Chapter 1

Descriptive statistics

Contents

Descriptive statistics is the set of scientific methods used to collect, describe and analyze observed data

1.1 Statistical vocabulary

❶ **Population**: is the set of individuals or objects of the same nature on which the study relates.

❷ **Individuals**: or statistical units are the elements of the population.

❸ **Sample**: is a subset of the population.

❹ **Statistical variable**: or character X is the subject under statistical study .

❺ **Statistical modality**: or category the different possible situations (levels) of a statistical variable.

There are two types of statistical variables

Quantitative variables

Are the variables that can be measured, they are characterized by numerical values. Variables whose modalities are numbers.

A quantitative statistical variable can be:

continuous: when it can take numbers from an interval of real numbers (measurement results).

Discrete: if it takes isolated values.

Temporal: These are particular quantitative variables that use units of measurement of time. There are two types, date type (date of birth: 04/26/1994) and time type (study hours: 6h).

Example 1.1.1.

variable	*possible modalities*	*type of variable*
height	*1.70m, 1.60m, 1.65m, 1.75m*	*continuous quantitative*
the number of students	*30, 50, 60, 80*	*discrete quantitative*

Qualitative variables

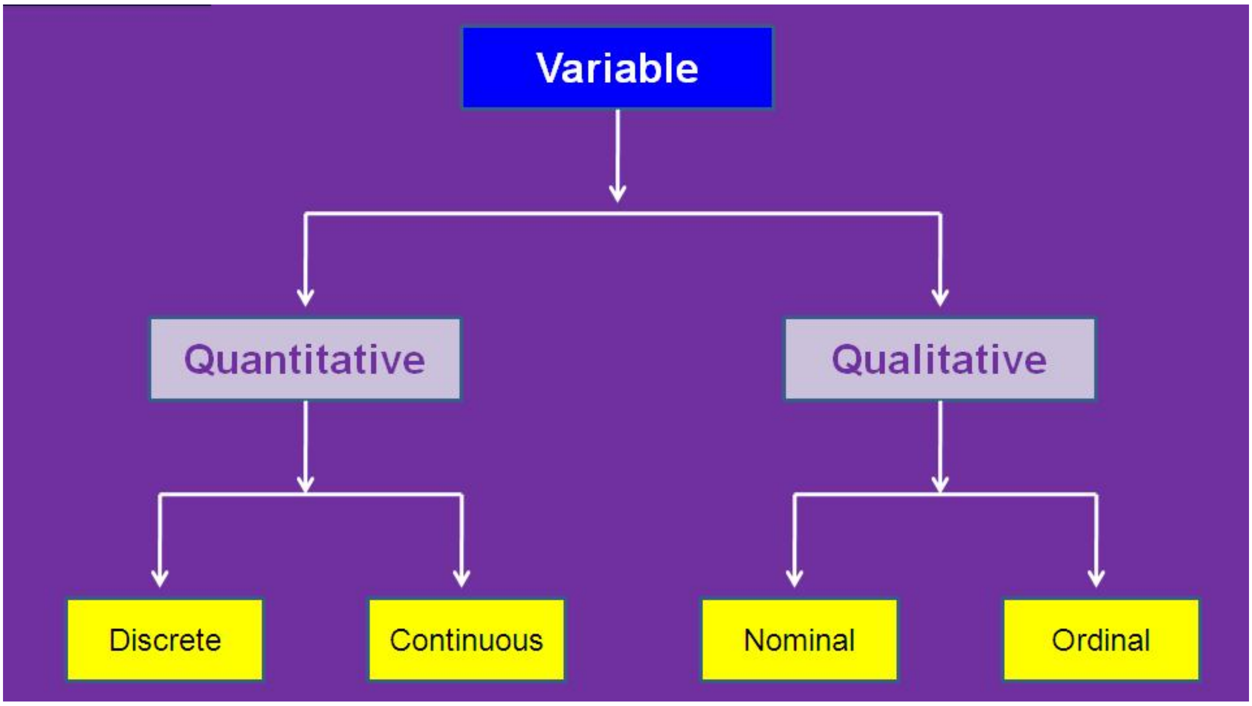
These are variables that are not measurable (do not have numerical values).

Variables whose modalities are words.

Qualitative statistical variables can be:

Ordinal: these are variables whose modalities are ordered according to their meaning.

Nominal: these are variables whose modalities cannot be ordered according to their meaning.



Example 1.1.2.

variable	*possible modalities*	*type of variable*
eye color	*black, blue, green, brown*	*nominal qualitative*
degree of satisfaction with one's standard of living	*very satisfied, satisfied, dissatisfied*	*ordinal qualitative*

❻ **Statistical series**: The simplest form of presenting statistical data relating to a single character or variable consists of a simple enumeration of the values taken by the character.

❼ **Absolute frequency** n_i: is the number of statistical elements relating to a given modality.

❽ **cumulative absolute frequency** $n_i^c \uparrow$: the number of individuals which correspond to the same modality and to the previous modality.

❾ **Relative frequency** f_i : the ratio $\frac{n_i}{n}$.

❿ **cumulative relative frequency** $f_i^c \uparrow$: the ratio $\frac{n_i^c \uparrow}{n}$.

Example 1.1.3. *The marks of 9 students in a group are as follows*

Notes	n_i	$n_i^c \uparrow$	f_i	$f_i^c \uparrow$
5	*2*	*2*	*2/9*	*2/9*
6	*1*	*3*	*1/9*	*1/3*
8	*3*	*6*	*1/3*	*2/3*
12	*2*	*8*	*2/9*	*8/9*
16	*1*	*9*	*1/9*	*1*
Total	$n = 9$		$\sum_{i=1}^{5} f_i = 1$	

① **Class (Interval)**: we call class a grouping of values of a variable according to intervals which can be equal or unequal. It is mainly used when the variable studied is continuous quantitative.

For each class we can define:

- A lower limit
- An upper limit
- Amplitude = upper limit - lower limit

- Class center $c_i = \dfrac{lower\ limit + upper\ limit}{2}$.

Example 1.1.4. *: The blood glucose level (glycemia) in 14 subjects in g/l*

class	c_i	n_i	$n_i^c \uparrow$	f_i	$f_i^c \uparrow$
[0,85 ; 0,91[	*0,88*	*3*	*3*	*3/14*	*3/14*
[0,91 ; 0,97[	*0,94*	*5*	*8*	*5/14*	*4/7*
[0,97 ; 1,03[	*1*	*3*	*11*	*3/14*	*11/14*
[1,03 ; 1,09[	*1,06*	*2*	*13*	*1/7*	*13/14*
[1,09 ; 1,15[	*1,12*	*1*	*14*	*1/14*	*1*
Total		*n=14*			$\sum_{i=1}^{5} f_i = 1$

1.2 Data description

Depending on the type of variable studied. There are two forms of presentation to describe a series of statistical data: tables and graphical representations.

1.2.1 Tables

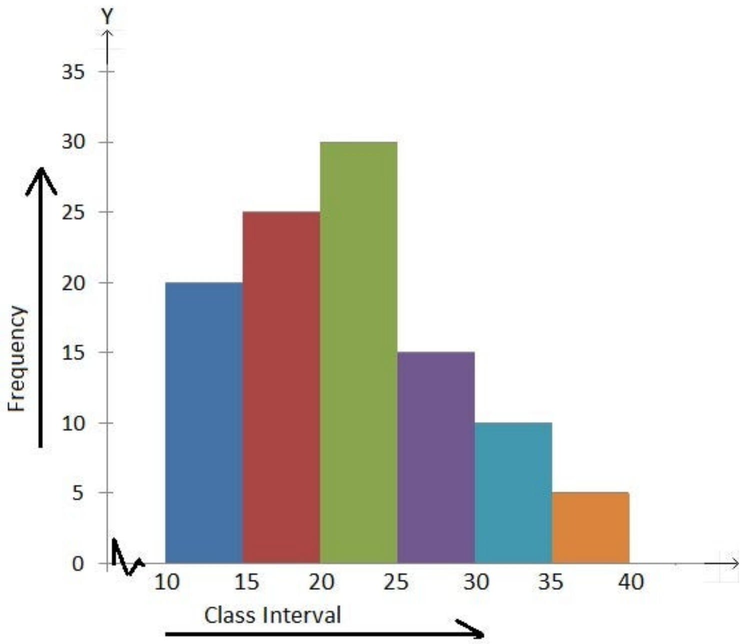
The table can be used whatever the nature of the data, it is used to present the data in an accurate and complete manner.

1.2.2 Graphics

The objective of the graphs is to bring out a systematic vision of the phenomenon studied by illustrating a general trend and giving an overall picture of the results.

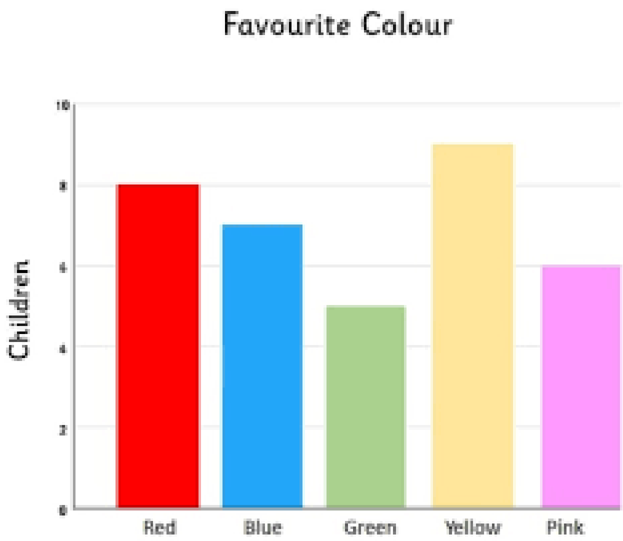
Histogram

Histograms are surfaces that allow the representation of a continuous quantitative variable.



Bar graphs

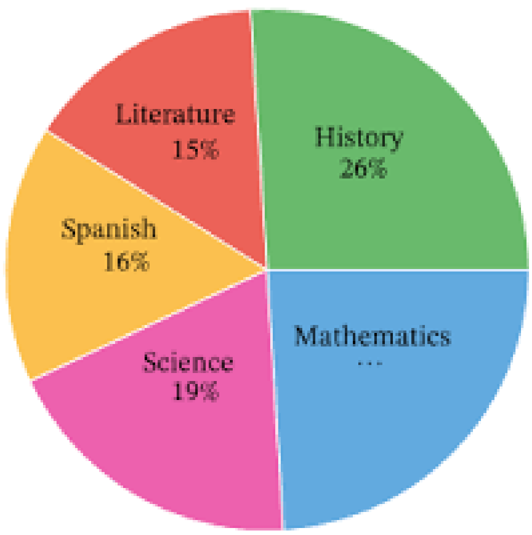
A bar graphs is a graphic representation reserved mainly for a qualitative variable using rectangles of the same width.



Circle graph or the Pie chart

We draw on a disk sections corresponding to the modalities of the character whose angles are proportional to the percentages.

$$\alpha_i = 360^0 * f_i = 360^0 * \frac{n_i}{n}$$



1.3 Position parameters

Central tendency or position parameters: values located in the center of the statistical distribution which are the mean, mode and median.

1.3.1 Mean

Case of a discrete statistical variable

Let X be a discrete statistical variable and $x_1, x_2, ..., x_k$ its values for which correspond the numbers $n_1, n_2, ..., n_k$, with $n = \sum_{i=1}^{k} n_i$

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

Example 1.3.1.

x_i	0	1	2	3	4
n_i	2	3	1	1	1

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i = \frac{1}{8}\Big(0\times 2 + 1\times 3 + 2\times 1 + 3\times 1 + 4\times 1\Big) = \frac{12}{8} = 1.5.$$

Case of a continuous statistical variable

Observations are grouped into classes, so

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i.$$

Example 1.3.2.

class	c_i	n_i
[1,2[	1.5	3
[2,3[	2.5	1
[3,4[	3.5	2

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{3} n_i c_i = \frac{1}{6}\Big(3 \times 1.5 + 1 \times 2.5 + 2 \times 3.5\Big) = \frac{14}{6} = 2.33.$$

1.3.2 Mode

Case of a discrete statistical variable

The mode Mo is the most commonly occurring value.

Example 1.3.3.

x_i	2	3	5	6	7	8	9	10
n_i	2	1	1	2	2	1	1	1

$Mo = 2, 6, 7$

Case of a continuous statistical variable

In this case the mode is calculated by the formula

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2}\right) a$$

• L_i: the lower limit of the modal class (the class that has the highest frequency)

• d_1 = the absolute frequency of the modal class- the absolute frequency of the previous class $(n_i - n_{i-1})$.

• d_2 =the absolute frequency of the modal class- the absolute frequency of the next class $(n_i - n_{i+1})$.

• a: the amplitude of the modal class.

Example 1.3.4.

class	n_i
[1,60-1,65[	3
[1,65-1,70[	8
[1,70-1,75[	2

• *The modal class is:* $[1,65 - 1,70[$.

• $L_i = 1,65$.

• $d_1 = 8 - 3 = 5$.

• $d_2 = 8 - 2 = 6$.

• $a = 1,70 - 1,65 = 0.05$ *then* $Mo = 1,65 + \left(\frac{5}{5+6}\right) 0.05 = 1,67$

1.3.3 Median

Case of a discrete statistical variable

The median Me is the value at the center of a series of numbers arranged in ascending order.

• If n is even, then

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

• If n is odd, then

$$Me = x_{\frac{n+1}{2}}$$

Example 1.3.5. *The number of children of 6 families is as follows*

$$7, \ 3, \ 1, \ 1, \ 5, \ 2$$

We first order the values:

$$\underbrace{1, \ 1, \ 2,}_{3} \ \underbrace{3, \ 5, \ 7}_{3}$$

We have $n = 6$ *is even so* $Me = \frac{x_3 + x_4}{2} = \frac{2+3}{2} = 2.5$.

Example 1.3.6. *The number of children of 7 families is as follows*

$$3, \ 2, \ 1, \ 0, \ 0, \ 1, \ 2$$

We first order the values:

$$\underbrace{0,\ 0,\ 1,}_{3}\ \underbrace{1}_{Me=x_4=1},\ \underbrace{2,\ 2,\ 3}_{3}$$

We have $n = 7$ *is odd so* $Me = x_4 = 1$.

Case of a continuous statistical variable

In this case the median is given by

$$Me = L_i + \left(\frac{\frac{n}{2} - \sum_{i=1}^{<Me} n_i}{n_{Me}} \right) a$$

• L_i: the lower limit of the median class

• $\sum_{i=1}^{<Me} n_i$ = the sum of the absolute frequencies corresponding to all classes below the median class.

• n_{Me}= the absolute frequency of the median class.

• a: the amplitude of the median class.

Example 1.3.7. *According to the example (1.1.4), we obtain*

• *The median class is:* $[0.91 - 0.97[$.

• $L_i = 0.91$.

• $n = 14$.

• $\sum_{i=1}^{<Me} n_i = 3$

• $n_{Me} = 5$.

• $a = 0.97 - 0.91 = 0.06$

then $Me = 0.91 + \left(\frac{7-3}{5}\right) 0.06 = 0.958$

1.3.4 Quartiles

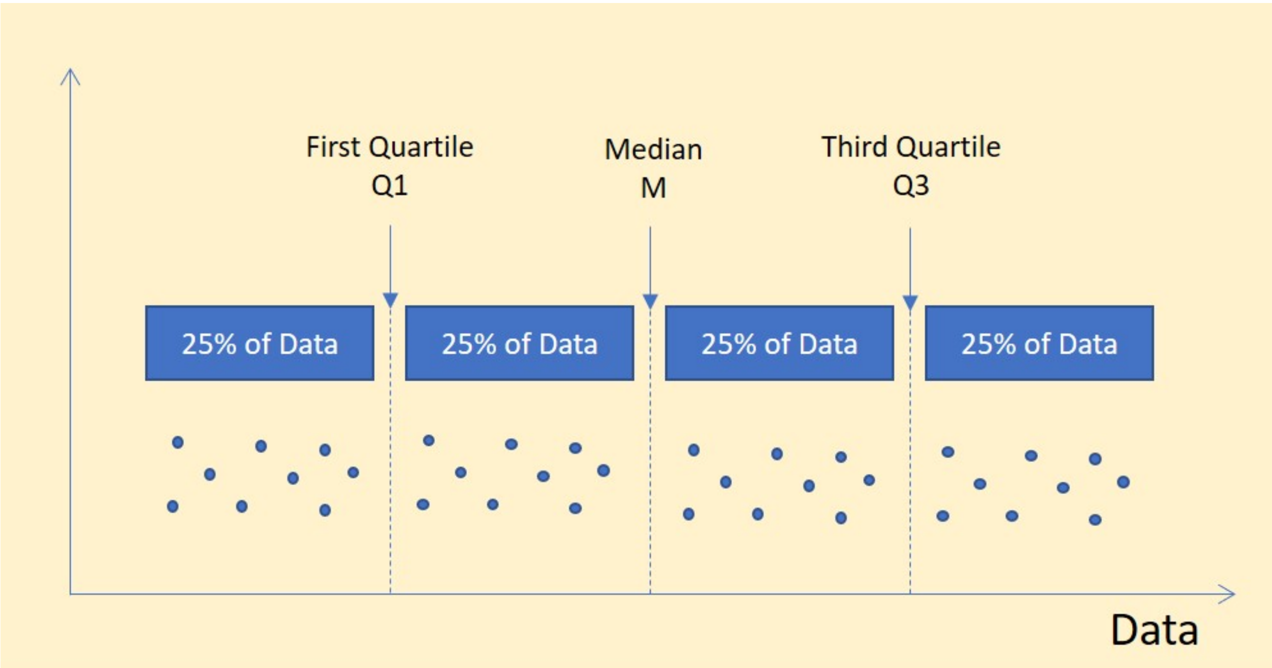
Case of a discrete statistical variable

Quartiles are the three values that divide the distribution into four equal parts.

• **The first quartile Q_1** represents 25% of the sample i.e. Q_1 is the value x_i whose position is the smallest integer following $\frac{n}{4}$.

• **The second quartile Q_2** represents 50% of the sample.

• **The third quartile Q_3** represents 75% of the sample i.e. Q_3 is the value x_i whose position is the smallest integer following $\frac{3n}{4}$.



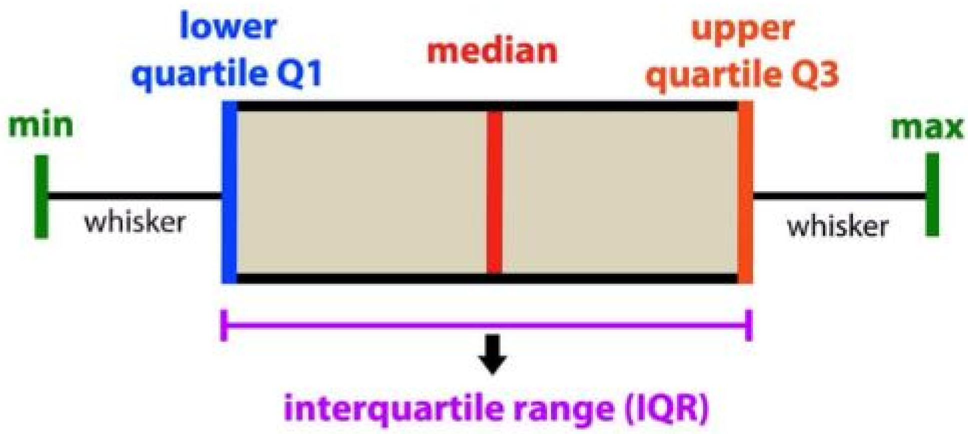
Interquartile range

The interquartile range is the difference between the third and first quartile:

$$I_Q = Q_3 - Q_1$$

.

Boxplot



Example 1.3.8. *In the example of the following observations*

x_i	1	3	5	7	9
n_i	1	2	1	2	2
n_i^c	1	3	4	6	8

- *We have* $n = 8$ *and* $\frac{n}{4} = 2$ *so* Q_1 *is the second value* $Q_1 = x_2 = 3$.

- *We have* $n = 8$ *and* $\frac{3n}{4} = 6$ *so* Q_3 *is the sixth value* $Q_3 = x_6 = 7$.

1.4 Dispersion parameters

Dispersion parameters are the parameters that summarize the dispersion of values around the central value

1.4.1 Range

The difference between the largest value and the smallest value observed is called the range e.

$$e = x_{max} - x_{min}$$

Example 1.4.1. *The marks of 10 students are as follows*

$$2,\ 3,\ 10,\ 10,\ 11,\ 12,\ 15,\ 18,\ 19,\ 20$$

then

$$e = x_{max} - x_{min} = 20 - 2 = 18$$

1.4.2 Variance

A variance is the arithmetic mean of the squares of the differences between the values of a variable and the arithmetic mean.

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \\ &= \sum_{i=1}^{k} f_i(x_i - \bar{x})^2. = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2. \end{aligned}$$

1.4.3 Standard deviation

We call standard deviation denoted σ_X the square root of the variance.

$$\sigma_X = \sqrt{V(X)}$$

1.4.4 Coefficient of variation

The coefficient of variation, CV, is defined by

$$CV = \frac{\sigma_X}{\bar{x}}$$

Example 1.4.2.

x_i	0	1	2	3	4
n_i	2	3	1	1	1

$$\bar{x} = 1.5$$

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \\ &= \frac{1}{8}\sum_{i=1}^{5} n_i x_i^2 - (1.5)^2 \\ &= \frac{1}{8}\left(2\times 0^2 + 3\times 1^2 + 1\times 2^2 + 1\times 3^2 + 1\times 4^2\right) - 2.25 \\ &= \frac{32}{8} - 2.25 \\ &= 1.75 \end{aligned}$$

The standard deviation

$$\sigma_X = \sqrt{V(X)} = \sqrt{1.75} = 1.3$$

and the coefficient of variation

$$CV = \frac{\sigma_X}{\bar{x}} = \frac{1.3}{1.5} = 0.87$$

1.5 Shape parameter

1.5.1 Skewness

There are several coefficients, the main ones are as follows:

• Pearson's skewness coefficient

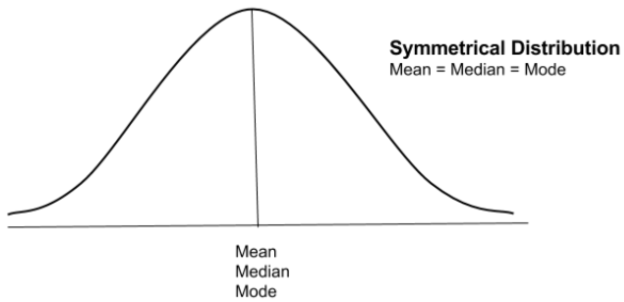
$$A_P = \frac{\bar{x} - Mo}{\sigma_X}.$$

• Yule's skewness coefficient:

$$A_Y = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}.$$

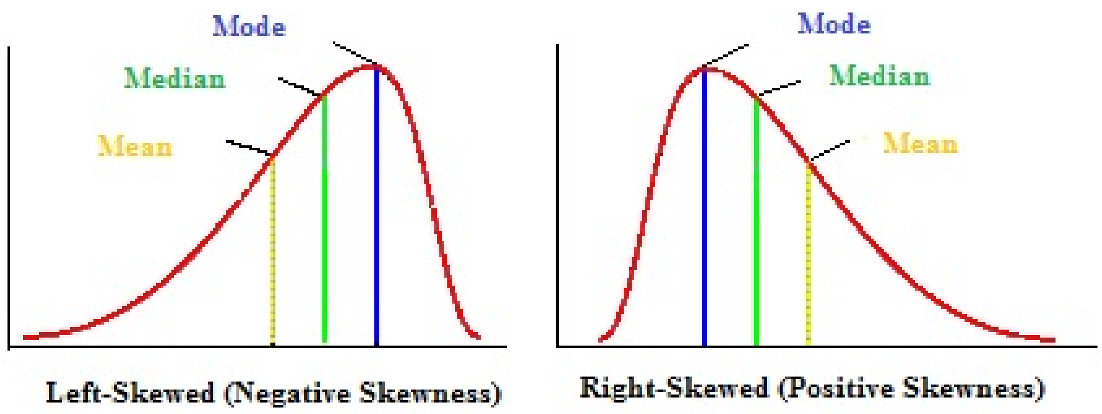
Remark

• A zero coefficient indicates that the distribution is symmetrical.



• A positive coefficient indicates a right-skewed distribution.

• A negative coefficient indicates a left-skewed distribution.



1.5.2 kurtosis

• Pearson′s kurtosis coefficient:

$$AP_P = \frac{m_4}{\sigma_X^4}$$

where m_4 is the centred moment of order 4 defined by

$$m_4 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^4$$

• Fisher′s kurtosis coefficient:

$$AP_F = \frac{m_4}{\sigma_X^4} - 3$$

Remark

• If $AP_F = 0$ then the distribution is called "normal" or "mesokurtic".

• If $AP_F < 0$ then the distribution is called "platykurtic".

• If $AP_F > 0$ then the distribution is called "leptokurtic".