Basic Descriptive Analysis

Numerical Variables

K. Gibert

Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group

Universitat Politècnica de Catalunya, Barcelona

karina.gibert@upc.edu

www.eio.upc.edu/homepages/karina

Guió

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

0. Descriptive analysis

Compact and Informative view of the variable structure

DATA= FIT+ ERROR

Characterizacion

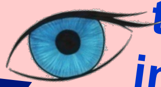
General Pattern **Deviations**

Structural Component Random Component

Tools

1. Graphical

   Visualitze variable’s distribution



the best inductor

2. Numerical

   Quantify what is observed in he graphs

Objectivate

© K. Gibert

1. Graphical tools

1. Performing the graph

   Mechanical (software)

2. Reading the graph

   Technical (statistitian or data miner)

3. Interpretation

   Conceptual (domain expert)

   Contextualization

Graphical tools

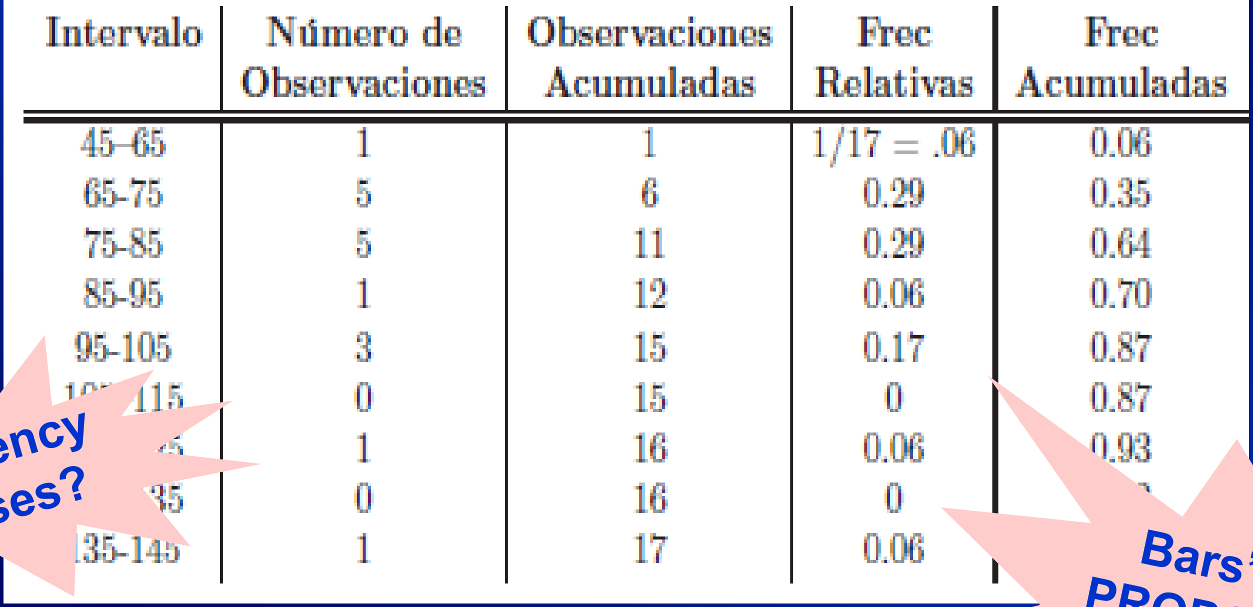
for numerical variables

1. Histogram
2. Boxplot

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

Histogram

Visualitzation of frequencies distribution table



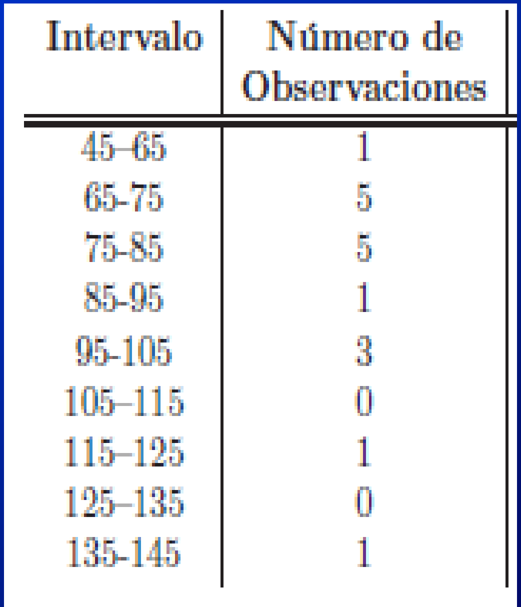
Intervalo	Número de Observaciones	Observaciones Acumuladas	Frec Relativas	Frec Acumuladas
45–65	1	1	1/17 = .06	0.06
65-75	5	6	0.29	0.35
75-85	5	11	0.29	0.64
85-95	1	12	0.06	0.70
95-105	3	15	0.17	0.87
[illegible]115	0	15	0	0.87
[illegible]	1	16	0.06	0.93
[illegible]35	0	16	0	[illegible]
[illegible]35-145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

Frequency Classes?

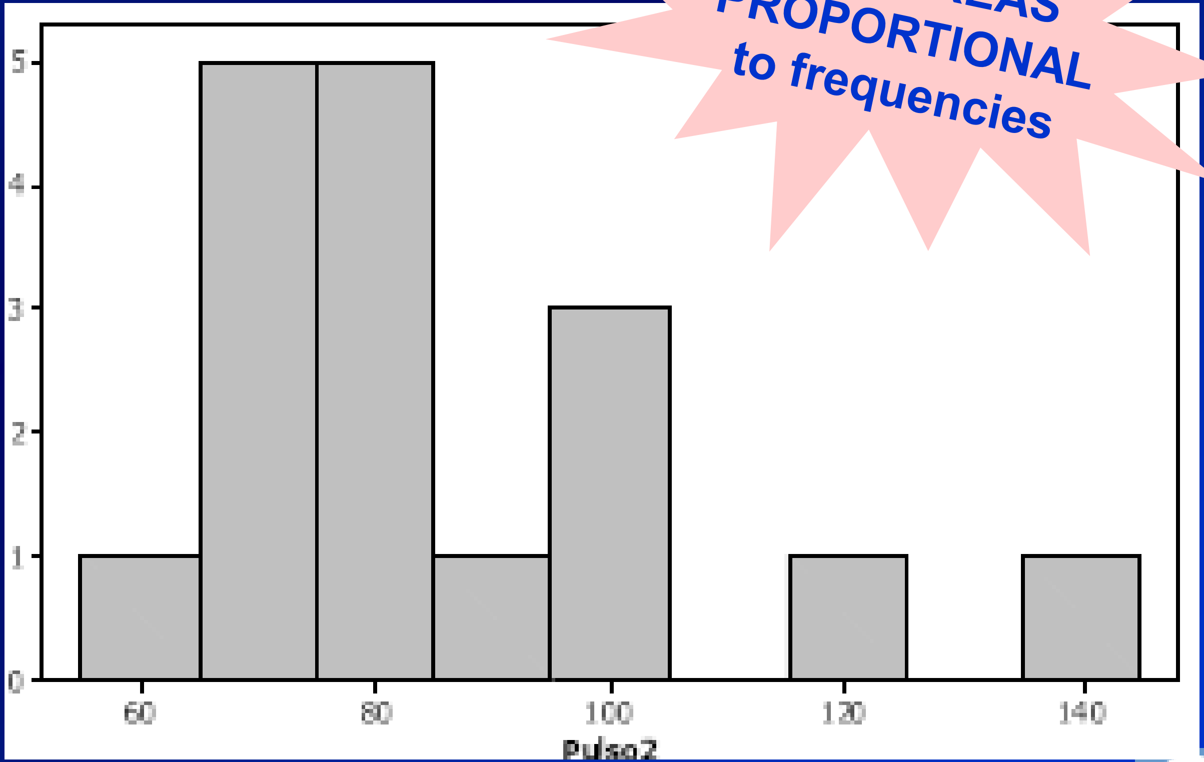
Bars' AREAS PROPORTIONAL to frequencies

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

Histogram



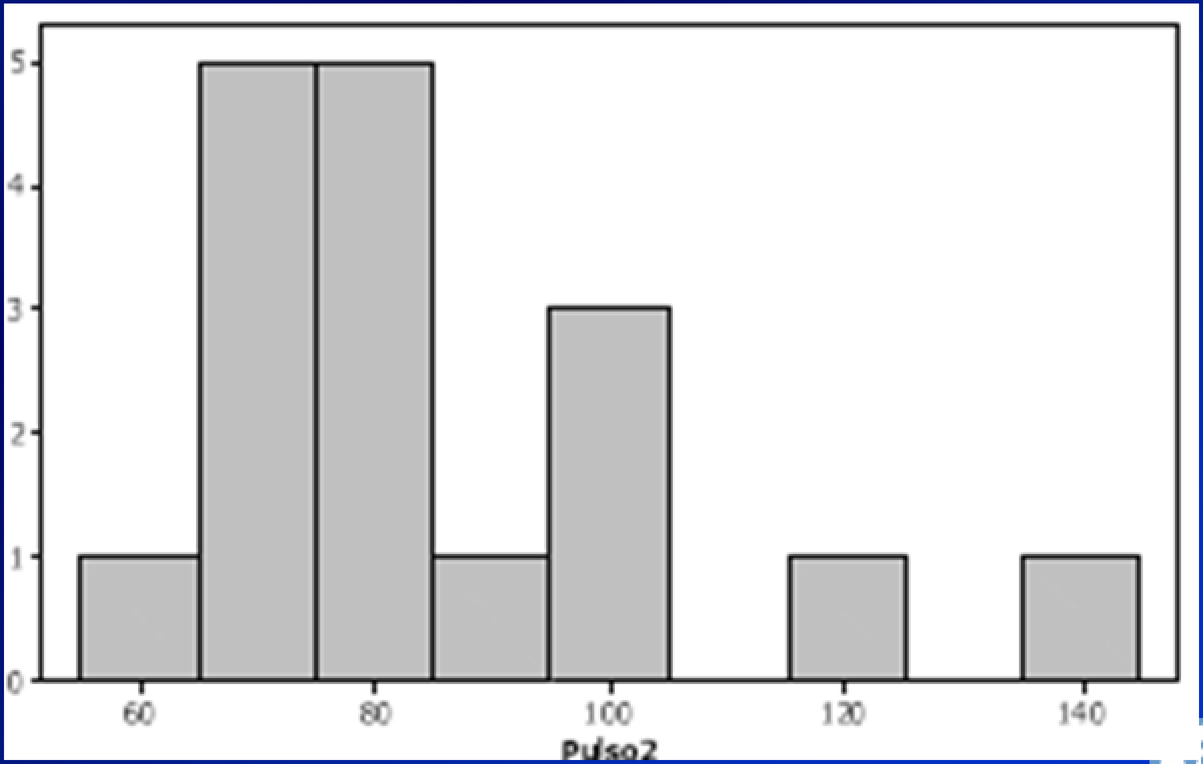
Intervalo	Número de Observaciones
45–65	1
65-75	5
75-85	5
85-95	1
95-105	3
105–115	0
115–125	1
125–135	0
135-145	1



Bars' AREAS PROPORTIONAL to frequencies

READING HISTOGRAMS

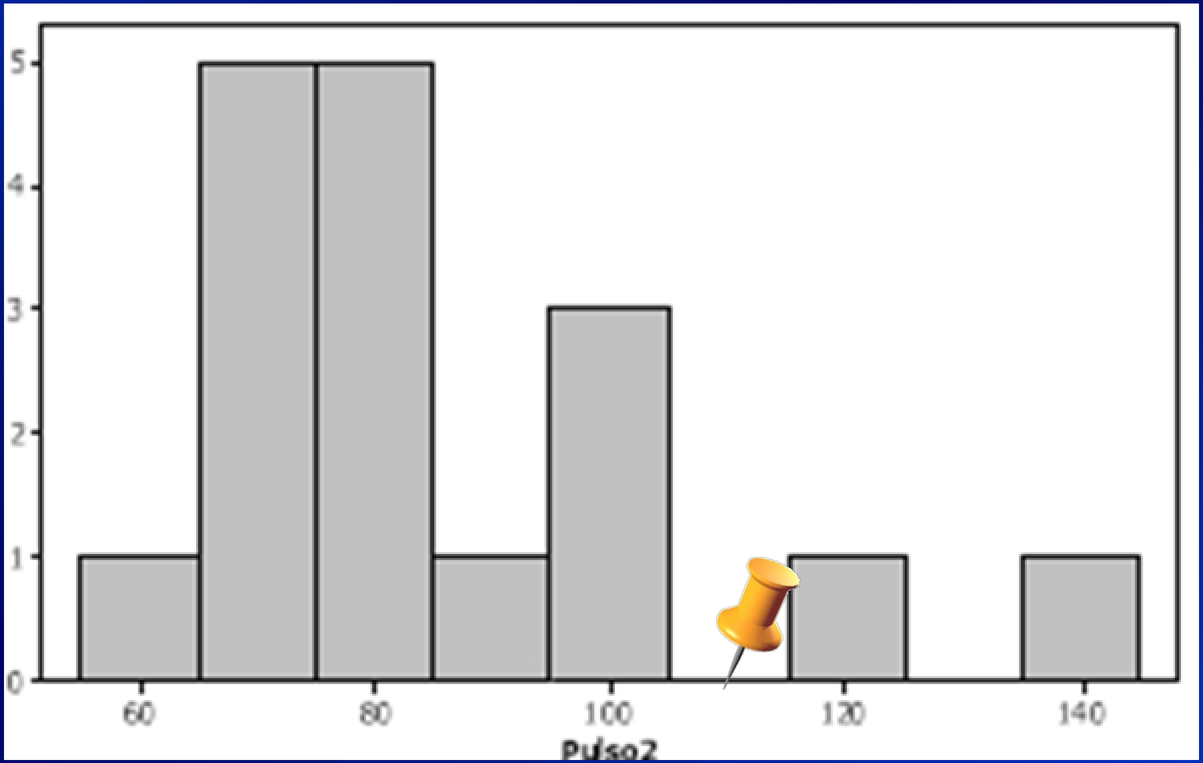
1. Range of variable (max-min)
2. Central trend
3. Dispersion
4. Simmetry
5. Anomalies



READING HISTOGRAMS

Central trend

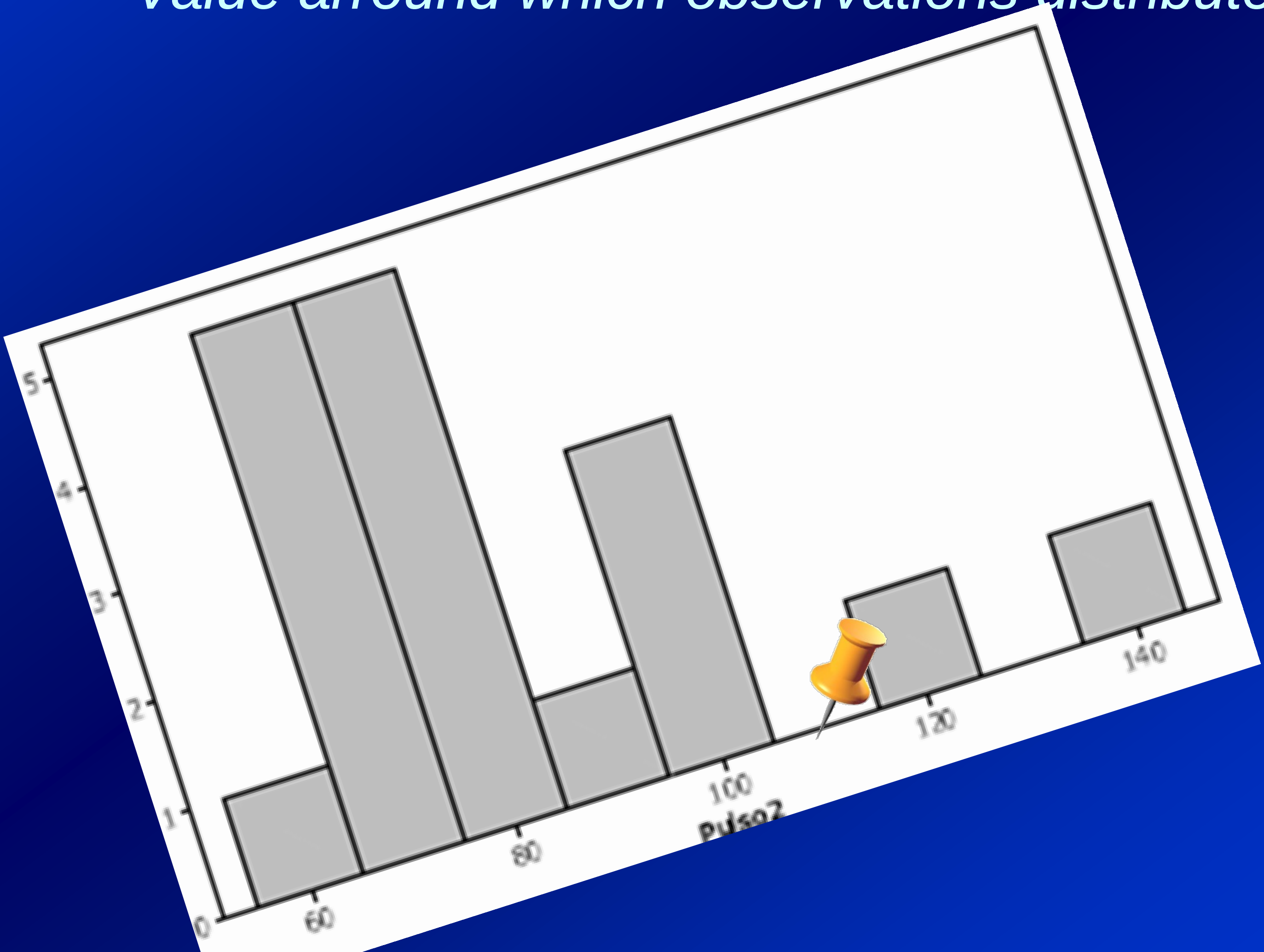
Value arround which observations distribute



READING HISTOGRAMS

Central trend

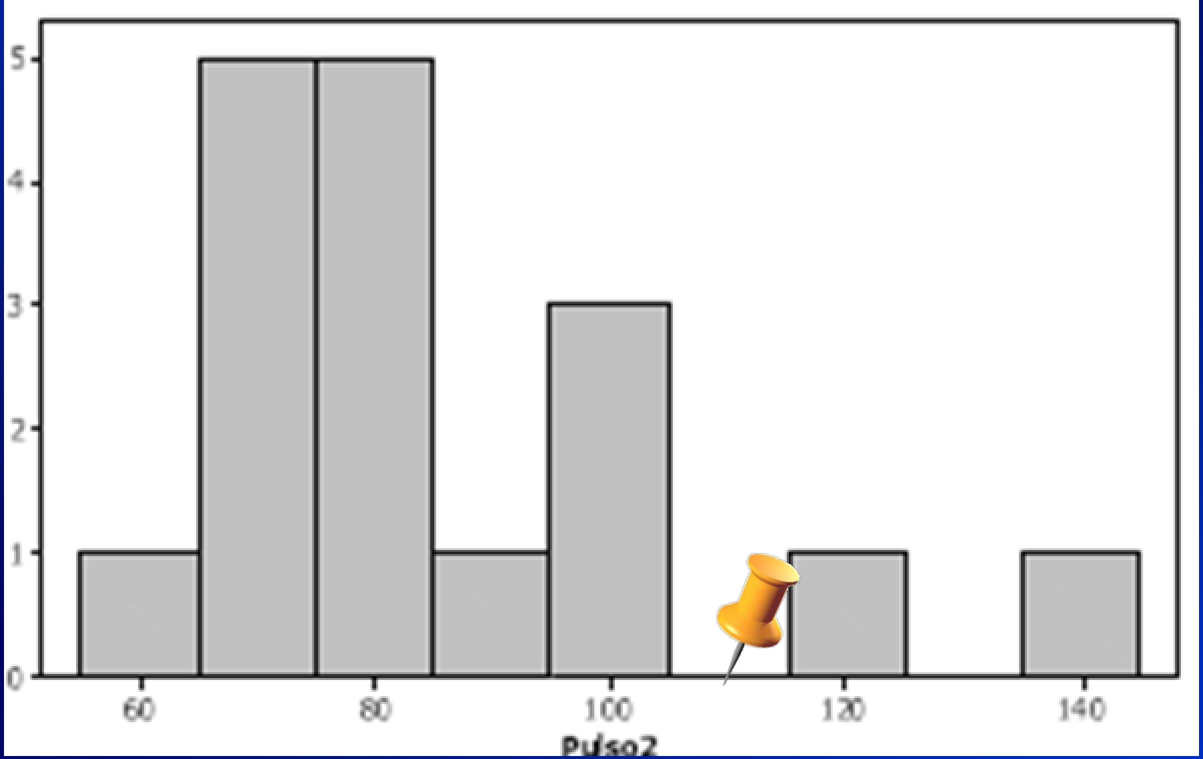
Value arround which observations distribute



READING HISTOGRAMS

Central trend

Value arround which observations distribute

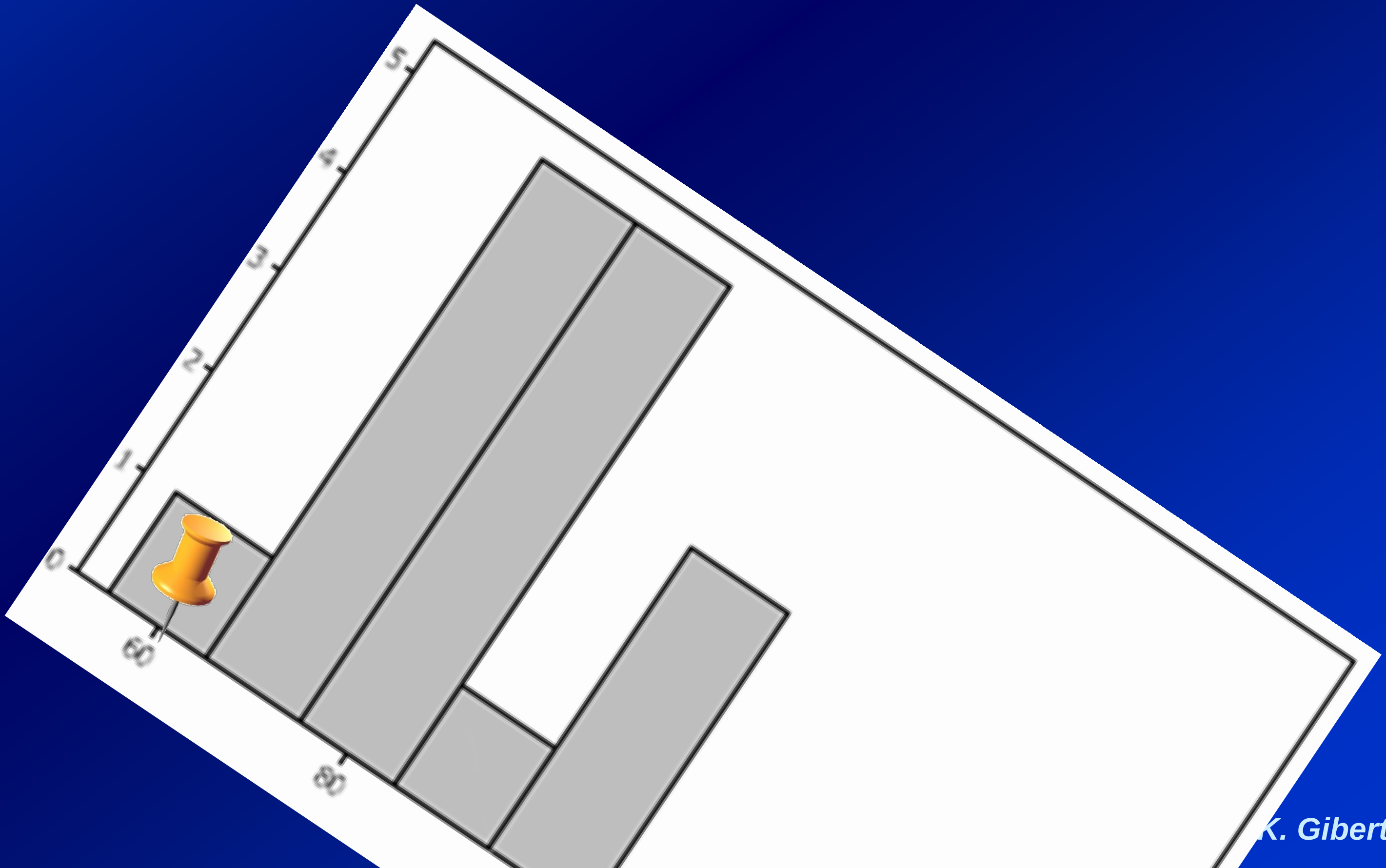


© K. Gibert

READING HISTOGRAMS

Central trend

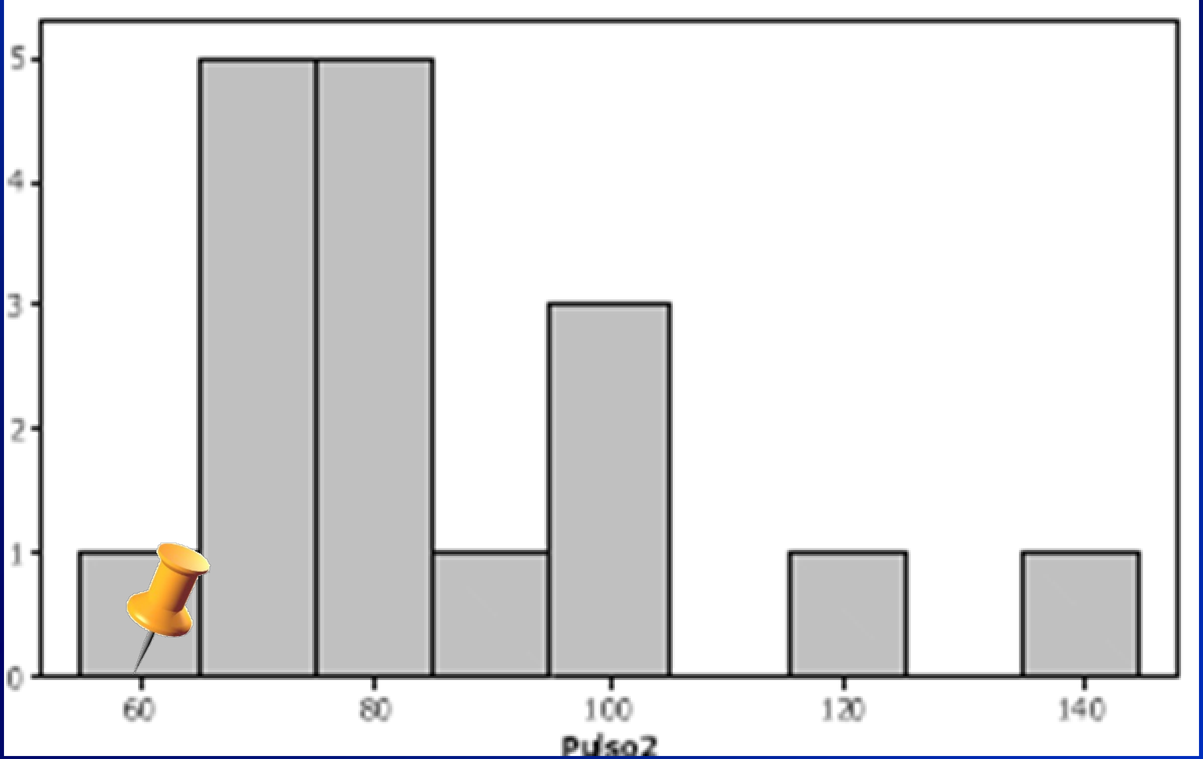
Value arround which observations distribute



READING HISTOGRAMS

Central trend

Value arround which observations distribute



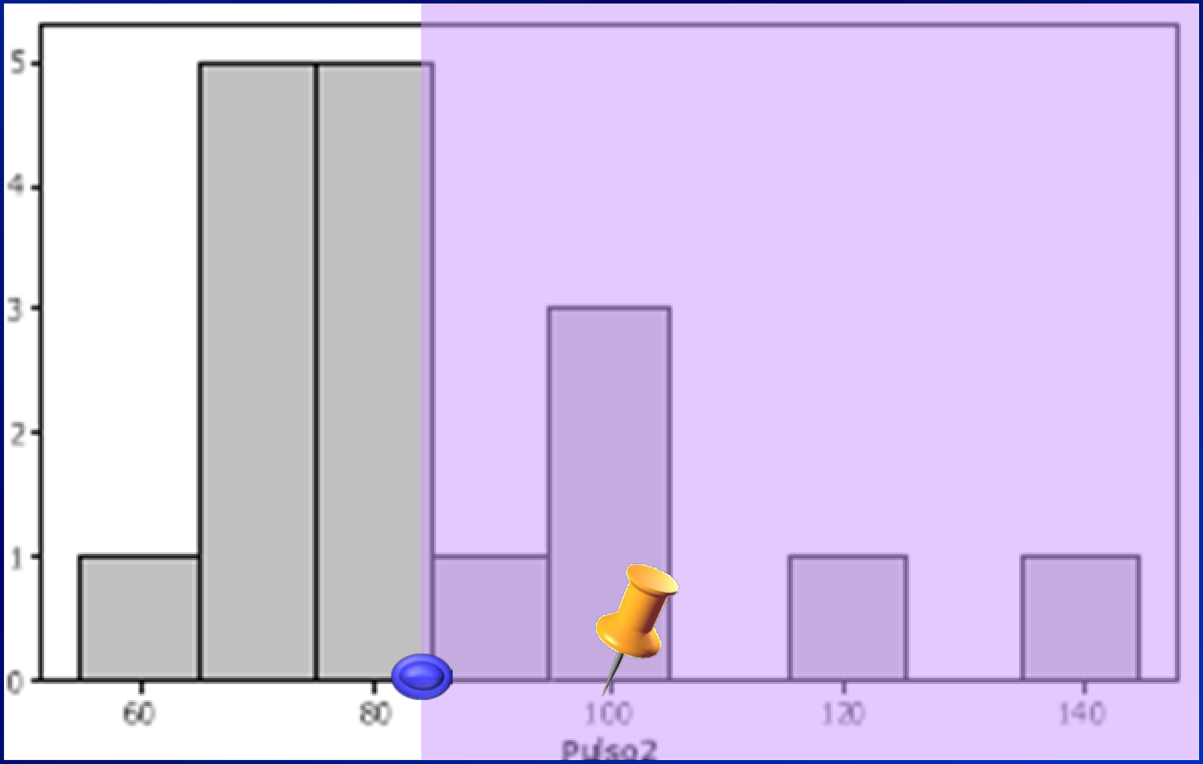
© K. Gibert

READING HISTOGRAMS

Dispersion/Variability

How observations concentrate arround central trend?

Mean distance to central trend

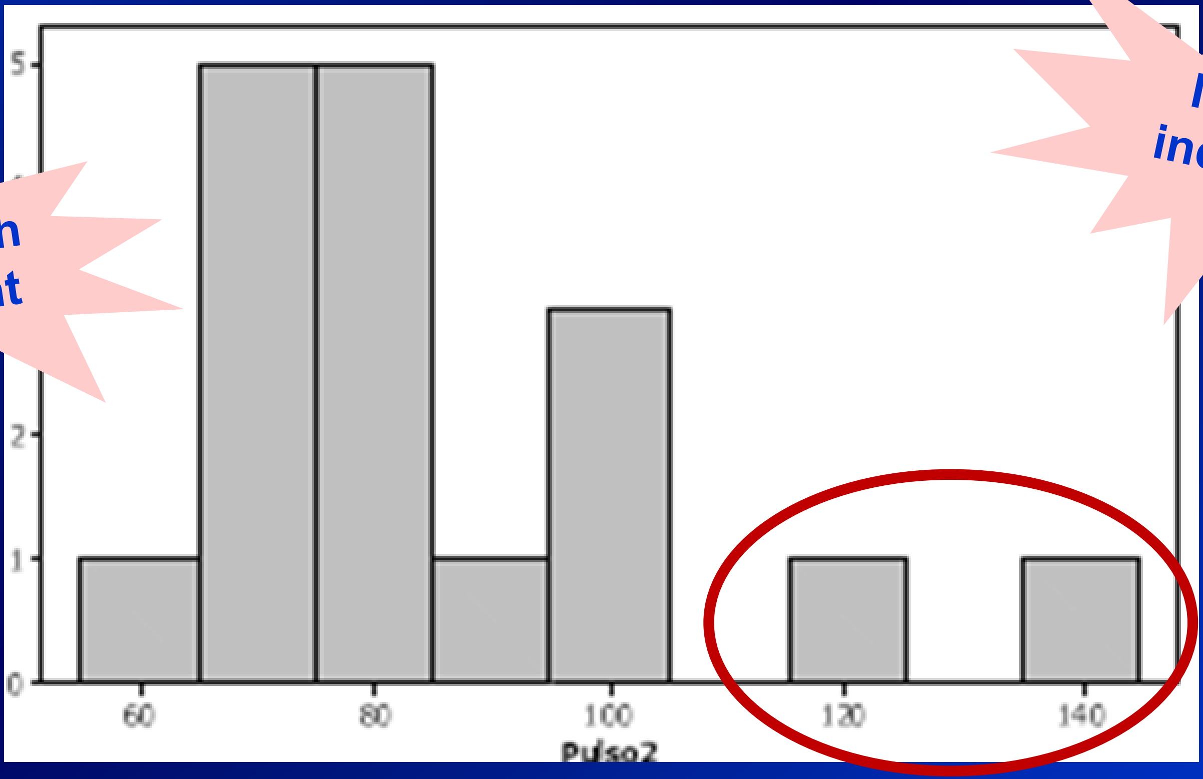


© K. Gibert

READING HISTOGRAMS

Anomalies

Outliers: Observations anormaly far from rest



CARE with Treatment

Identify individual

READING HISTOGRAMS

Main Patterns

[Gibert, JANO1996]

Albumine

Bilirrubine

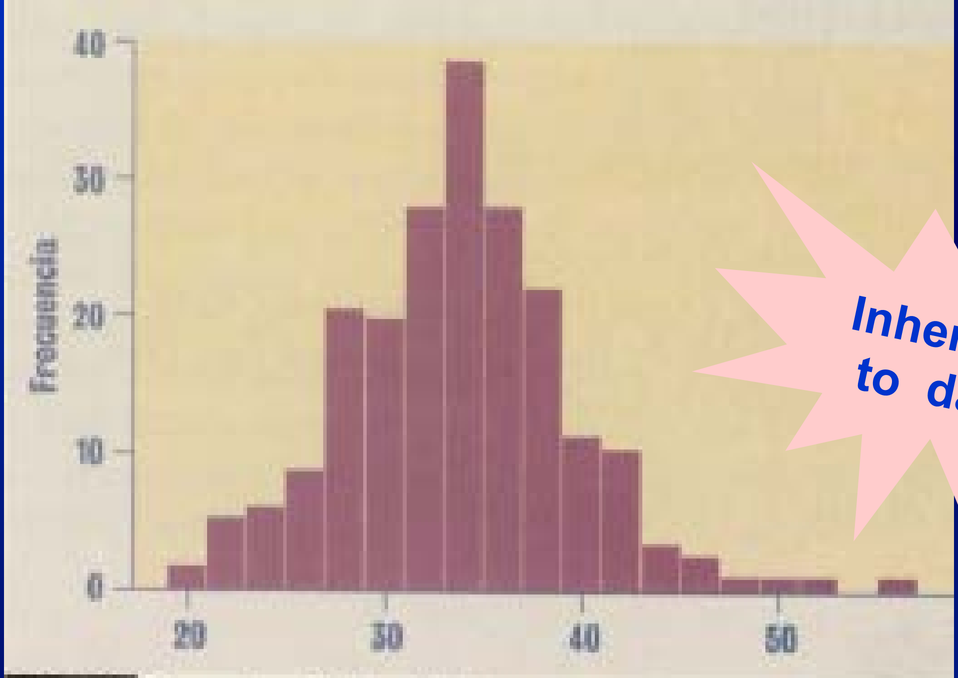
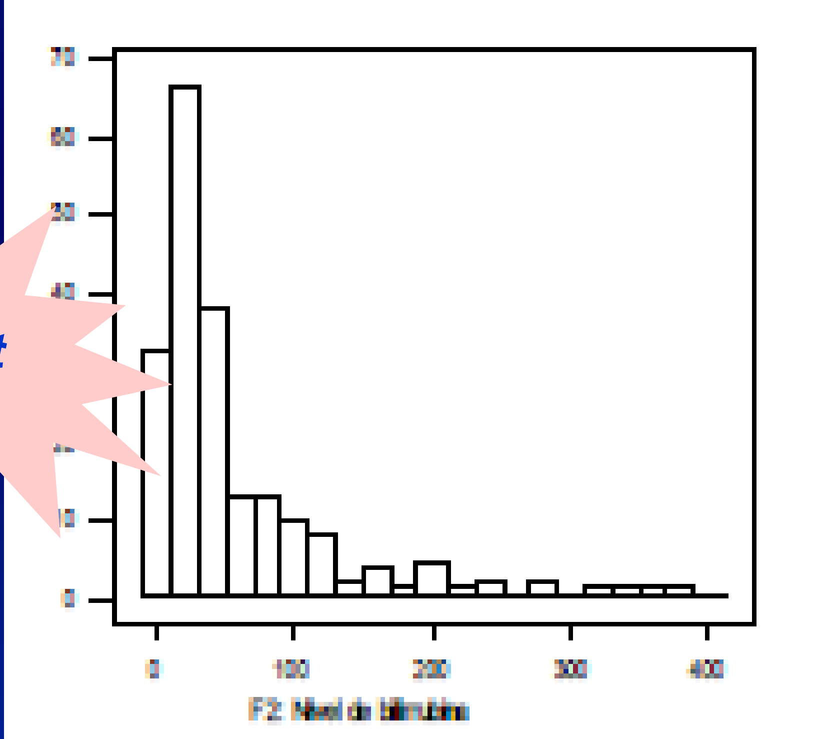


Figura 1 Concentración de albúmina en sangre.



Inherent to data

Symmetric

Asymmetric

© K. Gibert

READING HISTOGRAMS

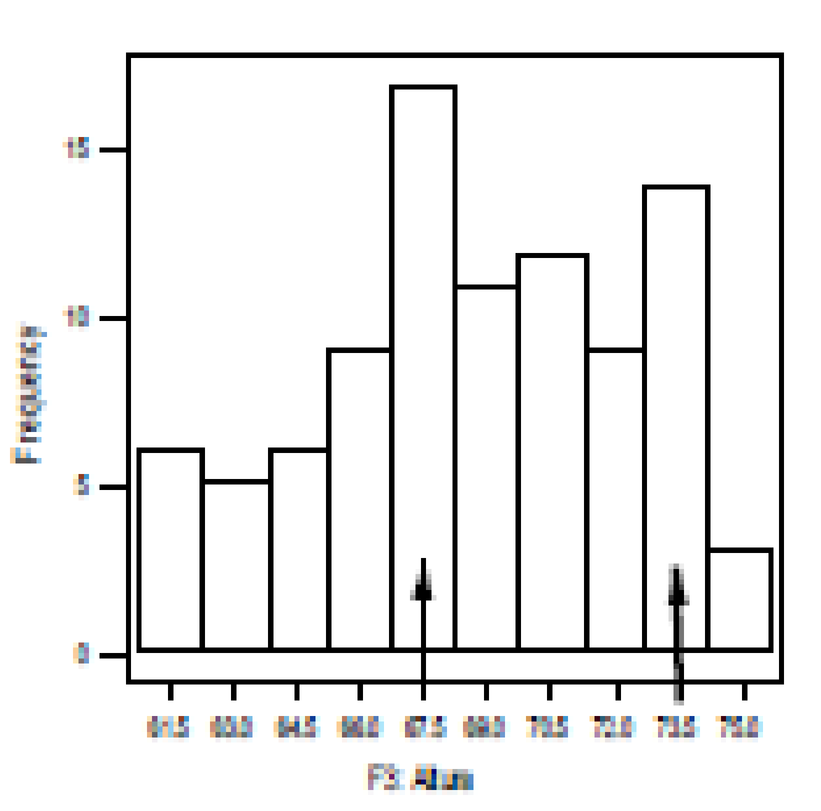
Main Patterns

Multimodality

Several central trends!!

Find discriminant factor

Height



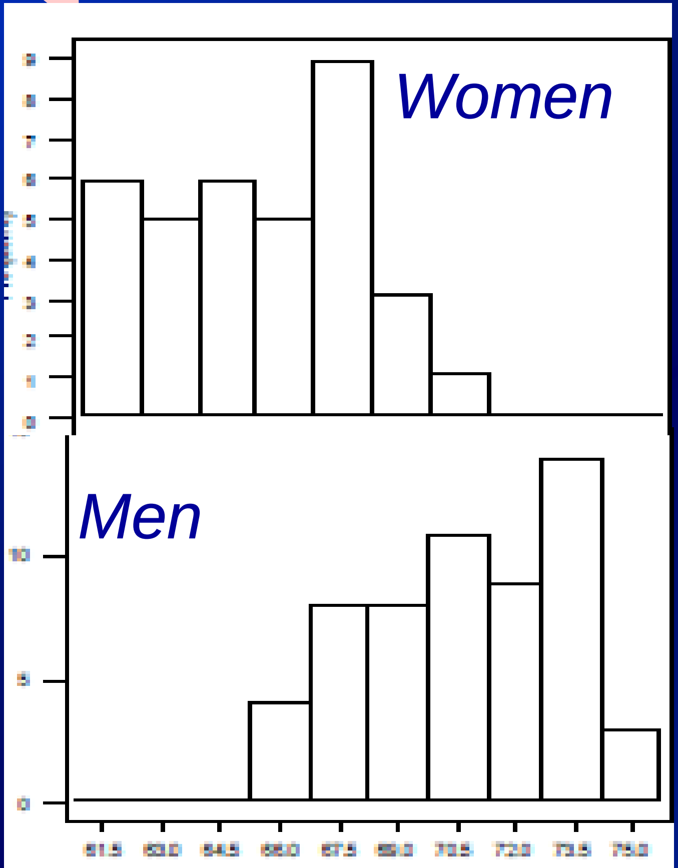
© K. Gibert

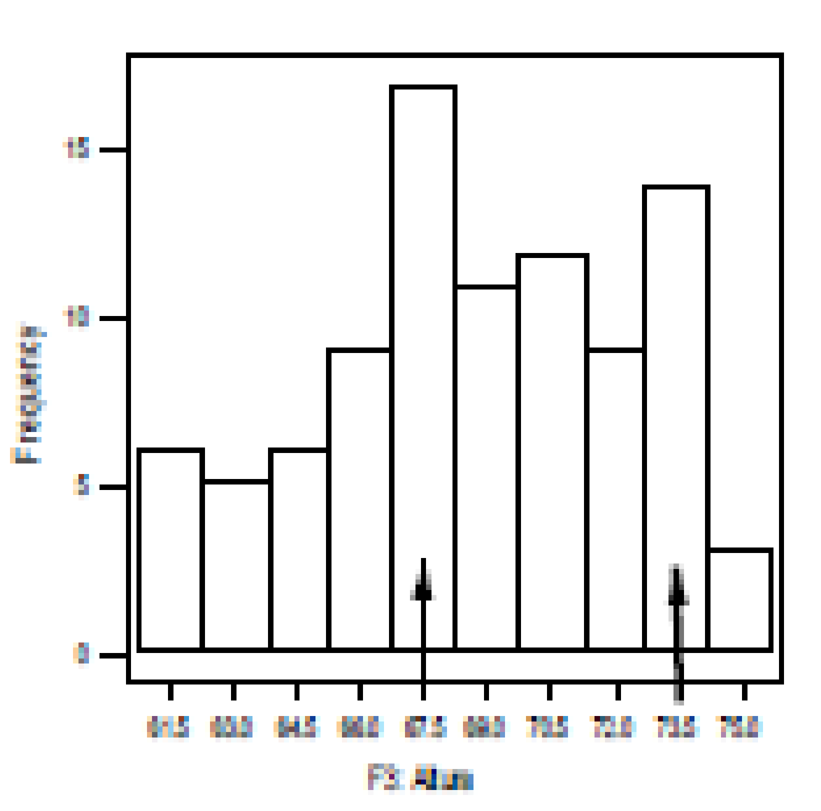
READING HISTOGRAMS

Find discriminant factor

Main Patterns

Height





© K. Gibert

READING HISTOGRAMS

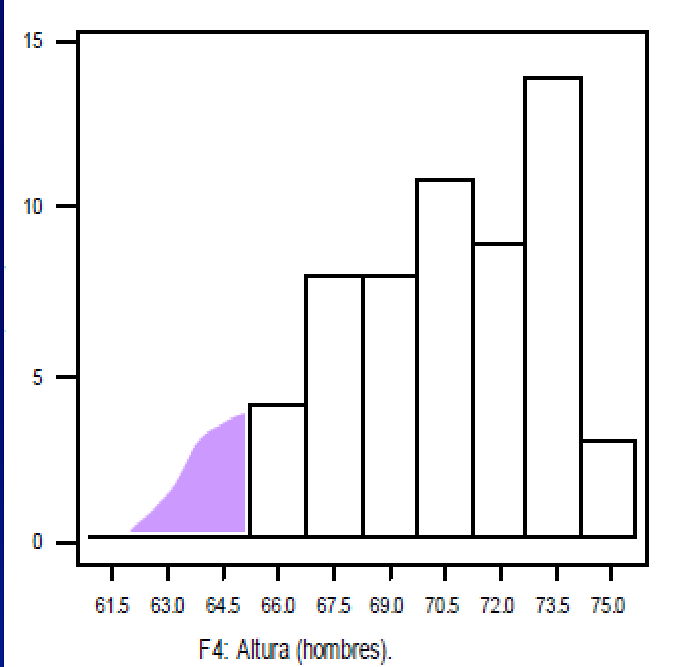
Main Patterns

Height of Men

Scarped

Part of distribution trunked!

(only adult men)

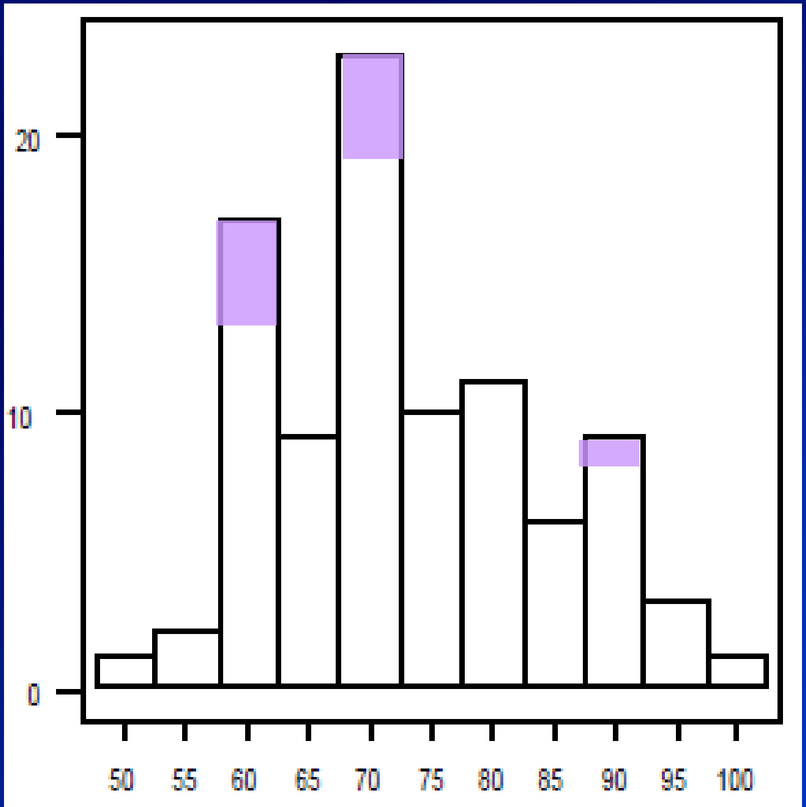


© K. Gibert

READING HISTOGRAMS

Main Patterns

Pulse per minute



Dentat

Measurement approximations!

Count one minute

Count 10 sec x 6

Count 25 sec x 4

…

Tools

1. Graphical

   Visualitze variable’s distribution

the best inductor

2. Numerical

   Quantify what is observed in he graphs

Objectivate

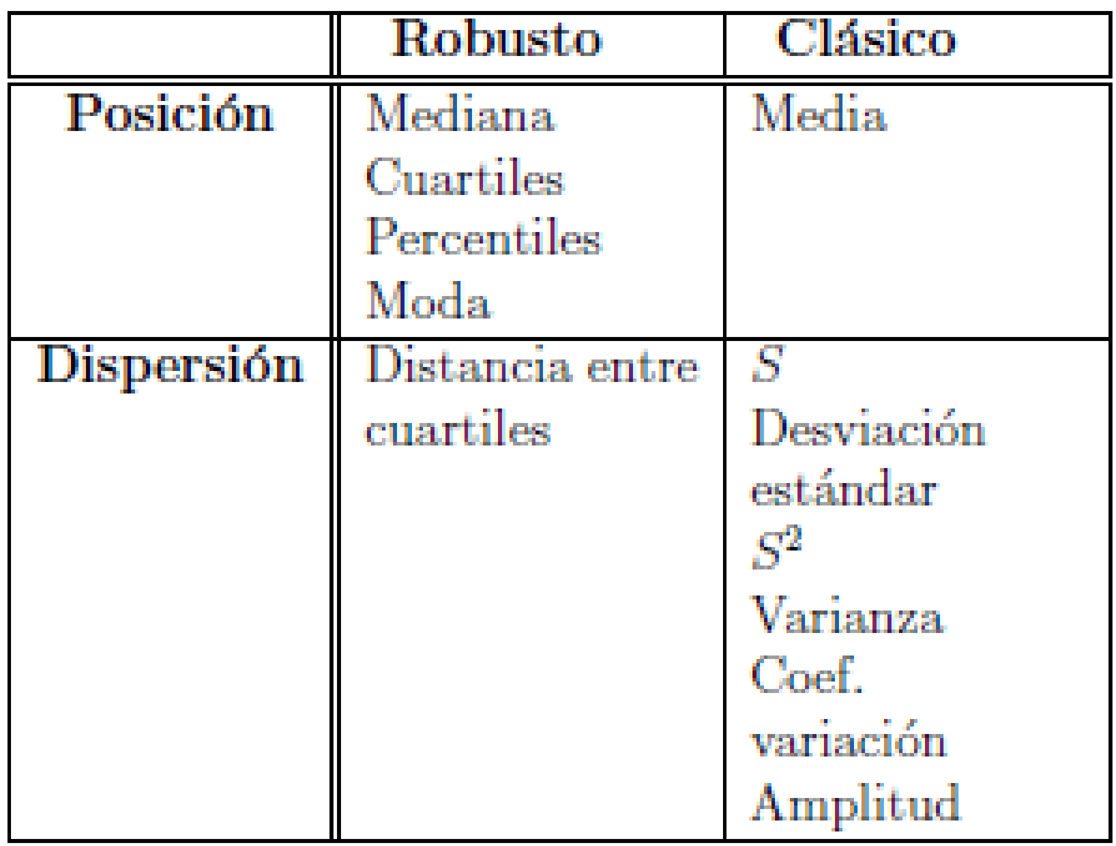
2. Numerical tools

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

Numerical tools

Mean

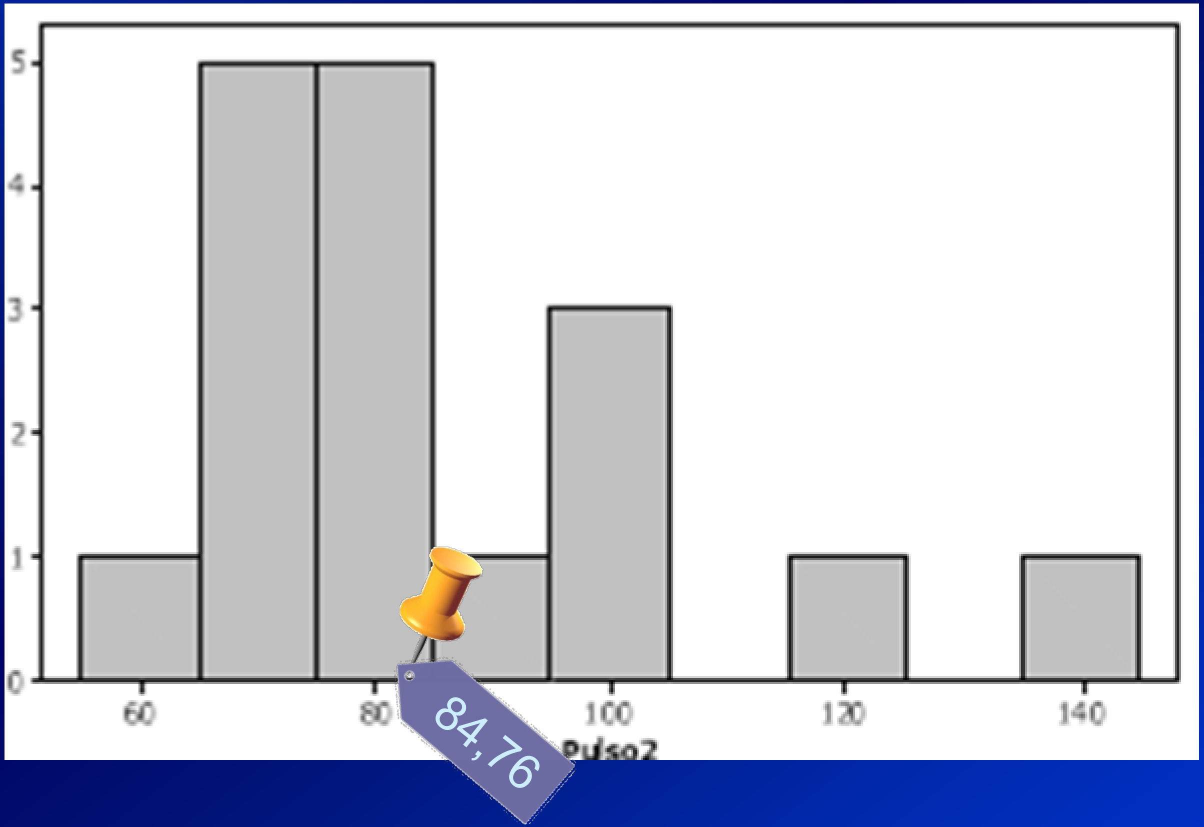
$x_1, x_{2,} \ldots x_n$ are n observatio ns of a variable X

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

READING HISTOGRAMS

Central trend

Value arround which observations distribute



© K. Gibert

Numerical tools

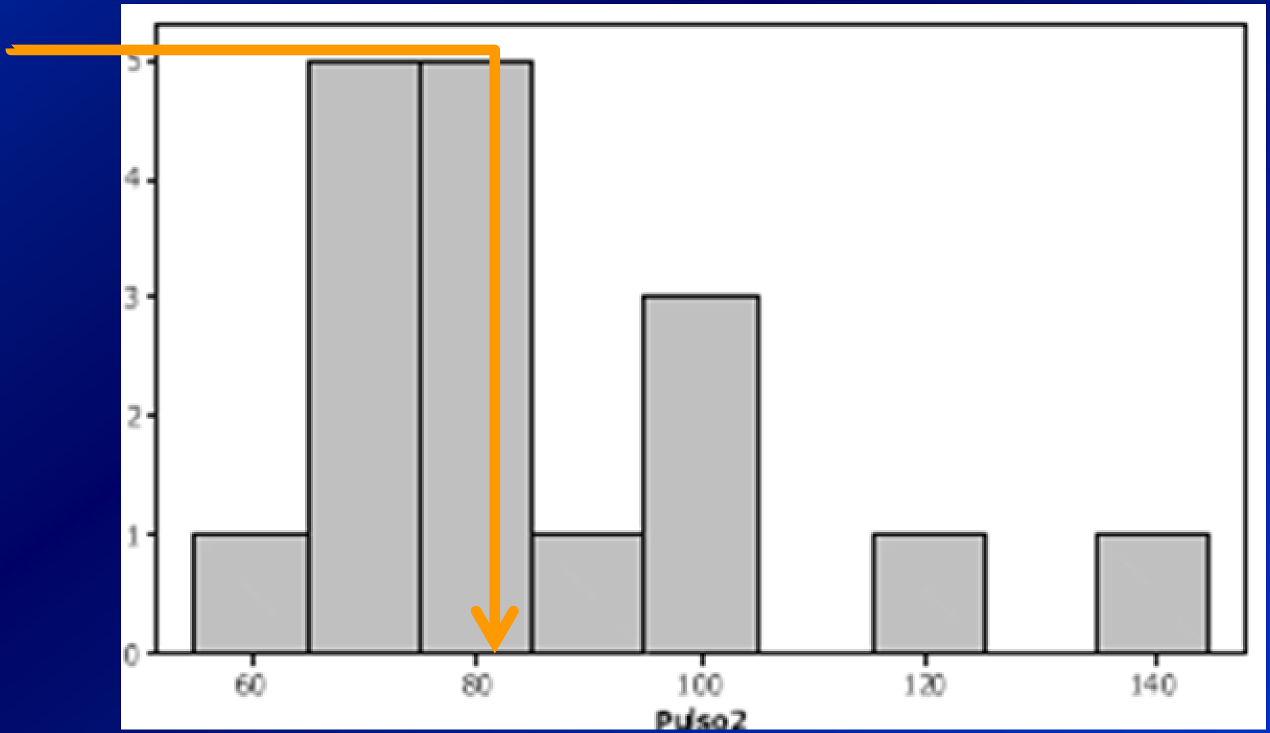
Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Numerical tools

Mode

The most frequent observation



© K. Gibert

Numerical tools

Dispersion Measures

$$V(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$s_X = \sqrt{V(X)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

$$qV(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$q\text{-}s_X = \sqrt{qV(X)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

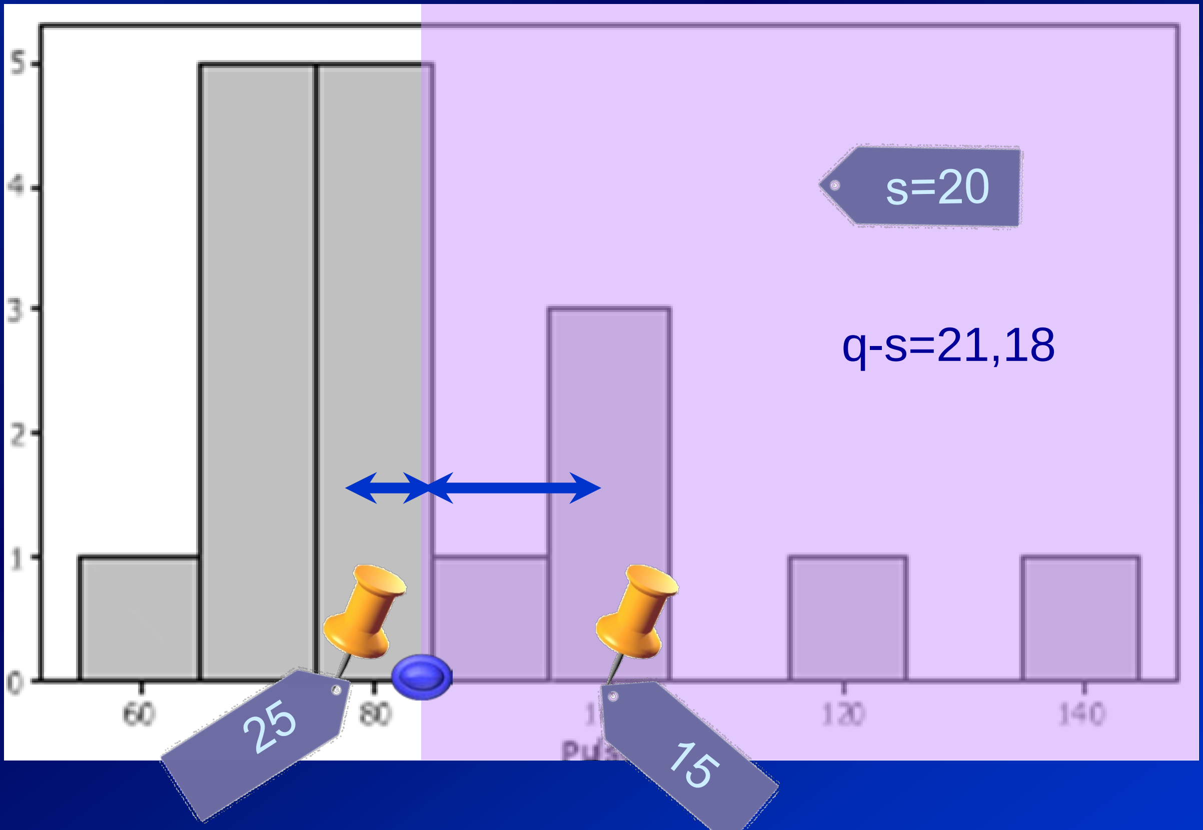
Mean distance to central trend

READING HISTOGRAMS

Dispersion/Variability

How observations concentrate arround central trend?

Mean distance to central trend



© K. Gibert

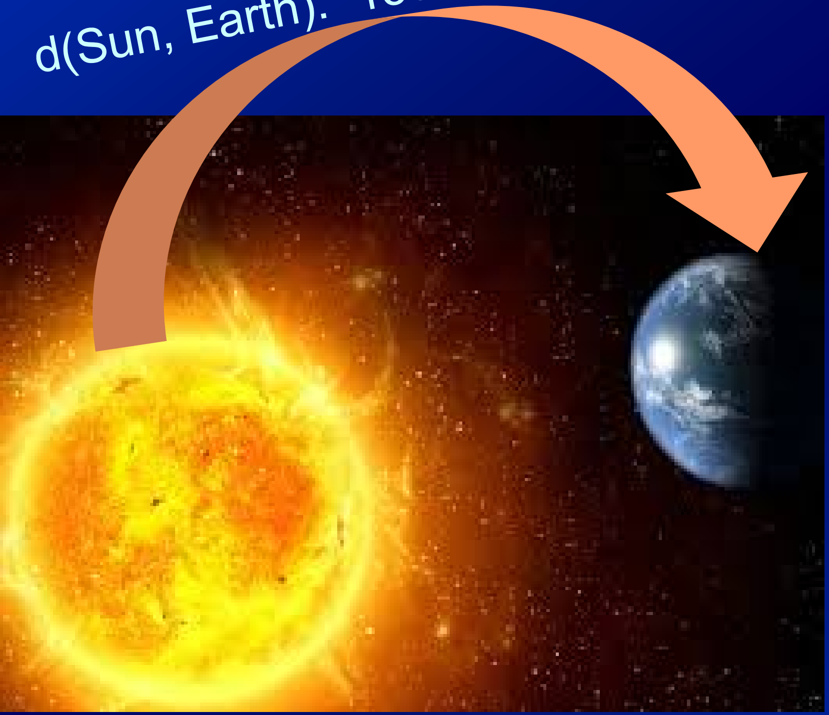
Numerical tools

Dispersion Measures

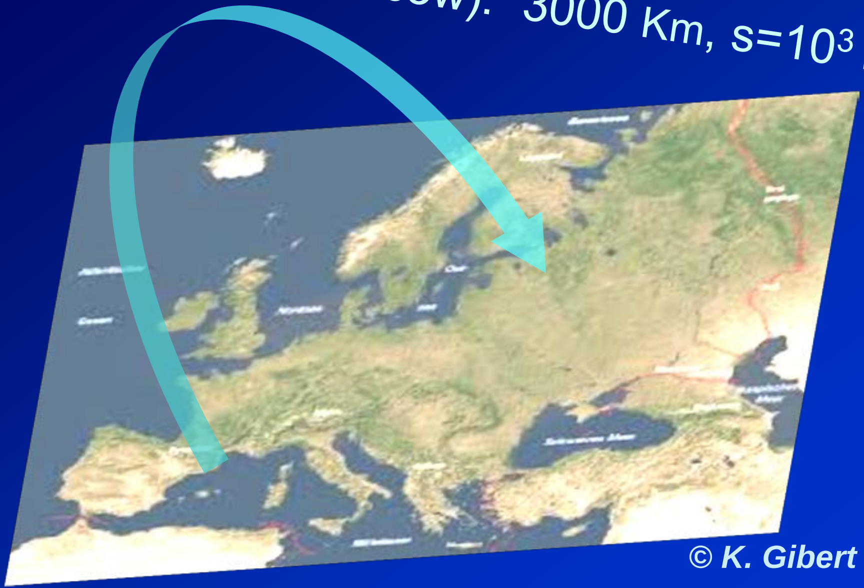
Variation Coeficient: $\frac{s}{\bar{x}} \times 100$

d_i/range

d(Sun, Earth): 150x10^6 Km, s=10^3 Km



d(BCN, Moscow): 3000 Km, s=10^3 Km



Numerical tools

5-Number Summary

Robust

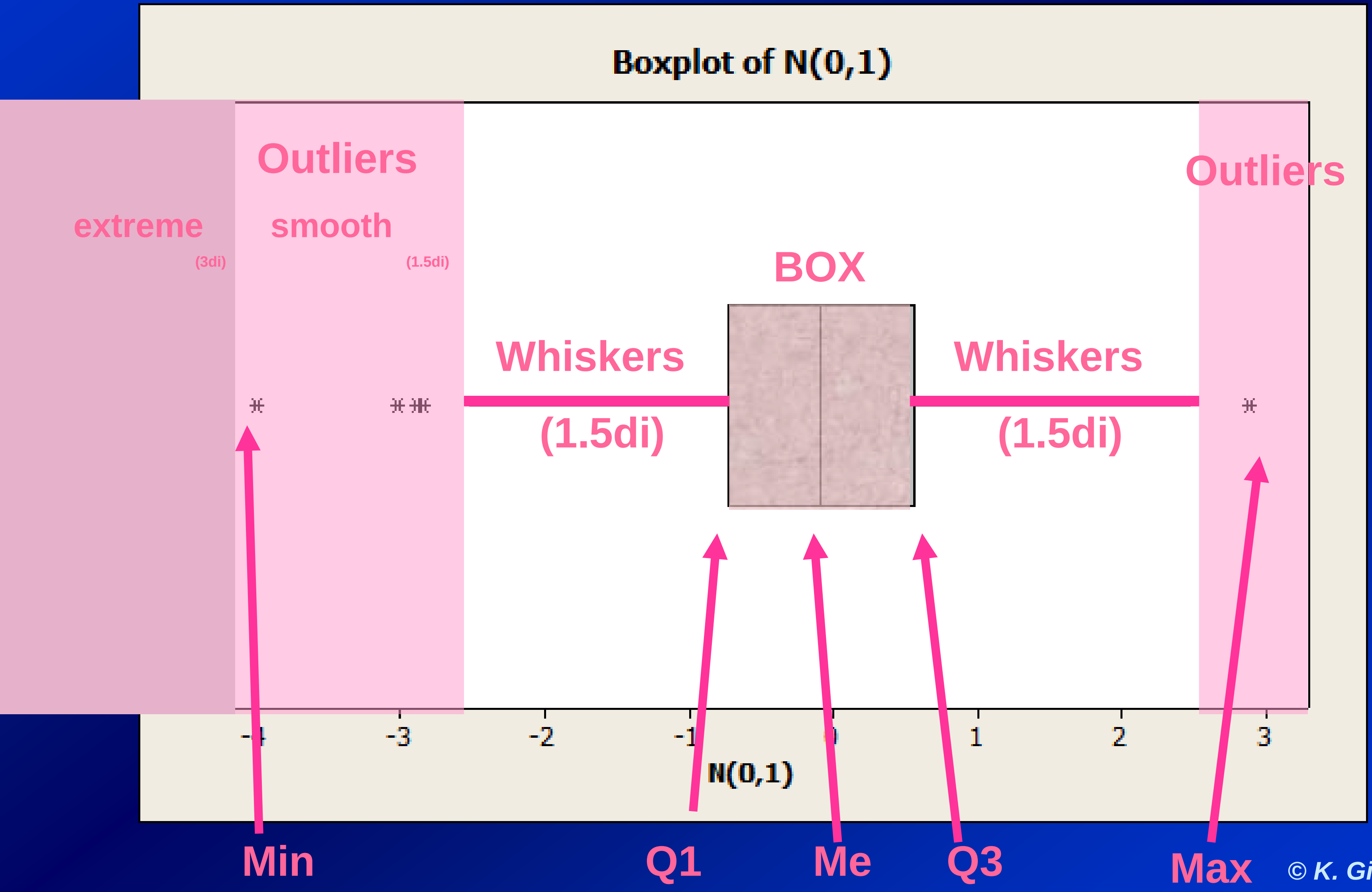
<min, Q1, Me, Q3, max>

Requires sorting

© K. Gibert

Boxplot *[Tukey 1956]*

Symbolic representation of 5-Number Summary



Boxplot of N(0,1)

Outliers

extreme (3di)

smooth (1.5di)

BOX

Whiskers (1.5di)

Whiskers (1.5di)

Outliers

Min

Q1

Me

Q3

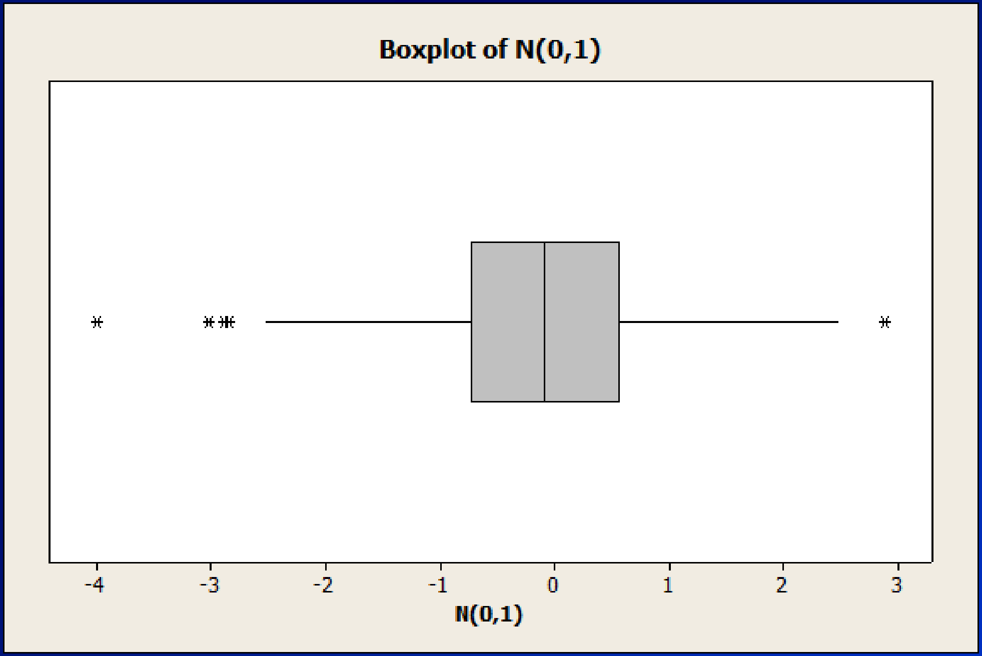
Max

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

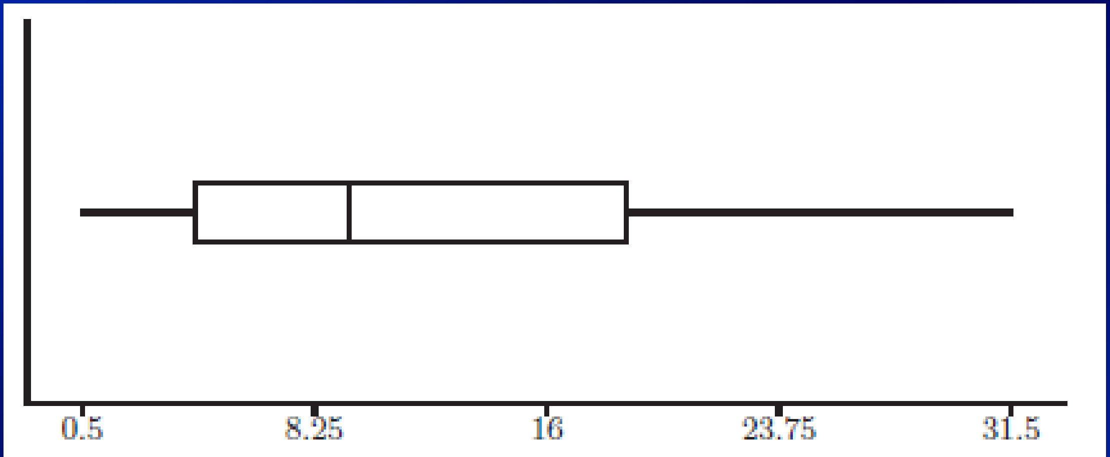


© K. Gibert

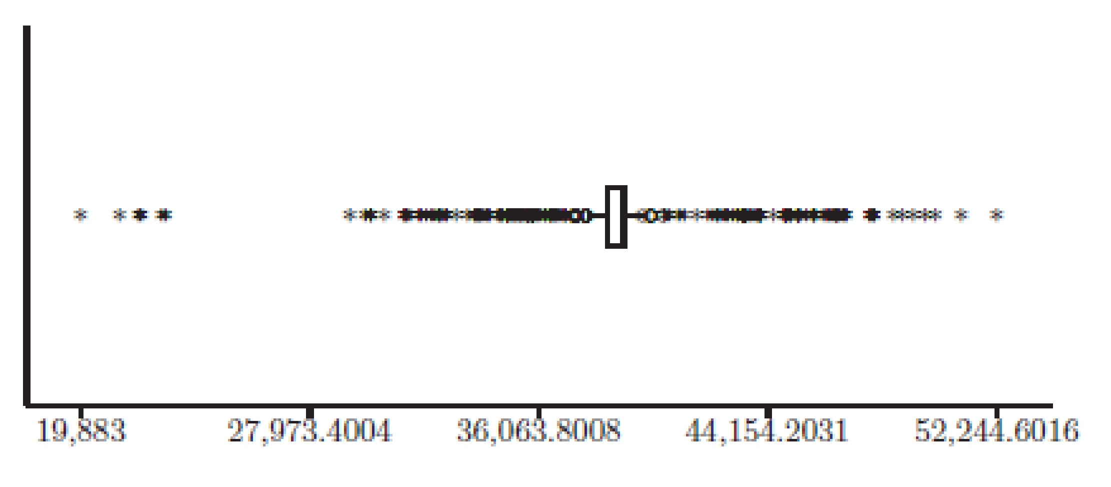
READING BOXPLOTS

Dispersion

Ammonium



QB-B

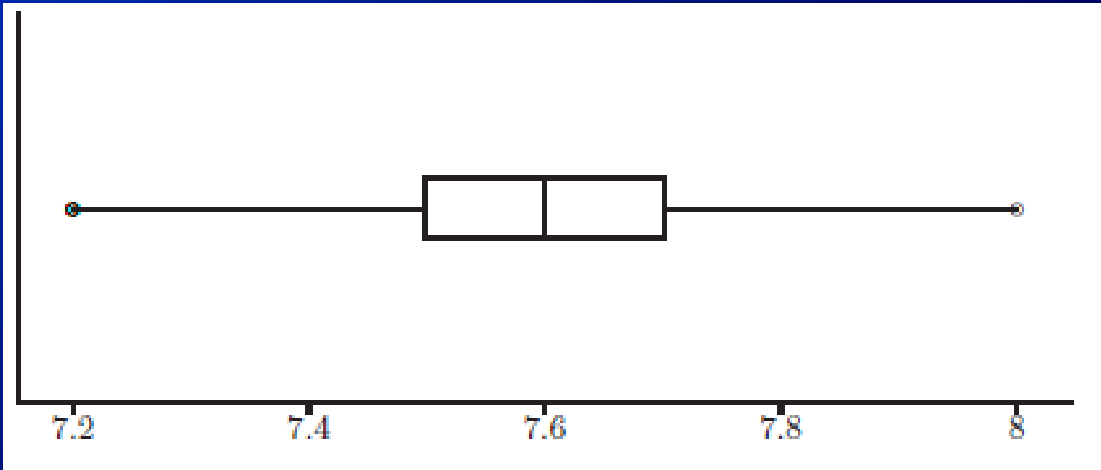


© K. Gibert

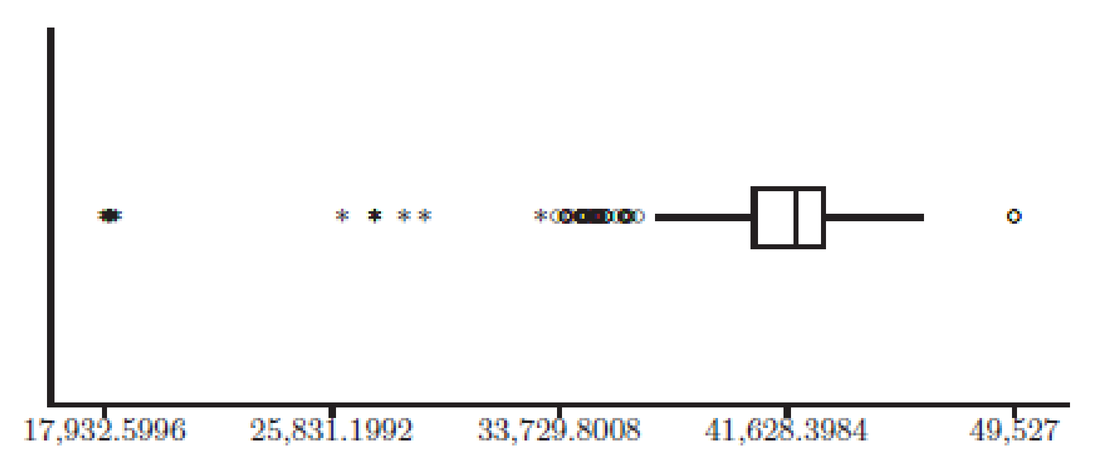
READING BOXPLOTS

Simmetry

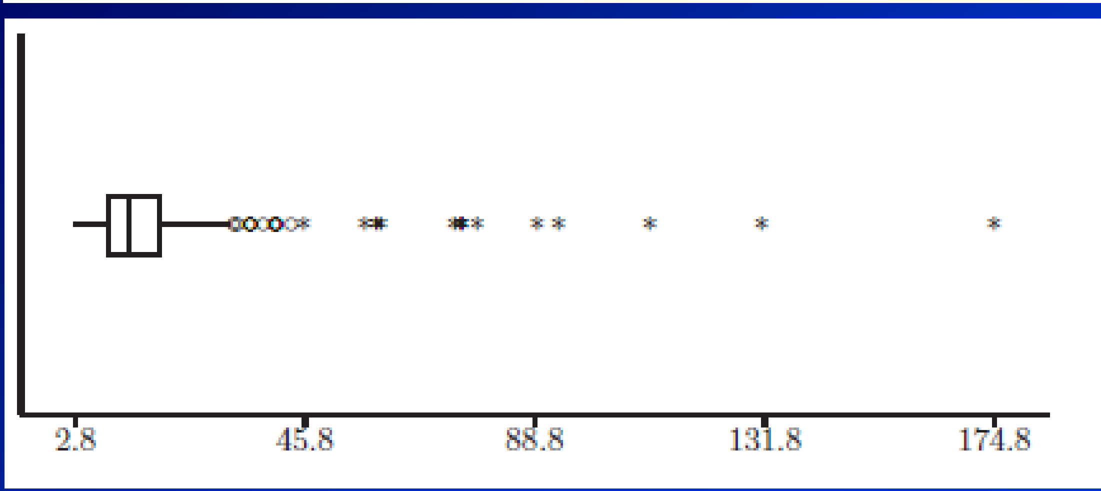
PH



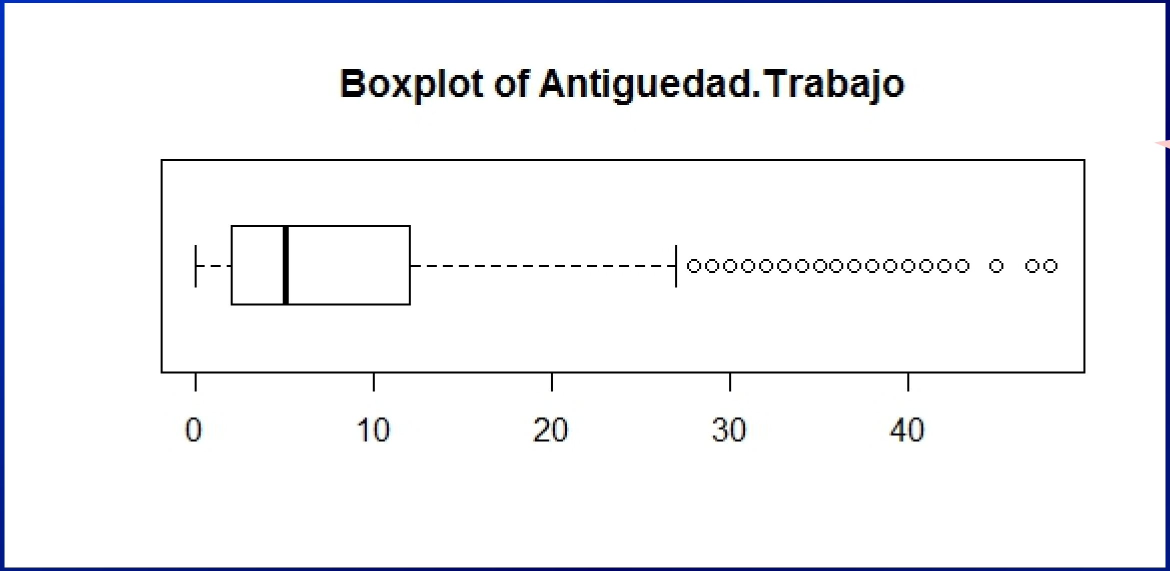
QR-G



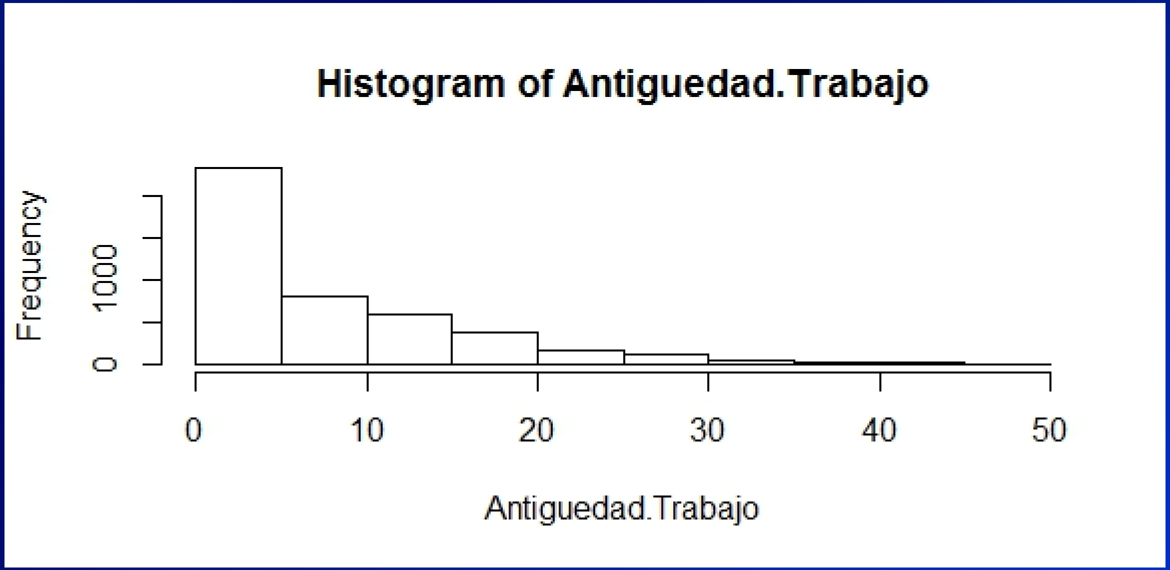
SS-S



READING BOXPLOTS



CARE With *



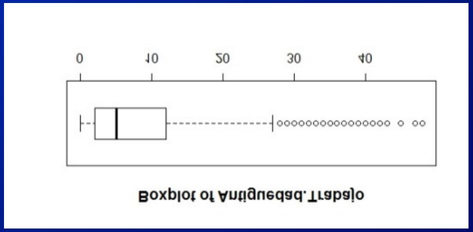
© K. Gibert

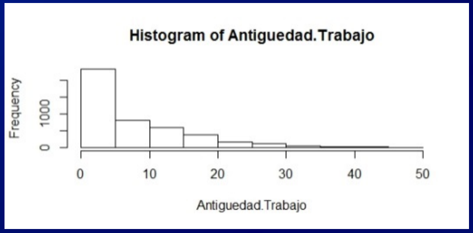
Symmetry

```
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

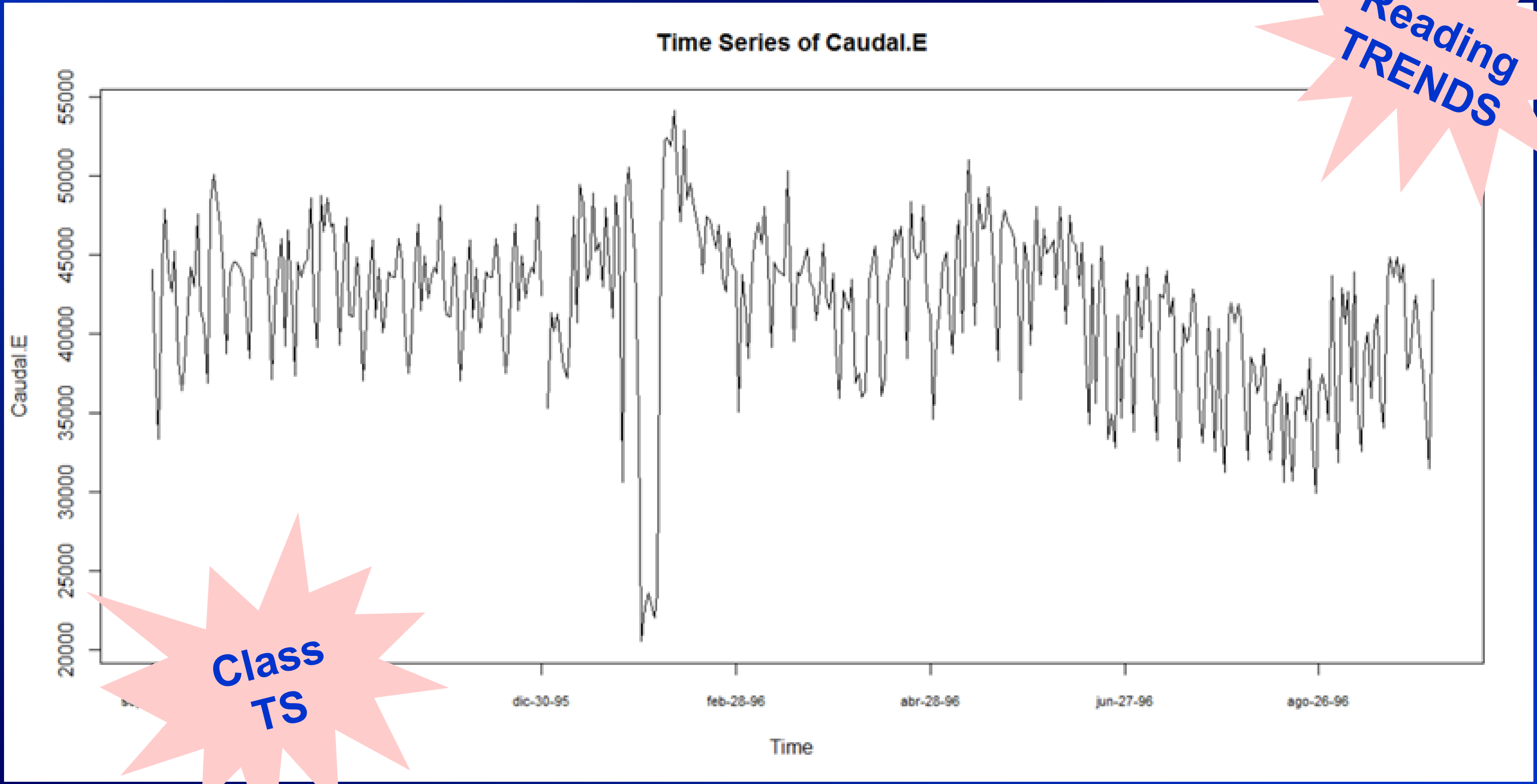
   3. Characterize the variable

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

3. Cronological data

Observations are sequentially sorted in the dataset

Time series plot



Reading TRENDS

Class TS

© K. Gibert

Cronological data

Observations are sequentially sorted in the dataset

Evenctually the Date is Available

Class Date

Cronological data

Date Objects in R

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

4. Assessing Normality

1. 68-97-99.5 Rule

   [x +-s], [x+- 2s] [x+-3s]

2. Normality plot (qq-plot, Henri line)

3. Normality assessment test: Shapiro Wilk

$$W = \frac{\left(\sum_{i=1}^{N} a_i y_i\right)^2}{\sum_{i=1}^{N} (y_i - m_1)^2}$$

yi= ith order statistic m1=sample mean

ai= computed as linear regression to the expected value of standard normal order statistics

5. Assessing Exponentiality

The rule of 70

Time of doubling : 70/R, R growing factor *[Moore, McCabe 93]*

X has exponential growth with constant factor R

if needs 70/R time to pass from X to 2X

Basic Descriptive Analysis Numerical Variables

Karina Gibert

Dpt. Statistics and Operation Research

Knowledge Engineering and Machine Learning Research group

Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu

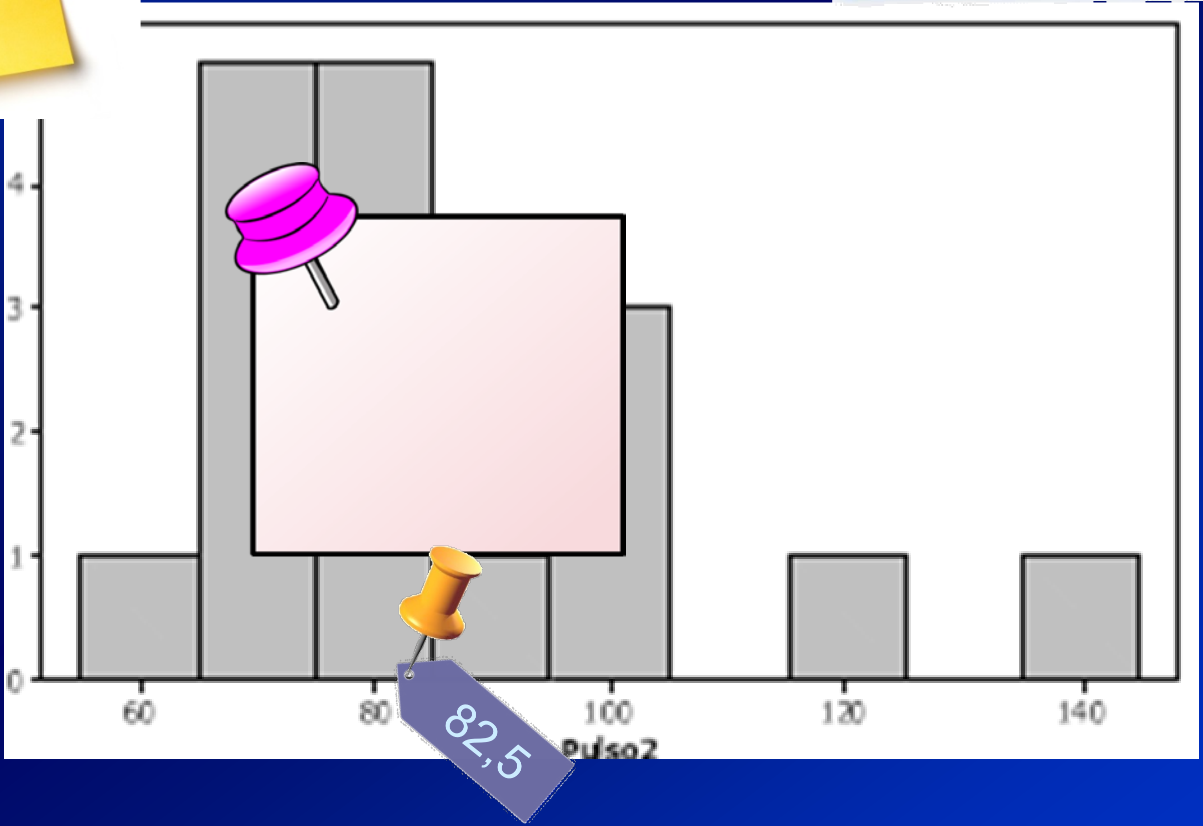
www.eio.upc.edu/homepages/karina

Are there any questions?...

READING HISTOG

Central trend

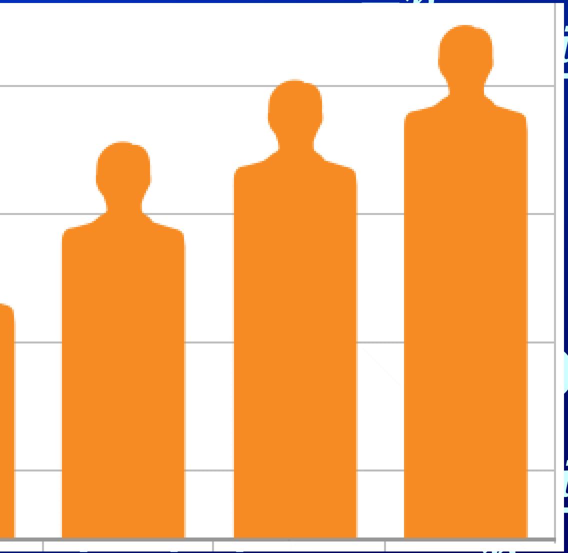
Value arround which observa



Numerical tools

Dispersio

$$\sum^{n} \quad i - \bar{x})^2$$



oeficient

$$\frac{i - \bar{x})^2}{n-1}$$

q

Mean dist
central