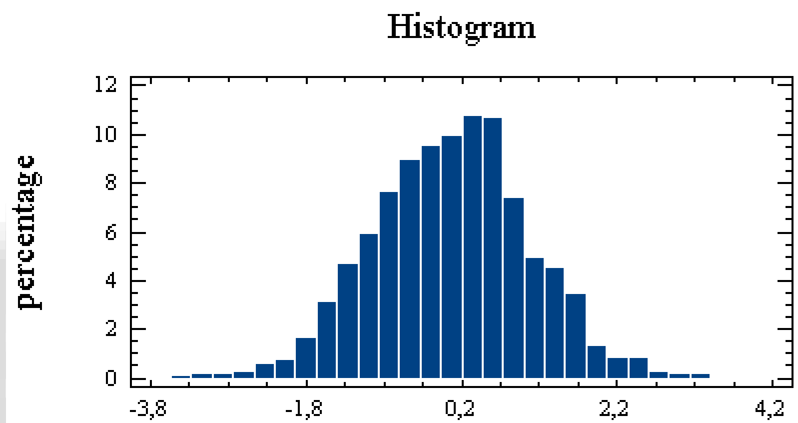
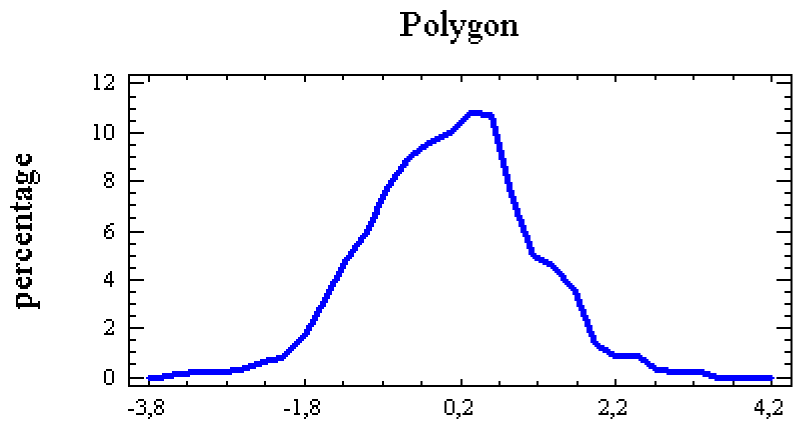
I. UNIVARIATE DESCRIPTIVE STATISTICS





Chapter I: Univariate Descriptive Statistics

1. Introduction. The purpose of Statistics.
2. Description of data by tables
3. Description of data by graphs
4. Characteristics measures of a variable

1. Introduction. The purpose of Statistics

What is Statistics? Why we study Statistics?

Gaining understanding through the observation

From a small number of data we obtain general conclusions

Real phenomenon

Actual Data → **Statistical analysis** → **Learning about the phenomenon**

Two alternative ways to get knowledge of the world

By theory

- Physics laws
- Mathematical rules
- Properties of ideal materials

From theoretical models DEDUCE the reality

DEDUCTION = to get consequences from a principle, a proposition or an assumptions.

By observation

- Data
- Statistics

From the data INDUCE or INFER a model (empirical)

INDUCTION = to draw, from specific observations or particular experiences, the underlying general principle.

Chapter I: Univariate Descriptive Statistics

1. Introduction. The purpose of Statistics.
2. Description of data by tables
3. Description of data by graphs
4. Characteristics measures of a variable

2. Description of data by tables

Objective: summarize the information to make easier its analysis

Univariate tables

They show the frequencies of each actual value

Example 1: number of cylinders of 155 cars (file cardata.sf)

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	3	1	0,0065	1	0,0065
2	4	104	0,6710	105	0,6774
3	5	3	0,0194	108	0,6968
4	6	30	0,1935	138	0,8903
5	8	17	0,1097	155	1,0000

2. Description of data by tables

Univariate tables

Example 2: month of birthday of 95 first-course students

Class	Value	Frequency	Relative Frequency
1	Enero	15	0,1579
2	Febrero	5	0,0526
3	Marzo	10	0,1053
4	Abril	9	0,0947
5	Mayo	10	0,1053
6	Junio	13	0,1368
7	Julio	9	0,0947
8	Agosto	7	0,0737
9	Septiembre	6	0,0632
10	Octubre	1	0,0105
11	Noviembre	3	0,0316
12	Diciembre	7	0,0737

2. Description of data by tables

univariate tables

If there are a lot of different values they are grouped into intervals/classes

Example: price of 155 cars (file cardata.sf)

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		0,0		0	0,0000	0	0,0000
1	0,0	2000,0	1000,0	1	0,0065	1	0,0065
2	2000,0	4000,0	3000,0	70	0,4516	71	0,4581
3	4000,0	6000,0	5000,0	60	0,3871	131	0,8452
4	6000,0	8000,0	7000,0	14	0,0903	145	0,9355
5	8000,0	10000,0	9000,0	8	0,0516	153	0,9871
6	10000,0	12000,0	11000,0	0	0,0000	153	0,9871
7	12000,0	14000,0	13000,0	0	0,0000	153	0,9871
8	14000,0	16000,0	15000,0	2	0,0129	155	1,0000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,0000
above	18000,0			0	0,0000	155	1,0000

How many classes we should use? Empirically $\sqrt{n}$, where n is the sample size

Chapter I: Univariate Descriptive Statistics

1. Introduction. The purpose of Statistics.
2. Description of data by tables
3. **Description of data by graphs**
4. Characteristics measures of a variable

3. Description of date by graphs

3.1 Bar Charts

3.2 Histograms and polygon of frequencies

3.3 Pie Charts

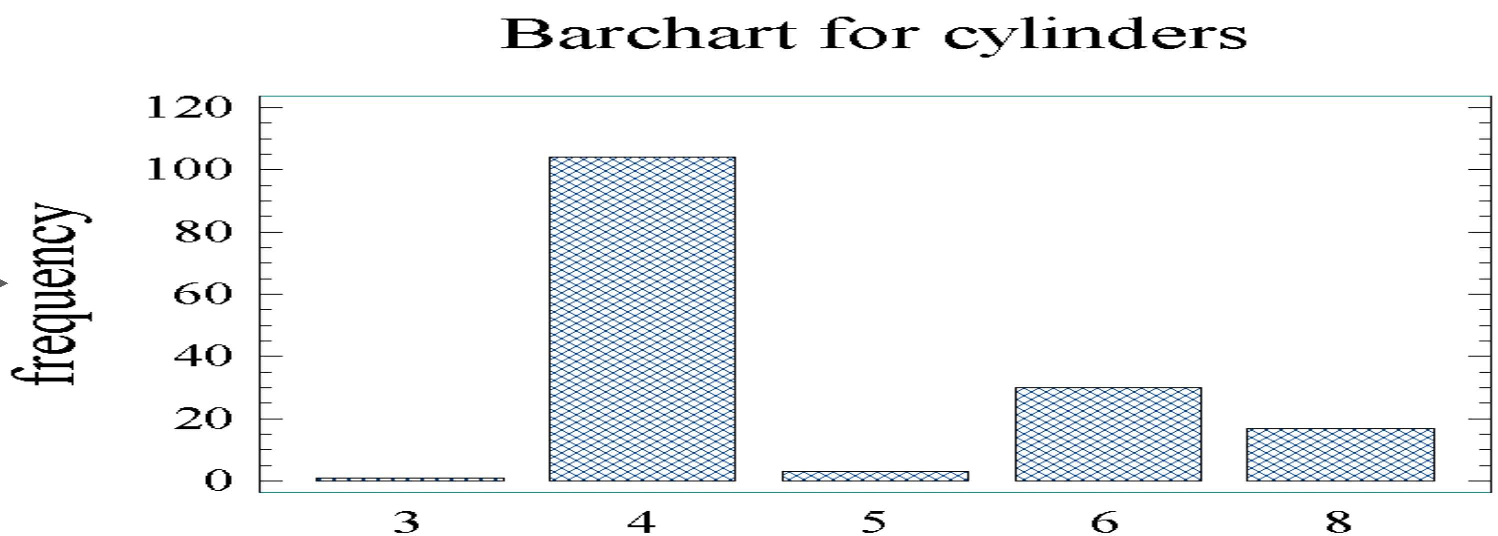
3.4 Time Series

3.1 Bar Chart

A Bar Chart is the graph representation of a frequency table containing categorical data types

Example: number of cylinders of 155 cars (file cardata.sf)

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	3	1	0,0065	1	0,0065
2	4	104	0,6710	105	0,6774
3	5	3	0,0194	108	0,6968
4	6	30	0,1935	138	0,8903
5	8	17	0,1097	155	1,0000

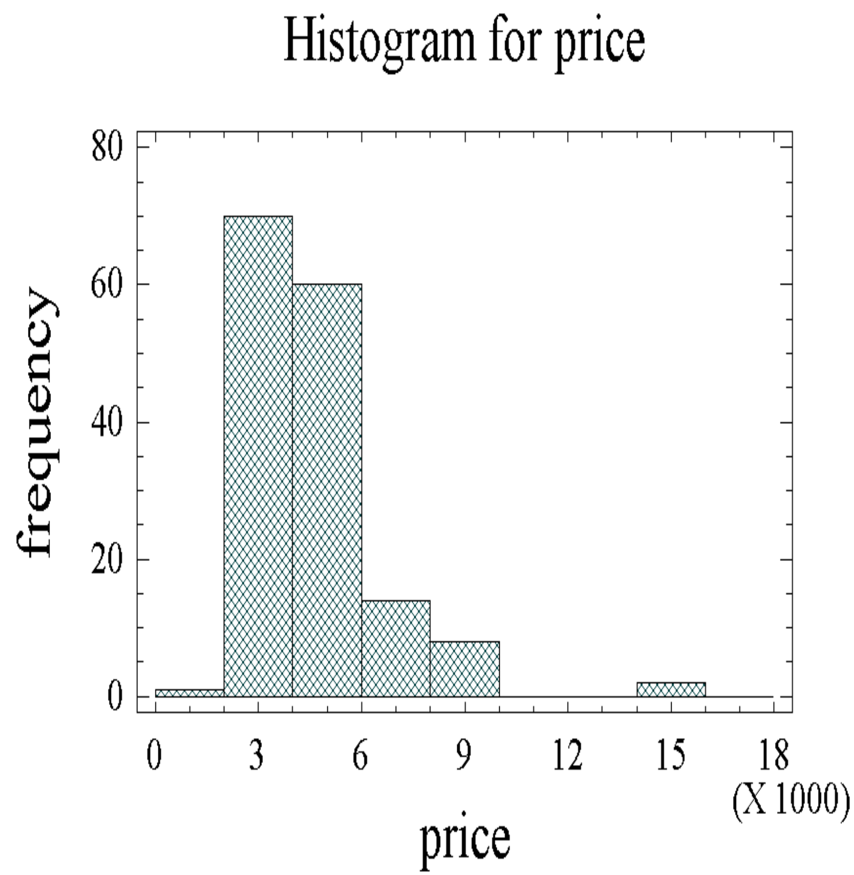


3.2 Frequency Histogram and Frequency Polygon

The Frequency Histogram is the graph representation of a frequency table whose data is grouped into intervals

Example: price of 155 cars (file cardata.sf)

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		0,0		0	0,0000	0	0,0000
1	0,0	2000,0	1000,0	1	0,0065	1	0,0065
2	2000,0	4000,0	3000,0	70	0,4516	71	0,4581
3	4000,0	6000,0	5000,0	60	0,3871	131	0,8452
4	6000,0	8000,0	7000,0	14	0,0903	145	0,9355
5	8000,0	10000,0	9000,0	8	0,0516	153	0,9871
6	10000,0	12000,0	11000,0	0	0,0000	153	0,9871
7	12000,0	14000,0	13000,0	0	0,0000	153	0,9871
8	14000,0	16000,0	15000,0	2	0,0129	155	1,0000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,0000
above	18000,0			0	0,0000	155	1,0000



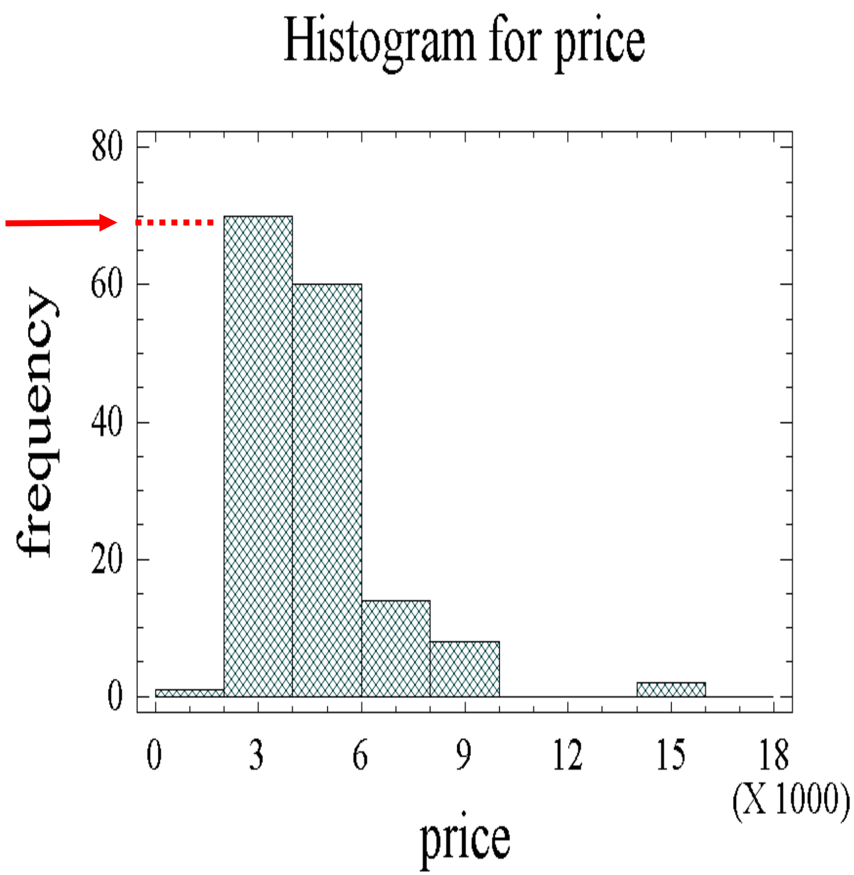
The histogram is one of the most useful graphical tools to summarize information

3.2 Frequency Histogram and Frequency Polygon

The Frequency Histogram is the graph representation of a frequency table whose data is grouped into intervals

Example: price of 155 cars (file cardata.sf)

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		0,0		0	0,0000	0	0,0000
1	0,0	2000,0	1000,0	1	0,0065	1	0,0065
2	2000,0	4000,0	3000,0	70	0,4516	71	0,4581
3	4000,0	6000,0	5000,0	60	0,3871	131	0,8452
4	6000,0	8000,0	7000,0	14	0,0903	145	0,9355
5	8000,0	10000,0	9000,0	8	0,0516	153	0,9871
6	10000,0	12000,0	11000,0	0	0,0000	153	0,9871
7	12000,0	14000,0	13000,0	0	0,0000	153	0,9871
8	14000,0	16000,0	15000,0	2	0,0129	155	1,0000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,0000
above	18000,0			0	0,0000	155	1,0000



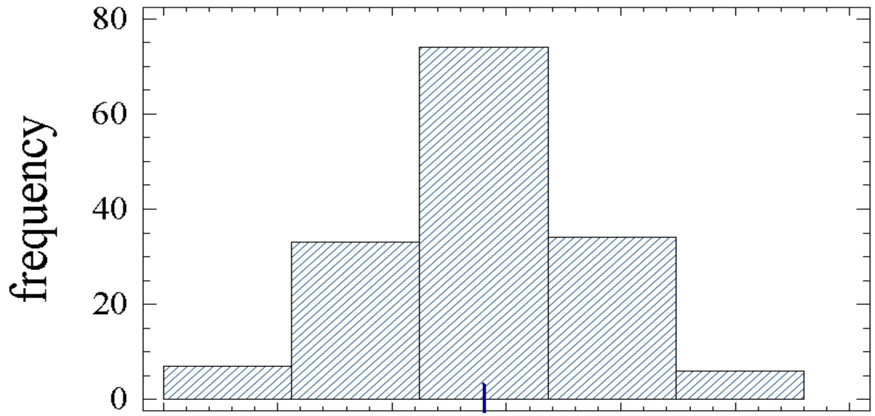
The histogram is useful to summarize the following information:

- Concentrations
- Asymmetries
- Gaps
- Outliers

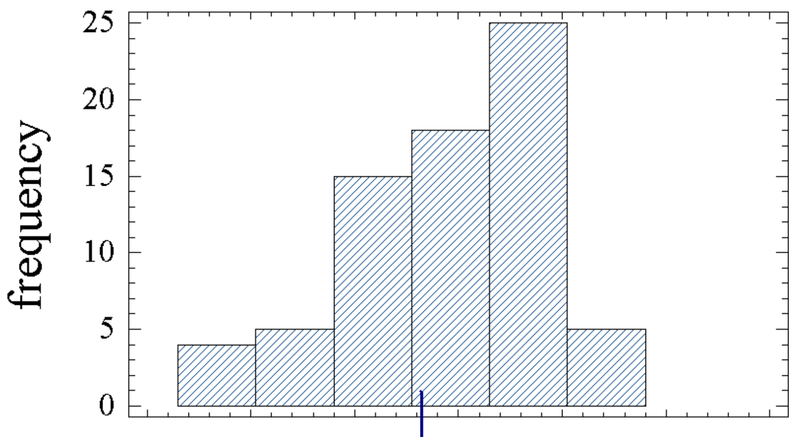
3.2 Frequency Histogram and Frequency Polygon

The Frequency Histogram is the graph representation of a frequency table whose data is grouped into intervals

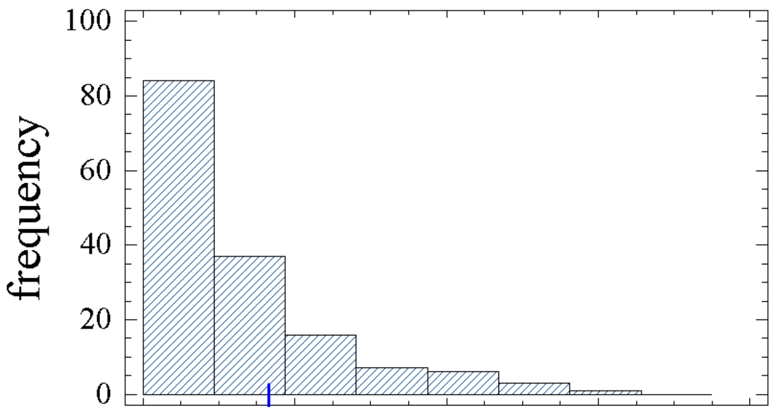
Symmetric distribution



Negative asymmetric distribution
Skewness to the left

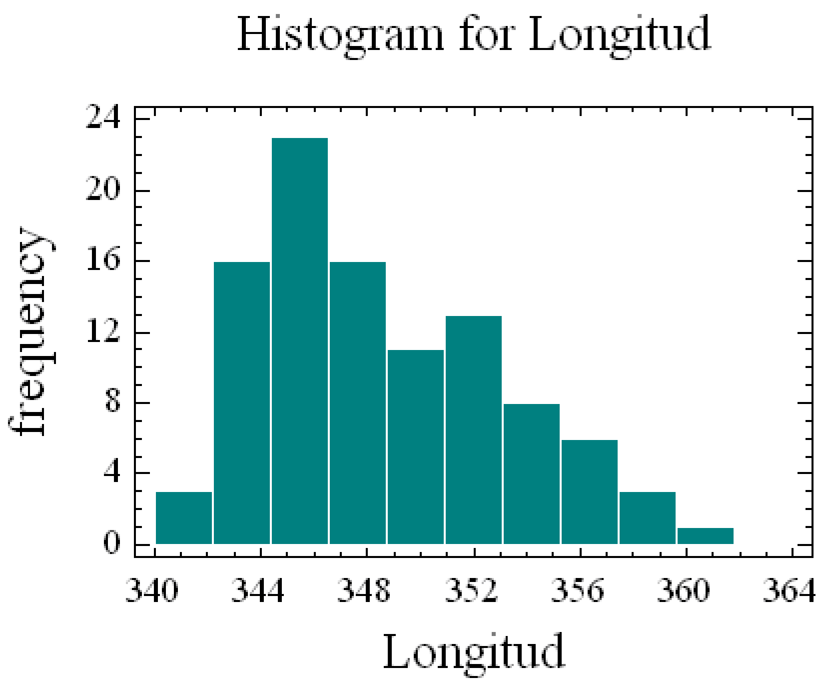


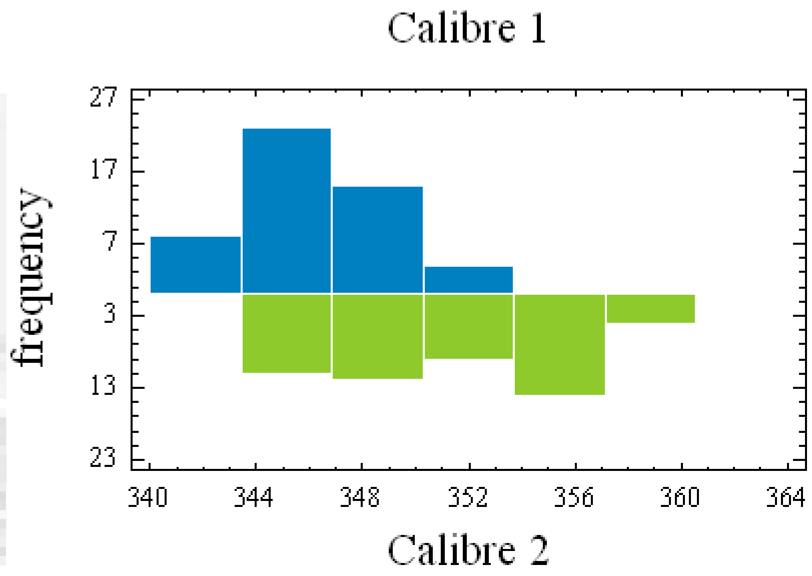
Positive asymmetric distribution
Skewness to the right



Example:

The lengths of 100 nails of same type, measured by two guys with different calibres, 50 nails each.





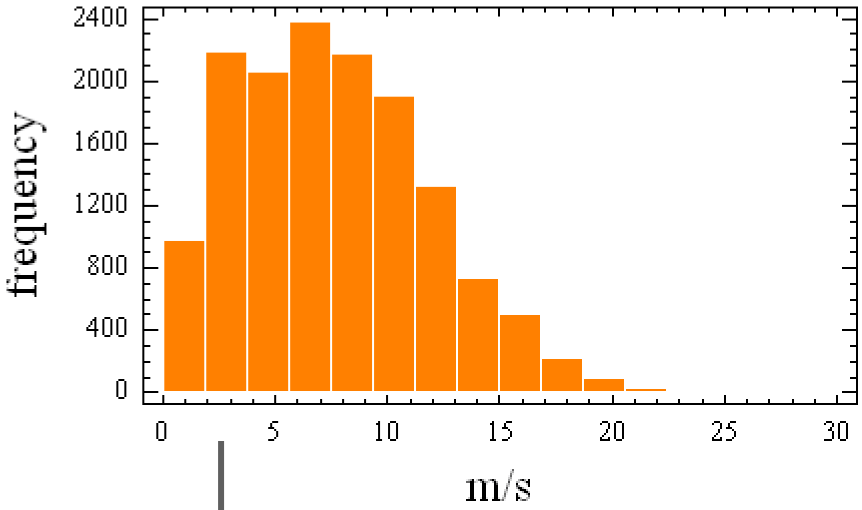
The two concentrations of values seem to be due to the two different calibres

What calibre seems better?

Example:

Speed values of the wind (m/s) registered in a wind power station for several months. Each data is the average speed registered during one hour. We have 14000 data.

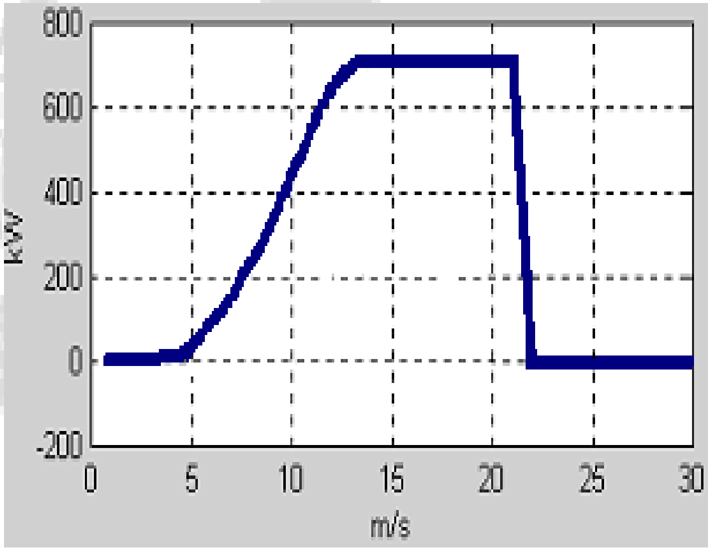
Speed Histogram



¿Is there a concentration around of 2.5 m/s?

(at 2.5 m/s the wind turbine does not produce any energy)

Power generated by a wind turbine like a function of the speed wind

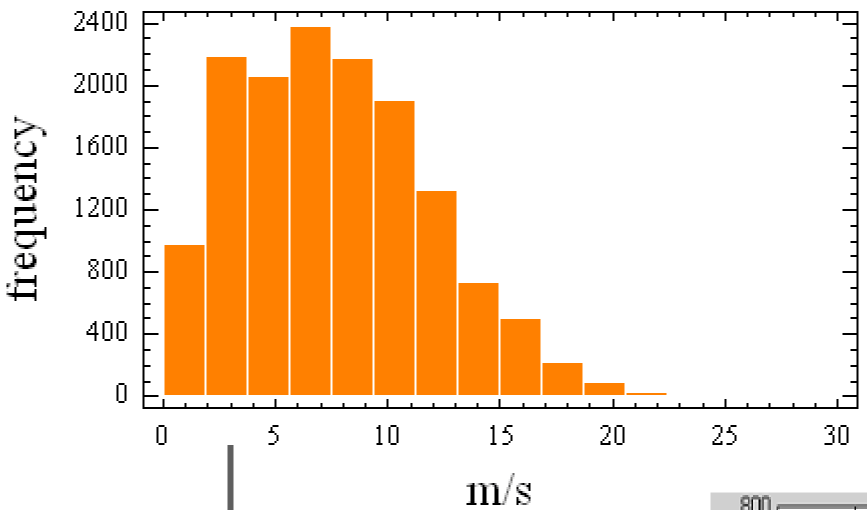


Example:

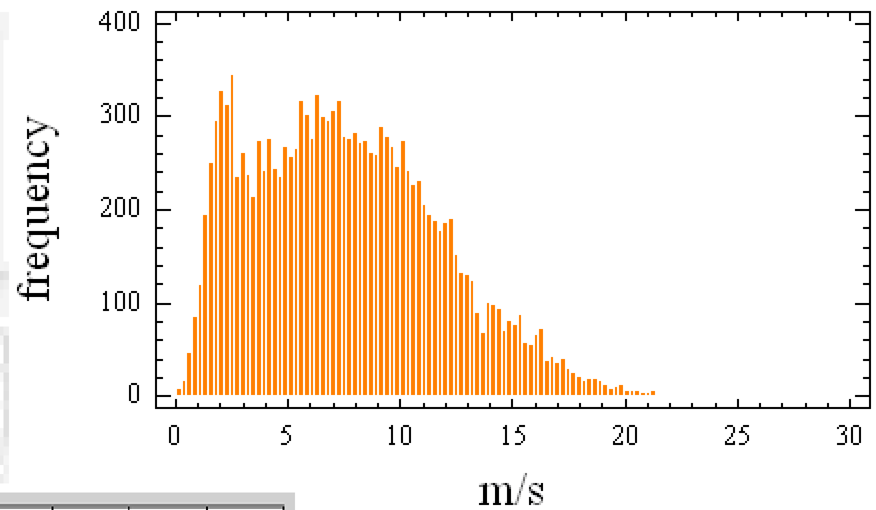
Speed values of the wind (m/s) were registered in a wind power station for several months. Each data is the average speed registered during one hour. We have $\boxed{14000}$ data.

$\sqrt{14000} \approx 118$ classes

Speed Histogram

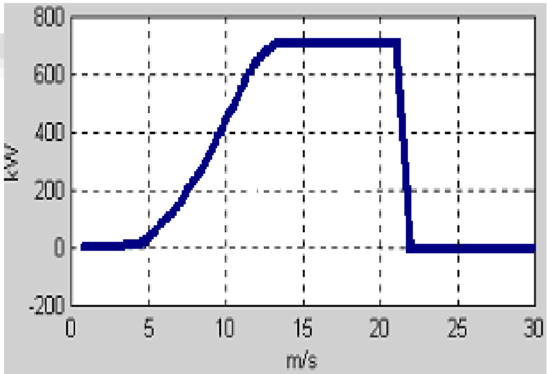


Speed Histogram



¿Is there a concentration around of 2.5 m/s?

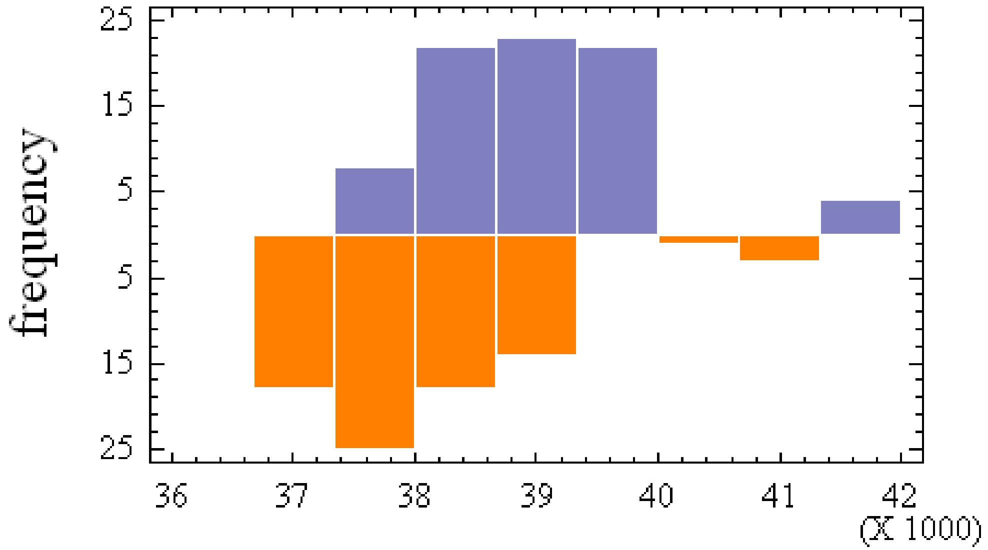
(a 2.5 m/s the wind turbine do not produce energy)



Example:

Time that a computer need to write a file of 300 Mb in its hard disk. Two experiments are doing; in one the hard disk is defragmented, in the other, the 40% of hard disk is fragmented. Each experiment is doing again 79 times

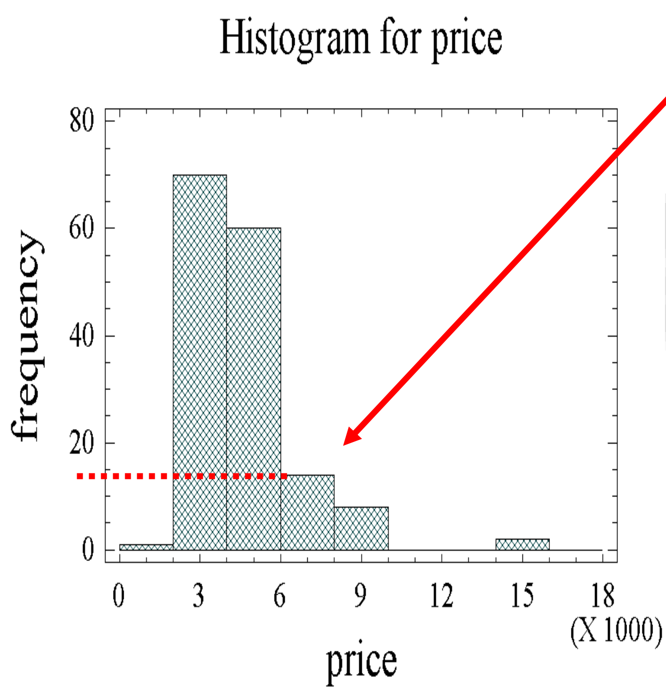
Fragmented Hard Disk

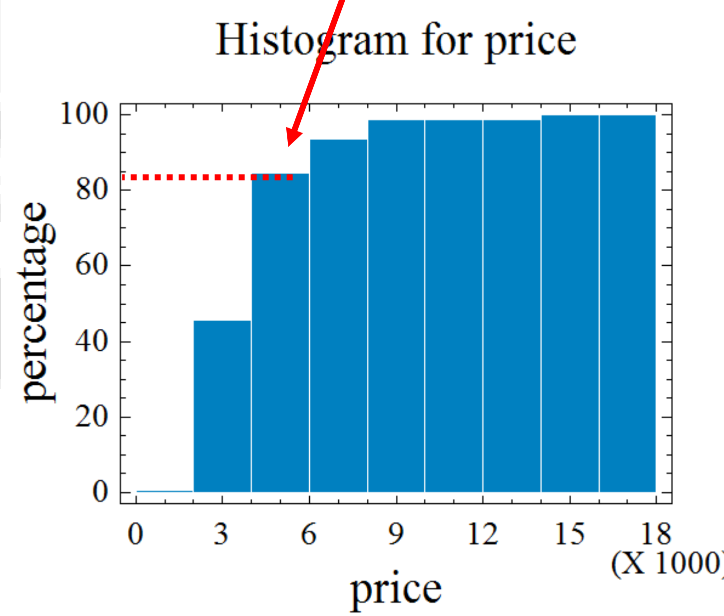


Defragmented Hard Disk

We cuse the histogram also to describe the cumulative frequancies. Also in this case it can be expressed in relative or absolute values

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		0,0		0	0,0000	0	0,0000
1	0,0	2000,0	1000,0	1	0,0065	1	0,0065
2	2000,0	4000,0	3000,0	70	0,4516	71	0,4581
3	4000,0	6000,0	5000,0	60	0,3871	131	0,8452
4	6000,0	8000,0	7000,0	14	0,0903	145	0,9355
5	8000,0	10000,0	9000,0	8	0,0516	153	0,9871
6	10000,0	12000,0	11000,0	0	0,0000	153	0,9871
7	12000,0	14000,0	13000,0	0	0,0000	153	0,9871
8	14000,0	16000,0	15000,0	2	0,0129	155	1,0000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,0000
above	18000,0			0	0,0000	155	1,0000





3.2 Frequency Histogram and Frequency Polygon

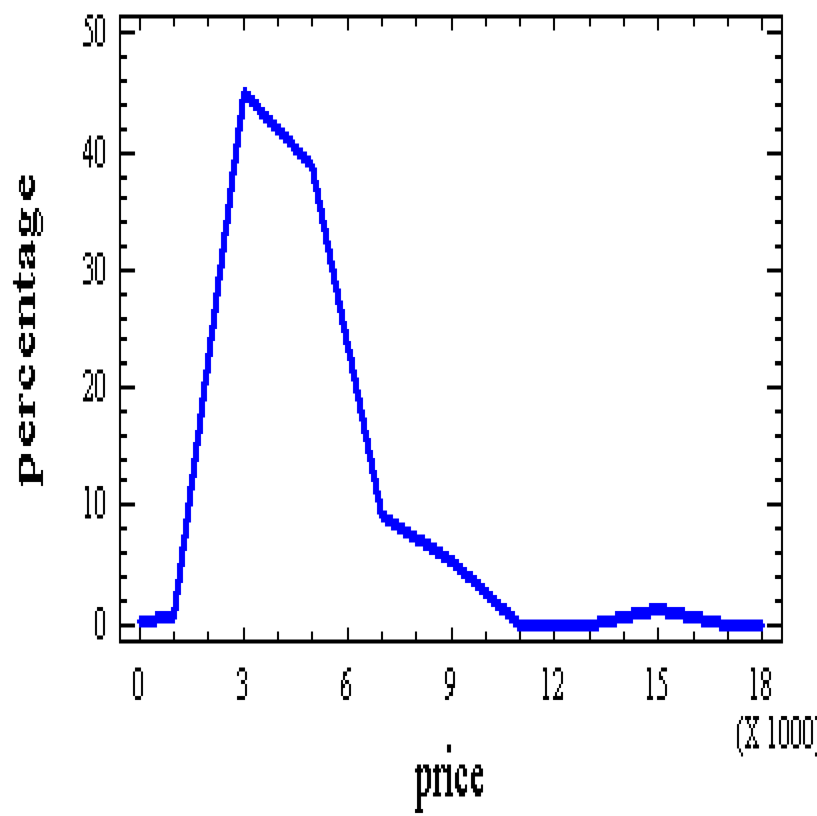
The Frequency Polygon is the graph representation of a frequency table whose data is grouped into intervals

Example: price of 155 cars (file cardata.sf)

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel Frequenc
at or below		0,0		0	0,0000	0	0,000
1	0,0	2000,0	1000,0	1	0,0065	1	0,006
2	2000,0	4000,0	3000,0	70	0,4516	71	0,458
3	4000,0	6000,0	5000,0	60	0,3871	131	0,845
4	6000,0	8000,0	7000,0	14	0,0903	145	0,935
5	8000,0	10000,0	9000,0	8	0,0516	153	0,987
6	10000,0	12000,0	11000,0	0	0,0000	153	0,987
7	12000,0	14000,0	13000,0	0	0,0000	153	0,987
8	14000,0	16000,0	15000,0	2	0,0129	155	1,000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,000
above	18000,0			0	0,0000	155	1,000

The polygon of frequencies is obtained by linking with lines the top midpoints of the histogram.

Frequency Polygon for Price



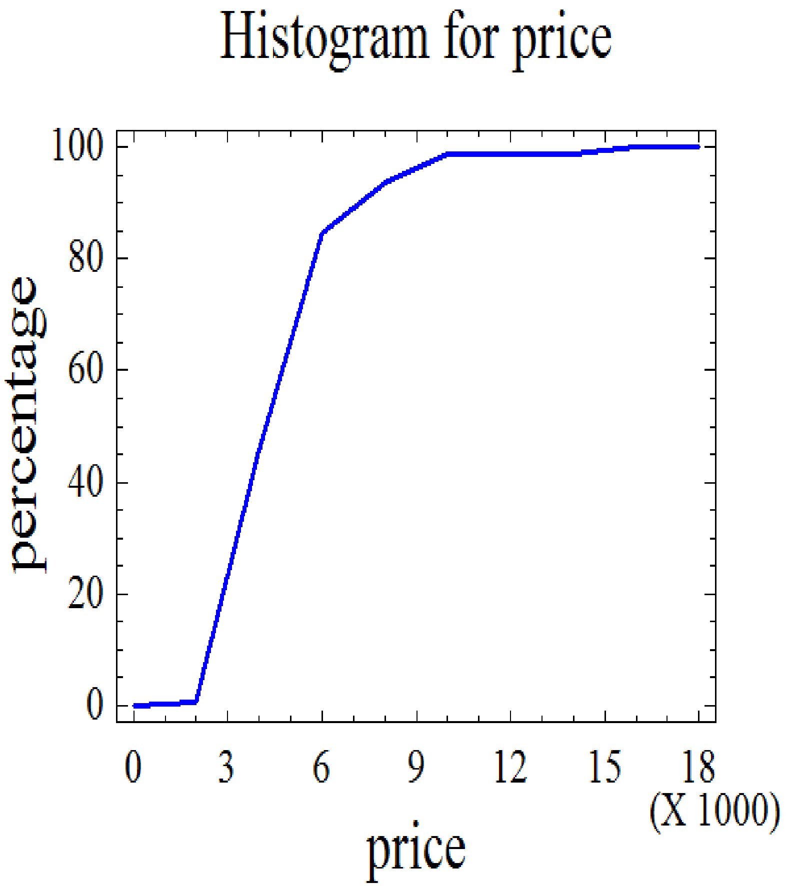
3.2 Frequency Histogram and Frequency Polygon

The Frequency Polygon is the graph representation of a frequency table whose data is grouped into intervals

Example: price of 155 cars (file cardata.sf)

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		0,0		0	0,0000	0	0,0000
1	0,0	2000,0	1000,0	1	0,0065	1	0,0065
2	2000,0	4000,0	3000,0	70	0,4516	71	0,4581
3	4000,0	6000,0	5000,0	60	0,3871	131	0,8452
4	6000,0	8000,0	7000,0	14	0,0903	145	0,9355
5	8000,0	10000,0	9000,0	8	0,0516	153	0,9871
6	10000,0	12000,0	11000,0	0	0,0000	153	0,9871
7	12000,0	14000,0	13000,0	0	0,0000	153	0,9871
8	14000,0	16000,0	15000,0	2	0,0129	155	1,0000
9	16000,0	18000,0	17000,0	0	0,0000	155	1,0000
above	18000,0			0	0,0000	155	1,0000

It is possible to draw the cumulative frequencies as well.

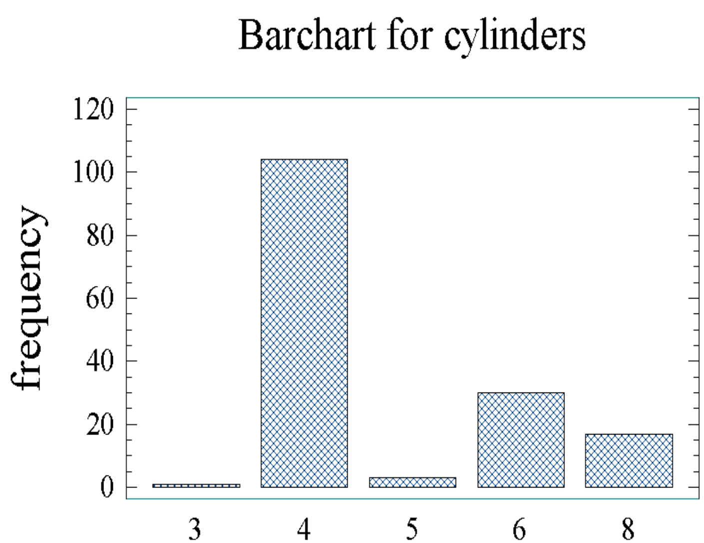


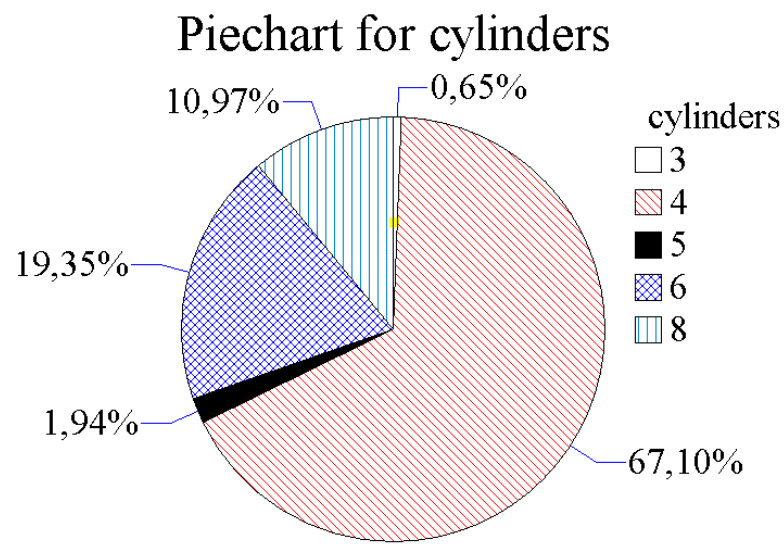
3.3 Pie Chart

The Pie Chart is a circle divided into proportional parts according to the relative frequencies

Example: number of cylinders of 155 cars (file cardata.sf)

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	3	1	0,0065	1	0,0065
2	4	104	0,6710	105	0,6774
3	5	3	0,0194	108	0,6968
4	6	30	0,1935	138	0,8903
5	8	17	0,1097	155	1,0000



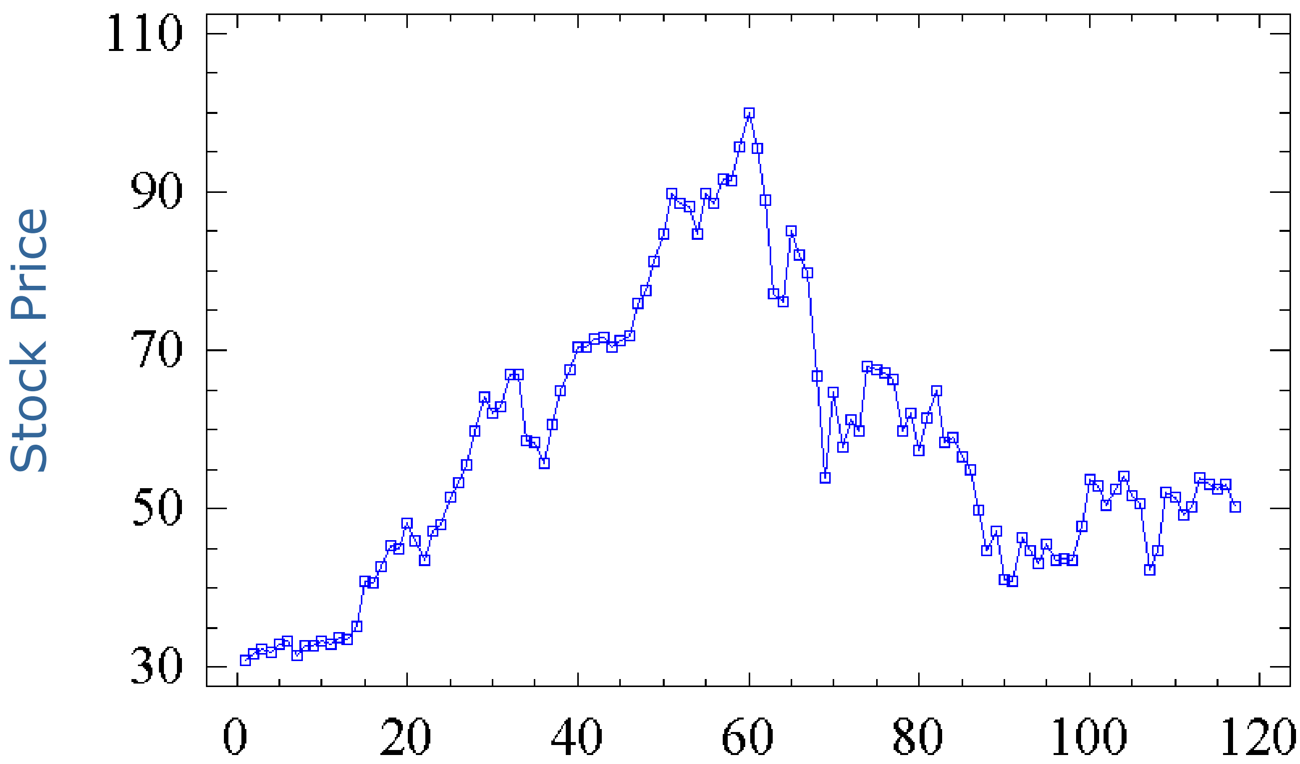


3.4 Time Series

Consider the plane (t,x). The axis T represents the time.

The Time Series represents the temporal evolution of the variable X(t).

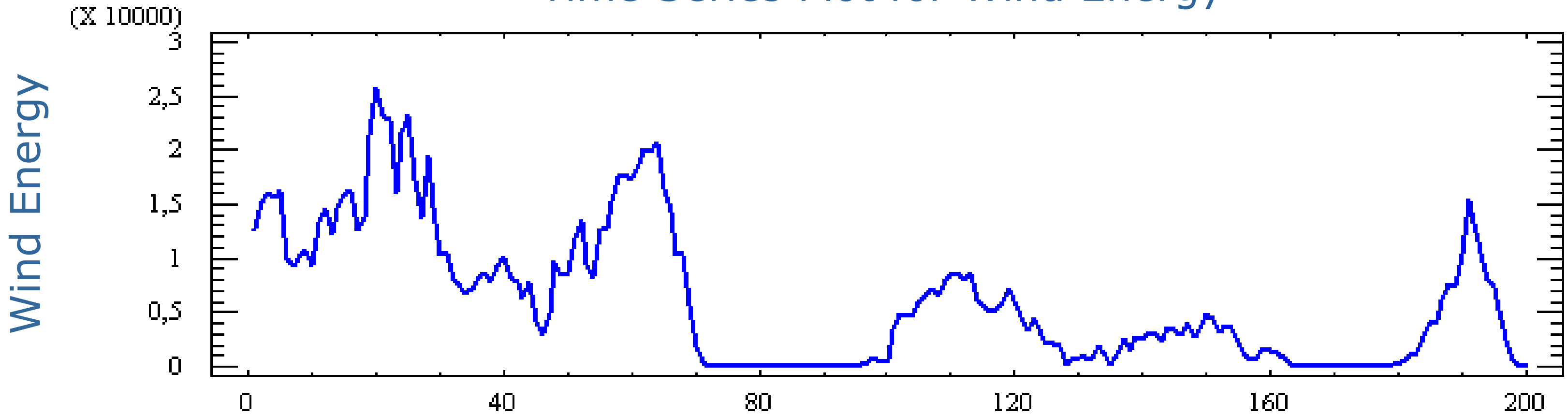
Time Series Plot for a Stock Price



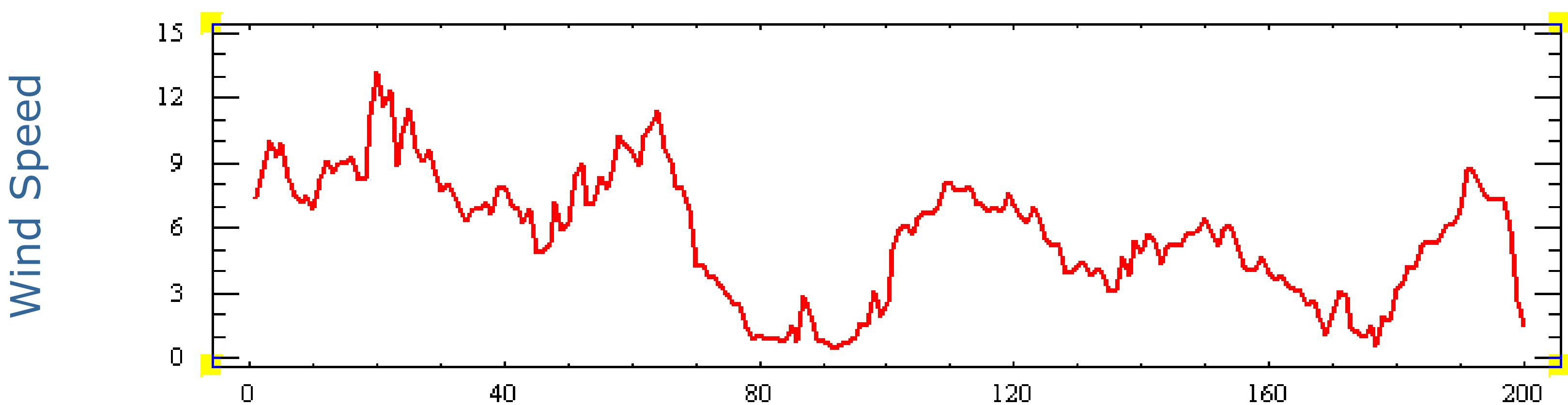
3.4 Time Series

In this example we plot the wind energy and its speed as functions of time.

Time Series Plot for Wind Energy



Time Series Plot for Wind Speed



Chapter I: Univariate Descriptive Statistics

1. Introduction. The purpose of Statistics.
2. Description of data by tables
3. Description of data by graphs
4. Characteristics measures of a variable

4 Characteristics measures of a dataset

Objective: We want to summarize the most important characteristics of data by using only few numbers.

Each feature → One number

4.1 Measures of position

Where is located the centre of the data?

There are a number of alternatives measures.

The most important are:

- The arithmetic mean or average
- The median
- The modes

4.1 Measures of position

- Arithmetic mean or sample mean

Given a set of observations $x_1, x_2, \ldots, x_n$

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

If there are J different values that appear repeated n_j times with j=1,2,...,J the sample mean can be written as

X_1, is repeated n_1 times

X_2, is repeated n_2 times

...

x_J, is repeated n_J times

$$\bar{x} = \sum_{j=1}^{J} x_j f_r(x_j).$$

where $fr(x_j)= n_j/N$ is the relative frequency of x_j and with $N=n_1+\ldots+n_J$ the sample size.

- Arithmetic mean or sample mean

Example: x={1,2,3,3,5,5,5,6,6}

$$\bar{x} = \frac{1+2+3+3+5+5+5+6+6}{9} = 4$$

or:

$$\bar{x} = \sum_{j=1}^{J} x_j f_r(x_j).$$

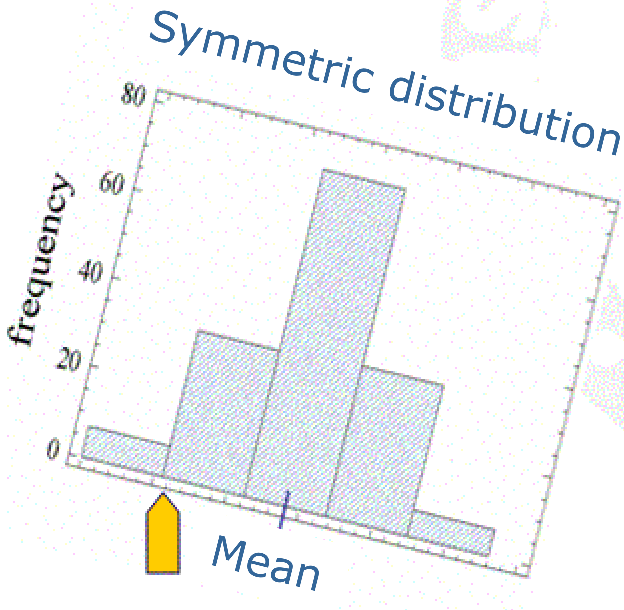
$$\bar{x} = 1\times\frac{1}{9} + 2\times\frac{1}{9} + 3\times\frac{2}{9} + 5\times\frac{3}{9} + 6\times\frac{1}{9} = 4$$

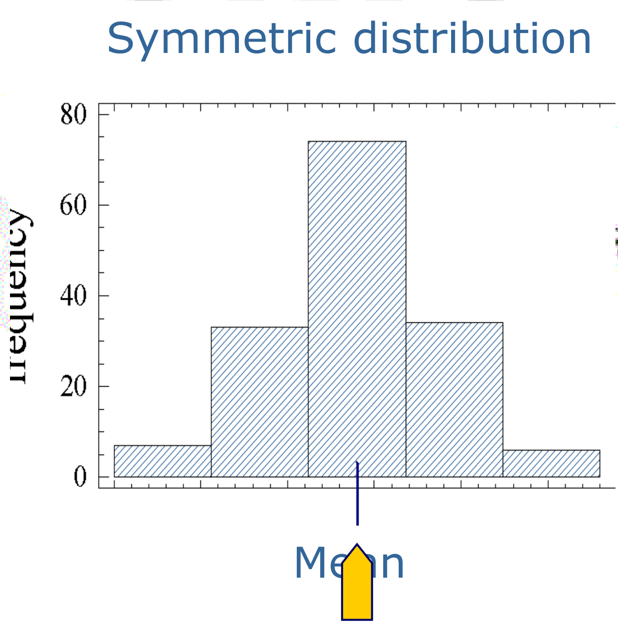
4.1 Measures of position

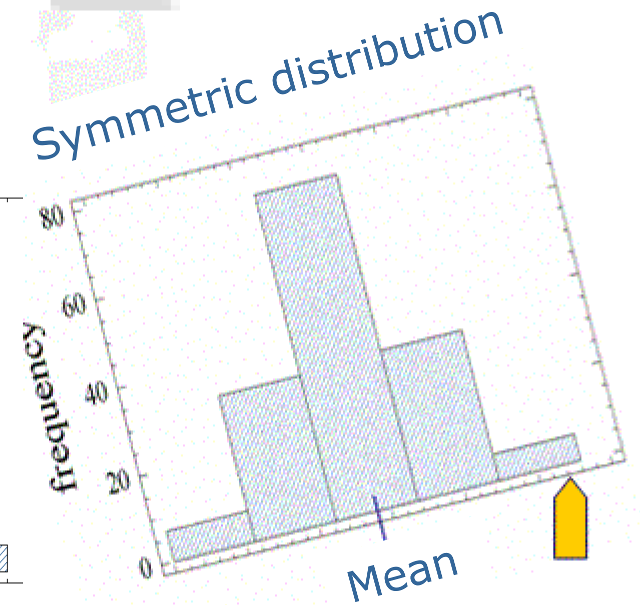
- Arithmetic mean or sample mean

The sample mean can be looked at as if it were the gravity centre of the dataset.

For instance, in a histogram, it is the balance point

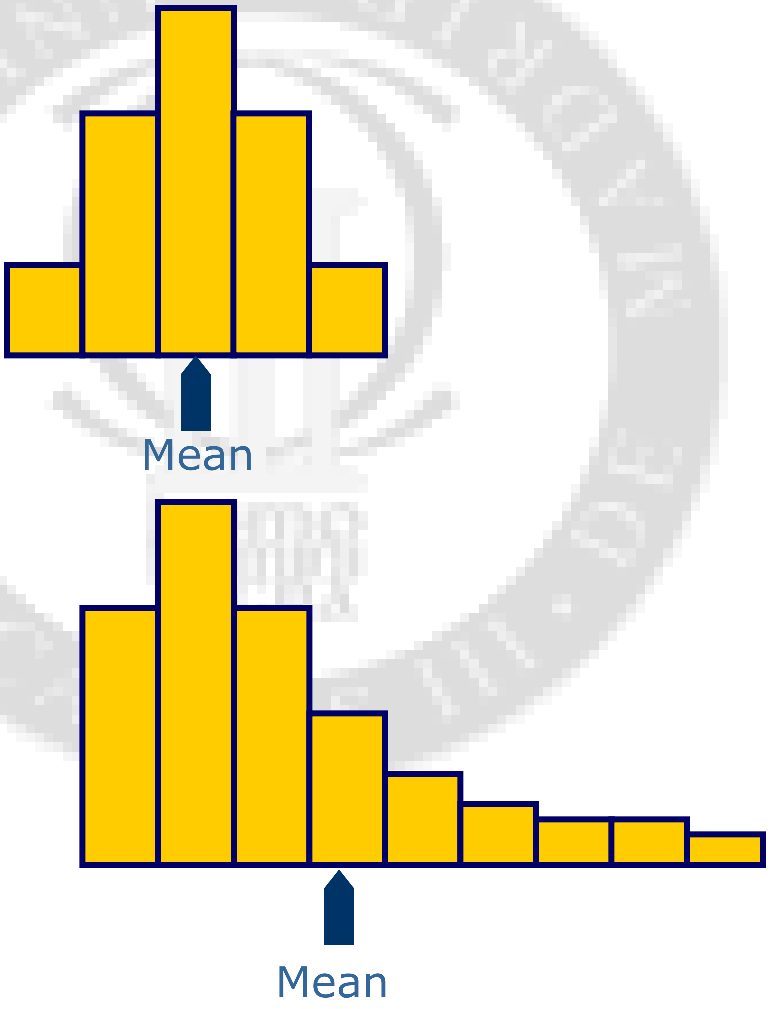






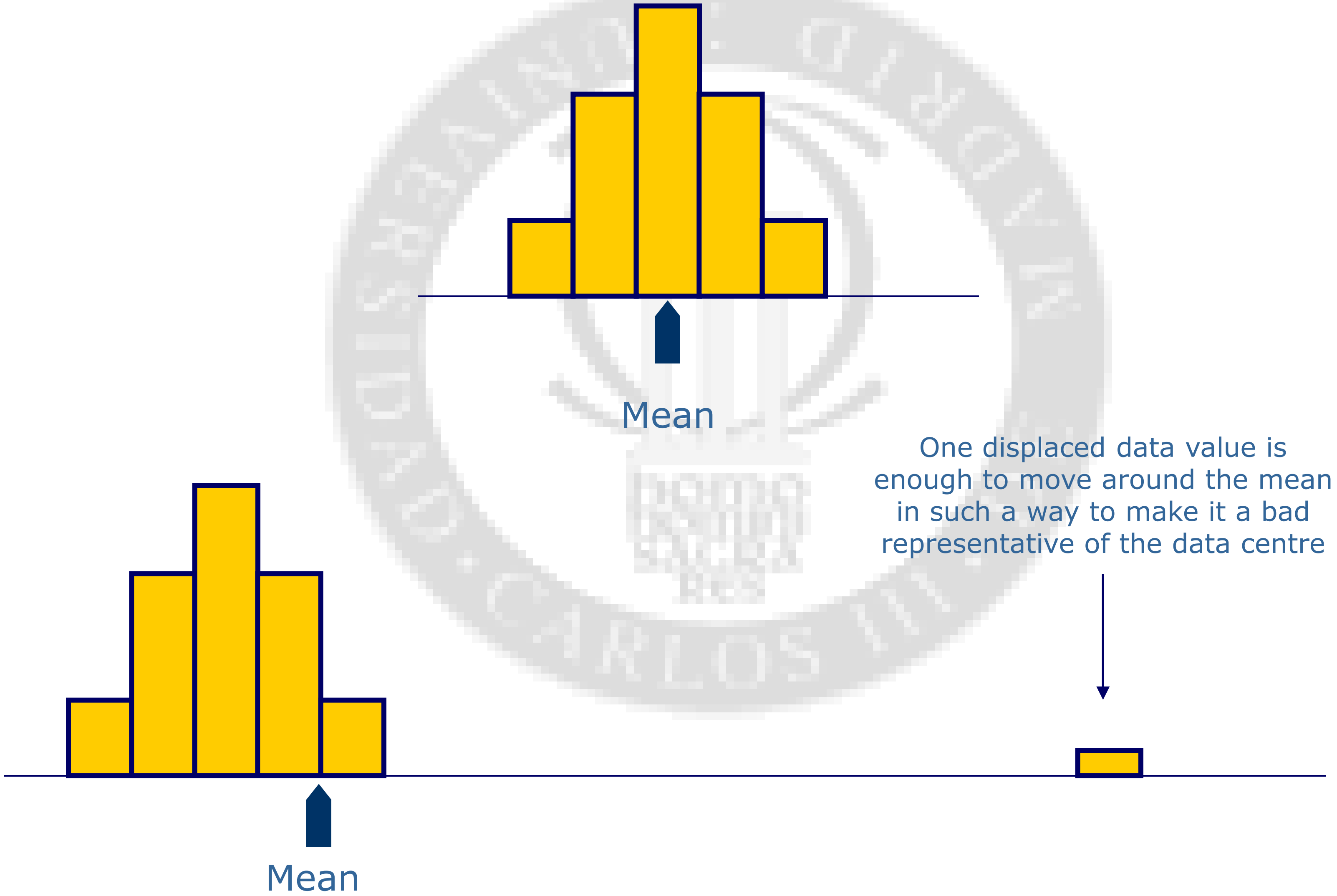
- Arithmetic mean or sample mean

The more asymmetric the distribution is the more the mean shifts to the tail



- Arithmetic mean or sample mean

It is very sensitive to outliers



- Median

The median is the value which leaves the 50% of the data to its left and to its right

It is not very sensitive to asymmetries

It is not sensitive to outliers

1 2 5 8 11 13 24 28 31

9 data

Median=11

With an odd number of data the median coincides with the central data value

1 2 3 5 8 11 13 24 28 31

10 data

Median=(8+11)/2=9,5

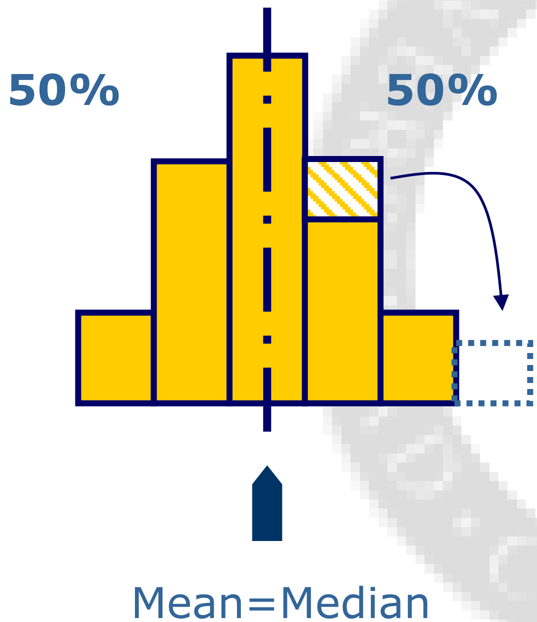
With a even number of data the median is given by the arithmetical mean of the two most central data values

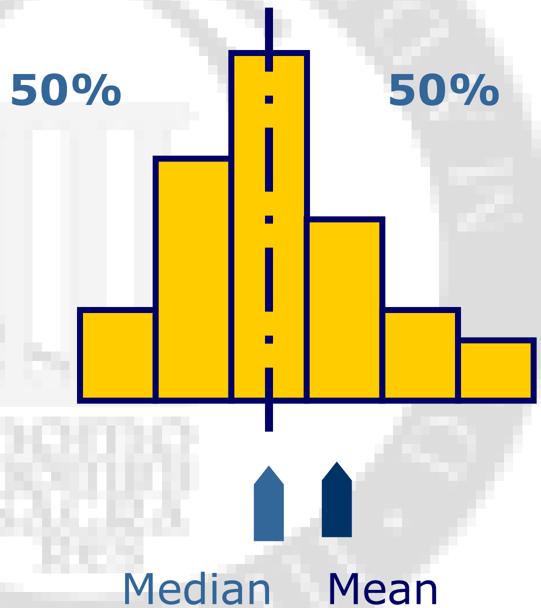
- Median

The median is the value which leaves the 50% of the data to its left and to its right

It is not very sensitive to asymmetries

It is not sensitive to outliers





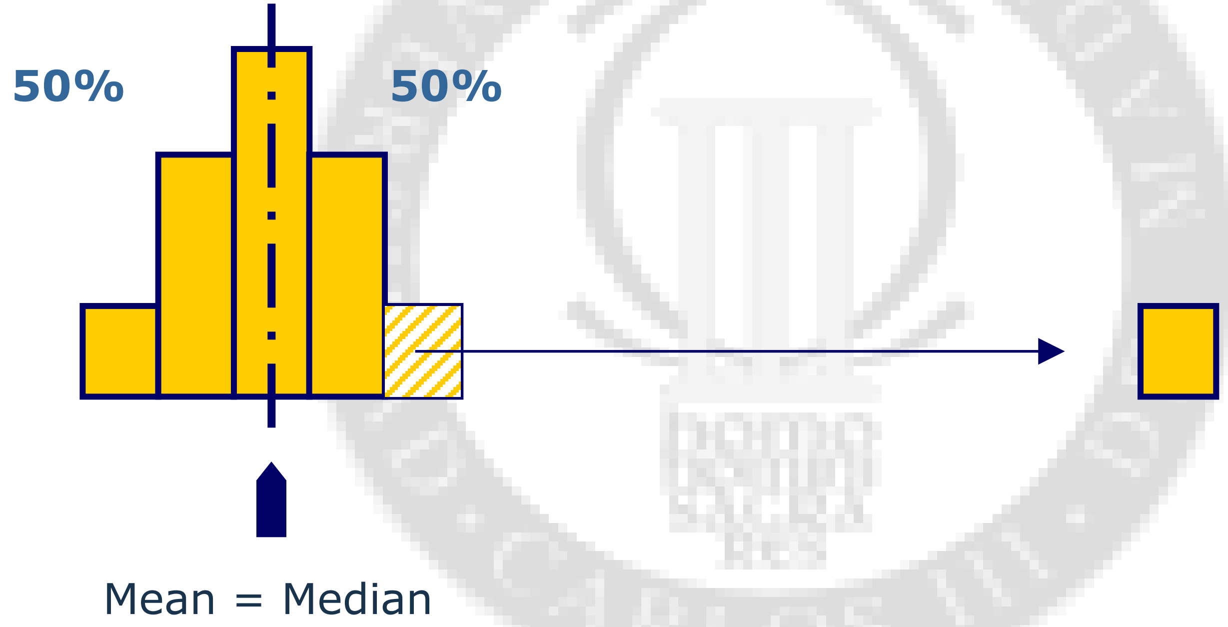
The median does not change while the mean does change

- Median

The median is the value which leaves the 50% of the data to its left and to its right

It is not very sensitive to asymmetries

It is not sensitive to outliers

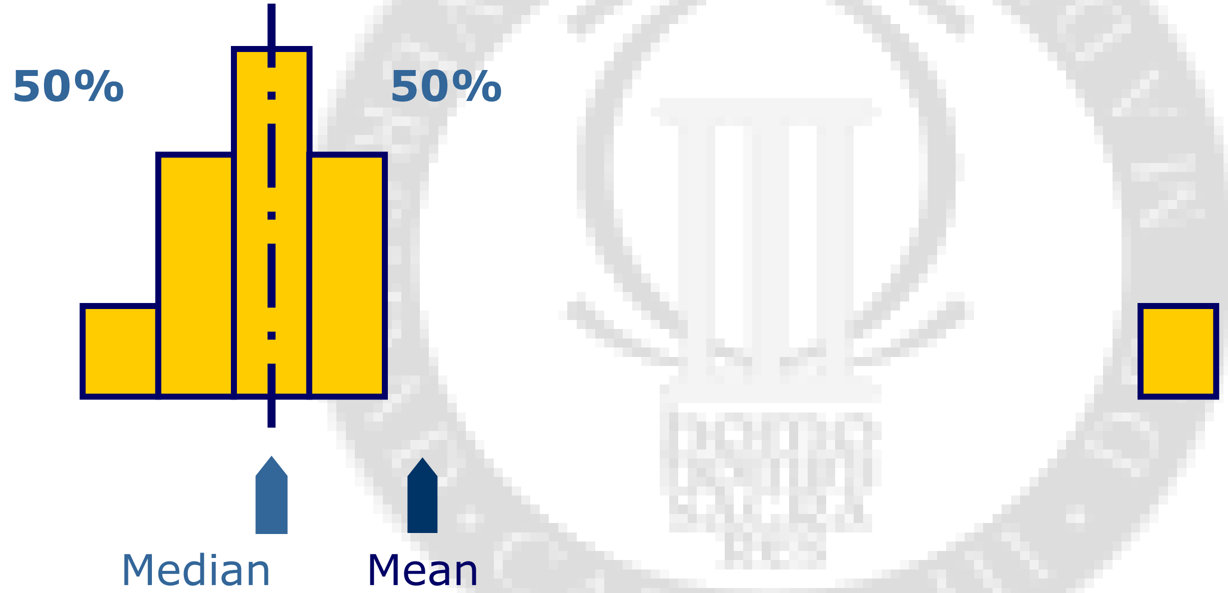


- Median

The median is the value which leaves the 50% of the data to its left and to its right

It is not very sensitive to asymmetries

It is not sensitive to outliers



The outliers do not modify the position of the median

In presence of outliers and strong asymmetries, the median is a more useful measure of position than the mean

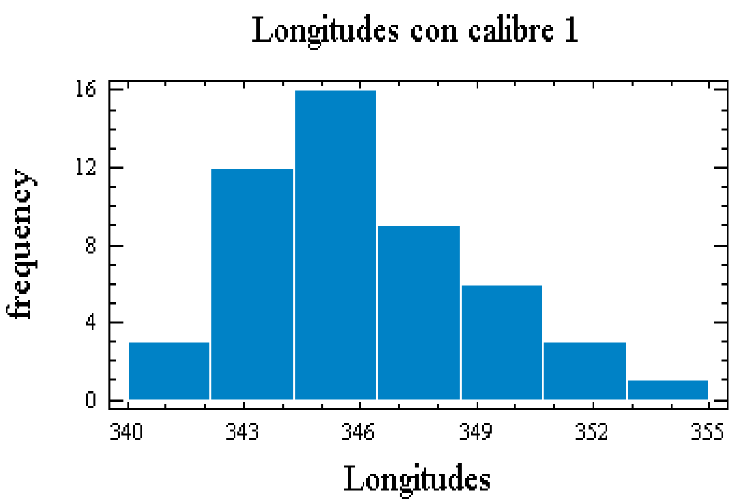
- Modes

Modes correspond to the values of data that show up with highest frequency

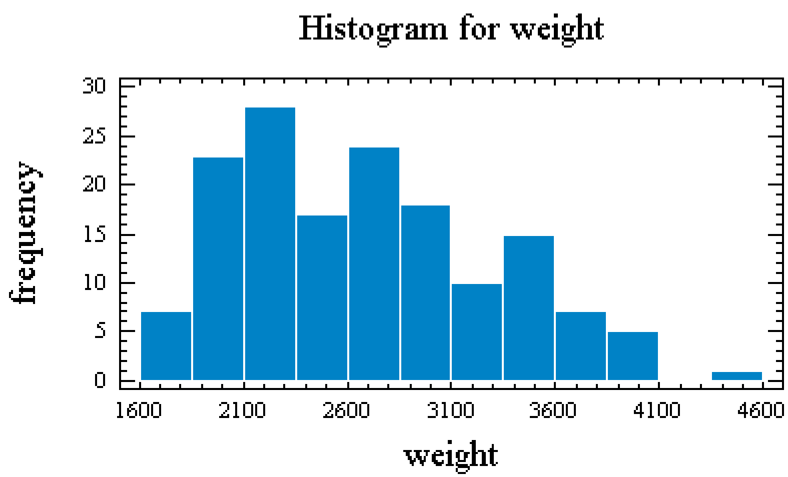
1 2 2 2 2 5 5 5 8 8 11 13

Mode=2

In case of grouped data, the mode is the class with highest frequency. There could be more than one mode suggesting the possible existence of heterogeneous groups



Unimodal distribution



Trimodal distribution

5.1 Measures of position → mean, median, mode

5.2 Measures of dispersion

- Variance (Standard deviation)
- Meda
- Range
- Quartiles
- Box-plot

- **Variance**

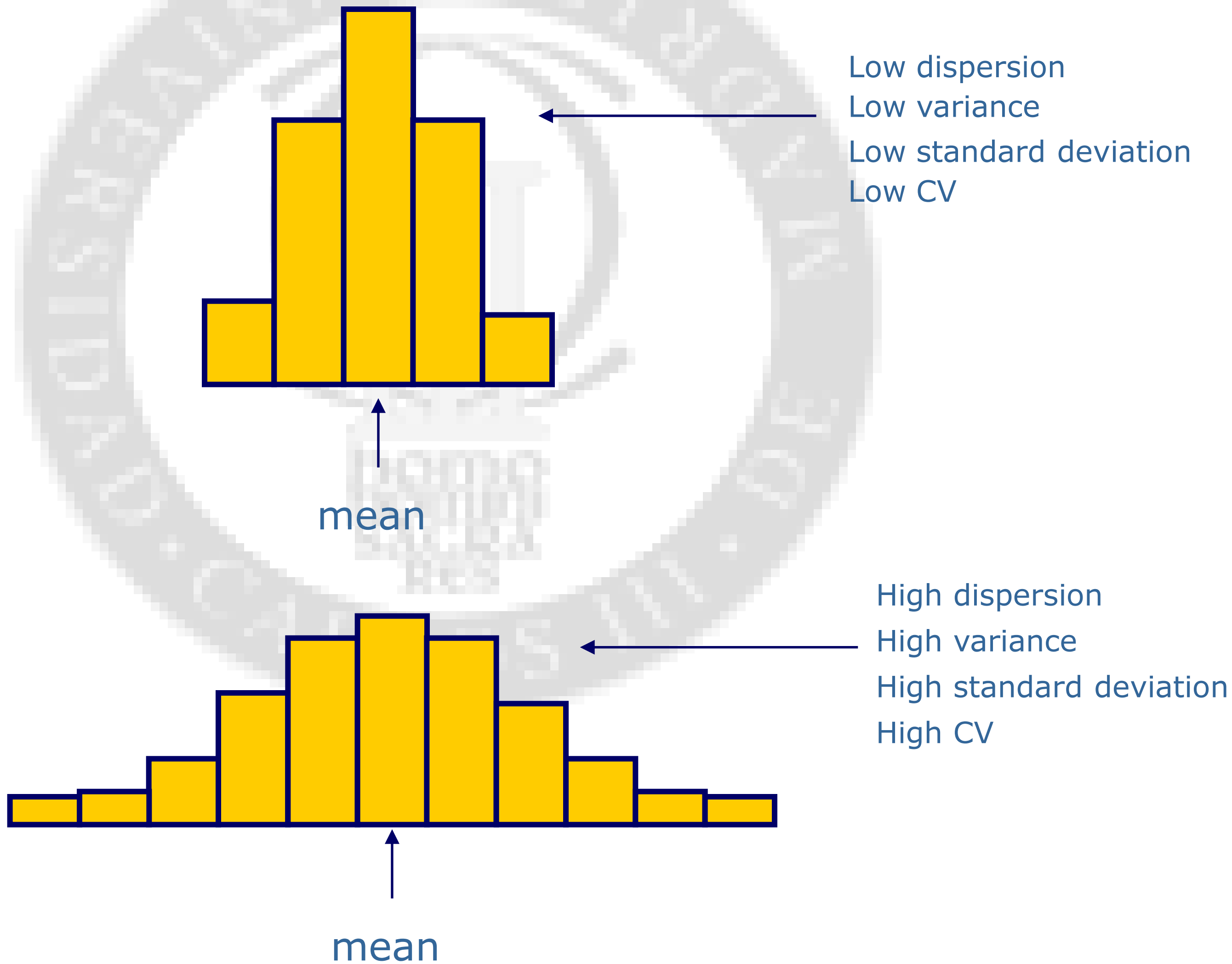
Measures the average (squared) deviation from the mean of the observations

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \qquad \underbrace{s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}.}_{\text{Standard deviation}} \qquad \underbrace{CV = \frac{s_x}{|\bar{x}|},}_{\text{Coefficient of variation}}$$

- **Variance**

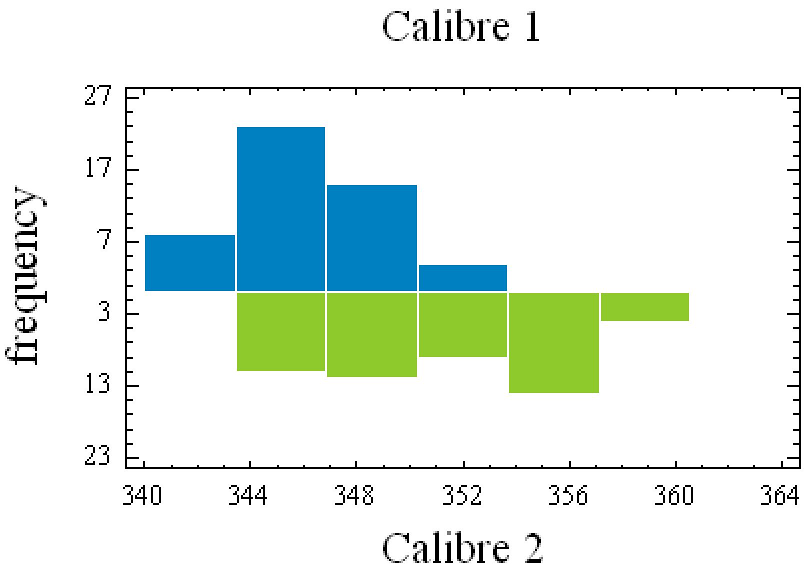
  Average (squared) deviation from the mean of the observations

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



Example:

The lengths of 100 nails of same type, measured by two guys with different calibres, 50 nails each.



What calibre is better?

Variance of Calibre 1: **7.25 mm²**

Variance of Calibre 2: **21.47 mm²**

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

- **MEDA**

Median of absolute deviations from the median

Data → $x_1, x_2, x_3, \ldots, x_n$

Median → $\mathbf{x_m}$

Absolute deviations from the median → $|x_1 - \mathbf{x_m}|, |x_2 - \mathbf{x_m}|, |x_3 - \mathbf{x_m}|, \ldots, |x_n - \mathbf{x_m}|$

Median of the previous values → MEDA

Less sensitive to outliers and asymmetries than the variance

Why?

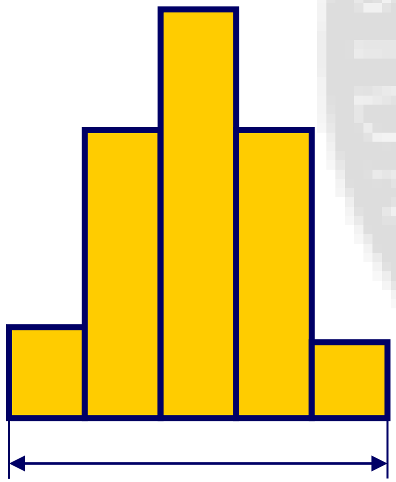
- **Range**

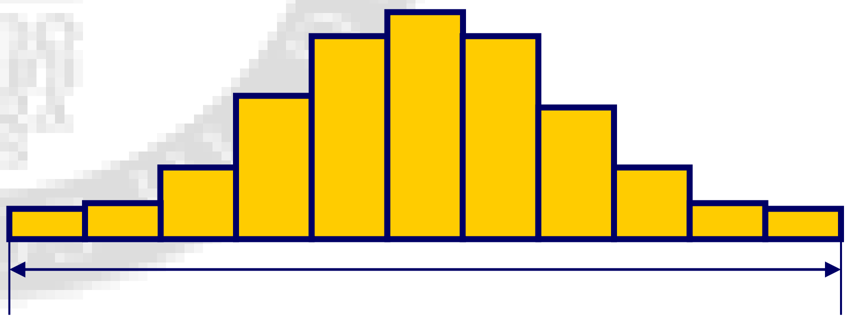
Maximum value minus minimum value

X: 1 2 5 8 11 13 24 28 31

Range: 31-1=30

The bigger the range the more is the dispersion





- **Quartiles Q_1, Q_2, Q_3**

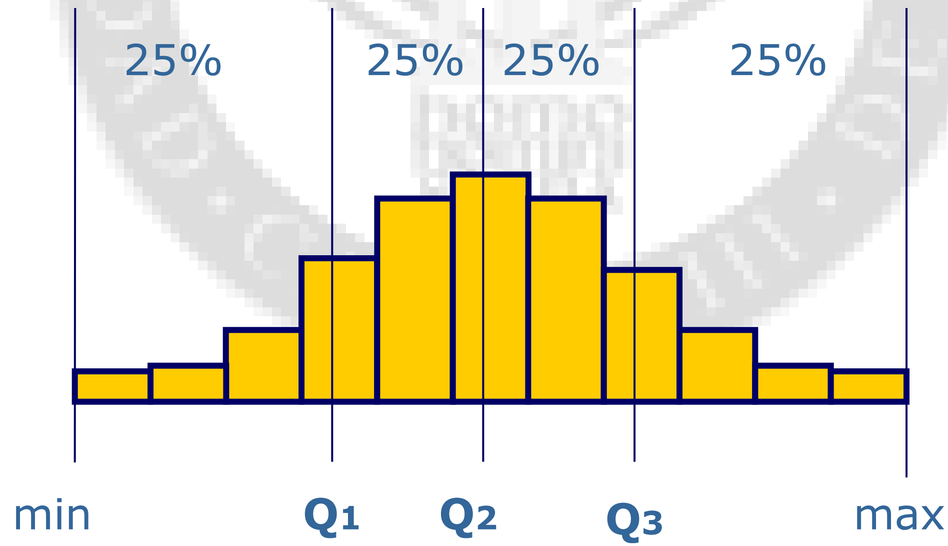
Divide the sample into 4 groups with similar frequencies, each one with 25% of the data (approximately)

Between the minimum and **Q_1** → 25% of the data
Between **Q_1** and **Q_2** → 25% of the data
50%

Between **Q_2** and **Q_3** → 25% of the data
Between **Q_3** and the maximum → 25% of the data
50%

Q_2 = median

Q_3-Q_1=Interquartile range (I.R.)



There are different methods to calculate Q_1 and Q_3 that give different results for small datasets.

- **Quartiles Q_1, Q_2, Q_3**

$x:\{1,1,3,3,5,9,11,14,15\}$

A simple method to calculate quartiles

1º: Obtain the median Q_2 → 5

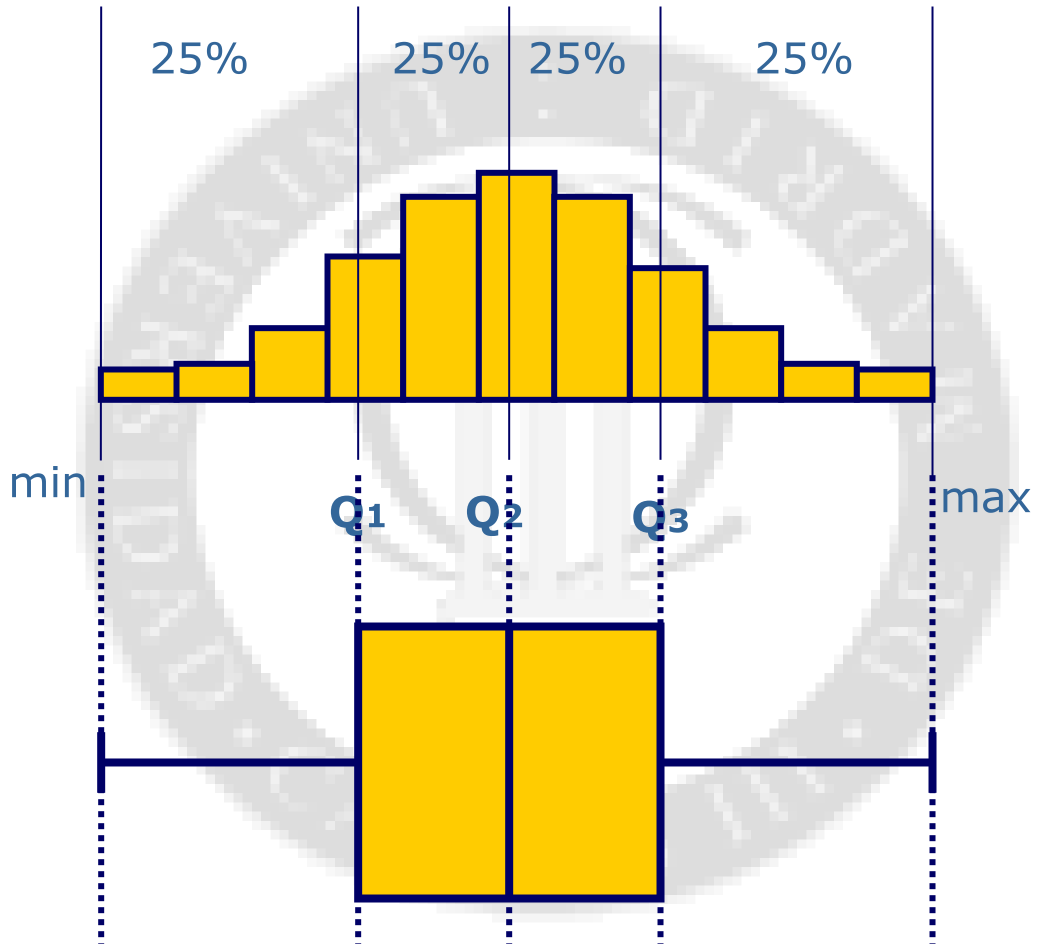
2º: Exclude this value and take two groups of data, one for each side of the median →
left.: $\{1,1,3,3,\}$
right.: $\{9,11,14,15\}$

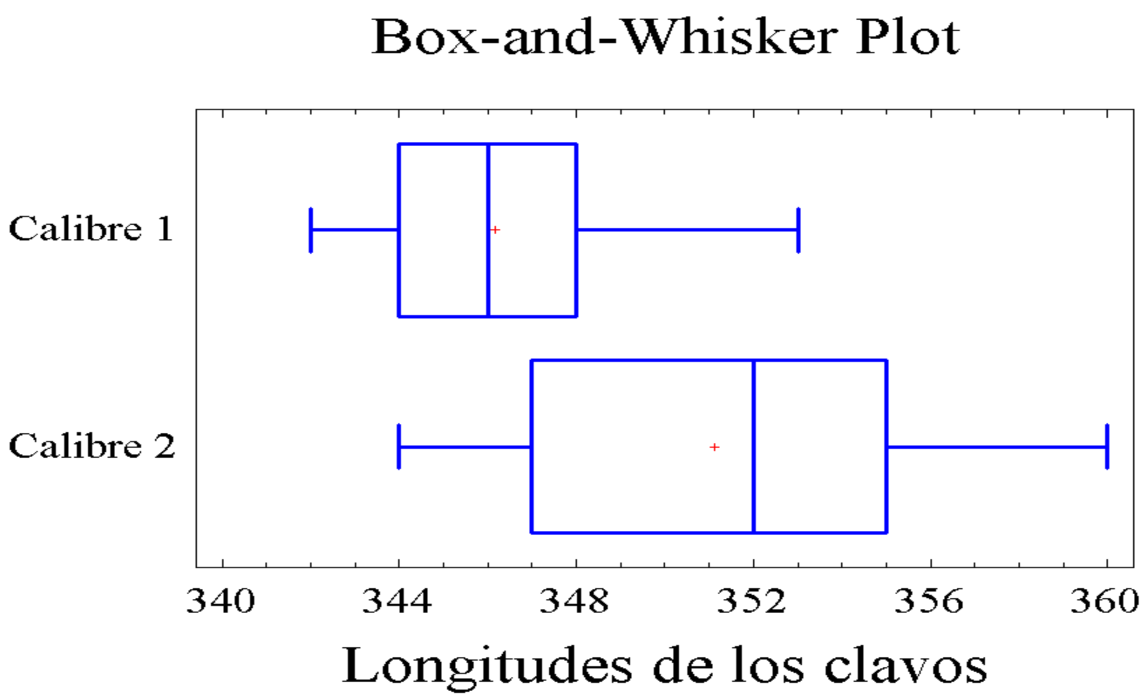
3º: Q_1 is the median of the left group → $Q_1=(1+3)/2=2$

4º: Q_3 is the median of the right group → $Q_3=(11+14)/2=12.5$

- **Boxplot (Box–and–Whisker plot)**

The boxplot is a graphical representation of the quartiles

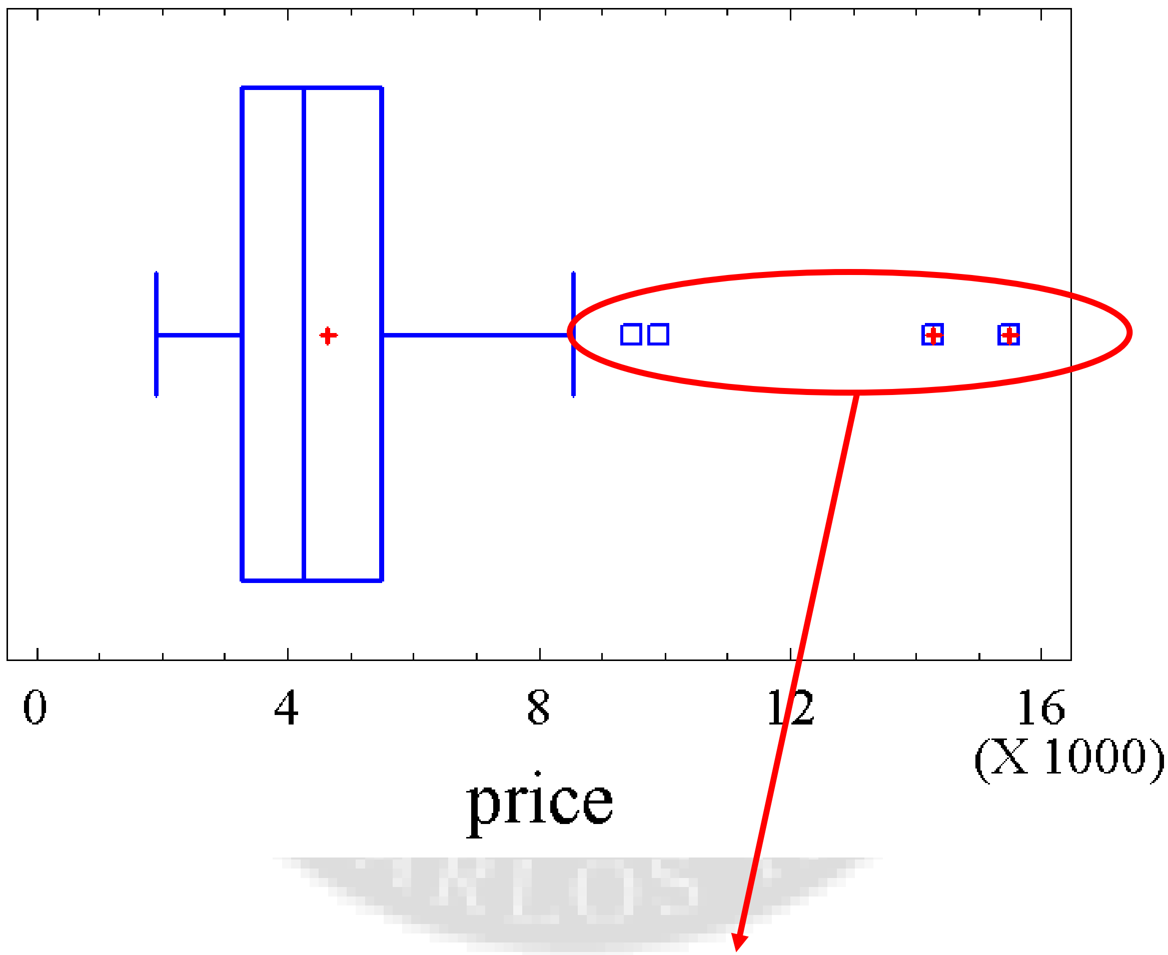




The boxplots are very useful to:

- **compare groups of data**
- observe asymmetries
- detect outliers **

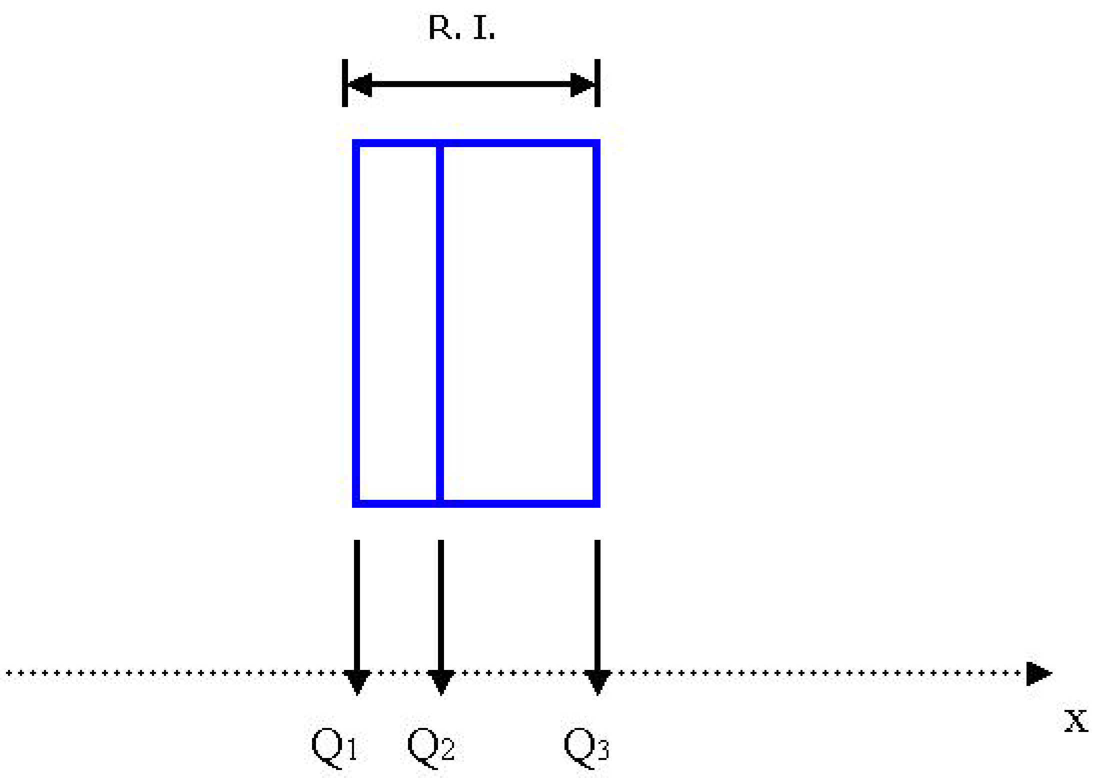
Box-and-Whisker Plot



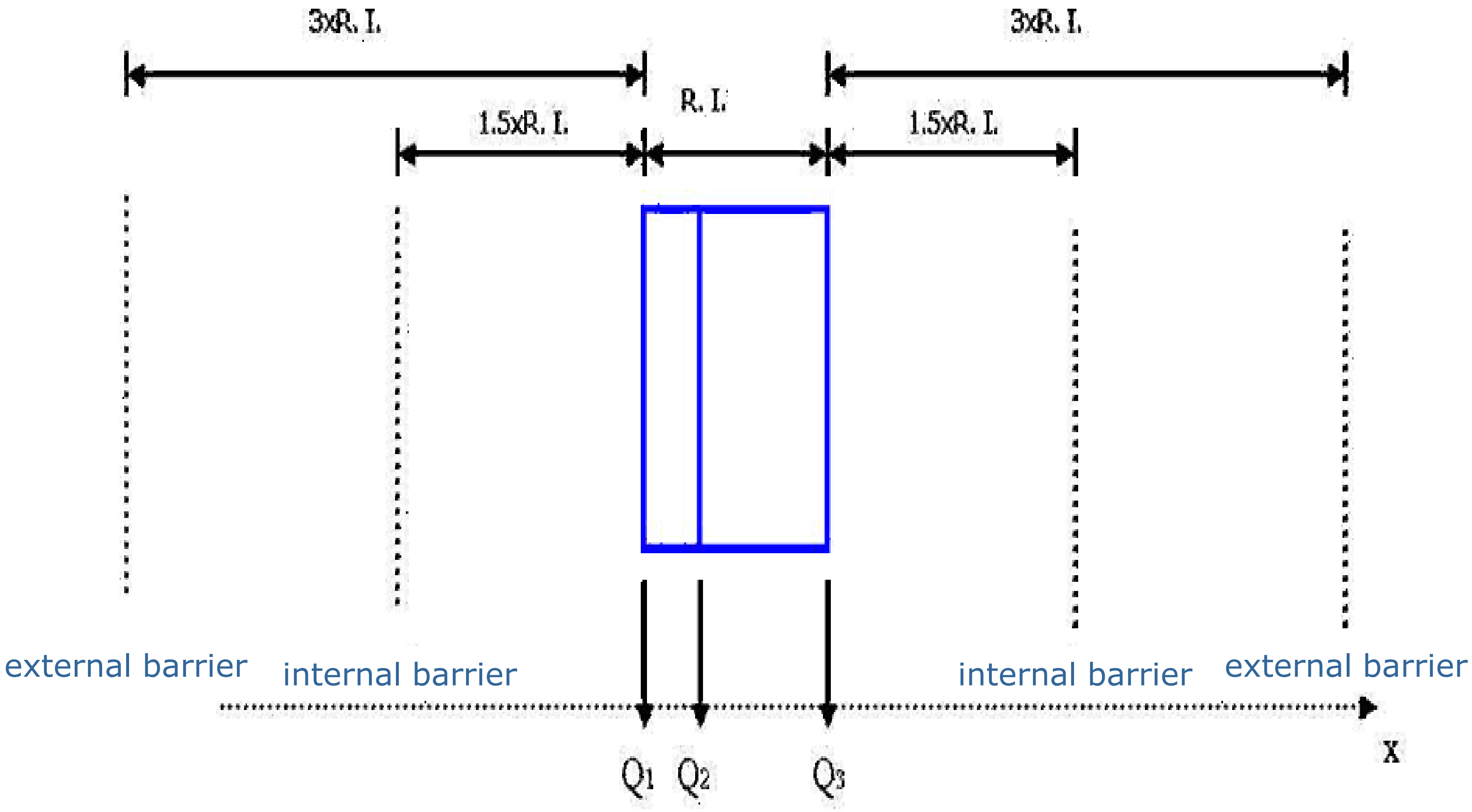
Extreme data (or 'outliers')

To make a Box-plot with marks of outliers

First step

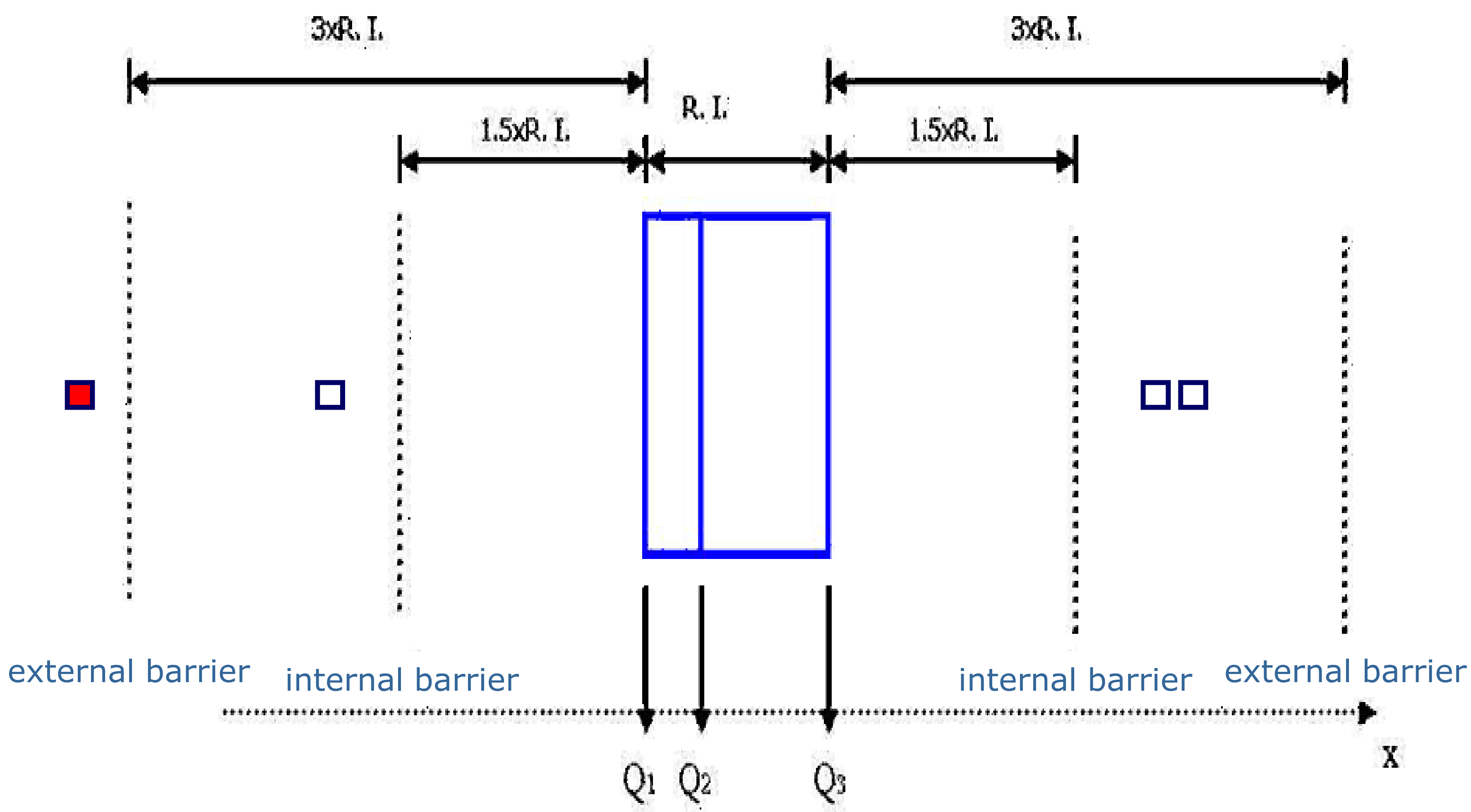


Second step



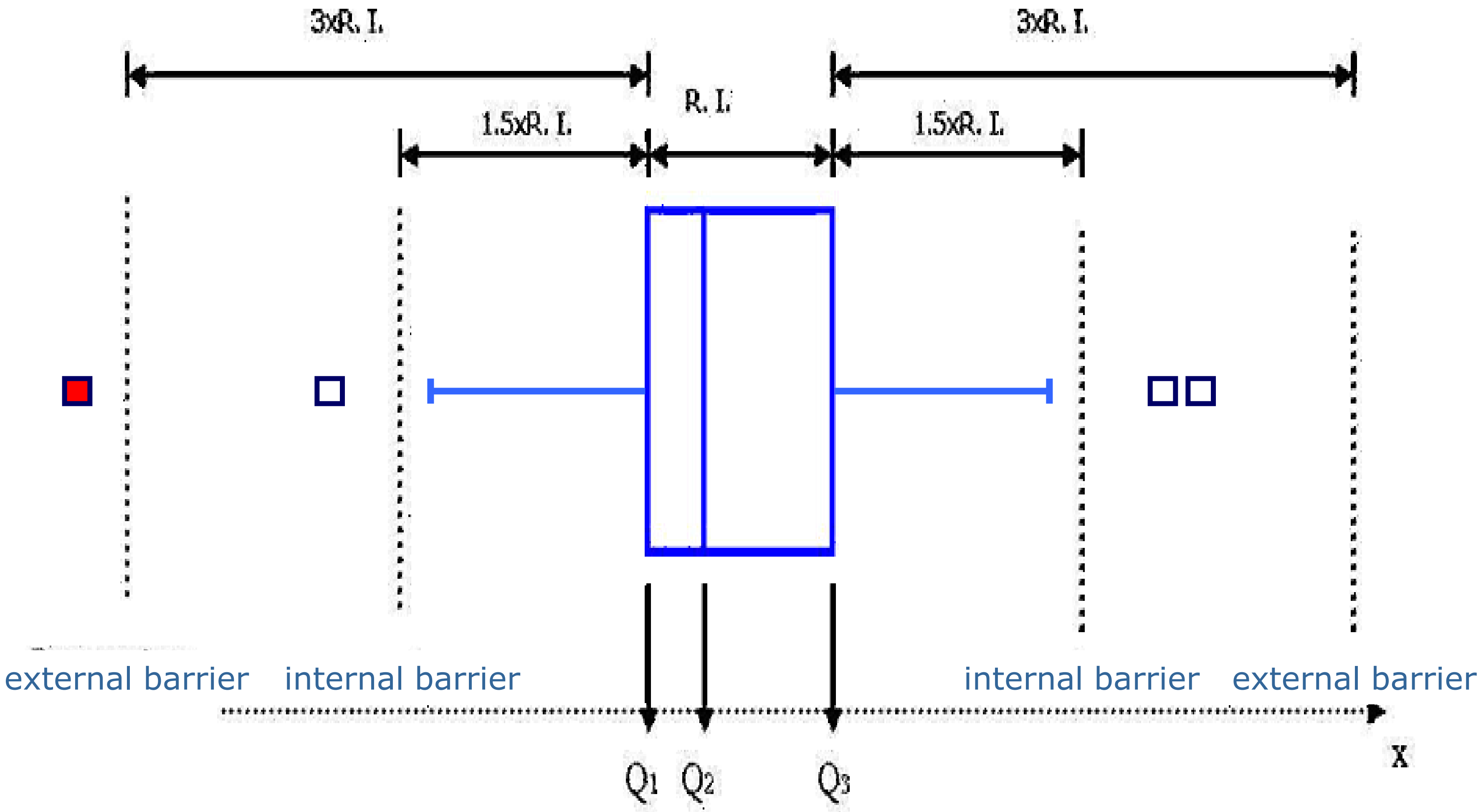
3xR. I.
3xR. I.
1.5xR. I.
R. I.
1.5xR. I.
external barrier
internal barrier
internal barrier
external barrier
X
Q1
Q2
Q3

Third step



The points which fall within of these zones are marker

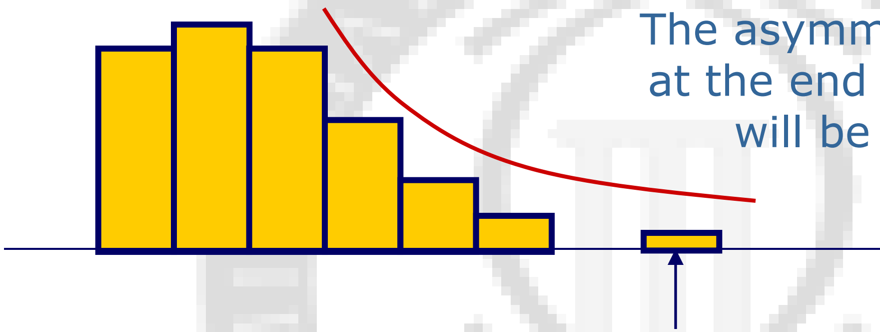
Third step



The lateral lines are spread only <u>to the last point</u> within of the internal barriers

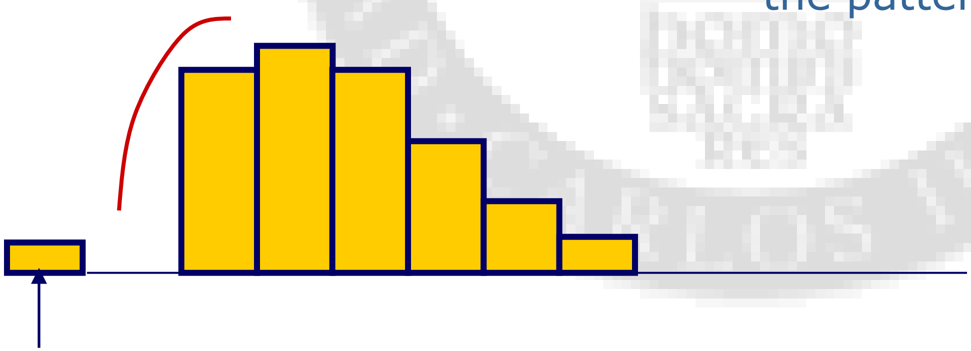
careful!! when there are asymmetries an extreme data is not necessarily an outlier

Outliers: data who is apart from the general pattern of data

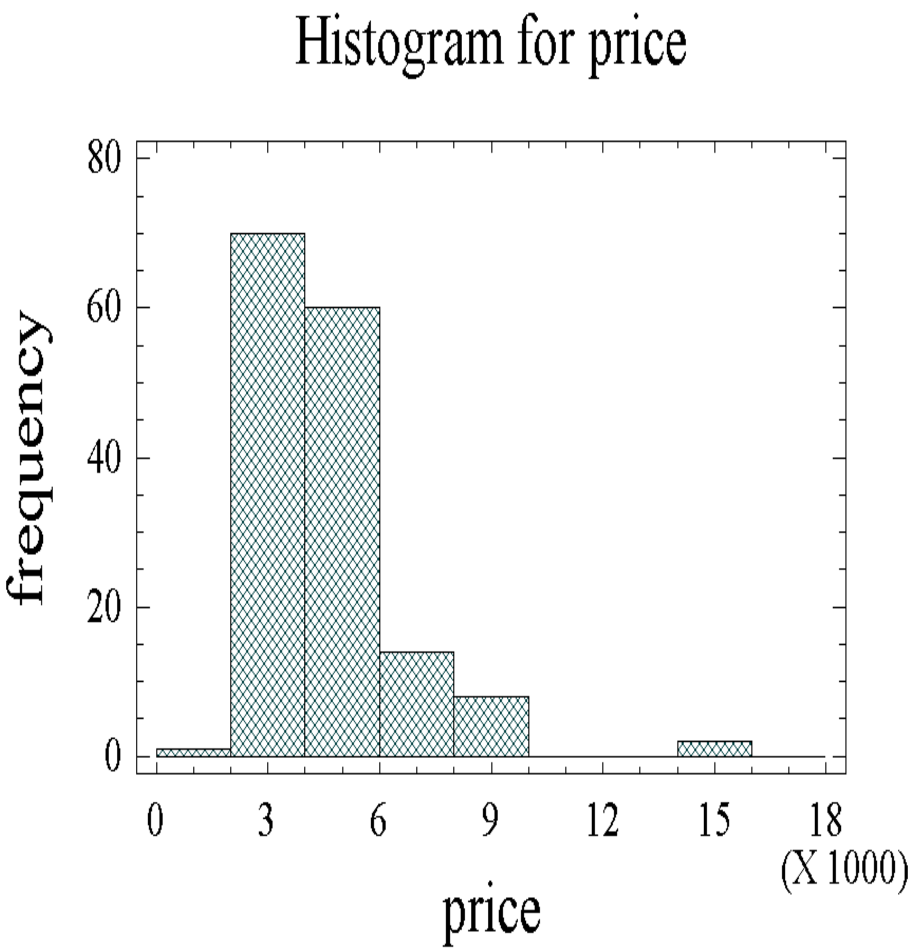


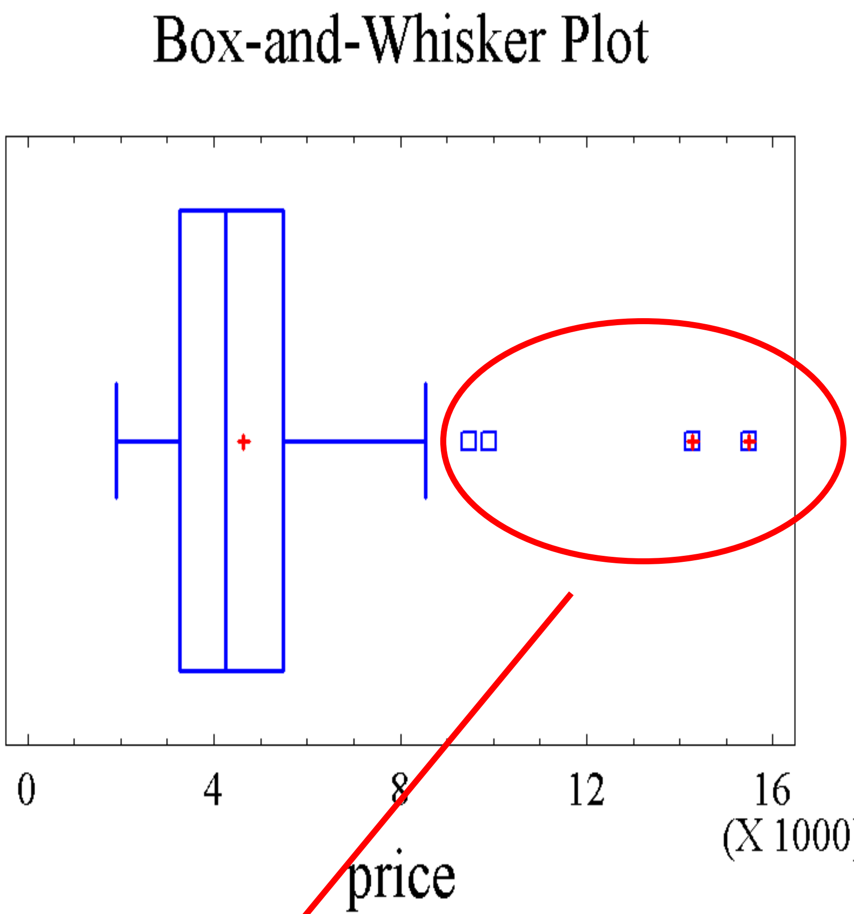
The asymmetry shows that at the end of the tail there will be fewer data.

It is not outlier. It agrees with the pattern of the data



Yes, it is an outlier





These values are compatible
with the positive skewness

5.1 Measures of centre → mean, median, mode

5.2 Measure of spread → variance, standard deviation, coefficient of variation, meda, range, quartiles, box-plot

5.3 Other measures of shape

- Measures of skewness
- Measures of kurtosis (flat or steep)

- **Measures of asymmetry**

Coefficient of skewness →

$$CA = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n s_x^3}.$$

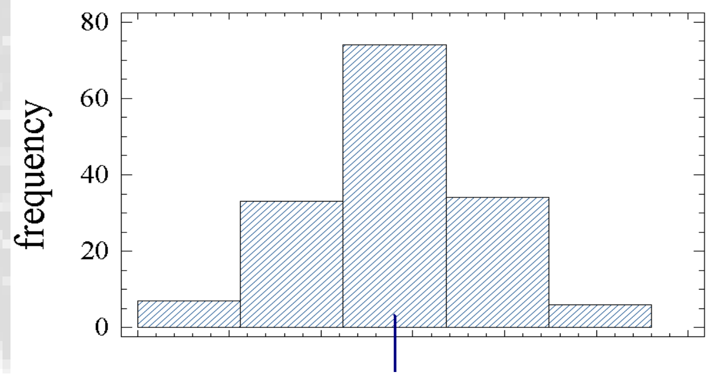
- CA = 0; if the distribution is perfectly symmetry
- CA > 0; if there is positive asymmetry, skewness to the right
- CA < 0: if there is negative asymmetry , skewness to the left

- **Measures of asymmetry**

Coefficient of skewness → $$CA = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n s_x^3}.$$

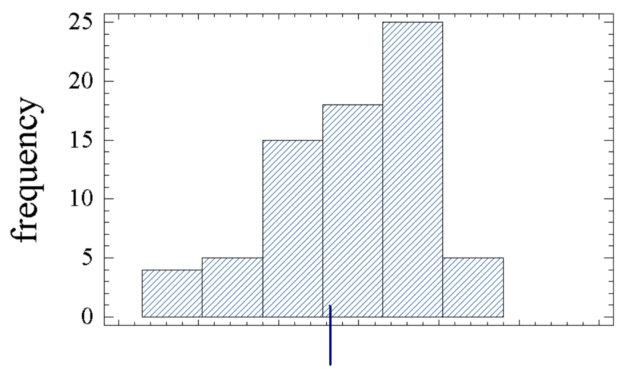
- CA = 0; if the distribution is perfectly symmetry
- CA > 0; if there is positive asymmetry, skewness to the right
- CA < 0: if there is negative asymmetry , skewness to the left

Symmetric distibution



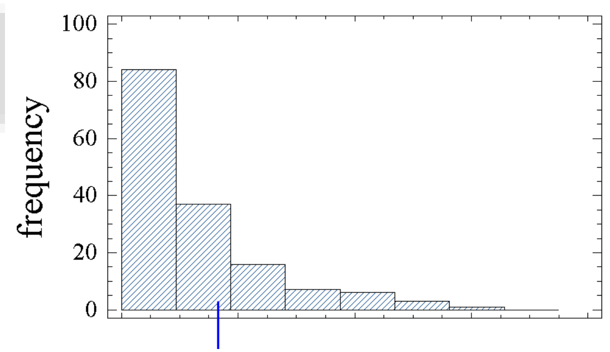
Mean

Distribution skewed to the left



Mean

Distribution skewed to the right

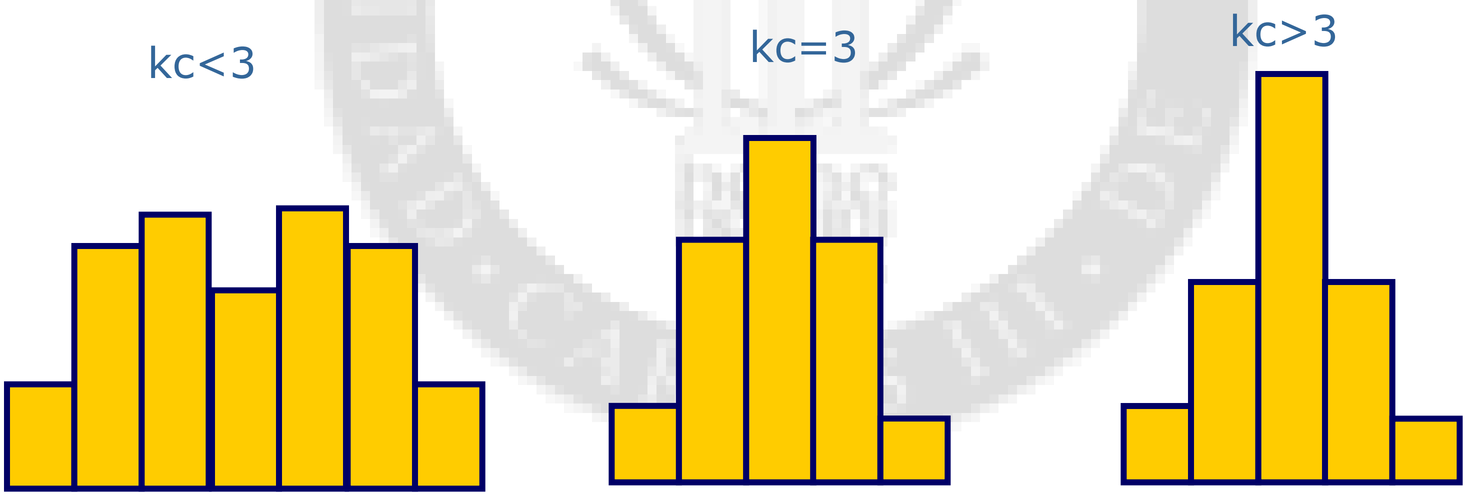


Mean

• Measures of kurtosis

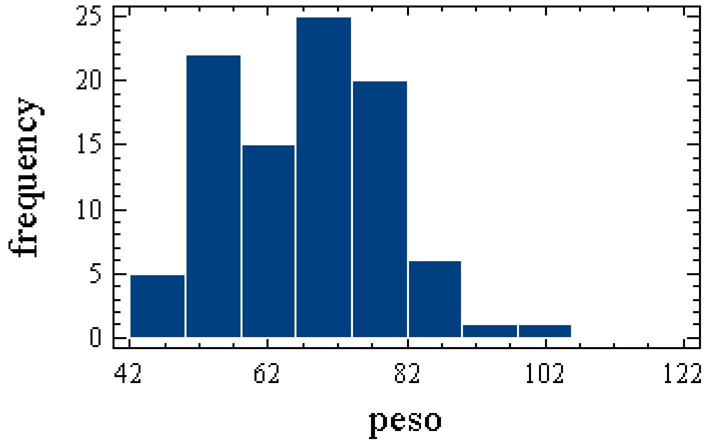
Coefficient of kurtosis → $$CAp = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n s_x^4}.$$

- Kc = 3; distribution with Gaussian bell shape
- Kc > 3; distribution is steeper than a Gaussian bell
- Kc < 3; distribution is flatter than a Gaussian bell



In many statistic software the kurstosis coefficient is defined as (Cap – 3)

Weight histogram

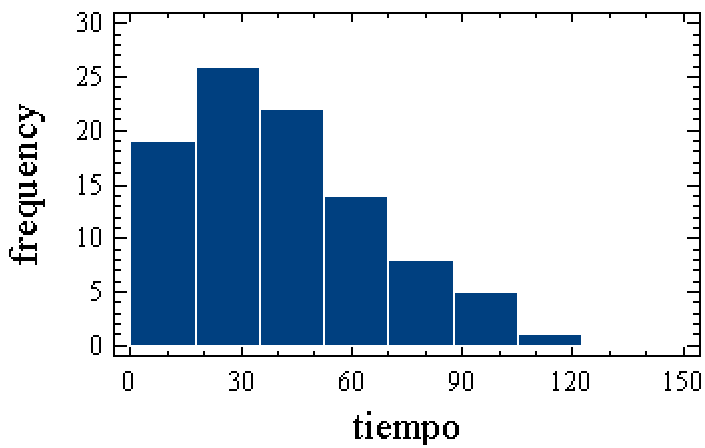



```
Summary Statistics for peso

Count = 95
Average = 67,7684
Median = 69,0
Skewness = 0,261155
Kurtosis = -0,502931
```

(Kurtosis-3)

Histogram for "Time to reach the University"




```
Summary Statistics for tiempo

Count = 95
Average = 41,4211
Median = 40,0
Skewness = 0,651076
Kurtosis = 0,0915265
```

Lower values of kurtosis can denote presence of 'multimodality'