

ELEMENTARY STATISTICS

A.Y. - 2025/2026

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.

Riassumendo: Indichiamo gli n valori osservati di una variabile numerica con

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

La **variabile è discreta** se le x_i **possono** assumere soltanto alcuni valori e non quelli intermedi: ad esempio soltanto valori **interi**.

La **variabile è continua** se le x_i **possono** assumere con continuità valori **reali**.

Quando si raccolgono i dati da una popolazione o da un campione i valori ottenuti si presentano come un insieme di valori disordinati che vengono chiamati **dati grezzi**.

I dati grezzi non forniscono una informazione *leggibile*: bisogna ordinarli ed organizzarli in modo da evidenziare le loro caratteristiche. A questo scopo si costruiscono tabelle e grafici.

STATISTICA DESCRITTIVA: VARIABILI DISCRETE

Riprendiamo l'esempio della pizzeria.

2 2 1 4 10 0 5 8 0 0

I dati sono numeri interi e sono “pochi” e vanno dal minimo 0 al massimo 10. Li rappresentiamo in una **tabella delle frequenze** che affianca ad ogni valore distinto la rispettiva frequenza di occorrenza.

Frequenza assoluta = numero di occorrenze di un dato valore.

Conviene ordinare i dati in ordine crescente

0 0 0 1 2 2 4 5 8 10

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

Valore	Frequenza assoluta
0	3
1	1
2	2
3	0
4	1
5	1
6	0
7	0
8	1
9	0
10	1
TOTALE	10

$$\text{Frequenza relativa} = \frac{\text{frequenza assoluta}}{\text{numero di osservazioni}}$$

Valore	Frequenza assoluta	Frequenza relativa
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
TOTALE	10	1

Frequenza percentuale = frequenza relativa \times 100

Valore	Freq. assoluta	Freq. relativa	Freq. percentuale
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
TOTALE	10	1	100

Un pò di formalizzazione (variabili numeriche **discrete**):

n = numero di osservazioni k valori distinti

$x_1 < x_2 < x_3 < \dots x_k$ valori distinti e messi in ordine crescente

Frequenze assolute: $n_1, n_2, n_3, \dots n_k$

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

Frequenze relative: $f_i = \frac{n_i}{n}, i = 1, 2, 3, \dots k$

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

Frequenze percentuali: $F_i = f_i \times 100, i = 1, 2, 3, \dots k$

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

GRAFICI DELLE DISTRIBUZIONI DI FREQUENZA

Torniamo all'esempio considerato finora.

$n =$ **numero di osservazioni** $= 10$

$k =$ **numero valori distinti** $= 11$

Le frequenze assolute (o quelle relative) di ogni valore vanno disposte su un grafico del tipo:

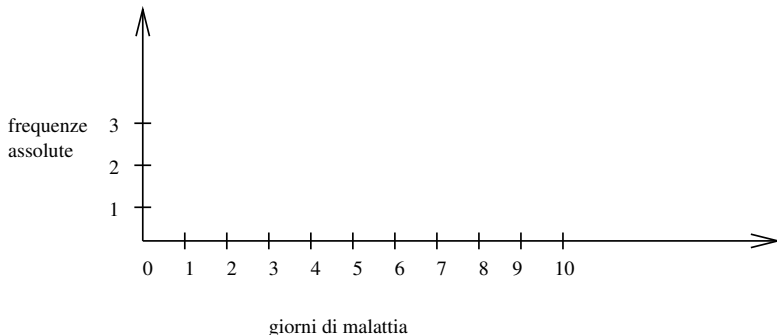


GRAFICO A BASTONCINI

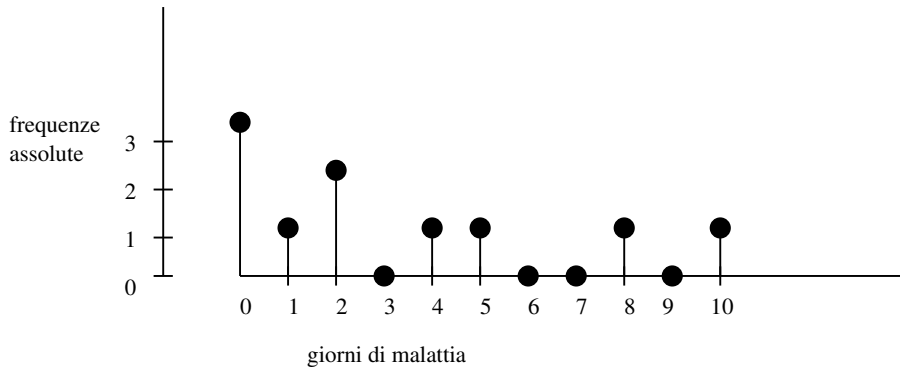


GRAFICO A BARRE

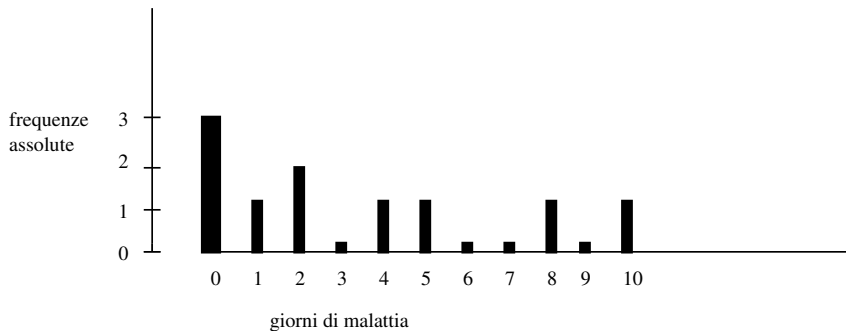
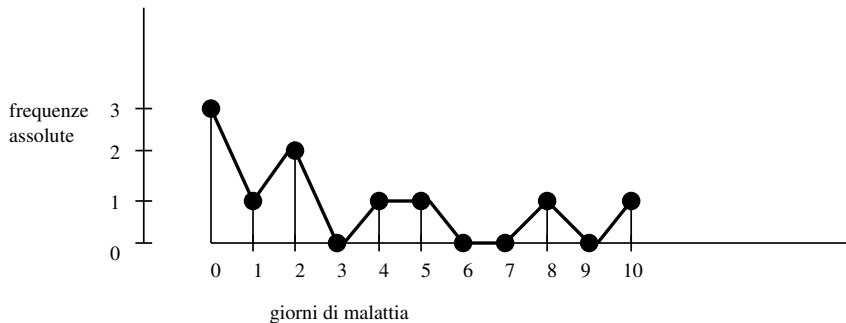


GRAFICO POLIGONALE

Si connettono con segmenti i valori delle frequenze.



STATISTICA DESCRITTIVA: VARIABILI CONTINUE

Variabili continue e variabili discrete che assumono molti valori distinti devono essere collocate in intervalli disgiunti detti **classi**.

Esempio 1. Risultati di 12 rilevazioni dell'emissione giornaliera di un gas inquinante da un impianto industriale in ordine crescente.

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

Consideriamo i seguenti intervalli (classi) di ampiezza 7:

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

I valori al bordo di una classe si chiamano **estremi** della classe. La parentesi quadra indica inclusione, quella tonda esclusione.

Le frequenze assolute e relative dei dati che cadono in ciascuna classe possono essere organizzate in una tabella

Classe	Freq. assoluta	Freq. relativa
[9, 16)	5	5/12
[16, 23)	2	1/6
[23, 30)	4	1/3
[30, 37)	1	1/12
TOTALE	12	1

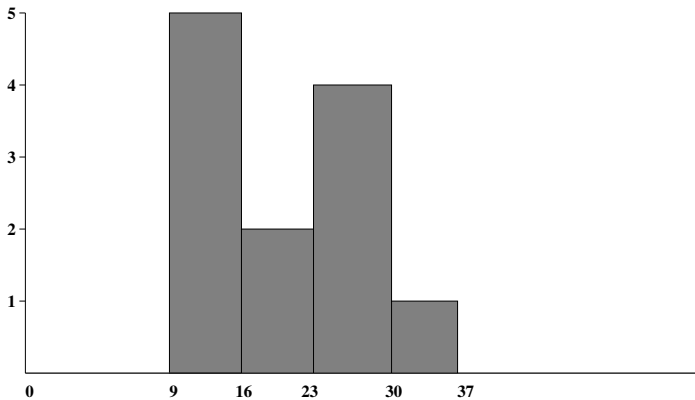
Il numero delle classi che si considerano è una scelta soggettiva. Troppe classi rendono la tabella poco leggibile, poche classi la rendono poco significativa. A volte conviene provare con diversi numeri di classi per capire quale è la tabella o il grafico più informativo. Di solito si scelgono da 5 a 10 classi.

ISTOGRAMMI

Il grafico più usato per i dati raggruppati è l'**istogramma**. Questo è un grafico a barre con le colonne sistemate adiacenti l'una all'altra. Come prima nell'asse verticale di norma si rappresentano o le frequenze assolute o quelle relative.

Un istogramma consiste di un insieme di rettangoli adiacenti aventi base sull'asse orizzontale: le basi sono gli intervalli che definiscono le classi (i punti medi degli intervalli sono i valori centrali delle classi). Se le classi hanno tutte la stessa ampiezza le altezze sono uguali alle corrispondenti frequenze assolute (o relative).

Classe	Freq. assoluta
[9, 16)	5
[16, 23)	2
[23, 30)	4
[30, 37)	1
TOTALE	12



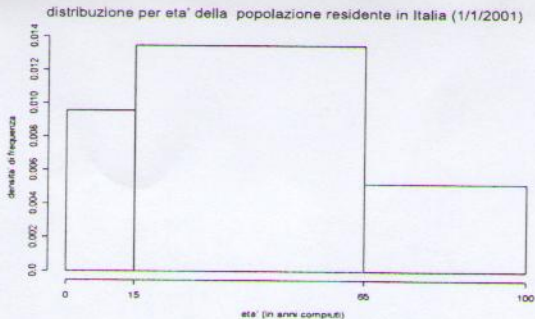
Nell'esempio 1 l'ampiezza degli intervalli (classi) è la stessa ma a volte conviene considerare classi di ampiezza diversa.

Esempio 2. La seguente tabella mostra l'età della popolazione residente in Italia al 1 gennaio 2001.

Classe	Freq. assoluta	Freq. rel.	Ampiezza
[0, 15)	8.303.904	0.144	15
[15, 65)	38.984.178	0.674	50
≥ 65	10.555.935	0.182	36 (max 100)
TOTALE	57.844.017	1	

Conviene costruire un istogramma con rettangoli che hanno **base = ampiezza della classe** e **area** (non più altezza!) **= frequenza**. Di conseguenza l'altezza degli istogrammi è pari alla frequenza divisa per l'ampiezza (chiamata anche densità di frequenza).

Classe	Frequenza	Ampiezza	Freq./Amp.
[0, 15)	0.144	15	0,0096
[15, 65)	0.674	50	0,01348
≥ 65	0.182	36	0,005055556



STATISTICA DESCRITTIVA: VARIABILI CATEGORICHE

I valori possibili di una variabile categorica vengono chiamati **categorie**.

Esempio. Indagine sullo stato occupazionale di 215 laureati della Facoltà di Scienze a 6 mesi dalla laurea.

Categoria	Freq. ass.	Freq. rel.	Freq. percent.
Lavora	127	0.5907	59.07
Cerca lavoro	54	0.2512	25.12
Tirocinio o stage	17	0.0791	7.91
Altro	17	0.0791	7.91
TOTALE	215	1	100

Per questo tipo di dati si usa di preferenza il **grafico a torta**.

Le frequenze sono rappresentate da settori circolari aventi angolo x che si ottengono dalla proporzione:

$$x = 360 \times \frac{\text{freq.percentuale}}{100} = 360 \times \text{freq.relativa}$$

