UD 2

DESCRIPTIVE STATISTICS

GENERAL CONCEPTS

POPULATION

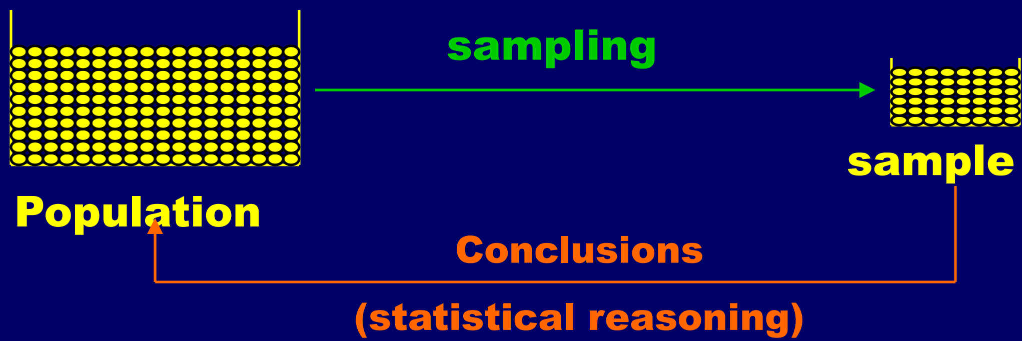
Set of objects that we are interested in, for which we intend to get conclusions.

Example: All pieces that are going to be assembled by means of a certain industrial process.

SAMPLE

Subset formed by parts of the objects (individuals) of a population.

Example: 20 pieces produced by the industrial process.



The sample must be "representative" of the population.

Only guarantee of "representativity": Random sampling.

OBJECT OF SAMPLING

To know the population, by analyzing one sample.

STATISTICAL INFERENCE

Process of reasoning to obtain conclusions (with a known margin of error) about the population, by analyzing samples extracted from the population.

EXAMPLES OF POPULATIONS

Does the population exist? YES

- Intention of voting of spaniards in a General Election in Spain.
- Development of a certain pathology in buildings in Valencia.

Partially

- No. of laptop batteries that are sold every day at a computer store.
- No. of errors in the invoices of the Account Department of the company.

No

- Resistance of a new type of polymer.
- Study to investigate if a dice is correct or not.
- In any experiment in a laboratory

The size of populations is usually large, but not always:

- Population of countries in the European Union

The results of any process always present VARIABILITY

All real populations have variability. That is, it is not possible to have two identical pieces.

RANDOM VARIABLE

It is any characteristic, that can be expressed numerically, that fluctuates among the individuals of the population.

Example: the length of a piece.

Types of RANDOM VARIABLES

- Nature
  - QUALITATIVE
  - QUANTITATIVE
- Number of characteristics
  - ONE-DIMENSIONAL
  - K-DIMENSIONAL
- Set of values
  - DISCRETE
  - CONTINUOUS

ONE-DIMENSIONAL FREQUENCY TABULATION

DISCRETE VARIABLE:

Absolute frequency

Digits chosen (X_i)	No. occurrences (n_i)	Relative frequency $f_i=n_i/N$
0	0	0
1	2	0.06
2	6	0.18
3	7	0.21
4	9	0.26
5	4	0.12
> 5	6	0.18

N=34

CONTINUOUS VARIABLE:

Frequency tabulation

Resistance of a polymer (Nw)

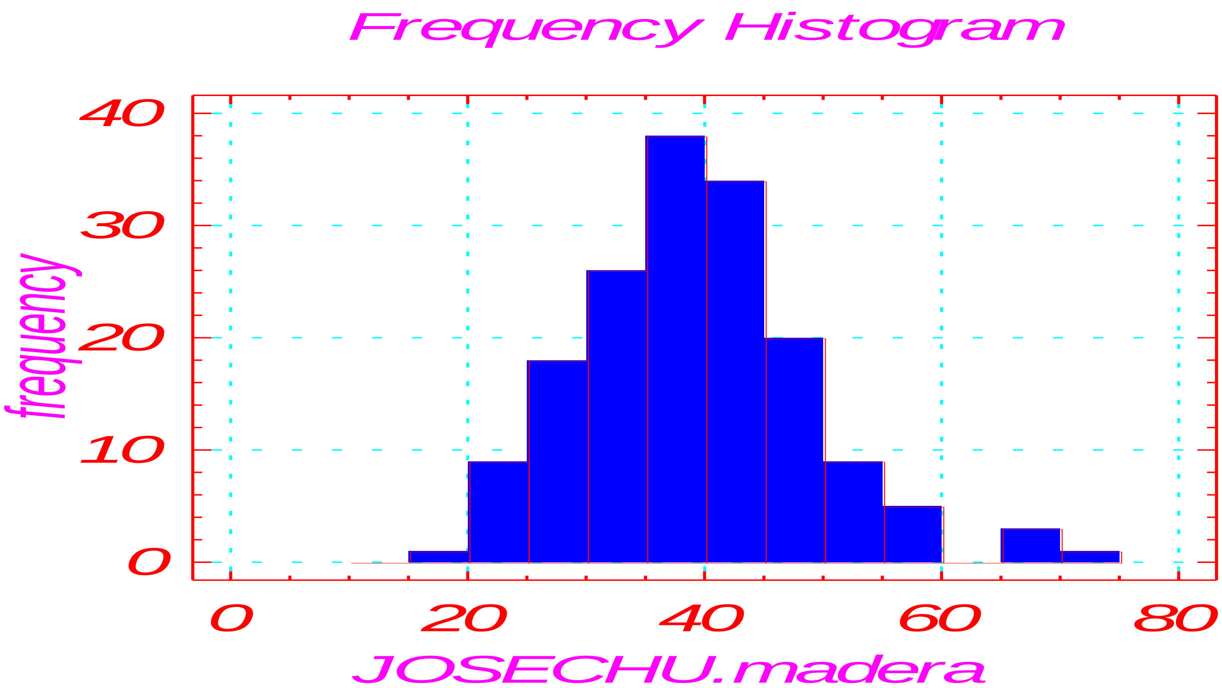
Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		10.00		0	.00000	0	.00000
1	10.00	15.00	12.50	0	.00000	0	.00000
2	15.00	20.00	17.50	1	.00610	1	.00610
3	20.00	25.00	22.50	9	.05488	10	.06098
4	25.00	30.00	27.50	18	.10976	28	.17073
5	30.00	35.00	32.50	26	.15854	54	.32927
6	35.00	40.00	37.50	38	.23171	92	.56098
7	40.00	45.00	42.50	34	.20732	126	.76829
8	45.00	50.00	47.50	20	.12195	146	.89024
9	50.00	55.00	52.50	9	.05488	155	.94512
10	55.00	60.00	57.50	5	.03049	160	.97561
11	60.00	65.00	62.50	0	.00000	160	.97561
12	65.00	70.00	67.50	3	.01829	163	.99390
13	70.00	75.00	72.50	1	.00610	164	1.0000

Mean = 39.3288 Standard Deviation = 9.46009 Median = 39.1 N=164

HISTOGRAMS

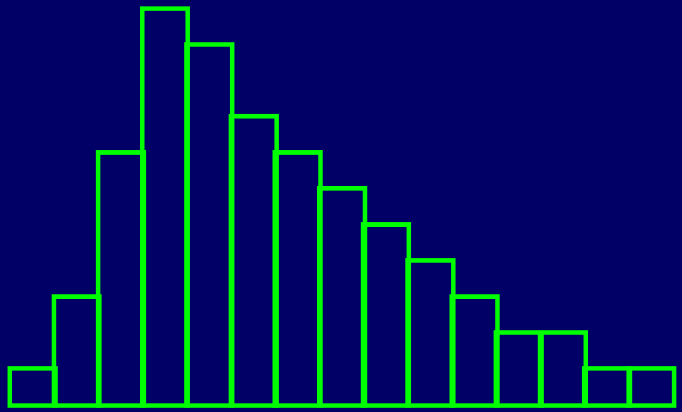
It is a graphical representation of one set of data (minimum 40-50 data) (frequency diagram)

Is this absolute or relative frequency?

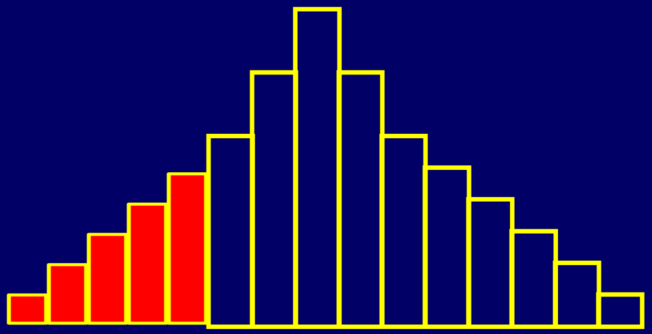


$$\text{Best number of intervals} \approx \sqrt{N} \in (5,15)$$

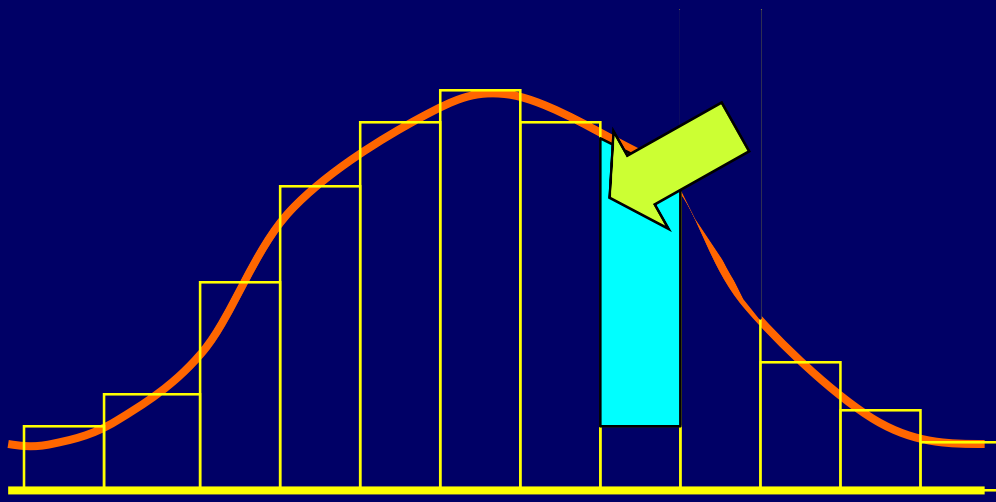
Is this a symmetric distribucion?



Asymmetric histogram



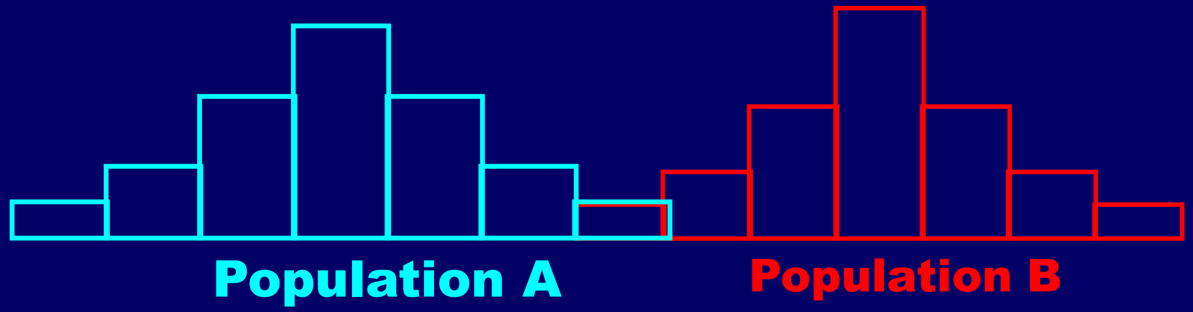
Truncated data

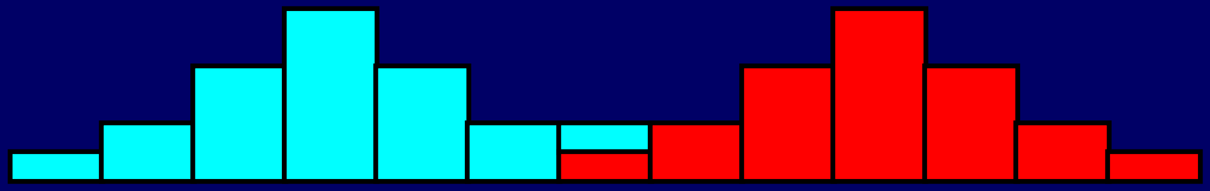


Abnormal frequency of one interval

(systematic error in data recording)

Mixture of populations



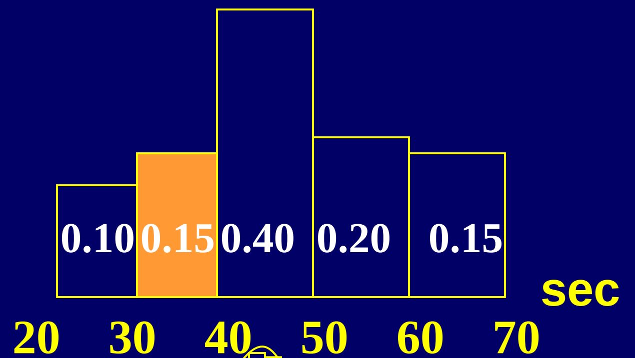


Histogram of 2 different populations

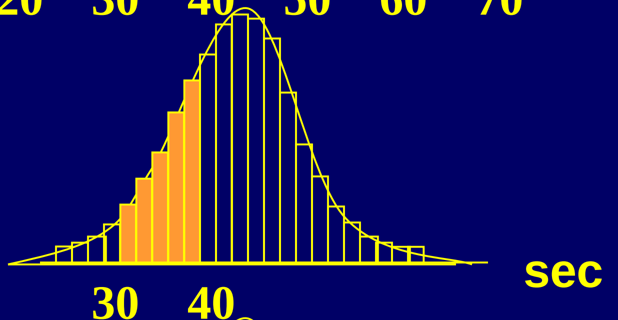
Continuous random variables: Density function.

Example: time (sec.) required by algorithm to invert a matrix

Sample = 40

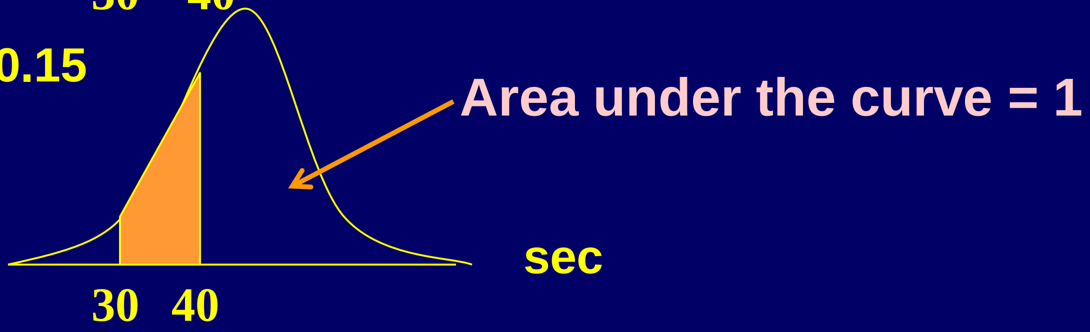


Dark area = 0.15

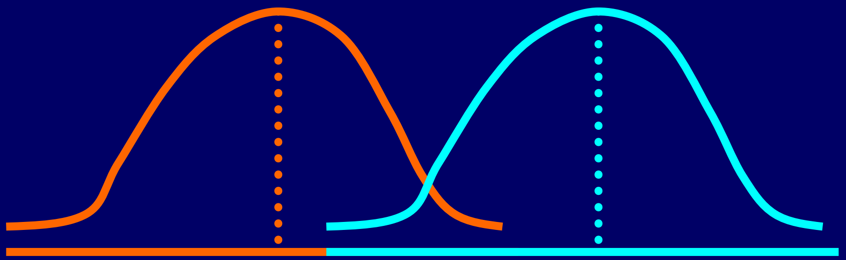


Probability (30<X<40) = 0.15

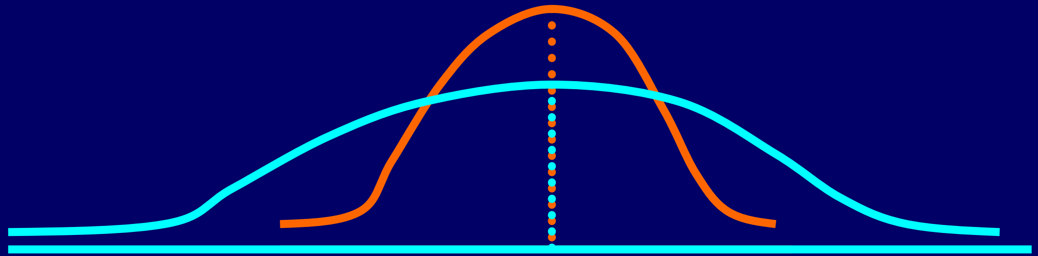
Dark area = 0.15



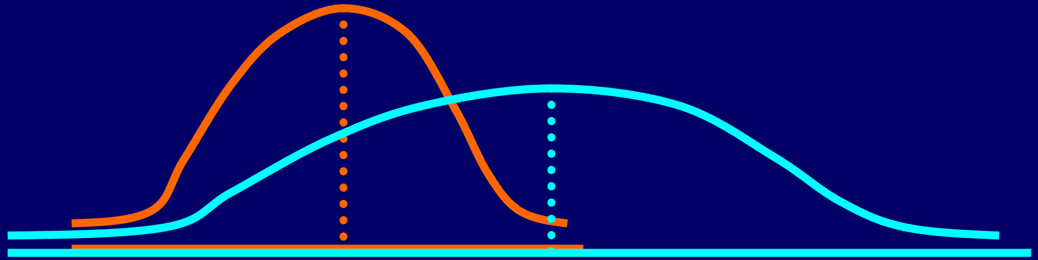
Parameters of Position and Dispersion of one random variable



Different position. Same dispersion.



Different dispersion. Same position.



Different dispersion. Different position.

PARAMETERS OF POSITION

AVERAGE

(mean)

$$\bar{x} = \frac{x_1 + \ldots\ldots + x_N}{N} = \frac{\sum x_i}{N}$$

Sample mean $\bar{x}$

Population mean: m (or μ)

In case of asymmetric data or outliers,

the MEDIAN is better parameter of position than the mean.

MEDIAN

$$\tilde{X} : (\text{No. values} < \tilde{X}) = (\text{No. values} > \tilde{X})$$

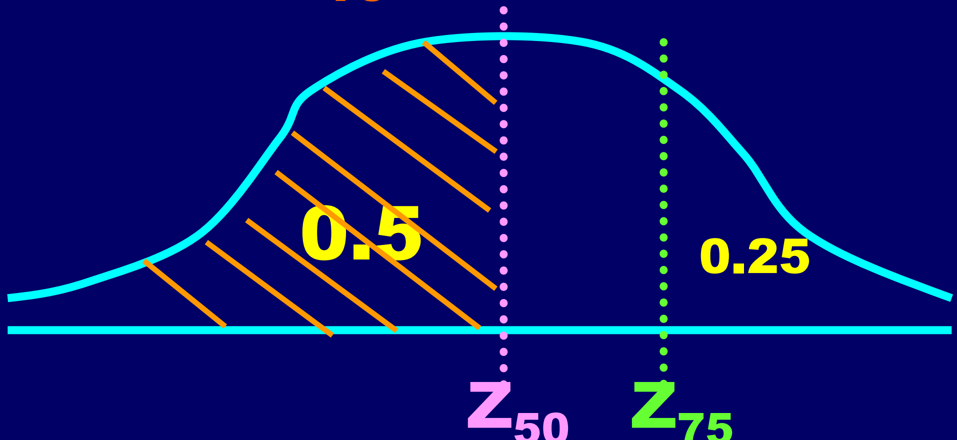
If N even: Average of values in the position $N/2$, $(N/2)+1$

If N odd: Value in the position $(N+1)/2$

Percentile 30 = Z_{30} $\longrightarrow$ $P(X<Z_{30}) = 0.3$

1st quartile (Q1) = Z_{25} $\longrightarrow$ $P(X<Q1) = 0.25$

3rd quartile (Q3) = Z_{75} $\longrightarrow$ $P(X<Q3) = 0.75$



CALCULATION OF MEDIAN AND QUARTILES

- Take all data
- Sort data in increasing order

(N+1)/2

$\bar{x} = 193.3$

order	1	2	3	4	5	6	7	8	9
Resistance	155	169	175	185	191	203	207	225	230
Nº cars	2	3	3	4	4	4	6	9	13
			Q_1		median		Q_3		

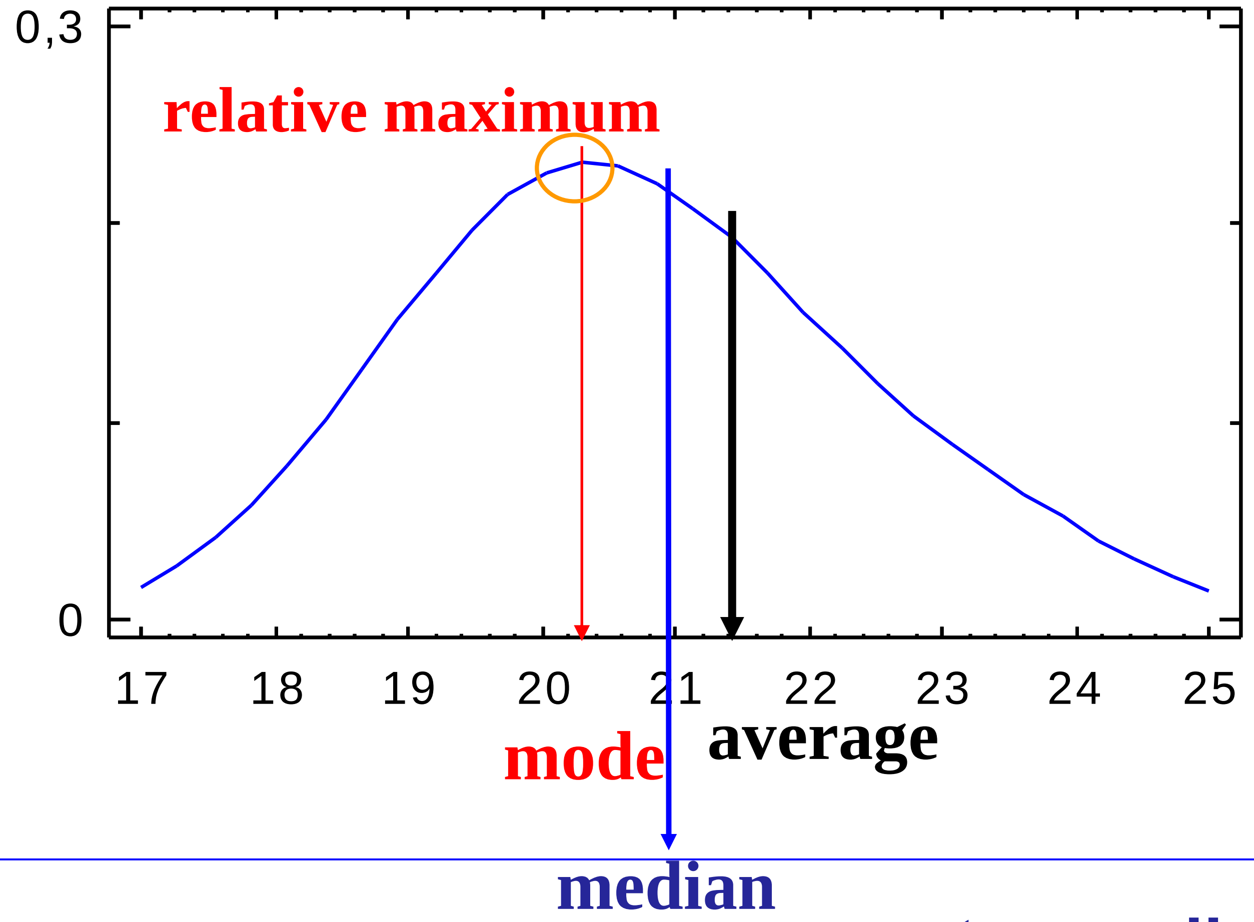
N/2

(N/2)+1

$\bar{x} = 214$

Order	1	2	3	4	5	6	7	8	9	10
Resistance	155	169	175	185	191	203	207	225	230	400
					197					
Nº cars	2	3	3	4	4	4	6	9	13	23
			Q_1		4			Q_3		

See formulary table for exact calculation of Q_1, Q_3



P (X<median) = 0.5

Same value in a Normal distribution

PARAMETERS OF DISPERSION

VARIANCE:

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{N-1} = \frac{\sum X_i^2 - N \cdot \overline{X}^2}{N-1}$$

STANDARD DEVIATION

$$s = \sqrt{s^2}$$ (Same units as data)

INTERQUARTILE RANGE:

$$Z_{75} - Z_{25}$$ = Q3 – Q1

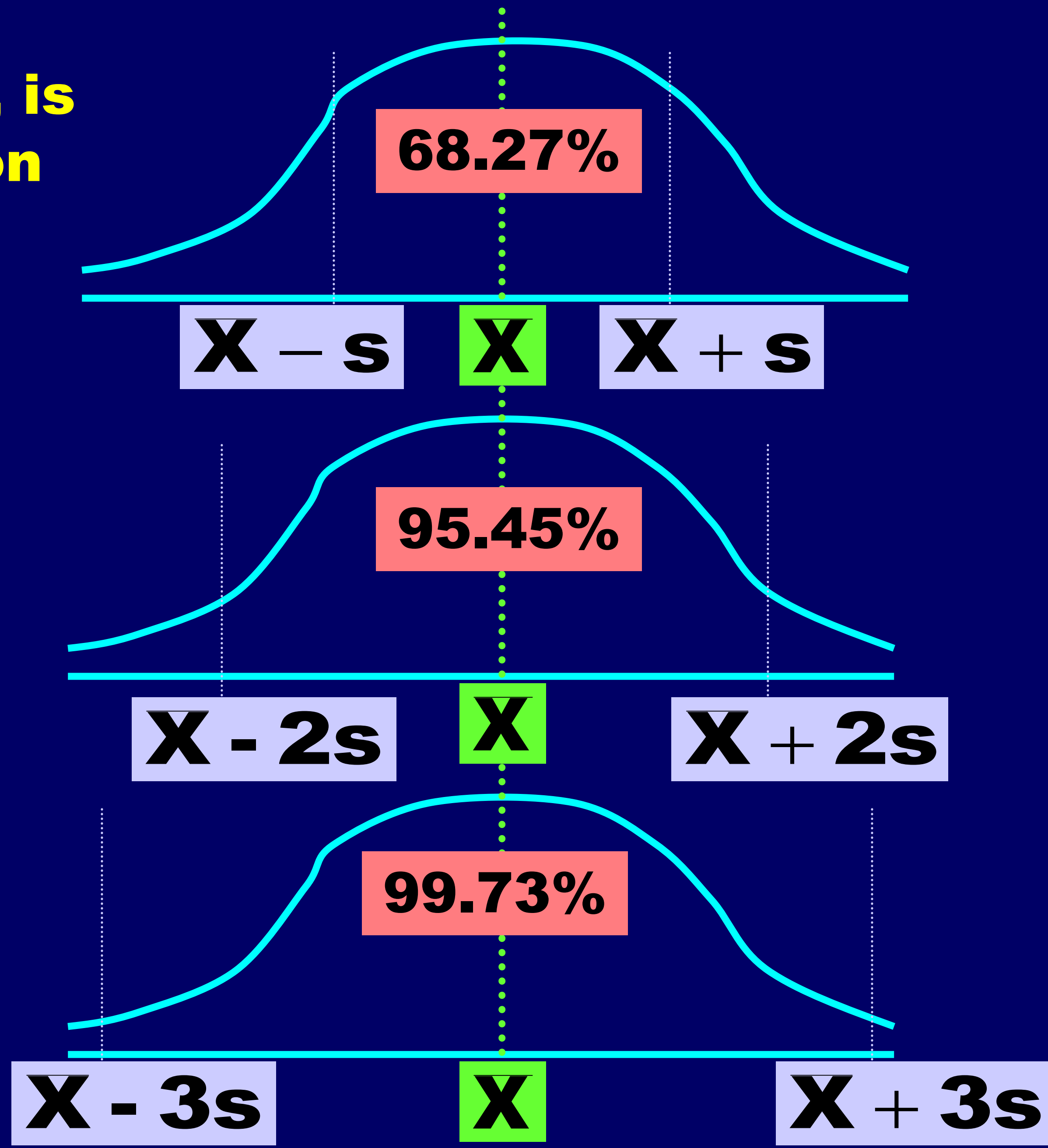
RANGE:

$$R = X_{max} - X_{min}$$

COEFFICIENT OF VARIATION

$$CV = \frac{s}{\overline{X}}$$ (non-dimensional)

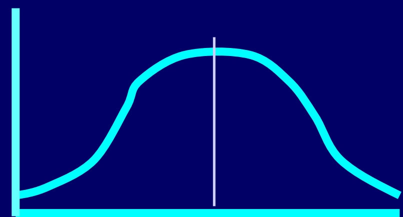
If m=10, s=3, is the dispersion high or low?



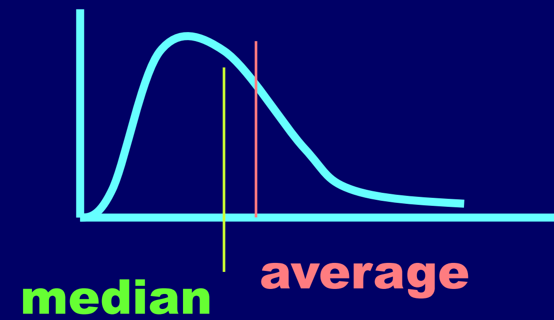
COEFFICIENT OF ASIMMETRY (SKEWNESS):

$$CA = \frac{\sum (X_i - \overline{X})^3 / (N-1)}{s^3}$$

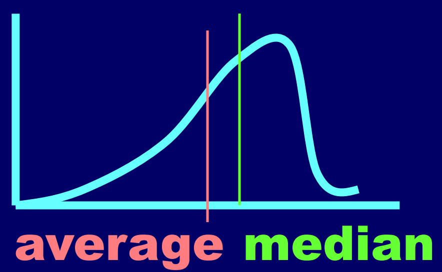
- CA = 0



- CA > 0



- CA < 0

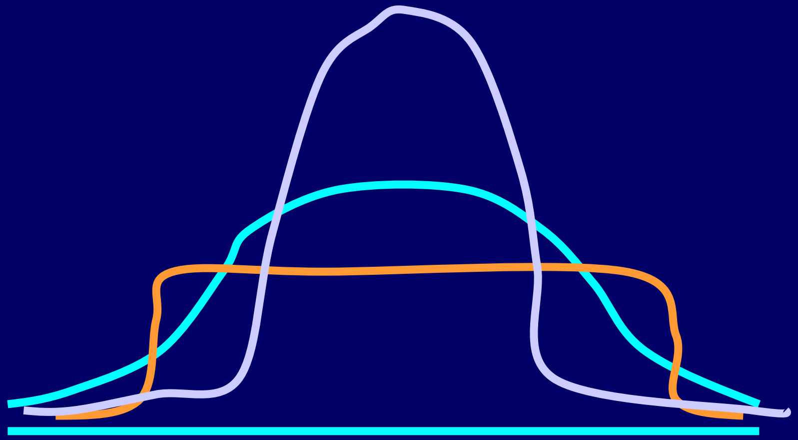


$$CA_{std} = \frac{CA}{\sqrt{6/n}} \Rightarrow \approx N(0;1) \;\; if \;\; n > 150$$

$$\text{If } CA_{std} \notin [-2,\ 2] \Rightarrow \textit{skewed distribution}$$

KURTOSIS COEFFICIENT

$$CC = \frac{\sum (X_i - \overline{X})^4 / (N-1)}{s^4} - 3$$



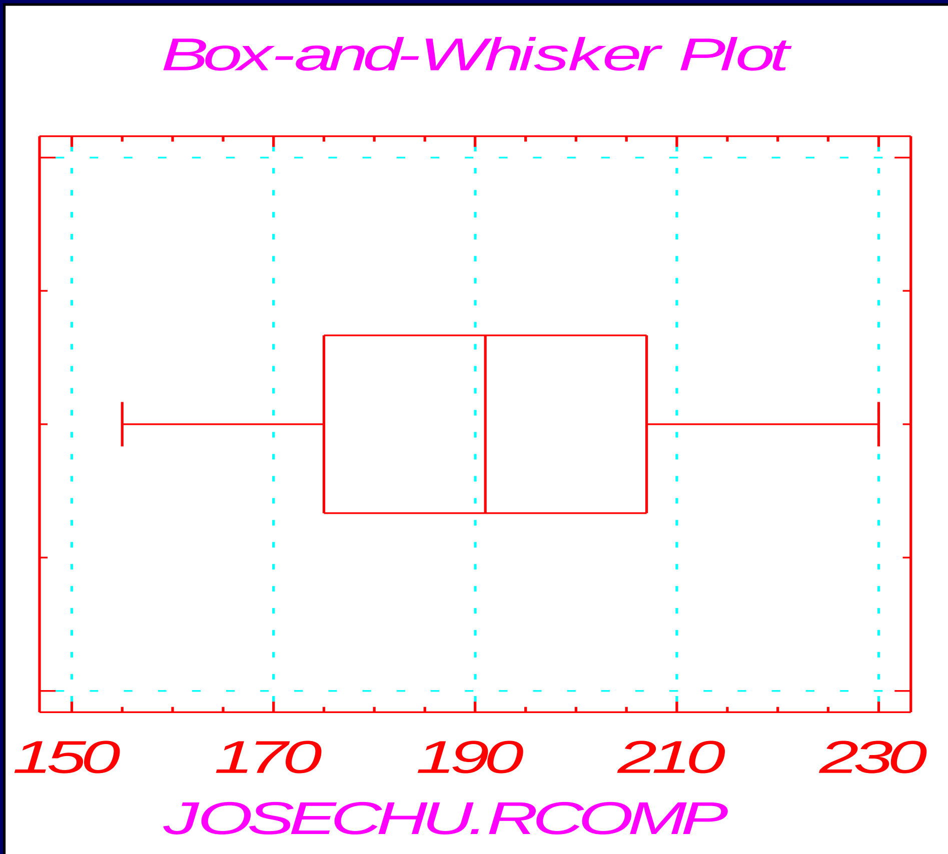
CC=3 (=0) NORMAL DISTRIBUTION

CC>3 (>0) LEPTOKURTIC DATA (e.g. Student's t); OUTLIERS?

CC<3 (<0) PLATIKURTIC DATA. CENSORED DATA?

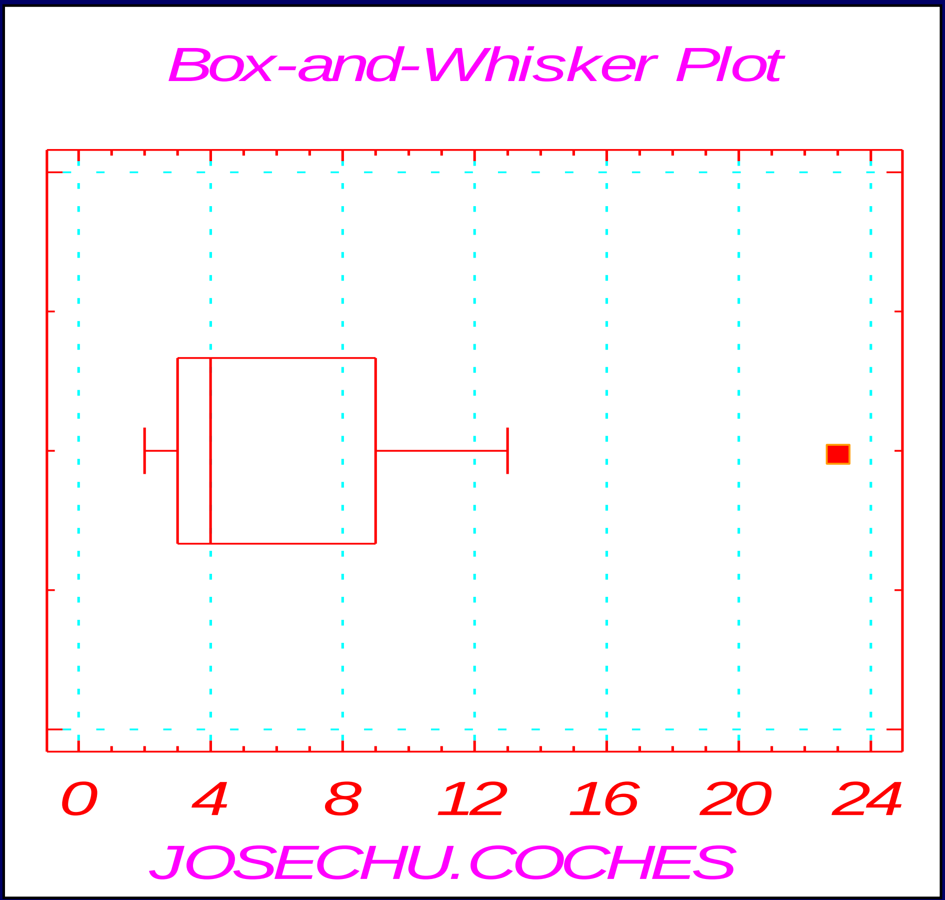
$$\text{If } CC_{std} \in [-2,\ 2] \Rightarrow \textit{Normal distribution}$$

BOX-WHISKER DIAGRAM



Sample size	9
Average	193.333
Median	191
Mode	191
Geometric mean	191.859
Variance	637.5
Standard deviation	25.2488
Standard error	8.4162
Minimum	155
Maximum	230
Range	75
Lower quartile	175
Upper quartile	207
Interquartile range	32
Skewness	0.0700
Standardized skewness	0.0858
Kurtosis	-0.9567
Standardized kurtosis	-0.5858

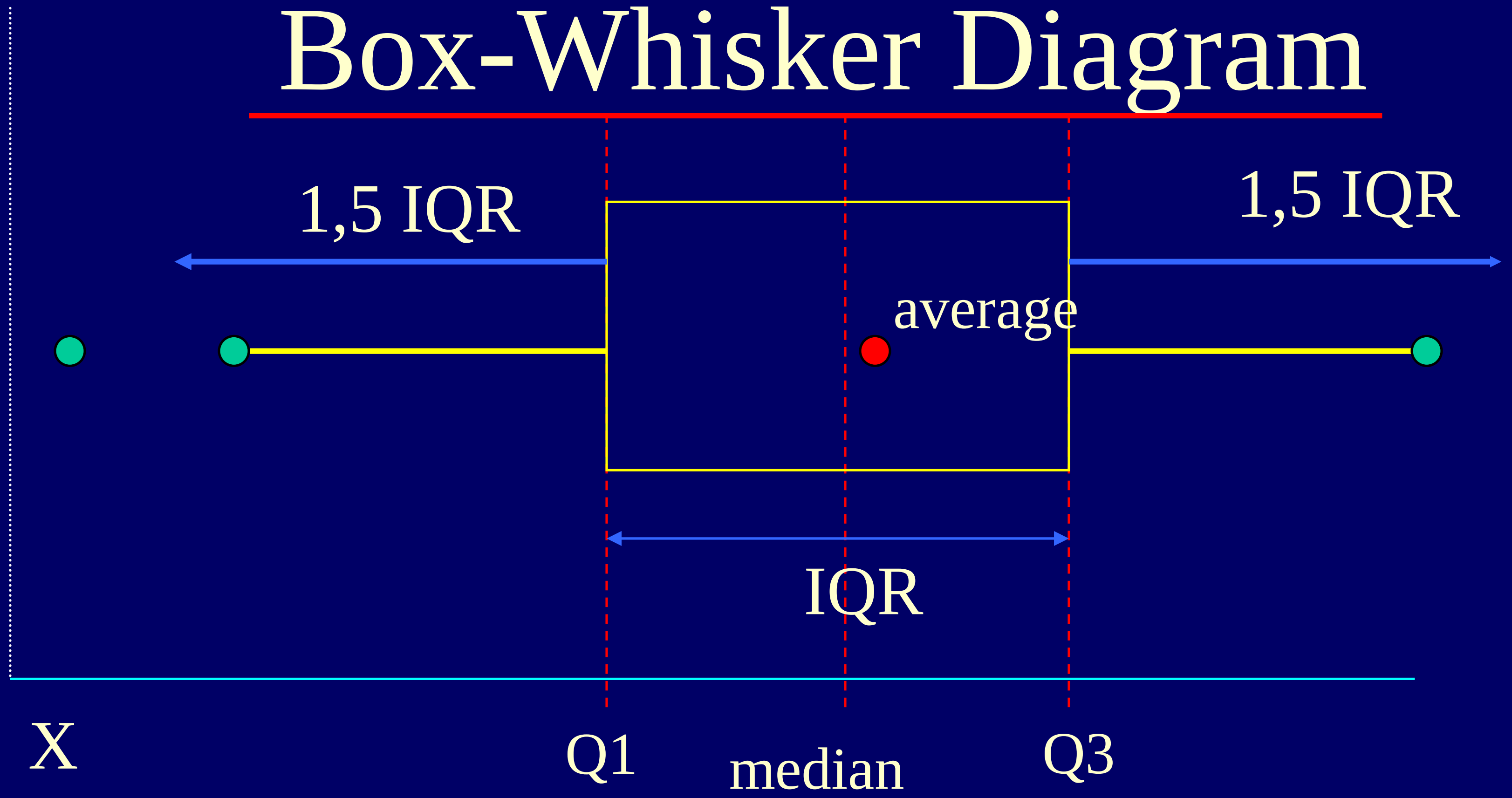
$$CA_{std}\, and\ CC_{std} \in \left[-2,\ 2\right] \Rightarrow Normal\ distribution$$



Variable:	JOSECHU.COCHES
Sample size	10
Average	7.1
Median	4
Mode	4
Geometric mean	5.33276
Variance	42.3222
Standard deviation	6.50555
Standard error	2.05724
Minimum	2
Maximum	23
Range	21
Lower quartile	3
Upper quartile	9
Interquartile range	6
Skewness	1.95257
Standardized skewness	2.52076
Kurtosis	3.78297
Standardized kurtosis	2.4419

$$\mathrm{CA}_{\mathrm{std}} > 2 \Rightarrow \textit{positively skewed distribution}$$

Box-Whisker Diagram



- The "box" comprises 50% of values, from the 1st to 3rd quartile
- The central line corresponds to the median
- The "whiskers" extend from the lowest to the highest observed value except if their distance to the nearest quartile is higher than 1.5 · IQR

Box-Whisker Diagram

- Those extreme values that differ from the nearest quartile more than 1.5 IQR are plotted as isolated points to highlight that they **might be outliers.**

Outlier: an abnormal datum that does not belong to the same population, "it lies out" of the rest. They are usually eliminated.

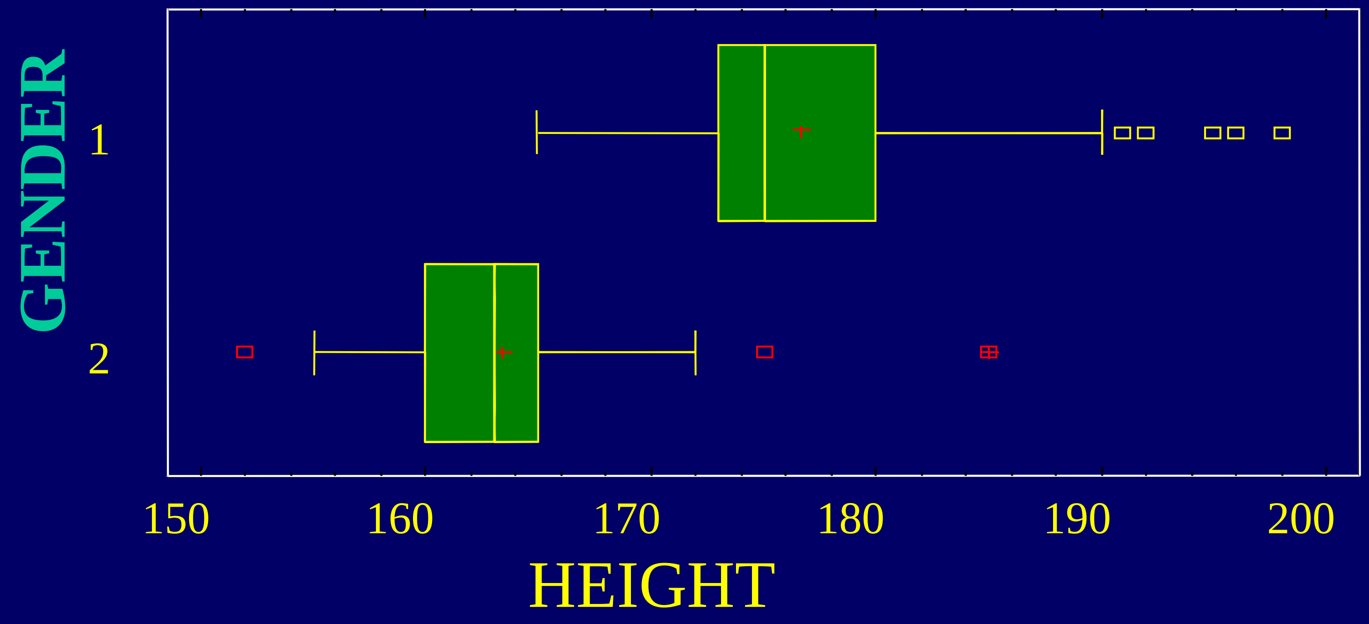
Not all isolated points are outliers !!

In a Normal distribution, isolated points in the box-whisker plot "quite close" to the end of a whisker are not outliers.

(check this by simulating 1000 Normal data with Statgraphics)

To check if a high value in a positive skewed distribution is an outlier: represent data on a Normal Probability Plot using transformations: $X^{0.5}$; $X^{0.25}$; log(x)

(check this by simulating 100 Chi^2 data with Statgraphics)



Is there any <u>outlier</u> in women's data?

Calculate the interquartile range of men's height

Calculate the range of women's height

Is the distribution of men's data asymmetric? Positive / negative?

EXERCISE: draw a box-whisker plot with the following data:
16; 8; 90; 22; 2; 50; 5; 30; 11

(check formula table for the exact value of Q1 and Q3)

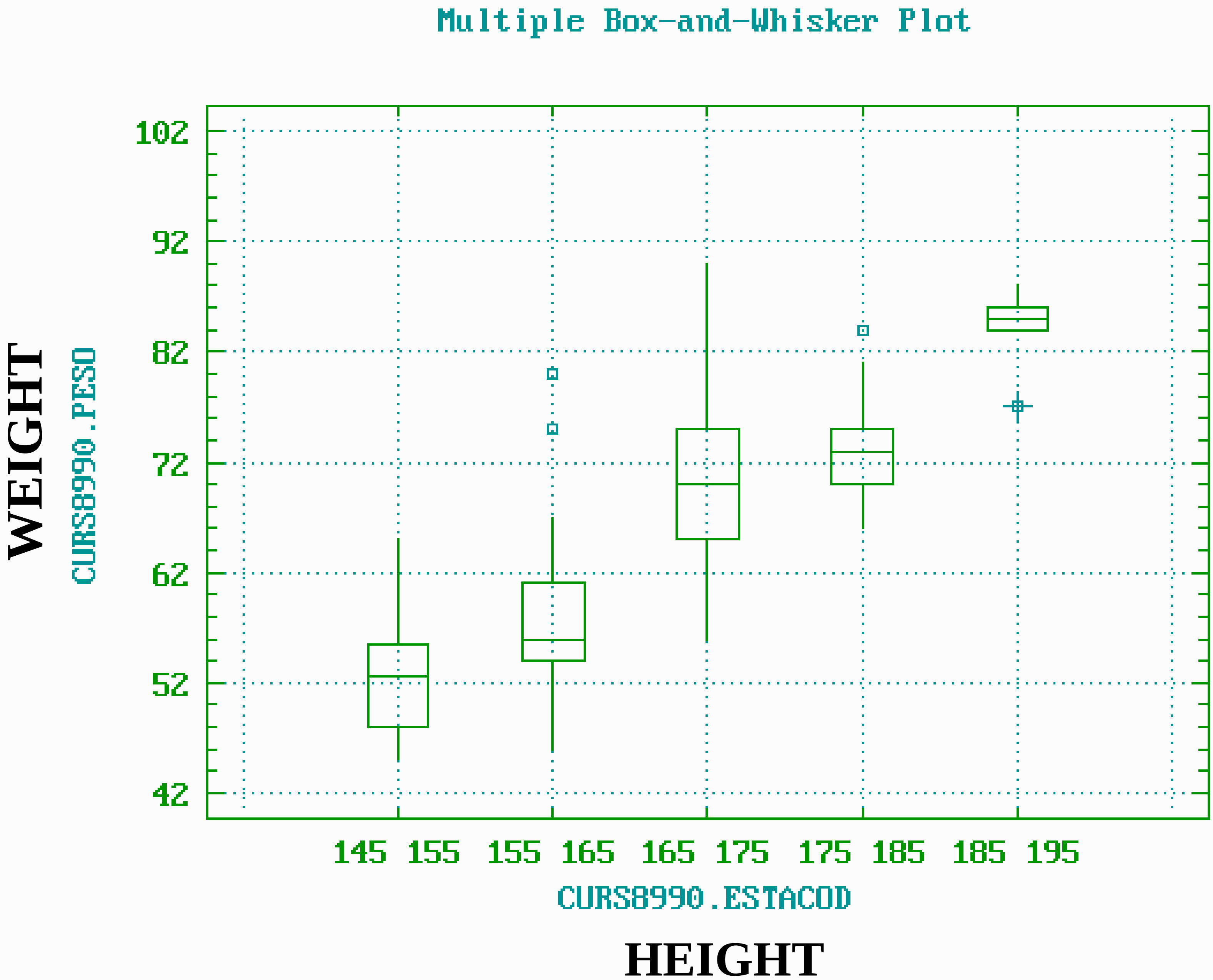
Calculate the range and the interquartile range

Describe the distribution (symmetric, CA>0, CA<0)

Is there any outlier that should be discarded?

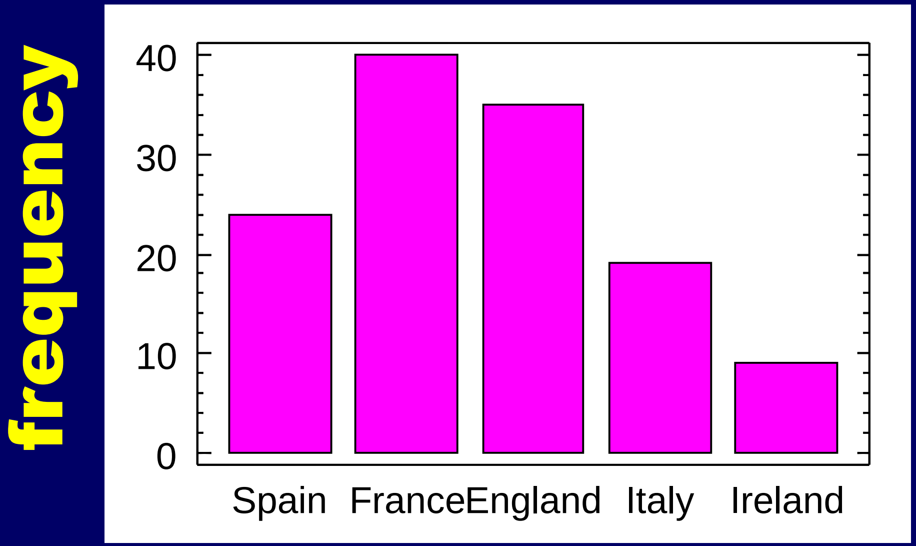
- Plot data on a Normal Probability Plot
- Use transformations: $X^{0.5}$; $X^{0.25}$; log(x)

Change 90 by 500; is it an outlier?

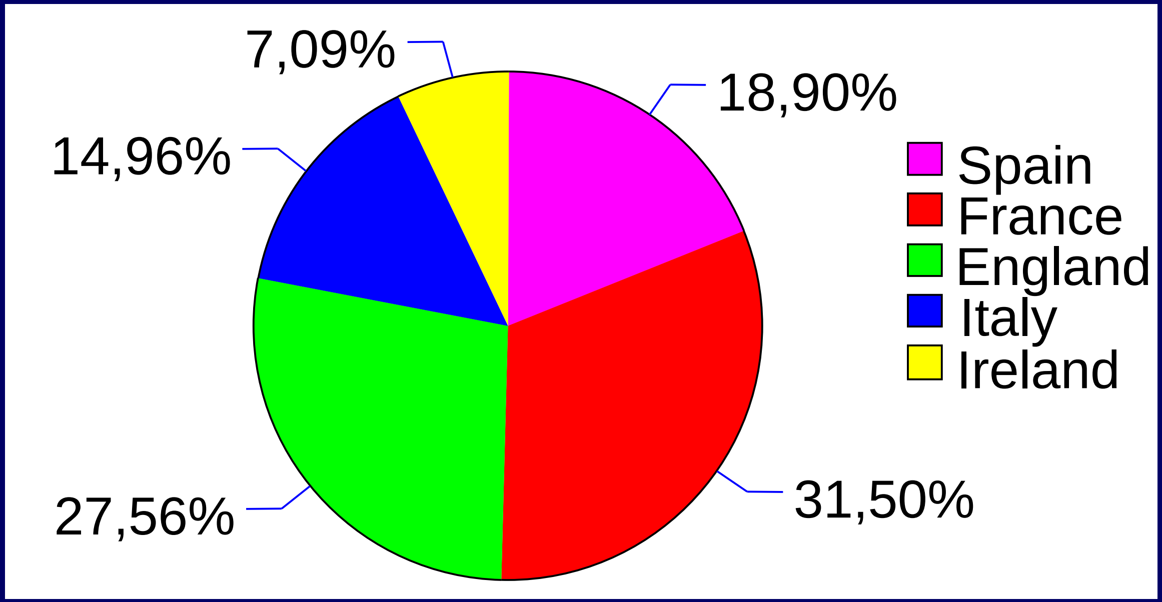


Multiple Box-and-Whisker Plot
102
92
82
72
62
52
42
WEIGHT
CURS8990.PESO
145 155
155 165
165 175
175 185
185 195
CURS8990.ESTACOD
HEIGHT

BARCHART



PIECHART



TWO-DIMENSIONAL DESCRIPTIVE STATISTICS

TWO-DIMENSIONAL RANDOM VARIABLES

WHEN TWO RANDOM NUMERIC CHARACTERISTICS ARE OBSERVED FROM EACH INDIVIDUAL, WE HAVE A TWO-DIMENSIONAL RANDOM VARIABLE.

X	Y
174	184
169	178
183	167
168	186

EXERCISE

Are these 2 one-dimensional variables or one two-dimensional variable?

- **length of pieces from supplier A (X) and supplier B (Y)**
- **In a married couple, the height of husband (X) and wife (Y)**
- **The height of students from Valencia (X) and Madrid (Y)**
- Time (ms) taken by algorithm X and Y to invert different matrixes

TWO-DIMENSIONAL VARIABLES:
CONTINGENCY TABLES

- THEY ALLOW TO STUDY THE RELATIONSHIP BETWEEN THE TWO COMPONENTS
- IF ONE OF THE VARIBLES IS CONTINUOUS, IT WILL BE REGROUPED IN INTERVALS.

gender	REPEAT: YES	REPEAT: NO	Row Total
MALE	5 10.9	41 89.1	
FEMALE	1 4.0	24 96.0	
COLUMN TOTAL	6 8.5	65 91.5	71

Relative frequency of repeat conditioned to gender

Marginal frequency of repeat

GENDER	REPEAT YES	REPEAT NO	Row Total
MALE	5 83.3	41 63.1	46 64.8
FEMALE	1 16.7	24 36.9	25 35.2
COLUMN TOTAL	6	65	71

Relative frequency of gender conditioned to repeat

Marginal frequency of gender

Marginal frequencies:

Frequency of each value of one variable without taking into account the other

Relative conditional frequencies:

Relative frequency of the value of one variable in relation to each value of the other

QUALITATIVE VARIABLES:

BY MEANS OF A CONTINGENCY TABLE.

REPEAT / GENDER	YES 1		NO 2		Row Total
MALE 1	5		41		46
	83.3	10.9	63.1	89.1	64.8
FEMALE 2	1		24		25
	16.7	4.0	36.9	96.0	35.2
COLUMN TOTAL	6		65		71
	8.5		91.5		

Marginal frequency of gender

Marginal frequency of repeat

Relative frequency of gender conditined to repeat

Relative frequency of repeat conditioned to gender

QUANTITATIVE VARIABLES:

BY MEANS OF A CONTINGENCY TABLE AFTER GROUPING THE DATA IN INTERVALS.

PROBLEM: SOME INFORMATION IS LOST IN THE TABULATION

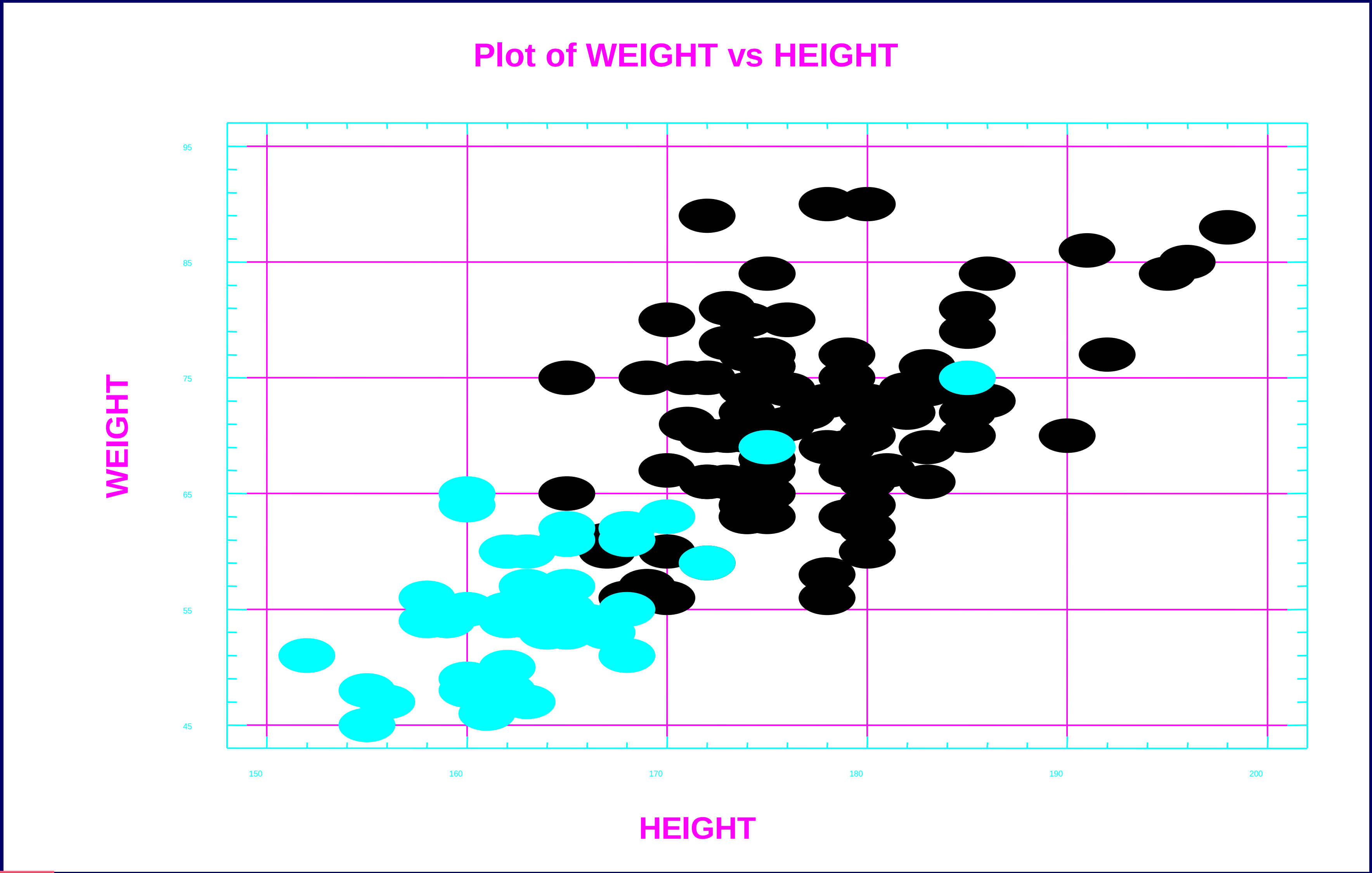
HEIGHT / WEIGHT	145 155	155 165	165 175	175 185	185 195	Row Total
40 55	9 75.0	17 44.7	0 .0	0 .0	0 .0	26 20.0
55 70	3 25.0	18 47.4	31 53.4	5 29.4	0 .0	57 43.8
70 85	0 .0	3 7.9	24 41.4	12 70.6	3 60.0	42 32.3
85 99	0 .0	0 .0	3 5.2	0 .0	2 40.0	5 3.8
Column Total	12 9.2	38 29.2	58 44.6	17 13.1	5 3.8	130 100

Marginal frequency of weight

Marginal frequency of height

Relative frequency of weight conditioned to height

SCATTERPLOT



EXERCISES:

in PoliformaT at:

recursos \ 04-ejercicios \ ejercicios resueltos \ ejercicios UD2.pdf