Probability and Statistics

Dr. ARRAR - Prof. REMITA
National Higher School of Artificial Inteligence

2022-2023

Program

Part 1 : Descriptive statistics

Part 2 : Introduction to probability calculus

The essential purpose of statistics is to facilitate decision-making under conditions of uncertainty.
There is no precise definition of statistics, but for us we will use the following definition:

Definition

We call statistics a set of methods or techniques used to analyze or process sets of observations that we will call data.

The methods used are based on mathematics and make extensive use of computer tools for their implementation.

Remark

We must not confuse statistics, which is the science that has just been defined, with a statistic, which is a set of numbers on a specific subject.

We distinguish three important phases in the evolution of statistics:

- From antiquity until the end of the 19th century, statistics remained mainly a set of enumeration.
- From the 19th century to the 1960s, mathematical statistics was built up school (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman ...).
- Since the 1960's, and with the development of computer and graphic tools, statistics graphics, statistics has undergone a considerable development.

- **Population:** set concerned by a statistical study (noted Ω).
- **Individual element or statistical unit (individu):** is one member of a set of entities being studied (noted $\omega \in \Omega$).
- **Sample (échantillon)**: subset of the population on which the observations are made.
- **Survey (enquête)**: operation consisting of observing (measuring, questioning, etc.) all the individuals in a sample.
- **Census (recensement)**: survey in which the sample observed is the entire population (exhaustive survey).
- **Poll (sondage):** survey in which the observed sample is a strict subset of the population (non-exhaustive survey).

- **Character**: It is a characteristic defined on the population and observed on the sample. The characteristic can be:
- **Modality or observations:** the different values taken by each character.

1. **Qualitative:** it is a character which is not measurable and it can be :
   - **nominal data** (gender, profession, family situations, ...)
   - **ordinal data** (military rank, grade in university, ...)
2. **Quantitative**: it is a characteristic that can be measured and we distiguish two cases :
   - **discrete** (number of children, number of rooms in an apartment, ...) also called **discrete statistical variable.**
   - **continuous** (height, age, speed, weight, rate, ...) also called **continuous statistical variable**.

To carry out a statistical study, it is advisable to follow the following procedure:
- First of all, we start by collecting the raw observations that we group in a table called data table.
- In many situations, the observations may be different and in large numbers, and the data table is not easily exploitable, so the observations are synthesized by grouping them in the statistical table.
- We complete the study by making graphical representations and numerical calculations in order to give a good interpretation of the observed data.

Schematically the statistical approach is summarized as follows:

Data table

Synthetize

Statistical table

Visualize

Graphical representations

Summarize

Numerical summaries

Example (1)

The study of the family situation of 30 employees of a company is summarized in the following table:

01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**
S	M	S	S	D	W	M	M	S	D
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
S	M	S	S	M	D	S	S	M	M
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
S	S	M	S	W	S	S	M	S	S

Example (2)

We study the number of children in a family in a city inhabited by 100 families. The following results were obtained:

01	**02**	**03**	**04**	**...**	**98**	**99**	**100**
2	0	4	2	...	6	3	1

Example (3)

We study the time taken by 30 workers to manufacture a given part. The following results were obtained:

01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**
56	60	60	65	67	70	71	73	73	74
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
75	76	77	77	77	78	78	78	78	78
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
79	81	81	82	82	83	84	85	87	90

We consider a sample of size n. Let X be the variable or the character studied,

- The **statistical series** is formed by the set of data $(x_i)_{i \leq n}$
- The **range (étendue)** of an ordered series is the number e defined by $e = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$; such that : $x_{\min}$ is the smallest observation that is $x_{(1)}$ and $x_{\max}$ is the greatest observation $x_{(n)}$.

Remark: A statistical series is always represented in increasing order.
In the discrete case, it often happens that observations are repeated, let n_i the number of occurences (or repetitions) of x_i;

- n_i is called **number or absolute frequency (effectif)** of x_i.
- Note $x_1, \cdots, x_k$ the observations of respective numbers $n_1, \cdots, n_k$.

- The **relative frequency (fréquence)** of the modality x_i is $f_i = \frac{n_i}{n}$.

Remark: *We have* $\sum_{i=1}^{k} n_i = n; 0 \leq f_i < 1$ and $\sum_{i=1}^{k} f_i = 1$

- The **percentage** of the modality x_i is $p_i = \frac{n_i}{n} \times 100$.
- The **increasing cumulative number** corresponding to the modality x_i is $n_i^c \nearrow = n_1 + n_2 + \cdots + n_i$.
- The **decreasing cumulative number** corresponding to the modality x_i is $n_i^c \searrow = n_k + n_{k-1} + \cdots + n_i$.

The presentation of a statistical series is done by a table, this presentation differs according to the nature of the studied character.

Example (1)

We take the example of the family situation of 30 employees.

X : Family situation	Numbers n_i
S	16
M	9
D	3
W	2
Total	30

Example (1)

We take the example of the family situation of 30 employees.

X : Family situation	Numbers n_i	Frequency f_i
S	16	0.5333
M	9	0.3000
D	3	0.1000
W	2	0.0667
Total	30	1

Example (1)

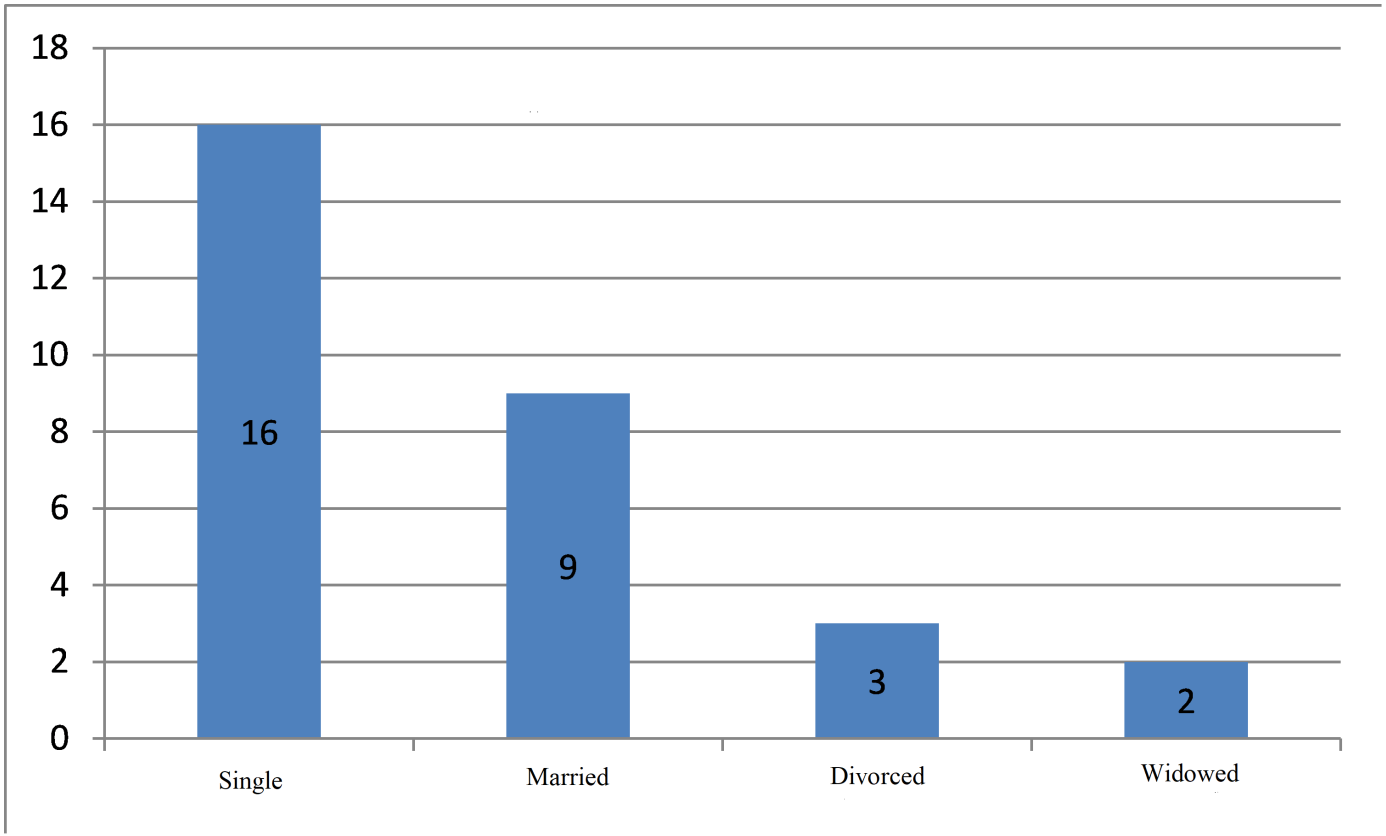
We take the example of the family situation of 30 employees.

X : Family situation	Numbers n_i	Frequency f_i	Percentage p_i
S	16	0.5333	53.33
M	9	0.3000	30
D	3	0.1000	10
W	2	0.0667	6.67
Total	30	1	100

Example (1)

We take the example of the family situation of 30 employees

Bar or pipe chart



Circular diagram (Pie chart)

Sector for "Single"

$$\begin{array}{c} \theta_S \longrightarrow 16 \\ 360 \longrightarrow 30 \end{array} \Longrightarrow \theta_S = \frac{16 \times 360}{30} = 192^\circ$$

Sector for "Married"

$$\begin{array}{c} \theta_M \longrightarrow 9 \\ 360 \longrightarrow 30 \end{array} \Longrightarrow \theta_M = \frac{9 \times 360}{30} = 108^\circ$$

Circular diagram (Pie chart)

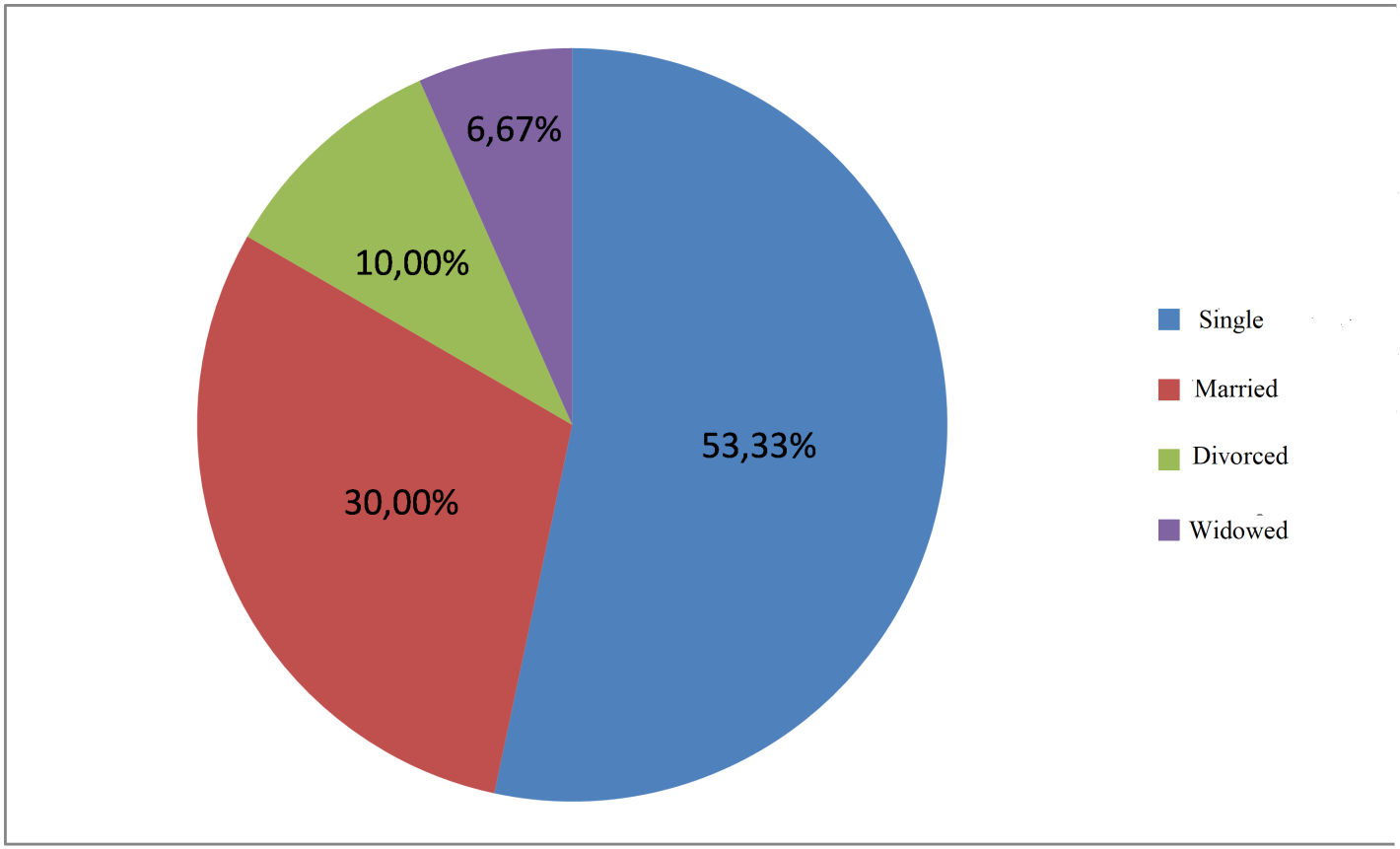
Sector for "Divorced"

$$\begin{array}{c} \theta_D \longrightarrow 3 \\ 360 \longrightarrow 30 \end{array} \Longrightarrow \theta_D = \frac{3 \times 360}{30} = 36^\circ$$

Sector for "Widow"

$$\begin{array}{c} \theta_W \longrightarrow 2 \\ 360 \longrightarrow 30 \end{array} \Longrightarrow \theta_W = \frac{2 \times 360}{30} = 24^\circ$$

Circular diagram or Pie chart



Example (2)

We take the example of the distribution of families according to the number of children.

X : Number of children	Numbers n_i
0	11
1	16
2	21
3	25
4	17
5	8
6 and more	2
Total	100

Example (2)

We take the example of the distribution of families according to the number of children.

X : Number of children	Numbers n_i	Frequency f_i
0	11	0.11
1	16	0.16
2	21	0.21
3	25	0.25
4	17	0.17
5	8	0.08
6 and more	2	0.02
Total	100	1

Example (2)

We take the example of the distribution of families according to the number of children.

X : Number of children	Numbers n_i	Frequency f_i	% p_i
0	11	0.11	11
1	16	0.16	16
2	21	0.21	21
3	25	0.25	25
4	17	0.17	17
5	8	0.08	8
6 and more	2	0.02	2
Total	100	1	100

Example (2)

We take the example of the distribution of families according to the number of children.

X : Number of children	Numbers n_i	Frequency f_i	% p_i	$n_i^c \nearrow$
0	11	0.11	11	11
1	16	0.16	16	27
2	21	0.21	21	48
3	25	0.25	25	73
4	17	0.17	17	90
5	8	0.08	8	98
6 and more	2	0.02	2	100
Total	100	1	100	–

Example (2)

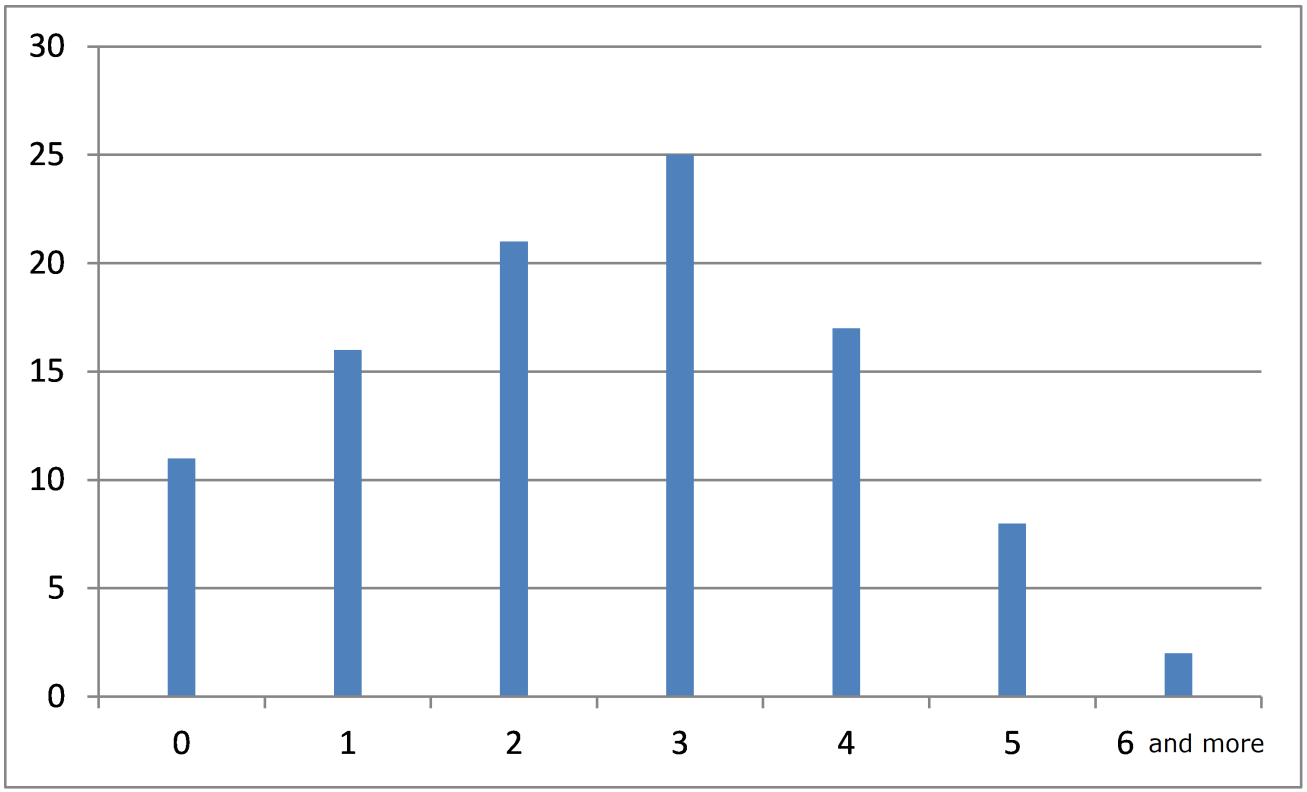
We take the example of the distribution of families according to the number of children.

X : Number of children	Numbers n_i	Frequency f_i	% p_i	$n_i^c \nearrow$	$n_i^c \searrow$
0	11	0.11	11	11	100
1	16	0.16	16	27	89
2	21	0.21	21	48	73
3	25	0.25	25	73	52
4	17	0.17	17	90	27
5	8	0.08	8	98	10
6 and more	2	0.02	2	100	2
Total	100	1	100	–	–

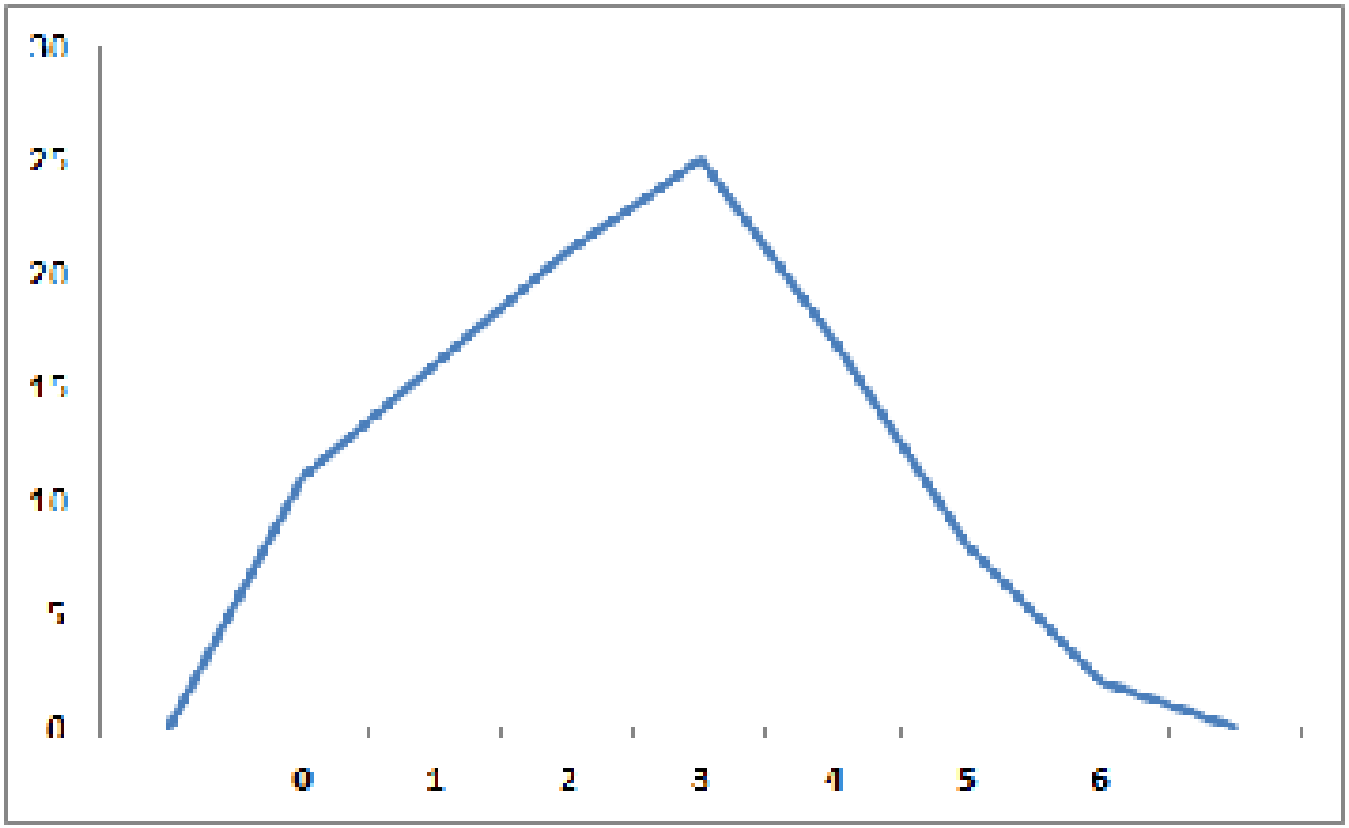
The example on the distribution of families according to the number of children is repeated.

X : Number of children	Numbers n_i
0	11
1	16
2	21
3	25
4	17
5	8
6 and more	2
Total	100

Bar chart



The polygon of numbers

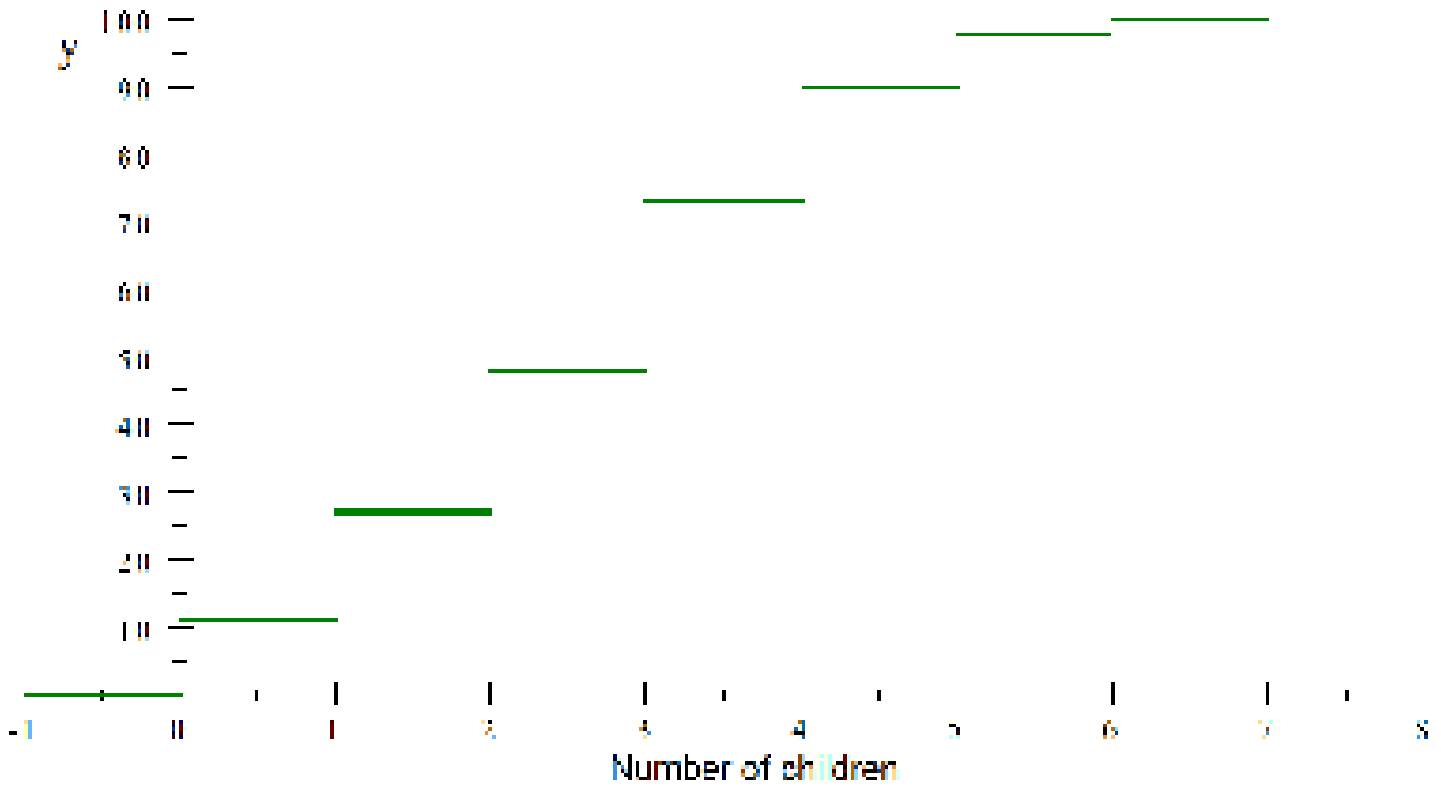


Example (2)

The example on the distribution of families according to the number of children is repeated.

X : Number of children	Numbers n_i	$n_i^c \nearrow$
0	11	11
1	16	27
2	21	48
3	25	73
4	17	90
5	8	98
6 and more	2	100
Total	100	

Cumulative absolute frequency curve



The data set of a statistical series is difficult to handle. It is therefore necessary to define a set of characteristic values (or parameters) that allow a condensed representation of the information contained in the statistical series. There are two categories of typical values:

1. The 1st order parameters or position parameters: **arithmetic mean, mode and median**.
2. The 2nd order parameters or dispersion parameters: **variance, standard deviation (écart-type), coefficient of variation and interquartile range (étendue intequartile).**

The mode, noted Mo, of a character (qualitative or quantitative) is the most observed modality, i.e. the one which has the greatest numbers (or the greatest frequency). The mode may not exist and if it does, it may not be unique.

- The set $2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18$ has for mode $Mo = 9$
- The set $3, 5, 8, 10, 12, 15, 16$ has no mode.
- The set $2, 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18$ has two modes $Mo_1 = 2$ and $Mo_2 = 9$, we say that it is a bimodal series.
- For example 1 the mode is $Mo =$Single.
- For example 2 the mode is $Mo = 3$.

The arithmetic mean of a statistical series $x_1, x_2, \cdots, x_k$, of a quantitative character X, and of respective numbers $n_1, n_2, \cdots, n_k$ is given by the real number X defined by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$$

where x_i is the modality i of the variable X and $n = \sum_{i=1}^{k} n_i$.

Example

Let's go back to example 2

X : Number of children	Numbers	$n_i x_i$
0	11	0
1	16	16
2	21	42
3	25	75
4	17	68
5	8	40
6 and more	2	12
Total	100	253

hence $\overline{X} = \frac{253}{100} = 2,53.$

Let be a statistical series of a variable X having for modalities (ordered in increasing order) $x_1 < x_2 < \cdots < x_n$. On The median of X is the number Me, if it exists, which divides the statistical series into two parts of equal numbers (containing the same number of observations) i.e. containing $\left[\frac{n}{2}\right]$ observations.

Consider the scores of 19 students on the statistics exam:

$$9, 15, 4, 8, 16, 10, 11, 10, 5, 12, 9, 10, 7, 12, 15, 11, 6, 13, 12.$$

The series is ordered in ascending order, so we have

$$4, 5, 6, 7, 8, 9, 9, 10, 10, 11, 11, 11, 12, 12, 12, 13, 15, 15, 16.$$

Half of the series is 9 so

$$\underbrace{4, 5, 6, 7, 8, 9, 9, 10, 10}_{9 \text{ values}}, \boxed{11}, \underbrace{11, 11, 12, 12, 12, 13, 15, 15, 16}_{9 \text{ values}}.$$

the median is the number that divides the series into two parts of equal numbers, the value that achieves this is 11 so $Me = 11$.

We now add the note 7 to the previous series, so we have a series of 20 notes:

$$4, 5, 6, 7, 8, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 15, 15, 16.$$

Half of the series is 10 so

$$\underbrace{4, 5, 6, 7, 7, 8, 9, 9, 10, 10}_{10 \text{ values}}, \underbrace{11, 11, 11, 12, 12, 12, 13, 15, 15, 16}_{10 \text{ values}}.$$

the median is the value that divides the series into two parts of equal numbers, it is found between the last 10 and the first 11, in this case we will take the average value of these two scores, then $Me = \frac{10+11}{2} = 10,5$.

In a general way let X be a discrete variable taking the ordered values $x_1, x_2, \cdots, x_p, x_{p+1}, \ldots, x_n$, then if

$$\begin{aligned} n &= 2p \Longrightarrow Me = \frac{x_{(p)} + x_{(p+1)}}{2} \\ n &= 2p+1 \Longrightarrow Me = x_{(p+1)}. \end{aligned}$$

Definition

The dispersion parameter measures the dispersion of the observations around a central value, these values tell us about the tendency of the observations to concentrate or disperse around the central values.

We have already determined a dispersion parameter, namely the range of a series, $e = x_{(n)} - x_{(1)}$.

The quartiles of a statistical series are the values noted Q_1, Q_2, Q_3 that divide the statistical series into four subseries containing the same number of observations. Then there are three quartiles, the first quartile Q_1, the second quartile Q_2 and the third quartile Q_3.
Remark. The second quartile Q_2 is the median Me.

Definition

The interquartile range is the difference between the third and first quartile and is denoted by

$$IQR = Q_3 - Q_1.$$

Example: Consider the marks of 21 students on the statistics exam

$$4, 5, 6, 7, 7, 8, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 13, 13, 15, 15, 16.$$

- Calculation of the second quartile: we have
  $n = 21 = 2 \times 10 + 1 \Longrightarrow p = 10$ hence
  $Q_2 = Me = x_{(p+1)} = x_{11} = 10.$

$$\underbrace{4, 5, 6, 7, 7, 8, 9, 9, 10, 10}, \boxed{10}, \underbrace{11, 11, 11, 12, 12, 13, 13, 15, 15, 16}.$$

- Calculation of the first quartile: the first sub-series contains $\frac{n-1}{2}$ observations : $\frac{n-1}{2} = 10 = 2 \times 5 \Longrightarrow p = 5$ hence $Q_1 = \frac{x_{(5)}+x_{(6)}}{2} = \frac{7+8}{2} = 7.5$.
- Calculation of the third quartile: The same goes for the second sub-series containing 10 observations, then $Q_3 = \frac{x_{(10+1+5)}+x_{(10+1+6)}}{2} = \frac{12+13}{2} = 12.5$

$$\underbrace{4, 5, 6, 7, 7}, \boxed{7.5}, \underbrace{8, 9, 9, 10, 10}, \boxed{10}, \underbrace{11, 11, 11, 12, 12}, \boxed{12.5}, \underbrace{13, 13, 15, 15, 16}.$$

Then $IQR = Q_3 - Q_1 = 12.5 - 7.5 = 5$.

The standard deviation of a statistical series $(x_1, x_2, \cdots, x_k)$, of a character X, and of respective numbers $(n_1, n_2, \cdots, n_k)$ is given by the real σ_X defined by

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i \left(x_i - \overline{X}\right)^2}$$

where $n = \sum_{i=1}^{n} n_i$.
x_i is the modality i of the variable X.

Remark 1. We can also describe σ_X in the following form

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \overline{X}^2}.$$

Remark 2. The variance of the variable X, noted $Var(X)$, is the square of the standard deviation,

$$Var(X) = \sigma_X^2.$$

Let's go back to example 2

X : Number of children	Numbers	$n_i x_i$	$n_i x_i^2$
0	11	0	0
1	16	16	16
2	21	42	84
3	25	75	225
4	17	68	272
5	8	40	200
6 and more	2	12	72
Total	100	253	869

hence $\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \overline{X}^2} = \sqrt{\frac{1}{100}\sum_{i=1}^{n} 869 - 2,53^2} \approx 1,51$. And $Var\,(X) = \sigma_X^2 = 2,2891$.

The coefficient of variation is a relative dispersion parameter expressed in percentage and defined by

$$CV_X = \frac{\sigma_X}{\overline{X}} \times 100.$$

The coefficient of variation in general if it is lower than 25% (for some authors it is 15% and for others 30%) one says that the series is homogeneous i.e. the observations are grouped around the average. Otherwise it is heterogeneous, i.e. the observations are scattered or far from the mean.

- $CV_X \leq 25\%$, we have a low dispersion,
- $25\% \leq CV_X \leq 80\%$, observations are quite scattered,
- $CV_X \geq 80\%$, a very strong dispersion.

For the previous example we have $CV_X = 100\frac{1,51}{2,53} \approx 59,68\%$.

In this case the number of possible values is in principle infinite. It is rare that observations are repeated, because the greater or lesser precision of the measurements will only allow us to discern a finite number of distinct values.
A statistical series is always presented in ascending order.
To build the statistical table we must group the data in intervals called **classes**.

Description of classes

- These classes are of the form $[a_i, b_i[$ or $]b_i, a_i]$.
- Their number is chosen in an arbitrary way but very often close to

$$K = \begin{cases} \sqrt{n} \text{ if } n < 50 \\ 1 + \frac{10}{3} \log_{10} n \text{ if } n \geq 50 \end{cases}$$

- It is recommended to choose an odd number.
- These classes are disjointed and their union must contain all the data obtained.
- In this case we call n_i the number of observations belonging to the i^{th} class.
- The amplitude of a class (a_i, b_i) is the length of the interval $[a_i, b_i[$: $b_i - a_i$, such that

$$a = \frac{e}{K}.$$

Remark

It is often recommended to choose the same amplitude for all classes.

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

For our example we have $x_{\min} = 56$ et $x_{\max} = 90$ et $n = 30 < 50$, then

$$e = 90 - 56 = 34 \text{ and } K = \sqrt{30} = 5.48$$

we can take 5 or 6 classes. The amplitude of the classes is

$$a = \frac{e}{K} = \frac{34}{5} = 6.8 \approx 7 \text{ (or } a = \frac{34}{6} \approx 5.66 \approx 6)$$

then we can built the following classes
$[56, 62[$; $[62, 68[$; $[68, 74[$; $[74, 80[$; $[80, 86[$; $[86, 92[$ with $a = 6$, or
$[56, 63[$; $[63, 70[$; $[70, 77[$; $[77, 84[$; $[84, 91[$ with $a = 7$.

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers
$[56; 63[$	59.5	3
$[63; 70[$	66.5	2
$[70; 77[$	73.5	7
$[77; 84[$	80.5	14
$[84; 91[$	87.5	4
Total	–	30

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency
$[56; 63[$	59.5	3	0.1
$[63; 70[$	66.5	2	0.0667
$[70; 77[$	73.5	7	0.2333
$[77; 84[$	80.5	14	0.4667
$[84; 91[$	87.5	4	0.1333
Total	–	30	1

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency	%
$[56; 63[$	59.5	3	0.1	10
$[63; 70[$	66.5	2	0.0667	6.67
$[70; 77[$	73.5	7	0.2333	23.33
$[77; 84[$	80.5	14	0.4667	46.67
$[84; 91[$	87.5	4	0.1333	13.33
Total	–	30	1	100

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency	%	$n_i^c \nearrow$
$[56;63[$	59.5	3	0.1	10	3
$[63;70[$	66.5	2	0.0667	6.67	5
$[70;77[$	73.5	7	0.2333	23.33	12
$[77;84[$	80.5	14	0.4667	46.67	26
$[84;91[$	87.5	4	0.1333	13.33	30
Total	–	30	1	100	–

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

classes	Center x_i	Numbers	Frequency	%	$n_i^c \nearrow$	$n_i^c \searrow$
$[56;63[$	59.5	3	0.1	10	3	30
$[63;70[$	66.5	2	0.0667	6.67	5	27
$[70;77[$	73.5	7	0.2333	23.33	12	25
$[77;84[$	80.5	14	0.4667	46.67	26	18
$[84;91[$	87.5	4	0.1333	13.33	30	4
Total	–	30	1	100	–	–

Remark

x_i represents the center of the class $[a_i, b_i[$ (in physics we assimilate any object to its center of gravity). This allows us to return to the discrete case. We say that we discretize a continuous case.

Remark

x_i represents the center of the class $[a_i, b_i[$ (in physics we assimilate any object to its center of gravity). This allows us to return to the discrete case. We say that we discretize a continuous case.

Remark

The table does not correspond to the series obtained, but it has the advantage of better representing the entire population by eliminating the outliers.

Example (3)

We take the example on the distribution of the workers according to the time taken to manufacture a given part.

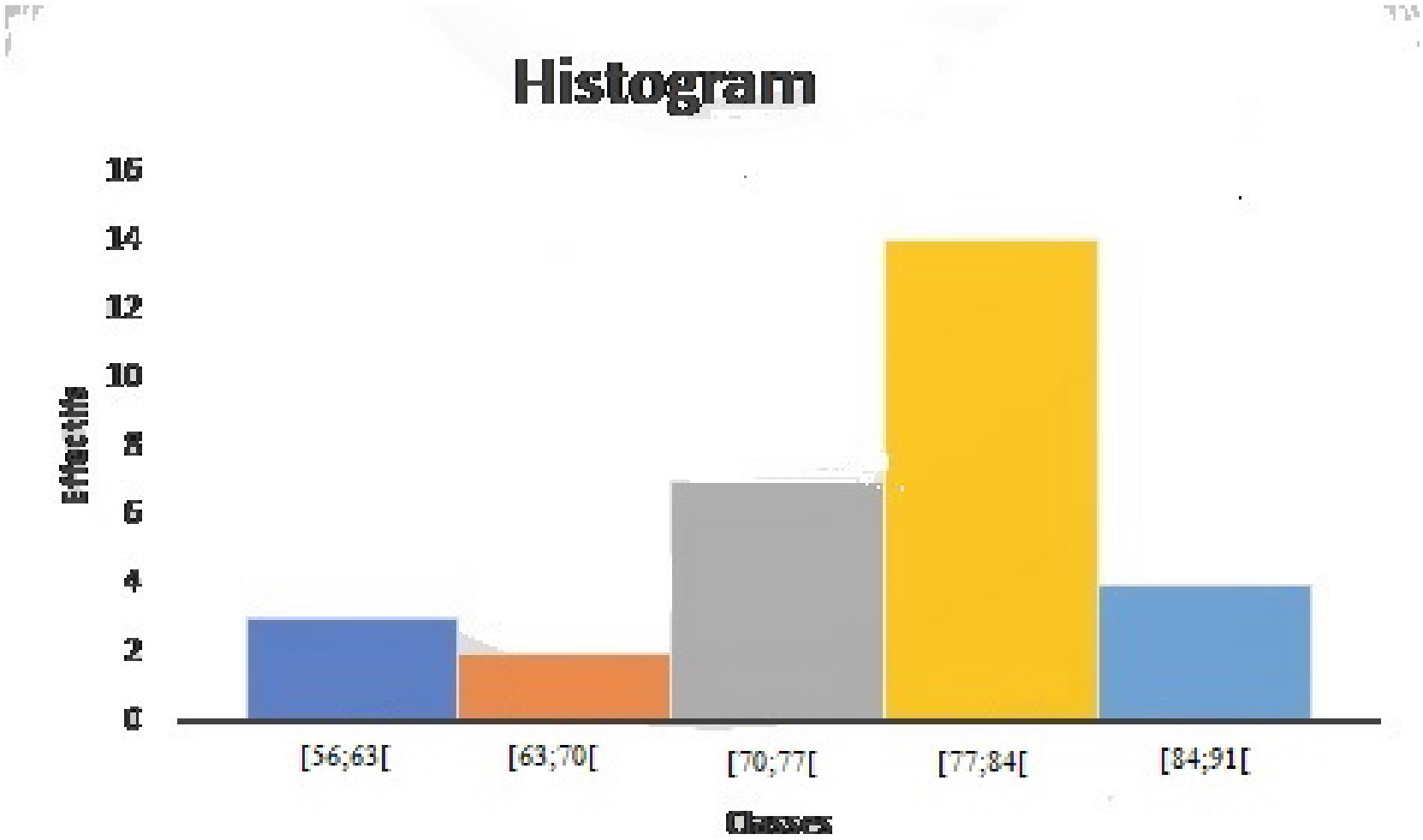
Histogram (case of classes of the same amplitude)

The histogram of a statistical series is the set of rectangles having for base the amplitude of a class and for height the number n_i of this class.

classes	Center x_i	Numbers	$n_i^c \nearrow$
$[56; 63[$	59.5	3	3
$[63; 70[$	66.5	2	5
$[70; 77[$	73.5	7	12
$[77; 84[$	80.5	14	26
$[84; 91[$	87.5	4	30
Total	–	30	–

Remark

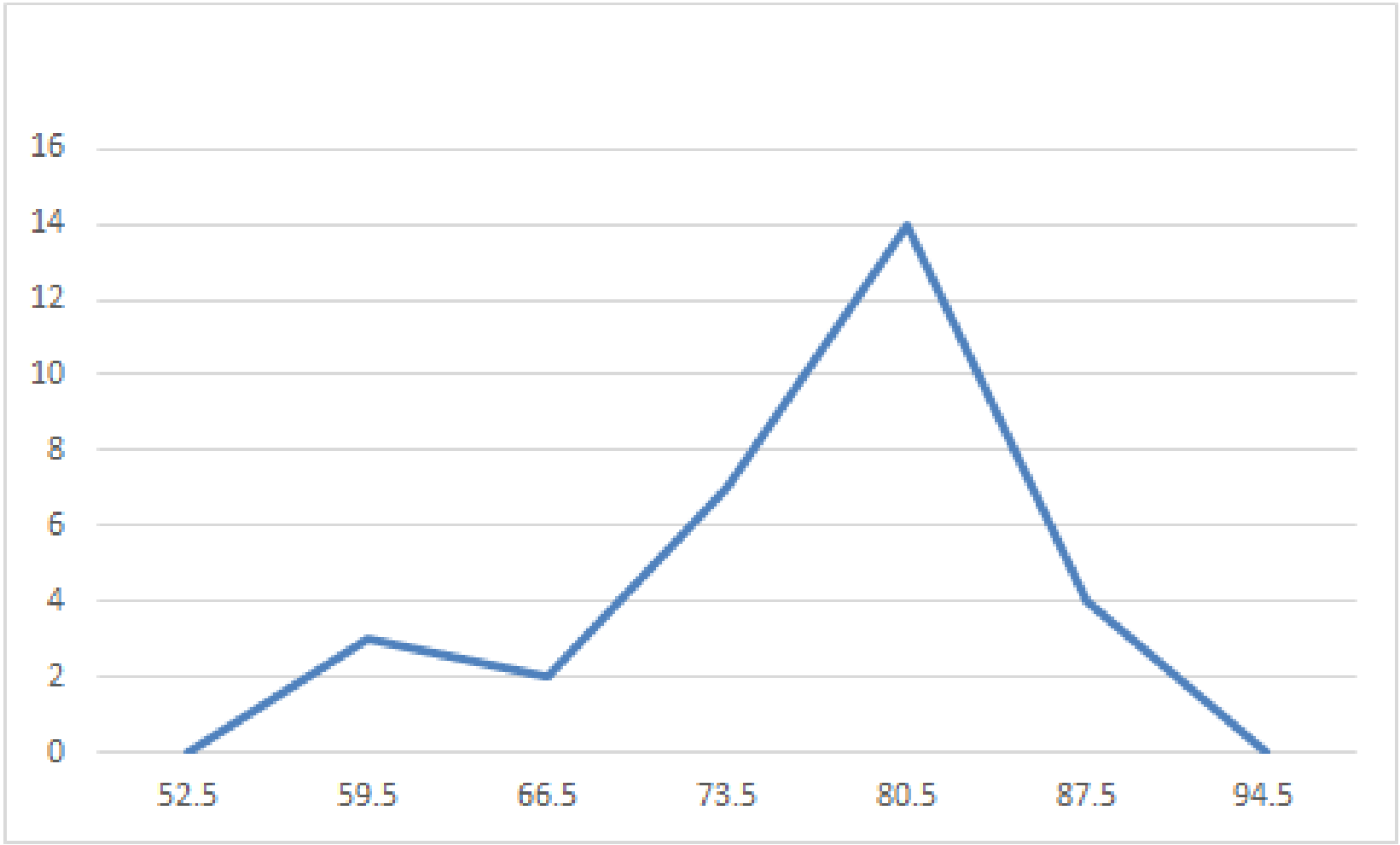
Histogram (case of classes with the same amplitude)



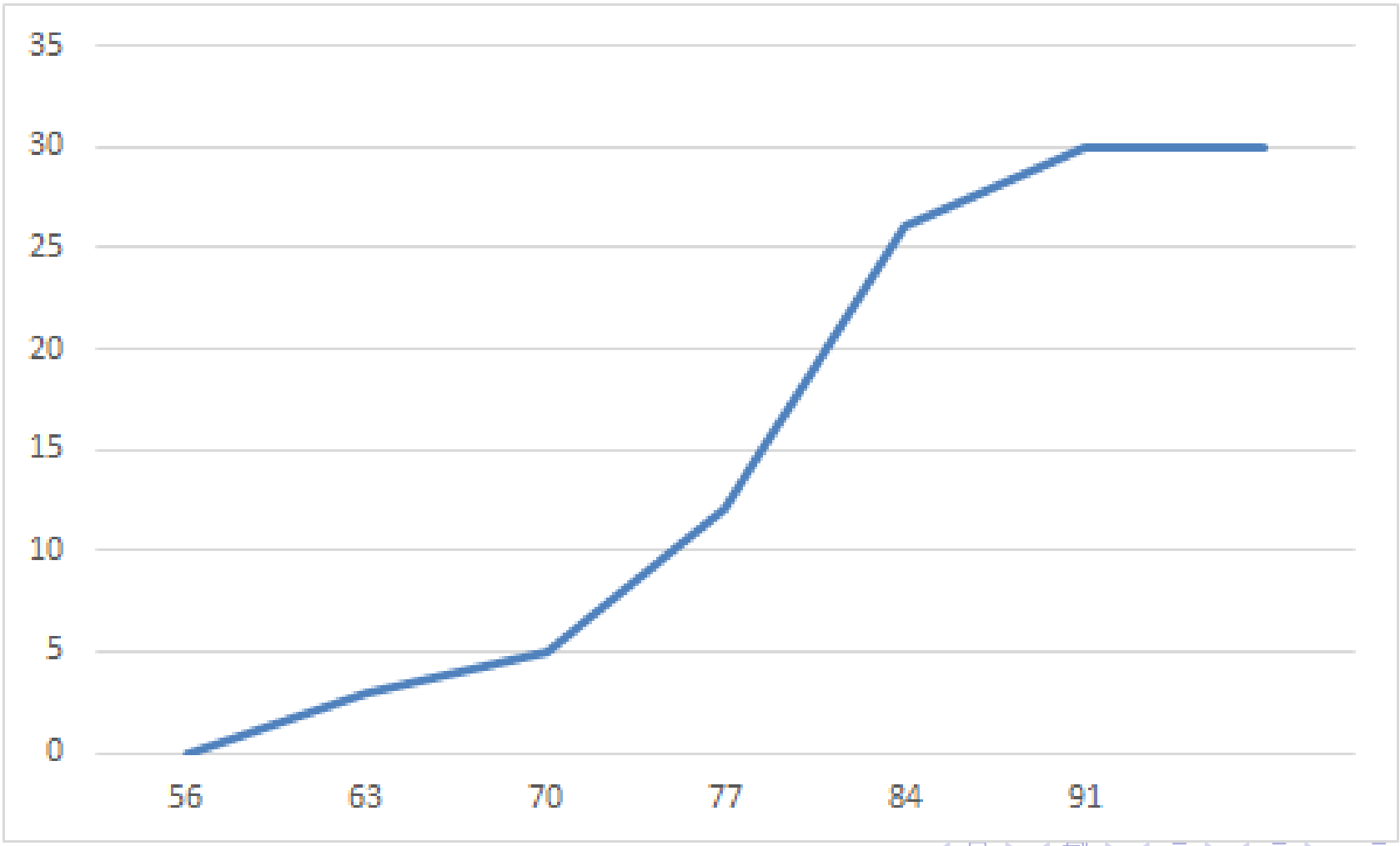
Histogram (case of classes of the same amplitude)
Remark: The total area of the Hitogram should always be equal to the sample size, considering the amplitude of the classes equal to the unit.
For the previous example we have

$$Area = (3 + 2 + 7 + 14 + 4) \times 1 = 30.$$

Absolute frequency polygon



Cumulative absolute frequency curve



Histogram in the case of classes of different amplitudes

Example (4)

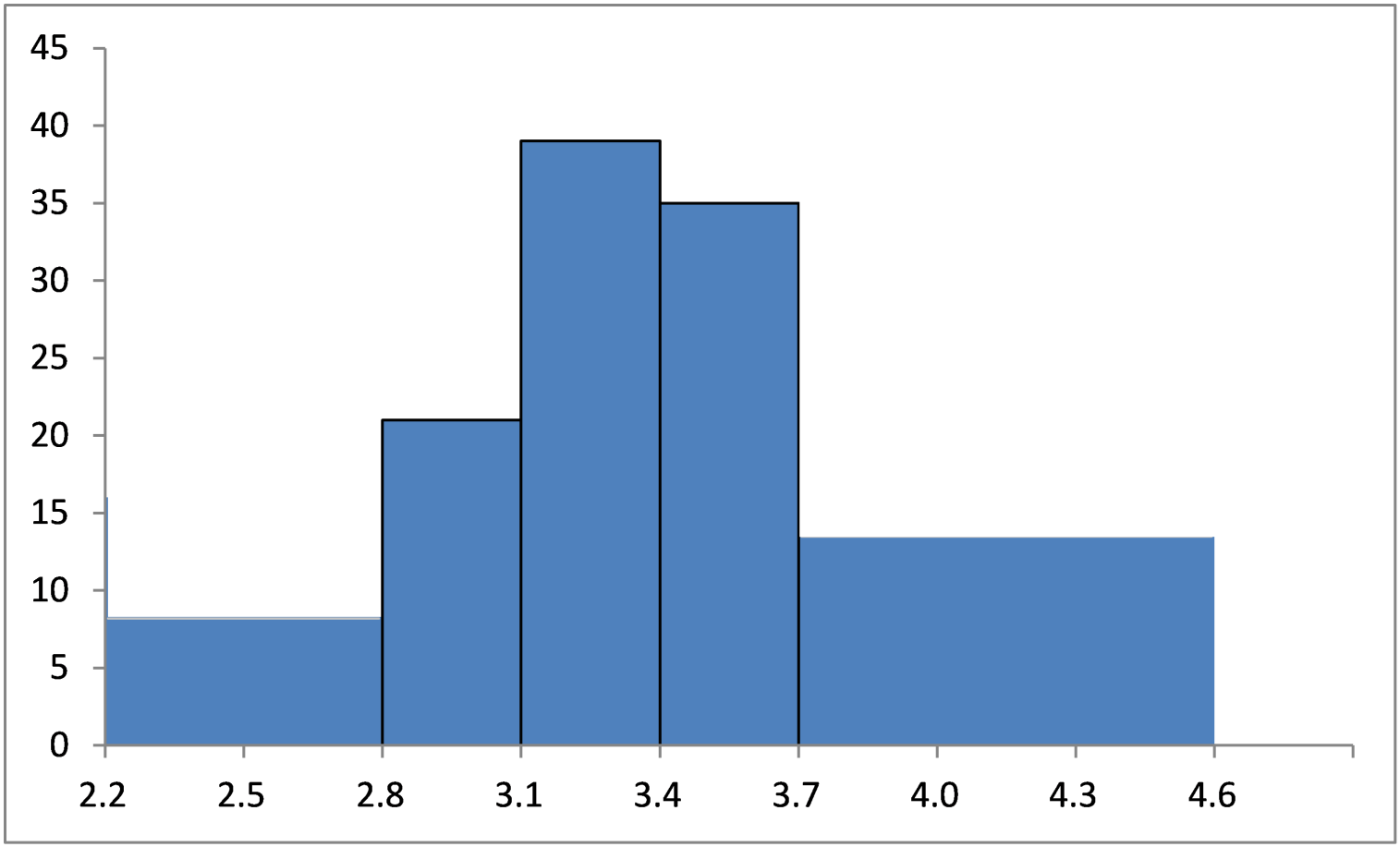
We consider the weight distribution in Kg of 150 newborns in a maternity hospital, the results are grouped in the following statistical table:

X Number of children	Center	Amplitude a	Numbers	Frequency
$[2.2; 2.8[$	2.50	$0,6$	16	0.1067
$[2.8; 3.1[$	2.95	$0,3$	21	0.1400
$[3.1; 3.4[$	3.25	$0,3$	39	0.2600
$[3.4; 3.7[$	3.55	$0,3$	35	0.2333
$[3.7; 4.6[$	4.15	$0,9$	39	0.2600
Total	–		150	1

Histogram in the case of classes with different amplitudes

To determine the histogram we determine the ratio between the different classes. For that, we consider the classes unit the classes having the smallest amplitude namely $[2.8; 3.1[$, $[3.1; 3.4[$ and $[3.4; 3.7[$ so we notice that the class $[2.2; 2.8[$ represent $\frac{0,6}{0,3} = 2$ unities and the classe $[3.7; 4.6[$ represent $\frac{0,9}{0,3} = 3$ unities. So to respect the previous remark we have to divide the number of students in the first class by 2 (i.e. $\frac{16}{2} = 8$) and the last class size by 3 (i.e. $\frac{39}{3} = 13$). We then obtain the following histogram:

Histogram in the case of classes with different amplitudes



Let's go back to example 3
first, we determine the modal class is $[77; 84[$ and then the mode will be the center of the class i.e. $Mo = 80,5$.
There is a second method which consists in estimating the mode by interpolation.
If $[a_i; b_i[$ is the modal class, then the mode is

$$Mo = a_i + \frac{e_1}{e_1 + e_2} a,$$

where
$a = b_i - a_i$ is the amplitude of the modal class;
e_1 is the difference between the numbers of the modal and previous classes;
e_2 is the difference between the numbers of the modal and following classes.

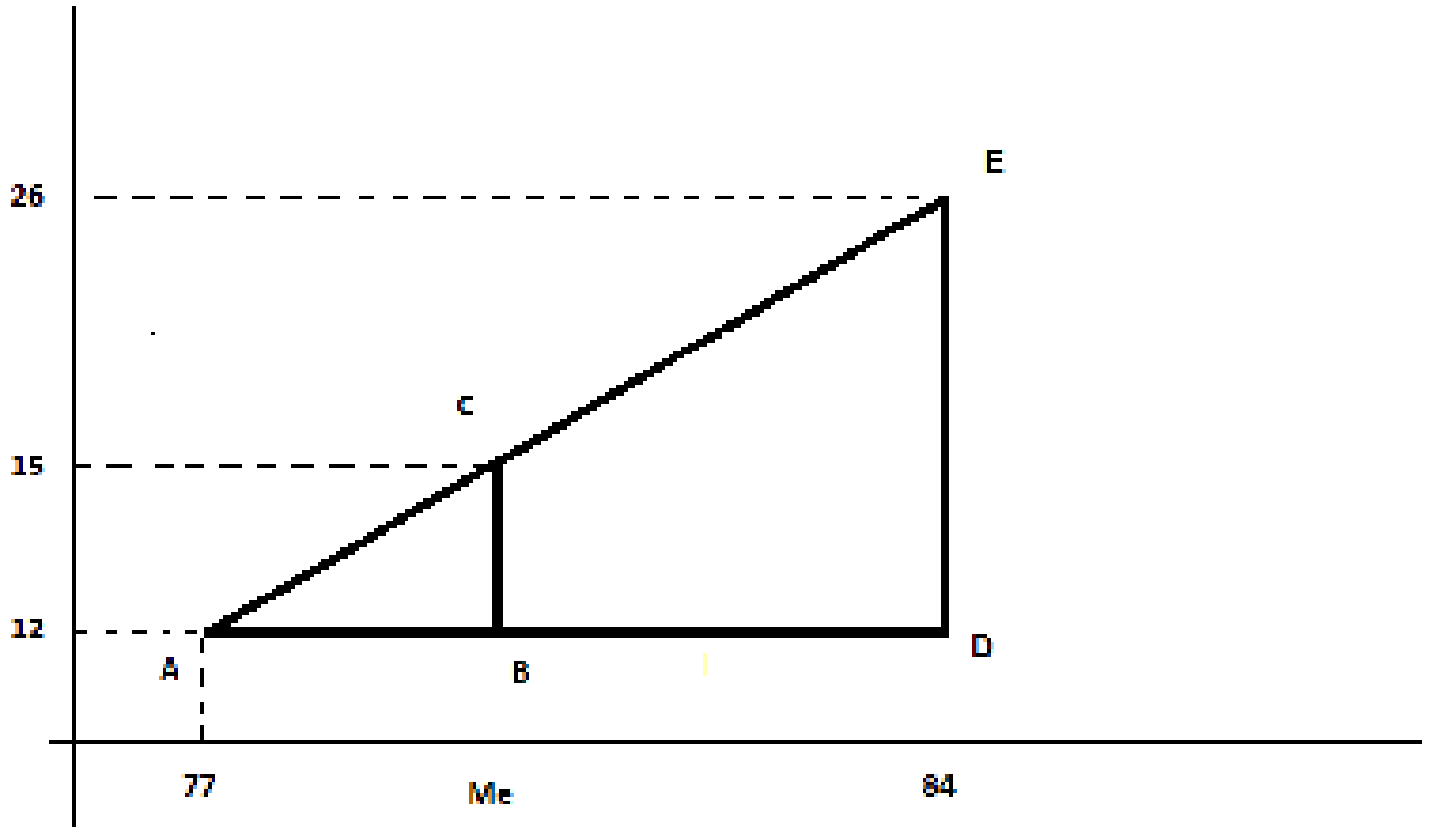
classes	Center x_i	Numbers	$n_i x_i$
$[56;63[$	59.5	3	178.5
$[63;70[$	66.5	2	133
$[70;77[$	73.5	7	514.5
$[77;84[$	80.5	14	1127
$[84;91[$	87.5	4	350
Total	–	30	2303

Then $\overline{X} = \frac{1}{30}\sum_{i=1}^{5} n_i x_i = \frac{2303}{30} \simeq 76.77.$

classes	Center x_i	Numbers	$n_i^c \nearrow$
$[56;63[$	59.5	3	3
$[63;70[$	66.5	2	5
$[70;77[$	73.5	7	12
$[77;84[$	80.5	14	26
$[84;91[$	87.5	4	30
Total	–	30	–

First, we determine the median class, which corresponds to half the numbers of people $\left(\frac{n}{2} = 15\right)$ hence $Me \in [77;84[$.

To determine *Me* we consider the curve of the increasing cumulative numbers but we will be interested only in the portion of the class $[77; 84[$



According to the theorem of Thales we have

$$\begin{aligned} \frac{AB}{AD} = \frac{BC}{DE} &\Longrightarrow \frac{Me-77}{84-77} = \frac{15-12}{26-12} \\ &\Longrightarrow Me = \frac{3}{14}7 + 77 \\ &\Longrightarrow Me = 78.5 \end{aligned}$$

Example

We take the example of the distribution of newborns according to their weight.

X : Number of children	Center	$n_i^c \nearrow$
$[2.2; 2.5[$	2.35	5
$[2.5; 2.8[$	2.65	16
$[2.8; 3.1[$	2.95	37
$[3.1; 3.4[$	3.25	76
$[3.4; 3.7[$	3.55	111
$[3.7; 4.0[$	3.85	131
$[4.0; 4.3[$	4.15	144
$[4.3; 4.6[$	4.45	150
Total	—	—

Determination of Q_1 : $\frac{n}{4} = 37.5 \leq 76$ then $Q_1 \in [3.1; 3.4]$ hence

$$Q_1 = \frac{37.5 - 37}{76 - 37} \times (3.4 - 3.1) + 3.1 = 3.104Kg$$

Determination of Q_3 : $\frac{3n}{4} = 112.5$ then $Q_3 \in [3.7; 4.0]$ hence

$$Q_3 = \frac{112.5 - 111}{131 - 111} \times (4.0 - 3.7) + 3.7 = 3.723Kg$$

hence

$$\begin{aligned} IQR &= Q_3 - Q_1 = 3.723 - 3.104 \\ &\Longrightarrow IQR = 0.619Kg. \end{aligned}$$

50% of the observations fall within an IQR length interval (between the values Q_1 and Q_3).

Example

We take the example of the distribution of newborns according to their weight.

X : Weight in Kg	x_i	n_i	$n_i^c \nearrow$	$n_i x_i$	$n_i x_i^2$
$[2.2; 2.5[$	$2,35$	5	5	$11,75$	$27,6125$
$[2.5; 2.8[$	$2,65$	11	16	$29,15$	$77,2475$
$[2.8; 3.1[$	$2,95$	21	37	$61,95$	$182,7525$
$[3.1; 3.4[$	$3,25$	39	76	$126,75$	$411,9375$
$[3.4; 3.7[$	$3,55$	35	111	$124,25$	$441,0875$
$[3.7; 4.0[$	$3,85$	20	131	77	$296,45$
$[4.0; 4.3[$	$4,15$	13	144	$53,95$	$223,8925$
$[4.3; 4.6[$	$4,45$	6	150	$26,7$	$118,815$
Total		150		$511,5$	$1779,795$

We have

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{8} n_i x_i = \frac{511,5}{150} \approx 3,41Kg \\
\sigma_X^2 &= \frac{1}{n}\sum_{i=1}^{8} n_i x_i^2 - \overline{X}^2 = \frac{1779,795}{150} - 3,41^2 \\
&\Longrightarrow \quad \sigma_X^2 \approx 0,2372 \Longrightarrow \sigma_X \approx 0,4870Kg.
\end{aligned}$$

And

$$\begin{aligned}
CV_X &= 100\frac{0,4870}{3,41} \\
&\Longrightarrow \quad 14,28\%.
\end{aligned}$$

We have

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{X}\right)^3}{\sigma_X^3}.$$

- The distribution is asymmetric if $|\gamma_1| > \frac{1}{2}$.
- The distribution spreads to the right if $\gamma_1 > 0$;
- The distribution spreads to the left if $\gamma_1 < 0$;
- The distribution is symmetric if $\gamma_1 = 0$;

We have

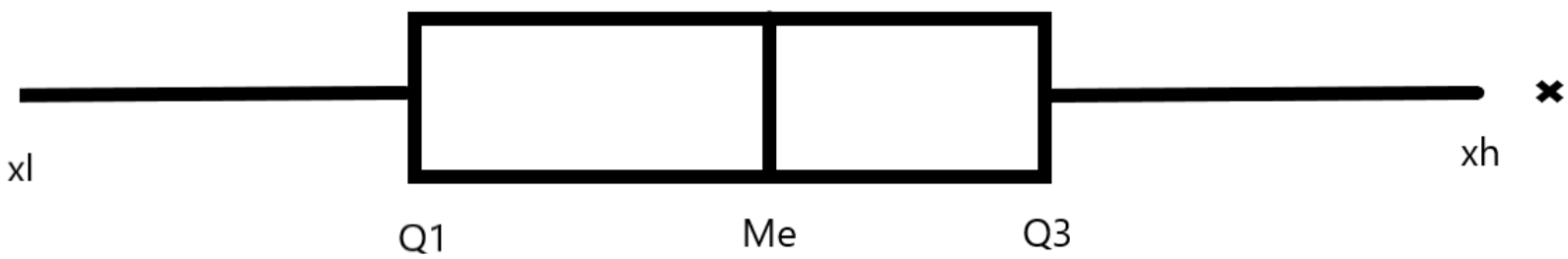
$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{X}\right)^4}{\sigma_X^4} - 3.$$

We say that

- the distribution is normal (the curve is mesokurtic) if $\gamma_2 = 0$;
- the distribution is less flattened than the normal distribution (the curve is leptokurtic) if $\gamma_2 > 0$;
- the distribution is more flattened than the normal distribution (the curve is platykurtic) if $\gamma_2 < 0$;

It is interesting to visualize concepts such as symmetry, dispersion or centrality of the distribution of values associated with a variable.
They are also very interesting to compare variables based on similar scales and to compare the values of observations of groups of individuals on the same variable.
To do that we want a graphical representation that answer all these questions and the box-plots is the method for graphically demonstrating the locality, dispersion and skewness groups of numerical data through their quartiles.
The box -plot or box-and-whisker plot was first introduced in 1970 by John Tukey, who later published on the subject in his book "Exploratory Data Analysis" in 1977.

1. We draw a rectangle of length IQR;
2. We locate the median by a line inside the box;
3. We calculate $Q_3 + 1,5 \cdot IQR$ and we look for the last observation $x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right)$;
4. We calculate $Q_1 - 1,5 \cdot IQR$ and look for the first observation $x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right)$;
5. We draw two lines from the midpoints of the rectangle widths to the values x_l and x_h which are called whiskers;
6. Any value that does not lie between the ends of the whiskers is usually represented by a star, which we will call the extreme value or outlier.


xl
Q1
Me
Q3
xh

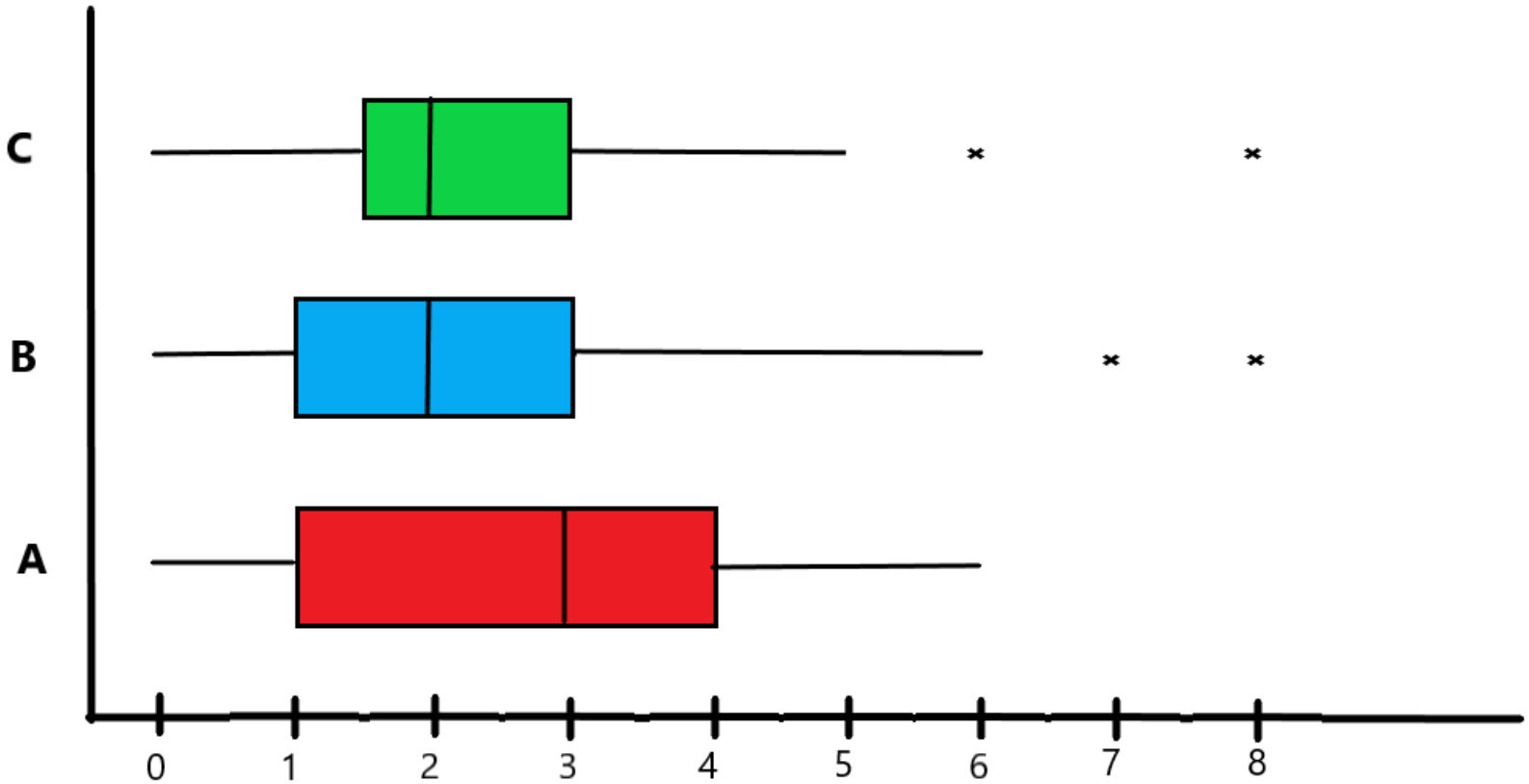
We consider the distribution of the number of children per family in three different neighborhoods A, B and C. Let X_A, X_B and X_C be the number of children per family in the neighborhoods A, B and C respectively.

X_A	n_i	$n_i^c \nearrow$
0	11	11
1	16	27
2	21	48
3	25	73
4	17	90
5	8	98
6	2	100
Total	100	

X_B	n_i	$n_i^c \nearrow$
0	12	12
1	28	40
2	46	86
3	38	124
4	20	144
5	2	146
6	1	147
7	1	148
8	2	150
Total	150	

X_C	n_i	$n_i^c \nearrow$
0	10	10
1	15	25
2	30	55
3	25	80
4	15	95
5	3	98
6	1	99
8	1	100
Total	100	

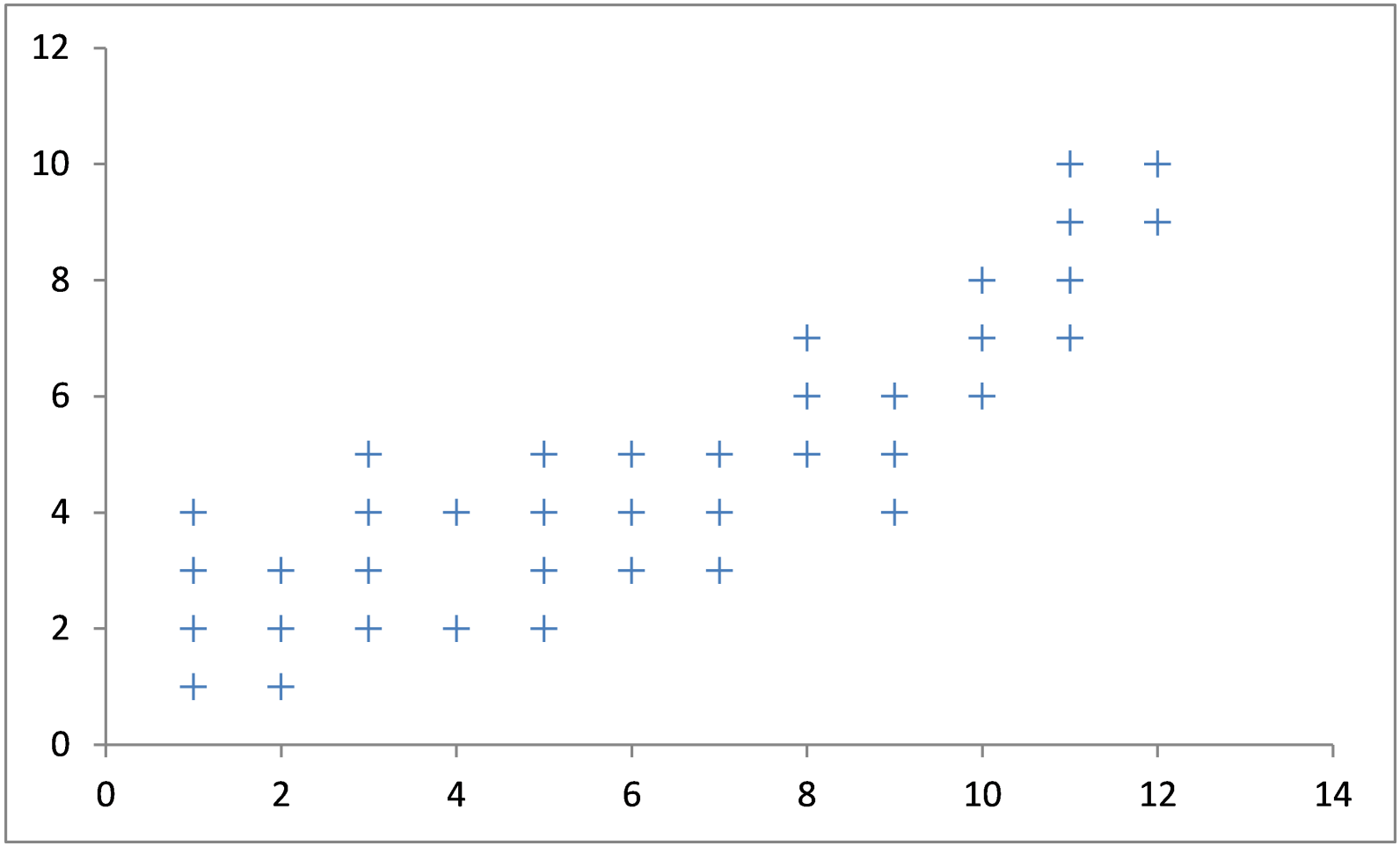
For the neighborhood A we have: $Me = 3; Q_1 = 1; Q_3 = 4$, then $IQR = 3$ so
$x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right) = \min\left(xi \middle| xi \geq -3,5\right) = 0$ and
$x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right) = \max\left(xi \middle| xi \leq 8,5\right) = 6.$
For the neighborhood B we have: $Me = 2; Q_1 = 1; Q_3 = 3$, then $IQR = 2$ so
$x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right) = \min\left(xi \middle| xi \geq -2\right) = 0$ and
$x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right) = \max\left(xi \middle| xi \leq 6\right) = 6$ and there is twi outliers 7 and 8.
For the neighborhood C we have: $Me = 2; Q_1 = 1,5; Q_3 = 3$, then $IQR = 1,5$ so
$x_l = \min\left(xi \middle| xi \geq Q_1 - 1,5 \cdot IQR\right) = \min\left(xi \middle| xi \geq -1,25\right) = 0$ and
$x_h = \max\left(xi \middle| xi \leq Q_3 + 1,5 \cdot IQR\right) = \max\left(xi \middle| xi \leq 5,25\right) = 5$ and there is two outliers 6 and 8.


C
B
A
0
1
2
3
4
5
6
7
8

Let $\mathcal{P}$ be a population of total size n, on which we study two quantitative characteristics X and Y, we are interested in the relation between these two variables.
We start by defining the double statistical series of $\mathcal{P}$ for the characters X and Y

$$\begin{array}{ccc} \mathcal{P} & \longrightarrow & \mathbb{R}^2 \\ e_{ij} & \longmapsto & (x_i, y_j) \end{array}$$

A first idea to try to show the relation between X and Y is to plot the scatter plot associated with the double statistical series.



The scatter plot can have various shapes and these shapes will guide us in defining the notion of correlation.

- **First situation:** the Scatter plot can be formed of aligned points so X and Y are linked by a functional relation of the form $y = f(x)$.
- **Second situation:** the scatter plot is dispersed, the two observed values do not depend on each other, we say that the two characters X and Y are **independent**.
- **Third situation:** intermediate situation between independence and functional relationship.

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_l	Marginal numbers
x_1	n_{11}	n_{12}		n_{1j}		n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2l}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}		n_{ij}		n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}		n_{kj}		n_{kl}	$n_{k\bullet}$
Marginal numbers	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

n_{ij} is the partial numbers of the couple (x_i, y_j) and $n = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$

$n_{i\bullet}$ is the marginal numbers of x_i and $n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$

$n_{\bullet j}$ is the marginal numbers of y_j and $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$.

$n = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j}$

Definition

The couple (X, Y) is statistically independent if we have $\forall i = 1, \cdots, k; j = 1, \cdots, l$

$$f_{ij} = \frac{n_{ij}}{n} = f_{i\bullet} \times f_{\bullet j} = \frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n}$$

Definition

We call covariance of the variables X and Y and we note $Cov(X, Y)$, the number

$$\begin{aligned} Cov(X, Y) &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \left(x_i - \overline{X}\right) \left(y_j - \overline{Y}\right) \\ &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i y_j - \overline{XY}. \end{aligned}$$

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \text{ et } \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \text{ et } \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

2. *If the variables* X *and* Y *are statistically independant then* $Cov(X, Y) = 0$.

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \text{ et } \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

2. *If the variables* X *and* Y *are statistically independant then* $Cov(X, Y) = 0$. *But the reciprocal is not true*

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \text{ et } \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

2. *If the variables* X *and* Y *are statistically independant then* $Cov\,(X,Y) = 0$. *But the reciprocal is not true*

Definition

If $Cov(X,Y) = 0$ we say that the variables X and Y are uncorrelated.

Definition

We call the linear coefficient of correlation of the variables X and Y the number

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

1. The linear coefficient of correlation is invariant by change of origin and unit of measurement.
2. We have $-1 \leq \rho(X, Y) \leq 1$
3. If $\rho(X, Y) = 0$, the variables X and Y are uncorrelated.

Remark

If $\rho(X, Y) > 0$: X and Y evolve in the same direction

Definition

We call the linear coefficient of correlation of the variables X and Y the number

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

1. The linear coefficient of correlation is invariant by change of origin and unit of measurement.
2. We have $-1 \leq \rho(X, Y) \leq 1$
3. If $\rho(X, Y) = 0$, the variables X and Y are uncorrelated.

Remark

If $\rho(X, Y) > 0$: X and Y evolve in the same direction
If $\rho(X, Y) < 0$: X and Y evolve in the opposite direction

Definition

We call the linear coefficient of correlation of the variables X and Y the number

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

1. The linear coefficient of correlation is invariant by change of origin and unit of measurement.
2. We have $-1 \leq \rho(X, Y) \leq 1$
3. If $\rho(X, Y) = 0$, the variables X and Y are uncorrelated.

Remark

If $\rho(X, Y) > 0$ *:* X *and* Y *evolve in the same direction*
If $\rho(X, Y) < 0$ *:* X *and* Y *evolve in the opposite direction*
If $\rho(X, Y)$ *is* $\pm$ *near of* 1 *: the correlation will be very good.*

Determine all the parameters for X and Y from the following contingency table

X \ Y	5	7	9	11	13	$n_{i\bullet}$
1				1	4	5
2			2	7	1	10
4			9	1		10
6	2	8	6	1		17
9	5	2	1			8
$n_{\bullet j}$	7	10	18	10	5	50

X \ Y	5	7	9	11	13	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\sum_j n_{ij}x_i y_j$
1				1	4	5	5	5	63
2			2	7	1	10	20	40	216
4			9	1		10	40	160	368
6	2	8	6	1		17	102	612	786
9	5	2	1			8	72	648	432
$n_{.j}$	7	10	18	10	5	**50**	**239**	**1465**	**1865**
$n_{.j}y_j$	35	70	162	110	65	**442**			
$n_{.j}y_j^2$	175	490	1458	1210	845	**4178**			
$\sum_i n_{ij}x_i y_j$	285	462	765	275	78	**1865**			

$$\begin{aligned}
\overline{X} &= \frac{239}{50} = 4.78; \overline{Y} = \frac{442}{50} = 8.84 \\
\sigma_X &= \sqrt{\frac{1465}{50} - 4.78^2} = \sqrt{6.4516} = 2.54 \\
\sigma_Y &= \sqrt{\frac{4178}{50} - 8.84^2} = \sqrt{5.4144} \approx 2.3269 \\
Cov\,(X, Y) &= \frac{1865}{50} - 4.78 \times 8.84 = -4.9552 \\
\rho\,(X, Y) &= \frac{-4.9552}{2.54 \times 2.3269} \approx -0.8384.
\end{aligned}$$

Since $\rho\,(X, Y) < 0$, then X and Y evolve in the opposite direction.

To have a general idea on the relation between two characters we study the conditional distributions, for that we are interested in the couples (x_i, y_j) where we fix one of the variables.

Definition

We call the conditional mean of the variable Y knowing x_i, the real number

$$\overline{Y}/x_i = \frac{1}{n_{i\cdot}} \sum_{j=1}^{l} n_{ij} y_j .$$

Definition

We call the regression curve of Y in X the broken curve that connects the points $\left(x_i, \overline{Y}/x_i\right)$.

In the same way we can construct the regression curve of X in Y.

Definition

We call the conditional mean of the variable X knowing y_j, the real number

$$\overline{X}/y_j = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i.$$

Definition

We call the regression curve of X in Y the broken curve that connects the points $\left(\overline{X}/y_j, y_j\right)$.

We can also determine the conditional variance .

Definition

We call the conditional variance of the vairable Y knowing x_i, the real number

$$\sigma^2_{Y/x_i} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{l} n_{ij} \left(y_j - \overline{Y}/x_i\right)^2 = \frac{1}{n_{i\cdot}} \sum_{j=1}^{l} n_{ij} y_j^2 - \left(\overline{Y}/x_i\right)^2 .$$

Definition

We call th conditional variance of the vairable X knowing y_j, the real number

$$\sigma^2_{X/y_j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} n_{ij} \left(x_i - \overline{X}/y_j\right)^2 = \frac{1}{n_{\cdot j}} \sum_{i=1}^{l} n_{ij} x_i^2 - \left(\overline{X}/y_j\right)^2 .$$

We will use the conditional variance or conditional standard deviation to measure the strength of the relationship of Y with X, by drawing the broken curves connecting the points $\left(x_i, \overline{Y}/x_i - \sigma_{Y/x_i}\right)$ et $\left(x_i, \overline{Y}/x_i + \sigma_{Y/x_i}\right)$.
The band between these two curves is called the regression corridor of Y in X.
In the same way we can draw the regression corridor of X in Y by joining the points $\left(\overline{X}/y_j - \sigma_{X/y_j}, y_j\right)$ and $\left(\overline{X}/y_j + \sigma_{X/y_j}, y_j\right)$.

Example

Let's take the previous example

X \ Y	5	7	9	11	13	$n_{i.}$	$\overline{Y}/x_i$	σ_{Y/x_i}
1				1	4	5	12.6	0.80
2			2	7	1	10	10.8	1.08
4			9	1		10	9.2	0.60
6	2	8	6	1		17	7.71	1.52
9	5	2	1			8	6	1.41
$n_{.j}$	7	10	18	10	5	50		
$\overline{X}/y_j$	8.14	6.6	4.72	2.5	1.2			
σ_{X/y_j}	1.36	1.20	1.48	1.32	1.33			

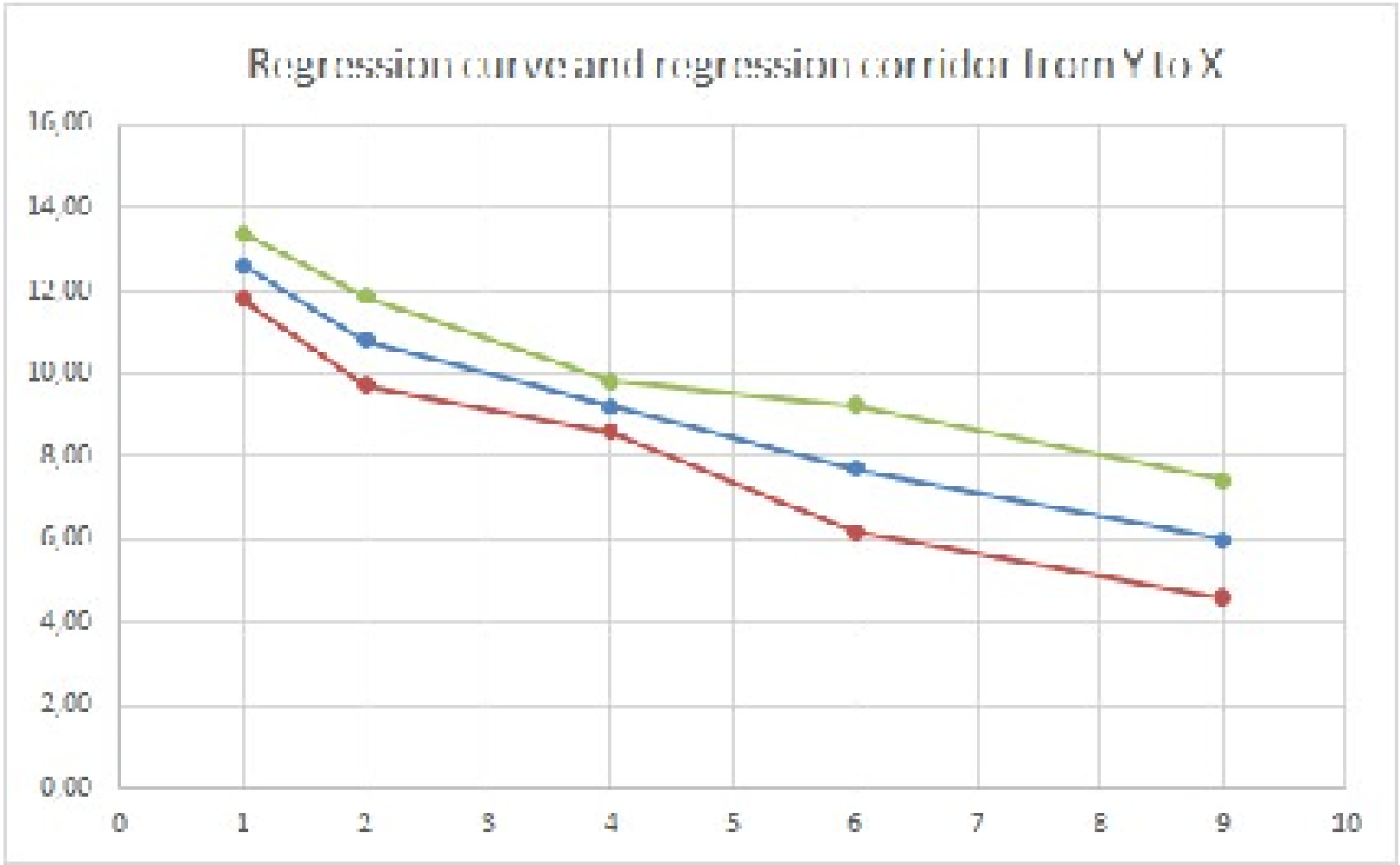
Example

To draw the regression corridor of Y in X we join the points

$$\left(x_i; \overline{Y}/x_i - \sigma_{Y/x_i}\right) = \{(1; 11.80), (2; 9.72), (4; 8.60), (6; 6.18), (9; 4.59)\}$$

then the points

$$\left(x_i; \overline{Y}/x_i + \sigma_{Y/x_i}\right) = \{(1; 13.40), (2; 11.88), (4; 9.80), (6; 9.23), (9; 7.41)\}.$$


Regression curve and regression curve from Y to X
16,00
14,00
12,00
10,00
8,00
6,00
4,00
2,00
0,00
0
1
2
3
4
5
6
7
8
9
10

We have the following relationship called variance decomposition

$$\sigma_Y^2 = \overline{\sigma_{Y/X}^2} + \sigma_{\overline{Y/X}}^2.$$

where

$$\overline{\sigma_{Y/X}^2} = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\sigma_{Y/x_i}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(y_j - \overline{Y}/x_i\right)^2$$

is the average conditional variance of Y with respect to X (also called the variance around the regression corridor) is the average of the conditional variances, and

$$\sigma_{\overline{Y/X}}^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\left(\overline{Y}/x_i - \overline{Y}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\left(\overline{Y}/x_i\right)^2 - \overline{Y}^2$$

is the variance of conditional means.

Example

For the previous example we have $\overline{Y} = 8.84$,

$$\begin{aligned}\overline{\sigma^2_{Y/X}} &= \left(5 \cdot 0.64 + 10 \cdot 1.16 + 10 \cdot 0.36 + 17 \cdot 2.33 + 8 \cdot 2.00\right)/50 \\ &= 1.48\end{aligned}$$

and

$$\begin{aligned}\sigma^2_{\overline{Y/X}} &= \left(5 \cdot 12.60^2 + 10 \cdot 10.80^2 + 10 \cdot 9.20^2 + 17 \cdot 7.71^2 + 8 \cdot 6^2\right)/50 - \\ &= 3.96\end{aligned}$$

then $\overline{\sigma^2_{Y/X}} + \sigma^2_{\overline{Y/X}} = 5.44 \simeq \sigma^2_Y = 5.41$

Definition

We call the Pearson correlation ratio the real number

$$\eta^2_{Y/X} = \frac{\sigma^2_{\overline{Y/X}}}{\sigma^2_Y}.$$

It's the percentage of variability (of the variable Y) due to the differences between the modalities (of the variable X).

Remark

1. $0 \leq \eta^2_{Y/X} \leq 1$.
2. If $\eta^2_{Y/X} = 0 \Longleftrightarrow \sigma^2_{\overline{Y/X}} = 0$ then the regression curve of Y in X is horizontal, this means that the variable Y is uncorrelated on average with the variable X.
3. If $\eta^2_{Y/X} = 1 \Longleftrightarrow \overline{\sigma^2_{Y/X}} = 0$ then Y is totally linked to X.

Theorem

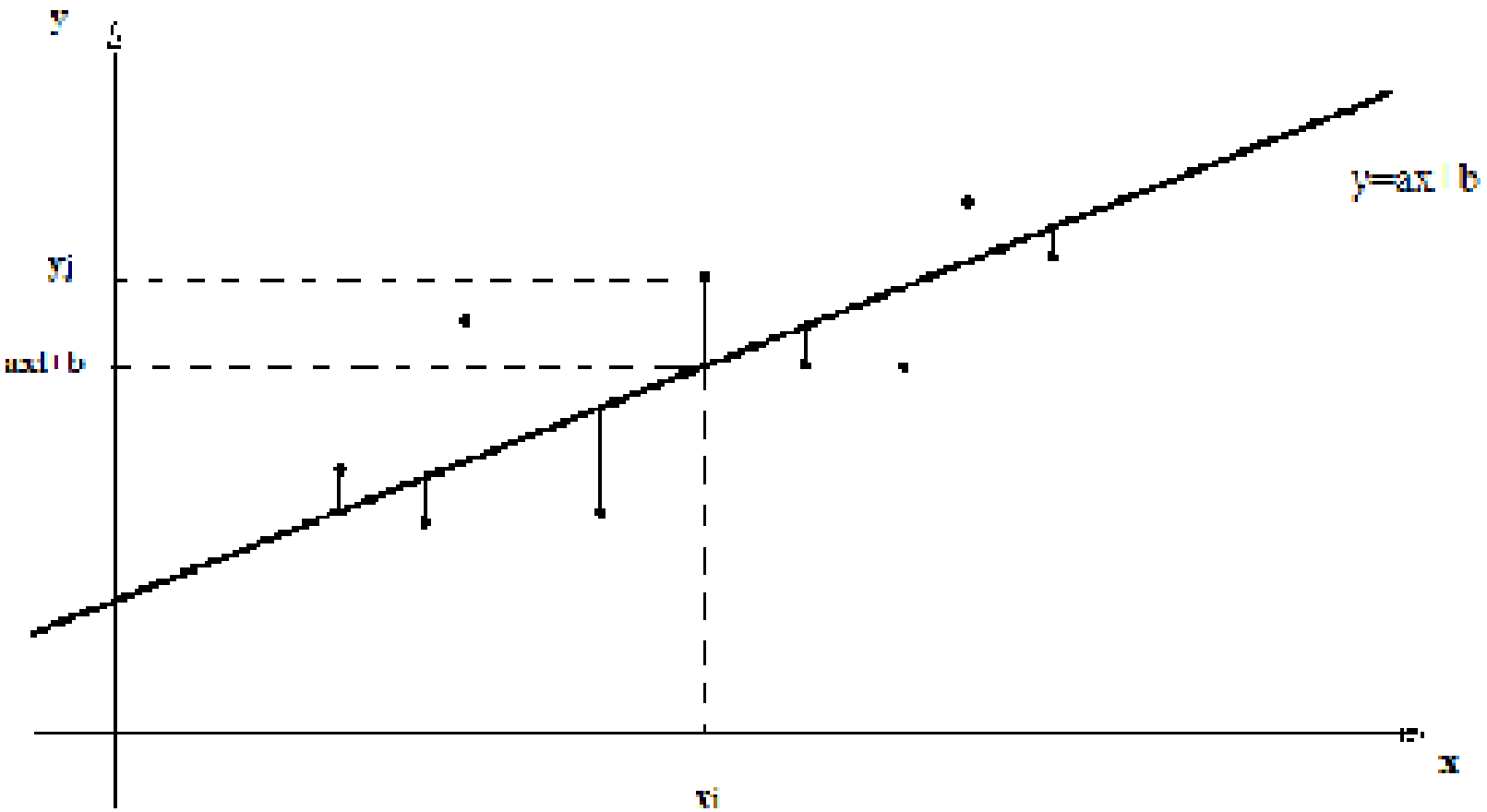
The linear correlation coefficient and the Pearson correlation ratio realize the following relationship $\rho^2(X,Y) \leq \eta^2_{Y/X}$.

Interpretation of results

1. If $\rho^2(X,Y) \leq \eta^2_{Y/X} \leq 0.1$ then Y and X are not correlated.
2. If $\rho^2(X,Y) \leq 0.1 < \eta^2_{Y/X} < 0.9$ then Y is partially linked to X but this link is not linear.
3. If $0.1 < \rho^2(X,Y) \leq \eta^2_{Y/X} < 0.9$ then Y is partially linked to X and this link is linear.
4. If $0.1 < \rho^2(X,Y) \leq 0.9 \leq \eta^2_{Y/X}$ then Y is linked to X and this link is functional but not linear.
5. If $0.9 < \rho^2(X,Y) \leq \eta^2_{Y/X}$ then Y is linearly related to X.

For the previous example we have $\eta^2_{Y/X} = \frac{3.96}{5.41} \simeq 0.7319$ **and** $\rho^2(X,Y) \simeq 0.7029$.

The linear fit consists in replacing the scatterplot by a line such that the estimated y-values along it, for the different x_i values are very close to the y_j values.



Definition

When $\rho^2(X,Y) > 0,9$ there exists a linear relationship between X and Y of the form $Y = aX + b$ which is called the regression line of Y in X and which minimize the sum

$$S = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \left(y_j - ax_i - b\right)^2 .$$

The condition $\rho^2(X,Y) > 0,9$ is mandatory to confirm the linearity of the relationship between X and Y but we can still determine a regression line for values less than $0,9$.

Theorem

The regression line from Y to X is the line of the form $Y = aX + b$ with

$$a = \frac{Cov(X,Y)}{\sigma_X^2} \text{ et } b = \overline{Y} - \frac{Cov(X,Y)}{\sigma_X^2}\overline{X}.$$

Remark

We can also determine the regression line from X to Y of the form $X = a'Y + b'$, où

$$a' = \frac{Cov(X,Y)}{\sigma_Y^2} \text{ et } b' = \overline{X} - \frac{Cov(X,Y)}{\sigma_y^2}\overline{Y}.$$

Theorem

The regression line from Y to X is the line of the form $Y = aX + b$ with

$$a = \frac{Cov(X,Y)}{\sigma_X^2} \text{ et } b = \overline{Y} - \frac{Cov(X,Y)}{\sigma_X^2}\overline{X}.$$

Remark

We can also determine the regression line from X to Y of the form $X = a'Y + b'$, où

$$a' = \frac{Cov(X,Y)}{\sigma_Y^2} \text{ et } b' = \overline{X} - \frac{Cov(X,Y)}{\sigma_y^2}\overline{Y}.$$

The two lines pass through the mean point $(\overline{X}, \overline{Y})$.

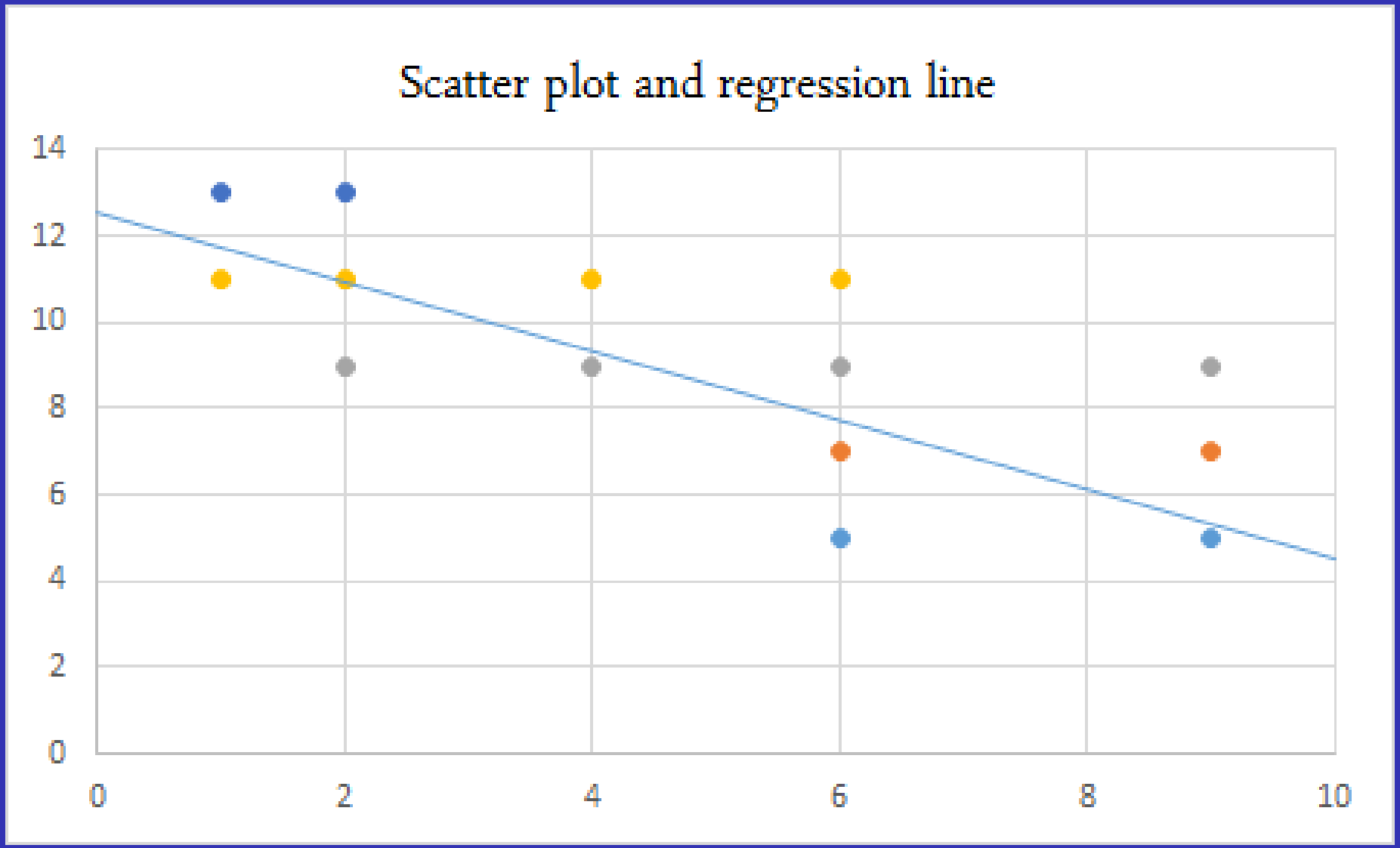
Example

We consider the previous example. The equation of the regression line from Y to X is determined.
We have $\rho^2(X,Y) \approx 0,7029 < 0,9$ the fit is not actually linear but there is a strong enough correlation that we can still fit it by a line of the form $(D_{Y/X}) : y = ax + b$ where $a = \frac{Cov(X,Y)}{\sigma_X^2}$ and $b = \overline{Y} - \frac{Cov(X,Y)}{\sigma_X^2}\overline{X}$, such that

$$\begin{aligned} a &= \frac{Cov(X,Y)}{\sigma_X^2} = \frac{-4,9552}{2,54^2} \approx -0,7681 \\ b &= \overline{Y} - \frac{Cov(X,Y)}{\sigma_X^2}\overline{X} = 8,84 + 0,7681 \cdot 4,78 = 12,5115 \end{aligned}$$

hence the equation of the regression line from Y to X is

$$(D_{Y/X}) : y = -0,7681x + 12,5115.$$

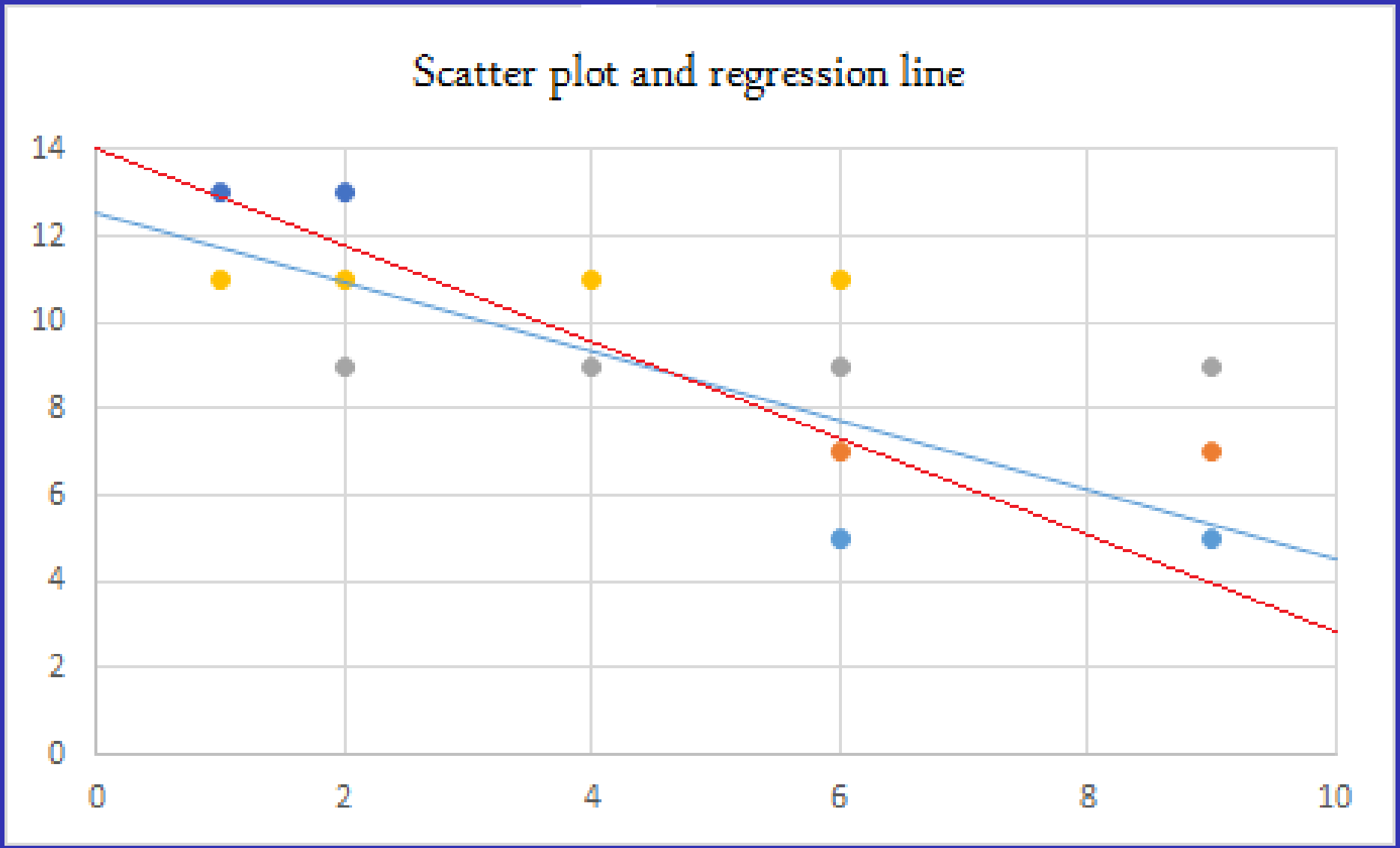

Scatter plot and regression line
14
12
10
8
6
4
2
0
0
2
4
6
8
10

We determine the equation of the regression line from X to Y.
$(D_{X/Y}) : x = a'y + b'$ such that:

$$\begin{aligned} a' &= \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{-4,9552}{2,3269^2} \approx -0,9152 \\ b' &= \overline{X} - \frac{Cov(X,Y)}{\sigma_Y^2}\overline{Y} = 4,78 + 0,9152 \times 8,84 \\ &= 12,8704 \end{aligned}$$

hence the equation of the regression line from X to Y is

$$(D_{X/Y}) : x = -0,9152y + 12,8704.$$


Scatter plot and regression line
14
12
10
8
6
4
2
0
0
2
4
6
8
10