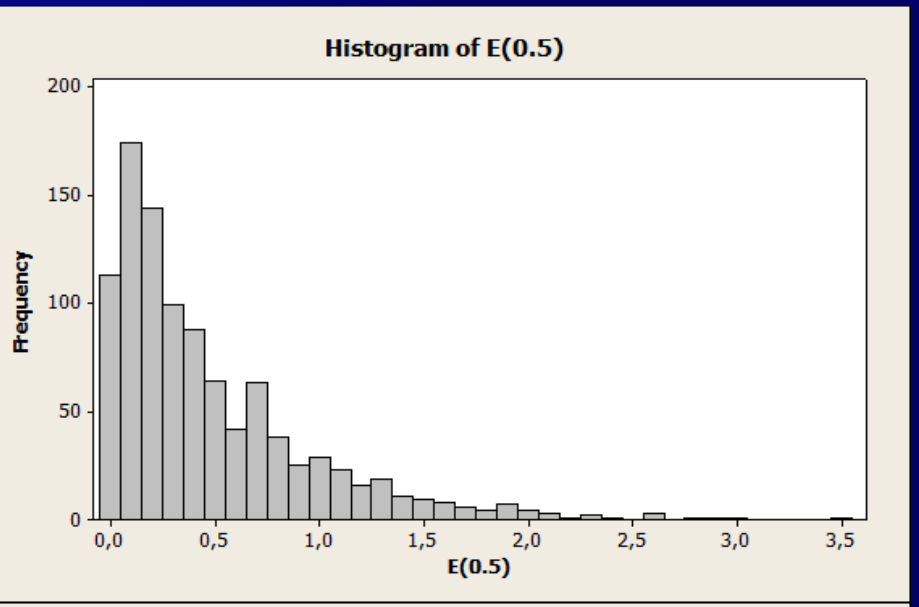
Boxplot *[Tukey 1956]*

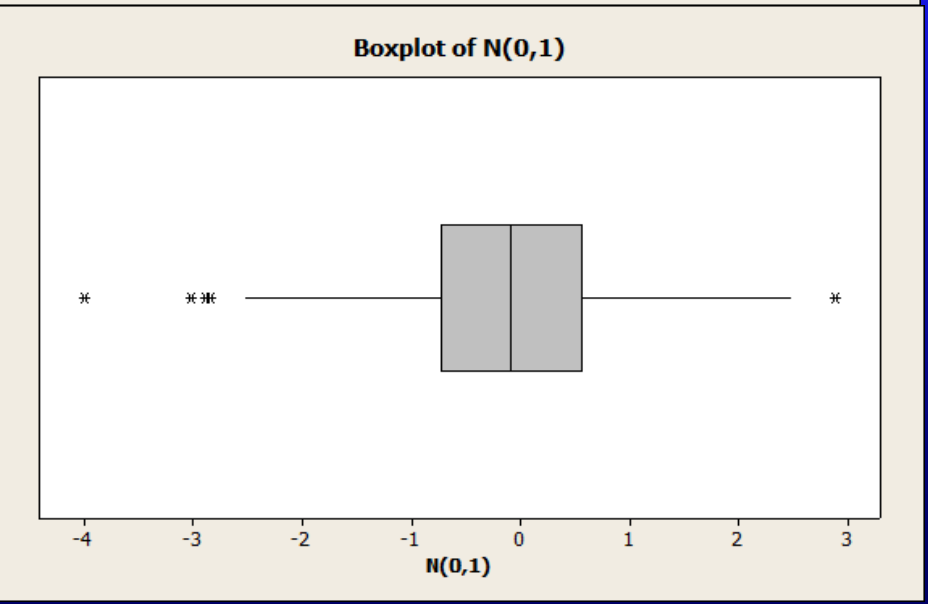
- *Symbolic representation of empirical distribution*

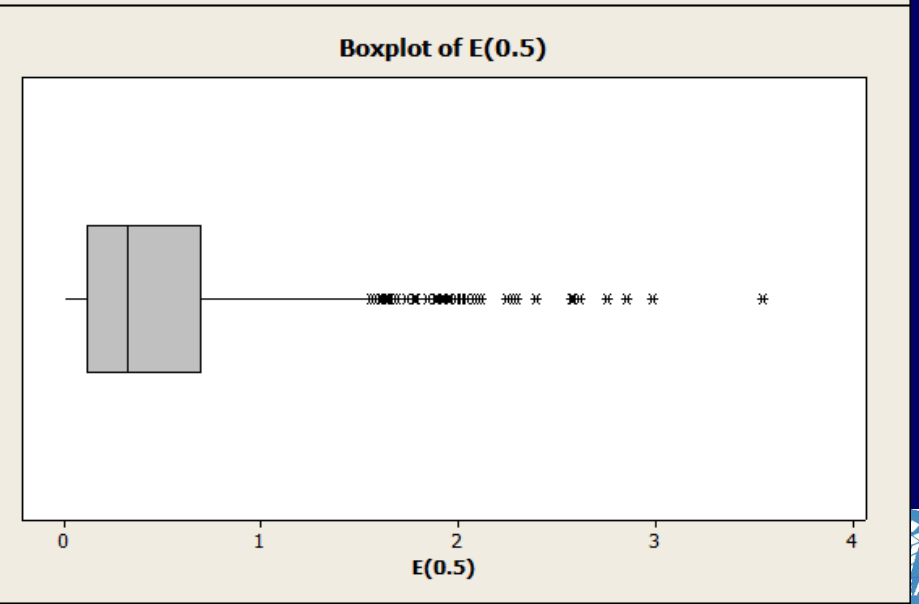


Boxplot [Tukey 1956]









Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Instance selection

- Evaluation of representative instances in a dataset
  - Elimination of irrelevant instances
  - Sampling
  - Resampling

- Reparing unbalanced datasets when required
  - oversampling
  - undersampling

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Feature selection

- Evaluation of relevant variables in a dataset
- Priorization and ranking under different criteria
  - Feature weighting (determine weights of variables in the analysis)
- Elimination of irrelevant variables
  - Feature selection

- Feature selection
- IA methods
- Statistical Feature selection: use statistical test for ranking
- Sometimes just use threshold on feature weighting ranks

Feature selection

- Goal: discard non-interesting variables
- Reduce data dimensionality
- Eliminate noise and redundancies
- Improve performance of algorithms
- Avoid spurious relationships in models
- Reduce curse of dimensionality
- Requires a response variable to be explained Y

- Rank relevance degree of Y wrt all other variables
- Discard less relevant

Statistical Feature selection

Guyon, I. (2008). Practical feature selection: from correlation to causality. NATO science for peace and security, 19, 27-43.

– Hypothesis test:

H_0 : There is no relation between the y and x

H_1 : There is a relation

- Get p-values for the dependence between Y and X
- Lower p-values imply strongest dependence
- Rank variables by ascending p-values
- Discard irrelevant variables (threshold over p-values)

- Specific tests depends on type of variables analyzed

Statistical Feature selection

– Hypothesis test:

- Y numerical
  - X numerical: Correlations test / Sheffer generalized coefficient
  - X qualitative: F test /Kruskal-Wallis
- Y qualitative
  - X numerical: F test/Kruskal-Wallis
  - X qualitative: chi-2 test

Feature selection

- Evaluation of relevant variables in a dataset
  - Priorization and ranking under different criteria
    - Feature weighting (determine weights of variables in the analysis)
  - Elimination of irrelevant variables
    - Feature selection

- Feature selection
  - IA methods (based on information theory)
  - Statistical methods (based on statistical tests)
  - Sometimes just use threshold on feature weighting ranks

AI Feature Selection

- **Wrappers:**

Rank subsets of features by accuracy in predicting Y (costly. Method specific oriented)

- **Filters:**

Rank subsets of features by some proxy measure (mutual information, statistical signifficance, Relief method)

- **Embedded methods**:

Implicit feature selection as part of the modelling algorithm, that penalizes less efficient variables internally (LASSO)

Relief

Score features

- For each feature X
  - For each object i
  - Find nearest neighbour from class A (j) and B (l), upon feature X
  - The score of the feature increases proportional to the distance of the neighbour in the same class of i and decreases proportional to the distance of the neighbour in the other class

    wk=wk+ d(i,j)-d(i,l) or wk=wk+d(i,l)+d(i,j) depending on class of I

  - Average final score modified by all objects

- Sort all X according to final scoring

Kira, Kenji and Rendell, Larry (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. AAAI-92 Proceedings.

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Variables Transformation

- Homogeneization
- Approaching to methods hypothesis
- Getting more interpretability

Variables Transformation

- Data cleaning reasons
  - Measurement units of Thyroids hormones from different laboratories

–1993
– Collaboration UPC, Barcelona, Spain
– Andrija Stampar School of Public Health, Zagreb, Croatia
– Setre Milordsnice Clinical Hospital, Zagreb, Croatia
–
–Find patterns of thyroids dysfunctions 1002 patients, 12 measurements

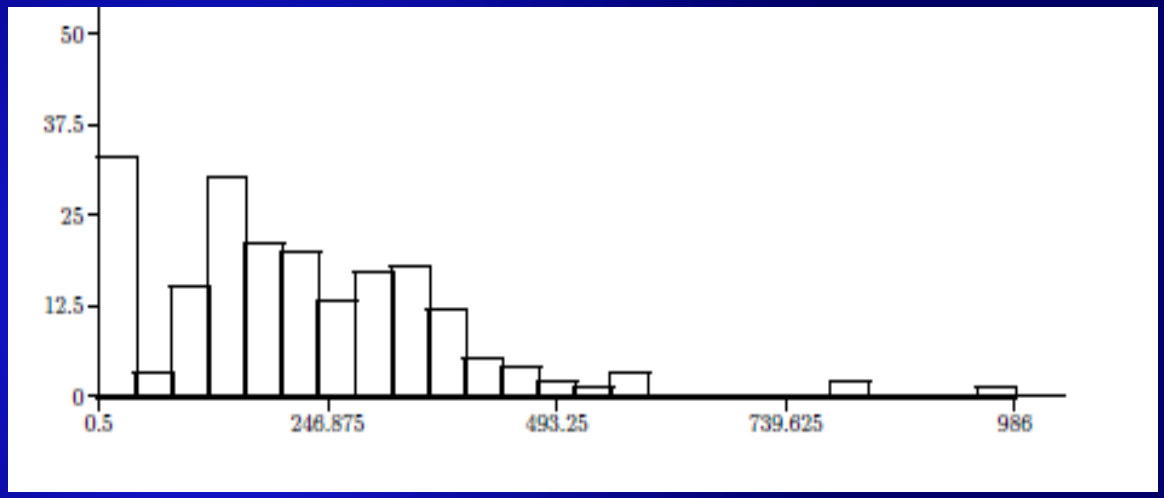
–2013
–Collaboration UPC, Atención Primaria ICS
– http://www.sidiap.org/

–Laboratory measuments in TSH

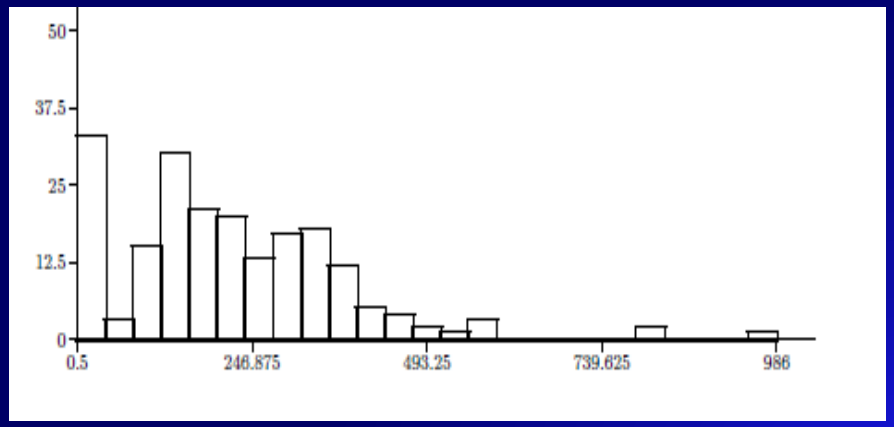
Laboratory Tests measurements

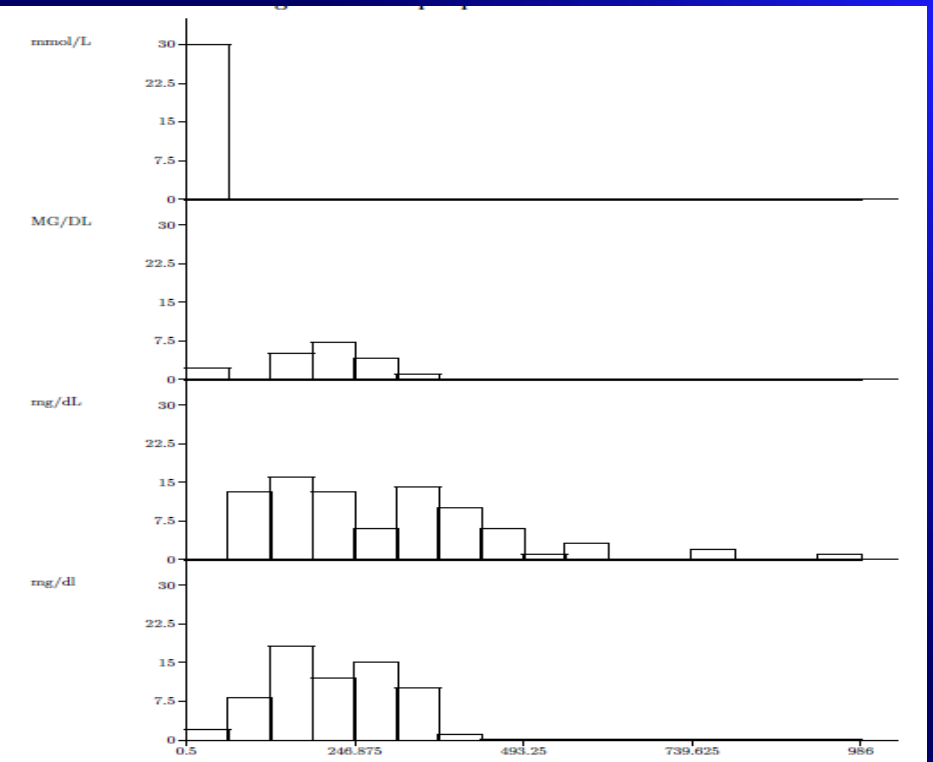
Measurement of Total Cholesterol from 200 pacs from Catalan Public Health System in 2013 (Primary Care)

Summary Statistics	
Number of objecs	200
Number of missing values	0
Number of useful values	200
Mean	213.2037
Median	193.9
First Quartile (Q1)	115.8
Third Quartile (Q3)	306.2
Minimum	0.5
Maximum	986
Quasi-standard deviation	156.4218
Variation Coeficient	0.7318



Frequency Table		
Modalities	Freq. absol.	Freq. relat.
mg/dl	66	0.33
mg/dL	85	0.425
MG/DL	19	0.095
mmol/L	30	0.15
missing data	0	0





Unitat	Count
(mg/dL)	12
0.0 - 240.0	1
55519	1
G/L	4
mg/dl	50
mg/dL	50
MG/DL	50
mg/dl^MMOL/L.	50
mmol/L	50
Mmol/L	1
MMOL/l	1
MMOL/L	1
mmol/L^mg/dL	50
N=	321

mmol/l = 38,669 mg/dl

Variables Transformation

- Data cleaning reasons
  - Measurement units of Thyroids hormones from different laboratories
  - Refer the whole set of variables to comparable units
    - *all concentration variables in mg/l*
    - *proportions instead of absolute numbers, ...*
  - Coertions: Information loss.
    - Discretization (h/week working)
    - Categorization (Thiroids levels)
    - Recategorizations (professions)

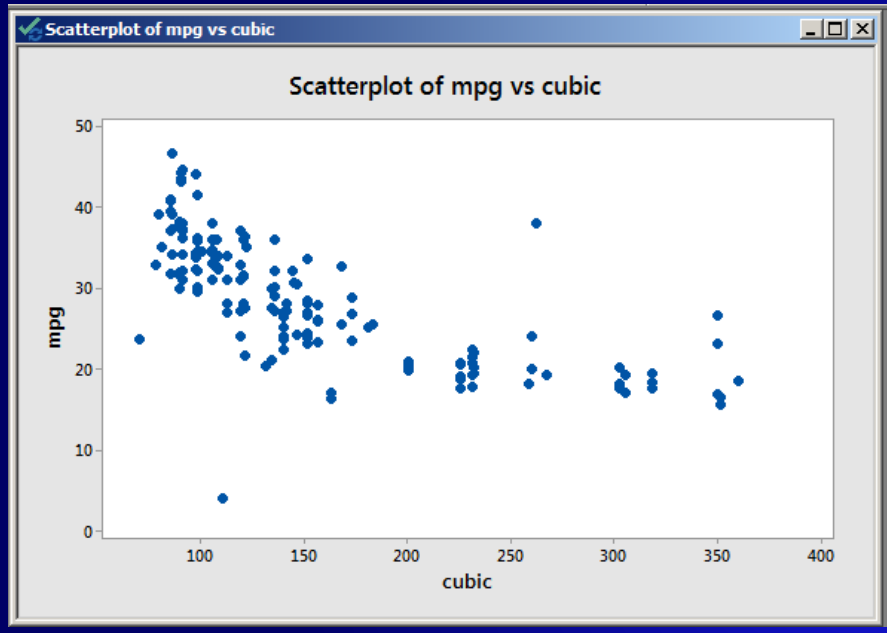
Better avoid

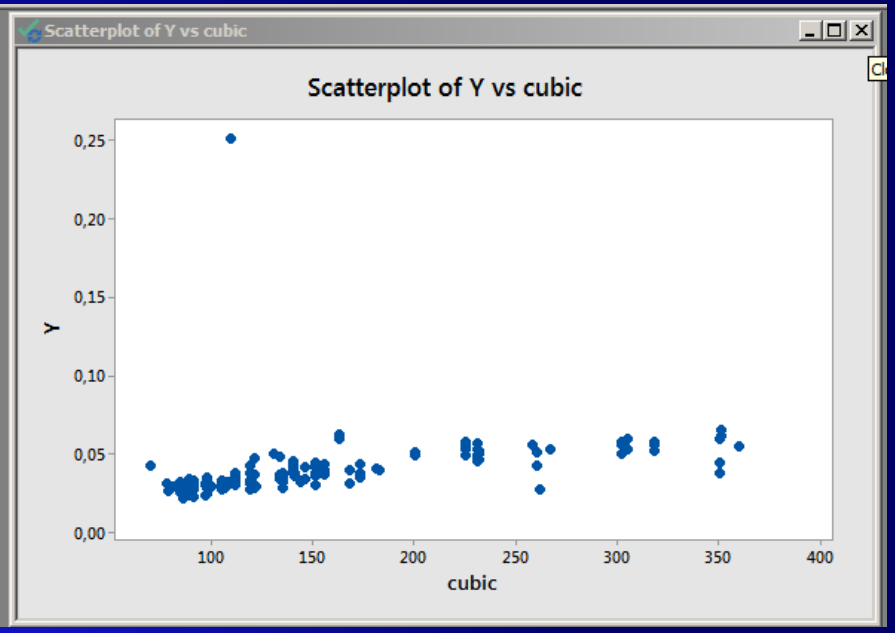
- Technical questions:
  - Estanderditzation, normalitzation o linealirization
  - Eventual logaritmic transformation
  - Required by data mining technique to apply

Select a technique respectfull with original data

Exceptional situations

where transforms make sense





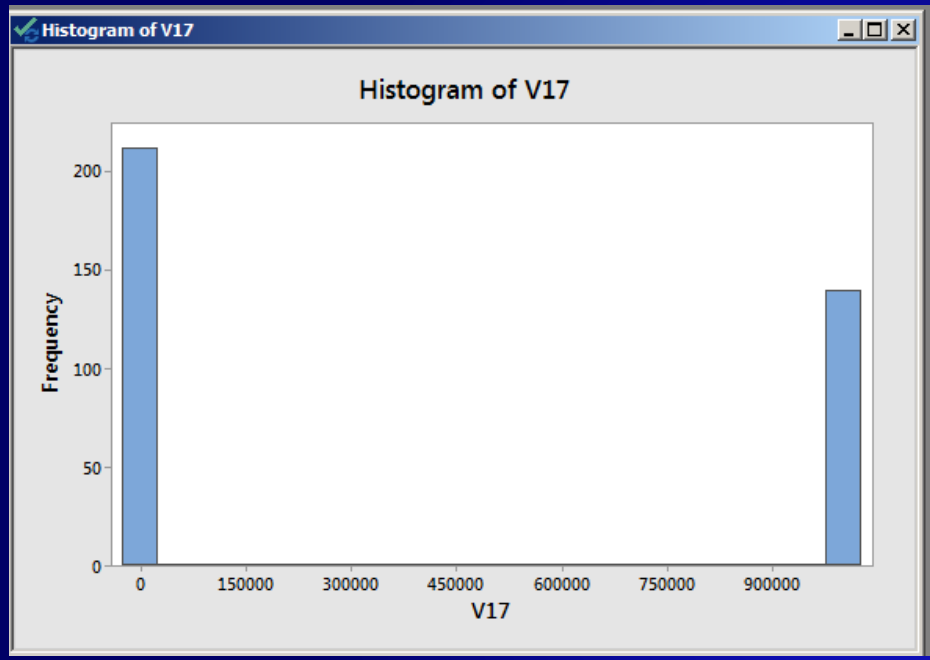
- Mpg: miles per gallon of a car
- Cubic: cubic capacity of the car engine
  Non linear relationship (regression non suitable)

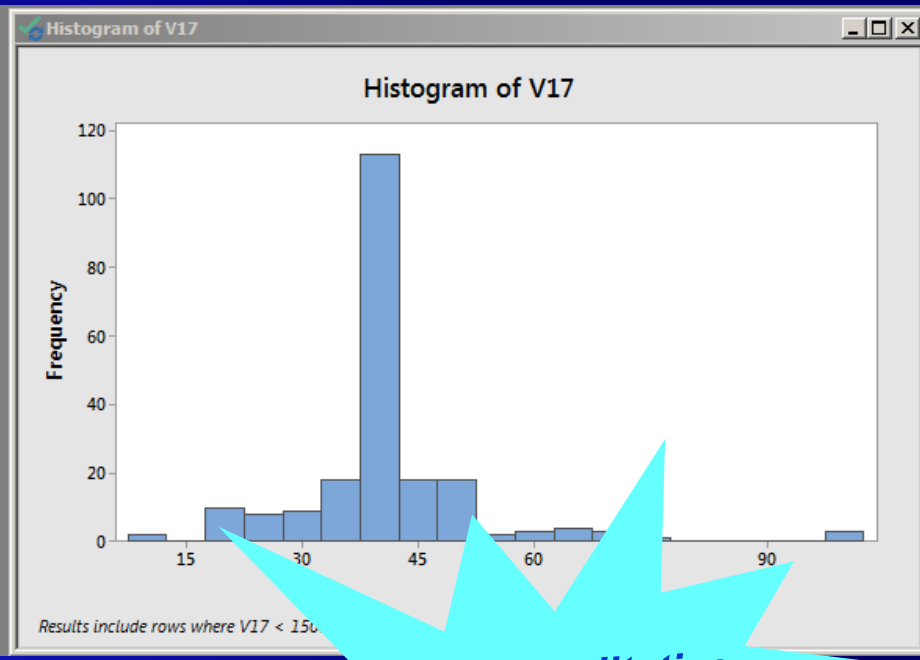
- Y = 1/mgp : Linearizes the relationship

Y is car Consumption!!!!

Exceptional situations

where transforms make sense





- Hours working per week
- 3-modal:
  - Arround 20 h/w
  - Arround 40 h/w
  - Arround 65 h/w

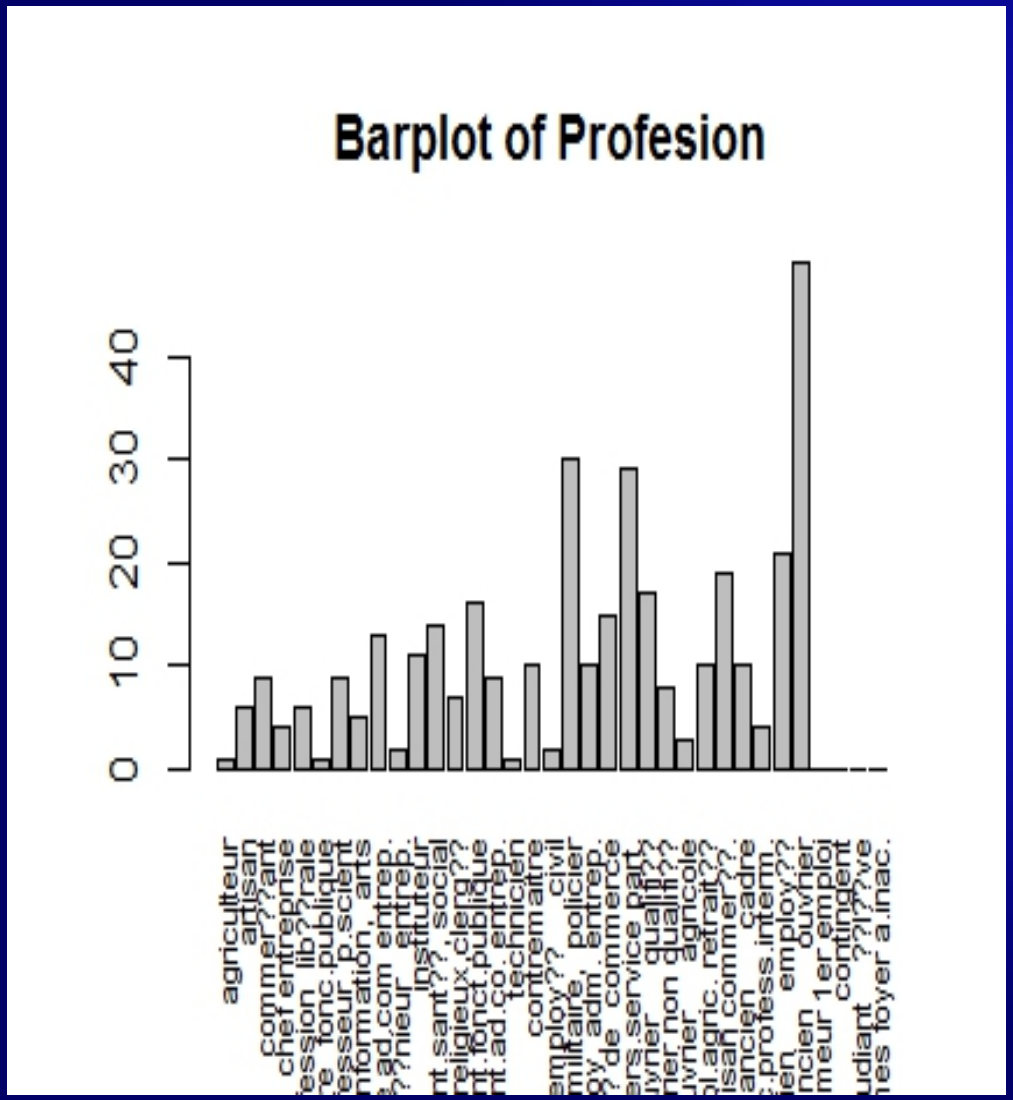
- Correspondence with part-time, full-time, extra turns works

Build a qualitative variable: Type of work (part-time, full, turn)

Exceptional situation

where transforms make sense

Regroup professions in a more general families



- Professions: 31 modalities unmanageable
- Families of professions:
  - agriculteur
  - ouvrier agricole
  - expl.agric. Retraitée

  Agriculture sector

  - Artisan
  - anc.artisan commerce
  - information, arts
  - commerçant
  - employée de commerce

  Arts and commerce

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Derivation of new variables

- Aggregates (additions of other variables)
  - Total household income

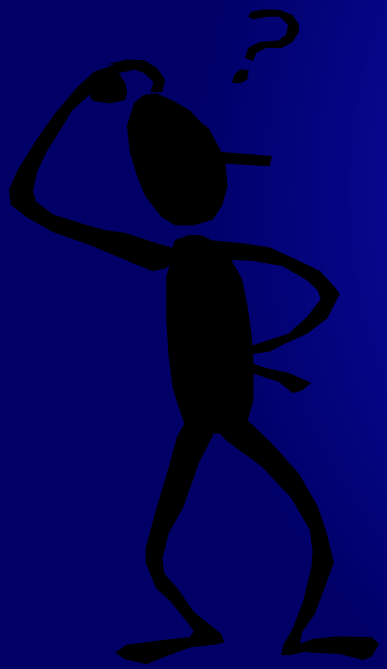
- Synthetic indicators
  - Classical generation of global score in psychometric scales
  - Indicators
  - *(Lund parameter =external contacts/days hospital indicator of "development of a health system")*
  - *Case Credit Scoring (saving capacity)*

Input missingsPreviously According to operation

- Binary indicators
  - *If condition regarding a combination of values*
  - *then indicator=1, else the indicator=0*

- *Dimensionality reduction techniques*

Datos, Descriptiva y Pre-processing



Karina Gibert

Dpt. Statistics and Operation Research

Knowledge Engineering and Machine Learning Research group at
Intelligent Data Science and Artificial Intelligence Specific Research Center

Institut Universitari de Recerca en Ciència y Tecnologia de la Sostenibilitat
Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu
www.eio.upc.edu/homepages/karina

Are there any questions?...