UNIT 2. BASIC CONCEPTS. SCALES OF MEASUREMENT

BASIC CONCEPTS

Statistical methods are a set of **models**, **methods** and **techniques** to collect, organize, analyse and interpret the information that **help** us to **make decisions** more **scientifically**.

Population, or universe is the set of objects or individuals of the same nature, defined without ambiguity, of which a series of characteristics or behaviours are to be studied. In statistics, the term **subject** refers to each one of the elements of the population. They can refer to such diverse things as human beings, provinces, companies, buildings, etc. Once the population under study has been precisely defined, who are the subjects that make it up, we must indicate the **characteristics**, called **variables**, that we are interested in. Given a set of individuals, the observation of whether or not a certain characteristic is realized in them, or the measurement of the intensity with which it is verified is called **statistical data**.

To be more precise, we will call **data** the set of raw values, without any given orientation (the figure 1.8); **information** is the organised or structured data, which makes it appropriate for a certain purpose (the height 1.8 m); finally, **knowledge** is the interpretation of the information at a specific domain (the height 1.8 m corresponding to a pygmy woman).

In order to study a certain characteristic of the individuals in a population, we can either collect information from all the individuals that make up the population, which is called a **census**, or from only a significant part of them by taking a **sample** of the population. The sampling alternative is chosen for reasons such as economy, speed, quality, impossibility or destructive observation, among others.

TYPES OF VARIABLES

It is obvious that the characteristics to be studied can be very different and all experiments examine some kind of variable. The reason why it is important to know what kind of variables we are dealing with, is because the statistical techniques to be used will depend on this, and therefore the validity of the conclusions to be drawn from their analysis.

A company's share price is not the same as the level of education of an unemployed person. While the share price reflects a characteristic that is naturally associated with a number, the level of education reflects a quality and its observation does not require the use of a number, but rather a range of categories or modalities. This example makes it possible to distinguish between the two main types of variables: quantitative and qualitative variables.

Qualitative variables or attributes are those that reflect a quality of the individual whose observation is not necessarily linked to a number. These variables are measured by categories or modalities that must meet the following three requirements: they must be well defined, they must be mutually exclusive, and each and every individual must be included in one of the categories. Although a number may sometimes be associated with each of the modalities, these numbers do not have the meaning of quantities.

Quantitative variables are those whose value in an individual is naturally indicated by a number, since they correspond to characteristics that are represented by quantities and for which arithmetic operations are completely meaningful. Examples are: height, age, salary, temperature, a stock market index or concentration of toxic substances in the environment (measured in parts per million). Quantitative variables can, in turn, be discrete or continuous.

A **discrete quantitative variable** takes a finite or infinite number of numerable values. For example, the number of students in a course, the number of members in the family unit and the prices of shares on the Madrid Stock Exchange.

A **continuous quantitative variable** takes an infinite number of values in a finite interval. The observation of a conceptually continuous variable is usually carried out in practice in a discrete form because of the limitations of the measuring instruments. In this sense, for example, people's heights can in principle take an infinite number of values, but in practice they are measured to the nearest centimetre. Other examples of continuous quantitative variables are weight and temperature.

SCALES OF MEASUREMENT

The procedure of assigning numbers to an elementary unit's attribute variables is referred to as **measurement**. A m**easurement scale** is the reference employed to measure the value or to observe the category of a variable in a subject. The four scales of measurement are: nominal, ordinal, interval and ratio.

Not all statistical methods are meaningful or valid when data are measured on certain scales, particularly on nominal and ordinal scales. Therefore, data analysts must recognize the measurement scale and choose the statistical methods appropriate for that scale.

A **nominal scale** assigns values to a list of categories, and thus, a variable measured in this type of scale is qualitative. A nominal scale enables us to tell when entities are the same or not with respect to a specific property. Data measured in a nominal scale may be identified by a name, letter code, arbitrary symbol, or a number. Numerical designations are merely identification tags but are generally preferred when preparing data for computer analysis, thus it is meaningless to use any arithmetic operation. A special case are binary variables that present only two categories. Examples of variables measured in a nominal scale are: *Gender*, *Corporate divisions*, *Country*, *Race*, *Nationality*, *Marital status*, ….

An **ordinal scale** assigns values to categories (then, being also qualitative) but also **rank** orders the categories according to some property, providing a **stronger** form of measurement than the nominal scale. An ordinal scale does not assume equal

intervals or distances between the possible numerical indicators. Examples of variables measured in an ordinal scale are: *Degree of agreement with a statement*, *Level of education*, *Socioeconomic status, Mohs' scale of mineral hardness, …*

An **interval scale** assigns a numerical value, thus being quantitative, in such a way that numerically equal distances on the scale represent equal changes in the property being measured. The zero point is arbitrary, because it does not represent the absence of the property being measured. Examples of variables measured in an interval scale are: *Temperature in degrees Celsius or Fahrenheit*, *Calendar dates*, *Scores obtained in an intelligence test*, …

Finally, a **ratio scale** assigns a numerical value (thus, also quantitative) and includes all the features of the previous three scales but it also includes the concept of an absolute zero point or origin. An absolute zero-point permits multiplication and division. Therefore, it makes sense to say that a distance of 6 metres is twice as long as 3 metres. Examples of variables measured in a ratio scale are: *Height*, *Weight*, *Total sales volume*, *Number of children in a family*, *Geographical distance*, etc…

DATA TYPES

Another categorization of data types reflects the process by which the data were gathered:

- **Cross-sectional data** are observations of many different individuals (subjects, objects) at a given time, each observation belonging to a different individual. One example are the *values of the gross annual income for each of 1000 randomly chosen households in New York City for the year 2000.*
- **Longitudinal data** or **time-series data** refer to observations of one individual (subject, object) made over time. Example: *values of the total gross annual income of one household from New York City for the years 1991-2000.*

- A **panel data** set combines cross-sectional and time-series data, containing observations of many different individuals made over time. Example: *values of the gross annual income for each of 1000 households in New York City, collected for each of 10 years from 1991 to 2000.*