

ELEMENTARY STATISTICS

A.Y. - 2026/2027

Teacher:

Maurizio Serva (serva@univaq.it)

<https://people.disim.univaq.it/~serva/teaching/teaching.html>
(no copy and paste)

Reference texts:

Barbara Illosky et al. *Introductory Statistics*, OpenStax, Rice University. Downloadable from (no copy and paste)
https://people.disim.univaq.it/~serva/teaching/intr_stat.pdf

Maria Garetto, *Statistica*, Quaderni didattici dell'Università di Torino. Downloadable from (no copy and paste)
https://people.disim.univaq.it/~serva/teaching/quad_stat.pdf

Goal of Statistics: to collect data from experiments or investigations, organize and describe them and make predictions.

- **Descriptive Statistics:** to collect data, organize and describe them.
- **Inferential Statistics:** to draw conclusions and make predictions.

To draw meaningful conclusions from data, it is often necessary to make assumptions about the probabilities of obtaining each value. The set of these assumptions is called **Probabilistic Model**. Therefore, **Probability Theory is necessary**.

POPULATION AND SAMPLES

In statistics, we are interested in obtaining information on a whole set of elements that we call **Population**.

Example: There are 47,800 residents in the municipality of Rieti, so the population is made up of 47,800 individuals (elements).

Problem: How many residents are in favor of building a subway? The population is too large to interview everyone.

Strategy: We decide to interview 1,000 residents to obtain partial information.

Let's define **Sample**: the subset of the population being surveyed.

The notion of population must be understood in a broad sense.

For example, suppose we want to measure the concentration of particulate matter (PM10) in the streets and squares of the municipality of L'Aquila at a certain time.

In this case, the population consists of all the streets and squares within the municipality.

Choosing a sample means selecting some streets and squares in which to place detectors.

Keep in mind that the **sample** can coincide with the entire **population**.

The sample should be chosen so that it is **representative**.

Bad choices:

- residents of the city center
- residents of the same neighborhood
- etc...

It is also incorrect to intentionally choose the sample so that it contains all types of individuals: the same percentage of men and women, all types of workers (bricklayers, civil servants, office workers, etc.), residents of every neighborhood, etc.

The correct methodology is choosing a **random sample**.

Example: Draw 1,000 balls from an urn containing 47,800.

The methods for selecting a representative sample are a broad field of research.

DESCRIPTIVE STATISTICS

Goal: to present concisely through tables, graphs, means, variances, or other forms of summary, the results of a research study, survey, or poll... so that the essential characteristics of the collected data can be easily perceived.

Data can refer to one or more measured quantities called **variables**.

Numerical variables (quantitative) can be discrete or continuous.

Categorical variables (qualitative) all others.

EXAMPLE OF A DISCRETE NUMERICAL VARIABLE

The owner of a pizzeria records the days his 10 employees were absent from work due to illness over the last 6 weeks.

Antonio 2, Cinzia 2, Dario 1, Elena 4, Francesco 10, Rita 0, Renato 5, Pasquale 8, Anna 0, Maria 0

This is an example of a discrete numeric variable (days of sick leave) recorded in 10 observations.

The data relating to a numeric variable are therefore presented as a sequence $x_1, x_2, \dots, x_n$ of values of the variable obtained in n observations.

In our example $n = 10$ and $x_1 = 2, x_2 = 2, x_3 = 1, \dots, x_{10} = 0$.

EXAMPLE OF A CONTINUOUS NUMERICAL VARIABLE

The maximum temperature (in degrees Celsius) recorded in Rome on August 1st for 10 years yielded the following results:

30.1, 28.2, 31.3, 22.6, 27.5, 29.4, 24.5, 27.8, 33.2, 22.8

This is an example of continuous numerical variable: the values are real numbers.

In summary: Let us denote the n observed values of a numerical variable by

$$x_1, \quad x_2, \quad x_3, \quad \cdots, \quad x_n$$

The **variable is discrete** if the x_i can only assume some **real** values and not intermediate ones: for example, only **integer** values.

The **variable is continuous** if the x_i can assume with continuity **real** values.

When data are collected from a population or a sample, the resulting values appear as a collection of disordered values called **raw data**.

Raw data do not provide a **readable** information: they must be ordered and organized to highlight their characteristics. Tables and graphs are created for this purpose.

DESCRIPTIVE STATISTICS: DISCRETE VARIABLES

Let's take the example of the pizzeria again:

2 2 1 4 10 0 5 8 0 0

The data are integers and are "few," ranging from 0 to 10. We represent them in a **frequency table** that places each distinct value next to its respective frequency of occurrence.

Absolute frequency = number of occurrences of a given value.

It is best to sort the data in ascending order:

0 0 0 1 2 2 4 5 8 10

0, 0, 0, 1, 2, 2, 4, 5, 8, 10

Value	Absolute frequency
0	3
1	1
2	2
3	0
4	1
5	1
6	0
7	0
8	1
9	0
10	1
Total	10

Total = Number of observations = 10

$$\text{Relative frequency} = \frac{\text{Absolute frequency}}{\text{Number of observations}}$$

Value	Absolute frequency	Relative frequency
0	3	3/10
1	1	1/10
2	2	1/5
3	0	0
4	1	1/10
5	1	1/10
6	0	0
7	0	0
8	1	1/10
9	0	0
10	1	1/10
Total	10	1

Percentage frequency = Relative frequency $\times$ 100

Value	Absolute freq.	Relative freq.	Percentage freq.
0	3	3/10	30
1	1	1/10	10
2	2	1/5	20
3	0	0	0
4	1	1/10	10
5	1	1/10	10
6	0	0	0
7	0	0	0
8	1	1/10	10
9	0	0	0
10	1	1/10	10
Total	10	1	100

A formalization step (discrete numerical variables):

n = number of observations k = number of distinct values

In our example $n = 10$ (the number of observation *i.e.*, the number of collected data) and $k = 11$ (the number of possible values ranging from 0 to 10, both included).

Absolute frequencies: $n_1, n_2, n_3, \dots, n_k$

$$n_1 + n_2 + \dots + n_k = n$$

Relative frequencies: $f_i = \frac{n_i}{n}, \quad i = 1, 2, 3, \dots, k$

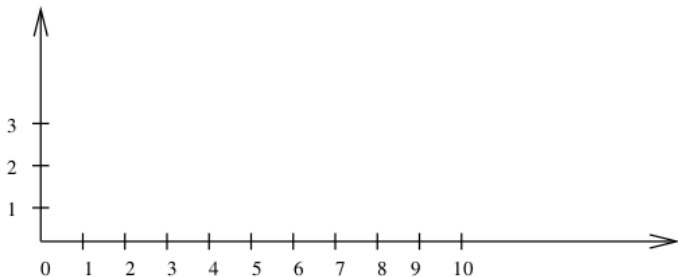
$$f_1 + f_2 + \dots + f_k = 1$$

Percentage frequencies: $F_i = f_i \times 100, \quad i = 1, 2, 3, \dots, k$

$$F_1 + F_2 + \dots + F_k = 100$$

FREQUENCY DISTRIBUTION GRAPHS

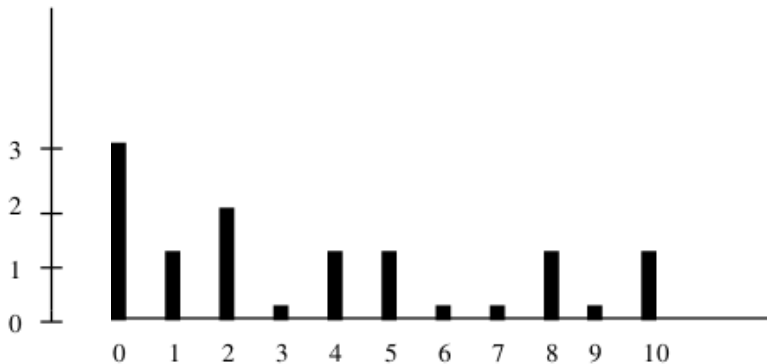
Let's go back to the example we've been considering so far with $n = 10$ and $k = 11$. The absolute (or relative) frequencies of each value should be arranged on a Cartesian plane of the type:



where the x-axis shows the number i of sick days and the y-axis shows the absolute frequencies n_i .

BAR GRAPH

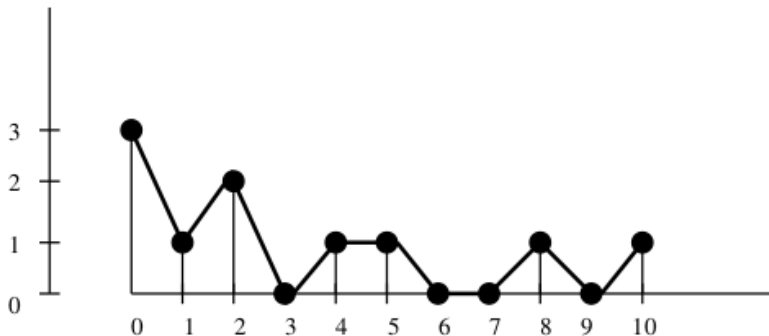
The most common representation is the bar graph shown below.



where, as above, the x-axis shows the number of sick days and the y-axis shows the absolute frequencies.

OTHER GRAPHS

Other graph representations can be drawn as, for example, the polygonal graph below:



DESCRIPTIVE STATISTICS: CONTINUOUS VARIABLES

Continuous variables and discrete variables that take many values must be placed in disjoint intervals called **classes**.

Example. Results of 12 daily emission measurements of a polluting gas from an industrial plant in ascending order.

9.0 11.2 13.2 13.2 15.8 17.3
18.7 23.9 24.8 26.4 29.6 31.8

Let us consider the following intervals (classes) of width 7:

$[9, 16)$, $[16, 23)$ $[23, 30)$ $[30, 37)$

The values at the extremes of the class are called **edges** of the class. The square bracket indicates inclusion, the round bracket indicates exclusion.

The absolute and relative frequencies of data falling into each class can be organized in a table

Class	Absolute frequency	Relative frequency
[9, 16)	5	5/12
[16, 23)	2	1/6
[23, 30)	4	1/3
[30, 37)	1	1/12
Total	12	1

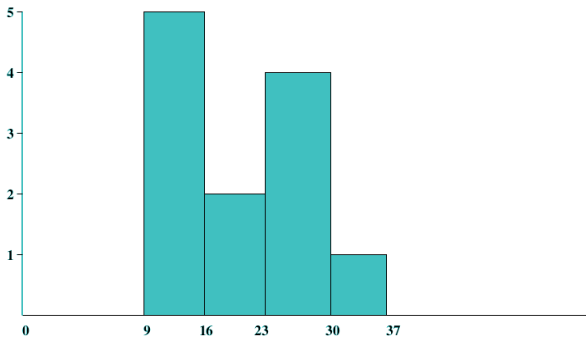
The number of classes considered is a subjective choice. Too many classes make the table difficult to read, while too few classes make it not very significant. Sometimes it's a good idea to experiment with different numbers of classes to determine which table or graph is the most informative. Usually, 5 to 10 classes are chosen.

HISTOGRAMS

The most used graph for grouped data is the **histogram**. This is a bar graph with columns arranged adjacent to each other. As before, the y-axis typically shows either absolute or relative frequencies.

A histogram consists of a set of adjacent rectangles with bases on the x-axis: the bases are the interval that define the classes and the midpoints of the intervals are the central values of the classes. If the classes all have the same amplitude, the pitches are equal to the corresponding absolute (or relative) frequencies.

Class	Absolute frequency
[9, 16)	5
[16, 23)	2
[23, 30)	4
[30, 37)	1
Total	12



DESCRIPTIVE STATISTICS: CATEGORICAL VARIABLES

The possible values of a categorical variable are called **categories**.

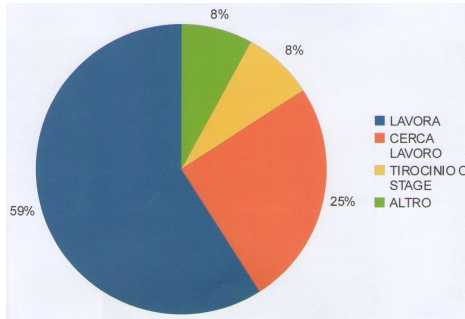
Esempio. Survey on the employment status of 215 graduates of the Faculty of Science 6 months after graduation.

Categories	Abs. freq.	Relative freq.	Percentage freq.
Works	127	0.5907	59.07
Looks for a work	54	0.2512	25.12
Training or stage	17	0.0791	7.91
Other	17	0.0791	7.91
Total	215	1	100

For this type of data the **pie graph** is preferably used. For any of the categories, indexed by i , the corresponding circular sector has an angle θ_i .

The angles θ_i , are obtained as follows:

$$\theta_i = 360 \times \text{Relative frequency} = 360 \times f_i$$



Where: blue = work, red = look for a work, yellow = training or stage, green = other.