Chapter 1

Preliminaries

1.1 Introduction

Statistics is the study of the collection of data, their analysis, their processing, the interpretation of the results and their presentation in order to make the data understandable by everyone. It is at the same time a science, a method and a set of techniques. Data analysis is used to describe the phenomena studied, make predictions and make decisions about them. In this way, statistics is an essential tool for understanding and managing complex phenomena. Statistics are useful in all disciplinary fields, from economics to biology through psychology and of course engineering sciences. The statistics consist of:

- Collect data.
- Present and summarize this data.
- Draw conclusions about the population studied and assist in decision-making.
- In the presence of time-dependent data, we try to make predictions.

1.2 Vocabulary

Descriptive statistics aim to study the characteristics of a set of observations such as the measurements obtained during an experiment. The experiment is the preliminary step to any statistical study. Generally, the statistical method is based on the following concept.

1.2.1 Statistical test

Definition 1.1. The **statistical test** is an experience that we provoke.

Example 1.1. A manufacturer of electric bulbs having the choice between 4 types of filaments intends to study the influence of the nature of the filament on the lifespan of the bulbs manufactured. To do this, he will have 4 samples of identical bulbs made, except for the filament, burn the bulbs until they go out, then compare the results obtained.

1.2.2 Population

In statistics, the term **population** applies to any statistical object studied, whether students (of a university or a country), households or any other group on which we make statistical observations.

Definition 1.2. We call **population** the group on which our statistical study relates. This set is denoted Ω.

Example 1.2. 1. We consider all the students in section A. We are interested in the number of brothers and sisters of each student. In this case Ω = all students.

2. If we now look at automobile traffic in a city, the population is then all the vehicles likely to be circulating in this city on a given date. In this case: Ω = all vehicles.

1.2.3 Individual (statistical unit)

A population is composed of individuals. The individuals who compose a statistical population are called **statistical units**.

Definition 1.3. We call **individual** any element of the population Ω, it is denoted $\omega \in \Omega$.

Example 1.3. If we study the annual production of a factory of metal beverage cans. The population is all the boxes produced during the year and a box constitutes an individual.

1.2.4 Character (statistical variable)

"Descriptive" statistics, as its name suggests, seeks to describe a given population. We are interested in the characteristic of units which can take different values.

Definition 1.4. We call **character** (or **statistical variable**, denoted S.V) any application

$$X : \Omega \to C$$

where C is the set of values of the character X (this is what is measured or observed on individuals).

Example 1.4. Height, temperature, nationality, eye color, professions.

1.2.5 Modalities

The modalities of a statistical variable are the different values that can take the statistical variable.

Example 1.5. Variable: "family situation". Modalities: single, married, divorced.

1.3 Types of Characters

We distinguish two categories of characters: qualitative characters and quantitative characters.

1.3.1 Qualitative character

Qualitative characters are those whose modalities cannot be ordered, that is to say if we consider two characters taken at random, we cannot say of one that it is inferior or equal to the other.

Definition 1.5. The elements of C cannot be represented by numbers.

Example 1.6. The condition of a house: we can consider the following modalities: ancient, degraded, new, renovated.

1.3.2 Quantitative character

Quantitative characters are those whose modalities can be ordered. Thus, the age, life size or salary of an individual are quantitative characteristics.

Definition 1.6. All values are represented by numbers. Likewise, it is divided into two kinds of characters, discrete and continuous.

Example 1.7.
1. The salary of factory employees. Modalities: 10000da, 20000da, etc. Type: discrete.
2. The rigidity of the springs. Modalities: [10, 20] N/m. Type: continuous.

1.4 Statistical Tables and Graphical Representations

Statistical tables and graphs are used to summarize and illustrate the results of data collection. They allow us to visualize how the values of a variable are distributed in a given population. We distinguish between the case of **qualitative variables** and **quantitative variables**.

1.4.1 Case of Qualitative Variables

Statistical tables and graphs are used to summarize and illustrate the results of data collection. They allow us to visualize how the values of a variable are distributed in a given population. We distinguish between the case of **qualitative variables** and **quantitative variables**.

1.4.2 Definition of a Statistical Table

Definition 1.7. A **statistical table** is an organized arrangement of data that summarizes the observed values of a variable and their corresponding frequencies. It allows a clear and structured presentation of information before any graphical representation or calculation of indicators (mean, variance, etc.).

A statistical table generally contains the following components:

- **Modalities (or classes):** the different values or categories taken by the studied variable.
- **Absolute frequency (n_i):** the number of times a modality or value occurs in the dataset.
- **Relative frequency (f_i):** the ratio of n_i to the total number of observations N:

$$f_i = \frac{n_i}{N}, \quad \text{with } \sum_i f_i = 1$$

- **Percentage frequency:** obtained by multiplying the relative frequency by 100:

$$\text{Percentage} = f_i \times 100$$

- **Cumulative frequency:** for ordered variables, it represents the sum of the frequencies of all modalities less than or equal to a given value. It can be expressed in absolute, relative, or percentage form.

We now distinguish between the case of **qualitative** and **quantitative** variables.

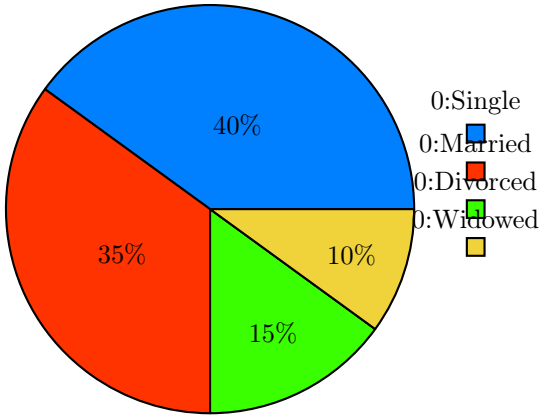
Example: Marital Status of 100 People

Modality	Absolute Frequency	Relative Frequency	Percentage (%)
Single	40	0.40	40%
Married	35	0.35	35%
Divorced	15	0.15	15%
Widowed	10	0.10	10%
Total	100	1.00	100%

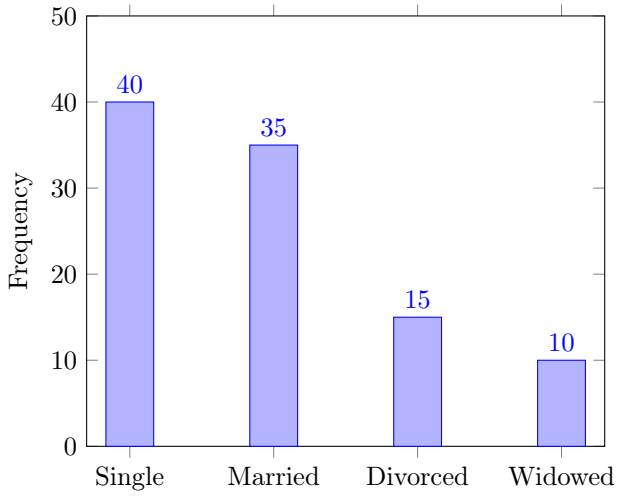
1. Circular Diagram (Pie Chart)

Each modality is represented by a circular sector whose angle is proportional to its frequency:

$$\text{Angle} = \frac{n_i}{N} \times 360^\circ$$



2. Bar Chart



1.4.3 Case of Quantitative Variables

Definition 1.8. A **qualitative variable** (or categorical variable) describes a quality, attribute, or category that cannot be measured numerically. Its possible values are called **modalities**.

Example: color of a car, marital status, nationality, type of housing, etc.

Definition 1.9. A **modality** is one of the possible categories or labels that a qualitative variable can take. Each individual in the population belongs to one and only one modality of the variable.

The statistical table for a qualitative variable usually includes:

- the list of modalities;
- the absolute frequencies (n_i);
- the relative frequencies (f_i);
- and optionally, cumulative or percentage frequencies.

Example: A survey of 50 people on their preferred drink gives the following results:

Drink	**Absolute frequency** (n_i)	**Relative frequency** (f_i)	**Percentage (%)**
Tea	20	0.40	40%
Coffee	15	0.30	30%
Juice	10	0.20	20%
Water	5	0.10	10%
Total	50	1.00	100%

Interpretation: Tea is the most preferred drink among respondents (40%).

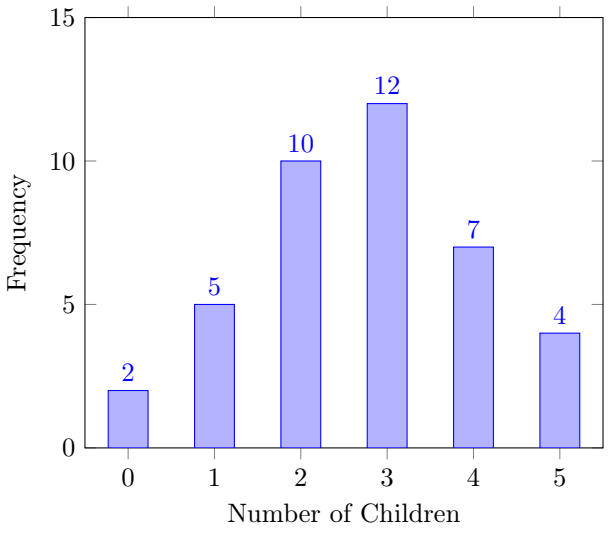
Graphical representations for qualitative variables include:

- Pie chart (circular diagram);
- Bar chart (column diagram);
- Band diagram (100% stacked bar).

Example: Number of Children per Family (Discrete Variable)

Number of Children (x_i)	**Frequency** (n_i)	**Relative Freq.** (f_i)	**Cumulative Freq.**
0	2	0.05	0.05
1	5	0.13	0.18
2	10	0.25	0.43
3	12	0.30	0.73
4	7	0.18	0.91
5	4	0.09	1.00
Total	40	1.00	

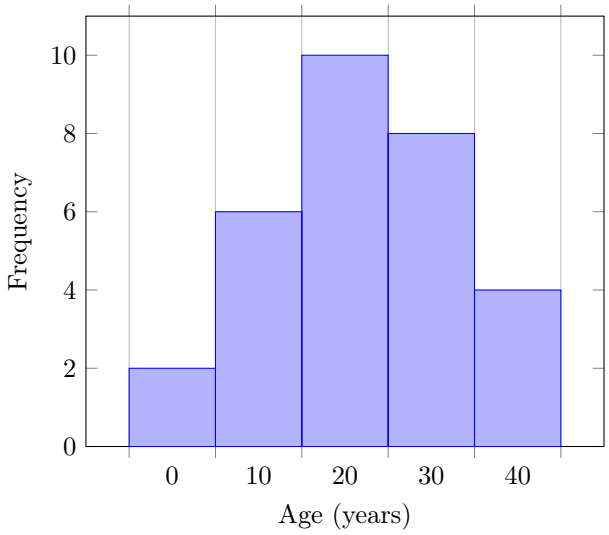
1. Bar Diagram



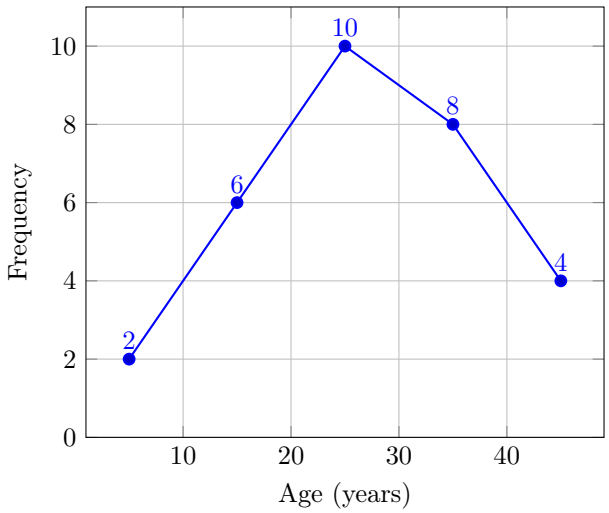
Example: Age Distribution (Continuous Variable)

Interval	Frequency	Midpoint	Cumulative
$[0, 10[$	2	5	2
$[10, 20[$	6	15	8
$[20, 30[$	10	25	18
		etc.	

2. Histogram



3. Frequency Polygon



1.5 Exercises with Solutions

1.5.1 Exercises on Basic Statistical Concepts

Exercise 1.1. School Survey

A study is carried out in a high school to determine:

- the number of siblings of each student,
- the favorite sport of each student,
- and the weight of each student.

1. Identify the **population**.
2. What is an **individual** in this study?
3. For each of the three characteristics above, give: