Chapter

Univariate Descriptive Statistics

Contents

Descriptive statistics is the set of scientific methods that allow the collection, description, and analysis of observed data.

1.1 Statistical vocabulary

(1) Population: is the set of individuals or objects of the same nature on which the study is conducted.
(2) Individuals: Individuals or statistical units are the elements of the population.
(3) Sample: is a subset of the population.
(4) Statistical variable: the characteristic is the property that one intends to observe in the population or the sample. A characteristic that is the subject of a study is also called a statistical variable X.
(5) 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; they are characterized by numerical values. Variables whose modalities are numbers.

A quantitative statistical variable can be:
Continuous: when it can take values from an interval of real numbers (measurement results).
Discrete: when it takes isolated values.
Temporal: These are particular quantitative variables that use units of time. There are two types: date type (date of birth: 26/04/1994) and time type (study hours: 6h).

Example 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 (do not have numerical values). Variables whose modalities are words.

Qualitative statistical variables can be:
Ordinal: variables whose modalities can be ordered according to their meaning.
Nominal: variables whose modalities cannot be ordered according to their meaning.

Example 1.2.

variable	possible modalities	type of variable
eye color	black, blue, green, brown	qualitative nominal
degree of satisfaction with standard of living	very satisfied, satisfied, dissatisfied	qualitative ordinal

(6) Statistical series: The simplest form of presenting statistical data related to a single characteristic or variable consists of a simple listing of the values taken by the characteristic.
(7) Total frequency: The total frequency n is the total number of individuals in the population.
(8) Frequency: the frequency or absolute frequency denoted n_i is the number of statistical elements corresponding to a given modality.
(9) Cumulative increasing frequency: The cumulative increasing frequency denoted $n_i^c \uparrow$ is the number of individuals corresponding to the same modality and the previous modalities.
(11) Cumulative decreasing frequency: The cumulative decreasing frequency denoted $n_i^c \downarrow$ is the number of individuals corresponding to the same modality and the subsequent modalities.
(1) Relative frequency: the relative frequency denoted f_i is the ratio between the frequency of a value and the total frequency $\frac{n_i}{n}$.

(2) Cumulative increasing relative frequency: denoted $f_i^c \uparrow$, is the ratio $\frac{n_i^c \uparrow}{n}$.
(3) Cumulative decreasing relative frequency: denoted $f_i^c \downarrow$, is the ratio $\frac{n_i^c \downarrow}{n}$.
Example 1.3. The grades of 9 students in a group

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

(4) Class (Interval): A class is a grouping of values of a variable into intervals that may be equal or unequal. It is mainly used when the studied variable is quantitative continuous.
For each class, one can define:

- A lower bound
- An upper bound
- Class interval (amplitude) = upper bound – lower bound
- Class center $c_i = \frac{\text{upper bound} + \text{lower bound}}{2}$.

Example 1.4.: Blood glucose level (glycemia) of 14 subjects in g/l

class	Class center c_i	n_i	$n_i^c \uparrow$	$n_i^c \downarrow$	Frequency f_i	$f_i^c \uparrow$	$f_i^c \downarrow$
[0,85; 0,91[	0.88	3	3	14	3/14	3/14	1
[0,91; 0,97[	0.94	5	8	11	5/14	4/7	11/14
[0,97; 1,03[	1	3	11	6	3/14	11/14	6/14
[1,03; 1,09[	1.06	2	13	3	1/7	13/14	3/14
[1,09; 1,15[	1.12	1	14	1	1/14	1	1/14
Total		$n = 14$			$\sum_{i=1}^{5} f_i = 1$		

1.2 Data description

According to the type of the studied variable, there are two forms of presentation used to describe a statistical data series: tables and graphical representations.

1.2.1 Tables

A table can be used regardless of the nature of the data; it is used to present data in an accurate and complete manner.

1.2.2 Graphs

The purpose of graphs is to highlight a systematic view of the studied phenomenon by illustrating a general trend and providing an overall picture of the results.

Histogram

Histograms are surfaces that allow the representation of a quantitative continuous variable. 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.

Example 1.5.

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 mainly used for the distribution of a qualitative variable using rectangles of equal width.

Example 1.6.

Marital status	Single	Divorced	Married	Widowed
Frequency	9	2	7	2

Pie chart

Sections corresponding to the modalities of the characteristic are drawn on a disk, with angles proportional to the percentages.

$$\alpha_i = 360^0 * f_i = 360^0 * \frac{n_i}{n}$$

Example 1.7.

Language	Number of students	f_i	α_i
English	500	0.5	180^0
French	200	0.2	72^0
German	150	0.15	54^0
Italian	100	0.1	36^0
Other	50	0.05	18^0

1.3 Measures of position

Measures of central tendency or position: values located at the center of the statistical distribution, namely the mean, the mode, and the median.

1.3.1 Arithmetic mean

Case of a discrete statistical variable

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

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

Example 1.8.

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

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

Case of a continuous statistical variable

The observations are grouped into classes, then

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

Example 1.9.

class	c_i	n_i
$[1, 2[$	1.5	3
$[2, 3[$	2.5	1
$[3, 4[$	3.5	2

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

1.3.2 Mode

Case of a discrete statistical variable

The mode Mo is the value x_i having the largest frequency.

Example 1.10.

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

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

Case of a continuous statistical variable

In this case, the mode is calculated using the formula

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

- L_i: the lower bound of the modal class (class corresponding to the largest frequency)
- d_1 = frequency of the modal class minus the frequency of the previous class $(n_i - n_{i-1})$.
- d_2 = frequency of the modal class minus the frequency of the following class $(n_i - n_{i+1})$.
- a: the amplitude of the modal class.

Example 1.11.

class	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[.
- $L_i = 1,65$.
- $d_1 = 8 - 3 = 5$.
- $d_2 = 8 - 2 = 6$.
- $a = 1,70 - 1,65 = 0.05$ therefore $Mo = 1,65 + \left(\frac{5}{5+6}\right) 0.05 = 1,67$

1.3.3 Median

Case of a discrete statistical variable

The median Me is the value located at the center of a series of numbers arranged in increasing order.

- If n is even, then

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

- If n is odd, then

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

Example 1.12. The number of children in 6 families is as follows

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

First, the values are ordered:

$$\underbrace{1, 1, 2}_{3}, \underbrace{3, 5, 7}_{3}$$

We have $n = 6$ (even), therefore $Me = \frac{x_3 + x_4}{2} = \frac{2+3}{2} = 2.5$.
Example 1.13. The number of children in 7 families is as follows

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

First, the values are ordered:

$$\underbrace{0,0,1}_{3},\ \underbrace{1}_{Me=x_4=1},\ \underbrace{,2,2,3}_{3}$$

We have $n = 7$ (odd), therefore $Me = x_4 = 1$.

Case of a 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 bound of the median class (the class that divides the total frequency into two equal parts)
- $\sum_{i=1}^{<Me} n_i$: the sum of frequencies corresponding to all classes preceding the median class
- n_{Me}: the frequency of the median class
- a: the amplitude of the median class

Example 1.14. According to example (1.4), we obtain

- The median class is: [0.91-0.97[.
- $L_i = 0.91$.
- $n = 14$.
- $\sum_{i=1}^{<Me} n_i = 3$
- $n_{Me} = 5$.
- $a = 0.97 - 0.91 = 0.06$
  therefore $Me = 0.91 + \left(\frac{7-3}{5}\right) 0.06 = 0.958$

1.3.4 Quartiles

Case of a discrete statistical variable

Quartiles are the three values that divide the distribution into four equal parts. They are respectively called:

- The first quartile $\mathbf{Q_1}$ represents 25% of the sample, i.e., Q_1 is the value x_i whose position is the smallest integer greater than $\frac{n}{4}$.
- The second quartile $\mathbf{Q_2}$ represents 50% of the sample.
- The third quartile $\mathbf{Q_3}$ represents 75% of the sample, i.e., Q_3 is the value x_i whose position is the smallest integer greater than $\frac{3n}{4}$.

Interquartile range

The interquartile range is the difference between the third and the first quartile:

$$I_Q = Q_3 - Q_1$$

Example 1.15. For the following observations

x_i	1	3	5	7	9
n_i	1	2	1	2	2
n_i^c	1	3	4	6	8

- We have $n = 8$ and $\frac{n}{4} = 2$, therefore Q_1 is the second value: $Q_1 = x_2 = 3$.
- We have $n = 8$ and $\frac{3n}{4} = 6$, therefore Q_3 is the sixth value: $Q_3 = x_6 = 7$.

1.4 Measures of dispersion

Measures of dispersion summarize the spread of values around the central value.

1.4.1 Range

The range e is defined as the difference between the largest and the smallest observed value.

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

Example 1.16. The grades of 10 students are as follows

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

thus

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

1.4.2 Variance

The variance is defined as the arithmetic mean of the squares of the deviations between the values of a variable and the arithmetic mean.

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$$
$$= \sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2.$$

1.4.3 Standard deviation

The standard deviation denoted σ_X (or root mean square deviation) is the square root of the variance.

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

1.4.4 Coefficient of variation

The coefficient of variation denoted CV is defined by

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

Example 1.17.

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

$$\bar{x} = 1.5$$

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \\ &= \frac{1}{8}\sum_{i=1}^{5} n_i x_i^2 - (1.5)^2 \\ &= \frac{1}{8}\left(2\times 0^2 + 3\times 1^2 + 1\times 2^2 + 1\times 3^2 + 1\times 4^2\right) - 2.25 \\ &= \frac{32}{8} - 2.25 \\ &= 1.75 \end{aligned}$$

The standard deviation

$$\sigma_X = \sqrt{V(X)} = \sqrt{1.75} = 1.3$$

and the coefficient of variation

$$CV = \frac{\sigma_X}{\bar{x}} = \frac{1.3}{1.5} = 0.87$$

1.5 Shape parameters

1.5.1 Skewness

There are several skewness coefficients; the main ones are:

- Pearson's skewness coefficient:

$$A_P = \frac{\bar{x} - Mo}{\sigma_X}$$

- Yule's skewness coefficient:

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

Remark

- A positive coefficient indicates that the distribution is more spread to the right.
- A negative coefficient indicates that the distribution is more spread to the left.
- A zero coefficient indicates that the distribution is symmetric.

1.5.2 Kurtosis

Kurtosis is measured by:

- Pearson's kurtosis coefficient:

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

where m_4 is the fourth-order central moment defined by

$$m_4 = \frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^4$$

- Fisher's kurtosis coefficient:

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

Remark

- If $AP_F = 0$, the distribution is said to be "normal" or "mesokurtic".
- If $AP_F < 0$, the distribution is said to be flatter than the "normal" distribution or "platykurtic".
- If $AP_F > 0$, the distribution is said to be less flat than the "normal" distribution or "leptokurtic".