LARBI BEN M'HIDI-OUM EL BOUAGHI UNIVERSITY

FACULTY OF EXACT SCIENCES AND NATURAL AND LIFE SCIENCES

DEPARTEMENT OF MATHEMATICS AND COMPUTER SCIENCE

First year licence

Second semester

Introduction to probability and descriptive statistics

Lecturer teacher:

Besma Bennour

2023-2024

Contents

Chapter 1

Bacis concepts and statistical vocabulary

- **Statistics :** is a branch of mathematics dealing with collecting, organizing, summarizing, analysis and making decision from data.
- **Fields :** it can be found in all sciences.
- **Statistical software** like : Excel, SPSS, R, Matlab, Python, ...
- **Types of statistics :** statistics is divided into two mains areas, wich are descriptive and inferential statistics.
  1. Descriptive statistics deals with methods for collecting, organizing, and describing data by using tables, graphs, and measures.
  2. Inferential statistics deals with methods that use sample results, to help in estimation or make decisions about the population.

In this book, we're going to study only the first one.

In the first chapter, we're going to present some statistical vocabulary (like population and variable) and some basic concepts (frequency, relative frequency, percentage, increasing cumulative frequency, and increasing cumulative relative frequency).

1.1 Statistical vocabulary : key terms

1.1.1 Population

A population is the set of all elements under study. It can be a collection of any persons, things, or objects, for examples: persons set, books, animals, machines, computer, or departement, ...

* The population size is noted by N.

1.1.2 Sample

A sample is the subset of the population. The sample size is noted by n.

1.1.3 Element or individuel

An element (or member) of a sample or population is a specific subject or object about which the information is collected. It's noted by w.

1.1.4 Caracter or variable

A statistical variable is all application denoted X such that:

$$\begin{aligned} X \ : \ & \mathbf{P} \rightarrow \mathbb{R} \\ & w \rightarrow X(w) \end{aligned}$$

* A variable is a characteristic of interest for each person or thing in a population or a sample.

1.1.5 Types of variable

We have two types of variables: *quantitative and qualitative.*

1. A variable is **quantitative** if the data set is a set of numbers. Quantitative variable may be discrete or continuous :

- ***Discrete*** variable assumes values that can be counted. For example : number of children, number of error, number of people living in a town, number of machines in a gym, number of accidents, ...
- ***Continuous*** variable assumes all values between any two specific values, i.e. they take all values in an interval. For example : distance, age, lifetime, height, weight, ...

2. A variable is **qualitative** if the data set is a set of names or labels (i.e. it takes non-numerical values).

   Qualitative variables can be ordinal or nominal :

   - A variable is ***ordinal*** when its values can be ordered. For example : level of stadies, mention au bac, classe d'âge, stade d'une maladie, ...
   - A variable is ***nominal*** when its values can not be ordered. For example : genders (male, female), nationality, profession, religious affiliation (muslim, Christian, ...), blood type (A, B, AB, O), hair color, ...

1.1.6 Data and values

* The value of a variable for an element is called **an observation** or **a measurement**. The value may be number or word.

* The set of all values is called data set.

1.1.7 Frequency of value

A frequency is the number of times a value x_i of the data set occurs. This number is denoted by n_i.

* The population size n is given by: $n = n_1 + ... + n_k = \sum_{i=1}^{k} n_i$

1.2 Basic concepts

1.2.1 Relative frequency of value

A relative frequency is the ratio (fraction or proportion) of the frequency n_i to the total number n.

$$f_i = \frac{n_i}{N}$$

* The relative frequency f_i is always between 0 and 1.

* We have : $f_1 + ... + f_k = \sum_{i=1}^{k} f_i = 1$.

* The percentage of a value x_i is the number $p_i = f_i \times 100$.

1.2.2 Cumulative frequency

We have two types :

Increasing Cumulative Frequency $N_x \uparrow$ and Decreasing Cumulative Frequency $N_x \downarrow$.

* Case of quantitative discrete data :

1. **Increasing Cumulative Frequency (ICF)** of a value $x \in \mathbb{R}$ is the sum of the frequencies n_i of values x_i such as $x_i \leq x$. It's noted by $N_x \uparrow$.

$$N_x \uparrow = \sum_{i\,:\,x_i \leq x} n_i \ , \quad x \in \mathbb{R}$$

Case particular: if $x = x_i$ we obtain $N_{x=x_i} \uparrow$.

2. **Decreasing Cumulative Frequency (DCF)** of a value $x \in \mathbb{R}$ is the sum of the frequencies n_i of values x_i such as $x_i > x$. It's noted by $N_x \downarrow$.

$$N_x \downarrow = \sum_{i:x_i > x} n_i \quad ou \quad N_x \downarrow = n - N_x \uparrow$$

Case particular: if $x = x_i$ we obtain $N_{x=x_i} \downarrow$.

* Case of quantitative continuous data :

1. **Increasing Cumulative Frequency (ICF)** of a value $x \in \mathbb{R}$ is given by :

$$N_x \uparrow = \sum_{i\,:\,x_i < x} n_i \ , \quad x \in \mathbb{R}$$

Case particular: if $x = e_i$ we obtain $N_{x=e_i} \uparrow$.

2. **Decreasing Cumulative Frequency (DCF)** of a value $x \in \mathbb{R}$ is given by :

$$N_x \downarrow = \sum_{i:x_i \geq x} n_i \quad ou \quad N_x \downarrow = n - N_x \uparrow$$

Case particular: if $x = e_i$ we obtain $N_{x=e_i} \downarrow$.

1.2.3 Relative frequency cumulative

We have also two types : increasing cumulative relative frequency $F_x \uparrow$ and decreasing cumulative relative frequency $F_x \downarrow$.

* Case of quantitative discrete data :

1. **Increasing Cumulative Relative Frequency (ICRF)** of a value $x \in \mathbb{R}$ is the sum of the relative frequencies f_i of values x_i such as $x_i \leq x$. It's noted by $F_x \uparrow$.

$$F_x \uparrow = \sum_{i\,:\,x_i \leq x} f_i \ , \quad x \in \mathbb{R}$$

Case particular: if $x = x_i$ we obtain $F_{x=x_i} \uparrow$.

2. **Decreasing Cumulative Relative Frequency (DCF)** of a value $x \in \mathbb{R}$ is the sum of the relative frequencies f_i of values x_i such as $x_i > x$. It's noted by $F_x \downarrow$.

$$F_x \downarrow = \sum_{i:x_i > x} f_i \quad ou \quad F_x \downarrow = 1 - F_x \uparrow$$

Case particular: if $x = x_i$ we obtain $F_{x=x_i} \downarrow$.

* Case of quantitative continuous data :

1. **Increasing Cumulative Relative Frequency (ICRF)** of a value $x \in \mathbb{R}$ is given by :

$$F_x \uparrow = \sum_{i\,:\,x_i < x} f_i \ , \quad x \in \mathbb{R}$$

Case particular: if $x = e_i$ we obtain $F_{x=e_i} \uparrow$.

2. **Decreasing Cumulative Relative Frequency (DCRF)** of a value $x \in \mathbb{R}$ is given by :

$$F_x \downarrow = \sum_{i:x_i \geq x} f_i \quad ou \quad F_x \downarrow = 1 - F_x \uparrow$$

Case particular: if $x = e_i$ we obtain $F_{x=e_i} \downarrow$.

1.3 Data set and frequency table

1.3.1 Case of a quantative discrete data

Example : Twenty students were asked " how many hours they worked per day ? ". Their responses, in hours, are as follows :

$$5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3$$

Based on this data, we find the following frequency table:

Values x_i	2	3	4	5	6	7	$\sum$
Frequencies n_i	3	5	3	6	2	1	$n = 20$

- The population studied is the set (group) of students.
- The population size : $n = 20$.
- The variable X studied is the number of working hours per day.
- The type of X is a quantitative discrete.

We add the lines for calculating $f_i, p_i, N_{x=x_i} \uparrow, F_{x=x_i} \uparrow$ as follows :

Values x_i	2	3	4	5	6	7	$\sum$
Frequencies n_i	3	5	3	6	2	1	$n = 20$
Relative frequencies f_i	0.15	0.25	0.15	0.3	0.1	0.05	1
Percentages $p_i\%$	15	25	15	30	10	5	100
ICF $N_{x=x_i} \uparrow$	3	8	11	17	19	20	///
DCF $N_{x=x_i} \downarrow$	17	12	9	3	1	0	///
ICRF $F_{x=x_i} \uparrow$	0.15	0.4	0.55	0.85	0.95	1	///
DCRF $N_{x=x_i} \downarrow$	0.85	0.6	0.45	0.15	0.05	0	///

1.3.2 Case of quantitative continuous data

Example : We measured the height in cm of a group of people and found the following results:

153	165	160	150	159	151	163	160	158	149
154	153	163	140	158	150	158	155	163	159
157	162	160	152	164	158	153	162	166	162
165	157	174	158	171	162	155	156	159	162
152	158	164	164	162	158	156	171	164	158

- The population studied is the set (group) of persons.
- The population size : $n = 50$.
- The variable X studied is the height per person.
- The type of X is a quantitative continuous.

- According to Sturge's rule, the number of classes is:

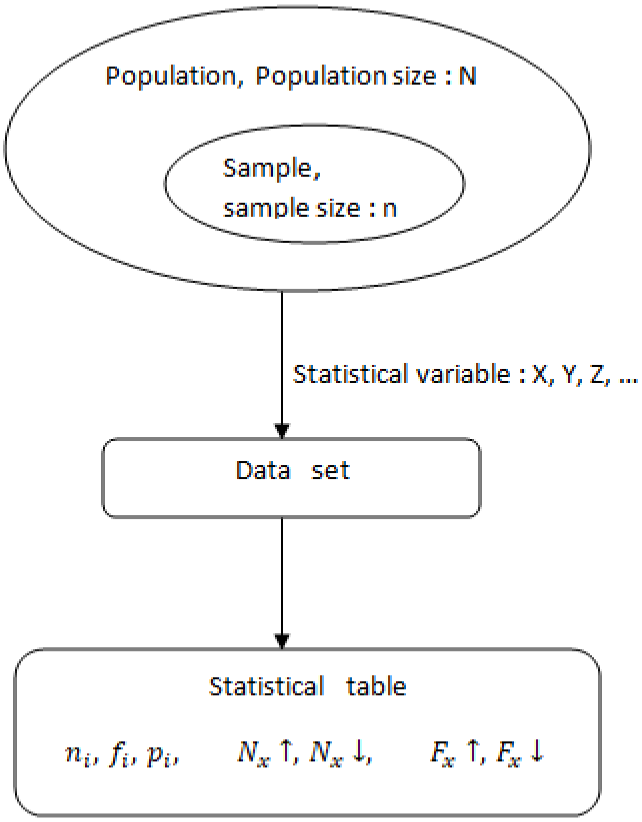
$$N_{classes} = 1 + 3.3\ log(n) = 6.61 \simeq 7$$

and according to Yule's rule we have :

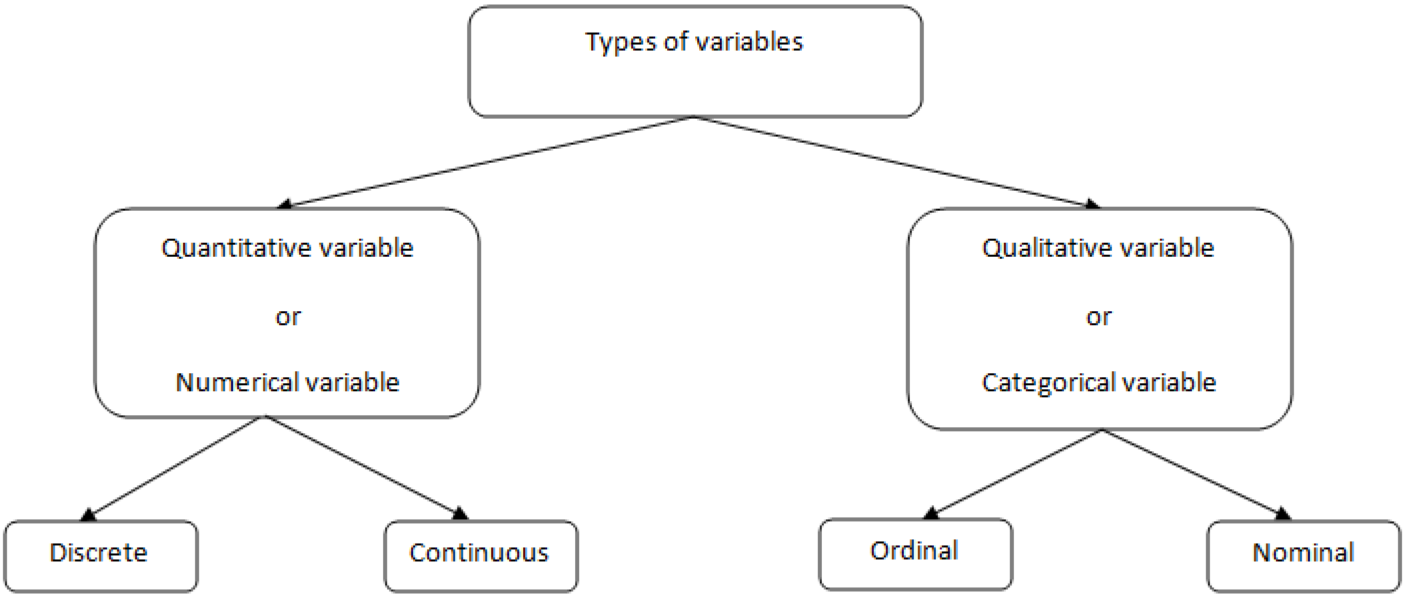
$$N_{classes} = 2.5 \quad \sqrt[4]{n} = 6.64 \simeq 7$$

and the amplitude class (or width) is : $a_i = \dfrac{max - min}{6.6} = \dfrac{174 - 140}{6.6} = 5.15 \simeq 5.$
So we find the following frequency table :

$[e_{i-1}, e_i[$	$[140, 145[$	$[145, 150[$	$[150, 155[$	$[155, 160[$	$[160, 165[$	$[165, 170[$	$[170, 175$
n_i	1	1	9	17	16	3	3
ICF $N_{x=e_i} \uparrow$	1	2	11	28	44	47	50
DCF $N_{x=e_i} \downarrow$	49	48	37	22	6	3	0
$f_i = \dfrac{n_i}{n}$	0.02	0.02	0.18	0.34	0.32	0.06	0.06
ICRF $F_{x=e_i} \uparrow$	0.02	0.04	0.22	0.56	0.88	0.94	1
DCRF $F_{x=e_i} \downarrow$	0.98	0.96	0.78	0.44	0.12	0.06	0


Population, Population size : N
Sample,
sample size : n
Statistical variable : X, Y, Z, ...
Data set
Statistical table
n_i, f_i, p_i,
$N_x \uparrow$, $N_x \downarrow$,
$F_x \uparrow$, $F_x \downarrow$

Conclusion 1 : Key terms of statistics.



Conclusion 2 : Types of variable

Chapter 2

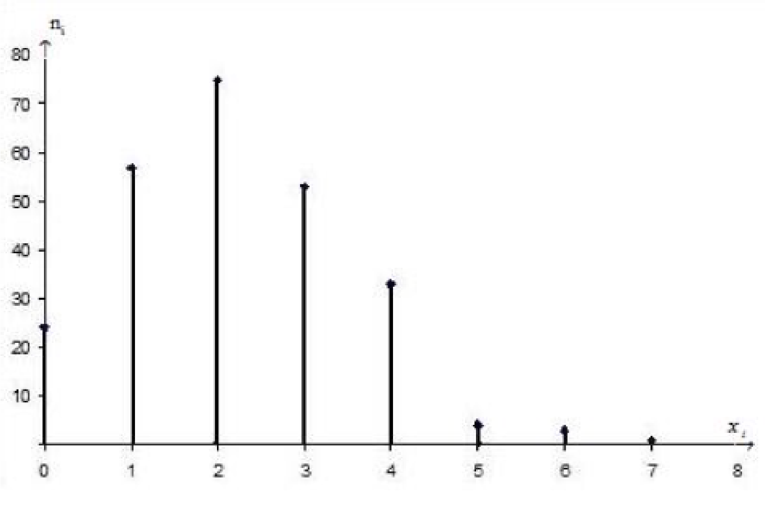
Graphical representation of data

2.1 Graphing ungrouped data

2.1.1 Bar chart

Example 1 : Consider the following fequency table :

Values x_i	0	1	2	3	4	5	6	7	$\sum$
Frequencies n_i	25	55	75	50	35	5	4	1	250
Relative frequencies f_i	0.1	0.22	0.3	0.2	0.14	0.02	0.016	0.004	1



The frequency diagram (the Bar chart)

2.1.2 Increasing cumulative frequency (or relative frequency) curve

- Step 1: calculate $N_x \uparrow$ or $F_x \uparrow$.

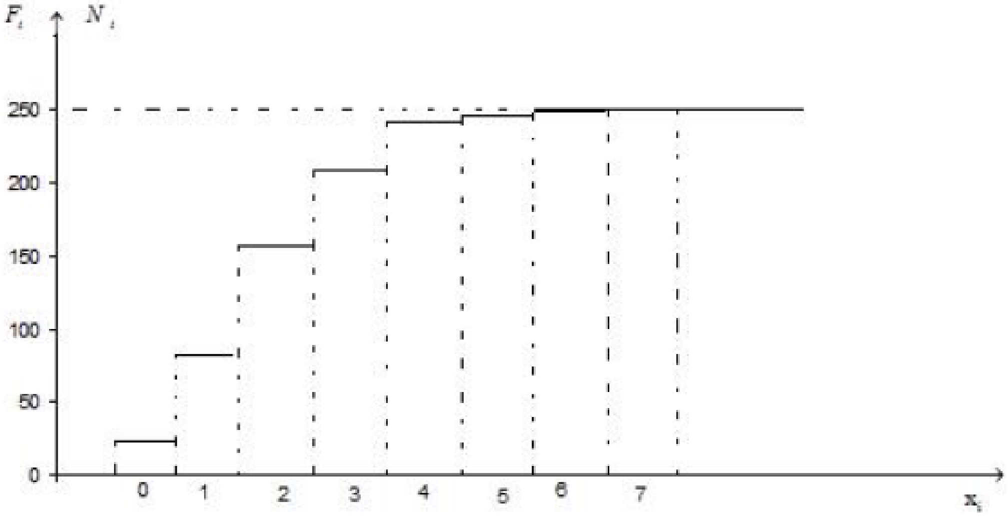
- Step 2: Place the x_i on the x-axis and $N_x \uparrow$ or $F_x \uparrow$ on the y-axis.

- Step 3: Determine the points $(x_i, N_{x_i} \uparrow)$ or $(x_i, F_{x_i} \uparrow)$ on the plan.

- Step 4: Draw the curve as follows :

Values x_i	0	1	2	3	4	5	6	7	Total
$N_{x=x_i} \uparrow$	25	80	155	205	240	245	249	250	/////

$$N_x \uparrow = \sum_{i\,:\,x_i \le x} n_i = \begin{cases} 0 & if\ \ x < 0 \\ 25 & if\ \ 0 \le x < 1 \\ 80 & if\ \ 1 \le x < 2 \\ 155 & if\ \ 2 \le x < 3 \\ 205 & if\ \ 3 \le x < 4 \\ 240 & if\ \ 4 \le x < 5 \\ 245 & if\ \ 5 \le x < 6 \\ 249 & if\ \ 6 \le x < 7 \\ 250 & if\ \ x \ge 7 \end{cases}$$



The frequency (or relative frequency) curve

2.2 Graphing grouped data

2.2.1 The frequency (or the relative frequency) histogram

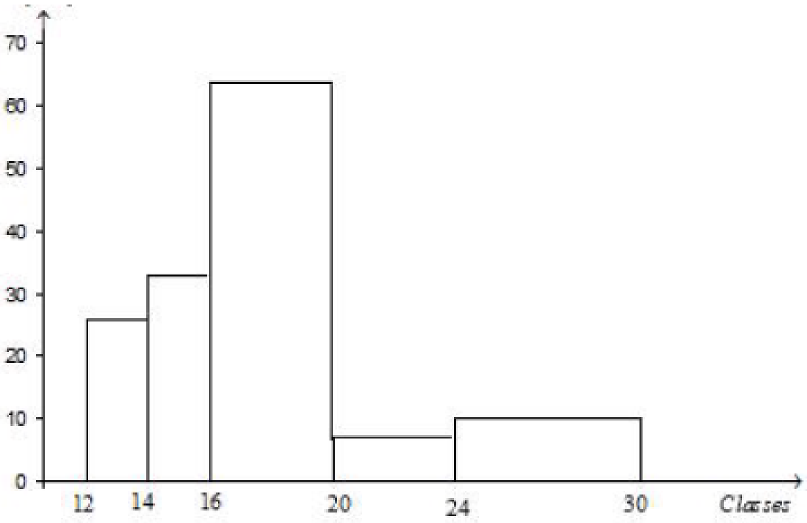
Case 01

- Step 1: Calculate the width of classes a_i. If the width a_i **are equal**, so
- Step 2: Draw the histogram such as the x-axis for the classes and the y-axis for the frequencies n_i or relative frequencies f_i.

Case 02

- Step 1: Calculate the width of classes a_i. If the width a_i **are not equal**, so,
- Step 2: Determine the unit u_i such as the minimum of width $a_i \rightarrow u_i = 1$, and determine the density $d_i = \dfrac{n_i}{u_i}$ (or $d_i = \dfrac{f_i}{u_i}$).
- Step 3: Draw the histogram such as the x-axis for the classes and the y-axis for the densities d_i.

Example : Find the frequency table of this frequency histogram



The frequency histogram

2.2.2 Increasing cumulative frequency (or relative frequency) curve

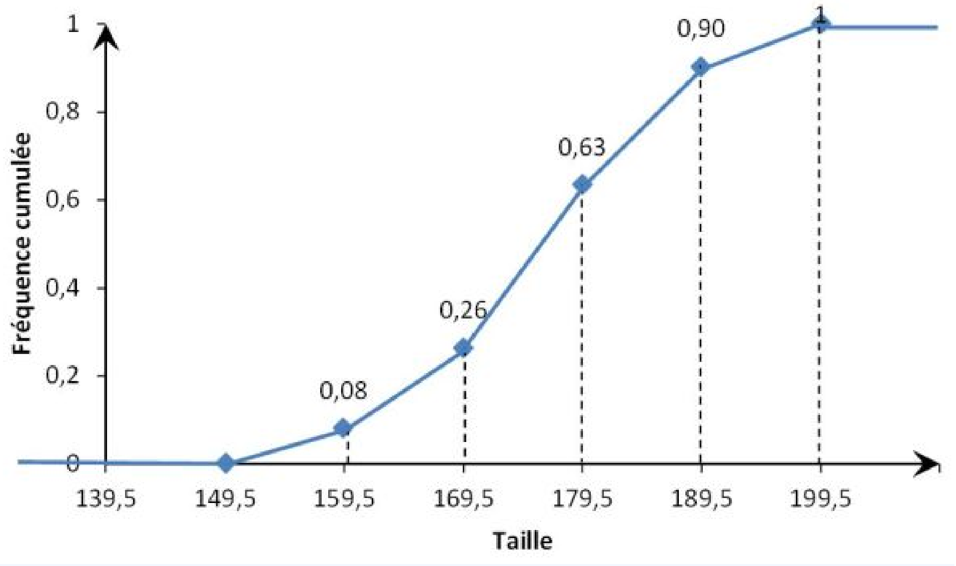
- Step 1: Calculate $N_x \uparrow$ or $F_x \uparrow$.

- Step 2: Place the calasses on the x-axis and the $N_x \uparrow$ or $F_x \uparrow$ on the y-axis.

- Step 3: Determine the points $(e_i, N_{e_i} \uparrow)$ or $(e_i, F_{e_i} \uparrow)$ on the plan.

- Step 4: Draw the curve.

  Exercise: From the following curve, find the relative frequency table.



The relative frequency curve.

Chapter 3

Numerical representations of data

3.1 Measures of position

3.1.1 Mean

The mean is given by the following formulas :

Quantitative discete data	Quantitative continuous data
$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\ x_i$ or $\overline{x} = \sum_{i=1}^{k} f_i\ x_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\ c_i$ or $\overline{x} = \sum_{i=1}^{k} f_i\ c_i$

3.1.2 Mode

The mode, noted Mo, is the most frequent number (value).

Quantitative discete data	Quantitative continuous data
	Case 1: the a_i are equals
	From the line of n_i (or f_i),
	note that the most frequent is n_i, so :
$Mo = x_i$ such as	The mode class : $[e_{i-1}, e_i[$,
$n_i = n_{max}$	$m_1 = n_i - n_{i-1}$
(from the line of n_i)	$m_2 = n_i - n_{i+1}$
or	so, the mode is given by :
$f_i = f_{max}$	$Mo = e_{i-1} + (e_i - e_{i-1})\,\dfrac{m_1}{m_1 + m_2}$
(from the line of f_i)	**Case 2: the a_i are not equals**
	We change n_i by d_i.

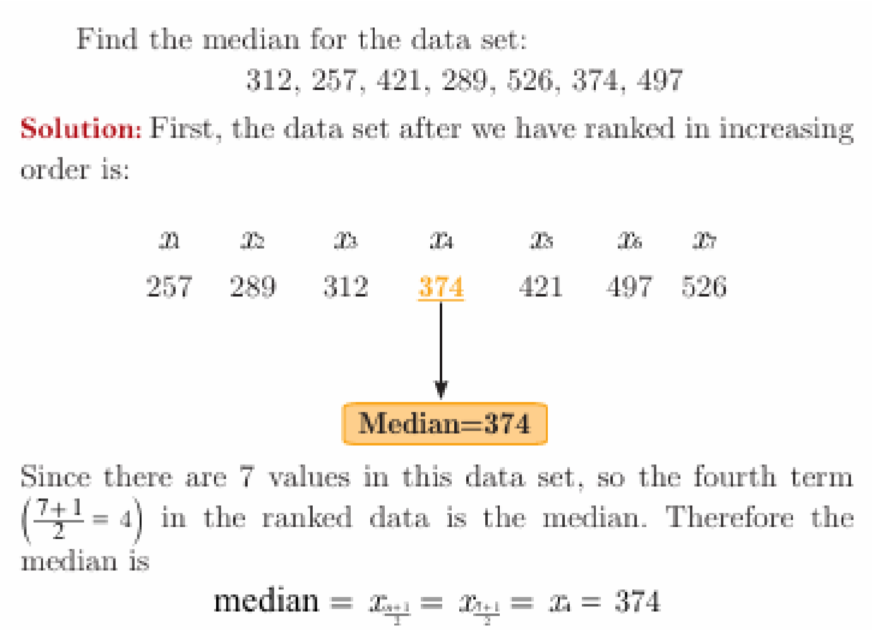
3.1.3 Median

The median is the midille value in a data set, noted by Me. So the median is the solution of equation :

$$N_{Me} \uparrow = \frac{N}{2} \;\; ou \;\; F_{Me} \uparrow = 0.5$$

Quantitative discrete data	Quantitative continuous data
•If N is an even number:	• $Me = e_{i-1} + a_i \dfrac{\frac{N}{2} - N_{e_{i-1}} \uparrow}{n_i}$
$Me = \frac{(\frac{N}{2})^{th} value + (\frac{N}{2}+1)^{th} value}{2}$	such as:
•If N is an odd number:	$N_{e_{i-1}} \leq \frac{N}{2} \leq N_{e_i}$
	and $e_{i-1} \leq Me \leq e_i$
$Me = (\frac{N+1}{2})^{th} value$	and $a_i = e_i - e_{i-1}$
	or:
	• $Me = e_{i-1} + a_i \dfrac{0.5 - F_{e_{i-1}} \uparrow}{f_i}$

Example 1:



Find the median for the data set:

312, 257, 421, 289, 526, 374, 497

Solution: First, the data set after we have ranked in increasing order is:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
257	289	312	374	421	497	526

Median=374

Since there are 7 values in this data set, so the fourth term $\left(\frac{7+1}{2}=4\right)$ in the ranked data is the median. Therefore the median is

$$\text{median} = x_{\frac{n+1}{2}} = x_{\frac{7+1}{2}} = x_4 = 374$$

The relative frequency curve.

Example 2: Find the median of the following data set:

$$8, 12, 7, 17, 14, 45, 10, 13, 17, 13, 9, 11$$

3.1.4 Quartiles

-The first quartile q_1 is the solution of equation :

$$N_{q_1}\uparrow = \frac{N}{4} \quad or \quad F_{q_1}\uparrow = 0.25$$

So:

Quantitative discrete data	Quantitative continuous data
$q_1 = \left(\frac{N}{4}\right)^{th}\ value$	$q_1 = e_{i-1} + (e_i - e_{i-1})\frac{\frac{N}{4} - N_{e_{i-1}}\uparrow}{n_i}$ such as : $N_{e_{i-1}} \leq \frac{N}{4} \leq N_{e_i}$ and $e_{i-1} \leq q_1 \leq e_i$
or	or
$F_{i-1}\uparrow < 0.25 < F_i\uparrow$ so: $q_1 = x_i$	$q_1 = e_{i-1} + a_i\,\frac{0.25 - F_{e_{i-1}}\uparrow}{f_i}$

-The third quartile q_3 is the solution of equation :

$$N_{q_3}\uparrow = \frac{3N}{4} \quad or \quad F_{q_3}\uparrow = 0.75$$

So:

Quantitative discrete data	Quantitative continuous data
$q_3 = \left(\frac{3N}{4}\right)^{th}\ value$	$q_3 = e_{i-1} + (e_i - e_{i-1})\frac{\frac{3N}{4} - N_{e_{i-1}}\uparrow}{n_i}$ such as : $N_{e_{i-1}} \leq \frac{3N}{4} \leq N_{e_i}$ and $e_{i-1} \leq q_3 \leq e_i$
or	or
$F_{i-1}\uparrow < 0.75 < F_i \uparrow$ so: $q_3 = x_i$	$q_3 = e_{i-1} + a_i\ \frac{0.75 - F_{e_{i-1}}\uparrow}{f_i}$

Particular cases:

- If $N_{x_i} \uparrow = 0.25$ so $q_1 = x_i$.
- If $N_{x_i} \uparrow = 0.5$ so $Me = x_i$.
- If $N_{x_i} \uparrow = 0.75$ so $q_3 = x_i$.

3.2 Measures of dispersion

3.2.1 Rang

The rang is given by : $E = x_{max} - x_{min}$

3.2.2 Variance

The variance, noted by $var(X)$, is given by the following formula :

Quantitative discrete data	Quantitative continuous data
$var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\ (x_i - \overline{x})^2$	$var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\ (c_i - \overline{x})^2$
$= \left(\frac{1}{N}\sum_{i=1}^{k} n_i\ x_i^2\right) - \overline{x}^2$	$= \left(\frac{1}{N}\sum_{i=1}^{k} n_i\ c_i^2\right) - \overline{x}^2$
or:	or:
$var(X) = \left(\sum_{i=1}^{n} f_i\ x_i^2\right) - \overline{x}^2$	$var(X) = \left(\sum_{i=1}^{n} f_i\ c_i^2\right) - \overline{x}^2$

3.2.3 Standard deviation

The standard deviation, noted σ_X, such as: $\sigma_X = \sqrt{var(X)}$.

The standard deviation characterizes the dispersion of a data set. The smaller the σ_X, the more the data are clustered around the mean $\overline{x}$, and the more homogeneous the population.

3.2.4 Coefficient of variation

The coefficient of variation is:

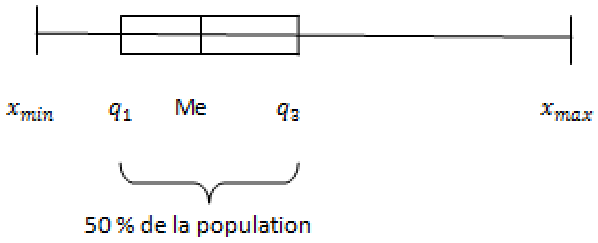
$$CV = \frac{\sigma_X}{\overline{x}}$$

3.3 Box Plots

Box plots give a good graphical image of the concentration of the data.

Step 1 Find x_{min}, x_{max}, Me, q_1, and q_3.

Step 2 Draw the box Plots as following :


x_{min}
q_1
Me
q_3
x_{max}
50 % de la population