Spaces

Before modeling or analyzing data, it is essential to fully understand the nature of the variables being manipulated. The **type of variable** determines:

- the mathematical space in which it exists;
- the distance measures that can be used to compare it to others;
- and the relevant models to use.

In this section, we present the most common types of variables, as well as the associated spaces.

1 Statistical unit

A **statistical unit** is the basic element on which an observation is made. Morally, it is the "carrier" of the information that is used to determine the level of aggregation of the analysis. The statistical unit is a **choice** made by the modeler.

Examples

- In the case of an income survey, the individual can be chosen as the unit.
- In the case of a study of high school classes, the class can be chosen as the unit.
- In the case of a medical imaging database, the image can be chosen as the unit.

Sometimes, the same database can be analyzed at several levels.

An image is made up of pixels, each of which can be described by numerical variables (e.g., RGB values, opacity, etc.). You can choose to analyze each pixel, and therefore take the pixel as the unit, or analyze each image as a whole, and therefore take the image as the unit.

2 Types of variables

There are generally four types of variables, which are identified at the level of the smallest statistical unit in the dataset.

Numerical (or quantitative) variable

A numerical (or quantitative) variable is a variable whose values are numbers representing a measurable quantity.
Examples: income in dollars, weight, age, etc.

Ordinal variable

An ordinal variable is a qualitative (or categorical) variable whose categories can be ordered naturally, without the difference between the categories being quantifiable.
Examples: income level (low, medium, or high), satisfaction level ("strongly disagree," "disagree," "no opinion," "agree," "strongly agree"), etc.

Symmetric nominal variable

A symmetric nominal variable is a qualitative (or categorical) variable in which all categories are equally informative.
Examples: nationality, field of study, etc.

Asymmetric nominal variable

An asymmetric nominal variable is a qualitative (or categorical) variable in which one of the categories has a special status, often being more frequent or considered the "default" value. Thus, having two observations with the "default" value of this asymmetric nominal variable does not tell us much about them, whereas we can glean much more information from two observations that do not have the "default" value.
Examples: presence or absence of a symptom, fraudulent transaction or not, etc.

Although these types of variables are the most common, there are many other types of variables. For example, we may be interested in comparing curves, texts, images, networks, etc. In these situations, the choice of representation depends on the level at which we wish to place ourselves, and therefore on the statistical unit.

3 Associated spaces

Once our data has been collected, the first step in statistical analysis is to choose a mathematical space in which to work. This space, sometimes called the **observation space** and denoted by $\mathcal{X}$, depends on the type of data observed. It constitutes the formal framework in which our variables take their values, and it guides the methodological choices that follow.

Case of a numerical variable

When observing a numerical variable (e.g., the temperature of a country), the natural space in which to work is the set of real numbers, $\mathcal{X} = \mathbb{R}$. In some cases, this space can be restricted to a specific interval. For example, if we are interested in a person's height, we can take $\mathcal{X} = [0, +\infty)$ because the variable in question cannot be negative.

Case of a nominal (or qualitative, or categorical) variable

For a nominal variable, the space is a finite set of modalities, the set of modalities taken by the variable. For example, if we study the results of a dice roll, the variable can take values from 1 to 6, and the associated space will therefore be $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.

When the data is more complex, more suitable spaces must be chosen. For curve or signal analysis, we can work in a function space. For example, we can consider the space of continuous functions on a closed interval $[a, b]$, denoted $\mathcal{X} = \mathcal{C}([a, b])$. For text analysis (viewed as a sequence of characters), the workspace can be an alphabet. For example, we can consider $\mathcal{X} = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}\}$.

Often, several variables are observed at the same time, e.g., the height, weight, and gender of an individual. In this case, the observation space will be the Cartesian product (also called the product set) of the spaces associated with each variable:

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \mathcal{X}_p,$$

where p is the number of variables. In the case where we observe p numerical variables, we will simply write $\mathcal{X} = \mathbb{R}^p$.