Introduction to Statistics and Probability

Rahmani Nacer

Laboratory

04 - 03 - 2025

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

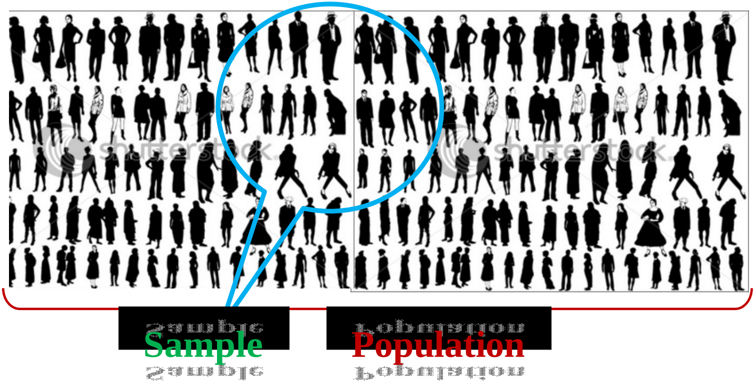
1. **Population:** All people or things you are stading."consists of all subjects (human or otherwise) that are studied."
2. **Sample** (échantillon): is a subset of the population(is a group selected from a population).
3. **Data:**are the values (measurements or observations)that the variables can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



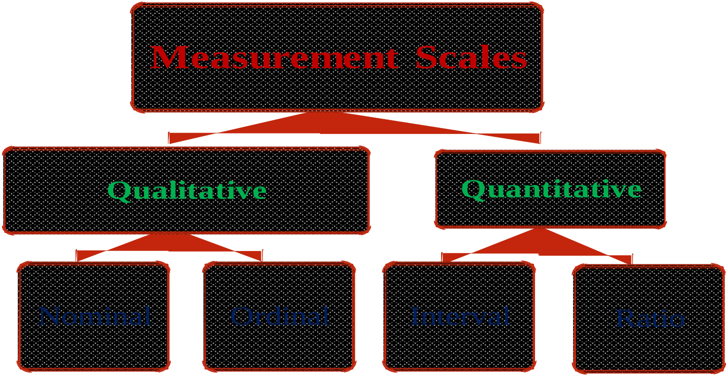
Statistics is the science of conducting studies to collect, organize, summarize, analyze, present, interpret and draw conclusions from data.

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

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.

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:

$$n_{icum} = \sum_{p=1}^{i} n_p$$

- 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:

$$n_{icum} = \sum_{p=1}^{i} n_p$$

- **relativ frequency :**

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

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:

$$n_{icum} = \sum_{p=1}^{i} n_p$$

- **relativ frequency :**

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

- **the cumulative relativ frequency** , the number $f_i cum$ where

$$f_{icum} = \sum_{p=1}^{i} f_p$$

´

- 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

´

- We then calculate **the proportions** or **relativ frequency** of each modality by dividing the number of each modality by the total number:

´

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

´

- We then calculate **the proportions** or **relativ frequency** 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

´

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

Qualitative Variables "Signalétique"

´

Quantitative variables

The raw table looks like this:

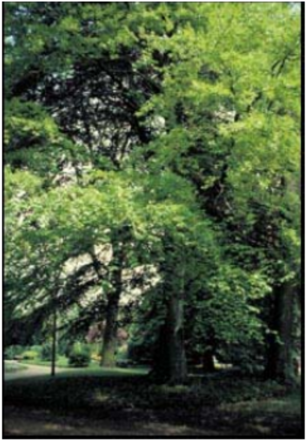
Data value	**Variable**
1	x_1
2	x_2
$\vdots$	$\vdots$
n	x_n

objectif: créer un tableau plus synthétique.

Cas des variables discrètes :
We study a discrete variable X with p modalities in a population of size n..

Modalities	**frequency**	**Centre of class**	**relativ Frequency:** $f_i = \frac{n_k}{n}$
x_1	n_1	c_1	f_1
x_2	n_2	c_2	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_p	c_p	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:



´

$\mathbf{x}_i$	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\mathbf{n}_i$	182	98	46	28	12	5	2	3	0
$\mathbf{f}_i = \frac{n_k}{n}$	0.485	0.261	0.123	0.075	0.032	0.013	0.005	0.006	0
$\mathbf{f}_i$**cum**	0.485	0.746	0.869	0.944	0.976	0.989	0.994	1	1

avec:

-x_i : the number of galls per leaf

-n_i : the number of leaves bearing x_i galls

´

- **Case of continuous variables:**

´

- **Case of continuous variables:**
- Individuals are grouped into classes. The range of possible values is divided into a partition of intervals.

´

- **Case of continuous variables:**
- Individuals are grouped into classes. The range of possible values is divided into a partition of intervals.
- Let p be the number of intervals. The data are presented in the following form:

Classes	**frequency**	**Class Centers**	**Relative Frequency:** $f_i = \frac{n_k}{n}$
$[e_0, e_1[$	n_1	c_1	f_1
$[e_1, e_2[$	n_2	c_2	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[e_3, e_4[$	n_p	c_p	f_p

´

Case of continuous variables:

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	$\frac{a_{m-1}+a_m}{2}$	$N_m = 0 + n_1 + ... + n_{m-1}$	$F_m = N_m/n$	$n - n_1 - ... - n_{m-1}$
Σ	n	$-$	n	1	0

Case of continuous variables:

in the case of a continuous quantitative variable, constructing the frequency table first requires grouping the data into classes. This involves determining the expected number of classes and the corresponding width of each class or class interval.

- Various empirical formulas can be used to determine the number of classes for a sample of size n.

Case of continuous variables:
in the case of a continuous quantitative variable, constructing the frequency table first requires grouping the data into classes. This involves determining the expected number of classes and the corresponding width of each class or class interval.

- Various empirical formulas can be used to determine the number of classes for a sample of size n.
- **Sturges' Rule**: The number of classes is given by::
  $k = 1 + 3,332\,(\log n)$

Case of continuous variables:
in the case of a continuous quantitative variable, constructing the frequency table first requires grouping the data into classes. This involves determining the expected number of classes and the corresponding width of each class or class interval.

- Various empirical formulas can be used to determine the number of classes for a sample of size n.
- **Sturges' Rule**: The number of classes is given by::
  $k = 1 + 3,332\,(\log n)$
- **The interval between each class** is then obtained as follows:
  $C = (X_{max} - X_{min})\,/\,k$

4.1 Statistical Tables

Case of continuous variables:
in the case of a continuous quantitative variable, constructing the frequency table first requires grouping the data into classes. This involves determining the expected number of classes and the corresponding width of each class or class interval.

- Various empirical formulas can be used to determine the number of classes for a sample of size n.
- **Sturges' Rule**: The number of classes is given by::
  $k = 1 + 3,332\,(\log n)$
- **The interval between each class** is then obtained as follows:
  $C = (X_{max} - X_{min})\,/\,k$
- avec X_{max} et X_{min}, Respectively, the largest and smallest values of X in the statistical series.

Case of continuous variables:

- **Yule's Rule**: The number of classes is given by:: $k = 2,5\left(\sqrt[4]{n}\right)$

Case of continuous variables:

- **Yule's Rule**: The number of classes is given by:: $k = 2,5\left(\sqrt[4]{n}\right)$
- **The interval between each class is** then obtained as follows: $C = (X_{max} - X_{min})/k$

Case of continuous variables:

- **Yule's Rule**: The number of classes is given by:: $k = 2,5\left(\sqrt[4]{n}\right)$
- **The interval between each class is** then obtained as follows: $C = \left(X_{max} - X_{min}\right)/k$
- avec X_{max} et X_{min}, Respectively, the largest and smallest values of X in the statistical series.

Exemple3:

- As part of the study of the ruffed grouse population (Bonasa umbellus), the values of the length of the main rectrix can be distributed as follows:

  $n = 50$ avec $X_{max} = 174$ et $X_{min} = 140$

Exemple3:

- As part of the study of the ruffed grouse population (Bonasa umbellus), the values of the length of the main rectrix can be distributed as follows:

  $n = 50$ avec $X_{max} = 174$ et $X_{min} = 140$The number of classes*:*

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

Exemple3:

- As part of the study of the ruffed grouse population (Bonasa umbellus), the values of the length of the main rectrix can be distributed as follows:

  $n = 50$ avec $X_{max} = 174$ et $X_{min} = 140$The number of classes:

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

  The interval between each class is**:**

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

´

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

Qualitative (or categorical) variables represent categories or groups rather than numerical values. The most common graphical representations for qualitative data are:

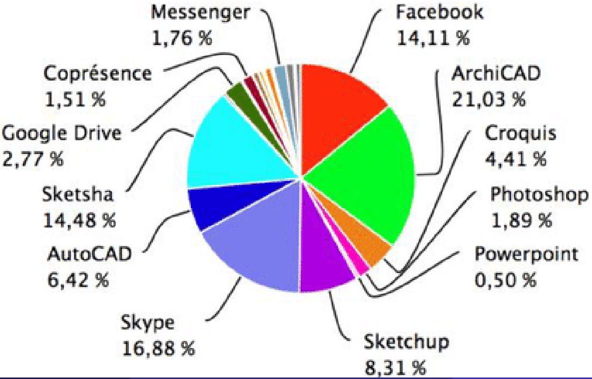
1. **Pie chart :**A bar chart is a graphical representation used to display and compare the frequency or proportion of different categories of a qualitative (categorical) variable.

- Represents the proportion of each category as a sector of a circle.
- Useful for showing the relative distribution of qualitative variables.

1. **Bar Chart**

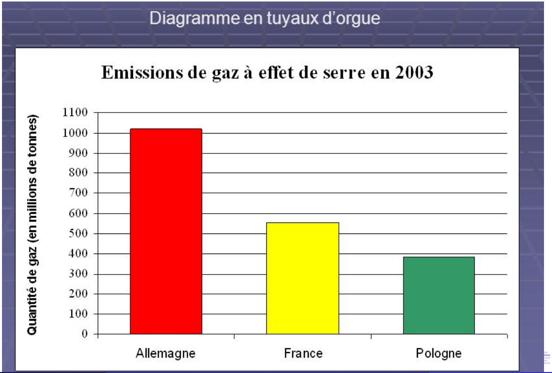
- Displays categories on the x-axis and their frequencies (or percentages) on the y-axis.
- Bars are separated, as qualitative data are not continuous.

1 Pie chart: Un diagramme en camembert (ou diagramme à secteurs): is a circular statistical graphic used to represent the proportions of different categories in a dataset. Each category is represented as a sector of the circle, with its size proportional to its relative frequency..



2- Bar chart:

- 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



:

Common Graphical Representations for Continuous Data

1. **Histogram**

- Divides the data into intervals (classes) and represents the frequency of values within each interval.
- Bars are adjacent (no gaps) since the data are continuous.

2. **Frequency Polygon**

- A line graph connecting the midpoints of histogram bars..

3. **Cumulative Frequency Curve (Ogive)**:

- A curve showing the cumulative sum of frequencies.
- Useful for identifying percentiles, medians, and quartiles.

:

Graphical Representations of Discrete Variables
Discrete variables take specific, countable values (e.g., number of children, number of cars in a household). Their graphical representations focus on showing the frequency of each value.

1. **Bar Chart**

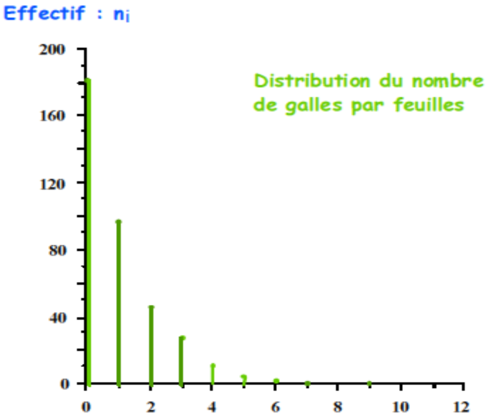
- Each discrete value is represented by a separate bar.
- The height of the bar corresponds to frequency or percentage.

2 **Line Graph** (Frequency Polygon for Discrete Data)

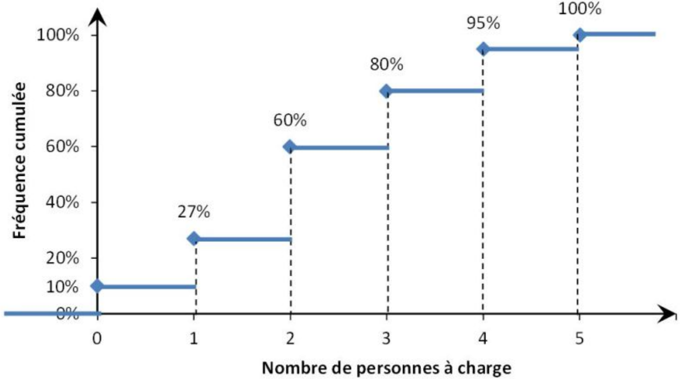
- Plots discrete values on the x-axis and their frequencies on the y-axis.
- Points are connected by straight lines to show trends.

3 . **the stepped curve:** la courbe en escalier

- **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)**

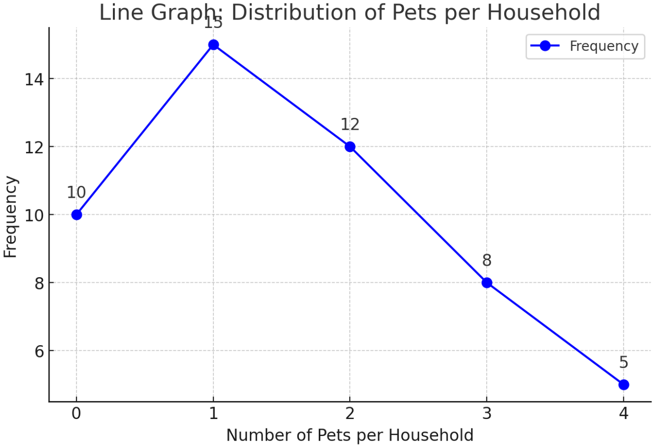


2. the stepped curve:l

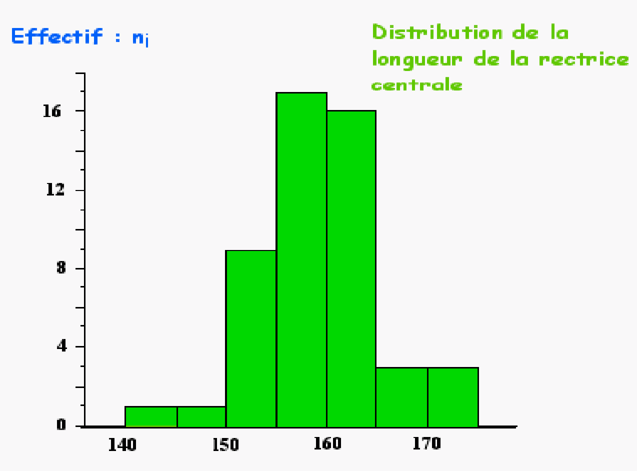


the stepped curve

:
3. the Line Graph:

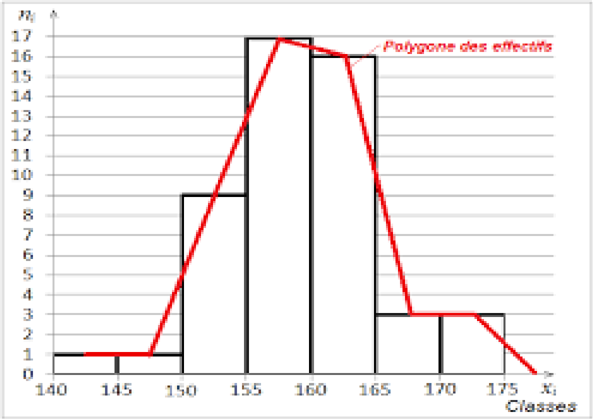


1. The Histogram:



X :longueur de la rectrice centrale en mm

2. Frequency polygons: Polygonne des effectifs: d'une **variable continue:**



3. Cumulative Frequency Curve (Ogive) : la courbe des effectifs cumulées croissant et décroissant sont présentés