Chapter1: Basic Concept of Statistics

I. Introduction

Statistics is a branch of mathematics that pertains to the collection, analysis, interpretation or explanation, and presentation of **data**. Statistics is a science generally concerned with the use of data in the context of uncertainty and decision-making in the face of uncertainty.

II. Definition

Statistics is the set of methods used to organise experiments providing observations leading to collecting data, analysing them and interpreting the results.

The statistical analysis is subdivided into two parts

1. Desciptive *Statistic*: aims to describe, i.e. to summarize or represent the data.
Typical questions:
*Raw representation (statistical series)
*Tabular Representation
*Graphic Representation
*Numerical Summaries or characteristics or indicators (Position, dispersion, relationship parameters.)

2. Inferential *Statistics* : The set of methods used to formulate a judgment. It requires more advanced mathematical tools (probability theory).

III. Basic Concepts

**POPULATION :* The studied collection of objects or people (students, computers, cars,…)

**INDIVIDUAL :* element of the studied population. (one student, one computer, one car,…).

*SAMPLE: a part of the studied population. The cardinality of a sample is called the sample size, denoted by n.

**VARIABLE (CHARACTER)* : Generaly denoted by a capital letter (X, Y, Z, ...) individuals common property that one aims to study it.
A character can be :

a) qualitative(categorical): we cannot associate to it a numerical value (eye color, processor, type of car, etc.).
A qualitative character includes:
*nominal: its data consist of labels or names (eyes color, type of car…)
*ordinal : designates the rank (convetional ordre) such as: level of importance (none, little, medium, quite, a lot).

b) quantitative : has numerical values (weight, the amount of RAM, processor speed, storage capacity, price, etc.).
A qualitative character includes:
*Continuous : can take all the numerical values of a determined interval (size, etc.), it results from a measurement.
*Discrete : can not take but isolated numerical values (number of residential rooms, number of damaged fruits, etc.), it involves a count or enumeration.

*MODALITY: one of the particular forms of a character. Eyes colour is a character, its modalities are: blue, green, brown, etc. In the context of a quantitative character we speak about VALUE. A modality or a value is generaly denoted by the small letter of that of the character i.e. if the character is denoted X its corresponding modality or value is denoted x.

*STATISTICAL SERIES: A statistical series is the sequence of modalities or values that a character takes within a sample.

*SAMPLE SIZE: the size of a sample is the number of its elements.

*ABSOLUTE FREQUENCY (denoted n_i): is the number of occurrence of a modality or a value assiciated to the character within the sample.

*RELATIVE FREQUENCY (dnoted f_i): The frequency of occurrence of a modality or value associated with a characteristic in a sample $f_i = \frac{n_i}{N}$

*CUMULATIVE FREQUENCY (noté n_i^{cum}): $n_i^{cum} = \sum_{j=1}^{i} n_j$. It is interpreted as the number of individuals who have the modality or value less than or equal to the corresponding modality.

*CUMULATIVE RELATIVE FREQUENCY (denoted by f_i^{cum}): $f_i^{cum} = \sum_{j=1}^{i} f_j = \frac{n_i^{cum}}{n}$. It is interpreted as the frequency of individuals who have the modality or value less than or equal to the corresponding modality.

*POURCENTAGE :(expresed in %): is a realive frequency multiplied by 100.

III. Illustrative Example: In order to conduct a statistical study on First year students at the departement of Mathematics, the teacher ask at the first lecture, his students to provide responses on:

- The color of their eyses.
- Their behaviour towrds morning coffee.
- The number of sisters and brothers they have.
- Their heights in cm.

In a such study the population studied is all the dpartement mathematics First year students subsribed for the current academic year. Yet not all of them was their at the first lecture only 20 were there, so those who attend to the first lecture formed a sample of size 20 indivduals to the study.

We can distinguish four statistical variables (characters) of four differents type:

- ✓ X: Eyes color.
- ✓ Y: Behaviour towards the coffee.
- ✓ Z: Number of brothers and sisters.
- ✓ H: the Height (in m). H is a quantitative continuous character.

Responses provided by the students are the data that will be studied. These data are given by in the four statistical series corresponding to the four character respectively:

X: Black, Blue, Blue, Black, Brown, Blue, Black, Blue, Green, Brown, Brown, Green, Brown, Brown, Brown, Black, Blue, Black, Brown, Green.

Y: Somtimes, Often, Somtimes, Always, Often, Always, Often, Always, Somtimes, Always, Often, Somtimes, Somtimes, Never, Often, Never, Somtimes, Always, Never, Somtimes.

Z: 4 3 5 6 1 3 7 4 5 4 2 2 3 3 2 5 3 3 0 4

H: 1.59 1.45 1.53 1.73 1.50 1.72 1.61 1.50 1.71 1.63 1.80 1.58 1.69 1.66 1.69 1.75 1.73 1.65 1.64 1.55.

Notice that:

- ✓ X is a qualitative nominal character of four modalitie: Black, Blue, Green and Brown.
- ✓ Y is a qualitative ordinal character of four modalities: Never, Often, Somtimes, Always.
- ✓ Z is a quantitative discrete character of values: 0 1 2 3 4 5 6 7.
- ✓ H is a quantitative continuous character of values ranging between 1.50m and 1.80m.