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)

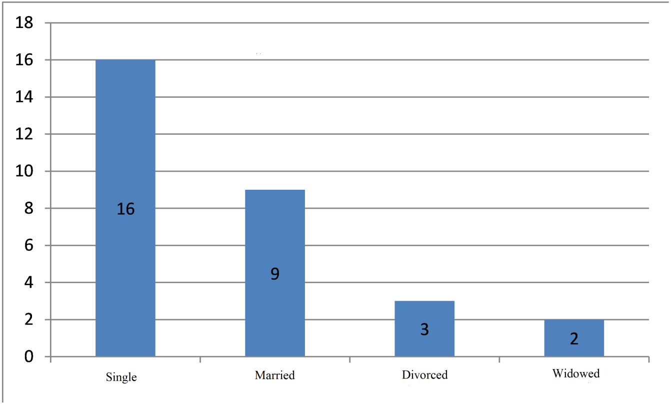
We take the example of the family situation of 30 employees.

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)

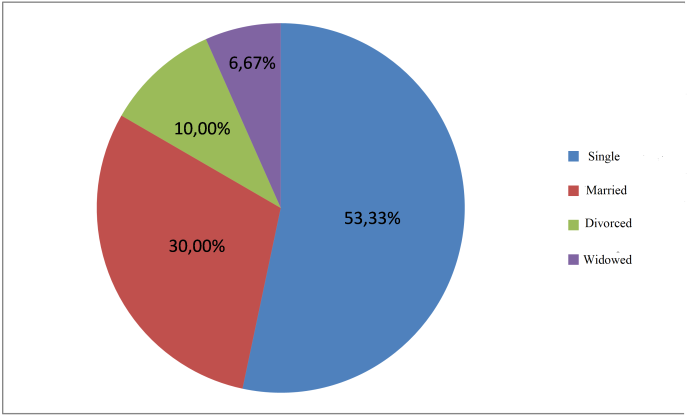
Sector for "Divorced"

$$\begin{array}{rcl} \theta_D & \longrightarrow & 3 \\ 360 & \longrightarrow & 30 \end{array} \implies \theta_D = \frac{3 \times 360}{30} = 36^\circ$$

Sector for "Widow"

$$\begin{array}{rcl} \theta_W & \longrightarrow & 2 \\ 360 & \longrightarrow & 30 \end{array} \implies \theta_W = \frac{2 \times 360}{30} = 24^\circ$$

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)

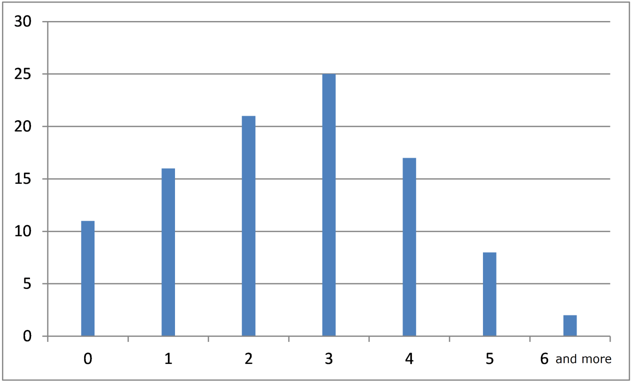
We take the example of the distribution of families according to the number of children.

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	—	—

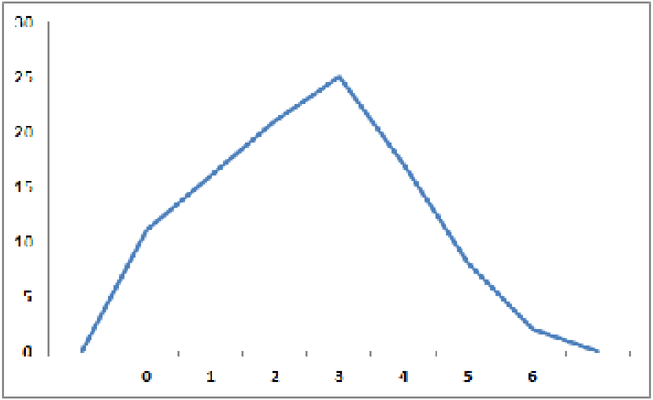
The example on the distribution of families according to the number of children is repeated.

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

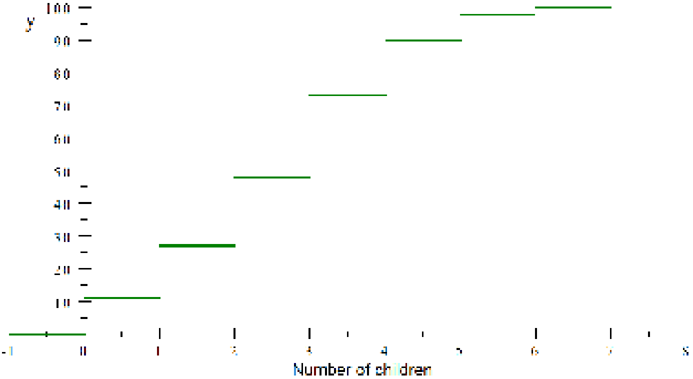


Example (2)

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

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

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 $Var\left(X\right) = \sigma_X^2 = 2,2891$.

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 $CV_X = 100\frac{1,51}{2,53} \approx 59,68\%$.