Chapter 1

Univariate Descriptive Statistics

Contents

Descriptive statistics is the set of scientific methods for collecting, describing, and analyzing observed data.

1.1 Statistical Vocabulary

- **Population:** The set of individuals or objects of the same nature on which the study is conducted.
- **Individuals:** Individuals or statistical units are the elements of the population.
- **Sample:** A subset of the population.
- **Statistical Variable:** A characteristic is the property we intend to observe in the population or sample. A characteristic that is the subject of a study is also called a statistical variable X.
- **Statistical Modality:** A modality (or category) refers to the different possible situations (levels) of a statistical variable.

Two types of statistical variables are distinguished:

- **Quantitative Variables:** These are variables that can be measured and are characterized by numerical values. Variables whose modalities are numbers.
- A quantitative statistical variable can be:
  - **Continuous:** When it can take values from a real-number interval (measurement results).
  - **Discrete:** When it takes isolated values.
  - **Temporal:** These are particular quantitative variables that use units of time measurement. There are two types: date (e.g., birth date: 26/04/1994) and time (e.g., study hours: 6h).

Examples 1.1.1.

Variable	*Possible Modalities*	*Type of Variable*
Height	$1.70m, 1.60m, 1.65m, 1.75m$	*Quantitative, Continuous*
Number of Students	$30, 50, 60, 80$	*Quantitative, Discrete*

- **Qualitative Variables:** These are variables that cannot be measured (they do not have numerical values). Variables whose modalities are words.
- Qualitative statistical variables can be:
  - **Ordinal:** These are variables whose modalities can be ordered by their meaning.
  - **Nominal:** These are variables whose modalities cannot be ordered by their meaning.

Example 1.1.1.

Variable	*Possible Modalities*	*Type of Variable*
Eye Color	*Black, Blue, Green, Brown*	*Qualitative, Nominal*
Satisfaction with Life	*Very Satisfied, Satisfied, Unsatisfied*	*Qualitative, Ordinal*

1. **Statistical Series**: The simplest form of presenting statistical data related to a single characteristic or variable is a simple enumeration of the values taken by the characteristic.
2. **Total Effectif**: The total number of individuals in the population, denoted as n.
3. **Effectif**: The effective or absolute frequency, denoted as n_i, is the number of elements corresponding to a given modality.
4. **Cumulative Increasing Effectif**: The cumulative effective , denoted as $n_i^c \uparrow$, is the number of individuals corresponding to the same characteristic (modality) and the previous characteristics.
5. **Cumulative Decreasing Effectif**: The cumulative frequency, denoted as $n_i^c \downarrow$, is the number of individuals corresponding to the same characteristic (modality) and the following characteristics.
6. **frequency** : denoted as f_i, is the ratio between the absolute frequency of a value and the total number of observations:

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

7. **Cumulative Increasing Frequency**: denoted as $f_i^c \uparrow$, is the ratio between the cumulative increasing absolute frequency of a value and the total number of observations:

$$f_i^c \uparrow = \frac{n_i^c \uparrow}{n}$$

8. **Cumulative Decreasing Frequency** denoted as $f_i^c \downarrow$, is the ratio between the cumulative decreasing absolute frequency of a value and the total number of observations:

$$f_i^c \downarrow = \frac{n_i^c \downarrow}{n}$$

Example 1.1.2. *Grades of 9 students in a group*

Grade	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	f_i	$f_i^c \uparrow$	$f_i^c \downarrow$	
5	2	2	9	$\frac{2}{9}$	$\frac{2}{9}$	1	
6	1	3	7	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{7}{9}$	
8	3	6	6	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{6}{9}$	
12	2	8	3	$\frac{2}{9}$	$\frac{8}{9}$	$\frac{3}{9}$	
16	1	9	1	$\frac{1}{9}$	1	$\frac{1}{9}$	
Total	$n = 9$			$\sum_{i=1}^{5} f_i = 1$			

14. **Class (Interval)**: A class is a grouping of values of a variable into intervals that may be equal or unequal. It is typically used when the variable studied is a continuous quantitative variable.

For each class, the following can be defined:

- A lower limit
- An upper limit
- Class interval (range) = upper limit - lower limit
- Class center $c_i = \frac{\text{upper limit+lower limit}}{2}$

Example 1.1.3. *Blood glucose levels (in g/l) of 14 subjects*

Class	c_i	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	f_i	$f_i^c \uparrow$	$f_i^c \downarrow$
[0.85; 0.91[	0.88	3	3	14	$\frac{3}{14}$	$\frac{3}{14}$	1
[0.91; 0.97[	0.94	5	8	11	$\frac{5}{14}$	$\frac{4}{7}$	$\frac{11}{14}$
[0.97; 1.03[	1.00	3	11	6	$\frac{3}{14}$	$\frac{11}{14}$	$\frac{6}{14}$
[1.03; 1.09[	1.06	2	13	3	$\frac{2}{14}$	$\frac{13}{14}$	$\frac{3}{14}$
[1.09; 1.15[	1.12	1	14	1	$\frac{1}{14}$	1	$\frac{1}{14}$
Total		$n = 14$			$\sum_{i=1}^{5} f_i = 1$		

1.2 Data Description

According to the type of variable being studied, there are two main ways to present and describe a series of statistical data: tables and graphical representations.

1.2.1 Tables

A table can be used regardless of the nature of the data, and it serves to present data in a precise and complete way.

1.2.2 Graphs

The objective of graphs is to provide a systematic view of the phenomenon being studied by illustrating a general trend and giving an overall image of the results.

Histogram: Histograms are surfaces used to represent continuous quantitative variables. The area of each surface is equal to the frequency corresponding to a class.

Bar Chart: A bar chart is a graphical representation of statistical data using segments.

Examples 1.2.1.

Diameter	12	13	14	15	16	17	18
Frequency	2	5	3	4	6	5	3

Bar Diagram: A bar diagram is a graphical representation primarily used for the distribution of a qualitative variable using rectangles of equal width.

Examples 1.2.2.

Marital Status	*Single*	*Divorced*	*Married*	*Widowed*
Frequency	9	2	7	2

Pie Chart: A pie chart displays sections on a disk corresponding to the modalities of the characteristic, where the angles are proportional to the percentages.

$$\alpha_i = \frac{360^\circ \times f_i}{n}$$

Examples 1.2.3.

Language	*Number of Students*	f_i	α_i
English	500	0.5	180°
French	200	0.2	72°
German	150	0.15	54°
Italian	100	0.1	36°
Other	50	0.05	18°

1.3 Position Parameters

Position parameters: values located at the center of the statistical distribution, which include the mean, mode, median and Quartiles.

1.3.1 Mean

For a discrete statistical variable:

Let X be a discrete statistical variable, and let $x_1, x_2, \ldots, x_k$ be its values with corresponding frequencies $n_1, n_2, \ldots, n_k$, where $n = \sum_{i=1}^{k} n_i$ is the total frequency.

The mean of X is given by:

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

Examples 1.3.1.

x_i	0	1	2	3	4
n_i	2	3	1	1	1

The mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i = \frac{1}{8}(0\times 2 + 1\times 3 + 2\times 1 + 3\times 1 + 4\times 1) = \frac{12}{8} = 1.5.$$

For a continuous statistical variable:

If the observations are grouped into classes, the mean is given by:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i,$$

where c_i is the class midpoint.

Examples 1.3.2.

Class:	c_i	n_i
[1.2, 1.3[	1.5	3
[2.0, 2.3[	2.5	1
[3.3, 3.4[	3.5	2

The mean is calculated as:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{3} n_i c_i = \frac{1}{6}(3\times 1.5 + 1\times 2.5 + 2\times 3.5) = \frac{14}{6} = 2.33.$$

1.3.2 Mode

For a discrete statistical variable:

The mode *Mo* is the value x_i with the highest frequency.

Examples 1.3.3.

x_i	2	3	5	6	7	8	9	10
n_i	2	1	1	2	2	1	1	1

There are three modes: $Mo = 2, 6, 7$.

For a continuous statistical variable:

The mode is calculated using the formula:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \times a$$

where:

- L_i is the lower boundary of the modal class (the class with the highest frequency),

- $d_1 = n_i - n_{i-1}$,
- $d_2 = n_i - n_{i+1}$,
- a is the amplitude of the modal class.

Example 1.3.1.

C_i	n_i
$[1.60, 1.65[$	3
$[1.65, 1.70$	8
$[1.70, 1.75[$	2

The modal class is $[1.65, 1.70[$*, so:*

$$L_i = 1.65, \quad d_1 = 8 - 3 = 5, \quad d_2 = 8 - 2 = 6, \quad a = 1.70 - 1.65 = 0.05$$

Thus,

$$Mo = 1.65 + \frac{5}{5+6} \times 0.05 = 1.67.$$

1.3.3 Median

For a discrete statistical variable:

The median *Me* is the value that lies in the center of an ordered list of numbers.

- If n is even, then:

$$Me = \frac{x_{n/2} + x_{n/2+1}}{2}$$

- If n is odd, then:

$$Me = x_{(n+1)/2}$$

Example 1.12:

The number of children in 6 families:

$$7, 3, 1, 1, 5, 2$$

Ordered values:

$$1, 1, 2, 3, 5, 7$$

Since $n = 6$ is even,

$$Me = \frac{x_3 + x_4}{2} = \frac{2+3}{2} = 2.5.$$

Example 1.3.2. *The number of children in 7 families:*

$$3, 2, 1, 0, 0, 1, 2$$

Ordered values:

$$0, 0, 1, 1, 2, 2, 3$$

Since $n = 7$ is odd,

$$Me = x_4 = 1.$$

For a continuous statistical variable:

The median is given by:

$$Me = L_i + \frac{n/2 - \sum_{i=1}^{Me} n_i}{n_{Me}} \times a$$

where:

- L_i is the lower boundary of the median class,
- $\sum_{i=1}^{Me} n_i$ is the cumulative frequency of all classes before the median class,
- n_{Me} is the frequency of the median class,
- a is the amplitude of the median class.

Example 1.3.3. *From* (1.1.3)*, we get:*

- *The median class is* $[0.91, 0.97[$,
- $L_i = 0.91$,
- $n = 14$,
- $\sum_{i=1}^{Me} n_i = 3$,
- $n_{Me} = 5$,
- $a = 0.97 - 0.91 = 0.06$.

Thus,

$$Me = 0.91 + \frac{7 - 3}{5} \times 0.06 = 0.958.$$

1.3.4 Quartiles

For a discrete statistical variable:

The quartiles are the three values that divide the distribution into four equal parts. They are:

- The first quartile Q_1 represents 25% of the sample, i.e., Q_1 is the value at the position of the smallest integer greater than or equal to $n/4$,
- The second quartile Q_2 represents 50% of the sample,
- The third quartile Q_3 represents 75% of the sample, i.e., Q_3 is the value at the position of the smallest integer greater than or equal to $3n/4$.

The **interquartile range** (IQR) is the difference between the third and first quartiles:

$$IQR = Q_3 - Q_1$$

Example 1.3.4. *Given the observations:*

x_i	1	3	5	7	9
n_i	1	2	1	2	2

$$n = 8, \quad n/4 = 2, \quad Q_1 = x_2 = 3$$

$$3n/4 = 6, \quad Q_3 = x_6 = 7$$

Thus, $Q_1 = 3$ *and* $Q_3 = 7$.

1.3.5 Dispersion Parameters

Dispersion parameters summarize the spread of values around the central value.

1.3.6 Range

The range e is the difference between the largest and smallest observed values:

$$e = x_{\max} - x_{\min}$$

Example 1.16:

The grades of 10 students:

$$2, 3, 10, 10, 11, 12, 15, 18, 19, 20$$

Thus,

$$e = x_{\max} - x_{\min} = 20 - 2 = 18.$$

1.3.7 Variance

The variance $V(X)$ is the arithmetic mean of the squared deviations between each value and the mean:

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2$$

1.3.8 Standard Deviation

The standard deviation σ_X is the square root of the variance:

$$\sigma_X = \sqrt{V(X)}$$

1.3.9 Coefficient of Variation

The coefficient of variation CV is defined as:

$$CV = \frac{\sigma_X}{\overline{x}}$$

Example 1.3.5. *For the data:*

x_i	0	1	2	3	4
n_i	2	3	1	1	1

$$\overline{x} = 1.5, \quad V(X) = 1.75, \quad \sigma_X = 1.3, \quad CV = \frac{1.3}{1.5} = 0.87$$

1.3.10 Shape Parameters

Skewness

There are several coefficients of skewness, the main ones are:

- **Pearson's skewness coefficient**:

$$AP = \frac{\overline{x} - Mo}{\sigma_X}$$

- **Yule's skewness coefficient**:

$$AY = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}.$$

Remark 1.3.1.
- *A positive coefficient indicates a distribution that is more spread out to the right.*
- *A negative coefficient indicates a distribution that is more spread out to the left.*
- *A zero coefficient indicates a symmetric distribution.*

Kurtosis

Kurtosis is measured by:

- **Pearson's kurtosis coefficient**:

$$APP = \frac{m_4}{\sigma_X^4}$$

where m_4 is the fourth central moment.

- **Fisher's kurtosis coefficient**:

$$APF = \frac{m_4}{\sigma_X^4} - 3$$

Remark 1.3.2.
- *If $APF = 0$, the distribution is said to be "normal" or "mesokurtic".*
- *If $APF < 0$, the distribution is said to be flatter than normal or "platykurtic".*
- *If $APF > 0$, the distribution is said to be more peaked than normal or "leptokurtic".*