Descriptive Statistics

1 Introduction

Descriptive statistics summarize and describe the main features of a dataset. They provide simple summaries and visualizations to better understand data.

Key Objectives:

- Organize data effectively.
- Identify patterns and trends.
- Summarize data numerically and visually.

2 Statistical Vocabulary

Statistics consist of various methods for organizing data, such as tables, histograms, and graphs, allowing for the arrangement of a large amount of data. Statistics developed in the second half of the 19th century in the field of social sciences (sociology, economics, anthropology, etc.). They have established their own specific vocabulary.

2.1 Statistical Test

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 in any statistical study. It involves "getting acquainted" with the observations. In general, the statistical method is based on the following concept.

Definition 1 *The statistical test is an experiment that is deliberately initiated.*

2.2 Population

In statistics, we work with populations. This term originates from the fact that demography, the study of human populations, played a central role in the early development of statistics, particularly through population censuses. However, in statistics, the term "population" applies to any statistical object under study, whether it involves students (from a university or a country), households, or any other group on which statistical observations are made. We define the concept of a population.

Definition 2 *A population is defined as the set on which our statistical study is based. This set is denoted by:* Ω.

2.3 Individual (Statistical Unit)

Definition 3 *The individuals that constitute a statistical population are called statistical units (they are elements of the population).*

2.4 Characteristic (Statistical Variable)

Descriptive statistics, aim to describe a given population. We focus on the characteristics of the units, which can take on different values.

2.5 Statistical modality

Definition 4 *The modalities of a statistical variable are the different values it can take.*

Example 1 *We summarize the different concepts in this example:*

Population: *All the employees of a factory.*

Individual: *Each employee of the factory.*

Characteristic: *Marital status.*

Modalities of the characteristic: *Married, single, divorced and widowed.*

3 Statistical variables

There are two types of characters: qualitative and quantitative.

Quantitative variables are the variables that can be measured, they are characterized by numerical valued, variables whose modalities are numbers.

- **Discrete:** Countable values (e.g., number of students).

  Example 2 *Number of brothers and sisters in a family.*

- **Continuous:** Any value within a range (e.g., weight, height...).

  Example 3 *Height of girls in a 1st year biology class.*

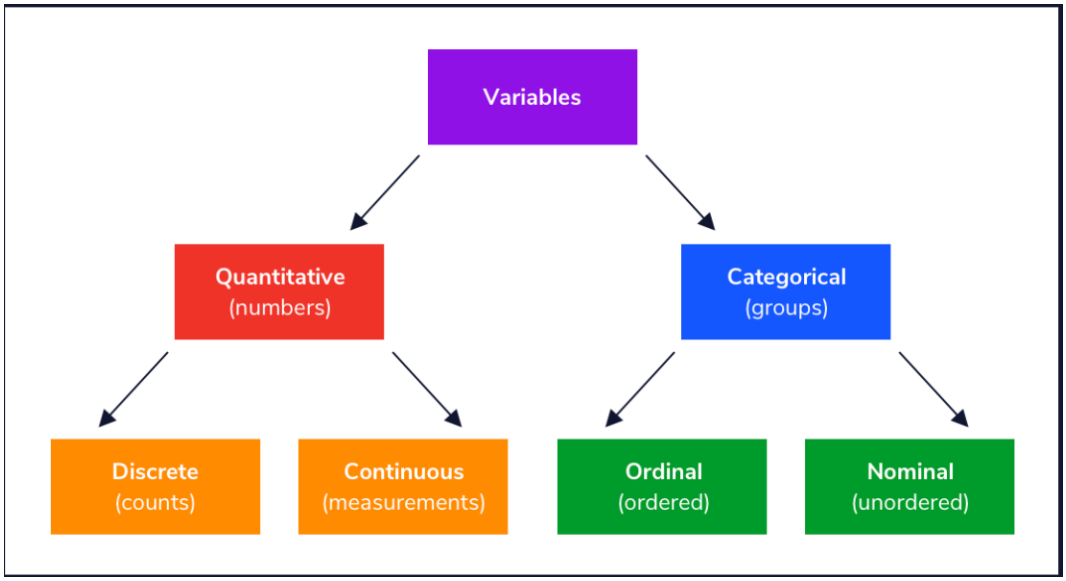
Qualitative variables (Categorical) are varaibles that are not measurable, whose modalities are words.

- Nominal: No inherent order (e.g., gender, colors).

  Example 4 *Coat colors: the set of modalities is: black, brown, white,...*

- Ordinal: Ordered categories (e.g., education levels).

  Example 5 *Degree of satisfaction with one's standard of living: the set of modalities is: very satisfied, satisfied, dissatisfied.*



4 Data Representation

Definition 5 (Statistical Series) *The statistical series is a correspondence that associates each individual in the studied population with a value of the characteristic being analyzed. The values of a statistical series for a characteristic X are denoted as: $x_1, x_2, x_3, ..., x_n$.*

4.1 Representation of Statistical Variables

In a population, we consider a sample of n individuals on which a variable X is observed.

- If X is quantitative discrete, we refer to the values x_i of the variable X.
- If X is qualitative nominal or ordinal, we refer to the modalities x_i of the variable X.
- If X is quantitative continuous, we refer to the classes c_i of the variable X.

Cumulative Frequency n: represents the total number of observations. It is calculated by successively adding the frequencies of the values in the dataset.

$$n = n_1 + n_2 + n_3 + \cdots + n_i.$$

Absolute frequency n_i: is the number of statistical elements relating to a given modality.

Relative frequency f_i: for each value x_i, we define

$$f_i = \frac{n_i}{n}.$$

f_i as the partial frequency of x_i. The frequency of a value is the ratio of the count of that value to the total count. We have,

$$\sum_{i=1}^{n} f_i = 1.$$

Cumulative absolute frequency $n_i^c \uparrow$: the number of individuals which correspond to the same modality and to the previous modality.

Cumulative relative frequency $n_i^c \uparrow$: the radio $\frac{n_i^c \uparrow}{n}$.

Example 6 *The notes of 9 students in a groupe are given in the following table:*

Notes	n_i	$n_i^c \uparrow$	f_i	$f_i^c \uparrow$
5	*2*	*2*	*2/9*	*2/9*
6	*1*	*3*	*1/9*	*1/3*
8	*3*	*6*	*1/3*	*2/3*
12	*2*	*8*	*2/9*	*8/9*
16	*1*	*9*	*1/9*	*1*
Total	$n = 9$		$\sum_{i=1}^{5} f_i = 1$	

- *The pupulation is: 9 students.*
- *Individual: one student.*
- *Characteristic: the notes.*
- *Modalities: 5, 6, 8, 12, 16.*

The case of interval (class): it is a groupe of values of a variable accordings to intervals which are equal. It is used when the variable studiesd is quantitative continuous.
For a classe define by: $[a, b]$ we have

- A lower limit: a

- An upper limit: b
- Amplitude: upper limit-lower limit $= b - a$
- Class center $c_i = \dfrac{upper\ limit + lower\ limit}{2} = \dfrac{b+a}{2}$.

Example 7 *The blood glucose level in 14 subjects in g/l is given as follows:*

class	c_i	n_i	$n_i^c \uparrow$	f_i	$f_i^c \uparrow$
[0,85 ; 0,91[	0,88	3	3	3/14	3/14
[0,91 ; 0,97[	0,94	5	8	5/14	4/7
[0,97 ; 1,03[	1	3	11	3/14	11/14
[1,03 ; 1,09[	1,06	2	13	1/7	13/14
[1,09 ; 1,15[	1,12	1	14	1/14	1
Total		n=14			$\sum_{i=1}^{5} f_i = 1$

- *The pupulation is: 14 subjects.*
- *Individual: one subject.*
- *Characteristic: the blood glucose level.*

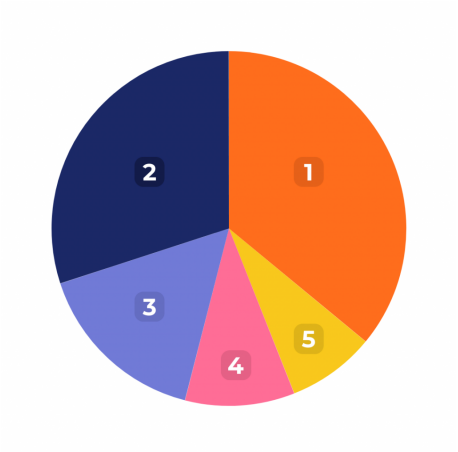
4.2 Data Description

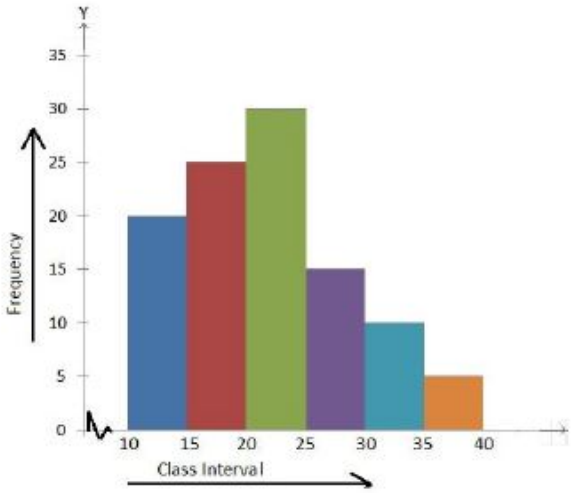
4.2.1 Tables

4.2.2 Graphics

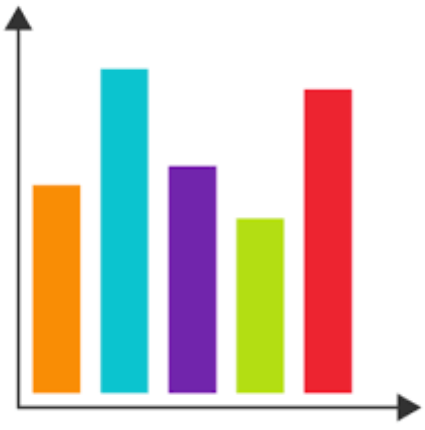
Circle graph: We draw on a disk sections corresponding to the modalities of the character (variable) whose angles are proportional to the porcentages.

$$\theta_i = \frac{360}{100}\frac{100 n_i}{n} = 360\frac{n_i}{n}.$$





Histogram



Bar graphs

5 Position parameters

5.1 Mean

1. **Case of discret statistical variable:**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i.$$

2. **Case of continuous statistical variable:**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i.$$

5.2 Mode

1. **Case of discret statistical variable:** The mode of a statistical variable is the value with the highest partial count (or the highest partial frequency), and it is denoted by Mo. For example, in Example 6, the mode is $Mo = 8$.

2. **Case of continuous statistical variable:**

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2}\right) a$$

- L_i: the lower limit of the modal class (the class that has the highest frequency)
- d_1 = the absolute frequency of the modal class- the absolute frequency of the previous class $(n_i - n_{i-1})$.
- d_2 =the absolute frequency of the modal class- the absolute frequency of the next class $(n_i - n_{i+1})$.
- a: the amplitude of the modal class.

In Example 7:

- The modale classe is $[0.91; 0.97[$.
- $L_i = 0.91$.
- $d_1 = 5 - 3 = 2$.
- $d_2 = 5 - 3 = 2$.
- $a = 0.97 - 0.91 = 0.06$. Then $Mo = 0.91 + \frac{2}{4}0.06 = 0.94$.

5.3 Median

1. **Case of discret statistical variable:** The median Me is the value at the center of a series of numbers arranged in ascending order.

   - if n is even, then
   $$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}.$$
   - if n is odd, then
   $$Me = x_{\frac{n+1}{2}}.$$

   In Example 6: $n = 9$ then, $Me = x_5 = 8$.

2. **Case of continuous statistical variable:**

   In this case the median is given by

   $$Me = L_i + \left(\frac{\frac{n}{2} - \sum_{i=1}^{<Me} n_i}{n_{Me}} \right) a$$

   - L_i: the lower limit of the median class
   - $\sum_{i=1}^{<Me} n_i$ = the sum of the absolute frequencies corresponding to all classes below the median class.
   - n_{Me}= the absolute frequency of the median class.
   - a: the amplitude of the median class.

   In Example 7:

   - The median classe is $[0.91; 0.97[$.
   - $L_i = 0.91$.
   - $n = 14$.
   - $\sum_{i=1}^{<Me} n_i = 3$.
   - $n_{Me} = 5$.
   - $a = 0.97 - 0.91 = 0.06$. Then, $Me = 0.91 + (\frac{7-3}{5})0.06 = 0.958$.

Dr. Kicha Abir