Title:

Probability-Statistics

	Course	*GW*
VHS	*1h30*	*1h30*

Mr Medjati . R

Email : medjati@yahoo.fr

Table of contents:

Part 1. Descriptive statistics.

Chapter 1 : *Statistical series of 1 variable.*

1.1- Introduction: Generalities, definitions, Types of variables.

1.2- Statistical series and their representations:

1.2.1- Discrete quantitative case.

1.2.2- Continuous quantitative case.

1.2.3- Qualitative case.

1.3- Measures of the statistical series

1.3.1- Measures of position :

Mode, arithmetic mean and median

1.3.2- Measures of Dispersion :

Variance and standard deviation.

Partie 1. Descriptive statistics.

Chapter 2 : Statistical series of 2 variables.

2.1- Introduction.

2.2- Distribution and characteristics :

2.2.1- Marginal distributions.

2.2.2- Marginal characteristics:
Means and marginal variances.

2.2.3- Conditional distribution.

2.2.4- Conditional characteristics.

2.3- Covariance of 02 variables

1.3.1- Definition and properties.

1.3.2- Correlation coefficient

2.4- Adjustments : Type $y = ax+b$, Type $y = Ba^x$

Part 2. *Probabilities.*

Chapter 3 : ***Introduction to Probability Calculus.***

3.1- Reminders about combinatorial analysis:
Permutation, arrangement, combination.

3.2- Basic concepts

3.3- Conditional probability
3.3.1- Theorem of total probability.
3.3.2- Bayes' theorem.
3.3.3- Independent events.

<u>Chapter 1</u> : *Statistical series of 1 variable.*

1.1- Introduction :

1.1.1- <u>The statistics</u>:

For a group of individuals or objects statistics is the study of :

1. *The data collection.*

2. *Their analysis, their treatment and the interpretation of the results.*

3. *Their presentation to make the data understandable to everyone.*

1.1.2- A statistical population :

A statistical population is the set on which observations are carried out.

Examples :

1. *Set of people interviewed for a survey.*
2. *Set of countries for which geographic or economic data are available, ...*

1.1.3- Individual (or statistical units):

The individuals are the elements of the statistical population studied. For each individual, we have one or more observations.

Examples :

1. *Each person interviewed for an investigation.*
2. *Each country for which we study socio-economic data, …*
3. *Every day of the year for which weather data is available, ...*

1.1.4- Statistical variable :

This is what is observed or measured on individuals in a statistical population.

It can be a qualitative or quantitative variable.

Examples *:*

1. *Size, weight, salary, gender, profession of a given group of individuals.*

2. *Maximum and minimum temperature, rainfall and sunshine, measured at a given location every day.*

A. Qualitative variable:

A statistical variable is qualitative if its values, or modalities, are expressed literally or through coding. (*i.e. an observation which is not measurable*).

Examples *:*

1. *Gender, family situation,…*

2. *Weather conditions observed at a given location each day* (*rainy, snowy, sunny, windy, etc*.)

B. Quantitative variable :

A statistical variable is quantitative if its values are numbers on which arithmetic operations such as sum, average, etc. have meaning.

Remark: *Quantitative variables can be discrete or continuous.*

B.1 Discrete quantitative variable:

It is a quantitative variable that can take a finite number by nature (or countable) of values by nature.

Exemples *:*

1. *Number of children per family.*
2. *Number of rooms in a flat.*

B.2 <u>Continuous quantitative variable</u>:

It is a quantitative variable that can take on an infinite number of values by nature, generally an entire real interval.

<u>Examples</u> *:*
Height, weight, wages, cultivated areas, temperature.

<u>Remark</u> **:** *In this case we use intervals (class)* $[a_i , b_i[$ *instead of* x_i *.*

1.1.5- <u>Modality</u>:

The modalities of a variable are the different results of the observation **(*numbers or properties*)***.*

Examples :

1. ***Qualitative case:***

The modalities of the variable $\boldsymbol{X}$ *= "family situation" are : M =* $\{$*single, married, widowed, divorced*$\}$.

2. ***Discrete quantitative case:***

The modalities of the variable $\boldsymbol{X}$ *= "Score on an exam" are : M =* $\{7; 9; 14; 16{,}5\}$.

3. ***Continuous quantitative case:***

The modalities of the variable $\boldsymbol{X}$ *= " Size" are the values belonging to the intervals (class)* $[150, 165[$, $[165, 180[$, *etc…*

Remark *: There are 2 types of qualitative statistical variables;*

1er - **Nominal qualitative variable**:

The variable is called qualitative nominal when the modalities cannot be ordered (cannot be classified).

Example 1 :

The variable X = «family situation» with the modalities noted : C, M, V, D.

Example 2 :

The variable X = "gender" with the modalities noted : M, F.

2eme - **Ordinal qualitative variable** : $\prec$ $<$

The variable is called ordinal qualitative when the modalities can be ordered. If, M $= \{x_1, x_2 \ldots, x_r\}$ *designates the set of modalities, these values are ordered, that is to say :*

$x_1 \prec x_2 \prec \cdots \prec x_r$. *The notation* $x_1 \prec x_2$ *is read as* x_1 *comes before* x_2.

Example 1 :

A satisfaction questionnaire asks consumers to evaluate a service by checking one of the following six categories :

(*a*) *poor,* (*b*) *average,* (*c*) *fairly good,* (*d*) *very good*, (*e*) *excellent*.

Example 2 :

La variable $\boldsymbol{X}$ *= « Educational level ».*

1.1.6 Frequency and Relative frequency of a modality :

- ***The frequency n_i of a modality*** *(or of class) is the number of times where the modality (resp class) n° **i** was observed.*

- ***The total frequency*** *N is the total number of observed individuals.*

$$\boxed{N = n_1 + n_2 + \cdots + n_r = \sum_{i=1}^{r} n_i}$$

- ***The relative frequency f_i of a modality*** *(or of class) is the frequency n_i divided by the total frequency N.*

$$\boxed{f_i = \frac{n_i}{N} = \frac{n_i}{\sum_{i=1}^{r} n_i}}$$

Remark : *The relative frequencies can be expressed as percentages, and we have the following result :*

$$\sum_{i=1}^{r} f_i = 1 \text{ car } \sum_{i=1}^{r} f_i = \sum_{i=1}^{r} \frac{n_i}{N} = \frac{1}{N} \sum_{1}^{r} n_i = \frac{N}{N} = 1$$

Exemple : *Out of 200 families, 50 have 2 children, we would say that the relative frequency f_i corresponding to the value $x_i = 2$ of the variable "number of children", is :*

$$f_i = \frac{n_i}{N} = \frac{50}{200} = \frac{1}{4} = 0{,}25 \quad soit \quad 25\%$$

1.1.7 Presentation in a statistical table:

A. Nominal qualitative case : *For a nominal qualitative statistical variable, if the set $M = \{M_1, M_2, \ldots, M_r\}$ designates all the modalities, then the statistical table associated with this variable is :*

Modalities (numbered) M_i	Frequencies n_i	Relative fréquencies f_i
M_1 (1)	n_1	f_1
M_2 (2)	n_2	f_2
.	.	.
.	.	.
.	.	.
M_r (r)	n_r	f_r
Total	N	1

Example 1:

We are interested in the values of the variable $\boldsymbol{X}$ *= «family situation» taken from 20 people whose coding is ;*

c : Single, m : married, v = widowed, d = widowed, divorced. So the variable domain $\boldsymbol{X}$ *is M =* $\{c, m, v, d\}$.

Consider the following results:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
m	m	d	c	c	m	c	c	c	m	c	m	v	m	v	d	c	c	c	m

And we obtain the following table:

M_i	n_i	f_i
c	9	0,45
m	7	0,35
v	2	0,10
d	2	0,10
Total	20	1,00

Remark 1 : *Before tackling other cases we define the following ;*

1- Increasing relative cumulative frequencies F_i **(**f_{ic}**) :** *is*

$$F_1 = f_1, F_2 = f_1 + f_2 \quad et$$

$$F_i = f_{ic} = f_1 + f_2 + \cdots + f_i = \sum_{p=1}^{i} f_p$$

2- Decreasing relative cumulative frequencies F'_i **(**f_{id}**) :** *C'est*

$$F'_r = f_r, F'_{r-1} = f_r + f_{r-1} \quad et$$

$$F'_i = f_{id} = f_r + f_{r-1} + \cdots + f_i = \sum_{p=i}^{r} f_p$$

Remark 2 : *In the same way we define the increasing cumulative frequencies* N_i **(**n_{ic}**)** *and the decreasing cumulative frequencies* N'_i **(**n_{id}**)**.

$$N_i = n_{ic} = \sum_{p=1}^{i} n_p \qquad et \quad N'_i = n_{id} = \sum_{p=i}^{r} n_p$$

B. Ordinal qualitative case :

If M = $\{x_1, x_2, \ldots, x_r\}$ *designates the set of modalities, these values are ordered, that is to say :* $x_1 \prec x_2 \prec \cdots \prec x_r$.

With $x_1 \prec x_2$ *is read as* x_1 *comes before* x_2*. then the statistical table associated with this variable is :*

x_i	Frequencies n_i	Increasing cumulative frequencies N_i	relative frequencies f_i	Increasing relative cumulative frequencies F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_r	n_r	N_r	f_r	F_r
Total	N		**1**	

Example :

20 shirts are classed by size :

x_1 = S, x_2 = M, x_3 = L, x_4 = XL, et x_5 = XXL.

The table associated is :

x_i	Frequencies n_i	Increasing cumulative frequencies N_i	relative frequencies f_i	Increasing relative cumulative frequencies F_i
x_1	4	4	0,20	0,20
x_2	2	6	0,10	0,30
x_3	5	11	0,25	0,55
x_4	8	19	0,40	0,95
x_5	1	20	0,05	1,00
Total	20		1,00	

Remark 3 :

The discrete quantitative case is done in the same way as the case, and we obtain a statistical table similar to that of the ordinal qualitative case.

And in the **continued quantitative case** we will have:

Classes $[b_{i-1}, b_i[$	Centers c_i	n_i	N_i	f_i	F_i
$[b_0, b_1[$	c_1	n_1	N_1	f_1	F_1
$[b_1, b_2[$	c_2	n_2	N_2	f_2	F_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$[b_{r-1}, b_r[$	c_r	n_r	N_r	f_r	F_r
Total		***N***		***1***	

Remarks :

1. The **center** of a class is : $$c_i = \frac{b_{i-1} + b_i}{2}, \quad c_i \approx x_i$$

2. The **amplitude** of a class is : $$a_i = b_i - b_{i-1}$$

Example :

The distribution of 100 households according to their monthly consumption expenses expressed ***in thousands of dinars*** *is as follows :*

Expense Classes	Number of households
[20-40[	15
[40-60[	20
[60-100[	20
[100-200[	45

And The table associated is :

Classes	Centers c_i	n_i	N_i	f_i	F_i
[20- 40[	30	15	15	0,15	0,15
[40-60[	50	20	35	0,20	0,35
[60-100[	80	20	55	0,20	0,55
[100-200[	150	45	100	0,45	1,00
Total		100		1,00	

Calculation example :

1- Decreasing relative cumulative frequencies F'_i and

2- Decreasing cumulative frequencies N'_i

$$F'_r = f_r, F'_{r-1} = f_r + f_{r-1} \quad ; \quad F'_i = f_{id} = f_r + f_{r-1} + \cdots + f_i = \sum_{p=i}^{r} f_p \quad et \quad N'_i = n_{id} = \sum_{p=i}^{r} n_p$$

Classes	Centers c_i	n_i	N'_i	f_i	F'_i
[20- 40[	30	15	100	0,15	1,00
[40-60[	50	20	85	0,20	0,85
[60-100[	80	20	65	0,20	0,65
[100-200[	150	45	45	0,45	0,45
Total		*100*		*1,00*	

Remark : *All the couples;*

1- $\{(x_i, n_i)\}$ or $\{(x_i, f_i)\}$ if the variable is discrete.

2- $\{([b_{i-1}, b_i[, n_i)\})$, or $\{([b_{i-1}, b_i[, f_i)\})$ if the variable is continuous.

Is called **the statistical series of the variable**.

1.2- Diagrammatic and Graphic presentation of data :

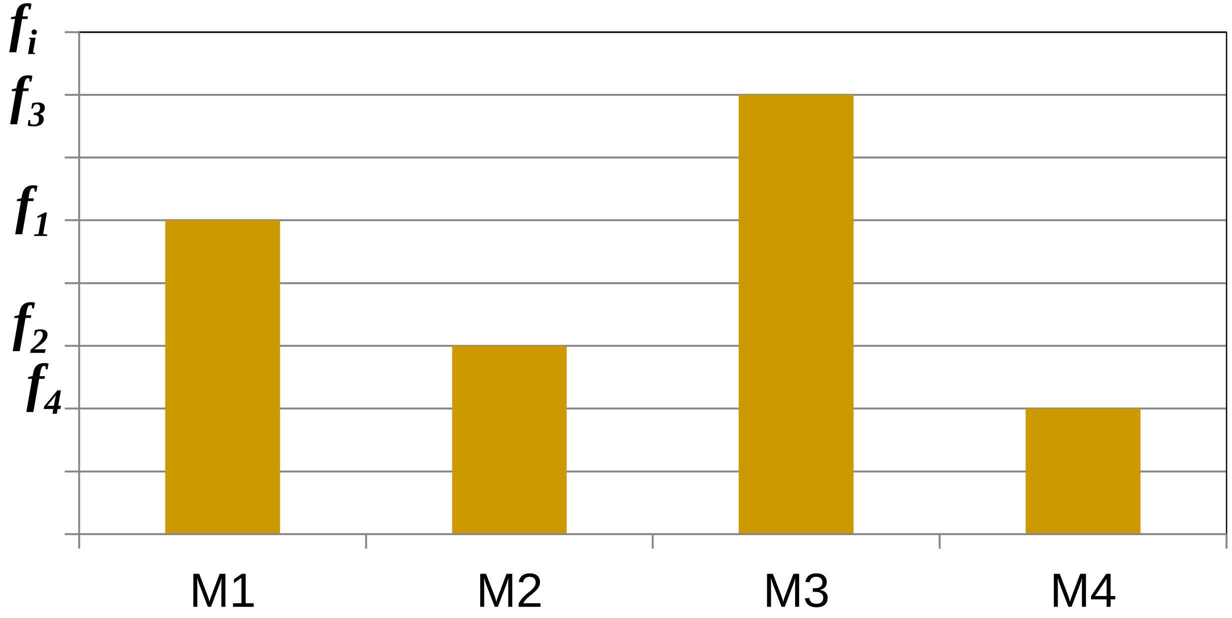
Equipe de probabilités
Coordinateur : Mr Medjati

1.2.1- Graphics representations- Qualitative case.

A- Bars Diagram :

It is a Cartesian benchmark such that :

for each modality M_i *we associate a rectangle with* ***constant base*** *whose height is the frequency* n_i *(The relative frequency* f_i*).*



Remark :

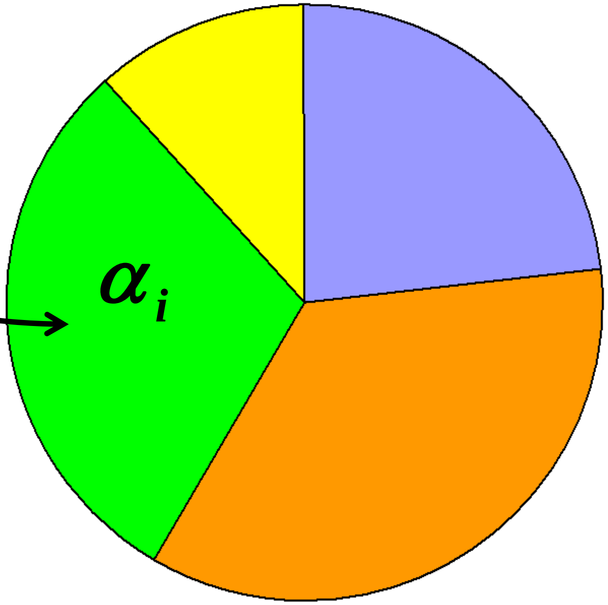
For the axis of frequencies (relative frequencies), we choose an arithmetic scale.

B- Pie chart :

It is a graph where the modalities are represented by portions of disk proportional to their frequencies, or to their relative frequencies.

In effect; for a modality $\mathbf{M_i}$*, of frequency* $\mathbf{n_i}$*, the angle at the center* $\boldsymbol{\alpha_i}$ *corresponding is given (in degree) by:*

$$\alpha_i = f_i \times 360° = \frac{n_i}{N} \times 360°$$



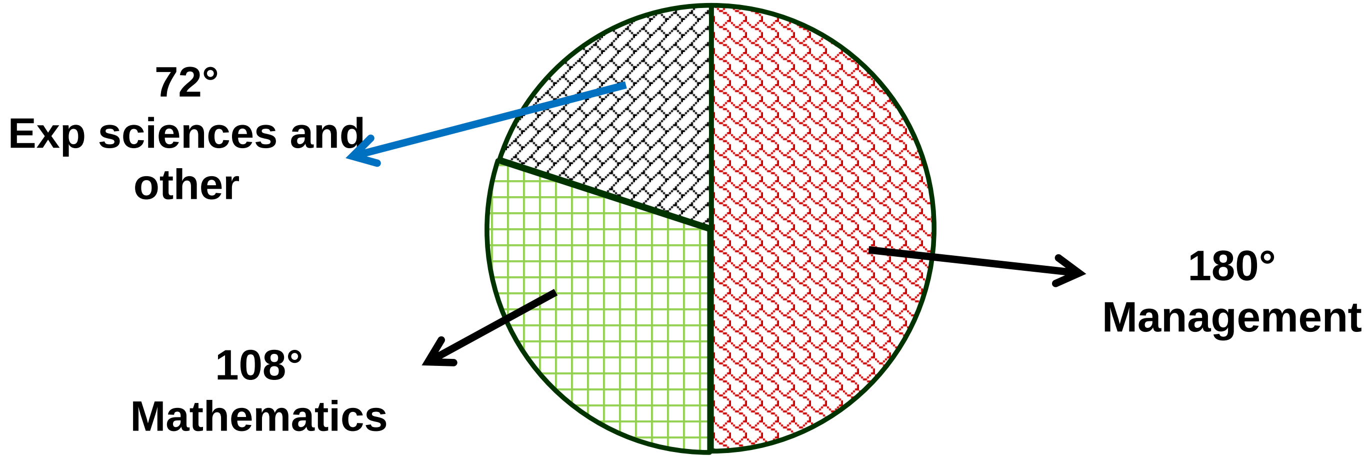
RemarK :

Bar and pie chart can be used in the quantitative case.

Example :

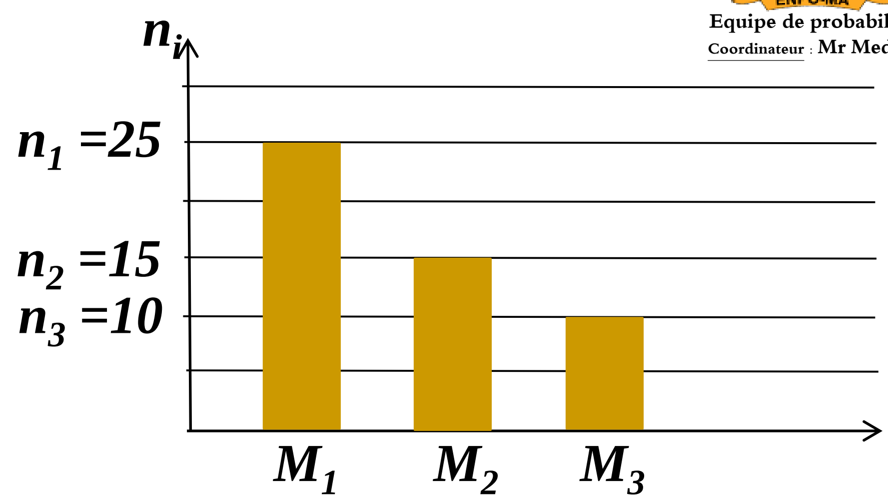
According to a study carried at the Oran business school, the distribution of 50 students according to the branch of the baccalaureate is reported in the following table:

the branch of bac M_i	n_i	f_i	***Angles α_i***
Management	***25***	***0,50***	***180°***
Mathematics	***15***	***0,30***	***108°***
Exp sciences and other	***10***	***0,20***	***72°***
Total	***50***	***1,00***	***360°***



And the associated bar chart is :

Branch of bac M_i	n_i	f_i
Management	25	0,50
Mathématics	15	0,30
Exp sciences and other	10	0,20
Total	50	1,00



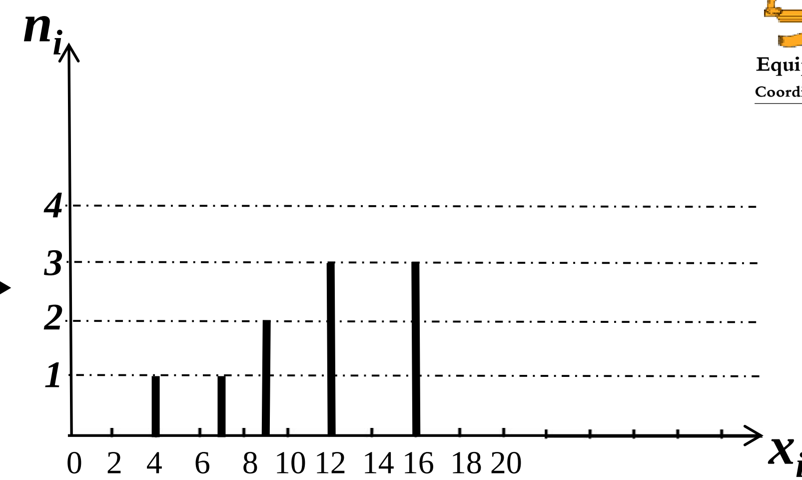
1.2.2- Discrete quantitative case :

- **Lines Diagram :**

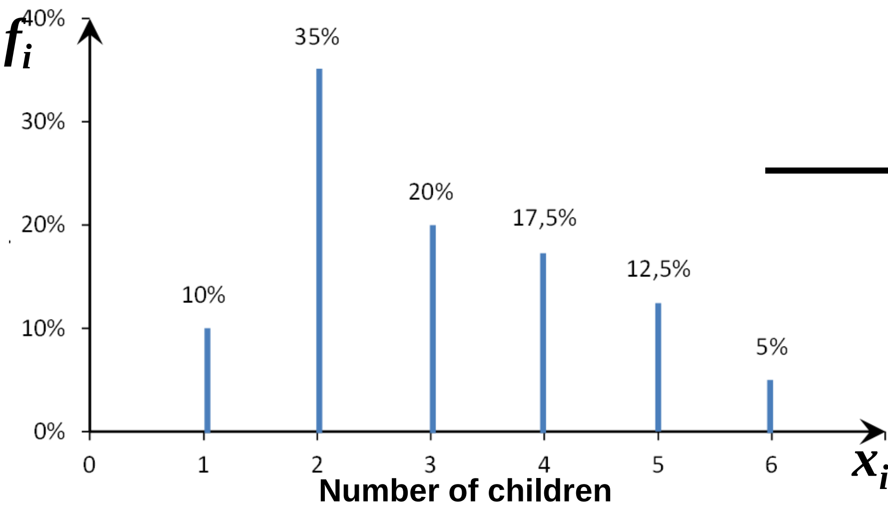
It is a Cartesian benchmark such that the values are placed on the abscissa, the frequencies (or relative frequencies) on the ordinate, and at every point (x_i, 0) we associate a vertical segment whose length is the frequency n_i (relative frequency f_i).

Example 1 :

x_i	n_i
4	1
7	1
9	2
12	3
16	3
Total	10



Example 2 :



x_i	f_i %
1	10
2	35
3	20
4	17,5
5	12,5
6	5
Total	100

1.2.3 - Continuous quantitative case:

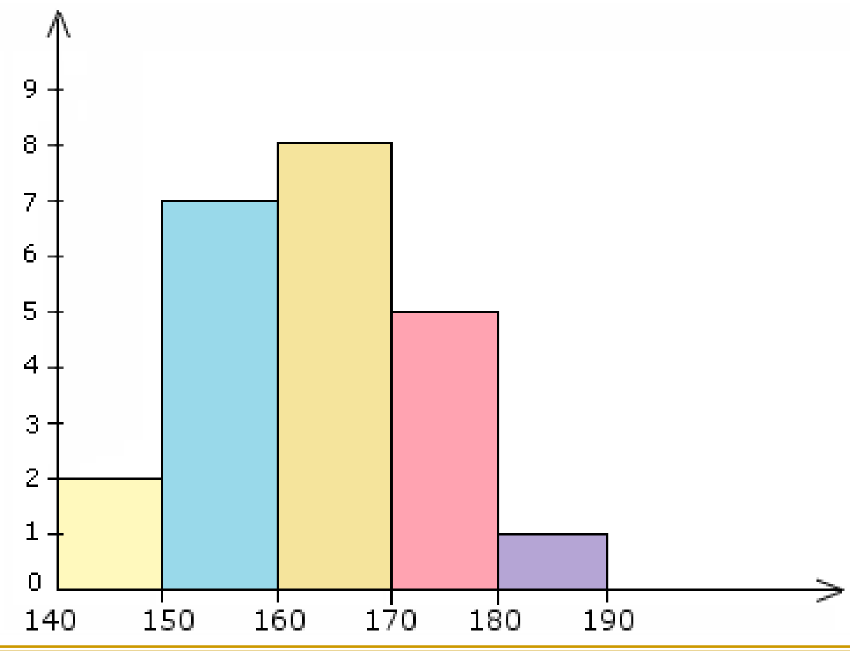
*We represent a continuous statistical series by **a histogram***

Equipe de probabilités
Coordinateur : Mr Medjati

Definition :

This is a figure obtained on a Cartesian coordinate system by representing for each class $[\, b_{i-1} \,,\, b_i \,[$ *a rectangle of area* $\underline{S_i}$ *proportional to the frequency* n_i *or the relative frequency* f_i*.*

*The rectangles of the histogram **are neighboring**.*



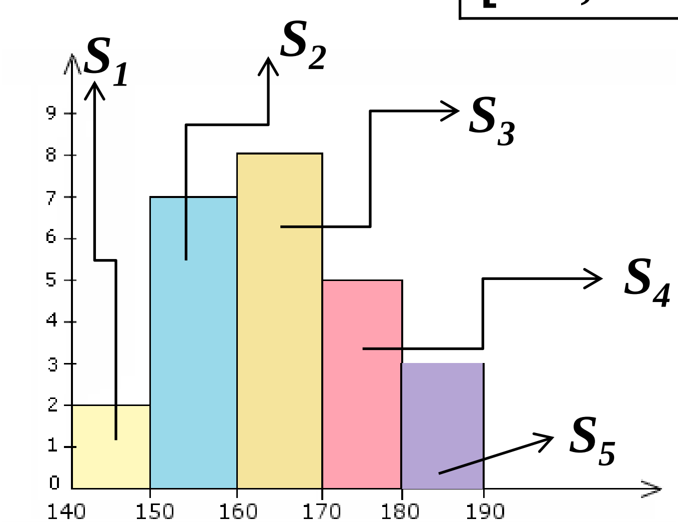
<u>Principle of histogram construction</u> : *there are two (02) cases.*

<u>1st case</u> :

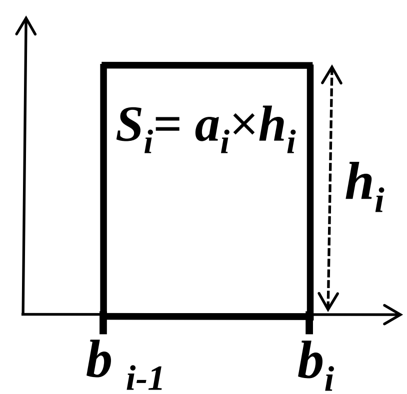
If the classes are of the same amplitude a_i *(ie* $a_i = a_j$*), we place the frequencies* n_i *on the ordinate (or relative frequencies* f_i *).*

<u>Example</u> :

$[b_{i-1}, b_i[$	n_i	a_i	f_i
[140 , 150[	2	10	0,08
[150 , 160[	7	10	0,28
[160 , 170[	8	10	0,32
[170 , 180[	5	10	0,20
[180 , 190[	3	10	0,12



<u>Principle</u>



2nd case :

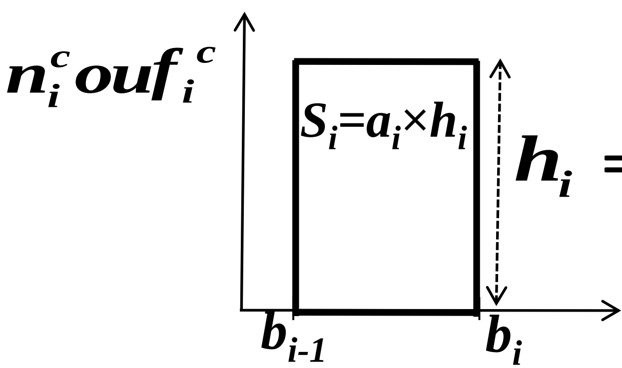
If the amplitudes a_i are different (ie $a_i \neq a_j$), we define;

- ***The density of a class*** *by* $d_i = \dfrac{n_i}{a_i}$ *and we put;*

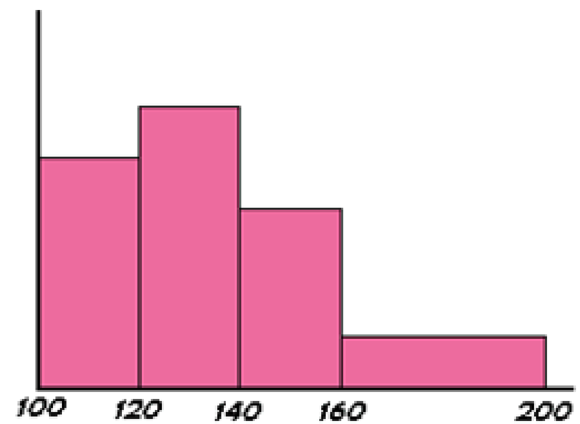
$$h_i = \frac{n_i}{a_i} \times a^* = d_i \times a^* = n_i^c$$

- *With a^* is called **the reference amplitude**. It is chosen arbitrarily so as to facilitate graphical representation (values on the ordinates axis).*

- *h_i is in this case called **corrected frequency** which we note n_i^c. So the rectangles S_i will be as follows :*



$$h_i = n_i^c \,(ou \;\; f_i^c)$$



Equipe de probabilités
Coordinateur : Mr Medjati

Example :

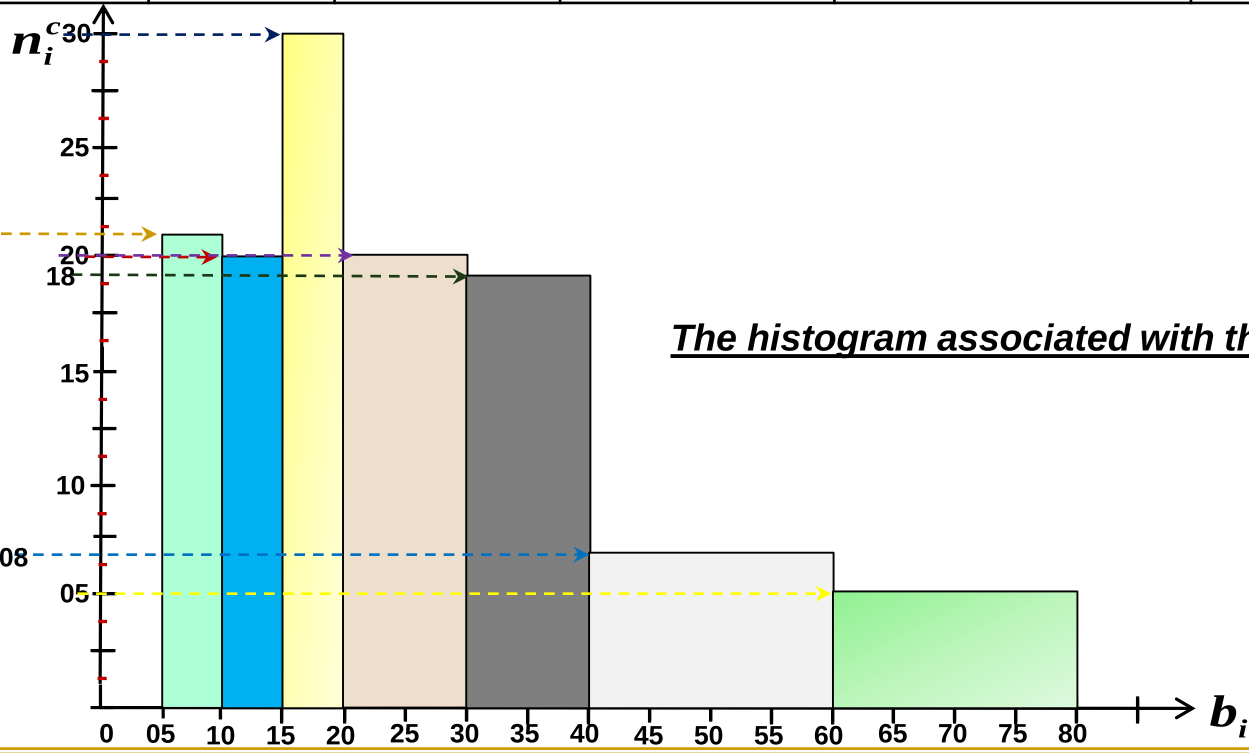
The distribution of 100 individuals by age classes is given by the next table ;

Classes $[b_{i-1}, b_i[$	n_i	Amplitude $a_i = b_i - b_{i-1}$	Density $d_i = \frac{n_i}{a_i}$	corrected frequency $n_i^c = d_i \times a^*$	f_i	Rel- Freq - cor f_i^c
[5 , 10[	11	5	2,2	22	0,11	0,22
[10 , 15[	10	5	2	20	0,10	0,20
[15, 20[	15	5	3	30	0,15	0,30
[20, 30[	20	10	2	20	0,20	0,20
[30 , 40[	18	10	1,8	18	0,18	0,18
[40 , 60[	16	20	0,8	08	0,16	0,08
[60, 80[	10	20	0,5	05	0,10	0,05
Total	100				1,00	

Remark :

In this example the reference amplitude a = 10.*

$[b_{i-1}, b_i[$	n_i	$a_i = b_i - b_{i-1}$	d_i	n_i^c	f_i	f_i^c
[5 , 10[	11	5	2,2	22	0,11	0,22
[10 , 15[	10	5	2	20	0,10	0,20
[15, 20[	15	5	3	30	0,15	0,30
[20, 30[	20	10	2	20	0,20	0,20
[30 , 40[	18	10	1,8	18	0,18	0,18
[40 , 60[	16	20	0,8	08	0,16	0,08
[60, 80[	10	20	0,5	05	0,10	0,05
Total	100				1,00	



The histogram associated with this statistical series

1.3- Cumulative Distribution Function :

We call Cumulative Distribution Function of a quantitative statistical variable any application defined by :

$$F : \mathbb{R} \to [0\,,\,1]$$

$$x \to F(x) = P(X \leq x)$$

$F(x)$ *proportion of individuals whose value of the variable is strictly less than or equal to* x, *that's to say* $X \leq x$.

1- Discrete statistical variable case :

$F(x)$ = *relative frequency of* $(X \leq x) = f_1 + f_2 + \ldots + f_p = F_p$ *such as :* $f_1, f_2, \ldots, f_p$ *are the relative frequencies of the values of the variable* $\leq x$, *Otherwise* $F(x) = 0$. *Therefore*

Equipe de probabilités
Coordinateur : Mr Medjati

$$F(x)=\begin{cases} 0 \quad si \quad x < x_1 \\ F_i \quad si \quad x_i \le x < x_{i+1} \\ 1 \quad si \quad x_r \le x \end{cases}$$

such as ***r*** *designates the order of* ***<u>the last value (modality)</u>***.

Example :

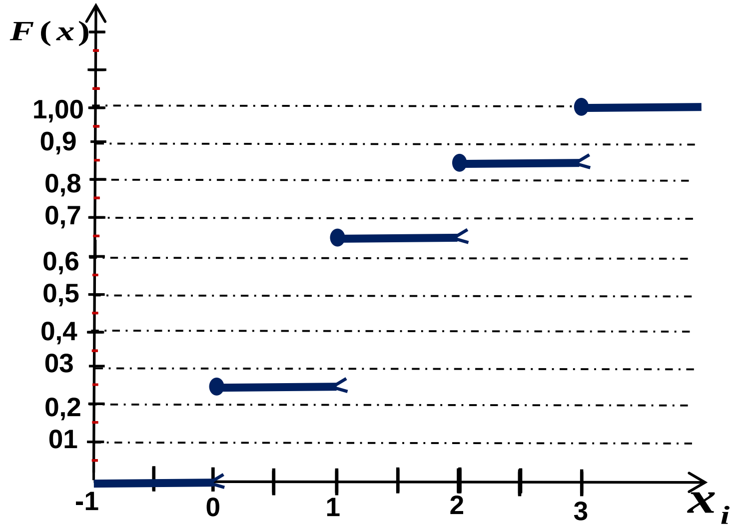
The following table gives the number of student absences from the analysis module.

number of absences x_i	n_i	f_i	F_i
0	5	0,25	0,25
1	8	0,40	0,65
2	4	0,20	0,85
3	3	0,15	1,00
Total	20	1,00	

$$F(x)=\begin{cases} 0 & si \quad x<0 \\ 0{,}25 & si \quad 0\leq x<1 \\ 0{,}65 & si \quad 1\leq x<2 \\ 0{,}85 & si \quad 2\leq x<3 \\ 1 & si \quad 3\leq x \end{cases}$$

$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ 0{,}25 & si \quad 0 \le x < 1 \\ 0{,}65 & si \quad 1 \le x < 2 \\ 0{,}85 & si \quad 2 \le x < 3 \\ 1 & si \quad 3 \le x \end{cases}$$

Thus we obtain the representation Cumulative Distribution Function, called a ***cumulative chart*** *.*



Remark : *In the discrete case we have a staircase function.*

Exercice **:** *A running association has a women's team. The following list is made up of the first names of athletes followed in parentheses from the last 10 Km times..*

Aicha**(*****51*****),** ***Ahlem*****(*****49*****),** ***Amel*****(*****50*****),** ***Badra*****(*****58*****),** ***Bouchra*****(*****55*****),** ***Dalia(64)*****,** ***Fadia*****(*****60*****),** ***Fahima*****(*****61*****),** ***Fatiha*****(*****46*****),** ***Fatima*****(*****56*****),** ***Fouzia*****(*****50*****),** ***Hajera*****(*****42*****),** ***Houria*****(*****54*****),** ***Ikram*****(*****48*****),** ***Ilham*****(*****45*****),** ***Imane*****(*****57*****),** ***Jamila*****(*****59*****),** ***Khadija*****(*****54*****),** ***Lamia*****(*****54*****),** ***Leila*****(*****46*****),*****Meriem*****(*****46*****),** ***Nabila*****(*****41*****),** ***Samia*****(*****39*****),** ***Samira(37)*****,** ***Wafaa*****(*****50*****),** ***Yamina*****(*****47*****),** ***Yasmine*****(*****50*****),** ***Zahira*****(*****44*****),** ***Zakia*****(*****51*****),** ***Zoulikha*****(*****59*****).**

The manager of the association decides to create in an order croissant five (05) teams (classes) of equivalent level such as :
*the **1st** team contains **3** athletes, the **2nd** team contains **3** athletes,*
*The **3rd** team contains **6** athletes, the **4th** team contains **9** athletes,*
*and the **5th** team contains **9** athletes.*

1. *Form the teams. (Make a table giving the minimum and maximum times for each team).*
2. *Give a graphical representation of relative frequencies in the form of a histogram (the reference amplitude a* = 1000).*
3. *Draw the relative frequency polygon.*

Equipe de probabilités
Coordinateur : Mr Medjati

Exercise solution :

1- Constitution of teams : *The list of athletes is;*

Aicha(51), Ahlem(49), Amel(50), Badra(58), Bouchra(55), Dalia(64), Fadia(60), Fahima(61), Fatiha (46), Fatima(56), Fouzia(50), Hajera(42), Houria(54), Ikram(48), Ilham(45), Imane(57), Jamila(59), Khadija(54), Lamia(54), Leila(46), Meriem(46), Nabila(41), Samia(39), Samira(37), Wafaa(50), Yamina(47), Yasmine(50), Zahira(44), Zakia(51), Zoulikha(59).

And we will arrange in ascending order of times : The 3 athletes with the shortest time (the best) will make up team 1;

Equipe 1	Equipe 2	Equipe 3	Equipe 4	Equipe 5
Samira (37)	Hajera (42)	Leila (46)	Amel (50)	Bouchra (55)
Samia (39)	Zahira (44)	Fatiha (46)	Wafaa (50)	Fatima (56)
Nabila (41)	Ilham (45)	Meriem (46)	Yasmine (50)	Imane (57)
		Yamina (47)	Fouzia (50)	Badra (58)
		Ikram (48)	Aicha (51)	Jamila (59)
		Ahlem (49)	Zakia (51)	Zoulikha (59)
			Khadija (54)	Fadia (60)
			Houria (54)	Fahima (61)
			Lamia (54)	Dalia (64)

Continuation of the exercise solution :

We thus created "classes". We write them in the form of an interval, for example the time interval of team 1 is :

$$[37,42[$$

[37

The time of 37 min is included in the interval

42[

The time of 42 min is not included in the interval

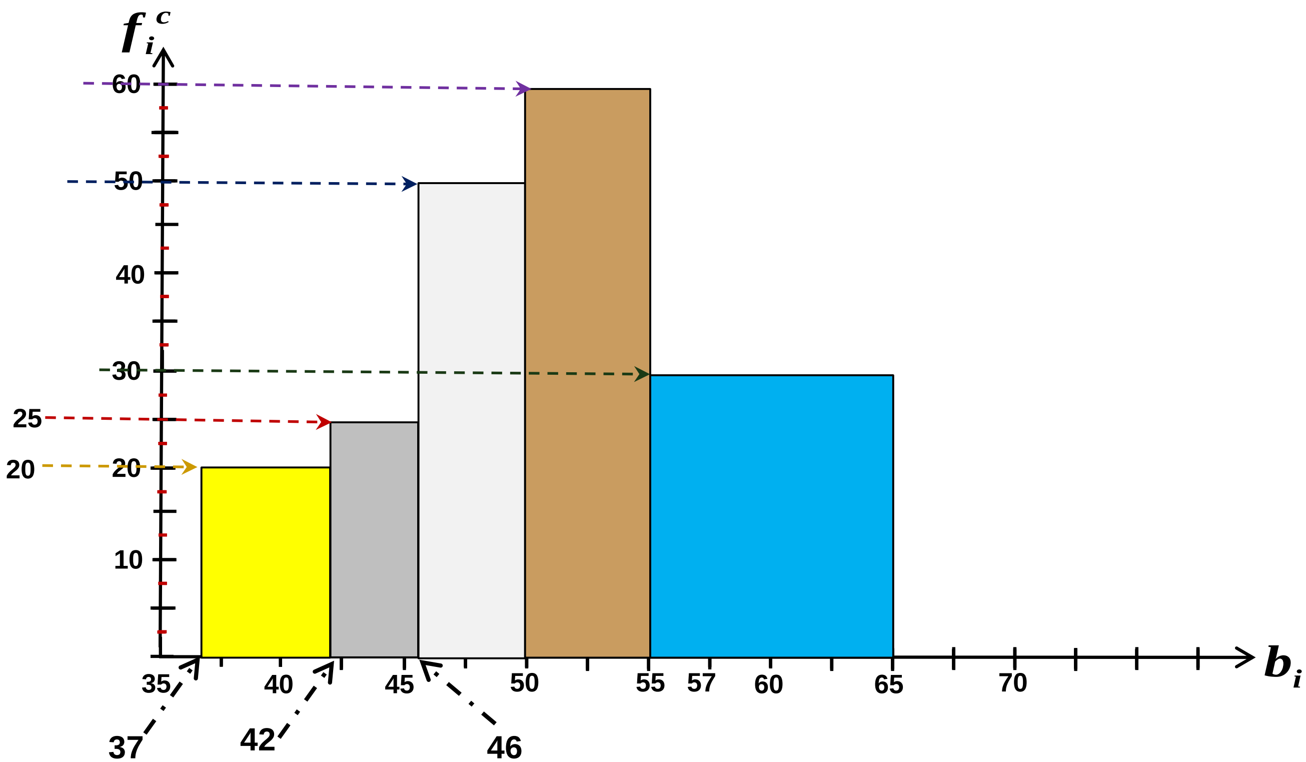
We can thus build a new table :

$[b_{i-1}, b_i[$	n_i
[37, 42[	3
[42 , 46[	3
[46 , 50[	6
[50 , 55[	9
[55, 65[	9

1. *2- Give a graphical representation of relative frequencies in the form of a histogram (the reference amplitude a* = 1000).*
   We will calculate the different amplitudes :

$[b_{i-1}, b_i[$	n_i	$a_i = b_i - b_{i-1}$	$d_i = \frac{n_i}{a_i}$	$n_i^c = d_i \times a^*$	f_i (%)	f_i^c
[37, 42[	3	5	0,6	600	0,10	20
[42 , 46[	3	4	0,75	750	0,10	25
[46, 50[	6	4	1,5	1500	0,20	50
[50 , 55[	9	5	1,8	1800	0,30	60
[55, 65[	9	10	0,9	900	0,30	30
Total	30				1,00	

So ;



Graphical representation of relative frequencies in the form of a histogram

3- *Draw the relative frequency polygon.*

***Definition of the relative frequency Polygon*:**

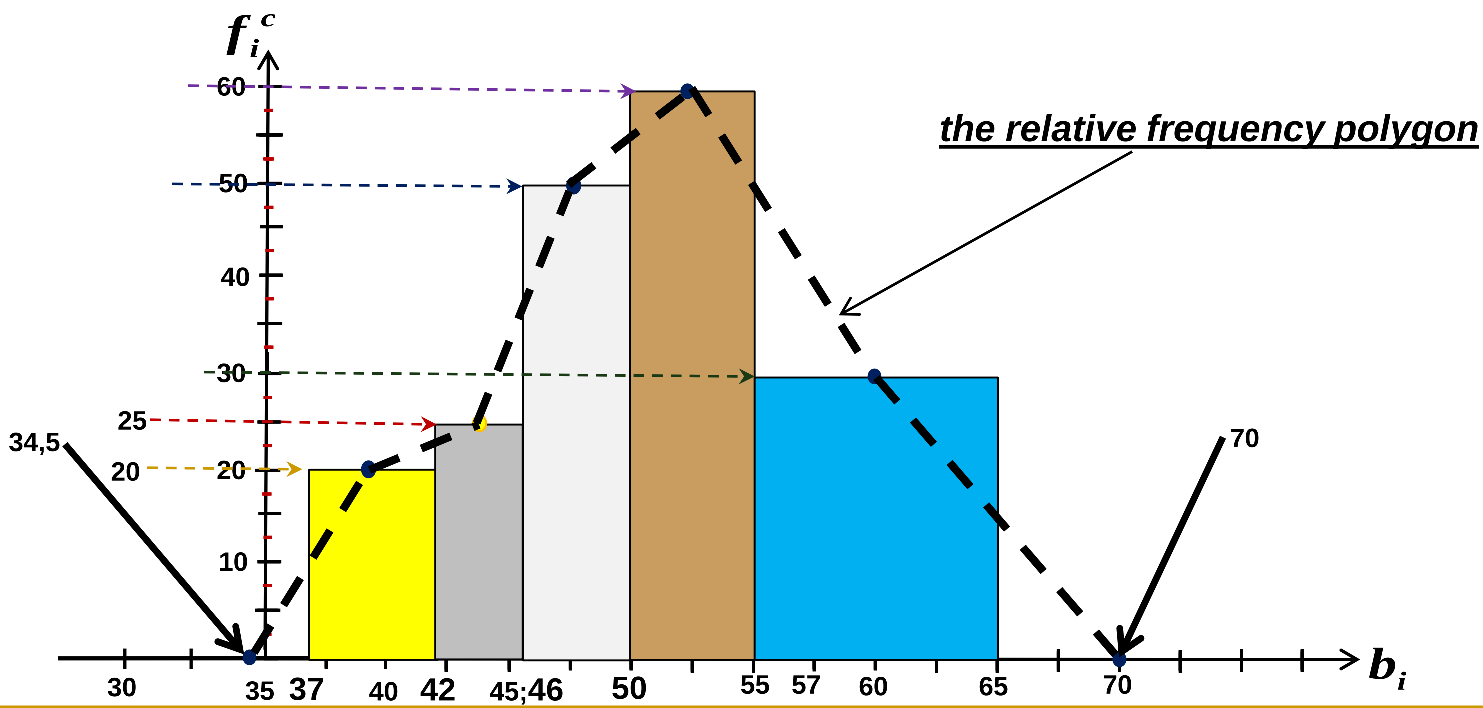
*It is **a broken line** connecting:*

***1- the midpoints of the vertices of the** rectangles of the histogram.*

***2- Closure is done by two points on the abscissa** axis located respectively **at half an interval of** :*

- The lower boundary of the first class

***- and the upper bound of the last class.** That's to say :*

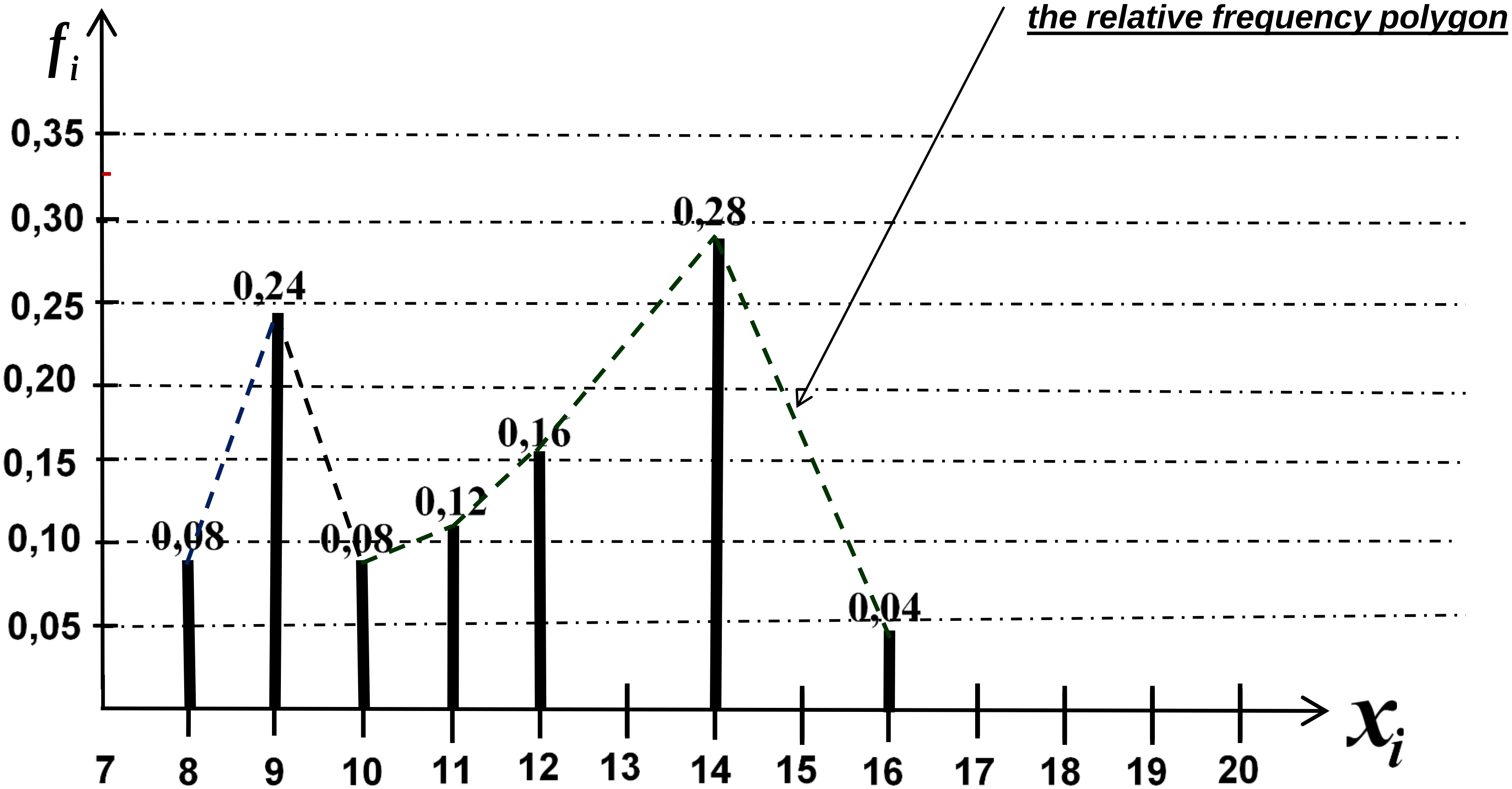


***Remarks* :**

1- The frequency polygon *is defined in the same way by associating it with a histogram of frequencies.*

2- For the case of a discrete statistical variable The frequency polygon (The relative frequency polygon respectively) is drawn on the lines diagram of the *frequencies* ***(The relative frequency respectively) by connecting the vertices of the lines .***

Example :



the relative frequency polygon
f_i
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,24
0,28
0,16
0,12
0,08
0,08
0,04
7
8
9
10
11
12
13
14
15
16
17
18
19
20
x_i

2- Cumulative Distribution Function for the case of a continuous statistical variable :

In this case we will just give the technique for obtaining the curve of the distribution function which is called ***cumulative chart*** *.*

In effect: $F : \mathbb{R} \to [0\ ,\ 1]$

$$x \to F(x) = P(X \leq x)$$

and its diagram (cumulative chart)*, is a broken line obtained by connecting*

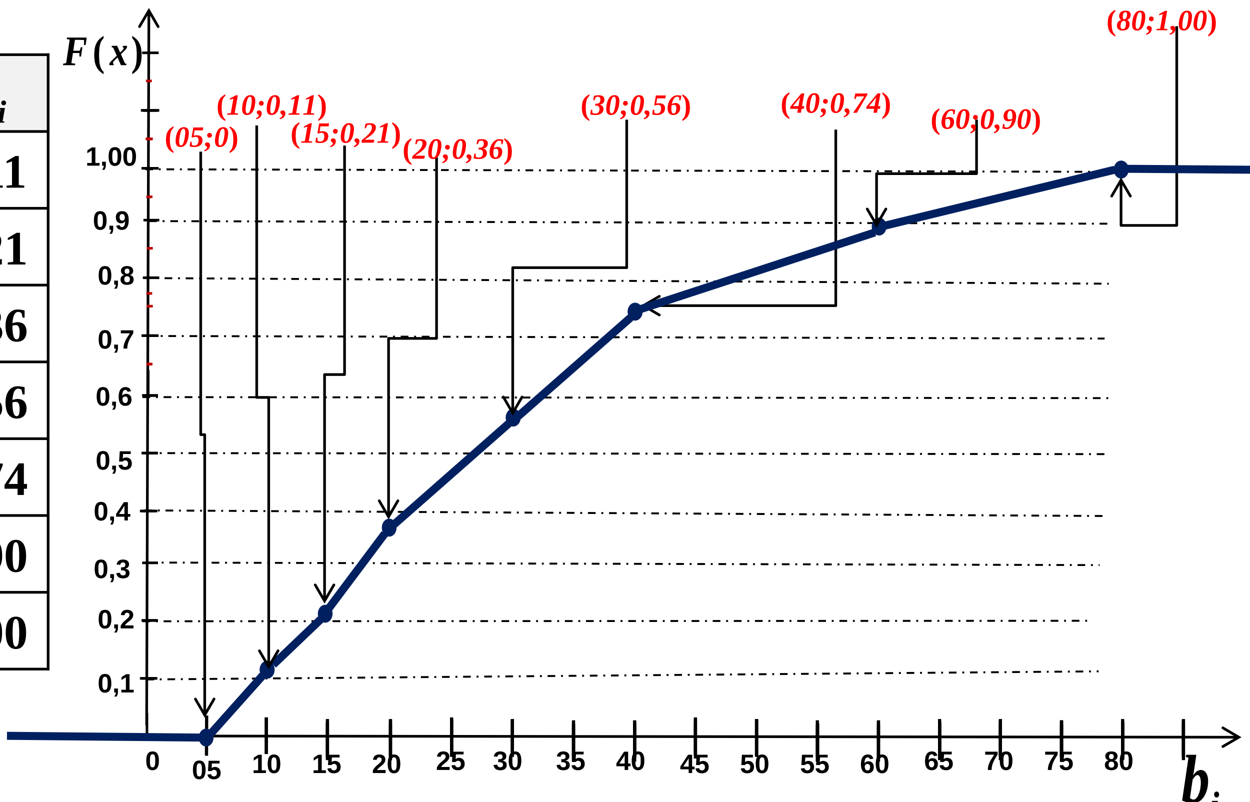
- *The different coordinate points* $(b_i\ ,F_i)$ *in increasing order with* $F_0 = 0.$
- *And by joining the left side of the point* $(b_0\ ,\ F_0)$ *the ½ line* $y = 0$ *and on the right side of the point* $(b_r\ ,\ F_r)$ *the ½ line* $y = 1.$

Example :

We take the example from ***page 32.***

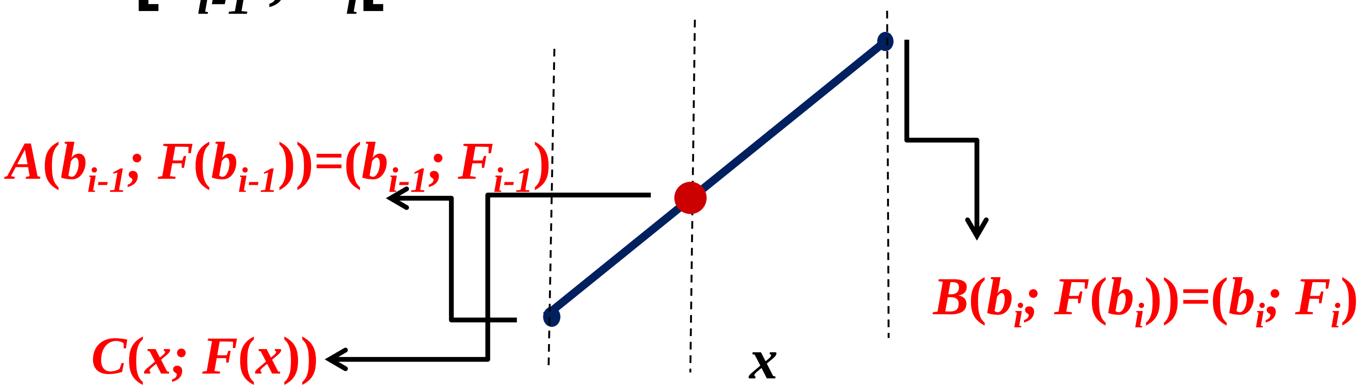
Thus the curve of the distribution function, called ***cumulative curve****, is drawn as follows:*

$[b_{i-1}, b_i[$	n_i	N_i	f_i	F_i
[5 , 10[	11	11	0,11	0,11
[10 , 15[	10	21	0,10	0,21
[15, 20[	15	36	0,15	0,36
[20, 30[	20	56	0,20	0,56
[30 , 40[	18	74	0,18	0,74
[40 , 60[	16	90	0,16	0,90
[60, 80[	10	100	0,10	1,00



Technique for calculating a value of $F(x)$ for $x \in \mathbb{R}$:

To calculate $F(x),\ \forall x \in \mathbb{R}$ *we will carry out* ***a linear interpolation*** *between the points* $A(b_{i-1};\ F(b_{i-1})) = (b_{i-1};\ F_{i-1})$ *et* $B(b_i;\ F(b_i)) = (b_i;\ F_i)$ *such as* $x \in [b_{i-1}\ ,\ b_i[$



The straight line equation (AB) *is of the form* $y = mx+p$ *such as :*

$$m = \frac{y_B - y_A}{x_B - x_A} = \frac{y_C - y_A}{x_C - x_A} = \frac{F(b_i) - F(b_{i-1})}{b_i - b_{i-1}} = \frac{F_i - F_{i-1}}{b_i - b_{i-1}} = \frac{F(x) - F_{i-1}}{x - b_{i-1}}$$

$$\Rightarrow F(x) = F_{i-1} + m(x - b_{i-1}) = F_{i-1} + \frac{f_i}{b_i - b_{i-1}}(x - b_{i-1})$$

Remark :

Otherwise we can calculate $\boldsymbol{x}$ *if we have the value of* $\boldsymbol{F(x)}$, *by using* ***a linear interpolation*** *between points* $A(b_{i-1};\ F(b_{i-1})) = (b_{i-1};\ F_{i-1})$ *and* $B(b_i;\ F(b_i)) = (b_i;\ F_i)$ *such as* $x \in [b_{i-1}, b_i[$ *and* $(\ F(x) \in [F_{i-1}, F_i[\ )$, *Indeed :*

$$\frac{F(b_i)-F(b_{i-1})}{b_i-b_{i-1}} = \frac{F_i-F_{i-1}}{b_i-b_{i-1}} = \frac{F(x)-F_{i-1}}{x-b_{i-1}}$$

$$\Rightarrow x = (F(x)-F_{i-1})\left(\frac{b_i-b_{i-1}}{F_i-F_{i-1}}\right)+b_{i-1} \Rightarrow x = b_{i-1}+a_i\left(\frac{F(x)-F_{i-1}}{F_i-F_{i-1}}\right)$$

Final formula

1.3- Measures of the statistical series

1.3.1- Measures of position:

1- The Arithmetic Mean : *Notée* $\bar{x}$ *est égale ;*

1st – *The case of an ungrouped statistical series; ie we have* N *observations :* $x_1, x_2, \ldots, x_N$ *, then the mean is given byr :*

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example 1 : *We consider the scores obtained in statistics by a group of students :* **14, 16, 12, 9, 11, 16, 7, 9, 7, 9.**

The mean of these scores is :

$$\bar{x} = \frac{14+16+12+9+11+16+7+9+7+9}{10} = 11$$

2eme – *Case of a grouped statistical series; the mean :*

A- *In the case of a discrete variable :*

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_r x_r}{N} = \sum_{i=1}^{i=r} \frac{n_i x_i}{N} = \sum_{i=1}^{i=r} f_i x_i$$

B- *In the case of a continuous variable :*

$$\bar{x} = \frac{n_1 c_1 + n_2 c_2 + \cdots + n_r c_r}{N} = \sum_{i=1}^{i=r} \frac{n_i c_i}{N} = \sum_{i=1}^{i=r} f_i c_i$$

With $c_i = \dfrac{b_{i-1} + b_i}{2}$, *the* ***Center*** *of the class* $[b_{i-1}, b_i[$.

Example 2 :

Classes	Centers c_i	n_i	f_i	$n_i c_i$
[20-40[	*30*	15	*0,15*	*450*
[40-60[	*50*	20	*0,20*	*1000*
[60-100[	*80*	20	*0,20*	*1600*
[100-200[	*150*	45	*0,45*	*6750*
Total		*100*	*1,00*	*9800*

$$\Rightarrow \bar{x} = \sum_{i=1}^{i=r} \frac{n_i c_i}{N} = \frac{9800}{100} = 98 \Rightarrow \bar{x} \in [60 - 100[$$

2- The Mode : *is that value of the variable which occurs with the greatest frequency in a data set (or the greatest relative frequency). We note the mode* M_o.

A. Case of a discrete variable :

Example : *We consider the scores obtained in statistics by a group of 20 students :*

7, 13, 5, 15, 12, 9, 7, 8, 14, 16, 13, 6, 13, 10, 13, 12, 10, 7,12, 13.

$\Rightarrow$

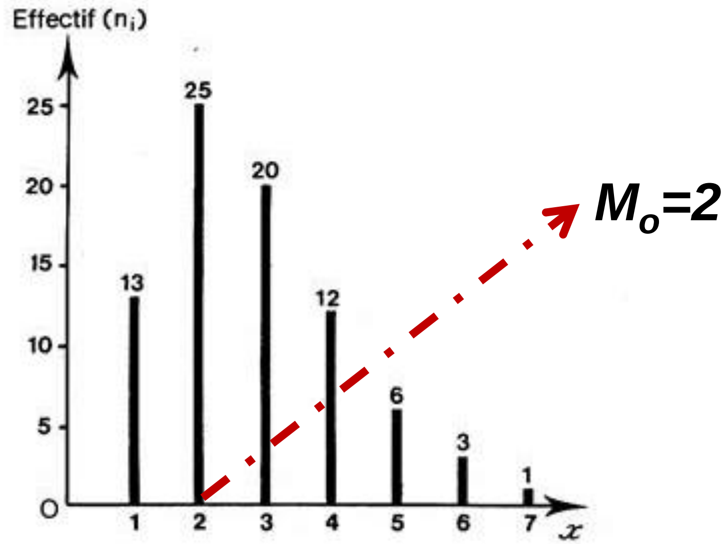
x_i	5	6	7	8	9	10	12	13	14	15	16
n_i	1	1	3	1	1	2	3	5	1	1	1

The mode of this series is $M_o = 13$*, value that appears* ***five times****.*

The interpretation *: The most common score is 13.*

Remark : *Graphically, in a lines chart the mode corresponds to* ***the abscissa of the highest lines****.*

That's to say :



B. <u>Case of a continuous variable</u> *: In this case, we are rather talking about* ***The modal class.*** *We have two cases.*

<u>B.1- Cases of identical amplitudes</u> *:(ie $a_i = a_j \; \forall \; i \neq j$)* ***The modal class*** *is the class of the greatest frequency n_i, either $[b_{i-1}, b_i[$, with the Mode $M_o \in [b_{i-1}, b_i[$ then :*

$$M_O = b_{i-1} + a_i \left(\frac{m_1}{m_1 + m_2} \right)$$

With;

- b_{i-1} *: lower bound of modal class.*
- b_i *: upper bound of modal class.*
- a_i *: amplitude of the modal class.*
- $m_1 = n_i - n_{i-1}$ *et* $m_2 = n_i - n_{i+1}$.

Example : *Let the distribution of the population of 20 households according to the income (in hundreds of DA) of both parents;*

Classes en HDA	*Amp a_i*	n_i	f_i
[200-300[	100	40	*0,20*
[300-400[	100	60	*0,30*
[400-500[	100	30	*0,15*
[500-600[	100	50	*0,25*
[600-700[	100	20	*0,10*
Total		200	*1,00*

The modal class is [*300 - 400*[. *The mode is calculated by* :

$$M_O = b_{i-1} + a_i\left(\frac{m_1}{m_1 + m_2}\right) = 300 + 100\left(\frac{60-40}{(60-40)+(60-30)}\right)$$

$$\Rightarrow M_O = 340 \quad HDA$$

Interpretation : The most common salary is said to be 340 HDA.

B.2- Case of uneven amplitudes :(ie $a_i \neq a_j$) ***The modal class*** *is the class of the greatest corrected frequency* n_i^c *(or the highest corrected relative frequency* f_i^c*), and the mode* M_o *is such that:*

$$M_O = b_{i-1} + a_i\left(\frac{m_1}{m_1+m_2}\right) \quad \text{With;} \quad \begin{cases} b_{i-1} : \text{lower bound of modal class.} \\ b_i : \text{upper bound of modal class.} \\ a_i : \text{amplitude of the modal class.} \\ m_1 = h_i - h_{i-1} = n_i^c - n_{i-1}^c \\ m_2 = h_i - h_{i+1} = n_i^c - n_{i+1}^c \end{cases}$$

Where h_i , h_{i-1} *and* h_{i+1} *are the corrected frequencies.*

Remark 1: *In both cases we can compute the mode* M_o *by using the relative frequencies instead of frequencies, by taking ;*

1- $m_1 = f_i - f_{i-1}$ *et* $m_2 = f_i - f_{i+1}$ *si* $(a_i = a_j \, \forall \, i \neq j)$

2- $m_1 = f_i^c - f_{i-1}^c$ **et** $m_2 = f_i^c - f_{i+1}^c$ *si* $(a_i \neq a_j)$

Example : *Let the distribution of 100 individuals according to their age; $a^* = 100$*

Classes	a_i	n_i	d_i	n_i^c
[20 , 30[	**10**	**20**	***2,00***	***200***
[30 , 40[	**10**	**25**	***2,50***	***250***
[40 , 60[	**20**	**35**	***1,75***	***175***
[60 , 80[	**20**	**20**	***1,00***	***100***

The modale class is [30 - 40[, and the mode is :

$$M_O = b_{i-1} + a_i\left(\frac{m_1}{m_1 + m_2}\right) = 30 + 10\left(\frac{250 - 200}{(250 - 200) + (250 - 175)}\right)$$

$$\Rightarrow M_O = 34$$

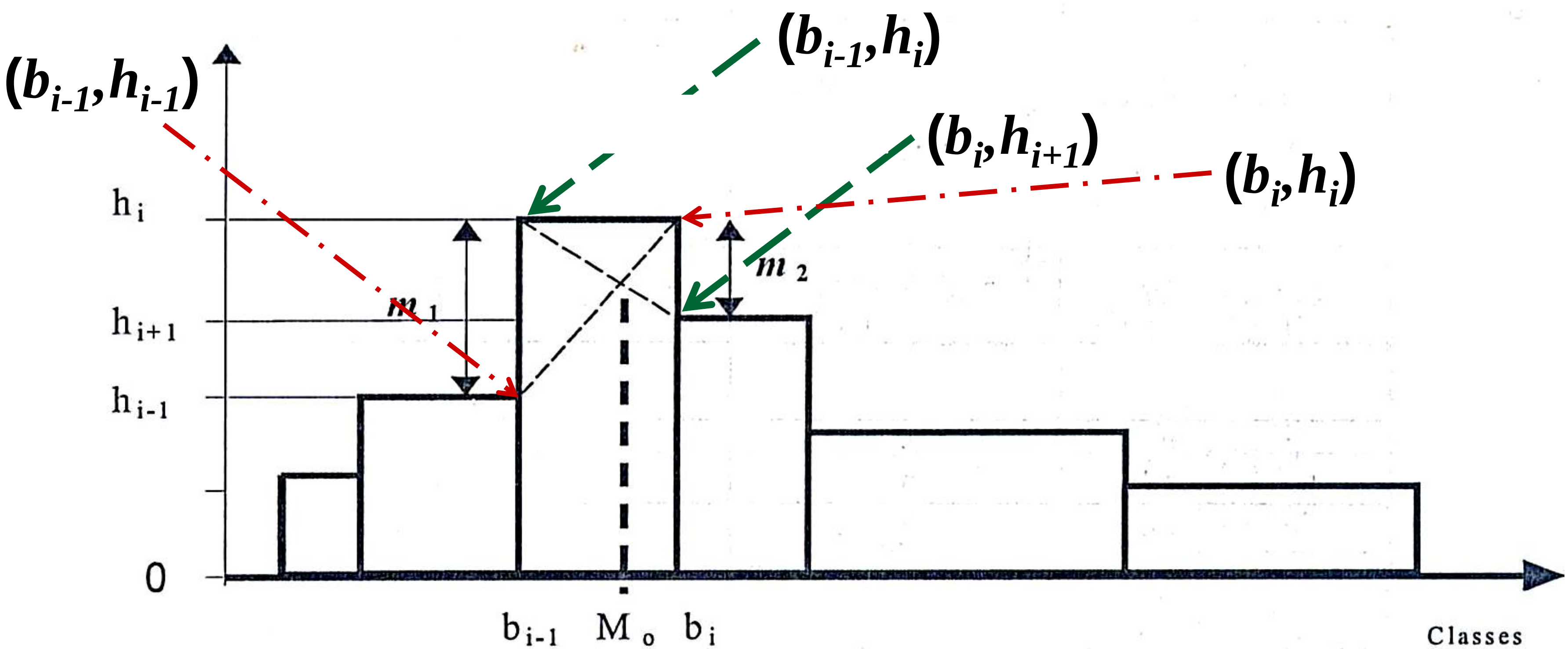
Interpretation: The most common age is 34 years.

Calculation of Mode by Graphical Method of a continuous variable :

Ifi ***the modale classe*** *is* $[b_{i-1}, b_i[$ *, with the Mode* $M_o \in [b_{i-1}, b_i[$ *then :*

$$M_o = b_{i-1} + a_i\left(\frac{m_1}{m_1 + m_2}\right)$$

And graphically the Mode $M_o \in [b_{i-1}, b_i[$ *the Mode on a histogram is the point where the two segments intersect* $[(b_{i-1}, h_i);(b_i, h_{i+1})]$ *and* $[(b_{i-1}, h_{i-1}); (b_i, h_i)]$*, see following figure :*



3- The Median : *For a statistical series arranged in ascending order the median* Me *is the value of the variable which divides the population into two groups of equal frequencies.*

A. Case of a discrete variable *: For a statistical series arranged in ascending order , that's to say :*

$v_1 \leq v_2 \leq \cdots \leq v_N$ *the median* Me *is the value of the middle which will depend on the total frequency* N*.*

1- If N *is odd* ($N = 2k+1$)*, then* $Me = v_{k+1}$.

2- If N *is even* ($N = 2k$)*, then* $Me = \dfrac{v_k + v_{k+1}}{2}$.

Example 1 *: Let the distribution of 9 households according to the number of children ;*

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

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

$\Rightarrow$

Number of children per household v_i	0	0	1	1	2	3	3	3	4
(ascending order) of individuals	1	2	3	4	5	6	7	8	9
	4 observation				v_5	4 observation			

We have $N = 9 = 2\times4+1 \;\Rightarrow\; Me = v_{k+1} = v_5 = 2.$

Example 2 : Let the distribution of 10 households according to the number of children ;

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

$\Rightarrow$

Number of children per household v_i	0	0	1	1	2	3	3	3	4	4
(ascending order) of individuals	1	2	3	4	5	6	7	8	9	10
	4 observation					v_5 4 v_6 observation				

We have $N = 10 = 2\times5$ (even) $\Rightarrow \; Me = \dfrac{v_k + v_{k+1}}{2} = \dfrac{2+3}{2} = 2{,}5.$

B. *Case of a continuous variable* *: We follow the following steps;*

1- Determining ***the median classe*** $[b_{i-1}, b_i[$ *, By looking for the class that contains the individual of order* $k+1$ $(resp\ k)$ *if* $N = 2k+1$ $(resp\ N = 2k)$.

2- By ***Linear interpolation****, we can calculate the median inside the median class which is given by :*

$$Me = b_{i-1} + a_i \left(\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \right) \quad With; \left\{ \begin{array}{l} N_i : \text{increasing cumulative frequency of the median class,} \\ N_{i-1} : \text{increasing cumulative frequency of the class before the median class} \\ N : \text{the total frequency.} \end{array} \right.$$

Remark :

We can determine the median in the same way using increasing cumulative frequencies.

And we will have the formula :

$$Mé = b_{i-1} + a_i \left(\frac{0{,}5 - F_{i-1}}{F_i - F_{i-1}} \right)$$ *With;*

F_i **:** *increasing cumulative relative frequency of the median class,*

$F_{i\text{-}1}$ **:** *increasing cumulative relative frequency of the class before the median class .*

Example 3 *: We take the example from page 32.*

$[b_{i-1}, b_i[$	n_i	N_i	f_i	F_i
[5 , 10[	11	11	0,11	0,11
[10 , 15[	10	21	0,10	0,21
[15, 20[	15	36	0,15	0,36
[20, 30[	20	56	0,20	0,56
[30 , 40[	18	74	0,18	0,74
[40 , 60[	16	90	0,16	0,90
[60, 80[	10	100	0,10	1,00

We have $\frac{N}{2}=50 \Rightarrow$ *the median classe is* $[20, 30[$ *and et we will have :*

$$Me = b_{i-1} + a_i\left(\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}\right)$$

$$\Rightarrow Me = 20 + 10\left(\frac{50-36}{56-36}\right) = 20 + 10\left(\frac{14}{20}\right) \Rightarrow Mé = 27 \quad years$$

Remarque **:** *The median can be defined as the inverse of the cumulative distribution function for the value* $x = 0{,}5$; $Me = F^{-1}(0{,}5)$.

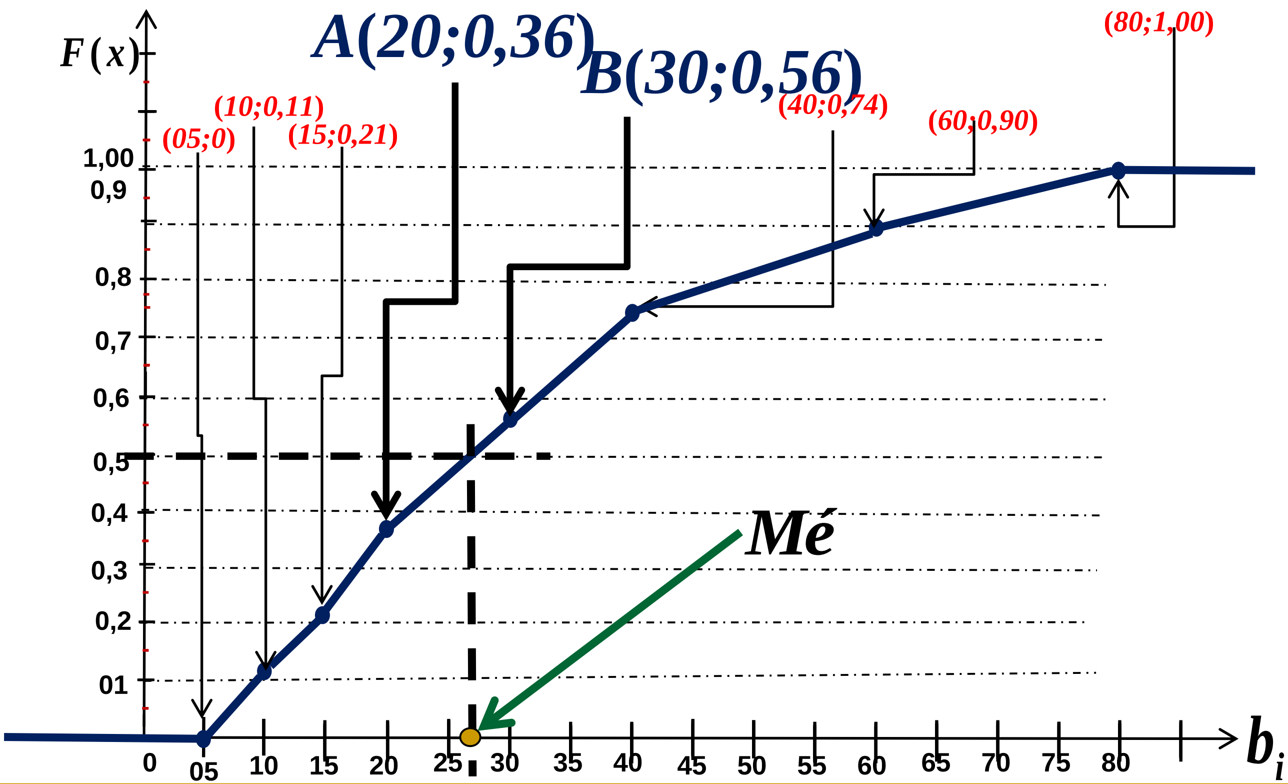
We say that of the median On dit que ***the order of the median*** *is* $p = F(Mé) = $ ***0,5****. And we can calculate the median graphically from* ***the cumulative curve.***

Example *: Using our example of the distribution of 100individuals according to their age.*

$[b_{i-1}, b_i[$	n_i	N_i	f_i	F_i
[5 , 10[	11	11	0,11	0,11
[10 , 15[	10	21	0,10	0,21
[15, 20[	15	36	0,15	0,36
[20, 30[	20	56	0,20	0,56
[30 , 40[	18	74	0,18	0,74
[40 , 60[	16	90	0,16	0,90
[60, 80[	10	100	0,10	1,00

- *the median classe is* **[20, 30[** *and the median* ***Me*** *is the abscissa of order*

$$F(Me) = 0{,}5 \left(Me = F^{-1}(0{,}5) \right)$$



So the equation of the straight line (**AB**) *is of the form* $\boldsymbol{y = mx+p}$ *such as :*

$$m = \frac{y_B - y_A}{x_B - x_A} = \frac{F_i - F_{i-1}}{b_i - b_{i-1}} = \frac{F(Me) - F_{i-1}}{Me - b_{i-1}}$$

$$\Rightarrow Me = b_{i-1} + a_i\left(\frac{F(Me) - F_{i-1}}{F_i - F_{i-1}}\right) = 20 + 10\left(\frac{0{,}50 - 0{,}36}{0{,}56 - 0{,}36}\right)$$

$$\Rightarrow Me = 27 \quad years$$

3- The Quartiles: *The Quartiles* Q_1, Q_2, Q_3 *are measures that divide the frequency distribution in to four equal parts :*

25 % *the values* ≤ Q_1, ***25 %*** *between* Q_1 *and* Q_2 ; ***25 %*** *between* Q_2 *and* Q_3 , *and* ***25 %*** *greater than* Q_3.

Remark : Q_1, Q_2 , Q_3 *are respectively the abscissa of the ordinate points 0,25 ; 0,5 ; 0,75 on the increasing cumulative curve.* Q_2 *is equal to the median.*

That's to say :

- *The order of* Q_1 *is* $p = F(Q_1) = 0{,}25$.
- *The order of* Q_2 *is* $p = F(Q_2) = F(Me) = 0{,}50$.
- *The order of* Q_3 *is* $p = F(Q_3) = 0{,}75$.

4 - Calculation of Quartiles :

4.1 - Discreet case : *We have* p *the order of quartile* Q_i *, with* $i = 1, 2, 3$ *then :*

- *If* $(N \times p)$ *is an integer number, then* $Q_i = \dfrac{v_{(N\times p)} + v_{(N\times p)+1}}{2}$.

- *If* $(N \times p)$ *is not an integer number, then* $Q_i = v_{\lceil N \times p \rceil}$

where $\lceil N \times p \rceil$ *represents the smallest integer number greater than or equal to* $N \times p$ **(** *which is called integer part with excess* ***integer part with excess*** **)**.

Example 1 *: Let the distribution of 12 households according to the number of children ;*

x_i	0	1	2	3	4
n_i	2	2	1	5	2

- *The first quartile Q_1* *: As $(N \times p) = 12 \times 0{,}25 = 3$ is an integer number, we have :*

$$Q_1 = \frac{v_{(N\times p)} + v_{(N\times p)+1}}{2} = \frac{v_3 + v_4}{2} = \frac{1+1}{2} \Rightarrow Q_1 = 1$$

- *The Second Quartile $Me = Q_2$* *: As $(N \times p) = 12 \times 0{,}50 = 6$ is an integer number, we have :*

$$Q_2 = Me = \frac{v_{(N\times p)} + v_{(N\times p)+1}}{2} = \frac{v_6 + v_7}{2} = \frac{3+3}{2} \Rightarrow Q_2 = 3$$

- *The third quartile Q_3* *: As $(N \times p) = 12 \times 0{,}75 = 9$ is an integer number, we have :*

$$Q_3 = \frac{v_{(N\times p)} + v_{(N\times p)+1}}{2} = \frac{v_9 + v_{10}}{2} = \frac{3+3}{2} \Rightarrow Q_3 = 3$$

Example 2 : *Let the distribution of 9 households according to the number of children :*

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

- The first quartile Q_1 : *As* $(N \times p) = 9\times0{,}25 = 2{,}25$ *is not an integer number, we have :* $Q_1 = v_{[2{,}25]} = v_3 = 1$.

- The Second Quartile $Me = Q_2$: *As* $(N \times p) = 9\times0{,}50 = 4{,}50$ *is not an integer number, we have :* $Q_2 = v_{[4{,}50]} = v_5 = 2$

- The third quartile Q_3: *As* $(N \times p) = 9\times0{,}75 = = 6{,}75$ *is not an integer number, we have :* $Q_3 = v_{[6{,}75]} = v_7 = 3$.

4.2 - The continue case :

For the calculation of Q_1, Q_2, Q_3 *: We follow the following steps;*

1- Determination the class $[b_{i-1}, b_i[$ *of* $Q_1 \in [b_{i-1}, b_i]$ *, By searching for the class that contains the order individual* $\lceil N \times p \rceil = \lceil N/4 \rceil$ *.*

2- If N_i *: increasing cumulative frequency of the class of* Q_1 *,*

N_{i-1} *: increasing cumulative frequency of the class before the class of* Q_1 *and* N *: the total frequency.*

F_i *: increasing relative cumulative frequency of the class of* Q_1 *,*

F_{i-1} *: increasing cumulative frequency of the class before the class of* Q_1*. the we have :*

$$Q_1 = b_{i-1} + a_i\left(\frac{N/4 - N_{i-1}}{N_i - N_{i-1}}\right) = b_{i-1} + a_i\left(\frac{0{,}25 - F_{i-1}}{F_i - F_{i-1}}\right)$$

Nb : *The calculation of* Q_2 *et* Q_3 *is done in the same way such that;*

$$Q_2 = Mé = b_{i-1} + a_i\left(\frac{N/2 - N_{i-1}}{N_i - N_{i-1}}\right) = b_{i-1} + a_i\left(\frac{0,5 - F_{i-1}}{F_i - F_{i-1}}\right)$$

$$Q_3 = b_{i-1} + a_i\left(\frac{N(3/4) - N_{i-1}}{N_i - N_{i-1}}\right) = b_{i-1} + a_i\left(\frac{0,75 - F_{i-1}}{F_i - F_{i-1}}\right)$$

1.3.2- Measures of Dispersion :

Remark 1 *: The quartiles already seen as measures of position can be considered as measures of dispersion.*

1- The range *: The range noted* E *(or* R*) is simply the difference between the largest and smallest observed value.*

$$E = x_{\max} - x_{\min}$$

2- The interquartile range *: This is the difference between the first and last quartiles. That's to say;*

$$IQ = Q_3 - Q_1$$

Remark 2 *: The interquartile range measures the range of the middle 50% of values in a classified data series.*

Example : We take the distribution of the 100 individuals according to their ages.

$[b_{i-1}, b_i[$	n_i	N_i	f_i	F_i
[5 , 10[	11	11	0,11	0,11
[10 , 15[	10	21	0,10	0,21
[15, 20[	15	36	0,15	0,36
[20, 30[	20	56	0,20	0,56
[30 , 40[	18	74	0,18	0,74
[40 , 60[	16	90	0,16	0,90
[60, 80[	10	100	0,10	1,00

Let's calculate the quartiles Q_1 , Q_3 and the interquartile range. On a :

$\left\lceil \frac{N}{4} \right\rceil = 25, \left\lceil \frac{3N}{4} \right\rceil = 75$, so the class of Q_1 is **[15, 20[** , that of Q_3 is **[40, 60[** :

$$\Rightarrow Q_1 = b_{i-1} + a_i \left(\frac{0{,}25 - F_{i-1}}{F_i - F_{i-1}} \right) \Rightarrow Q_1 = 15 + 5 \left(\frac{0{,}25 - 0{,}21}{0{,}36 - 0{,}21} \right) = 16{,}33 \text{ years}$$

- Which means that 25% of individuals are under the age of 16 years and 4 months. ($0{,}33 \times 12 = 3{,}96 \approx 4$). And for Q_3 we have :

- For Q_3 we have :

$$Q_3 = b_{i-1} + a_i\left(\frac{0{,}75 - F_{i-1}}{F_i - F_{i-1}}\right) \Rightarrow Q_3 = 40 + 20\left(\frac{0{,}75 - 0{,}74}{0{,}90 - 0{,}74}\right) = 41{,}25 \text{ years}$$

- Which means that 75% of individuals are under the age of 41 years and 3 months (0,25×12 = 3). So the interquartile range is:

$$\Rightarrow IQ = Q_3 - Q_1 = 24{,}92 \text{ years}$$

- Which means the age difference between Q1 and Q3 is 24 years, 11 months and 12 days (0,92×12 =11,04 and 0,4×30 =12).

Remark 3 **:**

If *$N \times p = N_i$, then the quartiles $x_p = b_i$ although $b_i \notin [b_{i-1} , b_i[$ and the class of the quartile is $[b_{i-1} , b_i[$* .

Example *: We take the distribution of the 100 individuals according to their ages.*

Classes	n_i	N_i	f_i	F_i
[20 , 30[	25	25	0,25	0,25
[30 , 40[	20	45	0,20	0,45
[40 , 60[	35	80	0,35	0,80
[60 , 80[	20	100	1,00	1,00

*1- Calculation of **1**st quartile Q_1 :*

*We have the order of **1**st quartile Q_1 is $p = 0,25$.*

As $\lceil N \times p \rceil = \lceil 100 \times 0,25 \rceil = \lceil 25 \rceil = 25$, and $N_1 = 25$, then :

The class of Q_1 is $[20, 30[$, that's to say $Q_1 \in [20, 30]$. Therefore :

$$Q_1 = b_0 + a_1\left(\frac{N(1/4) - N_0}{N_1 - N_0}\right) = 20 + 10\left(\frac{25-0}{25-0}\right) = 30 \in [20,30]$$

Remark 4 **:**

Approximate quartile values can be obtained graphically from ***the cumulative curve.***

3- The Variance : *The variance of a variable* X *noted* $V(x)$ *is the sum of the squares of the deviations from the mean divided by the number of observations (Total frequency* N*).*

A- *In the case* ***of a discrete variable*** *:*

$$V(x) = \frac{1}{N}\sum_{i=1}^{i=r} n_i\left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{i=r} f_i\left(x_i - \overline{x}\right)^2$$

B- *In the case* ***of a continuous variable*** *:*

$$V(x) = \frac{1}{N}\sum_{i=1}^{i=r} n_i\left(c_i - \overline{x}\right)^2 = \sum_{i=1}^{i=r} f_i\left(c_i - \overline{x}\right)^2$$

With $c_i = \dfrac{b_{i-1} + b_i}{2}$, *le* ***centre*** *de la classe* $[b_{i-1}, b_i[$.

Remark 5 **:** *The variance can be written in another form called a « developed formula » :*

- The developed formula for the variance is

1st **–** *Case of a non-grouped statistical series; i.e. we have N observations:*

$$V(x)=\left(\frac{1}{N}\sum_{i=1}^{i=r}x_i^{\,2}\right)-\bar{x}^2$$

A- *In the case* ***of a discrete variable*** *:*

$$V(x)=\left(\frac{1}{N}\sum_{i=1}^{i=r}n_i x_i^{\,2}\right)-\bar{x}^2=\left(\sum_{i=1}^{i=r}f_i x_i^{\,2}\right)-\bar{x}^2\cdots\cdots(*)$$

B- *In the case* ***of a continuous variable*** *:*

$$V(x)=\left(\frac{1}{N}\sum_{i=1}^{i=r}n_i c_i^{\,2}\right)-\bar{x}^2=\left(\sum_{i=1}^{i=r}f_i c_i^{\,2}\right)-\bar{x}^2$$

Proof of the developed formula : *We have;*

$$V(x)=\frac{1}{N}\sum_{i=1}^{i=r}n_i(x_i-\bar{x})^2=\frac{1}{N}\sum_{i=1}^{i=r}n_i\left(x_i^{\,2}-2x_i\bar{x}+\bar{x}^2\right)$$

$$\Rightarrow V(x)=\frac{1}{N}\sum_{i=1}^{i=r}n_i x_i^{\,2}-2\frac{\bar{x}}{N}\sum_{i=1}^{i=r}n_i x_i+\frac{\bar{x}^2}{N}\sum_{i=1}^{i=r}n_i$$

$$\Rightarrow V(x)=\left(\frac{1}{N}\sum_{i=1}^{i=r} n_i x_i^{\,2}\right)-2\bar{x}.\bar{x}+\bar{x}^2 \quad \Rightarrow \quad (*)$$

Remark 5 *: This developed formula for variance is easier to remember and faster to calculate.*

Remark 6 *: The variance is expressed in the square of the unit of the variable. For example, the variance of the age variable is expressed in «squared years (years²)» because;*

$$V(x)=\left(\frac{1}{N}\sum_{i=1}^{i=r} n_i x_i^{\,2}\right)-\bar{x}^2$$

4- The standard deviation : *We call standard deviation which we denote by* $\sigma(x)$ *, the square root of the variance:* $\sigma(x)=\sqrt{V(x)}$

Remark 7 :

i) The standard deviation is expressed in the same unit of measurement as the variable.

ii) It is used as an indicator of the dispersion of the statistical series, so that in an increasing order the mean $\bar{x}$ *divides the population into two parts such that the individuals having the value of the variable less than* $\bar{x}$ *will have approximately ,* $\bar{x}-\sigma(x)$ *the others* $\left(X>\bar{x}\right)$ *will have* $\bar{x}+\sigma(x).$

iii) More the larger the standard deviation, the dispersion of bservations around the mean of the variable is strong.

iv) A distribution will have a standard deviation near 0 if these values are collected around the mean.

Exemple 1 : *Consider the following statistics scores of a group of 20 students*:

x_i	n_i	$n_i.x_i$	$n_i.x_i^2$
2	2	4	8
3	2	6	18
7	4	28	196
8	2	16	128
12	3	36	432
17	2	34	578
18	5	90	1620
Total	20	214	2980

Then : $\bar{x} = \sum_{i=1}^{i=r} \frac{n_i x_i}{N} = \frac{214}{20} = 10{,}7$ *and* $V(x) = \left(\frac{1}{N} \sum_{i=1}^{i=r} n_i x_i^{\,2} \right) - \bar{x}^2$

$$\Rightarrow V(x) = \frac{2980}{20} - (10{,}7)^2 \Rightarrow V(x) = 149 - 114{,}49 = 34{,}51 \Rightarrow \sigma(x) = 5{,}87$$

So, some students (the good ones) will have approximately the mean score ***(10,7) plus (+) 5,87 (=16,57)*** *the others (the bad ones) will have the mean score* ***(10,7) less (-) 5,87 (= 4,83)***.