# Introduction to Statistics "PART1"

Rahmani Nacer

Laboratory

14 - 03 - 2024

# Objectives

By the end of this lecture the student should be able to:

1. Recognize the applications of statistics in real life

# Objectives

By the end of this lecture the student should be able to:

1. Recognize the applications of statistics in real life
2. Define the terms “Population” and “Sample”

# Objectives

By the end of this lecture the student should be able to:

1. Recognize the applications of statistics in real life
2. Define the terms “Population” and “Sample”
3. Define and calculate different statistical variables (mean, standard deviation, median, etc.)

# 1.Some Terminology

1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."

# 1.Some Terminology

1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population( is a group selected from a population).

# 1.Some Terminology

1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population( is a group selected from a population).
3. **Data:**are the values (measurements or observations)that the variables can assume.

# 1.Some Terminology

1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population( is a group selected from a population).
3. **Data:**are the values (measurements or observations)that the variables can assume.
4. **Data set** : Collection of data values

# 1.Some Terminology

1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population( is a group selected from a population).
3. **Data:**are the values (measurements or observations)that the variables can assume.
4. **Data set** : Collection of data values
5. **Datum Or a data value (**individu**)** Each value in the data set (FR:c'est un élément de la population).

# 1.Some Terminology

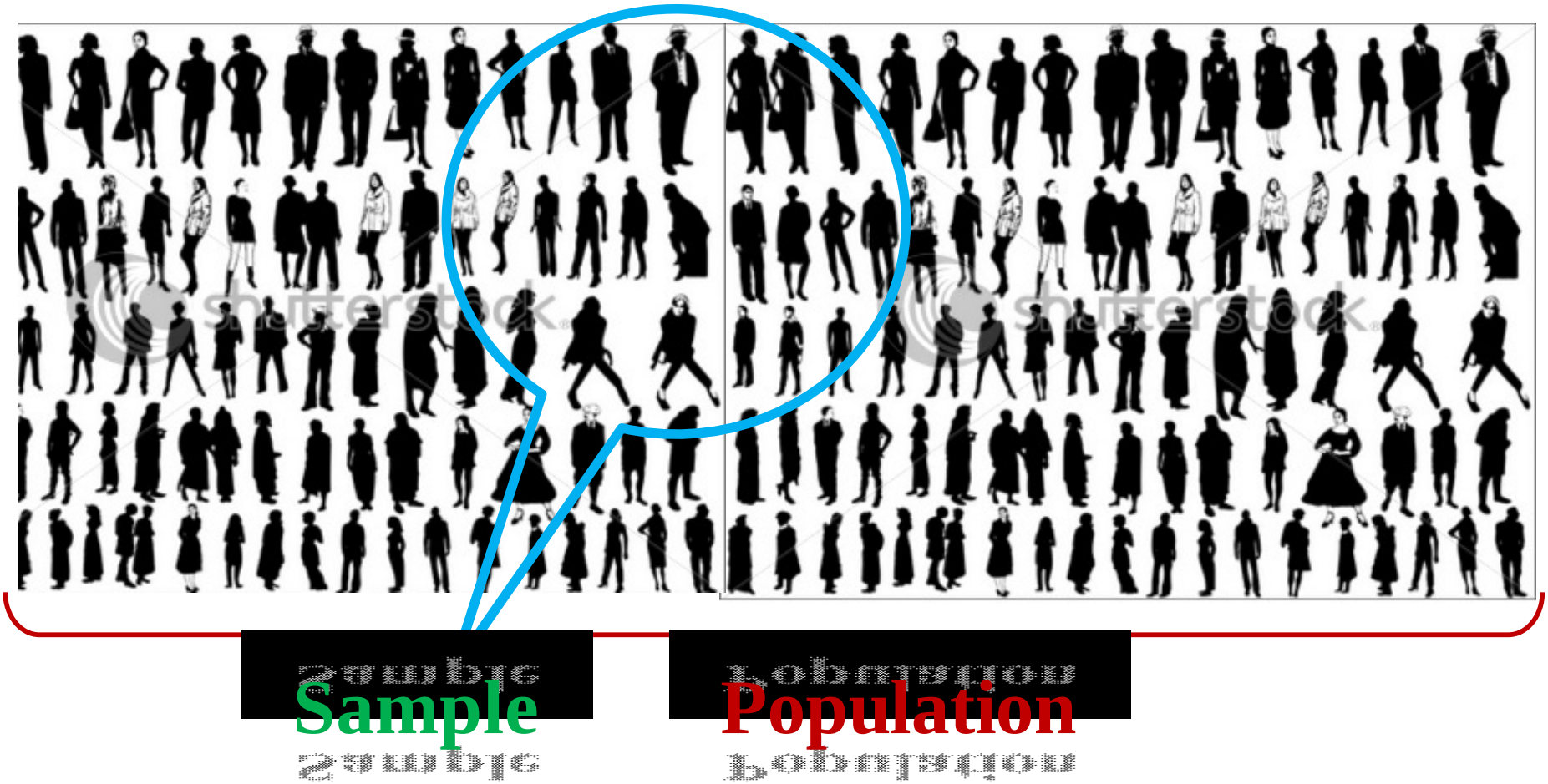
1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population( is a group selected from a population).
3. **Data:**are the values (measurements or observations)that the variables can assume.
4. **Data set** : Collection of data values
5. **Datum Or a data value (**individu**)** Each value in the data set (FR:c'est un élément de la population).
6. **Variable(**Caractère**)**: is characteristic or attribute that can assume different values.( FR: c'est la propriété étudiée).

# 1.Some Terminology

1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population( is a group selected from a population).
3. **Data:**are the values (measurements or observations)that the variables can assume.
4. **Data set** : Collection of data values
5. **Datum Or a data value (**individu**)** Each value in the data set (FR:c'est un élément de la population).
6. **Variable(**Caractère**)**: is characteristic or attribute that can assume different values.( FR: c'est la propriété étudiée).
7. **Parameter**: A numerical description measiring the variable in the sample.

# 2.1.Some Terminology

## POPULATION and Sample



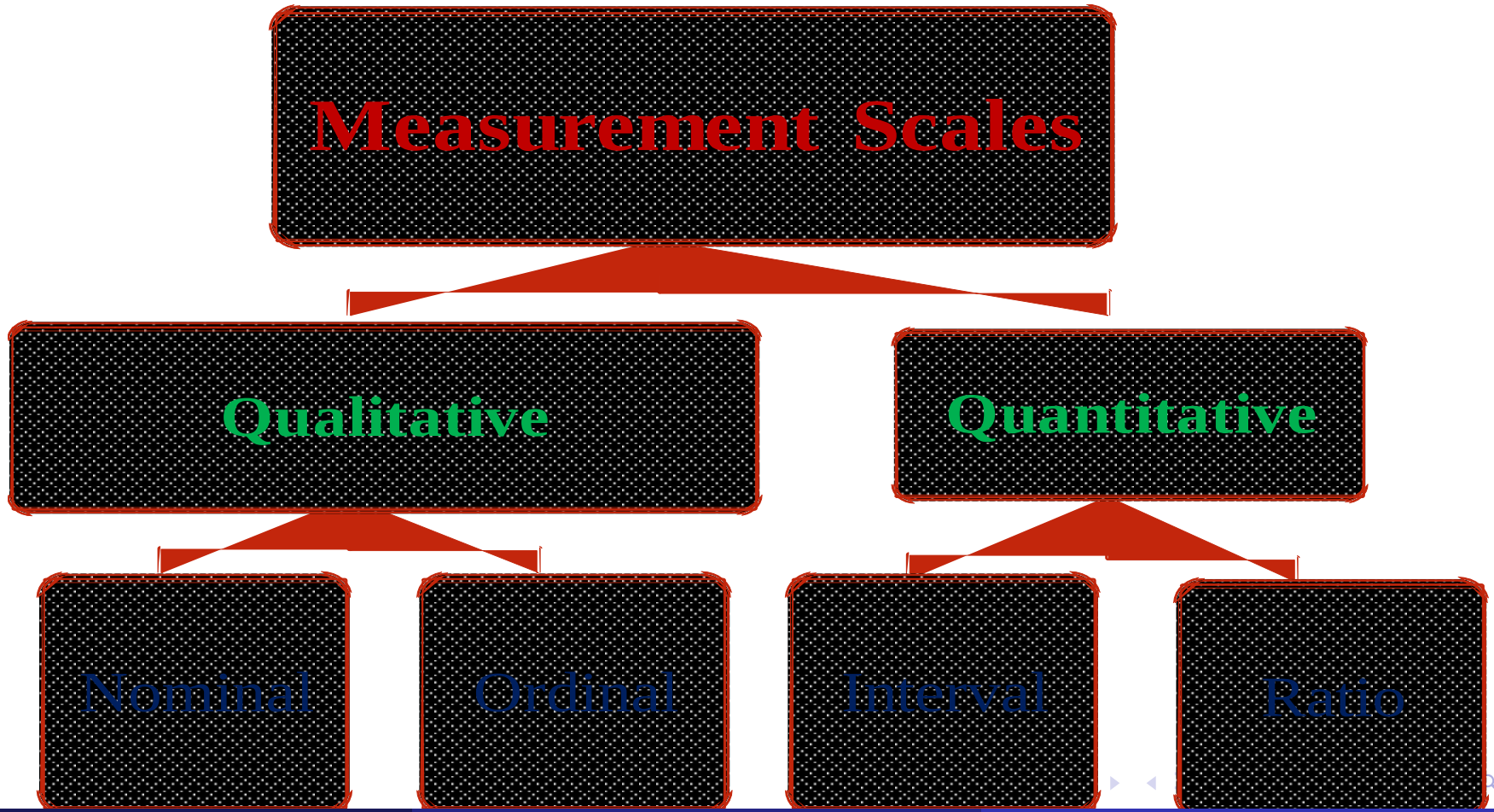

Statistics is the science of conducting studies to collect, organize,

# 3.Variables (caractères)

- A variable is a characteristic or condition that can change or take on different values.

# 3.Variables (caractères)

- A variable is a characteristic or condition that can change or take on different values.
- Most research begins with a general question about the relationship between two variables for a specific group of individuals.



# 3.1 Types of Variables

Variables can be classified as Qualitative Variables or Quantitative variables.

1. **Qualitative Variables:**are variables that have distinct categories , according tosome characteristic or attribute.
   *For example*: Gender ,Marital status ,Color. . . . . etc

2. **Quantitative variables:**are variables that can be counted or measured.
   *For example*: Age ,Height , Weight ,temperature . . . ..etc

# 3.1 Types of Variables

**Quantitative variables:** can be classified as discrete or continuous

1. **Discrete variables** (such as class size) consist of indivisible categories, and
2. **continuous variables** (such as time or weight) are infinitely divisible into whatever units a researcher may choose. For example, time can be measured to the nearest minute, second, half-second, etc.

**Qualitative Variables:** can be classified as Nominal or Ordinal level

1. **Nominal level:** classifies data into mutually exclusive , exhausting categories in which no order or ranking can be imposed on the data. For example: Eye color ,Gender ,Political party , blood types . . . etc
2. **Ordinal level:**classifies data into categories can be ranked .For example: Grade of course (A,B,C) ,Size( S,M,L) Rating scale (Poor ,Good ,Excellent )

# 4 Data representation:

- Suppose that you have collected some discrete data. It will be difficult to get a "feel" for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.

A frequency table is a list of possible values and their frequencies.

le nombre $f_i cum$ tel que $f_i cum = \sum^{i}_{} f_n$

# 4 Data representation:

- Suppose that you have collected some discrete data. It will be difficult to get a "feel" for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.
- **The frequency (**Effectifs**)** of a value is the number $n_i$ of observations taking that value. and the **cumulative frequency** (Effectifs cumulées**)** is:

$$n_i cum = \sum_{p=1}^{i} n_p$$

A frequency table is a list of possible values and their frequencies.

le nombre $f_i cum$ tel que $f_i cum = \sum_{p=1}^{i} f_p$

# 4 Data representation:

- Suppose that you have collected some discrete data. It will be difficult to get a "feel" for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.
- **The frequency (**Effectifs**)** of a value is the number $n_i$ of observations taking that value. and the **cumulative frequency** (Effectifs cumulées**)** is:

$$n_i cum = \sum_{p=1}^{i} n_p$$

A frequency table is a list of possible values and their frequencies.

- **Cumilative Proportion :**

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

le nombre $f_i cum$ tel que $f_i cum = \sum_{p=1}^{i} f_p$

# 4 Data representation:

- Suppose that you have collected some discrete data. It will be difficult to get a "feel" for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.
- **The frequency (**Effectifs**)** of a value is the number $n_i$ of observations taking that value. and the **cumulative frequency** (Effectifs cumulées**)** is:

$$n_i cum = \sum_{p=1}^{i} n_p$$

A frequency table is a list of possible values and their frequencies.

- **Cumilative Proportion :**

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

- On appelle fréquences cumulées ou fréquences relatives cumulées en $x_i$,

le nombre $f_i cum$ tel que $f_i cum = \sum_{p=1}^{i} f_p$

# 4.1 Frequency Distribution table

## Qualitative Variables

´

- A frequency table is a list of possible values and their frequencies.

| **Modalty** | **frequency** |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| $\vdots$ | $\vdots$ |
| $x_k$ | $n_k$ |

# 4.1 Proportion Distribution table

## Qualitative Variables

´

- We then calculate the proportions of each modality by dividing the number of each modality by the total number:

# 4.1 Proportion Distribution table

## Qualitative Variables

´

- We then calculate the proportions of each modality by dividing the number of each modality by the total number:
- $$f_k = \frac{n_k}{n}$$

# 4.1 Proportion Distribution table

## Qualitative Variables

´

- We then calculate the proportions of each modality by dividing the number of each modality by the total number:
- $$f_k = \frac{n_k}{n}$$
-

| Modality | Proportion |
|---|---|
| $x_1$ | $f_1$ |
| $x_2$ | $f_2$ |
| $\vdots$ | $\vdots$ |
| $x_k$ | $f_k$ |

´

**Example1:**if we have a table of the form

| Numéro de Cliente | Signalétique |
|---|---|
| 1 | M. |
| 2 | Mme |
| 3 | Mlle |
| ⋮ | ⋮ |
| 627630 | Mme |

Tab. 3 – Variable Signalétique

# 4.1 Distribution table

## Qualitative Variables

´

Using flat sorting, we will construct a table of the form:

| Signalétique | Nombre de Clientes | Proportions |
|---|---|---|
| M. | 60985 | 0,0972 |
| Mme | 424641 | 0,6766 |
| Mlle | 142004 | 0,2262 |
| Total | 627630 | 1 |

Tab. 4 – Distributions de la Variable Signalétique

# 4.1 Distribution table

:

´

**Quantitative variables**

The raw table looks like this:

| **Data value** | **Variable** |
|---|---|
| 1 | $x_1$ |
| 2 | $x_2$ |
| $\vdots$ | $\vdots$ |
| $n$ | $x_n$ |

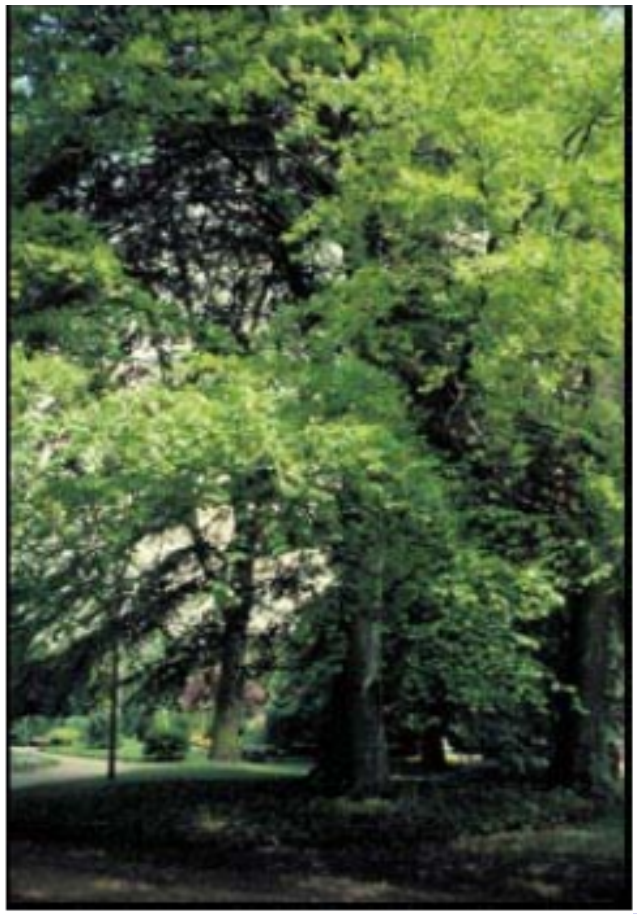
**objectif**: créer un tableau plus synthétique.

**Cas des variables discrètes** :
on étudie une variable discrète $X$ à $p$ modalités dans une population de taille $n$.

| **Modalités** | $x_1$ | $x_2$ | $\cdots$ | $x_p$ |
|---|---|---|---|---|
| **Effectifs** | $n_1$ | $n_2$ | $\cdots$ | $n_p$ |
| **Fréquences:** $f_i = \frac{n_k}{n}$ | $f_1$ | $f_2$ | $\cdots$ | $f_p$ |

´

**Exemple2:***La cécidomyie du hêtre* provoque sur les feuilles de cet arbre des galles dont la distribution de fréquences observées est la suivante:



# 4.1 Tableaux statistiques

## Quantitative variables

´

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 182 | 98 | 46 | 28 | 12 | 5 | 2 | 3 | 0 |
| $f_i = \frac{n_k}{n}$ | 0.485 | 0.261 | 0.123 | 0.075 | 0.032 | 0.013 | 0.005 | 0.006 | 0 |
| $f_i cum$ | 0.485 | 0.746 | 0.869 | 0.944 | 0.976 | 0.989 | 0.994 | 1 | 1 |

avec:

-$x_i$ :*le ombre de galles par feuille*

-$n_i$ : *nombre de feuilles portant* $x_i$ *galles*

´

- **Cas des variables continues** :

Soit $p$ le nombre d'intervalles. Les données se présentent sous la forme suivante:

| **Classes** | **Effectifs** | **Centres de classe** | **Fréquences:** $f_i = \frac{n_k}{n}$ |
|---|---|---|---|
| $[e_0, e_1[$ | $n_1$ | $c_1$ | $f_1$ |
| $[e_1, e_2[$ | $n_2$ | $c_2$ | $f_2$ |
| $[e_2, e_3[$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $[e_3, e_4[$ | $n_p$ | $c_p$ | $f_p$ |

´

- **Cas des variables continues** :
- on regroupe les individus par classes. On décompose l'intervalle des valeurs possibles en une partition d'intervalles.

Soit $p$ le nombre d'intervalles. Les données se présentent sous la forme suivante:

| **Classes** | **Effectifs** | **Centres de classe** | **Fréquences:** $f_i = \frac{n_k}{n}$ |
|---|---|---|---|
| $[e_0, e_1[$ | $n_1$ | $c_1$ | $f_1$ |
| $[e_1, e_2[$ | $n_2$ | $c_2$ | $f_2$ |
| $[e_2, e_3[$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $[e_3, e_4[$ | $n_p$ | $c_p$ | $f_p$ |

´

**Cas des variables continues** :

| $X$ | $n_i$ | $X_i$ | $N_i \nearrow$ | $F_i \nearrow$ | $N_i \searrow$ |
|---|---|---|---|---|---|
| $[a_0 \ , \ a_1]$ | $n_1$ | $\frac{a_0+a_1}{2}$ | $N_1 = 0$ | $F_1 = N_1/n$ | $n$ |
| $[a_1 \ , \ a_2]$ | $n_2$ | $\frac{a_1+a_2}{2}$ | $N_2 = 0 + n_1$ | $F_2 = N_2/n$ | $n - n_1$ |
| $[a_2 \ , \ a_3]$ | $n_3$ | $\frac{a_2+a_3}{2}$ | $N_3 = 0 + n_1 + n_2$ | $F_3 = N_3/n$ | $n - n_1 - n_2$ |
| $\vdots$ | | | | | |
| $[a_{i-1} \ , \ a_i]$ | $n_i$ | $\frac{a_{i-1}+a_i}{2}$ | $N_i = 0 + n_1 + ... + n_i$ | $F_i = N_i/n$ | $n - n_1 - ... - n_i$ |
| $\vdots$ | | | | | |
| $[a_{m-1} \ , \ a_m]$ | $n_m$ | $\frac{a_{m-1}+a_m}{2}$ | $N_m = 0 + n_1 + ... + n_{m-1}$ | $F_m = N_m/n$ | $n - n_1 - ... - n_{m-1}$ |
| $\Sigma$ | $n$ | $-$ | $n$ | $1$ | $0$ |

# 4.1 Tableaux statistiques

´

**Cas des variables continues** :
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe

- Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille $n$.

# 4.1 Tableaux statistiques

´

**Cas des variables continues** :
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe

- Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille $n$.
- La règle de **STURGE** : *Nombre de classes:* $k = 1 + 3,332\,(\log n)$

´

**Cas des variables continues** :
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe

- Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille $n$.
- La règle de **STURGE** : *Nombre de classes:* $k = 1 + 3,332\,(\log n)$
- La règle de **YULE** : *Nombre de classes:* $k = 2,5\left(\sqrt[4]{n}\right)$

# 4.1 Tableaux statistiques

´

**Cas des variables continues** :
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe

- Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille $n$.
- La règle de **STURGE** : *Nombre de classes:* $k = 1 + 3,332\,(\log n)$
- La règle de **YULE** : *Nombre de classes:* $k = 2,5\left(\sqrt[4]{n}\right)$
- L'intervalle entre chaque classe est obtenu ensuite de la manière suivante:

# 4.1 Tableaux statistiques

´

**Cas des variables continues** :
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe

- Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille $n$.
- La règle de **STURGE** : *Nombre de classes:* $k = 1 + 3,332\,(\log n)$
- La règle de **YULE** : *Nombre de classes:* $k = 2,5\left(\sqrt[4]{n}\right)$
- L'intervalle entre chaque classe est obtenu ensuite de la manière suivante:
- **Intervalle de classe:** $C = (X_{max} - X_{min})/k$

# 4.1 Tableaux statistiques

´

**Cas des variables continues** :
Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe

- Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille $n$.
- La règle de **STURGE** : *Nombre de classes:* $k = 1 + 3,332\,(\log n)$
- La règle de **YULE** : *Nombre de classes:* $k = 2,5\left(\sqrt[4]{n}\right)$
- L'intervalle entre chaque classe est obtenu ensuite de la manière suivante:
- **Intervalle de classe:** $C = (X_{max} - X_{min})/k$
- avec $X_{max}$ et $X_{min}$, respectivement la plus grande et la plus petite valeur de $X$ dans la série statistique.

´

**Exemple3:**

Dans le cadre de l'étude de la population de gélinottes huppées (Bonasa umbellus), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

$n = 50$ avec $X_{max} = 174$ et $X_{min} = 140$*Nombre de classes:*

$$k = 1 + 3.332(\log n) = 1 + 3,332\,(\log 50) = 7$$

´

**Exemple3:**

Dans le cadre de l'étude de la population de gélinottes huppées (Bonasa umbellus), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

$n = 50$ avec $X_{max} = 174$ et $X_{min} = 140$*Nombre de classes:*

$$k = 1 + 3.332(\log n) = 1 + 3,332\,(\log 50) = 7$$

**Intervalle de classe:**

$$c = (X_{max} - X_{min})/k = \frac{174 - 140}{7} = 5$$

´

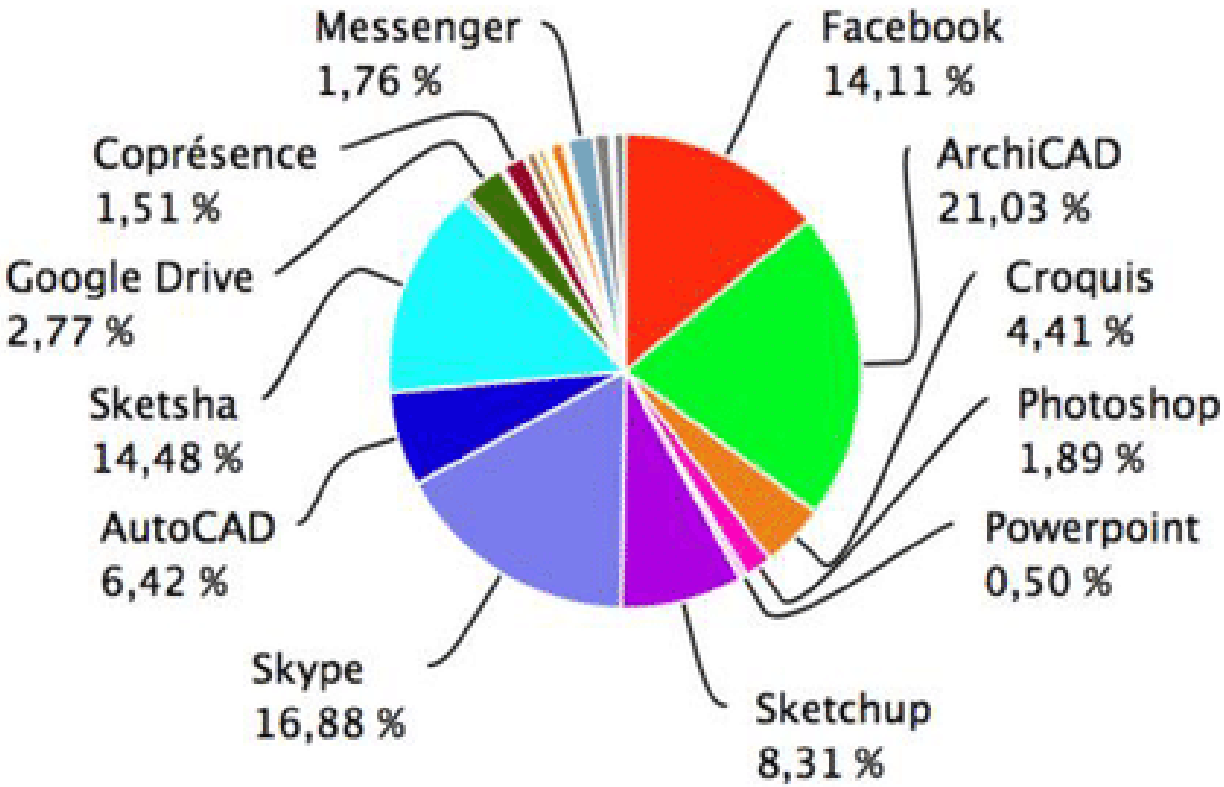
**Exemple3:**

| Caractère $X$ : $x_i$ : longueur de la rectrice bornes des classes | [140-145[ | [145-150[ | [150-155[ | [155-160[ | [160-165[ | [165-170[ | [170-175[ |
|---|---|---|---|---|---|---|---|
| Valeur médiane des classes, $x_i'$ | 142,5 | 147,5 | 152,5 | 157,5 | 162,5 | 167 ,5 | 172,5 |
| $n_i$ : nombre d'individu par classe de taille $x_i$ | 1 | 1 | 9 | 17 | 16 | 3 | 3 |
| $f_i$ : fréquence relative | 0,02 | 0,02 | 0,18 | 0,34 | 0,32 | 0,06 | 0,06 |
| $f_i cum.$ : fréquence relative cumulée | 0,02 | 0,04 | 0,22 | 0,56 | 0,88 | 0,94 | 1 |

Graphical representations have the advantage of immediately providing information on the general appearance of the distribution. They facilitate the interpretation of the data collected.
**Case of a qualitative variable:** can be represented using two types of diagrams:

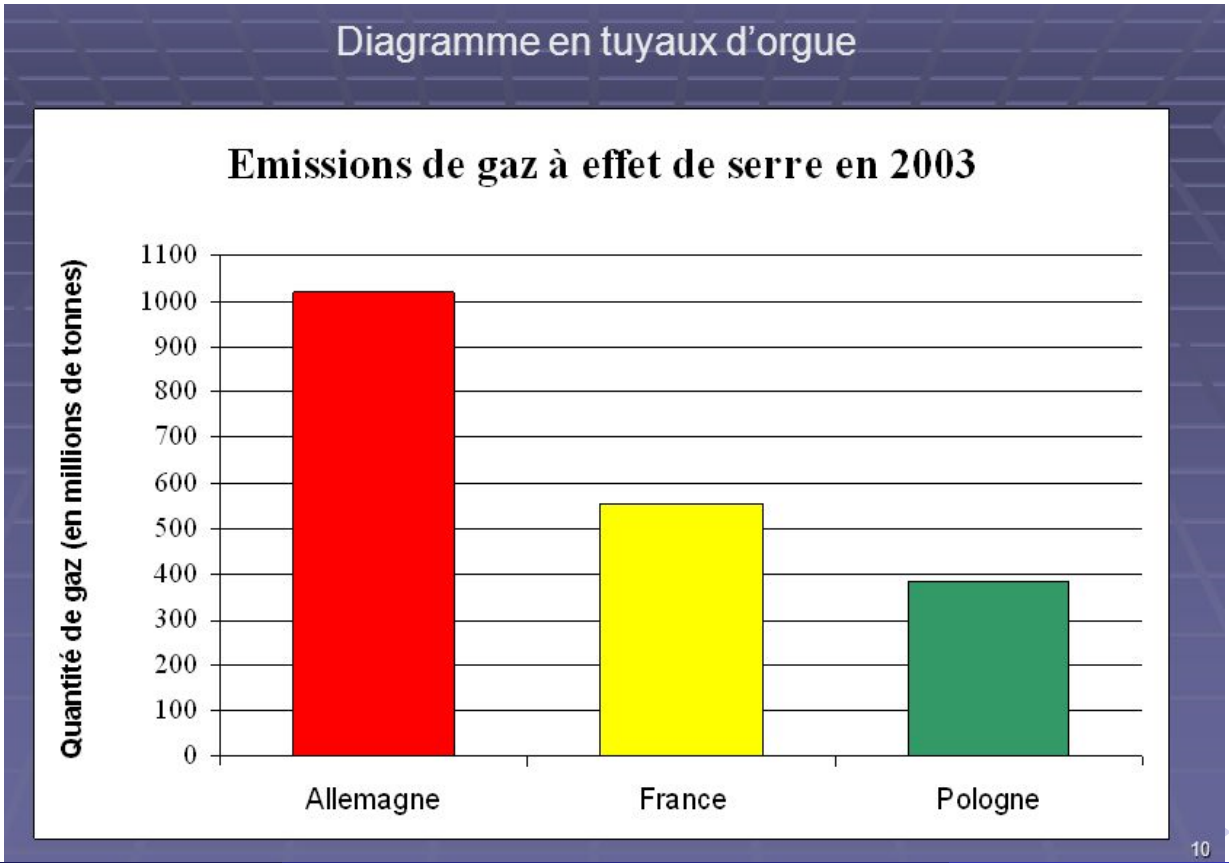
1. Un diagramme en camembert (ou diagramme à secteurs ) : **Pie chart**
2. Tuyaux d'orgue: **Bar graph**.

**1 Pie chart: Un diagramme en camembert (ou diagramme à secteurs ):**diagrams consist of dividing a disk or half-disk, into slices, or sectors, corresponding to the modalities observed and whose surface area is proportional to the effectiveness, or frequency, of the modality .



## 2- Bar graph:

- We plot the modalities on the abscissa, arbitrarily.
- We carry rectangles on the ordinates whose length is proportional to the numbers, or frequencies, of each modality



:

**Case of a Quantitative variable:** can be represented using five types of diagrams:

1.**Bar Graphs** if the **variable is discrete** "Diagramme en bâton"
2. **the stepped curve:** la courbe en escalier ( **variable discrète)**

3. **The histogram** of densities if the distribution is **continuous**

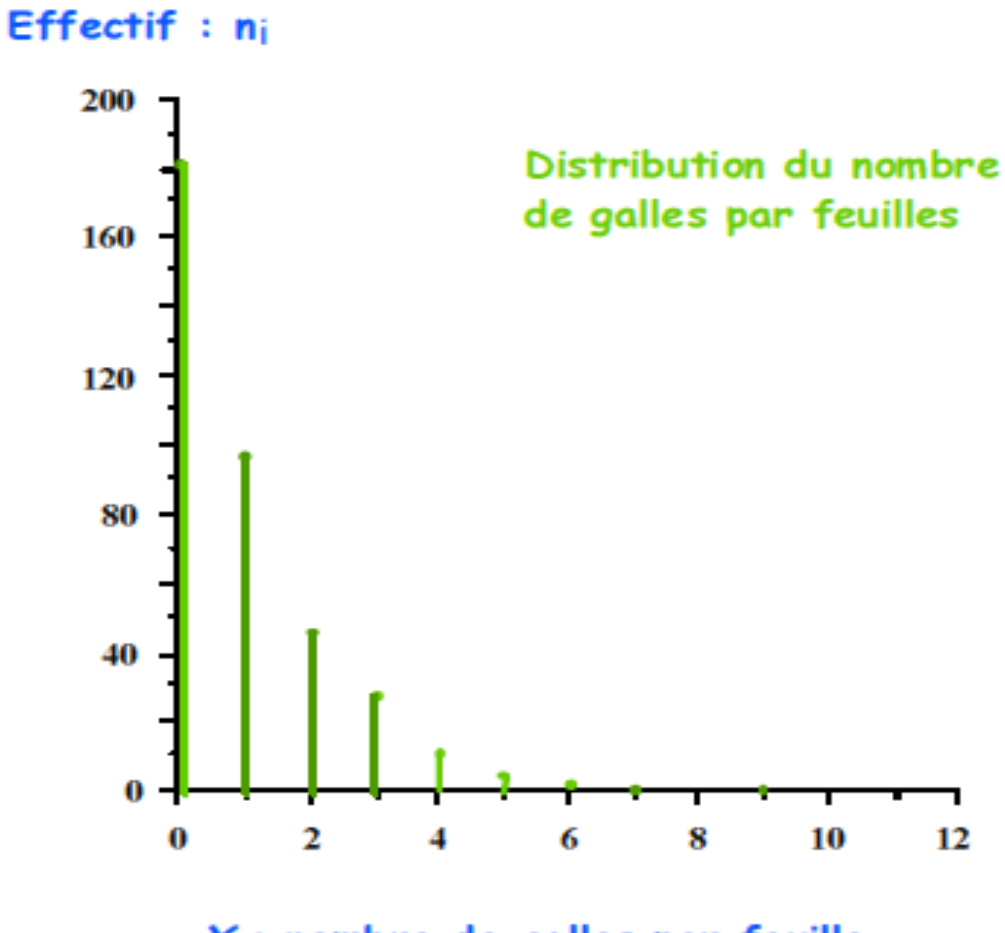
4.**Frequency polygons**:"Polygonne de fréquence"if the **variable is continuous**

5. **An ogive** "cumulative frequency "la courbe des fréquences cumulées (ou des effectifs cumulés).

# 5. Graphical representations

## Case of a quantitative discrete variable

- **1. Bar Graphs**, des effectifs ou des fréquences: La différence avec le cas qualitatif consiste en ce que les abscisses ici sont les valeurs de la variable statistique.**(voir exemple2)**



## 2. the stepped curve:la courbe en escalier:

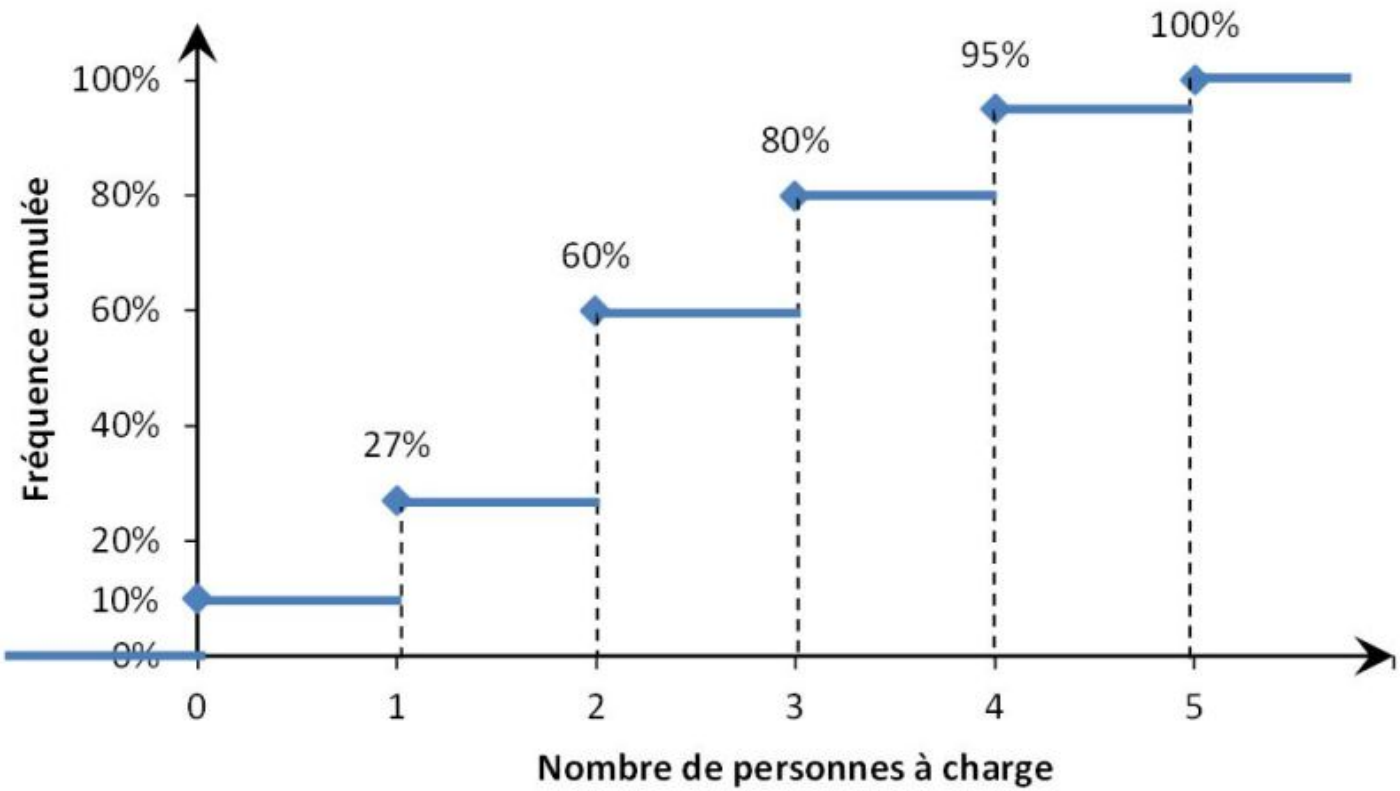
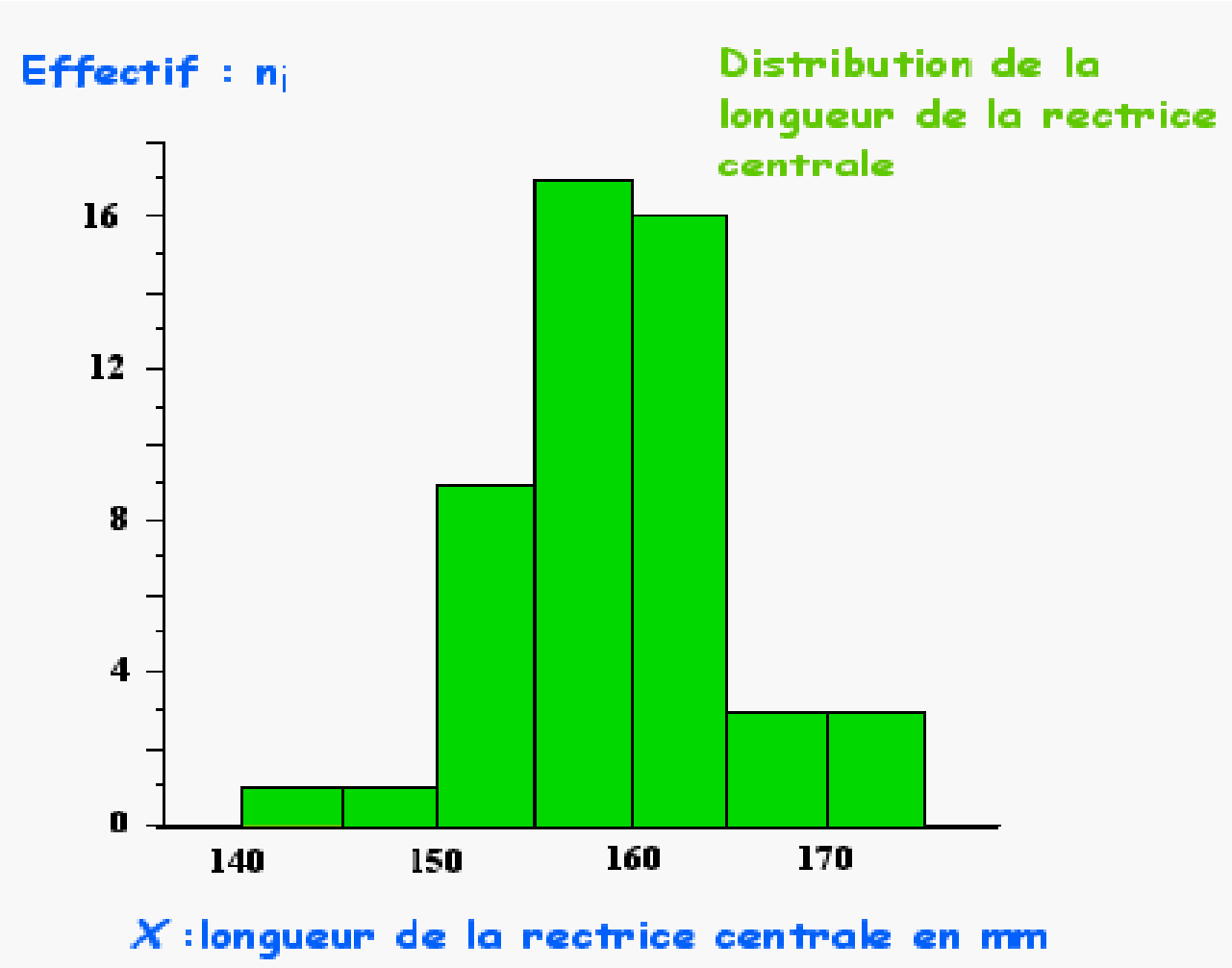
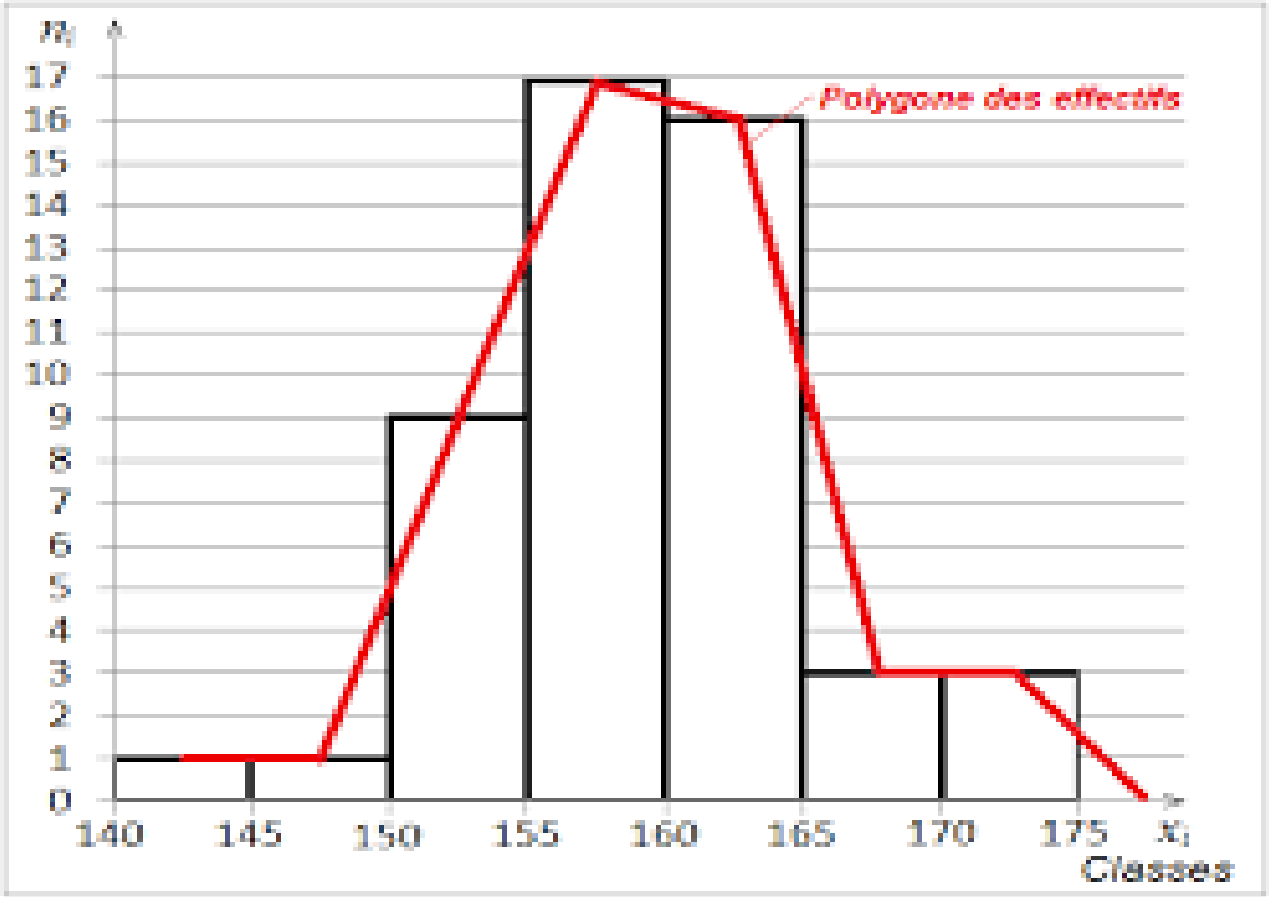



Figure: La courbe en escalier

**1.The histogram: l'histogramme** des densités si la distribution est **continue (exemple3)**



**2. Frequency polygons: Polygonne des effectifs:** d'une **variable continue:** On obtient le polygone des effectifs (ou des fréquences) en reliant les milieux des bases supérieures des rectangles.



**3. An ogive: la courbe des effectifs cumulées croissant et décroissant** sont présentés