Descriptive Statistics and Probability

AMEZIANE BACHIR

2024/2025

Contents

Chapter 1

Study of one variable

1.1 Usual vocabulary

Descriptive statistics is a mathematical science pertaining collection, presentation, analysis and interpretation of data.

- Population. denoted by Ω : a well-defined collection of objects.
- Variable. denoted by X : characteristics of the objects.

If Ω is finite we denote the number of elements of Ω by $|\Omega|$ or $\#\Omega$.

- Observation. denoted by x_i : an observed value of a variable.
- Data. a collection of observations (x_i) .

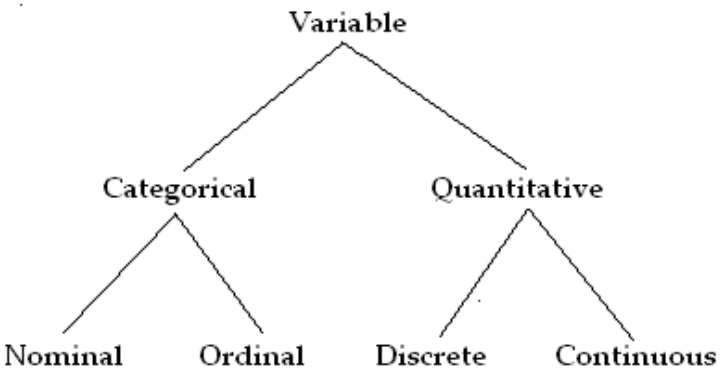
This is represented in the usual mathematical formalism as follows : $X : \Omega \longrightarrow E$
$\omega \longrightarrow X(\omega)$

$X(\Omega)$ is called the set of modalities.
descriptive statistics $\longrightarrow$ study data $\longrightarrow$ understand the population

The variable can be :

- Qualitative(categorial). described by a word or phrase wich can be nominal or ordinal e.g. gender of a person (male, female), blood group, color
- Quantitative (numerical). described by a number wich can be
- Discrete. can take a finite or countable number e.g. $X = \#$ of students in a class

- Continuous. can take any value in an interval e.g. X = height of a student



1.1.1 Frequencies

Suppose that we have collected data from Ω, $|\Omega| = n$ and $X(\Omega) = \{x_1, \ldots, x_r\}$ such that
$x_1 \leq x_2 \leq \cdots \leq x_r$.
Frequency. frequency of a value x_i denoted by n_i is the number of observations taking that value
Relative Frequency. denoted by f_i, $f_i = \dfrac{n_i}{n}$
Percent Frequency. denoted by p_i, $p_i = f_i \times 100\%$
Cumulative Frequency (Relative Cumulative Frequency). denoted by N_i (F_i) and defined by $N_i = \sum_{j \leq i} n_j = n_1 + n_2 + \cdots + n_i$ $\left(F_i = \sum_{j \leq i} f_j = \dfrac{N_j}{n} \right)$
We can from now present a data using the following table called frequency table

X	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_r	f_r	N_r	F_r
	n	1		

Observe that, we have always $N_r = n$ and $F_r = 1$.
Example 1 : In a survey of 20 families in a village, the number of children per family was recorded and the following data obtained

$$4; 3; 2; 1; 3; 0; 2; 3; 2; 4; 5; 2; 0; 2; 3; 4; 5; 2; 3; 2$$

This data is organized in the following table :

X	n_i	f_i	N_i	F_i
0	2	$0,1$	2	$0,1$
1	1	$0,05$	3	$0,15$
2	7	$0,35$	10	$0,5$
3	5	$0,25$	15	$0,75$
4	3	$0,15$	18	$0,9$
5	2	$0,1$	20	1
	20	1		

1.1.2 Grouping by classes

Sometimes we need to group the modalities of a variable into a certain number of intervals called classes or bins. This is the case when the variable is continuous or when it is discrete and only some of its modalities are of interest for the study.
In this case, we divide the set E of modalities into intervals $[a_0, a_1[$, $[a_1, a_2[$, ..., $[a_{r-2}, a_{r-1}[$, $[a_{r-1}, a_r]$ which form a partition of E:

$$E = [a_0, a_1[\cup[a_1, a_2[\cup... \cup [a_{r-2}, a_{r-1}[\cup[a_{r-1}, a_r].$$

The interval $[a_{i-1}, a_i[$ is called class i.
In this case, the number of individuals n_i is the number of individuals whose modality belongs to the interval $[a_{i-1}, a_i[$. We then define f_i, N_i and F_i in the same way as in the previous section.
In the case of grouping by classes, we define the center (class mark) and the amplitude (length or width) of class i respectively by

$$c_i = \frac{a_{i-1} + a_i}{2} \text{ and } \quad l_i = a_i - a_{i-1}.$$

Finally, in order to take into account the amplitude of class i when evaluating its frequecy, we define the frequency density (or relative frequency density) by:

$$h_i = \frac{n_i}{l_i} \text{ (or } \frac{f_i}{l_i}) \ .$$

This is the case, for example, in the definition of the empirical distribution function (see section 1.2.2) below.

1.2 Usual graphic representations.

1.2.1 discrete case

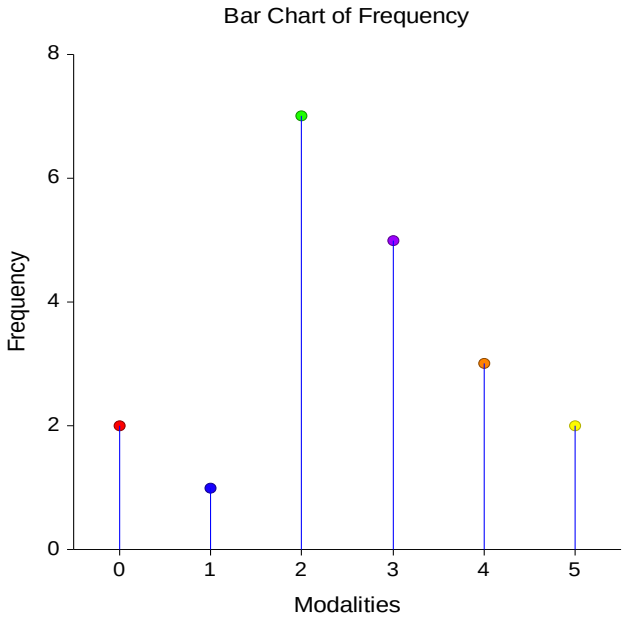
Consider a finite, quantitative or qualitative ordinal variable $X : \Omega \longrightarrow E$, with $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $E = \{x_1, ..., x_r\}$. We order the modalities of X in increasing order:

$$x_1 < x_2 < \cdots < x_r.$$

Bar chart.

On the the $x-$axis we place the modalidies in ascending order, and on the ordinate we place the frequencies or the relative frequencies. The height of the bar from x_i is equal to n_i (or f_i). For each case, we specify whether the diagram is proportional to the frequencies or to the relative frequencies.

The bar chart (relative to frequencies) corresponding to the example 1 is as follows:



Cumulative curve.

On the the $x-$axis we place the modalidies in ascending order, and on the ordinate we place the cumulative frequencies or the relative cumulative frequencies . As with the bar chart, in each case we specify whether the cumulative curve is proportional to the cumulative frequencie or to the relative cumulative frequencies.

Definition : The cumulative curve relating to the relative cumulative frequencies is the graph of the function F_X defined by

$$F_X(x) = \begin{cases} 0 & \text{if } x < x_1 \\ F_i & \text{if } x \in [a_i, a_{i+1}[\,,\ i = 1, \ldots, r-1 \\ 1 & \text{if } x \geq x_r \end{cases}$$

F_X is called the empirical distribution function of the variable X. We define in the same way the cumulative curve relating to the cumulative frequencices.

The cumulative diagram (relative to the relative cumulative frequencies) corresponding

to the example 1 is as follows:

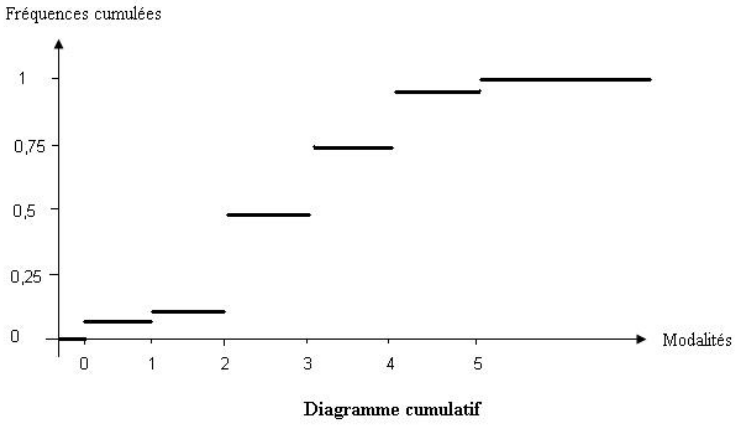


Diagramme cumulatif

1.2.2 Continuous case

Here we consider the case of a continuous variable for which the modalities have been grouped into the following classes:

$$[a_0, a_1[,\ [a_1, a_2[, ...,\ [a_{r-1}, a_r[$$

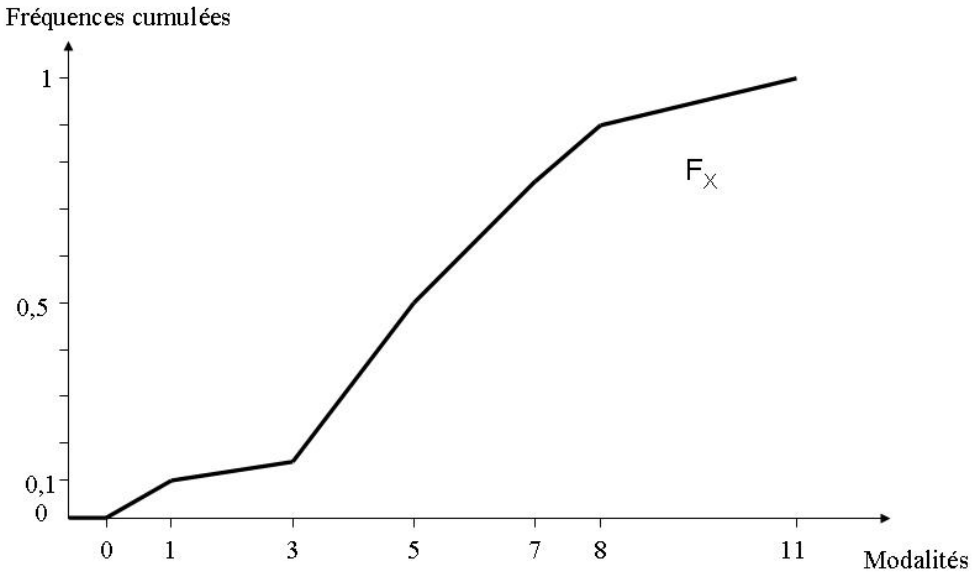
Cumulative curve.
This is the equivalent of the cumulative diagram (relative to relative cumulative frequencies) in the discrete case. In this case, the empirical distribution function can be represented using linear interpolations.
Definition**.** The cumulative curve of the variable X is the graph of the function F_X defined by $F_X(0) = 0$ and :

$$F_X(x) = \begin{cases} 0 & \text{if } x < a_0 \\ F_{i-1} + h_i\ (x - a_{i-1}) & \text{if } x \in [a_{i-1}, a_i[\ ,\ i = 1, \ldots, r-1 \\ 1 & \text{if } x \geq a_r \end{cases}$$

F_X is called the empirical distribution function of the variable X.
We recall that $h_i = \dfrac{f_i}{a_i - a_{i-1}}$ is the frequency density associated with the class $[a_{i-1}, a_i[$ defined in section 1.1.2.
Example 2 **:** We give the following relative cumulative frequency table

Modalities	F_i
$[0, 1[$	$0,1$
$[1, 3[$	$0,15$
$[3, 5[$	$0,5$
$[5, 7[$	$0,75$
$[7, 8[$	$0,9$
$[8, 11[$	1

The cumulative diagram (relative to the relative cumulative frequencies) corresponding to this example is as follows:



Histogram.

How to create a histogram for a continuous data set?

On the the $x-$axis we place the classes, and on the ordinate we place the cumulative frequencies or the relative cumulative frequencies. Draw a rectangle above each class

- For equal class width case, rectangle height = relative frequency.

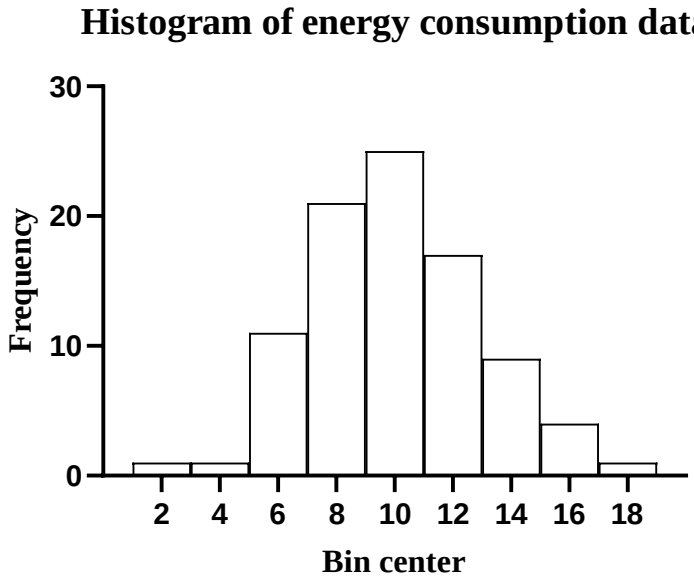
- For unequal class width case :

$$\text{rectangle height} = \text{frequency density } h_i = \frac{\text{relative frequency of the class}}{\text{class width}}.$$

The area of the rectangle is proportional to the relative frequency.

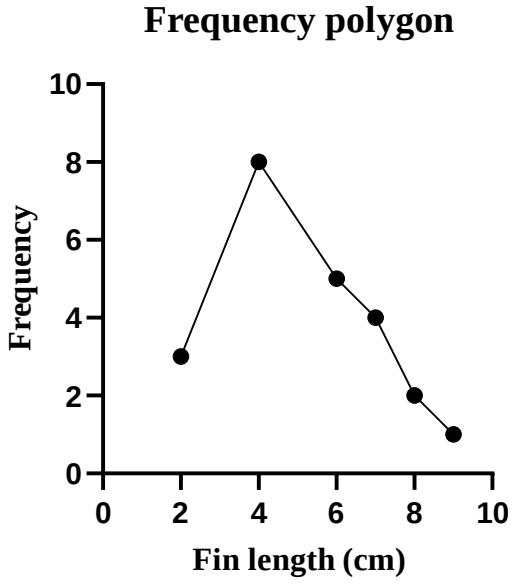
Example 3 : Adjusted energy consumption during a particular period for a 90 gas-heated homes are recorded as follows :

Bin	$[1,3[$	$[3,5[$	$[5,7[$	$[7,9[$	$[9,11[$	$[11,13[$	$[13,15[$	$[15,17[$	$[17,19[$
n_i	1	1	11	21	25	17	9	4	1
f_i	$0,011$	$0,011$	$0,122$	$0,233$	$0,278$	$0,189$	$0,100$	$0,044$	$0,011$



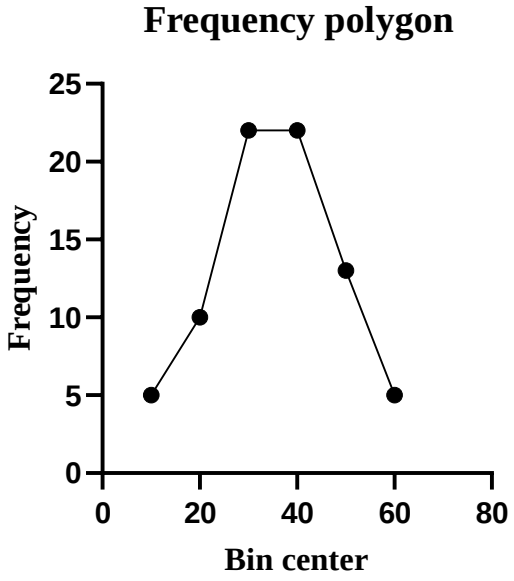
Frequency polygons.
A frequency distribution may be displayed as a frequency polygon.
A frequency polygon may be superimposed on a histogram by joining the midpoints of the tops of the rectangles. This is in grouped data.
(a) Ungrouped data : The fin length in (cm) of particular type of fish is given. Draw a frequency polygon to illustrate this information.

Fin length (cm)	2	4	6	7	8	9
Frequency	3	8	5	4	2	1



(b) Grouped data: The following table shows the age distribution of cases of a certain disease reported during a year in a particular state. Prepare a frequency polygon.

Age	Number of cases
$[5, 15[$	5
$[15, 25[$	10
$[25, 35[$	22
$[35, 45[$	22
$[45, 55[$	13
$[55, 65[$	5

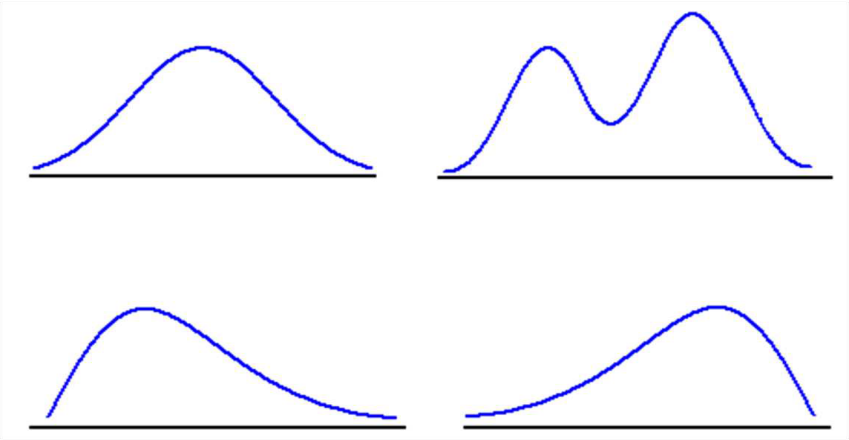


Polygons shapes.

Polygons have a variety of shapes, the shape of a polygon conveys important information about the distribution of data.

- Unimodal: Single peak
- Bimodal: Two peaks
- Multimodal: Two more peaks
- Symmetric: Left $\simeq$ right
- Positively skewed: Right tail stretching out

• Negatively skewed: Left tail stretching out



1.3 Measures of central tendency

One of the most important objectives of statistical analysis is to get one single value that describes the characteristic of the entire mass of data.
There are three main statistical measures which attempt to locate a 'typical'value. These are
1. Mean
2. Median
3. Mode

Mean.
This only concerns quantitative variables. Consider a variable $X : \Omega \longrightarrow E$, with $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $E = \{x_1, \ldots, x_r\}$.
Definition : The mean of the variable X is denoted by $\overline{X}$. This is defined by :
For ungrouped data :

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

Grouped data

$$\overline{X} = \frac{\sum_{j=1}^{r} n_i c_i}{n}$$

where c_i is a class center.
property : Consider a variable $Y : \Omega \longrightarrow F$ defined on the same population as X (which modalities set F is not necessarily equal to E). Let a and b be two real numbers, then $aX + bY$ defines a third variable on Ω having mean :

$$\overline{aX + bY} = a\overline{X} + b\overline{Y}.$$

Example 4 : Find the mean of set of numbers $63, 65, 67, 68, 69, 70, 71, 72, 74, 75$.

$n = 10$

$$\begin{aligned} \sum x &= 63 + 65 + 67 + 68 + 69 + 70 + 71 + 72 + 74 + 75 = 694 \\ \overline{X} &= \frac{\sum x}{n} = \frac{694}{10} = 69,4. \end{aligned}$$

Frequency distribution

No. of flowers (X)	1	2	3	4	5
No. of plants (n)	11	10	5	3	1

$\overline{X} = \dfrac{\sum n_i x_i}{n} = \dfrac{63}{30} = 2,1.$

Grouped frequency distribution

The lengths of 32 leaves were measured correct to the nearest mm. Find the mean length.

Length (mm)	$[20, 22[$	$[22, 24[$	$[24, 26[$	$[26, 28[$	$[28, 30[$
Frequency	3	6	12	9	2

Length (mm)	Midpoint (x_j)	n_j	$n_j x_j$
$[20, 22[$	21	3	63
$[22, 24[$	23	6	138
$[24, 26[$	25	12	300
$[26, 28[$	27	9	243
$[28, 30[$	29	2	58

$\overline{X} = \dfrac{\sum n_i x_i}{n} = \dfrac{802}{32} = 25,06\ mm.$

Median. This concerns quantitative or qualitative ordinal variables. It is a value noted Q_2 or M which divides the population into 2 groups of equal size : that of individuals whose modality value is less than M and that of individuals whose modality value is greater than M.

Example 5 : The median of the following series $1, 5, 6, 7, 34, 50, 176$ is equal to 7.

We will give how to calculate the median in the next section.

Mode : Mode is the value which occurs most frequently in a set of observations.

The mode may or may not exist, and even if it does exist, it may not be unique. A distribution having a unique mode is called 'unimodal' and one having more than one is called 'multimodal'.

Example 6 :

The set $2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18$

has mode 9.

The set $3, 5, 8, 10, 12, 15, 16$ has no mode.

The set $2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9$ has two modes, 4 and 7, and is called bimodal.

1.4 Measures of dispersion

These indicate how the data is distributed relative to each other.
Range. This is the difference between the minimum value and the maximum value of modalities

$$R = \max\{x_1, \ldots, x_r\} - \min\{x_1, \ldots, x_r\}.$$

Quartiles.
Median. separates the data into two parts: lower sub-data and upper sub-data. n odd : $\left\{x_{(1)}, \ldots, x_{\left(\frac{n+1}{2}\right)}\right\}$ and $\left\{x_{\left(\frac{n+1}{2}\right)}, \ldots, x_{(n)}\right\}$ n even : $\left\{x_{(1)}, \ldots, x_{\left(\frac{n}{2}\right)}\right\}$ and $\left\{x_{\left(\frac{n}{2}\right)}, \ldots, x_{(n)}\right\}$.
where $\left(x_{(i)}\right)$ are the data arranged in ascending ordre.
Quartiles divide the lower and upper sub-data into two parts:
– 1st Quartile: Q_1 = median of the lower sub-data, also called the lower fourth.
– 2nd Quartile: Q_2 = median of the entire data.
– 3rd Quartile: Q_3 = median of the upper sub-data, also called the upper fourth.
– The first Q_1, second Q_2 and third Q_3 quartiles divide the distribution into four equal parts.
– Inter Quartile Range:$IQR = Q_3 - Q_1$, also called fourth spread.
Calculating Quartiles.
For discrete data.
Method 1. From data arranged in ascending order of magnitude

$$Q_j = \frac{1}{2}\left(x_{\left[\frac{jn}{4}\right]} + x_{\left[\frac{jn+1}{4}\right]}\right), \quad j = 1, 2, 3.$$

where $[x]$. denotes the smallest integer greater than or equal to x.
Method 2. From frequency table we refer to the values of relative cumulative frequency : If $0,5$ belongs to the values of F_i, let $F_j = 0,5$ then $M = \dfrac{x_j + x_{j+1}}{2}$, if it doesn't belong let $F_{j-1} < 0,5 < F_j$ then $M = x_j$.
The method for Q_1 and Q_3 is the same, with Q_1 we use $0,25$ and with Q_3 we use $0,75$.
Example 7 : calculate the quartile for data in example 1

$$4; 3; 2; 1; 3; 0; 2; 3; 2; 4; 5; 2; 0; 2; 3; 4; 5; 2; 3; 2$$

Using Method 1 **:** we arrange the data in ascending order

$$0, 0, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5$$

$$Q_1 = \frac{1}{2}\left(x_{\left[\frac{n}{4}\right]} + x_{\left[\frac{n+1}{4}\right]}\right) = \frac{1}{2}\left(x_{\left[\frac{20}{4}\right]} + x_{\left[\frac{21}{4}\right]}\right) = \frac{1}{2}\left(x_{[5]} + x_{[5,25]}\right)$$
$$= \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(2 + 2) = 2.$$
$$Q_2 = \frac{1}{2}\left(x_{\left[\frac{2n}{4}\right]} + x_{\left[\frac{2n+1}{4}\right]}\right) = \frac{1}{2}\left(x_{\left[\frac{40}{4}\right]} + x_{\left[\frac{41}{4}\right]}\right) = \frac{1}{2}\left(x_{[10]} + x_{[10,25]}\right)$$
$$= \frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(2 + 3) = 2,5.$$

$$Q_3 = \frac{1}{2}\left(x_{\left[\frac{3n}{4}\right]} + x_{\left[\frac{3n+1}{4}\right]}\right) = \frac{1}{2}\left(x_{\left[\frac{60}{4}\right]} + x_{\left[\frac{61}{4}\right]}\right) = \frac{1}{2}\left(x_{[15]} + x_{[15,25]}\right)$$
$$= \frac{1}{2}\left(x_{15} + x_{16}\right) = \frac{1}{2}\left(3 + 4\right) = 3,5.$$

Method 2 :

$0,25 \notin F_i$, we have $\underbrace{0,15}_{F_{j-1}} < 0,25 < \underbrace{0,5}_{F_j}$
then $Q_1 = x_j = 2$.
$0,5 \in F_i$, then $M = \dfrac{x_j + x_{j+1}}{2} = \dfrac{2+3}{2} = 2,5.$
$0,75 \in F_i$ then $Q_3 = \dfrac{x_j + x_{j+1}}{2} = \dfrac{3+4}{2} = 3,5.$

X	n_i	f_i	N_i	F_i
0	2	0,1	2	0,1
1	1	0,05	3	0,15
2	7	0,35	10	$0,5 \longleftarrow F_j$
3	5	0,25	15	$0,75 \longleftarrow F_j$
4	3	0,15	18	0,9
5	2	0,1	20	1
	20	1		

.

For continuous data.
The quartiles M (Q_1) and (Q_3) are the values of the variable X defined by : $F_X(M) = 0,5$ $(0,25)$ and $(0,75)$ where F_X is the relative cumulative frequency Function of X in continuous case.
To calculate the Median : we first identify the class containing the median, then we apply the following formula :

$$M = L_1 + \frac{\frac{n}{2} - C}{n_m}\left(L_2 - L_1\right)$$

Or equivalently

$$\left(M = L_1 + \frac{0,5 - C_r}{f_m}\left(L_2 - L_1\right)\right)$$

where

L_1 is the lower bound class containing the median;
n is the total frequency;
C is the cumulative frequency just before the median class;
n_m is the frequency of the median class;
L_2 is the upper bound class containing the median.

How to determine the class containing the median :
First : if $\dfrac{n}{2}$ belongs to the values of N_i then $M =$ the upper bound class face to $\dfrac{n}{2}$.
Second : If $\dfrac{n}{2}$ doesn't belong, let $N_{j-1} < \dfrac{n}{2} < N_j$, then the class median is the class face to N_j and we apply the above formula.
We proceed similarly with Q_1 and Q_3, by replacing $\dfrac{n}{2}$ by $\dfrac{n}{4}$ and $\dfrac{3n}{4}$ in the formula.

Example 8 **:** calculate the quartiles of the data

Length (mm)	$[20,22[$	$[22,24[$	$[24,26[$	$[26,28[$	$[28,30[$
Frequency	3	6	12	9	2

Length (mm)	n_i	N_i
$[20,22[$	3	3
$[22,24[$	6	9
$[24,26[$	12	21
$[26,28[$	9	30
$[28,30[$	2	32

$\frac{n}{2} = 16$, $9 < 16 < 21$, then $M \in [24,26[$ and

$$\begin{aligned} M &= L_1 + \frac{\frac{n}{2} - C}{n_m}(L_2 - L_1) = 24 + \frac{16-9}{12}(26-24) \\ &= 25,16. \end{aligned}$$

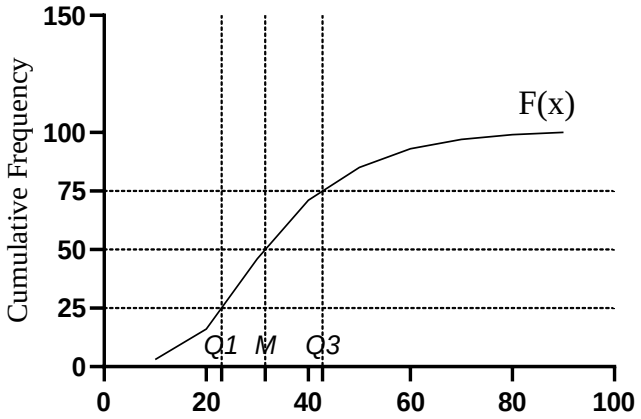
$\frac{n}{4} = 8$, $3 < 8 < 9$, then $Q_1 \in [22,24[$ and

$$\begin{aligned} Q_1 &= L_1 + \frac{\frac{n}{4} - C}{n_m}(L_2 - L_1) = 22 + \frac{8-3}{6}(24-22) \\ &= 23,66. \end{aligned}$$

$\frac{3n}{4} = 24$, $21 < 24 < 30$, then $Q_3 \in [26,28[$ and

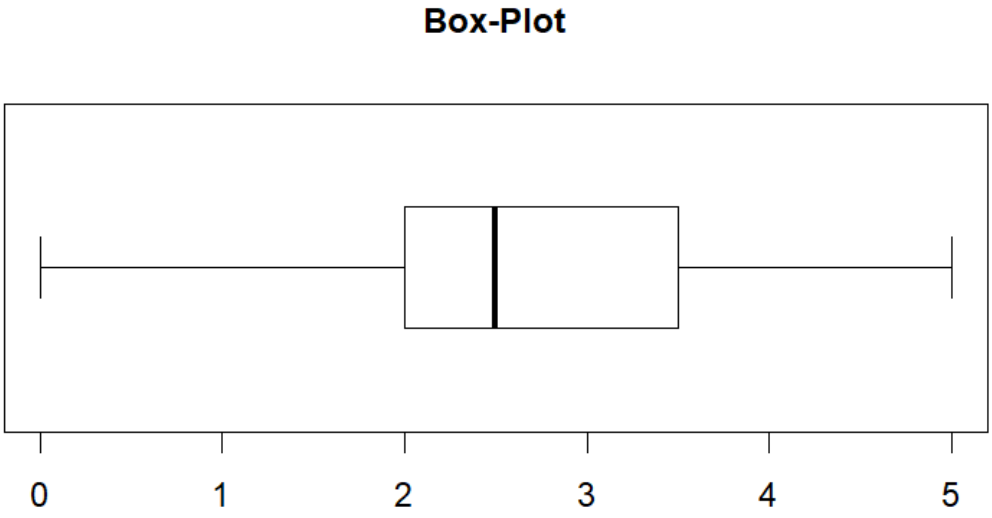
$$\begin{aligned} Q_3 &= L_1 + \frac{\frac{3n}{4} - C}{n_m}(L_2 - L_1) = 26 + \frac{24-21}{9}(28-26) \\ &= 26,66. \end{aligned}$$

Remark : We can use The F_X-graph to calculate the quartiles as below



Box-plots.
Boxplot is very useful in describing several of a data set's important features such as: center, spread and symmetry.
1. Draw a horizontal axis, find Q_1, Q_2 and Q_3 and calculate IQR.
2. Place a rectangle above the axis, with the left edge at Q_1, right edge at Q_3.
3. Place a vertical line segment inside the rectangle at the location of Q_2.
4. Draw whiskers out from each end of the rectangle to the smallest and largest obs.

Example 9 : Draw the Box-plot for data in Example 1
We have $\min x_i = 0,\ \max x_i = 5,\ Q_1 = 2,\ M = 2,5,\ Q_3 = 3,5$



Variance. We consider a quantitative variable $X : \Omega \longrightarrow E$, with $|\Omega| = n$ and $E = \{x_1, \ldots, x_r\}$.
Definition : The variance of the variable X is denoted by $Var(X)$ or σ^2. This is defined by

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

There is another way to express variance which is in fact the most commonly used form

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

Properties :
– We have always $Var\left(X\right) \geq 0$.
– If a and b are two real numbers then : $Var\left(aX + b\right) = a^2 Var\left(X\right)$.

Standard deviation. This is noted by $\sigma(X)$ and is defined by

$$\sigma(X) = \sqrt{Var\,(X)}.$$

Properties :
– We have always $\sigma\,(X) \geq 0$.
– If a and b are two real numbers, we have : $\sigma\,(aX + b) = |a|\,\sigma\,(X)$.
Example 10 : The mean, variance and standard deviation of the following statistical data :
3, 5, 8, 3, 16, 5, 5, 1, 0, 10, 10 are respectively

$$\overline{X} \simeq 5,82, \quad Var\,(X) \simeq 19,85, \quad \sigma\,(X) \simeq 4,46.$$