Descriptive statistics

— Efficiently describe large datasets.

— Search for the existence of a relationship (correlation) between two variables.

— Interpolate and extrapolate data.

I - One-variable statistical series

1) Some vocabulary...

1. A **character**, or **Statistical variable**, is a property common to **individuals** of a **population**.

2. A **sample** is a part of the complete population.

3. The **Total frequency**, noted N of a population or sample is the number of individuals in it.

4. A character can be **quantitative**, if it can be expressed by a number or **qualitative** (nationality eye colour,...) in the opposite case. otherwise.

5. We can also distinguish between quantitative character **discrete** which only take isolated numerical values (e.g. number of pupils per class) and quantitative character **continuous** where all the values of an interval can be taken (within an interval) (e.g. pupil size, life of a component).

2) Description by distributions

◇ **Frequency** : For a characteristic value (modality or class), the number of individuals in the population with this value is called the frequency. We often note n_1, $n_2, \dots n_p$ the respective frequencies of modalities x_1, $x_2, \dots x_p$.

◇ **Total frequency** : Total number of individuals in the population (or sample). It is equal to $n_1 + n_2 + \ldots + n_p$, often noted N.

◇ **Relative Frequency** : For a character value (modality or class), the quotient of the frequency n_i of this value x_i by the total frequency N is called the relative frequency noted by f_i. We often note f_1, $f_2, \dots,$ f_p the respective relative frequencies of the modalities x_1, $x_2, \dots x_p$, so :

$$f_1 = \frac{n_1}{N},\ f_2 = \frac{n_2}{N}, \dots,\ f_p = \frac{n_p}{N}\,.$$

We deduce that : $0 \leq f_k \leq 1$ for $0 \leq k \leq p$, and $\sum_{k=1}^{p} f_k = 1$.

◇ **Extreme values** : Minimum and maximum values of a quantitative character noted by $x_{\min}$ and $x_{\max}$ respectively.

◇ **increasing Cumulative frequency** : For an x_i value in a quantitative statistical series, its increasing cumulate frequency, noted by $N_i \nearrow$, is the sum of its frequency n_i and the frequencies of all the modalities preceding it. So,

$$N_i \nearrow = \sum_{k=1}^{i} n_k$$

◇ **decreasing Cumulative frequency** : For an x_i value in a quantitative statistical series, its increasing cumulate frequency, noted by $N_i \searrow$, is the sum of its frequency n_i and the frequencies of all the modalities following it. So,

$$N_i \searrow = \sum_{k=i}^{N} n_k$$

◇ **increasing Cumulative relative frequency** : For an x_i value in a quantitative statistical series, its increasing cumulate frequency, noted by $F_i \nearrow$, is the sum of its frequency f_i and the frequencies of all the modalities preceding it. So,

$$F_i \nearrow = \sum_{k=1}^{i} f_k$$

◇ **decreasing Cumulative relative frequency** : For an x_i value in a quantitative statistical series, its increasing cumulate frequency, noted by $F_i \searrow$, is the sum of its frequency f_i and the frequencies of all the modalities following it. So,

$$F_i \searrow = \sum_{k=i}^{N} f_k$$

3) Description by mean

(I)-Ungrouped data

Définition *Consider N values of a character* **Statistical series** *x_1, x_2,...,x_N.* **The mean or arithmetic mean**, *denoted by $\overline{X}$, is :*

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

If the value x_i is taken n_i times by the character for $1 \leq i \leq k$, then

$$\overline{X} = \frac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{N} = \frac{1}{N}\sum_{i=1}^{k} n_ix_i \quad avec \quad N = \sum_{i=1}^{k} n_i$$

This is known as a **weighted mean**.

The mean of a series can be used to locate its overall level : it is a **positional characteristic**, but does not give any information about the distribution, or **dispersion**, of the values around this central position.

For example, the statistical series : $10, 10, 10, 10, 10, 10, 10$ and $2, 2, 2, 10, 18, 1818$ have the same number and the same average, even though they are clearly different.
There are three other means.

Définition *For a statistical series of size N, we have :*

1. *Geometric mean, noted by G or G_X, and is defined by*

$$G = (\prod_{i=1}^{N} x_i)^{\frac{1}{N}}$$

Note that : $\log G = \frac{1}{N}\sum_{i=1}^{N} \log x_i$

2. *Harmonic mean, noted by H or H_X, and is defined by*

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

3. *Quadratic mean, noted by Q or Q_X, and is defined by*

$$Q = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

By using frequencies distribution $\{n_1, n_2, ..., n_k\}$ and $N = \sum_{i=1}^{k} n_i$, we have :

1. $$G = (\prod_{i=1}^{k} x_i^{n_i})^{\frac{1}{N}}$$

2. $$H = \frac{N}{\sum_{i=1}^{N} \frac{n_i}{x_i}}$$

3. $$Q = \frac{1}{N}\sum_{i=1}^{N} n_i x_i^2$$

By using relative frequencies distribution $\{f_1, f_2, ..., f_k\}$ and $\sum_{i=1}^{k} f_i = 1$, we have :

1. $$\overline{X} = \sum_{i=1}^{k} f_i x_i$$

2. $$G = (\prod_{i=1}^{k} x_i^{f_i})$$

3. $$H = \frac{1}{\sum_{i=1}^{N} \frac{f_i}{x_i}}$$

4. $$Q = \sum_{i=1}^{N} f_i x_i^2$$

(II)-Grouped data Here the data set is grouped in classes $[a_i, b_i[$ of centers c_i for $1 \leq i \leq k$. Then,

1. $$\overline{X} = \frac{\sum_{i=1}^{k} n_i c_i}{N} \text{ or } \overline{X} = \sum_{i=1}^{k} f_i c_i$$

2.
$$G = (\prod_{i=1}^{k} c_i^{n_i})^{\frac{1}{N}} \text{ or } G = (\prod_{i=1}^{k} c_i^{f_i})$$

3.
$$H = \frac{N}{\sum_{i=1}^{N} \frac{n_i}{c_i}} \text{ or } H = \frac{1}{\sum_{i=1}^{N} \frac{f_i}{c_i}}$$

4.
$$Q = \frac{1}{N} \sum_{i=1}^{N} n_i c_i^2 \text{ or } Q = \sum_{i=1}^{N} f_i c_i^2$$

Propriété *We have,*
$$H \leq G \leq X \leq Q$$

4) Description by mode

Définition
- *The value that occurs most frequently in a distribution is referred to as mode.*
- *It is symbolised as M_0. Mode is a measure that is less widely used compared to mean and median.*
- *There can be more than one type mode in a given data set.*

5) Computing Mode for Ungrouped Data

While computing mode from the given data sets all measures are first arranged in ascending or descending order. It helps in identifying the most frequently occurring measure easily.

Example 1 : Calculate mode for the following test scores in geography for ten students : 61, 10, 88, 37, 61, 72, 55, 61, 46, 22

Computation : To find the mode the measures are arranged in ascending order as given below : 10, 22, 37, 46, 55, 61, 61, 61, 72, 88.

The measure 61 occurring three times in the series is the mode in the given data set. As no other number is in the similar way in the data set, it possesses the property of being unimodal.

Example 2 : Calculate the mode using a different sample of ten other students, who scored : 82, 11, 57, 82, 08, 11, 82, 95, 41, 11.

Computation : Arrange the given measures in an ascending order as shown below : 08, 11, 11, 11, 41, 57, 82, 82, 82, 95

It can easily be observed that measures of 11 and 82 both are occurring three times in the distribution.

The dataset, therefore, is bimodal in appearance. If three values have equal and highest frequency, the series is trimodal.

Similarly, a recurrence of many measures in a series makes it multimodal.

However, when there is no measure being repeated in a series it is designated as without mode.

6) Computing Median and the mode for grouped Data

Let be data set grouped in the classes $[a_i, b_i[$ of frequencies n_i or relative frequencies f_i for $i = 1, 2, ..., k$

1. First, we locate the median M_e by determining the median class which class contains it.

2. The median class $[a_m, b_m[$ coincides with the first class when its cumulative increasing frequency $N_m \nearrow$ is at least half the total frequency N (so, its cumulative increasing relative frequency $F_m \nearrow$ is at least 0.5).
3. Hence,

$$M_e = a_m + \overbrace{(b_m - a_m)}^{h}\frac{\frac{N}{2} - N_{m-1}\nearrow}{n_m} \text{ and } M_e = a_m + \overbrace{(b_m - a_m)}^{h}\frac{0.5 - F_{m-1}\nearrow}{f_m}$$

1. First, we locate the mode M_o by determining the modal class which class contains it.
2. The modal class $[a_m, b_m[$ coincides with a class of the highest frequency n_m class and the highest relative frequency f_m.
3. Hence,

$$M_o = a_m + \overbrace{(b_m - a_m)}^{h}\frac{\Delta_i}{\Delta_s + \Delta_i}$$

where

$$\Delta_i = n_m - n_{m-1} \text{ and } \Delta_s = n_m - n_{m+1} \text{ or } \Delta_i = f_m - f_{m-1} \text{ and } \Delta_s = f_m - f_{m+1}$$

7) Description by the median and quantiles

Définition **The median, noted by M_e, of a statistical series ordinate is a value which divides this statistical series into two statistical sub-series of equal size.**

If the total number in the series is odd : $N = 2p+1$ the median is the $(p+1)^{\text{th}}$ value.
If the number is even : $N = 2p$, the median is generally taken to be the mean of the p^{th} and the $(p+1)^{\text{th}}$ values.

The mode of a statistical series is the most frequent value of the most frequent character.

Exercice 1 1) In a small company, the boss earns €10,000 a month and his 9 employees earn €1,500. What is the average salary in the company ? The median salary ?

2) Rechercher les montants des salaires moyen et médian en France. Commenter.

As with the mean, the median is a **characteristic of position** and does not account for the **dispersion**of values. To describe a statistical series, we must therefore also characterise the dispersion of the values around this position.

Définition

1. *The* **range** *of a series is the difference between the extreme values of the series.*
2. *The* **quartiles** *Q_1, Q_2 and Q_3 of a series are three values of the ordered series which divide it into four series of the same (25% of the total number.*
3. *The second quartile is the median : $Q_2 = M_e$.*
4. *The inter-quartile range is the number $Q_3 - Q_1$.*
5. *The* **deciles** *D_1, D_2,...,D_9 of a series are defined in the same way, by dividing the series into ten series of the same size (10 % of the total size.*
6. *The inter-decile gap is the number $D_9 - D_1$.*

Note : In the case of a continuous statistical series, the values are grouped into classes (or intervals). The statistical indicators are then calculated using the centres of the classes.

Exercice 2 Let be the statistical series :

Notes x_i	6	8	10	12	15	18
Nombre d'élèves n_i	1	5	3	4	2	2

The mean for this series is $\overline{X} = \ldots$

The variance is : $V(X) = \ \ldots$

and the standard deviation : $\sigma_X = \ldots$

The total number is $N = \ \ldots.$

The median is therefore the $\ldots^{textth}$.

value of the ordered series, i.e. $M_e = \ldots$

Its mode is 8.

8) Comparison of Mean, Median and Mode

For a symmetric data set, the mean median and mode will be approximately equal.

For Skew data set, the median is less sensitive than the mean to extreme observations. The mode ignores them.

The three measures of the central tendency could easily be compared with the help of normal distribution curve. The normal curve refers to a frequency distribution in which the graph of scores often called a bell shaped curve.

The mode is dependent on the choice of class intervals and is therefore not favoured for sophisticated work.