Descriptive Statistics

Antoine Ayache [*]

University of Lille

Contents

1 A brief introduction

The main goal of Descriptive Statistics is to describe observed data in synthetic ways which make their analyses to be more convenient. The word "statistics" comes from the latin word "statisticum", that is "of the state". It seems that this word was used for the first time in the 17-th century when Colbert was the French minister of finances for the King Louis XIV. Indeed, it was used at this period by the intendant of Burgundy Claude Bouchu in an administrative document entitled "Declaration of assets, charges, debts and statistics of communities of generals of Burgundy from 1666 to 1669".

However the need for countries of having "statistics", that is numerical data (describing their populations, their ressources, their military strength, and so on), had appeared (for instance in pharaonic Egypt) several millennia before the 17-th century.

[*] The French version of this course has been done in collaboration with Julien Hamonier.

2 Univariate descriptive analysis

2.1 Terminology

1. A **population** is a set of homogeneous elements. For instance a group of students, French taxpayers, the households in Lille, and so on.

2. The elements of a population are called **the individuals** or **the statistical units**.

3. **Observations** concerning some topic have been made on these individuals. The series of these observations is called **a statistical variable**. For instance, "the students grades (or scores) for the statistics examination", "their honors (mentions) for the bachelor degree", "their sexes", "the colors of their eyes", "the turnovers of some entreprises", "the number of children per household", and so on.

4. A statistical variable is said to be:

   (i) **Quantitative**: when it consists in numbers ("the students grades for the statistics examination", "the turnovers of some entreprises", "the number of children per household", etc.). There are 2 types of quantitative variables: the **discrete** quantitative variables and the **continuous** quantitative variables. A discrete (or discontinuous) variable can only take isolated values. For example, "the number of children in a household" can only be 0, or 1, or 2, or 3, . . . ; it can never be a value strictly between 0 and 1, or 1 and 2, or 2 and 3, ... This is also the case of "the grade of a student in the statistics examination" (one even assumes that the grades are integer numbers). A continuous quantitative variable can take any value in an interval. For example, "the turnover of a small entreprise" can be 29,1 thousand euros, 29,12 thousand euros, 29,127 thousand euros, . . . , even if in practice it has to be rounded. Two very classical examples of continuous quantitative variables are "temperature" and "duration".

   (ii) **Qualitative**: when it is not numerical; it describes individuals that fit into categories. For examples, the two categories of the variable "Sexes" are: Male (H) and Female (F); the four categories of the variable "colors of eyes" are Blue (B), Brown (M), Black (N), and Green (V); the four categories of the variable "honors for the bachelor degree" are: Very Good (TB), Good (B), Fairly Good (AB), and Acceptable (P). They are two types of qualitative variables: the **ordinal** qualitative variables and the **nominal** qualitative variables. More precisely, a qualitative variable is said to be ordinal when its categories **can** be ordered in a natural way (this is for instance the case for the variable "honors for the bachelor degree"); a qualitative variable is said to be nominal when its categories **can not** be ordered in a natural way (this is for instance the case of the variable "colors of eyes").

2.2 Graphical representations of a variable

Data table

The population corresponding to this data table is a group of 15 students.

Individuals	Colors of eyes	Sexes	Honors for the bachelor degree	Grades in the statistics examination
Michel	V	H	P	12
Jean	B	H	AB	8
Stéphane	N	H	P	13
Charles	M	H	P	11
Agnès	B	F	AB	10
Nadine	V	F	P	9
Étienne	N	H	B	16
Gilles	M	H	AB	14
Aurélie	B	F	P	11
Stéphanie	V	F	B	15
Marie-Claude	N	F	P	4
Anne	B	F	TB	18
Christophe	V	H	AB	12
Pierre	N	H	P	6
Bernadette	M	F	P	2

2.2.1 Qualitative variables (ordinal and nominal)

One can represent the variables "Colors of eyes", "Sexes", and "Honors for the bachelor degree" by **bar charts**. It is worth mentioning that every individual belongs to a unique category of each of the three variables. Indeed, there is no individual whose eyes have several different colors (one excludes heterochromia). There is no individual who is both male and female (one excludes hermaphroditism). There is no individual having at the same time several different honors for his bachelor degree.

Remark 2.1. *Generally speaking an individual always belongs to* a unique *category of a qualitative variable. One mentions that very often, one of the category of a qualitative variable is called* ***Other*** *(non-respondents, missing values or something like that); one can put in the latter category the individuals which fail to fall in any other category of the variable.*

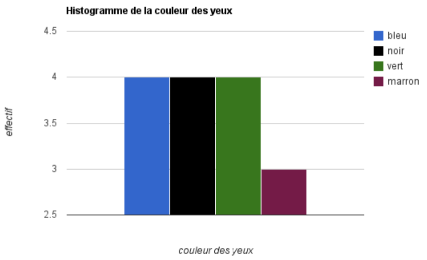
Let us now study the example of the variable "Colors of eyes". First, one counts the number of individuals belonging to each category of this variable: $n_B = 4$ individuals have blue eyes, $n_M = 3$ have brown eyes, $n_N = 4$ have black eyes, and $n_V = 4$ have green eyes. All of this is summarized in the following table:

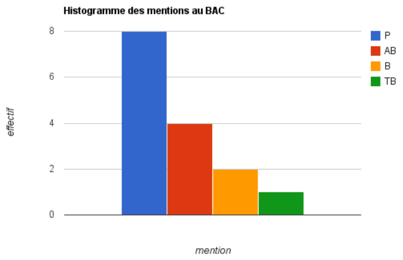
Color	Blue (B)	Brown (M)	Black (N)	Green (V)
Size	4	3	4	4

One obtains in the same way the following table for the variable "Honors for the bachelor degree":

Honor	Acceptable (P)	Fairly Good (AB)	Good (B)	Very Good (TB)
Size	8	4	2	1

Bar charts of "Colors of eyes" and "Honors for the bachelor degree"





One notices that the students are unevenly distributed among the categories of the variable "Honors for the bachelor degree". The **repartition table** of a variable illustrates how the individuals are distributed among its categories. Generally speaking, the frequency of any category "M" of a qualitative variable is given by the following formula:

$$f_M = (\text{frequency of a category "M" of a qualitative variable}) = \frac{(\text{size of "M"})}{(\text{total size})}$$

and its percentage by the formula:

$$p_M = (\text{percentage of the individuals belonging to the category "M"}) = f_M \times 100$$

It can easily be seen that

$$(\text{sum of the frequencies of all the categories of a qualitative variable}) = 1$$

and

$$(\text{sum of the percentages of all the categories of a qualitative variable}) = 100$$

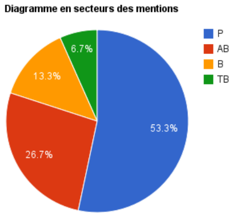
Repartition table of the variable "Honors for the bachelor degree"

Honors	Sizes	Frequencies	Percentages
P	$n_P = 8$	$f_P = 8/15 = 0,533$	$53,3\%$
AB	$n_{AB} = 4$	$f_{AB} = 4/15 = 0,267$	$26,7\%$
B	$n_B = 2$	$f_B = 2/15 = 0,133$	$13,3\%$
TB	$n_{TB} = 1$	$f_{TB} = 1/15 = 0,067$	$6,7\%$
	total size $N = 15$	$f_P + f_{AB} + f_B + f_{TB} = 1$	Total $= 100\%$

One mentions in passing that in this table the percentages are rounded to the nearest tenth (there is only one digit after the decimal comma).

Before ending this subsection, one mentions that distribution of individuals among the categories of a qualitative variable can also be illustrated by a **pie chart**.

Pie chart of the variable "Honors for the bachelor degree"



2.2.2 Discrete quantitative variable

Generally speaking, one associates to each value k of a discrete quantitative variable **a size denoted by** n_k; this is in fact the number of the individuals for which the variable takes the value k. The frequency and percentage corresponding to this same value k are denoted by f_k and p_k, and defined as:

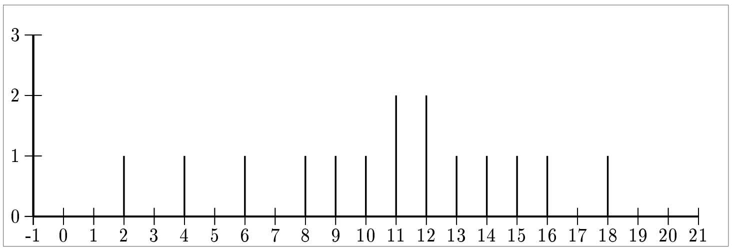
$$f_k = \frac{n_k}{N} \quad \text{and} \quad p_k = f_k \times 100$$

where N is the total size (the total number of individuals).

Repartition table of the variable "Students grades for the statistics examination"

Grades	Sizes	Frequencies
k=0	0	0
k=1	0	0
k=2	1	1/15
k=3	0	0
k=4	1	1/15
k=5	0	0
k=6	1	1/15
k=7	0	0
k=8	1	1/15
k=9	1	1/15
k=10	1	1/15
k=11	2	2/15
k=12	2	2/15
k=13	1	1/15
k=14	1	1/15
k=15	1	1/15
k=16	1	1/15
k=17	0	0
k=18	1	1/15
k=19	0	0
k=20	0	0

Such a table can be represented by a bar chart:

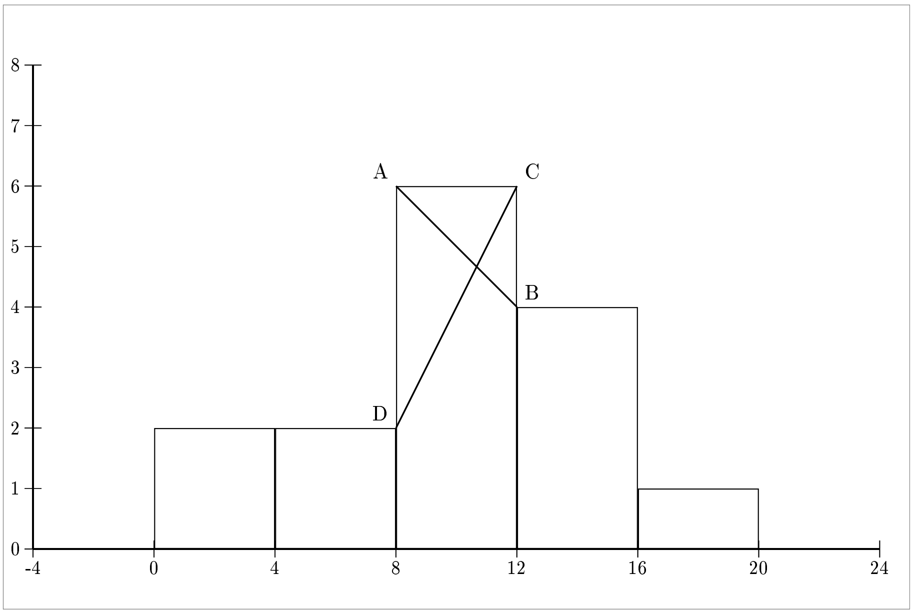


Yet, this representation can hardly be analyzed because it goes too much into details. Therefore, it is more relevant to use **classes** of grades, for instance the five classes: $[0, 4]$; $]4, 8]$; $]8, 12]$; $]12, 16]$ and $]16, 20]$. This approach allows to obtain a new variable which is said to be **classified**. It is important to precisely determine whether or not a bound of a class is included in it.

Repartition table of the classified variable "Students grades for the statistics examination"

Classes of grades	Sizes	Frequencies
$[0, 4]$	2	$2/15$
$]4, 8]$	2	$2/15$
$]8, 12]$	6	$6/15$
$]12, 16]$	4	$4/15$
$]16, 20]$	1	$1/15$

Histogram of the classified variable "Students grades for the statistics examination"



Notice that for this histogram the ordinate axis concerns the sizes of the classes which is rather uncommon; indeed, it is more usual that the ordinate axis of a histogram concerns the frequencies of classes. Also notice that a simple modification of graduation allows to transform sizes in frequencies and vice versa.

By creating classes *one synthesizes* information; even if some information is lost this is balanced by the fact that the structure of the underlying *statistical distribution* (that is the law of probability which can serve as a model) becomes more clear. The above histogram seems to be bell-shaped, this would mean that the "Students grades for the statistics examination" could be modeled by a normal law with appropriate parameters.

2.2.3 Continuous quantitative variable

A continuous quantitative variable can not be represented by a bar chart because it may take any value in some interval I in which there are infinitely many numbers. In order to build a repartition table for such a variable one needs to divide the interval I in sub-intervals $[x_0, x_1],]x_1, x_2], \dots,$

$]x_{k-1}, x_k]$ called **classes**. As far as possible, the bound of the classes $x_0, x_1, \ldots, x_k$ should be chosen in such a way that each class corresponds to a homogeneous group of individuals which is distinguishable from the other groups of individuals associated with the other classes. Usually, the number k of the classes should not be large (no more then ten).

The amplitude of the class $[x_0, x_1]$, that is its "width", is equal to $a_1 = x_1 - x_0$. Similarly, for each $i = 2, \ldots, k$, **the amplitude of the class** $]x_{i-1}, x_i]$ **is equal to** $a_i = x_i - x_{i-1}$**.** When the last class is defined by "more than ... ", then its amplitude is undetermined.

The histogram of a continuous quantitative variable is formed by the juxtaposition of the rectangles whose bases are the classes of the variable and whose areas are proportional to the frequencies of these classes. Thus, to the i-th class corresponds a rectangle based on the interval $]x_{i-1}, x_i]$ (or the interval $[x_0, x_1]$ if $i = 1$) and whose area is proportional to the frequency f_i.

If all the classes are of the same amplitude then the heights of the rectangles are proportional to their areas, and consequently these heights are proportional to the frequencies of the classes.

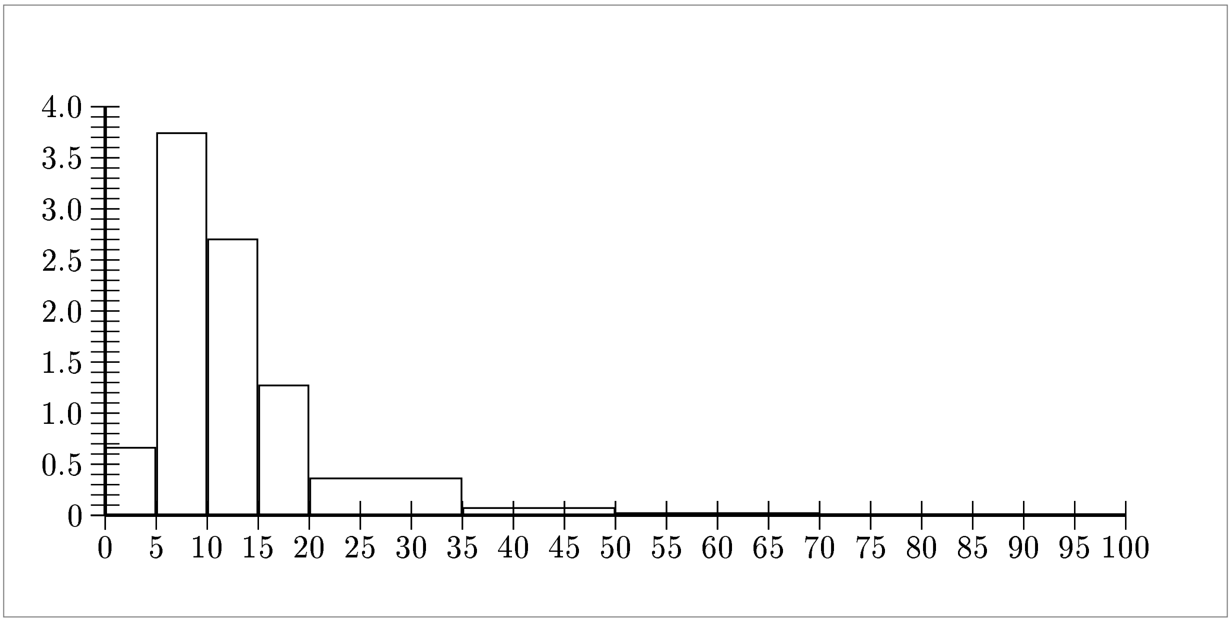
In the opposite case where there are some classes having different amplitudes, the height h_i of the rectangle corresponding to the i-th class is given by $h_i = f_i/a_i$ which corresponds to **the frequency per unit of amplitude**.

Let us now study a concret example of a continuous quantitative variable.

Repartition table of the continuous quantitative variable "Incomes of the French taxpayers in 1965"

Class of income	Size numbered		Amplitude	Height × 500
in	in the thousands	Percentage	in	
Francs	of individuals		Francs	$= \frac{\text{Percentage}}{\text{Amplitude}} \times 500$
$[0, 5000]$	549,3	$6,67\%$	5000	0,67
$]5000, 10000]$	3087,4	$37,51\%$	5000	3,75
$]10000, 15000]$	2229,0	$27,08\%$	5000	2,71
$]15000, 20000]$	1056,7	$12,84\%$	5000	1,28
$]20000, 35000]$	925,0	$11,24\%$	15000	0,37
$]35000, 50000]$	211,0	$2,56\%$	15000	0,09
$]50000, 70000]$	90,8	$1,1\%$	20000	0,03
$]70000, 100000]$	81,6	$0,99\%$	30000	0,02
	Total Size $= 8230,8$	Total $= 100\%$		

Histogram of the variable
"Incomes of the French taxpayers"
(the scale on the horizontal axis is 1 thousand of francs
and the scale on the vertical axis is 1/500)



The histogram seems to be bell-shaped, this would mean that the "Incomes of the French taxpayers in 1965" could be modeled by a normal law with appropriate parameters.

2.3 Measures of central tendency

2.3.1 The mode

a) Discrete quantitative variable (without classes)

The **mode** is the value of the variable having the largest size (or frequency).

Example 2.1. *A statistical survey was made on a population of 12398 families. Its goal was to determine how they were distributed in terms of their number of children having less than 14 years old. It led to the following table:*

Number of children	*Number of families*
0	*2601*
1	*6290*
2	*2521*
3	*849*
4	*137*
	Total = 12398

The mode of the variable "Number of children" is 1 child.

Remark 2.2. *Some variables can have more than just one mode. For example, for the variable "Students grades for the statistics examination" the largest size (or frequency) corresponds to the two values 11 and 12 of the variable.*

b) Continuous quantitative variable and discrete classified quantitative variable

The **modal class** is the one having the largest frequency per unit of amplitude; in other words the one having the highest rectangle of the histogram. For instance, for the continuous variable "Incomes of the French taxpayers" the modal class is $]5000, 10000]$. One mentions in passing that there are some variables having more than just one modal class.

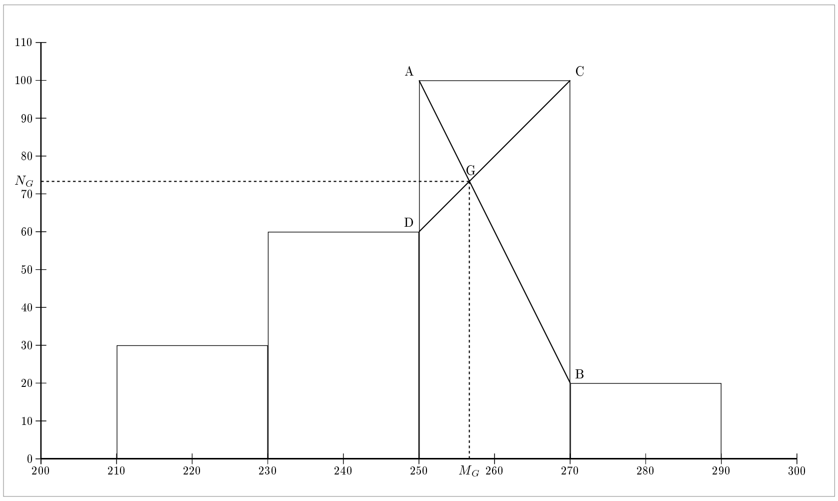
For being more precise, one can determine inside of the modal class **the accurate value of the mode**; the method for finding this accurate value is presented in the following example.

Example 2.2. *The accurate value of the mode of the variable described in this example allows to estimate the psychological price of a new product which can attract a maximal number of consumers. More precisely a statistical survey for determining the "best" price of a new product was made on a sample of 210 consumers. It led to the following table which provides for each price range in Euros the number of consumers ready to buy the product.*

Class (price range)	*Size (number of consumers)*
$[210, 230]$	*30*
$]230, 250]$	*60*
$]250, 270]$	*100*
$]270, 290]$	*20*
	Total = 210

The classes (of prices) being of the same amplitude (equals to 20), their sizes coincide with the heights of the corresponding rectangles of the histogram.

Histogram of the continuous variable "Price"



The modal class is $]250, 270]$ *since this is the price range which attracts the largest number of consumers. In order to determine inside of it the accurate value of the mode one proceeds by interpolation as shown on the above figure. One focuses on G the point of intersection of the two segments [AB] and [CD]. The abscissa* M_G *of G is the accurate value of the mode of the variable price, one can see on the above figure that* $M_G \simeq 257$ *Euros. The ordinate* N_G *is the* *number of consumers who are ready to buy the product at the price M_G, one can see on the above figure that $N_G \simeq 73$ consumers. Apart from this graphical method M_G and N_G can be more precisely determined by using computations. To this end, first one has to find the equations of the lines (AB) and (CD). Recall that, generally speaking, the equation of a non-vertical line can be expressed in the form $y = ax + b$, where a and b are two parameters. In our setting, for determining the values of a and b for the line (AB), one has to solve the linear system of the two equations*

$$\begin{cases} 250a + b = 100 \\ 270a + b = 20 \end{cases}$$

which reflects the fact that this line passes through the point A with coordinates $(250, 100)$ *and the point B with coordinates* $(270, 20)$. *One has*

$$\begin{cases} 250a + b = 100 \\ 270a + b = 20 \end{cases} \Leftrightarrow \begin{cases} 250a + b = 100 \\ -20a = 80 \end{cases} \Leftrightarrow \begin{cases} b = 100 - 250 \times (-4) = 1100 \\ a = -4 \end{cases}$$

thus the equation of the line (AB) is $y = -4x + 1100$. For determining the values of a and b for the line (CD), one has to solve the linear system of the two equations

$$\begin{cases} 250a + b = 60 \\ 270a + b = 100 \end{cases}$$

which reflects the fact that this line passes through the point D with coordinates $(250, 60)$ *and the point C with coordinates* $(270, 100)$. *One has*

$$\begin{cases} 250a + b = 60 \\ 270a + b = 100 \end{cases} \Leftrightarrow \begin{cases} 250a + b = 60 \\ 20a = 40 \end{cases} \Leftrightarrow \begin{cases} b = 60 - 250 \times 2 = -440 \\ a = 2 \end{cases}$$

thus the equation of the line (CD) is $y = 2x - 440$. At last the coordinates (M_G, N_G) *of the point G are obtained by solving the linear system of the two equations*

$$\begin{cases} N_G = -4M_G + 1100 \\ N_G = 2M_G - 440 \end{cases}$$

which reflects the fact that these coordinates (M_G, N_G) *satisfy simultaneously the equations of (AB) and (CD). One has*

$$\begin{cases} N_G = -4M_G + 1100 \\ N_G = 2M_G - 440 \end{cases} \Leftrightarrow \begin{cases} -6M_G + 1540 = 0 \\ N_G = 2M_G - 440 \end{cases} \Leftrightarrow \begin{cases} M_G = \frac{770}{3} \simeq 256,66 \\ \\ N_G = 2 \times \frac{770}{3} - 440 \simeq 73,33 \end{cases}$$

2.3.2 Median and Quantiles

The **median** (denoted by M_e) of a quantitative variable is its value which divides the population into two sub-populations of the same size. More precisely, the number of the individuals for which the values of the variable are greater than M_e is equal to the number of the individuals for which the values of the variable are less than M_e.

a) Discrete quantitative variable (without classes)

We are going to present a method[1] for computing the median. One first orders the individuals decreasingly that is by beginning by the individual having the highest value and by ending by

[1] This method, as well as other methods for computing the median, should be used with caution, especially when some values of the variable are repeated. Indeed, in this particular case, the method may not give a very satisfactory result.

the individual having the lowest value. Then, one orders the individuals increasingly that is by beginning by the individual having the lowest value and by ending by the individual having the highest value. At last, one associates to each individual a quantity called the depth and defined as the minimum of his two ranks provided by the decreasing and the increasing orders. **The median of the variable is the arithmetic mean of its values for the individuals having the maximal depth.**

For illustrating this method let us study two concrete examples:

Example 2.3.

Individuals	*Grades*	*Ranks (decreasingly)*	*Ranks (increasingly)*	*Depths*
Michel	*12*	*6*	*9*	*6*
Jean	*8*	*12*	*4*	*4*
Stéphane	*13*	*5*	*11*	*5*
Charles	*11*	*8*	*7*	$\boxed{7}$
Agnès	*10*	*10*	*6*	*6*
Nadine	*9*	*11*	*5*	*5*
Étienne	*16*	*2*	*14*	*2*
Gilles	*14*	*4*	*12*	*4*
Aurélie	*11*	*8*	*7*	$\boxed{7}$
Stéphanie	*15*	*3*	*13*	*3*
Marie-Claude	*4*	*14*	*2*	*2*
Anne	*18*	*1*	*15*	*1*
Christophe	*12*	*6*	*9*	*6*
Pierre	*6*	*13*	*3*	*3*
Bernadette	*2*	*15*	*1*	*1*

The median is equal to

$$M_e = \frac{11+11}{2} = 11\,.$$

Example 2.4. *This is the same example as the previous one except that Bernadette has been cancelled from the data table.*

Individuals	*Grades*	*Ranks (decreasingly)*	*Ranks (increasingly)*	*Depths*
Michel	*12*	*6*	*8*	$\boxed{6}$
Jean	*8*	*12*	*3*	*3*
Stéphane	*13*	*5*	*10*	*5*
Charles	*11*	*8*	*6*	$\boxed{6}$
Agnès	*10*	*10*	*5*	*5*
Nadine	*9*	*11*	*4*	*4*
Étienne	*16*	*2*	*13*	*2*
Gilles	*14*	*4*	*11*	*4*
Aurélie	*11*	*8*	*6*	$\boxed{6}$
Stéphanie	*15*	*3*	*12*	*3*
Marie-Claude	*4*	*14*	*1*	*1*
Anne	*18*	*1*	*14*	*1*
Christophe	*12*	*6*	*8*	$\boxed{6}$
Pierre	*6*	*13*	*2*	*2*

The median is equal to

$$M_e = \frac{11+11+12+12}{4} = 11,5\,.$$

Problem 2.1. *(a) Determine the median in the case where Agnès and Stéphanie are cancelled from the data table in Example 2.4. (b) Determine the median in the case where Agnès and Jean are cancelled from the data table in Example 2.4.*

b) Continuous quantitative variable and discrete classified quantitative variable

Let us first introduce the notions of **cumulative size**, **cumulative frequency**, and **cumulative function**. X denotes a continuous quantitative variable or a discrete classified quantitative variable whose range of values has been divided into k disjoint classes $[x_0, x_1],]x_1, x_2], \ldots,]x_{k-1}, x_k]$. Their sizes are denoted by n_1, n_2, $\ldots$, n_k. **The cumulative size of the 1-st class** (that is the class $[x_0, x_1]$) is the number N_1 of individuals for which *the value of the variable X is less than or equal to x_1*; thus one has

$$N_1 = n_1.$$

The cumulative size of the 2-nd class (that is the class $]x_1, x_2]$) is the number N_2 of individuals for which *the value of the variable X is less than or equal to x_2*; thus one has

$$N_2 = n_1 + n_2 = N_1 + n_2$$

The cumulative size of the 3-rd class (that is the class $]x_2, x_3]$) is the number N_3 of individuals for which *the value of the variable X is less than or equal to x_3*; thus one has

$$N_3 = n_1 + n_2 + n_3 = N_2 + n_3$$

More generally, **the cumulative size of the i-th class** (that is the class $]x_{i-1}, x_i]$), where $i = 1, 2, \ldots, k$, is the number N_i of individuals for which *the value of the variable X is less than or equal to x_i*; thus one has

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{l=1}^{i} n_l = N_{i-1} + n_i\,,$$

where the last equality holds when $i \geq 2$. **The cumulative frequency of the i-th class** is denoted by F_i and defined as

$$F_i = \frac{N_i}{N} = \sum_{l=1}^{i} f_l,$$

where f_l is the frequency of the l-th class and N the total size. Thus one has $F_1 = f_1$ and $F_i = F_{i-1} + f_i$ for all $i = 2, \ldots, k$.

Example 2.5. *The cumulative sizes and frequencies of the continuous variable "Incomes of the French taxpayers" are:*

Class of income	*Size*	*Cumulative size*	*Frequency*	*Cumulative frequency*
$[0, 5000]$	*549,3*	*549,3*	*0,0667*	*0,0667*
$]5000, 10000]$	*3087,4*	*3636,7*	*0,3751*	*0,4418*
$]10000, 15000]$	*2229,0*	*5865,7*	*0,2708*	*0,7126*
$]15000, 20000]$	*1056,7*	*6922,4*	*0,1284*	*0,841*
$]20000, 35000]$	*925,0*	*7847,4*	*0,1124*	*0,9534*
$]35000, 50000]$	*211,0*	*8058,4*	*0,0256*	*0,979*
$]50000, 70000]$	*90,8*	*8149,2*	*0,011*	*0,99*
$]70000, 100000]$	*81,6*	*8230,8*	*0,0099*	$0,9999 \simeq 1$

Problem 2.2. *Compute the cumulative size and the cumulative frequency of each class of the discrete classified variable "Students grades for the statistics examination".*

Solution of the Problem 2.2

Classes of grades	*Sizes*	*Cumulative sizes*	*Frequencies*	*Cumulative frequencies*
$[0,4]$	*2*	*2*	*0,133*	*0,133*
$]4,8]$	*2*	*4*	*0,133*	*0,266*
$]8,12]$	*6*	*10*	*0,4*	*0,666*
$]12,16]$	*4*	*14*	*0,267*	*0,933*
$]16,20]$	*1*	*15*	*0,067*	*1*

The cumulative function is often denoted by F. **Its value $F(t)$ at an arbitrary real number t is the percentage of individuals in the population for which the values of the variable X is less than or equal to t.**

Remark 2.3. (Four important properties of the cumulative function F)

1. *It is non-decreasing, that is for all real numbers t_1 and t_2 satisfying $t_1 \leq t_2$ one has $F(t_1) \leq F(t_2)$.*
2. *Let x_0 be the lower bound of the first class $[x_0, x_1]$, then one has $F(t) = 0$, for any real number $t \leq x_0$.*
3. *Let x_k be the upper bound of the last class $]x_{k-1}, x_k]$, then one has $F(t) = 1$, for any real number $t \geq x_k$.*
4. *For all $l = 1, \ldots, k$, one has $F(x_l) = F_l$, where F_l is the cumulative frequency of the l-th class.*

Remark 2.4. *When X is a continuous variable, using the fact that its classes are homogeneous entities, the values of its cumulative function F inside each one of them are obtained by linear interpolation: For all $l = 1, \ldots, k$, and for any real number t satisfying $x_{l-1} \leq t \leq x_l$, one has*

$$F(t) = \Big(\frac{t - x_{l-1}}{x_l - x_{l-1}}\Big) F_l + \Big(1 - \frac{t - x_{l-1}}{x_l - x_{l-1}}\Big) F_{l-1}\,, \tag{2.1}$$

with the convention that $F_0 = 0$.

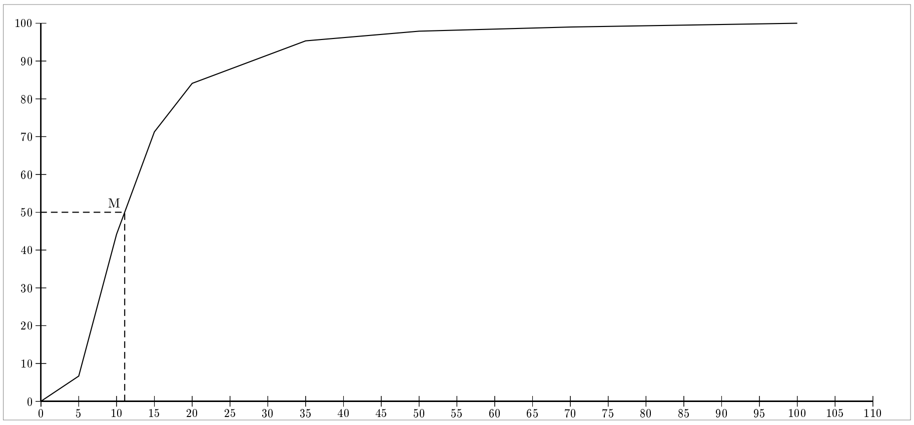
Remark 2.5. *Generally speaking, the median M_e of a continuous variable X with cumulative function F is such that*

$$F(M_e) = 50\%\,. \tag{2.2}$$

Thus M_e can be determined by using the equation (2.2).

Example 2.6. *We draw the graph of the cumulative function F of the continuous variable "Incomes of the French taxpayers". Then we determine the median M_e of this variable.*

Graph of the cumulative function F of the continuous variable "Incomes of the French taxpayers"



The unit on the horizontal axis is 1 thousand of Francs.
The vertical axis corresponds to percentages.

Using (2.2) and a graphical method we find $M_e \simeq 11,1$ thousands of Francs.

The following computational method allows to determine M_e in a more precise way. First one computes the equation of the line part of the graph of F to which the point M with coordinates $(M_e\,;\,50)$ belongs. Using the fact that this line passes through the two points with coordinates $(10\,;\,44,18)$ and $(15\,;\,71,26)$, one obtains the following linear system of two equations

$$\begin{cases} 10a+b=44,18 \\ 15a+b=71,26 \end{cases}$$

One has

$$\begin{cases} 10a+b=44,18 \\ 15a+b=71,26 \end{cases} \Leftrightarrow \begin{cases} 10a+b=44,18 \\ 5a=71,26-44,18=27,08 \end{cases} \Leftrightarrow \begin{cases} b=44,18-10\times 5,416=-9,98 \\ \\ a=\frac{27,08}{5}=5,416 \end{cases}$$

Thus the equation of this line is $y = 5,416x - 9,98$. Finally, using the fact the coordinates $(M_e\,;50)$ of the point M satisfy the latter equation, one gets that

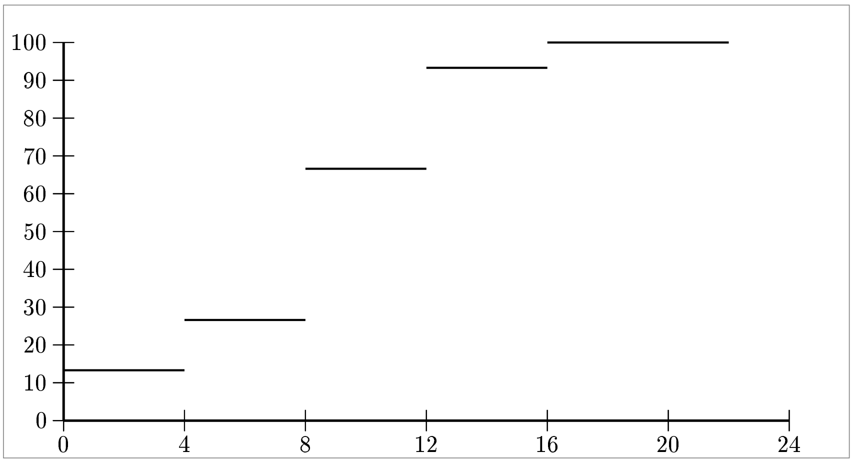
$$50 = 5,416M_e - 9,98 \Leftrightarrow M_e = \frac{50+9,98}{5,416} \simeq 11,075 \textit{ thousands of Francs.}$$

Another computational method for determining M_e consists in using Thales Theorem (the basic proportionality theorem):

$$\frac{50-44,18}{71,26-44,18} = \frac{M_e-10}{15-10} \Leftrightarrow M_e = (15-10)\times\Big(\frac{50-44,18}{71,26-44,18}\Big)+10 \simeq 11,075 \textit{ thousands of Francs.}$$

Remark 2.6. *When X is* **a discrete classified variable** *(for example the variable "Students grades for the statistics examination" in Problem 2.2), the graph of its cumulative function looks like a staircase and the jumps in it are very usually of different sizes. Generally speaking, in such a situation,* **the equation (2.2) characterizing the median fails to have a solution.** *Thus, other measures (arithmetic mean, etc.) of central tendency of X have to be used.*

Graph of the cumulative function of the discrete classified variable "Students grades for the statistics examination"



The notion of **quantile of order** α **(**$0 \leq \alpha \leq 1$**)** generalizes the notion of median. The quantile of order α of a quantitative variable X is its value x_α which divides the corresponding population of size N into two sub-populations of sizes αN and $(1-\alpha)N$. The first sub-population is formed by the individuals having values less than x_α, while the second one is formed by those having values greater than x_α. When X is a continuous variable of cumulative function F, one can determine x_α by using the equation

$$F(x_\alpha) = \alpha. \tag{2.3}$$

The quartiles of X are the three quantiles $x_{0,25}$, $x_{0,5}$ and $x_{0,75}$. $Q_1 = x_{0,25}$ is called the first quartile; one quarter of the values of X are less than Q_1. $Q_2 = x_{0,5} = M_e$ is the median. $Q_3 = x_{0,75}$ is called the third quartile; one quarter of the values of X are greater than Q_3.

The interquartile range (IQR) is defined as the difference between the third quartile and the first quartile:

$$IQR = Q_3 - Q_1.$$

The IQR is a measure of dispersion or spread of X in an absolute way or in comparison with another quantitative variable. Indeed, 50% of the values of X belong to the interval $[Q_1, Q_3]$; thus, the wider is IQR the more spread is X.

Problem 2.3. *Determine Q_1, Q_3 and IQR for the continuous variable "Incomes of the French taxpayers".*

2.3.3 Different notions of mean

Throughout this sub-section X is a discrete quantitative variable related with a population of N individuals; the values of X for these individuals are denoted by $x_1, x_2, \ldots, x_N$. The number of the *distinct* [2] values of X is denoted by K; for the sake of simplicity, one assumes that $x_1, \ldots, x_K$ are the distinct values of X.

[2] The number K of the distinct values of X should not be confused with the number N of the values of X; for instance, for the variable "Grades" in Example 2.3 one has $N = 15$ and $K = 13$.

a) Arithmetic mean

The arithmetic mean of X is denoted by $\overline{x}$ and defined as:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i\,. \tag{2.4}$$

Example 2.7. *For instance, the two arithmetic means of the two variables in Examples 2.3 and 2.4 are respectively* $\frac{161}{15} \simeq 10,73$ *and* $\frac{159}{14} \simeq 11,36$*. Recall that the only difference between these two examples is that in the second one Bernadette is cancelled from the data table. It turns out that the latter fact (the cancellation of Bernadette) has a more significant impact on the arithmetic mean than on the median which increases from* 11 *to* 11,5*. Generally speaking, arithmetic mean is more sensitive than median to extreme values.*

One denotes by n_i the number of occurrences of the value x_i of the variable X. For instance for the variable "Grades" the number of occurrences of the value 18 equals 1 since the value 18 is observed for only one individual; while the number of occurences of the value 11 equals 2 since the value 11 is observed for two different individuals.

Observe that using the fact that $\underbrace{x_i + x_i + \ldots + x_i}_{n_i \text{ times}} = n_i x_i$ it follows from (2.4) that

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{K} n_i x_i = \sum_{i=1}^{K} f_i x_i\,; \tag{2.5}$$

recall that $x_1, \ldots, x_K$ are assumed to be the distinct values of the variable X, and that $f_i = n_i/N$ is the frequency of the value x_i. The expression $\sum_{i=1}^{K} f_i x_i$ is called **weighted arithmetic mean of** X since the distinct values of X are weighted by the corresponding frequencies.

Example 2.8. *A statistical survey on a population of households has shown that* 30% *of them have 1 child,* 40% *2 children,* 15% *3 children,* 10% *4 children, and* 5% *5 children.*

The weighted arithmetic mean of the discrete variable "Number of children per household" is:

$$0,3 \times 1 + 0,4 \times 2 + 0,15 \times 3 + 0,1 \times 4 + 0,05 \times 5 \simeq 2,2 \textit{ children.}$$

b) Quadratic mean

The quadratic mean of X is denoted by $m_{2,X}$ and defined as:

$$m_{2,X} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \sqrt{\sum_{i=1}^{K} f_i x_i^2}\,.$$

Thus, the weighted quadratic mean of the variable "Number of children per household" of the Example 2.8 is:

$$m_2 = \left(0,3 \times 1^2 + 0,4 \times 2^2 + 0,15 \times 3^2 + 0,1 \times 4^2 + 0,05 \times 5^2\right)^{1/2} \simeq 2,47.$$

Problem 2.4. *Calculate the quadratic means of the two variables "Grades" already studied in Examples 2.3 and 2.4.*

c) Harmonic mean

Caution: The harmonic mean of the variable X can only be defined when the values $x_1, \dots, x_N$ of X are strictly positive real numbers.

The harmonic mean of X is denoted $m_{-1,X}$ and defined as:

$$m_{-1,X} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} = \frac{1}{\sum_{i=1}^{K} \frac{f_i}{x_i}} \,.$$

The harmonic mean can typically be used when a "concrete" meaning can be given to $1/x_1, \dots, 1/x_N$ the inverses of the values of X; this is for instance the case when the latter values are exchange rates, speeds of vehicles, and so on.

Example 2.9. *In a first transaction* 100 *Euros were converted in Dollars at the exchange rate* $0,87$ *Euro per Dollar. In a second transaction other* 100 *Euros were converted in Dollars at the exchange rate* $0,71$ *Euro per Dollar. The total amount of Dollars bought in both of these transactions is:*

$$\frac{100}{0,87} + \frac{100}{0,71} \simeq 255,79 \text{ Dollars.}$$

The average exchange rate for both of these transactions *is defined as the exchange rate of* c_m *Euro per Dollar which would have allowed to buy in a single transaction* $255,79$ *Dollars for* 200 *Euros. Therefore* c_m *satisfies the equation*

$$\frac{200}{c_m} = \frac{100}{0,87} + \frac{100}{0,71} \simeq 255,79$$

which implies that

$$c_m = \frac{200}{\frac{100}{0,87} + \frac{100}{0,71}} = \frac{2}{\frac{1}{0,87} + \frac{1}{0,71}} \simeq 0,78$$

Thus it turns out that c_m *is the harmonic mean of the exchange rates* $0,87$ *and* $0,71$ *of the two transactions. Also notice that* c_m *is strictly less than* $(0,87+0,71)/2 = 0,79$ *the arithmetic mean of these two rates.*

Problem 2.5. *A motorist has driven 40 kilometers at the speed 60 km/h and 40 other kilometers at the speed 120 km/h. One denotes by* v_m *his average speed in km/h for the whole journey of 80 kilometers. Calculate* v_m *and comment your result.*

d) Geometric mean

Caution: The geometric mean of the variable X can only be defined when the values $x_1, \dots, x_N$ of X are strictly positive real numbers.

The geometric mean of X is denoted by $M_{g,X}$ and defined as:

$$\begin{aligned} M_{g,X} &= \big(x_1 \times x_2 \times \ldots \times x_N\big)^{1/N} = \exp\Big(\frac{1}{N} \ln \big(x_1 \times x_2 \times \ldots \times x_N\big)\Big) \\ &= x_1^{f_1} \times \ldots \times x_K^{f_K} \,, \end{aligned}$$

where exp is the exponential function and ln the natural (Napierian) logarithm function.

Example 2.10. *During a decade (a period of ten years) salaries were multiplied by* 2, *and during the following decade they were multiplied by* 4. *Thus, for the whole period of these two decades the multiplying factor of salaries is* $2 \times 4 = 8$. The average multiplying factor per decade *for this period of twenty years is denoted by* μ *and defined as the multiplying factor which does not* *change from one decade to another and allows for a multiplication by* 8 *of salaries between the beginning and the end of this period of two decades. Therefore one has* $\mu^2 = 8 = 2 \times 4$ *that is* $\mu = \sqrt{2 \times 4} \simeq 2,83$. *Thus it turns out that* μ *is the geometric mean of the two multiplying factors* 2 *and* 4 *associated with the two decades. Also notice that* μ *is strictly less than* $(2+4)/2 = 3$ *the arithmetic mean of these two multiplying factors.*

Remark 2.7. *When the values* $x_1, \ldots, x_N$ *of* X *are strictly positive real numbers, one has:*

$$\min_{1 \leq i \leq N} x_i \leq m_{-1,X} \leq M_{g,X} \leq \overline{x} \leq m_{2,X} \leq \max_{1 \leq i \leq N} x_i \,.$$

In other words:

(Minimum of the values of X)
$\leq$ *(Harmonic mean of X)*
$\leq$ *(Geometric mean of X)*
$\leq$ *(Arithmetic mean of X)*
$\leq$ *(Quadratic mean of X)*
$\leq$ *(Maximum of the values of X)*

Thanks to these inequalities one can detect some miscalculations concerning some of these means.

Remark 2.8. *In one or other of the following two cases:*

- Y *is a continuous variable whose range of values has been divided into* k *classes* $[y_0, y_1]\,,]y_1, y_2]\,, \ldots\,,]y_{k-1}, y_k]\,;$
- Y *is a discrete classified variable whose classes are* $[y_0, y_1]\,,]y_1, y_2]\,, \ldots\,,]y_{k-1}, y_k]$.

The arithmetic mean of Y*, denoted by* $\overline{y}$*, is defined as* the arithmetic mean of the centers of the classes of Y weighted by the corresponding frequencies*; more precisely:*

$$\overline{y} = \sum_{i=1}^{k} f_i \Big(\frac{y_{i-1} + y_i}{2} \Big) = \frac{1}{N} \sum_{i=1}^{k} n_i \Big(\frac{y_{i-1} + y_i}{2} \Big),$$

where, for each i*,* f_i *and* n_i *respectively denote the frequency and the size of the* i*-th class. Notice that* $N = \sum_{i=1}^{k} n_i$ *is the total size.*

In the same vein, $m_{2,Y}$ *the quadratic mean of* Y *is defined as:*

$$m_{2,Y} = \sqrt{\sum_{i=1}^{k} f_i \Big(\frac{y_{i-1} + y_i}{2} \Big)^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{k} n_i \Big(\frac{y_{i-1} + y_i}{2} \Big)^2}$$

Moreover, when Y *is with strictly positive values (that is* $y_0 > 0$*) its harmonic and geometric means* $m_{-1,Y}$ *and* $M_{g,Y}$ *are defined as*

$$m_{-1,Y} = \bigg(\sum_{i=1}^{k} f_i \Big(\frac{y_{i-1} + y_i}{2} \Big)^{-1} \bigg)^{-1} = \bigg(\frac{1}{N} \sum_{i=1}^{k} n_i \Big(\frac{y_{i-1} + y_i}{2} \Big)^{-1} \bigg)^{-1}$$

and

$$M_{g,Y} = \Big(\frac{y_0 + y_1}{2} \Big)^{f_1} \times \ldots \times \Big(\frac{y_{k-1} + y_k}{2} \Big)^{f_k}$$

Problem 2.6. (a) *Calculate the arithmetic mean of the continuous variable "Incomes of the French taxpayers" (the answer is* 14292, 5 *Francs).* (b) *Calculate the arithmetic mean of the classified variable "Students grades for the statistics examination" which has already appeared in Problem 2.2 (the answer is* 10, 008*).*

2.4 Measures of dispersion

Throughout this sub-section X is a discrete quantitative variable related with a population of N individuals; the values of X for these individuals are denoted by $x_1, x_2, \ldots, x_N$. The number of the *distinct* values of X is denoted by K; for the sake of simplicity, one assumes that $x_1, \ldots, x_K$ are the distinct values of X.

a) The range

The range r_X of the variable X is *the difference between its largest and smallest values:*

$$r_X = \max_{1 \leq i \leq N} x_i - \min_{1 \leq i \leq N} x_i .$$

For the variable "Grades" in Example 2.3 the range is $18 - 2 = 16$.

b) Variance and standard deviation

The variance of the variable X is denoted by $\mathrm{Var}(X)$ and defined as *the arithmetic mean of the squares of the differences between the values of X and $\overline{x}$ its arithmetic mean*:

$$\mathrm{Var}(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2 . \tag{2.6}$$

The formula (2.6) can also be rewritten as:

$$\mathrm{Var}(X) = \sum_{i=1}^{K} f_i (x_i - \overline{x})^2;$$

recall that $x_1, \ldots, x_K$ are assumed to be the distinct values of the variable X, and that $f_i = n_i/N$ is the frequency of the value x_i. Another important formula (sometimes called Huygens formula) for calculating the variance is:

$$\begin{aligned}\mathrm{Var(X)} &= \left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i^2}\right) - (\overline{\mathrm{x}})^2 = \left(\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{f_i x_i^2}\right) - (\overline{\mathrm{x}})^2 \\ &= \big(\text{Quadratic mean of } X\big)^2 - \big(\text{Arithmetic mean of } X\big)^2 \end{aligned} \tag{2.7}$$

The standard deviation of the variable X is denoted by σ_X and defined as *the square root of the variance of this variable*:

$$\sigma_X = \sqrt{\mathrm{Var(X)}}.$$

One mentions in passing that standard deviation is the most commonly used measure of dispersion.

Example 2.11. *We are going to calculate the variance and the standard deviation of the variable "Grades" already studied in Example 2.3. Here this variable is denoted by X, its values are denoted by x_i with $i = 1, \ldots, 15$, and its arithmetic mean by $\overline{x}$; recall that $\overline{x} = 10,73$*

Individuals	*Grades*	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	x_i^2
Michel	*12*	*1,27*	*1,61*	*144*
Jean	*8*	*-2,73*	*7,45*	*64*
Stéphane	*13*	*2,27*	*5,15*	*169*
Charles	*11*	*0,27*	*0,07*	*121*
Agnès	*10*	*-0,73*	*0,53*	*100*
Nadine	*9*	*-1,73*	*2,99*	*81*
Étienne	*16*	*5,27*	*27,77*	*256*
Gilles	*14*	*3,27*	*10,69*	*196*
Aurélie	*11*	*0,27*	*0,07*	*121*
Stéphanie	*15*	*4,27*	*18,23*	*225*
Marie-Claude	*4*	*-6,73*	*45,29*	*16*
Anne	*18*	*7,27*	*52,86*	*324*
Christophe	*12*	*1,27*	*1,61*	*144*
Pierre	*6*	*-4,73*	*22,37*	*36*
Bernadette	*2*	*-8,73*	*76,21*	*4*
			Total=272,9	*Total=2001*

There are two methods for calculating Var(X)*, the first one relies on the formula (2.6) and the second one on the formula (2.7).* **The first method** *consists in what follows. According to the previous table, the sum of the squares of the differences between the values of X and its arithmetic mean equals to* $272,9$*; thus using (2.6), one obtains*

$$\mathrm{Var(X)} = \frac{272,9}{15} \simeq 18,19 \tag{2.8}$$

The second method *consists in what follows. According to the previous table, the sum of the square of the values of X equals to* 2001*; thus*

$$\big(\textit{Quadratic mean of } X\big)^2 = \frac{2001}{15} = 133,4$$

then one can derive from (2.7) that

$$\mathrm{Var(X)} = 133,4 - (10,73)^2 \simeq 18,27 \tag{2.9}$$

One mentions in passing that the slight difference between the result (2.8) and the result (2.9) can be explained by rounding errors. Moreover, this slight difference almost disappears when one calculates the standard deviation corresponding to each of these two results; indeed, one has $\sqrt{18,19} \simeq 4,26$ *and* $\sqrt{18,27} \simeq 4,27$*.*

A question: *According to you what is the value of sum of the numbers which are on the third column of the previous table ? (justify your answer)*

Problem 2.7. *The incomes in thousands of euros of four employees are: 3,1 ; 2,6 ; 1,7 ; 1,9. The corresponding statistical variable is denoted by Z.* (*a*) *Calculate the variance of Z by means of the two methods we have just presented.* (*b*) *Calculate the standard deviation of Z.*

Example 2.12 (Illustration of usefulness of standard deviation)**.** *The 25 students of a Master are divided into two groups, 13 students are in the group 1 and the 12 remaining students are in the group 2. These 25 students have passed an exam. The following table provides a description of the repartition of the students grades in each group as well as for the two groups together:*

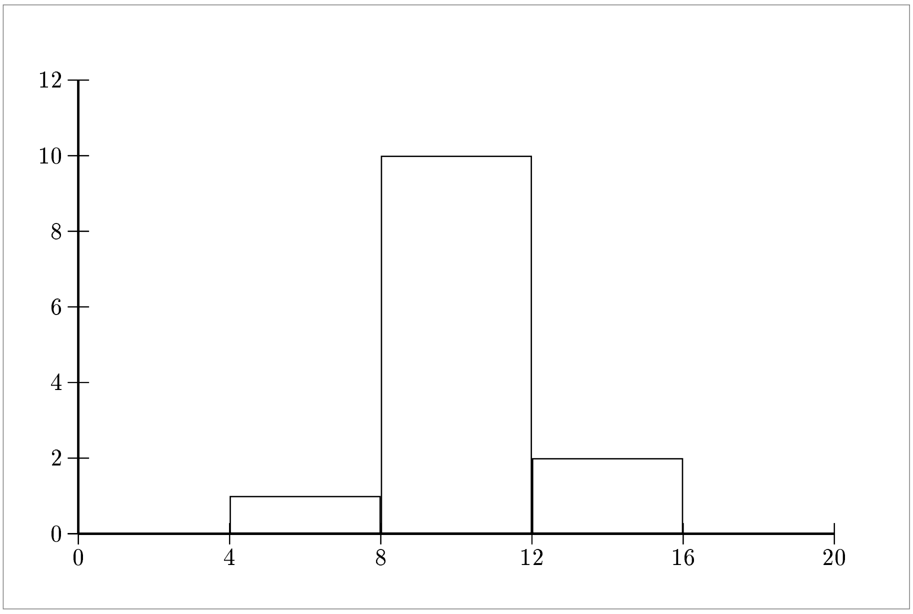
Repartition table of the students grades in each group

Centers of classes	*Classes of grades*	*Sizes for the group 1*	*Sizes for the group 2*	*Sizes for the two groups together*
2	*[0,4]*	*0*	*2*	*2*
6	*]4,8]*	*1*	*2*	*3*
10	*]8,12]*	*10*	*3*	*13*
14	*]12,16]*	*2*	*3*	*5*
18	*]16,20]*	*0*	*2*	*2*
		Total $= N_1 =$*13*	*Total* $= N_2 =$*12*	*Total* $= N =$*25*

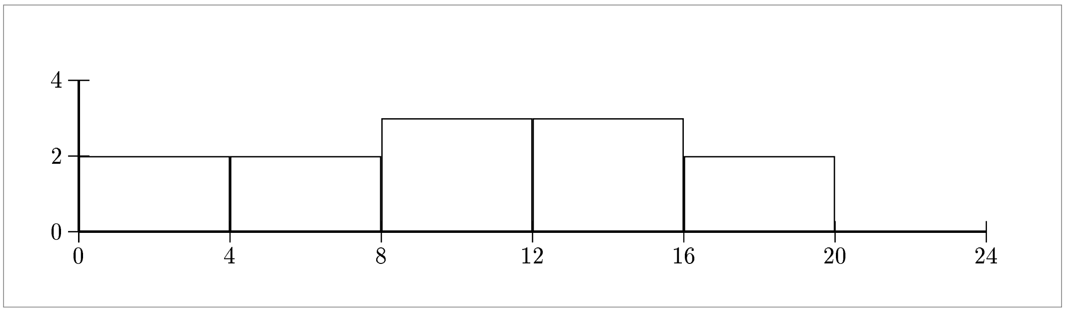
The discrete classified quantitative variable corresponding to the students grades in the group 1 is denoted by X_1*, that corresponding to the students grades in the group 2 is denoted by* X_2*, and that corresponding to the students grades in the two groups together is denoted by* X*.*

First, one wishes to compare the way the grades are distributed in the group 1 to that in the group 2. To this end one draws the histograms corresponding to the variables X_1 *and* X_2*.*

Histogram of the variable X_1 (group 1)



Histogram of the variable X_2 (group 2)



A comparison of the two histograms shows that the grades are much more spread in the group 2 than in the group 1. The calculation of the arithmetic means $\overline{x}_1$ and $\overline{x}_2$ of the variables X_1 and X_2, and of their standard deviations σ_{X_1} and σ_{X_2} will bring precisions concerning this claim. Using Remark 2.8 and (2.7) one has

$$\overline{x}_1 = \frac{1\times 6 + 10\times 10 + 2\times 14}{13} = \frac{134}{13} \simeq 10,31$$

$$\overline{x}_2 = \frac{2\times 2 + 2\times 6 + 3\times 10 + 3\times 14 + 2\times 18}{12} = \frac{124}{12} \simeq 10,33$$

$$\mathrm{Var(X_1)} = \frac{1\times 6^2 + 10\times 10^2 + 2\times 14^2}{13} - \Big(\frac{134}{13}\Big)^2 \simeq 3,60$$

$$\mathrm{Var(X_2)} = \frac{2\times 2^2 + 2\times 6^2 + 3\times 10^2 + 3\times 14^2 + 2\times 18^2}{12} - \Big(\frac{124}{12}\Big)^2 \simeq 27,96$$

$$\sigma_{X_1} = \sqrt{3,60} \simeq 1,90 \quad \textit{and} \quad \sigma_{X_2} = \sqrt{27,96} \simeq 5,29$$

Conclusion : *The standard deviation σ_{X_1} is rather small which means that the grades in the group 1 are rather homogeneous and concentrated around the mean $\overline{x}_1$. Though the mean $\overline{x}_2$ is almost equal to $\overline{x}_1$, for the group 2 the situation is different; the grades are heterogeneous and rather away from $\overline{x}_2$ since the standard deviation σ_{X_2} is nearly equal to three times σ_{X_1}.*

c) Total Variance, Within-groups Variance, Between-groups Variance

Let us continue to study Example 2.12. Recall that the discrete classified variable X concerns the 25 students grades of the groups 1 and 2 together. The arithmetic mean of X is denoted by $\overline{x}$ and called the *total arithmetic mean.* Generally speaking, **an important formula which connects the total arithmetic mean $\overline{x}$ to $\overline{x}_1$ and $\overline{x}_2$ the arithmetic means inside of two underlying groups of sizes N_1 and N_2 is:**

$$\overline{x} = \Big(\frac{N_1}{N_1+N_2}\Big)\overline{x}_1 + \Big(\frac{N_2}{N_1+N_2}\Big)\overline{x}_2 \tag{2.10}$$

Thus in the case of Example 2.12 one gets that

$$\overline{x} \simeq \frac{13}{25}\times 10,31 + \frac{12}{25}\times 10,33 \simeq 10,32$$

Notice that $\overline{x}$ can also be computed in a direct way by using Remark 2.8 and the the last column of the repartition table in Example 2.12.

Var(X) the variance of X is called the *total variance*. Generally speaking, **an important formula which connects the total variance** Var(X) **to** Var(X$_1$) **and** Var(X$_2$) **the variances inside of two underlying groups of sizes** N_1 **and** N_2 **is:**

$$\begin{aligned}\mathrm{Var(X)} &= \overbrace{\underbrace{\Big(\frac{N_1}{N_1+N_2}\Big)\mathrm{Var(X_1)} + \Big(\frac{\mathrm{N_2}}{\mathrm{N_1+N_2}}\Big)\mathrm{Var(X_2)}}_{\textbf{Within-groups Variance}}}^{\textbf{weighted arithmetic mean of } \mathrm{Var(X_1)} \textbf{ and } \mathrm{Var(X_2)}} \qquad (2.11)\\ &+ \overbrace{\underbrace{\Big(\frac{N_1}{N_1+N_2}\Big)\big(\overline{x}_1-\overline{x}\big)^2 + \Big(\frac{N_2}{N_1+N_2}\Big)\big(\overline{x}_2-\overline{x}\big)^2}_{\textbf{Between-groups Variance}}}^{\textbf{variance of the variable with values } \overline{x}_1 \textbf{ and } \overline{x}_2}\end{aligned}$$

Thus in the case of Example 2.12 one gets that

$$\mathrm{Var(X)} \simeq \left(\frac{13}{25}\times 3,60 + \frac{12}{25}\times 27,96\right) + \left(\frac{13}{25}\big(10,31-10,32\big)^2 + \frac{12}{25}\big(10,33-10,32\big)^2\right) \simeq 15,30$$

and consequently that the standard deviation $\sigma_X = \sqrt{15,30} \simeq 3,91$. Notice that Var(X) can also be computed in a direct way by using the formula (2.7), or the formula (2.6), and the last column of the repartition table in Example 2.12.

d) Average absolute deviation

The average absolute deviation from the mean of the discrete quantitative variable X is the arithmetic mean of the absolute values of the differences between the values of X and its arithmetic mean:

$$d_{\overline{x}} = \frac{1}{N}\sum_{i=1}^{N}|x_i-\overline{x}| = \sum_{i=1}^{K} f_i|x_i-\overline{x}|;$$

recall that $x_1,\ldots,x_K$ are assumed to be the distinct values of the variable X, and that $f_i = n_i/N$ is the frequency of the value x_i.

Example 2.13. *We are going to calculate the average absolute deviation from the mean $d_{\overline{x}}$ in the case of the variable "Grades" which was already studied in Example 2.3; recall that for this variable the arithmetic mean $\overline{x} \simeq 10,73$. One has*

Individuals	*Grades*	$(x_i - \overline{x})$	$\|x_i - \overline{x}\|$
Michel	*12*	*1,27*	*1,27*
Jean	*8*	*-2,73*	*2,73*
Stéphane	*13*	*2,27*	*2,27*
Charles	*11*	*0,27*	*0,27*
Agnès	*10*	*-0,73*	*0,73*
Nadine	*9*	*-1,73*	*1,73*
Étienne	*16*	*5,27*	*5,27*
Gilles	*14*	*3,27*	*3,27*
Aurélie	*11*	*0,27*	*0,27*
Stéphanie	*15*	*4,27*	*4,27*
Marie-Claude	*4*	*-6,73*	*6,73*
Anne	*18*	*7,27*	*7,27*
Christophe	*12*	*1,27*	*1,27*
Pierre	*6*	*-4,73*	*4,73*
Bernadette	*2*	*-8,73*	*8,73*
			Total=50,81

Thus, one obtains that

$$d_{\overline{x}} \simeq \frac{50,81}{15} \simeq 3,39$$

Remark 2.9. *One always has $d_{\overline{x}} \leq \sigma_X$. In other words, one always has*

(Average absolute deviation **from the mean***) $\leq$ (Standard deviation).*

The average absolute deviation from the median of the discrete quantitative variable X is the arithmetic mean of the absolute values of the differences between the values of X and its median M_e:

$$d_{M_e} = \frac{1}{N}\sum_{i=1}^{N} |x_i - M_e| = \sum_{i=1}^{K} f_i |x_i - M_e|;$$

recall that $x_1, \dots, x_K$ are assumed to be the distinct values of the variable X, and that $f_i = n_i/N$ is the frequency of the value x_i.

Problem 2.8. *Calculate the average absolute deviation from the median d_{Me} in the case of the variable "Grades" which was already studied in Example 2.3; recall that for this variable the median $Me = 11$.*

3 Relationship between two quantitative variables

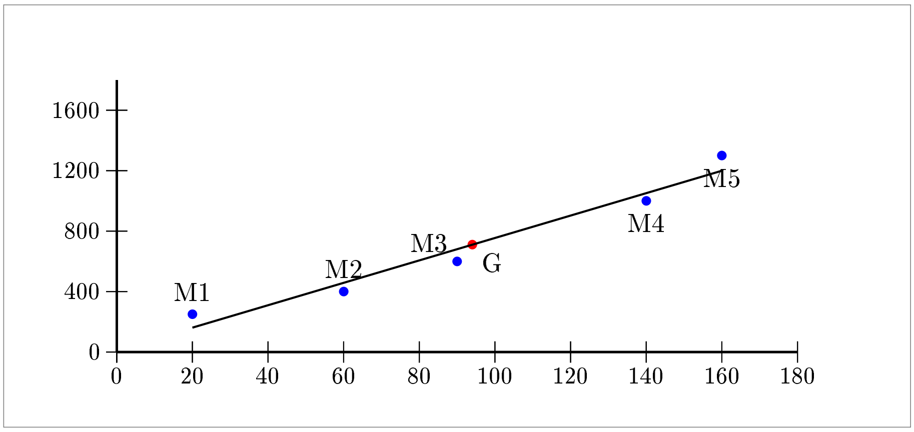
3.1 Simple linear regression

Example 3.1. *One wishes to study the relationship surface-price of* 5 *apartments in Paris. The quantitative variable X denotes the surface in m^2 (square meters), and the quantitative variable Y the selling price in thousands of Euros. The values of these two variables for the* 5 *apartments are given in the following table:*

Data table

X (in m^2)	$x_1 = 20$	$x_2 = 60$	$x_3 = 90$	$x_4 = 140$	$x_5 = 160$
Y (in thousands of Euros)	$y_1 = 250$	$y_2 = 400$	$y_3 = 600$	$y_4 = 1000$	$y_5 = 1300$

First one represents the two variables X and Y by **a cloud of points**: in an orthogonal system of axes, each couple of observed values (x_i, y_i) corresponding to an apartment (or individual) i is represented by a point M_i with coordinates (x_i, y_i). The shape of the cloud gives information on the type of the possible link between the variables X and Y.



In our present case the cloud seems more or less to follow a straight line. Thus, one can consider that there is **a linear relationship** between the surface of an apartment and its selling price. More precisely, it seems reasonable that the relationship between the surface x_i of an apartment and its price y_i be nearly of the form $y_i = ax_i + b$. The coefficients (or parameters) a and b, which are respectively called the **slope** and the **y-intercept**, have to be chosen so that the line of equation $y = ax + b$ passes **"as nearest as possible of a majority of points of the cloud"**. Now, we are going to formalize this idea.

Let D be an oblique line of equation $y = ax + b$ and let Δ be the line parallel to the vertical axis and passing through the point M_i. The two lines Δ and D intersect in a point $M_i^{'}$; the distance between M_i and $M_i^{'}$ equals $|y_i - ax_i - b|$. The coefficients a and b are chosen so that the quantity

$$(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + (y_3 - ax_3 - b)^2 + (y_4 - ax_4 - b)^2 + (y_5 - ax_5 - b)^2$$

be the smallest possible.

More generally, let $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$ be the observed values of two quantitative variables X and Y on a sample of N individuals. The coefficients of the **the least squares regression line (or the least squares best-fit line)**, that is the line allowing to match as closely as possible, in the sense of the least error squares criterion, the cloud of points $M_1 = (x_1, y_1)\,;\, M_2 = (x_2, y_2)\,;\ldots;\, M_N = (x_N, y_N)$ are the two real numbers a and b which minimize the quantity

$$(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + \ldots + (y_N - ax_N - b)^2\,.$$

They are given by the two formulas:

$$a = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_N - \overline{x})(y_N - \overline{y})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_N - \overline{x})^2} \tag{3.1}$$

and

$$b = \overline{y} - a\overline{x}\,. \tag{3.2}$$

Notice that one knows from the formula (3.2) that the regression line passes through the **center of gravity** of the cloud of points $M_1 = (x_1, y_1)\,;\, M_2 = (x_2, y_2)\,;\ldots;\, M_N = (x_N, y_N)$, that is the point G with coordinates $(\overline{x}, \overline{y})$ where $\overline{x}$ and $\overline{y}$ are the arithmetic means of the variables X and Y.

Having determined a and b, for each $i = 1, 2, \ldots, N$, one sets:

$$\widehat{y_i} = a x_i + b\,. \tag{3.3}$$

This quantity $\widehat{y_i}$ is called the **estimated value of Y by the regression equation when X equals** x_i. When the approximation of a cloud of points by a regression line is of good quality, for a majority of individuals i the estimated value $\widehat{y_i}$ is close to y_i the observed (or real) value of Y when X equals x_i.

Problem 3.1. *Prove that the arithmetic mean of $y_1, y_2, \ldots, y_N$ is always equal to that of $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_N$.*

Let us now show how the formulas (3.1), (3.2) and (3.3) can be applied to the data in Example 3.1. The arithmetic mean of the surfaces of the 5 apartments is $\overline{x} = \frac{470}{5} = 94\,m^2$, and the arithmetic mean $\overline{y}$ of their prices is $\overline{y} = \frac{3550}{5} = 710$ thousands of Euros. Thus, the coordinates of the center of gravity G are $(94\,, 710)$.

In order to calculate the values of the two coefficients a and b it is convenient to use the following table.

$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
-74	-460	34040	5476	211600
-34	-310	10540	1156	96100
-4	-110	440	16	12100
46	290	13340	2116	84100
66	590	38940	4356	348100
		Total = 97300	Total = 13120	Total = 752000

(3.4)

Thus it follows from the formulas (3.1) and (3.2) that:

$$a = \frac{97300}{13120} \simeq 7,416 \quad \text{and} \quad b = 710 - 7,416 \times 94 \simeq 12,896\,. \tag{3.5}$$

Hence the equation of the regression line is:

$$y = 7,416\,x + 12,896\,.$$

At last, let us calculate $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_5$ the estimated prices (in thousands of Euros) by the regression equation of the 5 apartments. Using (3.3) and (3.5), one obtains that: $\widehat{y_1} = 7,416 \times 20 + 12,896 \simeq 161$; $\widehat{y_2} = 7,416 \times 60 + 12,896 \simeq 458$; $\widehat{y_3} = 7,416 \times 90 + 12,896 \simeq 680$; $\widehat{y_4} = 7,416 \times 140 + 12,896 \simeq 1051$ and $\widehat{y_5} = 7,416 \times 160 + 12,896 \simeq 1199\,.$

The values of X and the corresponding observed and estimated values of Y are recorded in the following table:

X (in m^2)	$x_1 = 20$	$x_2 = 60$	$x_3 = 90$	$x_4 = 140$	$x_5 = 160$
Observed values of Y (in thousands of Euros)	$y_1 = 250$	$y_2 = 400$	$y_3 = 600$	$y_4 = 1000$	$y_5 = 1300$
Estimated values of Y (in thousands of Euros)	$\widehat{y_1} = 161$	$\widehat{y_2} = 458$	$\widehat{y_3} = 680$	$\widehat{y_4} = 1051$	$\widehat{y_5} = 1199$

Remark 3.1. *Generally speaking $TV(Y)$, $EV(Y)$ and $RV(Y)$,* **the total variation, the explained variation and the residual variation of the variable** Y *are defined as:*

$$TV(Y) = (y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \ldots + (y_N - \overline{y})^2 = N \times \mathrm{Var(Y)}\,, \tag{3.6}$$

$$EV(Y) = (\widehat{y}_1 - \overline{y})^2 + (\widehat{y}_2 - \overline{y})^2 + \ldots + (\widehat{y}_N - \overline{y})^2\,, \tag{3.7}$$

and

$$VR(Y) = (y_1 - \widehat{y}_1)^2 + (y_2 - \widehat{y}_2)^2 + \ldots + (y_N - \widehat{y}_N)^2\,. \tag{3.8}$$

The following important equality, which is reminiscent of the Pythagorean theorem, is always satisfied:

$$TV(Y) = EV(Y) + RV(Y)\,. \tag{3.9}$$

The coefficient of determination *is denoted by $R^2(X,Y)$ (pronounced "R squared") and defined as the ratio:*

$$R^2(X,Y) = \frac{EV(Y)}{TV(Y)}\,. \tag{3.10}$$

It results from (3.9) that $R^2(X,Y)$ is **always between** 0 **and** 1**.**

Problem 3.2. *Calculate $TV(Y)$, $EV(Y)$, $RV(Y)$ and $R^2(X,Y)$ for the data in Example 3.1.*

3.2 Covariance and linear correlation coefficient

It is always possible to draw the regression line whatever the shape of the cloud of points $M_1 = (x_1, y_1)\,;\, M_2 = (x_2, y_2)\,;\ldots;\, M_N = (x_N, y_N)$ might be. Yet, this line does not always provide a good approximation of the cloud, as for instance when it has a rather circular form. The first thing one has to make for studying the quality of the approximation of a cloud of points by a regression line is to check the **linear correlation coefficient of the variables** X **and** Y denoted by $R(X,Y)$. In order to define the latter coefficient, one first needs to introduce the **covariance of** X **and** Y denoted by $\mathrm{Cov(X,Y)}$ and defined as:

$$\mathrm{Cov(X,Y)} = \frac{(\mathrm{x_1} - \overline{\mathrm{x}})(\mathrm{y_1} - \overline{\mathrm{y}}) + (\mathrm{x_2} - \overline{\mathrm{x}})(\mathrm{y_2} - \overline{\mathrm{y}}) + \ldots + (\mathrm{x_N} - \overline{\mathrm{x}})(\mathrm{y_N} - \overline{\mathrm{y}})}{\mathrm{N}}\,, \tag{3.11}$$

where $\overline{x}$ and $\overline{y}$ denote the arithmetic means of the variables X and Y, whose values are $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$. Observe that it follows from (3.11) and (2.6) that

$$\mathrm{Cov(X,X)} = \mathrm{Var(X)}\,.$$

Similarly to the variance, the covariance of X and Y can as well be calculated by using the following formula (sometimes called Huygens formula):

$$\mathrm{Cov(X,Y)} = \left(\frac{\mathrm{x_1y_1} + \mathrm{x_2y_2} + \ldots + \mathrm{x_Ny_N}}{\mathrm{N}}\right) - \overline{\mathrm{x}}\,\overline{\mathrm{y}}\,. \tag{3.12}$$

In fact the formula (2.7) is nothing else than the formula (3.12) in the particular case where $X = Y$.

Example 3.2. *Let X and Y be the variables "Surfaces" and "Prices" issued from the data in Example 3.1. We are going to present two methods for calculating* $\mathrm{Cov(X,Y)}$*, the first consists in using the formula (3.11), and the second the formula (3.12).*

First method: *One knows from the table (3.4) that*

$$(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+(x_3-\overline{x})(y_3-\overline{y})+(x_4-\overline{x})(y_4-\overline{y})+(x_5-\overline{x})(y_5-\overline{y})=97300\,;$$

thus, it results from (3.11) that:

$$\mathrm{Cov(X,Y)}=\frac{97300}{5}=19460\,.$$

Second method: *For calculating the sum* $x_1y_1+x_2y_2+x_3y_3+x_4y_4+x_5y_5$ *, one uses the following table:*

x_i	y_i	x_iy_i
20	*250*	*5000*
60	*400*	*24000*
90	*600*	*54000*
140	*1000*	*140000*
160	*1300*	*208000*
		Total = 431000

which allows to find that: $x_1y_1+x_2y_2+x_3y_3+x_4y_4+x_5y_5=431000$*. Thus, one obtains that:*

$$\frac{x_1y_1+x_2y_2+x_3y_3+x_4y_4+x_5y_5}{5}=\frac{431000}{5}=86200\,. \tag{3.13}$$

On the other hand, one knows from Sub-section 3.1 that $\overline{x}=94$ *and* $\overline{y}=710$*, which entails that:*

$$\overline{x}\,\overline{y}=94\times 710=66740\,. \tag{3.14}$$

Finally, putting together (3.12), (3.13) and (3.14), one gets that:

$$\mathrm{Cov(X,Y)}=86200-66740=19460\,.$$

Remark 3.2. (Cauchy-Schwarz inequality) *The absolute value of the covariance of two quantitative variables* X *and* Y *is always less than or equal to the product of their standard deviations:*

$$|\mathrm{Cov(X,Y)}|\leq\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}\,.$$

This inequality can equivalently be expressed as:

$$-\sigma_X\sigma_Y\leq\mathrm{Cov(X,Y)}\leq\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}.$$

Let us check that the Cauchy-Schwarz inequality is satisfied in the particular case of the data of Example 3.1. In this particular case, one already knows that $\mathrm{Cov(X,Y)}=19460$, and it remains to us to calculate the standard deviations σ_X and σ_Y. One knows from the table (3.4) that:

$$(x_1-\overline{x})^2+(x_2-\overline{x})^2+(x_3-\overline{x})^2+(x_4-\overline{x})^2+(x_5-\overline{x})^2=13120$$

and

$$(y_1-\overline{y})^2+(y_2-\overline{y})^2+(y_3-\overline{y})^2+(y_4-\overline{y})^2+(y_5-\overline{y})^2=752000\,.$$

Thus, one can derive from the formula (2.6) that $\mathrm{Var(X)}=\frac{13120}{5}=2624$ and $\mathrm{Var(Y)}=\frac{752000}{5}=150400$, hence $\sigma_X=\sqrt{2624}\simeq 51,22$ and $\sigma_Y=\sqrt{150400}\simeq 387,81$. Combining the previous results, one can check that the Cauchy-Schwarz inequality is satisfied in the particular case of the data in Example 3.1; indeed one has:

$$19460 = |\mathrm{Cov(X, Y)}| \leq \sigma_{\mathrm{X}}\sigma_{\mathrm{Y}} \simeq 51,22 \times 387,81 \simeq 19863,63\,.$$

The linear correlation coefficient of the two variables X **and** Y is denoted by $R(X,Y)$ and defined as:

$$R(X,Y) = \frac{\mathrm{Cov(X, Y)}}{\sigma_X \sigma_Y}\,. \tag{3.15}$$

Thus, for the data in Example 3.1 one has:

$$R(X,Y) \simeq \frac{19460}{51,22 \times 387,81} \simeq 0,979\,.$$

Remark 3.3. (Important properties of linear correlation coefficient)

(i) *It results from the Cauchy-Schwarz inequality that* $R(X,Y)$ *is always between* -1 *and* $+1$.

(ii) *When one multiplies with itself the linear correlation coefficient (defined in (3.15)) one obtains the coefficient of determination (defined in (3.10)). This is why the coefficient of determination is denoted by* $R^2(X,Y)$.

(iii) *The coefficient* a *(the slope) of the regression line is given by:*

$$a = \frac{\mathrm{Cov(X, Y)}}{\mathrm{Var(X)}} = \frac{\sigma_Y}{\sigma_X} R(X,Y)\,. \tag{3.16}$$

It is worth mentioning that the last equality in (3.16) implies that **the slope** a **of the regression line and the linear correlation coefficient** $R(X,Y)$ **are always of the same sign**.

Remark 3.4. (Interpretation of linear correlation coefficient $R(X,Y)$**)**

(i) ***When*** $R(X,Y)$ ***is close to*** 0*, there is no correlation between* X *and* Y*; the regression line does not provide a good approximation of the cloud of points associated with* X *and* Y.

(ii) ***When*** $R(X,Y)$ ***is close to*** $+1$*, there is a direct correlation between* X *and* Y*; which roughly speaking means that* Y *increases when* X *increases, and that* X *increases when* Y *increases.*

(iii) ***When*** $R(X,Y)$ ***is close to*** -1*, there is an inverse correlation between* X *and* Y*; which roughly speaking means that* Y *increases when* X *decreases, and that* X *decreases when* Y *increases.*

Before ending this section let us emphasize that:

Remark 3.5. *In order that the approximation of a cloud of points by a regression line to be of good quality, a necessary condition is that the corresponding linear correlation coefficient be close to* $+1$ *or* -1*. But this is not a sufficient condition. In other words, the approximation might be of bad quality even when the linear correlation coefficient is close to* $+1$ *is* -1*. For being nearly sure that it is of good quality, one has to make use of tests (Student and Fisher-Snedecor) whose presentation falls beyond the scope of this course.*

4 Problems on univariate statistical analysis

Problem 4.1. *Results of a statistical survey on the monthly incomes in hundreds of euros of* 105 *French households are summarized in the following table:*

Classes of Incomes	$[10;13]$	$]13;16]$	$]16;19]$	$]19;22]$	$]22;25]$	$]25;28]$
Sizes	$n_1 = 15$	$n_2 = 21$	$n_3 = 28$	$n_4 = 21$	$n_5 = 12$	$n_6 = 8$

1) Calculate in terms of percentages the frequencies and the cumulative frequencies of the classes of the variable "Incomes".
2) Draw the histogram of the variable "Incomes".
3) Determine the modal class of this variable and the accurate value of the mode in a graphical way.
4) Determine the accurate value of the mode by means of calculations.
5) Draw the graph of the cumulative function F of this variable and explain the method allowing to draw it.
6) Determine the median of the variable "Incomes" by using three different methods.
7) Calculate the quartiles of this variable and determine the interquartile range.
8) Calculate the arithmetic mean of this variable.
9) Calculate the variance and the standard deviation of this variable.

Problem 4.2. *The following table concerns the audit times in minutes for 50 balance sheets.*

Classified variable "Audit times" (in minutes)	*[10, 20]*	*]20,30]*	*]30,40]*	*]40,50]*	*]50,60]*
Number of balance sheets	*3*	*5*	*10*	*12*	*20*

1) Calculate in terms of percentages the frequencies and the cumulative frequencies of the classes of the variable "Audit times".
2) Draw the histogram of the variable "Audit times".
3) Determine the modal class of this variable and the accurate value of the mode in a graphical way.
4) Determine the accurate value of the mode by means of calculations.
5) Draw the graph of the cumulative function F of this variable and explain the method allowing to draw it.
6) Determine the quartiles of the variable "Audit times" by using three different methods.
7) Calculate the interquartile range of this variable and explain why it is useful.
8) Calculate the arithmetic mean of this variable.
9) Calculate the variance and the standard deviation of this variable.

Problem 4.3. *The following table concerns the distribution of the net monthly incomes in hundreds of euros of* 153 *former students of some business school.*

Classified variable "Incomes"	*[20, 25]*	*]25,30]*	*]30,35]*	*]35,40]*	*]40,45]*	*]45,50]*
Number of individuals	*13*	*33*	*45*	*39*	*14*	*9*

1) Calculate in terms of percentages the frequencies and the cumulative frequencies of the classes of the variable "Incomes".

2) Draw the histogram of the variable "Incomes".
3) Determine the modal class of this variable and the accurate value of the mode in a graphical way.
4) Determine the accurate value of the mode by means of calculations.
5) Draw the graph of the cumulative function F of this variable and explain the method allowing to draw it.
6) Determine the median of the variable "Incomes" by using three different methods.
7) Calculate the quartiles of this variable and determine the interquartile range.
8) Calculate the arithmetic mean of this variable.
9) Calculate the variance and the standard deviation of this variable.

Problem 4.4. *The three questions of the problem can be treated independently of one another.*
1) Calculate the geometric mean and the harmonic mean of the following four numbers: $x_1 = 2,1$; $x_2 = 4,3$; $x_3 = 5,7$ and $x_4 = 7,3$.
2) The geometric and arithmetic means of six strictly positive numbers have been calculated by a student during his statistic exam. He have found that the geometric mean equals $10,22$ and the arithmetic mean equals $7,32$. What do you think of that ? (justify your answer)
3) The turnover of the entreprise "Durand" increased by 80% during the year 2002 and it increased by 30% during the year 2003. Calculate the average rate of increase per year of this turnover for the period of the two years 2002 and 2003.

Problem 4.5. (issued from an exam of L2 (2014)) *A statistical survey was realized on a sample of 170 travelers. It concerned the waiting time in minutes of each one of them for accessing to a desk of a train station during rush hours. The results of the survey are summarized in the following table in which the classified variable T denotes waiting time.*

Classes of T	$0 \leq T \leq 5$	$5 < T \leq 10$	$10 < T \leq 15$	$15 < T \leq 20$
Number of travelers	$n_1 = 45$	$n_2 = 51$	$n_3 = 54$	$n_4 = 20$

1) a) What is the modal class of the variable T ?
b) Draw the histogram of the variable T.
c) Determine the accurate value of the mode in a graphical way.
2) Calculate the arithmetic mean of the variable T, its variance and its standard deviation.
3) a) Calculate the frequencies of the classes of the variable T.
b) Calculate the cumulative sizes of the classes of the variable T.
c) Calculate the cumulative frequencies of the classes of the variable T.
4) Explain the method which allows to draw the graph of the cumulative function of the variable T, and draw this graph.
5) Determine in a graphical way approximative values for the median Me and the third quartile Q_3 of the variable T.
6) Determine more precise values for Me and Q_3 by using calculations.

Problem 4.6. (issued from an exam of L2 (2015)) *The two questions of the problem can be treated independently of one another (you are asked to give your numerical results with not more than two decimal places).*
1) A motorist has first driven 50 kilometers at the speed 60 km/h, then 90 kilometers at the speed 120 km/h, and finally 10 kilometers at the speed 50 km/h. One denotes by v_m his average speed in km/h for the whole journey of 150 kilometers. Calculate v_m.
2) Let X be a discrete quantitative variable whose arithmetic mean and standard deviation are respectively equal to $10,34$ and $4,51$. Calculate the quadratic mean of X.

Problem 4.7. (issued from an exam of L2 (2016)) *A home appliances manufacturing company has made a statistical study on lifetimes (measured in years) on a sample of 133 washing machines of some model. A summary of this study is given in the following table in which the classified variable D denotes lifetime.*

Classes of D	$0 \leq D \leq 2$	$2 < D \leq 4$	$4 < D \leq 6$	$6 < D \leq 8$	$8 < D \leq 10$
Numbers of machines	$n_1 = 12$	$n_2 = 15$	$n_3 = 61$	$n_4 = 27$	$n_5 = 18$

1) Calculate the cumulative sizes of the classes of the variable D.
2) Calculate the frequencies and the cumulative frequencies of the classes of the variable D (the results are to be given in terms of percentages).
3) Calculate the arithmetic mean of the variable D.
4) Calculate the variance of the variable D and its standard deviation.

Problem 4.8. (issued from an exam of L2 (2017)) *A home appliances manufacturing company has made a statistical study on lifetimes (measured in years) on a sample of 233 washing machines of some model. A summary of this study is given in the following table in which the classified variable V denotes lifetime.*

Classes of V	$0 \leq V \leq 2$	$2 < V \leq 4$	$4 < V \leq 6$	$6 < V \leq 8$	$8 < V \leq 10$
Numbers of machines	$n_1 = 32$	$n_2 = 35$	$n_3 = 81$	$n_4 = 47$	$n_5 = 38$

1) Determine the modal class of V.
2) Calculate the quartiles of this variable and determine the interquartile range.

Problem 4.9. (issued from an exam of L2 (2019)) *Between the 1-st of January and the 30-th of April 2018 John recorded each day at 9 am the temperature in degrees Celsius displayed by the thermometer at the entrance of his garden. The repartition of his 120 temperature recordings is presented in the following table in which the classified variable T denotes the temperature in degrees Celsius.*

Classes	$-5 \leq T \leq 0$	$0 < T \leq 5$	$5 < T \leq 10$	$10 < T \leq 15$	$15 < T \leq 20$
Number of days	*11*	*20*	*46*	*24*	*19*

1) Calculate the frequencies, the cumulative sizes and the cumulative frequencies of the classes of the variable T.
2) Explain the method which allows to draw the graph of the cumulative function of the variable T, and draw this graph.
3) Determine in a graphical way approximative values for the three quartiles Q_1, Q_2 and Q_3 of T.
4) Determine more precise values for Q_1, Q_2 and Q_3 by means of calculations.
5) Calculate the interquartile range and explain why it is useful.
5) Calculate the arithmetic mean, the variance and the standard deviation of T.

Problem 4.10. (issued from an exam of L3M2S (2019)) *The distribution of the amounts in euros of the invoices of 1756 persons for their purchases in an hypermarket is presented in the following table:*

Classified variable "Amounts"	*[0 ; 30]*	*]30 ; 60]*	*]60 ; 100]*	*]100 ; 200]*	*]200 ; 300]*
Number of persons	*503*	*640*	*414*	*135*	*64*

1) Calculate the three quartiles and the interquartile range of the classified variable "Amounts".
2) Calculate the arithmetic mean, the variance and the standard deviation of this variable.

3) One respectively denotes by h_1 and h_5 the heights of the two rectangles of the histogram of this variable which correspond to the two classes $[0\,;30]$ and $[200\,;300]$. Calculate the ratio h_1/h_5.

Problem 4.11. *We are interested in a population of 8 persons which is divided in two groups: a group of five women whose weights in kilograms are 77 56 65 55 62, and a group of three men whose weights in kilograms are 88 102 77.*
1) Calculate the arithmetic mean of the variable "Weights" inside of each group, and its total arithmetic mean.
2) Calculate the within-groups variance and the between-groups variance related with the variable "Weights".
3) Calculate the total variance of the variable "Weights" by making use of the within-groups variance and the between-groups variance.
4) Calculate the total variance of the variable "Weights" by using another method.

5 Problems on simple linear regression

Problem 5.1. *In order to determine the selling price (expressed in Euros) of a new product a store conducted a survey on a sample of its customers. Its results are summarized in the following table.*

X	3	$3,5$	4	$4,5$	5
Y	20	18	13	9	7

The variable X denotes the price (expressed in Euros) proposed to the customers, and the variable Y denotes the number of customers ready to buy the product at this price.
1) Calculate $\overline{x}$ and $\overline{y}$ the arithmetic means of X and Y.
2) Calculate the covariance of X and Y.
3) a) Calculate the quadratic means of X and Y.
b) Calculate the variances of X and Y.
c) Calculate the standard deviations of X and Y.
d) Calculate the linear correlation coefficient of X and Y and comment your result.
3) Determine the equation of the regression line.

Problem 5.2. *In the following table are recorded the volumes of sales (expressed in thousands of products) of a product P for 4 consecutive trimesters, and the numbers of visits to trade customers made during these same trimesters by sales representatives for advertising the product.*

	Trimester 1	*Trimester 2*	*Trimester 3*	*Trimester 4*
Numbers of visits	26	27	31	30
Volumes of sales	53	68	79	69

The statistical variable corresponding to the "Numbers of visits" is denoted by X and the one corresponding to the "Volumes of sales" is denoted by Y.
1) a) Draw the cloud of points associated to the variables X and Y.
b) Determine the coordinates of the center of gravity (denoted by G) of this cloud of points and represent G.
2) a) Calculate $Cov(X,Y)$ the covariance of X and Y.
b) Calculate σ_X and σ_Y the standard deviations of X and Y.
c) Calculate $R(X,Y)$ the linear correlation coefficient of X and Y, and comment your result.
3) a) Does the regression line necessarily pass through the center of gravity G of the cloud of points associated to X and Y ?

b) Determine the equation of the regression line and draw it.

Problem 5.3. *A farmer wishes to quantify for some piece of land of* 1 *hectare (that is* 10000 m^2*) the relationship between the quantity (measured in kilograms) of fertilizer denoted by X and the quantity (measured in quintals,* 1 *quintal equals to* 100 *kilograms) of agricultural production denoted by Y. He has the following data table:*

X	100	200	300	400	500	600	700
Y	40	50	50	70	65	65	80

1) a) Draw the cloud of points associated to the variables X and Y.
b) Determine the coordinates of the center of gravity (denoted by G) of this cloud of points and represent G.
2) a) Calculate the covariance of X and Y.
b) Calculate the variance of X and that of Y.
c) Calculate the standard deviation of X and that of Y.
d) Calculate the linear correlation coefficient of X and Y and interpret your result.
e) Determine the equation of the regression line and represent it. Does the regression line necessarily pass through G ?
f) For each value x_i of the variable X calculate $\widehat{y}_i$ the corresponding estimated value of Y.
3) a) Calculate the average absolute deviation from the mean of the variable X.
b) Calculate the average absolute deviation from the mean of the variable Y.

Problem 5.4. *Eight married couples are denoted by the letters A, B, C, D, E, F, H and I. The variable X corresponds to age of husband and the variable Y to age of wife, X and Y are expressed in numbers of years. The values of these 2 variables for each of the 8 couples are given in the following table:*

couple	A	B	C	D	E	F	H	I
X (age of the husband)	33	43	23	37	35	40	28	26
Y (age of the wife)	30	39	22	34	31	32	25	30

1) a) Calculate the median of the variable X.
b) Calculate the average absolute deviation from the median of the variable X.
2) a) Draw the cloud of points associated to the variables X and Y.
b) Determine the coordinates of the center of gravity (denoted by G) of this cloud of points and represent G.
3) Calculate the standard deviations of X and Y.
4) Calculate the linear correlation coefficient of X and Y and interpret your result.
5) Determine the equation of the regression line and represent it. Does the regression line necessarily pass through G ?
6) Calculate the estimated value of Y for the couple D.

Problem 5.5. (issued from an exam of L2 (2014)) *In order to determine the optimal selling price of a toothpaste a statistical survey was conducted by a department store on a sample of 100 consumers. Its results are summarized in the following table in which the variable X corresponds to the price in euros of the toothpaste, and the variable Y corresponds to the number of consumers ready to buy this product at this price.*

X	3,5	2,5	1	1,5	3	2
Y	3	21	96	63	8	31

1) Determine the median of the variables X and Y.
2) Determine the ranges r_X and r_Y of the variables X and Y.
3) Calculate the average absolute deviations from the medians of the variables X and Y.
4) a) Draw the cloud of points associated to the variables X and Y.
b) Determine the coordinates of the center of gravity (denoted by G) of this cloud of points and represent G.
5) Calculate the quadradic means of the variables X and Y.
6) Calculate the variances and the standard deviations of the variables X and Y.
7) Calculate the covariance of the variables X and Y.
8) Calculate the linear correlation coefficient of the variables X and Y, and interpret your result.
9) Determine the equation of the regression line and represent it. Does the regression line necessarily pass through G ?
10) Imagine that the price of the toothpaste is 1,75 euros, then what would be the expected number of consumers (in the sample) who are ready to buy the toothpaste at this price ? (justify your answer in a precise way)

Problem 5.6. (issued from an exam of L2 (2015)) *The variable X denotes the "Number of the years of study after the bachelor degree", and the variable Y denotes the "salary" expressed in thousands of euros. The values of these two variables for a sample of eight employees, denoted by the letters A, B, C, D, E, F, H and I, are given in the following table:*

Employee	*A*	*B*	*C*	*D*	*E*	*F*	*H*	*I*
X	*4*	*5*	*3*	*0*	*8*	*2*	*1*	*5*
Y	*1,93*	*2,35*	*1,75*	*1,39*	*2,99*	*1,64*	*1,54*	*2,13*

You are asked to give your numerical results with not more than two decimal places.
1) a) Find the median and the arithmetic mean of X.
b) Find the median and the arithmetic mean of Y.
2) Calculate the covariance of X and Y.
3) Determine the equation of the regression line.
4) One denotes by M the point whose abscissa is the median of X and whose ordinate is the median of Y. Does the regression line pass through M ?
5) Draw the cloud of points associated to X and Y, and draw the regression line.
6) Calculate for each of the eight employees the estimated value of his salary resulting from the regression.
7) Calculate the linear correlation coefficient of X and Y, and interpret your result.

Problem 5.7. (issued from an exam of L2 (2016)) *The time (measured in months) between the date of birth of a baby and the date at which he begins to speak (a little bit) is called* the age of the first word of a baby. *In order to determine whether their exists a relationship between the age of the first word of a baby and his score at an IQ test called the* Gesell test, *one focuses on a sample of 9 babies denoted by A, B, C, D, E, F, H, I and J. In the following table, the values of the quantitative variable X are the ages of the first word of these babies, and the values of the quantitative variable Y are their scores at the Gesell test.*

Baby	*A*	*B*	*C*	*D*	*E*	*F*	*H*	*I*	*J*
X	*15*	*26*	*20*	*9*	*12*	*10*	*17*	*8*	*18*
Y	*95*	*71*	*87*	*96*	*105*	*100*	*121*	*104*	*93*

1) Determine the medians of the variables X and Y.
2) Calculate the average absolute deviations from the medians of the variables X and Y.
3) Calculate the arithmetic means of the variables X and Y.

4) Calculate the covariance of the variables X and Y.
5) Calculate the quadratic means of the variables X and Y.
6) Calculate the variances and the standard deviations of the variables X and Y.
7) Calculate the linear correlation coefficient of the variables X and Y
8) Determine the equation of the regression line.

Problem 5.8. (issued from an exam of L2 (2017)) *Seven bus connections proposed by a bus company are denoted by the letters A, B, C, D, E, F and K. In the following table, the quantitative variable X consists in the distances measured in kilometers corresponding to these connections, and the quantitative variable Y consists in the prices of the corresponding bus tickets.*

Connection	*A*	*B*	*C*	*D*	*E*	*F*	*K*
X	*75*	*15*	*117*	*93*	*41*	*25*	*35*
Y	*25*	*5*	*35*	*30*	*20*	*10*	*15*

1) a) Draw the cloud of points associated to the variables X and Y.
b) Determine the coordinates of the center of gravity (denoted by G) of this cloud of points and represent G.
2) a) Calculate the covariance of X and Y.
b) Calculate the quadratic means of X and Y.
c) Calculate the variances of X and Y.
d) Calculate the standard deviations of X and Y.
e) Calculate the linear correlation coefficient of X and Y, and interpret your result.
3) Determine the equation of the regression line and represent it. Does the regression line necessarily pass through G ?
4) For each value x_i of the variable X calculate $\widehat{y}_i$ the corresponding estimated value of Y.
5) Determine the medians of X and Y.
6) Calculate the average absolute deviations from the medians of the variables X and Y.

Problem 5.9. (issued from an exam of L2 (2018)) *Eight electronic technicians are denoted by the letters A, B, C, D, E, F, H and I. In the following table, the variable X corresponds to their numbers of weeks work experience and the variable Y to the numbers of improper assemblies which were mistakenly made by each one of them.*

Thecnician	*A*	*B*	*C*	*D*	*E*	*F*	*H*	*I*
X	*7*	*9*	*6*	*14*	*4*	*2*	*1*	*8*
Y	*26*	*20*	*28*	*16*	*26*	*38*	*32*	*25*

1) Calculate the arithmetic means of the variables X and Y.
2) Calculate the quadratic means of these two variables.
3) By using the notion of "depth", calculate the median of these two variables.
4) Calculate the variances and the standard deviations of these two variables.
5) a) Calculate the average absolute deviation from the mean of the variable X.
b) Calculate the average absolute deviation from the median of the variable X.
c) Calculate the average absolute deviation from the mean of the variable Y.
d) Calculate the average absolute deviation from the median of the variable Y.
6) a) Calculate the covariance of the two variables X and Y.
b) Calculate the linear correlation coefficient of the two variables X et Y, and interpret your result.
7) a) Draw the cloud of points associated to the variables X and Y, and calculate the coordinates of the center of gravity of this cloud.

b) Determine the equation of the regression line and draw this line.
8) For each value x_i of the variable X calculate $\widehat{y}_i$ the corresponding estimated value of Y.
9) Calculate the total variation of Y by using as less calculations as possible.
10) Calculate the explained variation of Y.
11) Calculate the residual variation of Y by using as less calculations as possible.

Problem 5.10. ((issued from an exam of L2 (2019)) *The following table concerns 8 countries in Central America. The quantitative variable X consists in their urbanization rates* [3] *for the year 1985, and the variable Y consists in their birth rates* [4] *for the same year.*

Countries	Mexico	Cuba	Salvador	Haiti	Honduras	Trinidad and Tobago	Panama	Nicaragua
X	43,2	33,3	11,5	13,9	19,0	6,8	37,7	28,5
Y	33,9	16,9	40,2	41,3	43,9	24,6	28,0	44,2

1) Draw the cloud of points associated to the variables X and Y, and calculate the coordinates of the center of gravity of this cloud.
2) Calculate the quadratic means of these two variables.
3) Calculate the variances and the standard deviations of these two variables.
4) Calculate the covariance of these two variables.
5) Calculate the linear correlation coefficient of these two variables and interpret your result.
6) Determine the equation of the regression line and draw this line.
7) Calculate the coefficient of determination by using the linear correlation coefficient.
8) Calculate the explained and the residual variations of Y by using the coefficient of determination and the variance of Y.

[3] That is the percentage of the population living in cities of more than 100000 habitants
[4] That is the number of births per 1000 habitants