



Cofinanziato
dall'Unione europea



DATA LAB

GUARDA AVANTI

Big Data, nuove competenze
per nuove professioni.



Università
degli Studi
di Ferrara



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



UNIVERSITÀ
DI PARMA



POLITECNICO
MILANO 1863
POLO TERRITORIALE DI
PIACENZA



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

“Anticipare la crescita con le nuove competenze sui Big Data” Operazione Rif. PA 2023-19167/RER approvata con

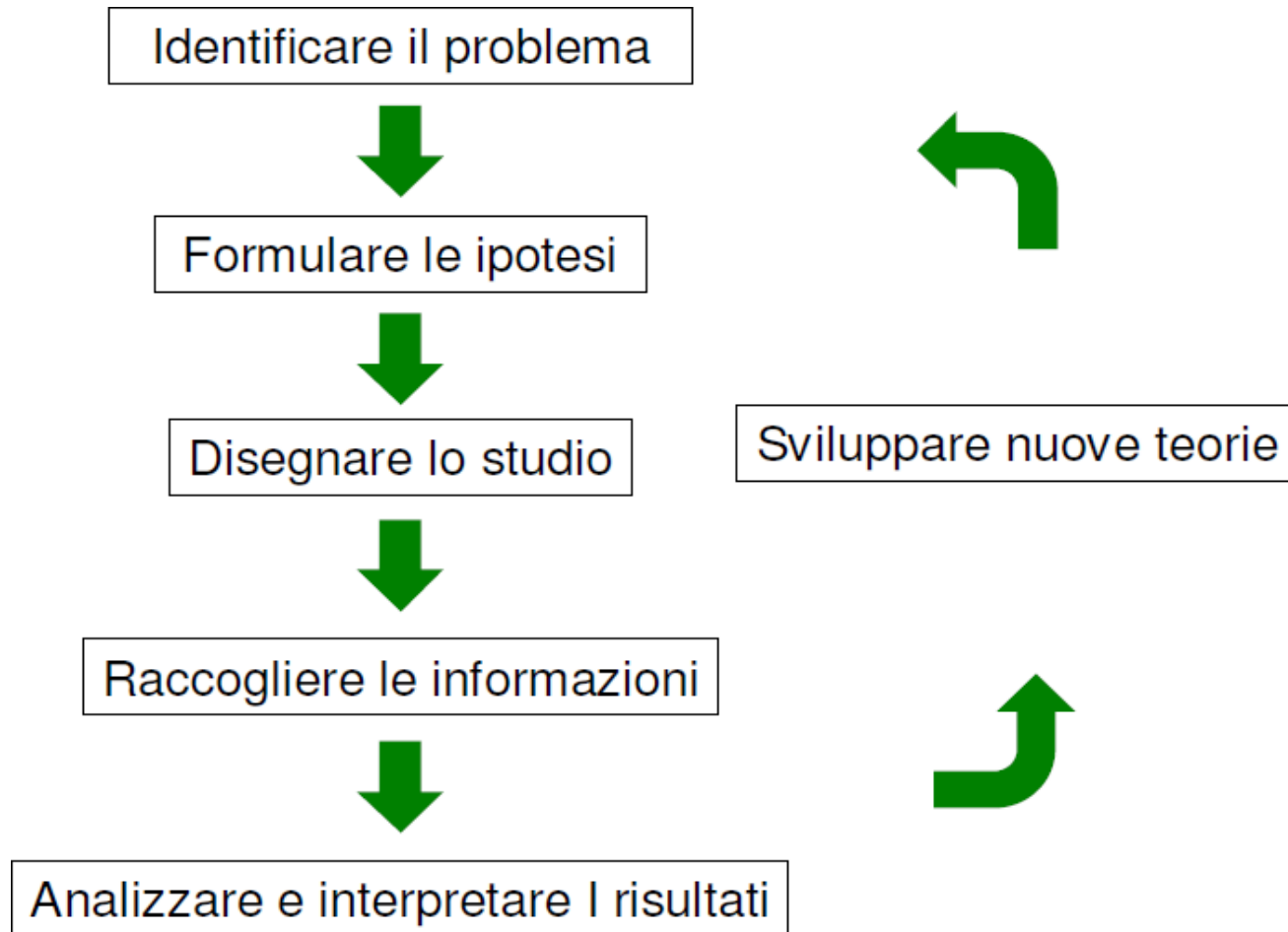
DGR

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Operazione Rif. PA 2023-19167/RER/10/1, "ANTICIPARE LA CRESCITA CON LE NUOVE
COMPETENZE SUI BIG DATA", approvata dalla Regione Emilia-Romagna con DGR n° 843
del 29/05/2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027



Informazioni, nuove conoscenze, decisioni



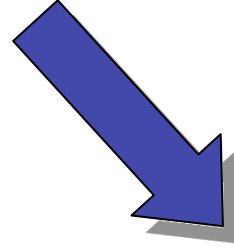
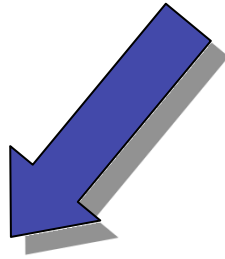
LA STATISTICA

- La Statistica ha come scopo la *conoscenza quantitativa dei fenomeni collettivi*. L'analisi statistica mira ad individuare **modelli** di interpretazione della realtà, attraverso canoni e tecniche che sono astrazioni, semplificazioni di una moltitudine di aspetti e di manifestazioni del reale.
- E' costituita da un insieme dei metodi che consentono di raccogliere, ordinare, riassumere, presentare ed analizzare dati e informazioni, trarne valide conclusioni e prendere decisioni sulla base di tali analisi e risultati.

Perché la Statistica è necessaria?

- per leggere, comprendere ed analizzare criticamente relazioni e risultati di analisi

Statistica



Funzione Descrittiva

Offre il metodo per riassumere le informazioni in modo da renderle utilizzabili più facilmente. Riduce i dati in forma maneggevole, sostituendo a molti numeri poche misure.

Funzione Inferenziale

Permette di generalizzare le informazioni, ricavando proprietà e leggi generali sulla base di dati rilevati solamente su una parte (campione) della popolazione.

UNITA' STATISTICA

DEFINIZIONE: Ogni elemento o caso appartenente alla popolazione oggetto diretto della osservazione da cui si raccolgono i dati.

Esempi: un singolo paziente o soggetto;
uno studente;
un occhio;
un campione di sangue;
etc

CAMPIONE E UNIVERSO

Un **campione rappresentativo** è un **insieme di unità statistiche** tratte da un **universo** (o **popolazione**). Un **universo** consiste della **totalità delle unità statistiche** che posseggono determinate caratteristiche. (Non necessariamente una popolazione deve essere composta da un numero elevato di elementi).

Esempi:

- **Universo:** tutti i pazienti adulti con una determinata malattia.
- **Campione:** 120 pazienti con quella malattia, inclusi in una sperimentazione clinica.
- **Universo:** un lotto di 5000 impianti.
- **Campione:** 10 impianti di quel lotto sottoposti a controllo di qualità.

Il campione è soltanto una parte del tutto.

VARIABILE STATISTICA

DEFINIZIONE: Ciascuna informazione, caratteristica che verrà rilevata sulle unità statistiche sarà chiamata **VARIABILE**.

Le “categorie” della variabile statistica in osservazione costituiscono le MODALITA' della variabile.

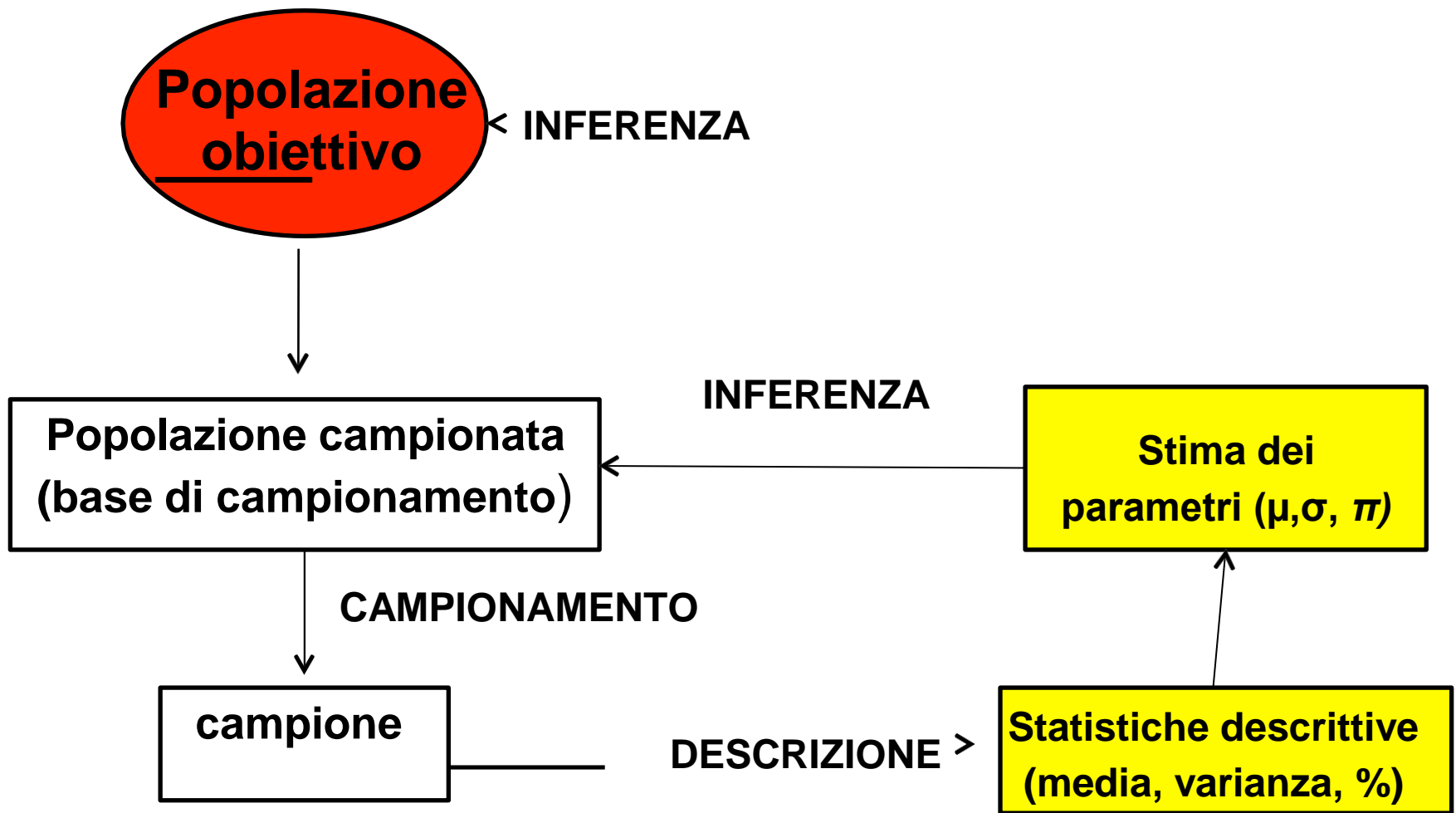
I dati sperimentali (variabili) si presentano sotto differenti forme, essi possono essere sia di tipo quantitativo sia di tipo qualitativo, ed essere espressi o con scale continue o con scale discrete.

GLOSSARIO

- **POPOLAZIONE**: l'insieme di tutte le unità statistiche oggetto dell'osservazione (es.: medici, paramedici, studenti, diabetici, obesi, addetti all'agricoltura...).
- **CAMPIONE**: la parte delle unità statistiche sottoposte all'osservazione, all'esperimento, etc.
- **UNITA' STATISTICA**: per ogni elemento o caso appartenente alla popolazione oggetto diretto della osservazione da cui si raccolgono i dati.

GLOSSARIO

- **CARATTERE (O VARIABILE):** la caratteristica (attributo o misura) osservata sulla unità statistica.
- **MODALITA':** ogni diversa presentazione del carattere o variabile osservata su ciascuna unità statistica.
- **FREQUENZA:** numero di volte che si presenta una data modalità.



Esempi

Le *variabili* sesso, età, peso, pressione arteriosa, etc (di pazienti inclusi in uno studio) hanno come modalità:

- *maschio* o *femmina* per la variabile "sesso";
- *anni*, per la variabile "età";
- *kg*, per il "peso corporeo",
- *mmHg*, per la "pressione arteriosa"
- *A, AB, B, 0* per il "gruppo sanguigno",
- *elementare, media inferiore, media superiore, università*,
,
per la variabile "titolo di studio"

VARIABILI STATISTICHE

VARIABILI QUALITATIVE

NOMINALI

Date due qualsiasi modalità, è possibile solo affermare se esse sono uguali o diverse.

Sesso; professione;
diagnosi medica; ...

ORDINALI O PER RANGHI

Esiste un criterio predeterminato per ordinare le modalità

ordine di nascita;
giorni della settimana;
indice di severità di una
malattia;...

VARIABILI QUANTITATIVE

DISCRETO

L'insieme delle modalità assumibili può essere messo in " corrisp. biunivoca " con un sottoinsieme dei numeri naturali.

Num. componenti
famiglia;
num. di figli;
num. di denti;
num. colonie batteriche in
una piastra;...

CONTINUO

(la variabile può assumere qualsiasi valore all'interno di intervalli di numeri reali.

statura;
peso;
glicemia;



LA SINTESI DEI DATI

Esempio 1. Su un campione di pazienti si rilevino le caratteristiche: sesso, età, altezza, peso, pressione arteriosa sistolica (PAS), tasso glicemico.

| | |
|---------------------|---------------------|
| nome: Rossi Amerigo | nome: Bianchi Paolo |
| sesso: maschio | sesso: maschio |
| età: 32 | età: 47 |
| altezza: 172 cm. | altezza: 170 cm. |
| peso: 64 Kg. | peso: 80 Kg. |
| PAS: 140 mm Hg. | PAS: 148 mm Hg. |
| glicemia: 190 mg/dl | glicemia: 180 mg/dl |

| | |
|------------------------|-----------------------|
| nome: Valenzi Alberica | nome: Alinori Alfonso |
| sessò: femmina | sessò: maschio |
| età: 45 | età: 27 |
| altezza: 168 cm. | altezza: 183 cm. |
| peso: 51 Kg. | peso: 85 Kg. |
| PAS: 125 mm Hg. | PAS: 138 mm Hg. |
| glicemia: 150 mg/dl | glicemia: 170 mg/dl |

Le informazioni raccolte per essere "trattate" da un computer devono essere organizzate in strutture chiamate comunemente

Data Base o File Dati.

Le informazioni vengono, comunemente, organizzate per riga, cioè su ogni riga, consecutivamente, vengono elencati i dati relativi ad un soggetto.

| A | B | C | D | E | F | G | H | I | J |
|----|-------|------------|--------------|-----------------------|--------------------------|------------------|-----------------|-------------|-------------|
| ID | Sesso | Età (anni) | Numero figli | Fumatore (0=NO; 1=SI) | BMI (kg/m ²) | Massa grassa (%) | Massa magra (%) | PAS (mm/Hg) | PAD (mm/Hg) |
| 1 | M | 58 | 0 | 0 | 22 | 3.83 | 29.94 | 130 | 100 |
| 2 | M | 57 | 0 | 0 | 25 | 6.44 | 21.68 | 131 | 90 |
| 3 | M | 51 | 1 | 0 | 40 | 4 | 27.62 | 128 | 90 |
| 4 | F | 36 | 3 | 0 | 38 | 14.03 | 32.91 | 110 | 80 |
| 5 | M | 50 | 2 | 0 | 35 | 9.43 | 28.6 | 136 | 85 |
| 6 | M | 64 | 2 | 1 | 29 | 9.42 | 26.43 | 149 | 100 |
| 7 | F | 39 | 1 | 1 | 24 | 8.41 | 26.99 | 120 | 80 |
| 8 | F | 38 | 1 | 0 | 18 | 20.23 | 33.85 | 115 | 75 |
| 9 | F | 70 | 0 | 1 | 19 | 3.99 | 24.49 | 95 | 50 |
| 10 | M | 49 | 4 | 1 | 31 | 7.29 | 30.37 | 100 | 40 |
| 11 | M | 53 | 2 | 1 | 30 | 11.18 | 26.83 | 105 | 40 |
| 12 | F | 56 | 3 | 1 | 22 | 5.84 | 26.53 | 100 | 60 |
| 13 | F | 54 | 5 | 1 | 39 | 4.67 | 28.91 | 95 | 45 |
| 14 | F | 45 | 0 | 0 | 24 | 13.24 | 27.08 | 120 | 90 |
| 15 | M | 51 | 0 | 0 | 25 | 3.79 | 29.52 | 120 | 80 |
| 16 | F | 54 | 1 | 0 | 18 | 3.28 | 30.82 | 130 | 70 |
| 17 | F | 39 | 1 | 0 | 20 | 7.22 | 27.33 | 135 | 70 |
| 18 | M | 68 | 2 | 1 | 29 | 25.72 | 31.84 | 90 | 60 |
| 19 | M | 61 | 3 | 0 | 21 | 5.12 | 24.82 | 95 | 60 |
| 20 | M | 59 | 2 | 1 | 22 | 5.07 | 25.7 | 100 | 80 |
| 21 | M | 60 | 2 | 0 | 20 | 5.03 | 30.02 | 115 | 75 |

DISTRIBUZIONI SEMPLICI DI FREQUENZE

I dati (cioè le informazioni raccolte) spesso sono di non immediata lettura.

Per questo si procede ad una sistematizzazione e sintesi delle informazioni raccolte, cioè alla loro **tabulazione**. Per ogni variabile si calcolano le **frequenze assolute (f.a.)** che rappresentano il numero di u.s. che presentano una stessa modalità del carattere.

Alcune distribuzioni semplici di frequenze

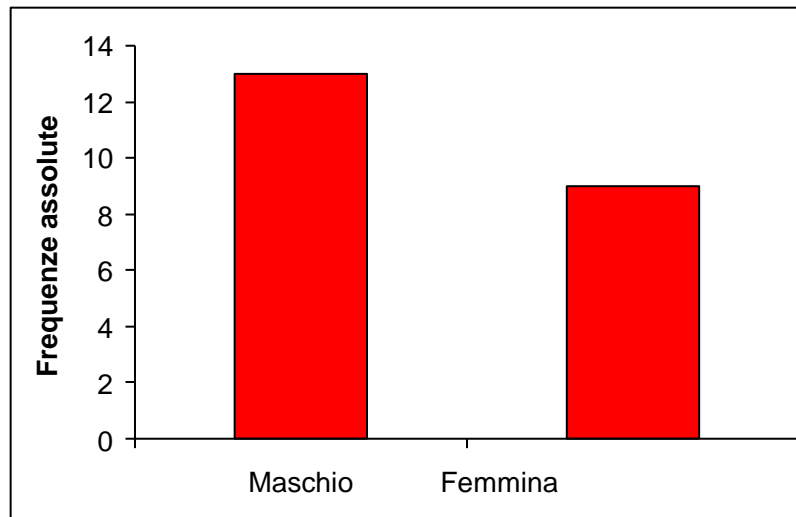
| Sesso | f.a. |
|--------------|-------------|
| M | 120 |
| F | 90 |
| Tot | 210 |

| N figli | f.a. |
|--------------------|-------------|
| 0 | 60 |
| 1 | 40 |
| 2 | 60 |
| >2 | 50 |
| Tot | 210 |

| Età (anni) | f.a. |
|-----------------------|-------------|
| <40 | 40 |
| 41-45 | 10 |
| 46-50 | 20 |
| 51-55 | 60 |
| 56-60 | 50 |
| >60 | 30 |
| Tot | 210 |

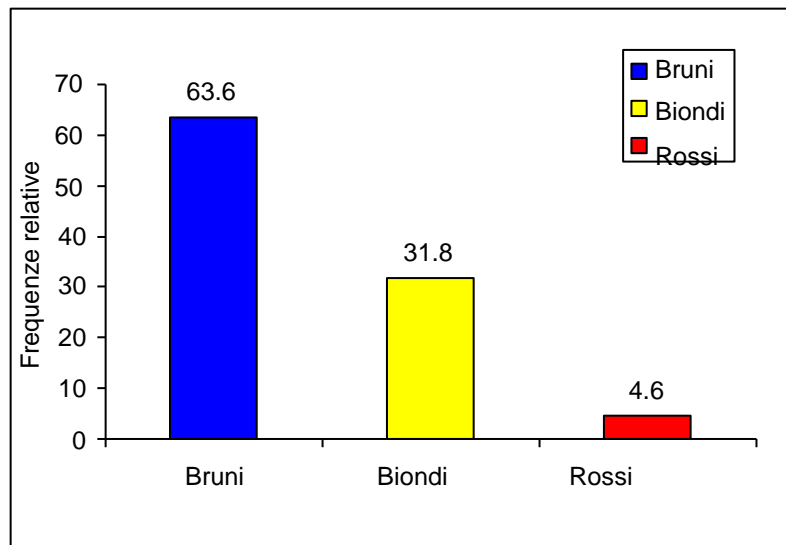
Variabile qualitativa nominale: SESSO

| SESSO | Frequenza assoluta | Frequenza % |
|---------|--------------------|-------------|
| Maschio | 12 | 57.1 |
| Femmina | 9 | 42.9 |
| Totale | 21 | 100 |



Variabile qualitativa nominale: COLORE DEI CAPELLI

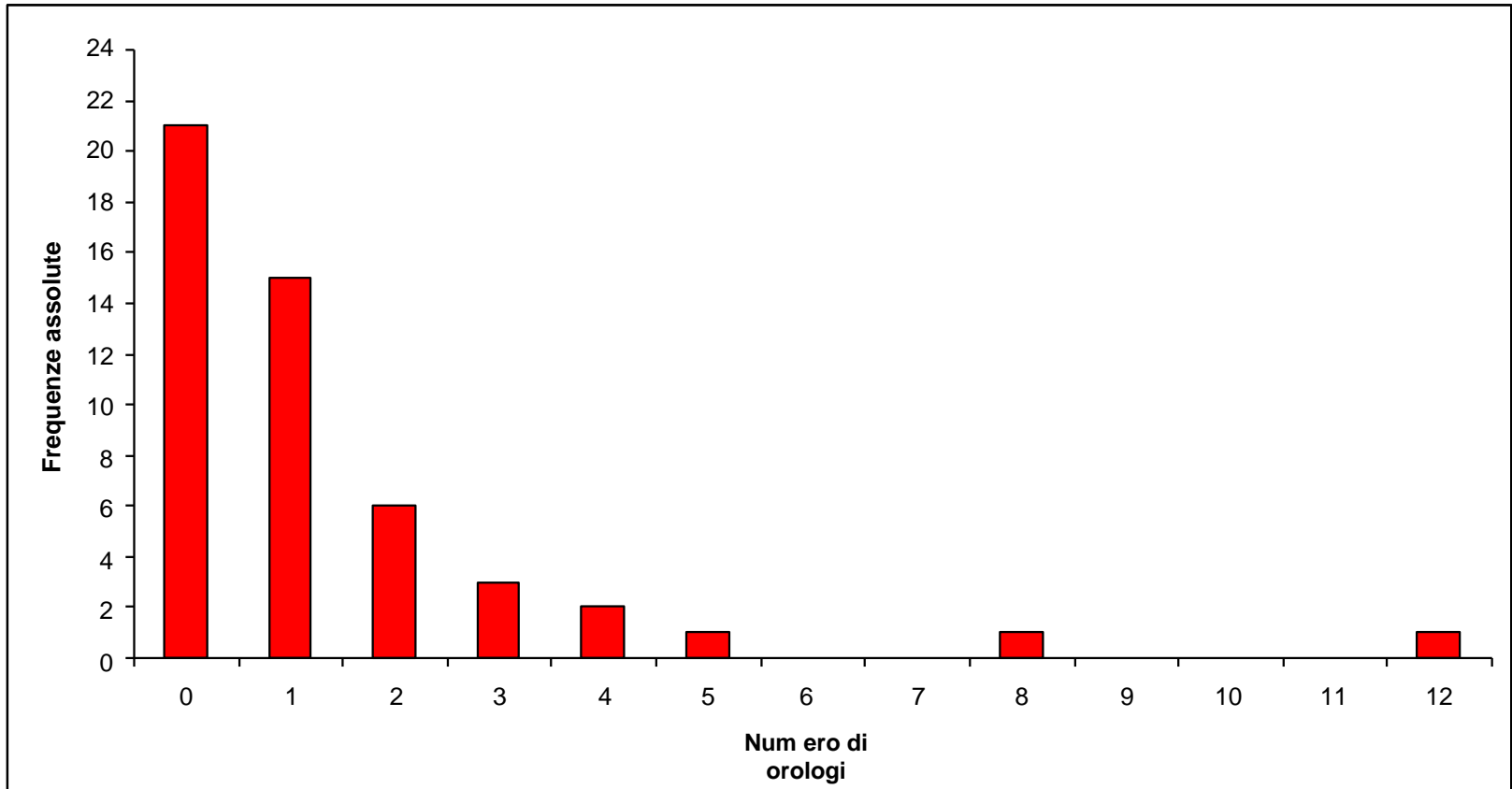
| COLORE CAPELLI | Frequenza assoluta | Frequenza % |
|----------------|--------------------|-------------|
| Bruni | 14 | 63.6 |
| Biondi | 7 | 31.8 |
| Rossi | 1 | 4.6 |
| Totale | 22 | 100 |



Variabile quantitativa ordinale: NUMERO DI OROLOGI

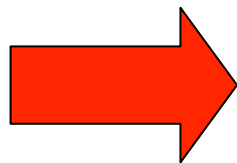
| NUMERO DI OROLOGI | Frequenza assoluta | Frequenza % | Frequenza cumulata |
|-------------------|--------------------|-------------|--------------------|
| 0 | 21 | 42.0 | 21 |
| 1 | 15 | 30.0 | 36 (21+15) |
| 2 | 6 | 12.0 | 42 (36+6) |
| 3 | 3 | 6.0 | 45 (...) |
| 4 | 2 | 4.0 | 47 |
| 5 | 1 | 2.0 | 48 |
| 8 | 1 | 2.0 | 49 |
| 12 | 1 | 2.0 | 50 |
| Totale | 50 | 100 | |

BAR CHART



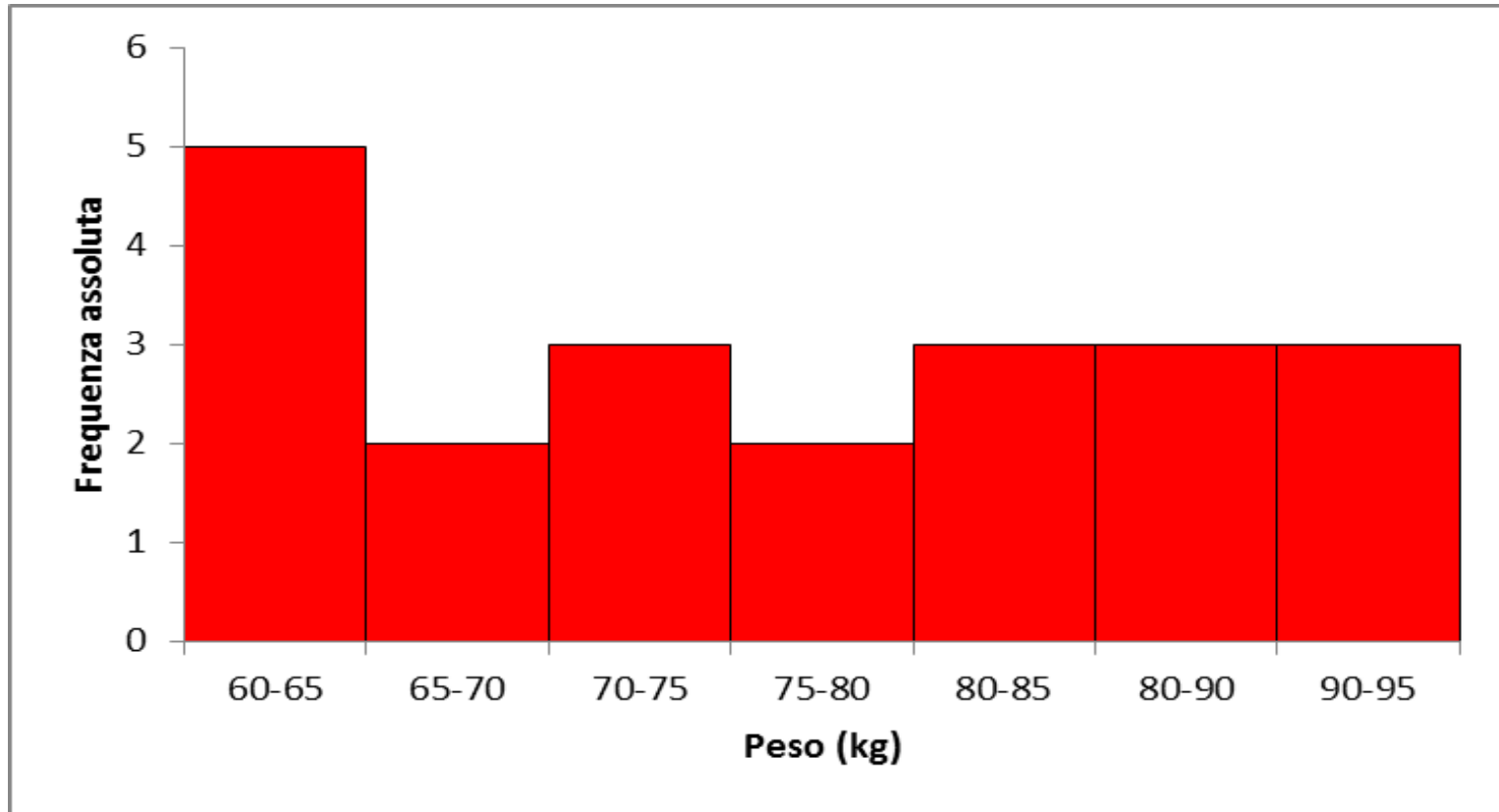
| PESO CORPOREO |
|------------------|
| 54 |
| 50 |
| 57 |
| 68 |
| 63 |
| 51 |
| 47 |
| 64 |
| 62 |
| 110 |
| 60 |
| 68 |
| 76 |
| 70 |
| 74 |
| 75 |
| 47 |
| 74 |
| 53 |
| 70 |
| 65 |
| 65 |

Variable quantitativa: PESO CORPOREO



| Peso corporeo | Freq. assoluta | Freq. relativa | Freq. cumulata |
|---------------|----------------|----------------|----------------|
| 40-50 | 3 | 0.14 | 0.14 |
| 51-60 | 5 | 0.23 | 0.37 |
| 61-70 | 9 | 0.41 | 0.78 |
| 71-80 | 4 | 0.18 | 0.96 |
| 81-90 | 0 | 0 | 0.96 |
| >90 | 1 | 0.04 | 1 |
| Totale | 22 | 1 | |

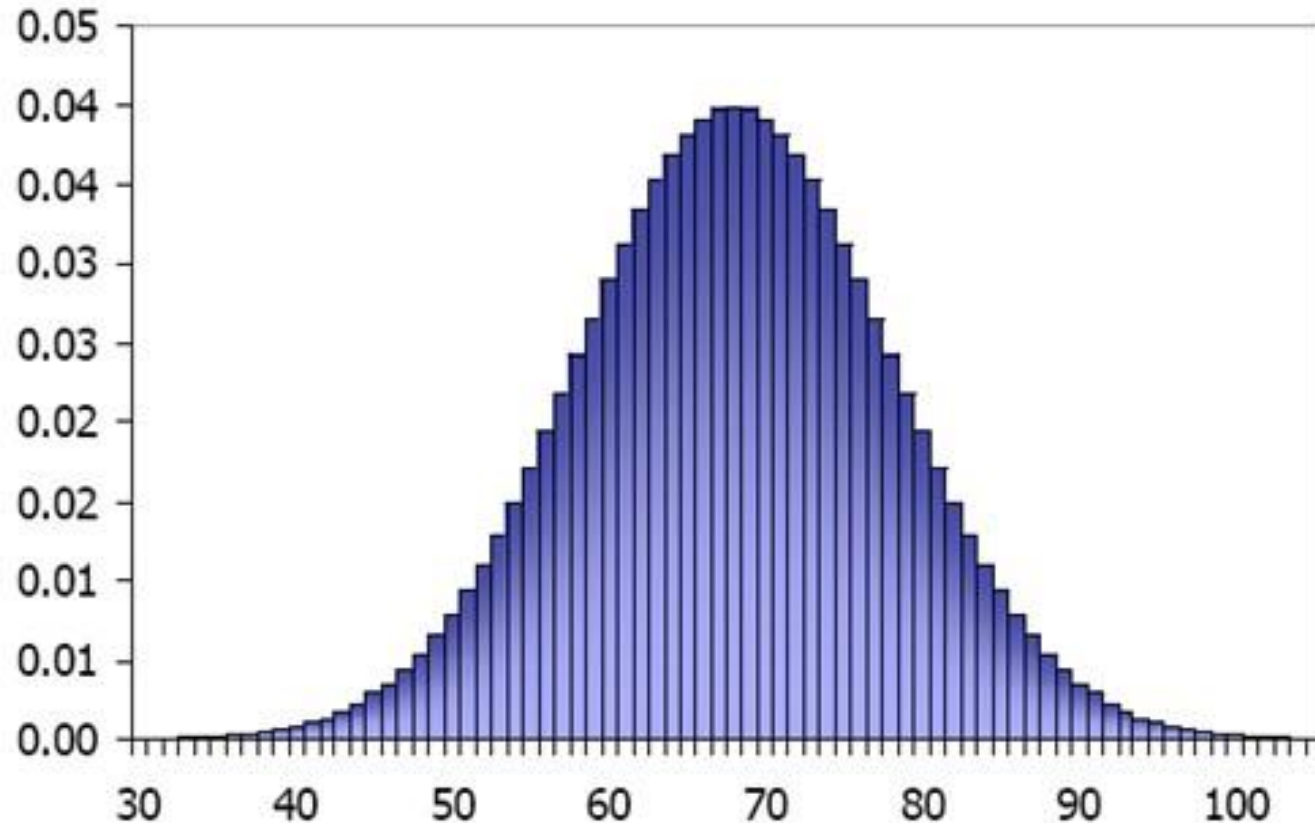
L'ISTOGRAMMA



In presenza di un carattere con molte modalità è possibile e conveniente formare delle classi di valori (BINS)

L'ISTOGRAMMA

Peso di 150.000 studenti



Distribuzione doppia di frequenze assolute

| BCO | Fumatori | Non fumatori | Totale |
|--------|----------|--------------|--------|
| SI | 160 | 100 | 260 |
| NO | 120 | 70 | 190 |
| Totale | 280 | 170 | 450 |



Ci accorgiamo che il confronto **non** può essere effettuato solo con le f.a. in quanto esse si riferiscono a collettivi di numerosità diversa.



Se vogliamo confrontare le frequenze le dobbiamo “depurare” dalla numerosità del collettivo; ciò lo si fa dividendo le f.a. per la numerosità (N) della popolazione e moltiplicando per 100 (cioè facendo riferimento ad una ipotetica popolazione di 100 unità).

Le frequenze così calcolate sono le **frequenze percentuali (f.%)**

Distribuzione doppia di frequenze percentuali

| BCO | Fumatori | | Non fumatori | |
|--------|----------|-------|--------------|-------|
| | f.a. | f.a.% | f.a. | f.a.% |
| SI | 160 | 57.1 | 100 | 58.8 |
| NO | 120 | 42.8 | 70 | 41.2 |
| Totale | 280 | 100 | 170 | 100 |

Distrib. di frequenze assolute, relative e cumulate

| Età | f.a. | f% | f.a.cum | f% cum |
|--------|------|------|---------|--------|
| 17 | 30 | 13.6 | 30 | 13.6 |
| 18 | 60 | 27.3 | 90 | 40.9 |
| 19 | 120 | 54.6 | 210 | 95.5 |
| 20 | 10 | 4.5 | 220 | 100 |
| Totale | 220 | 100 | | |

I GRAFICI STATISTICI

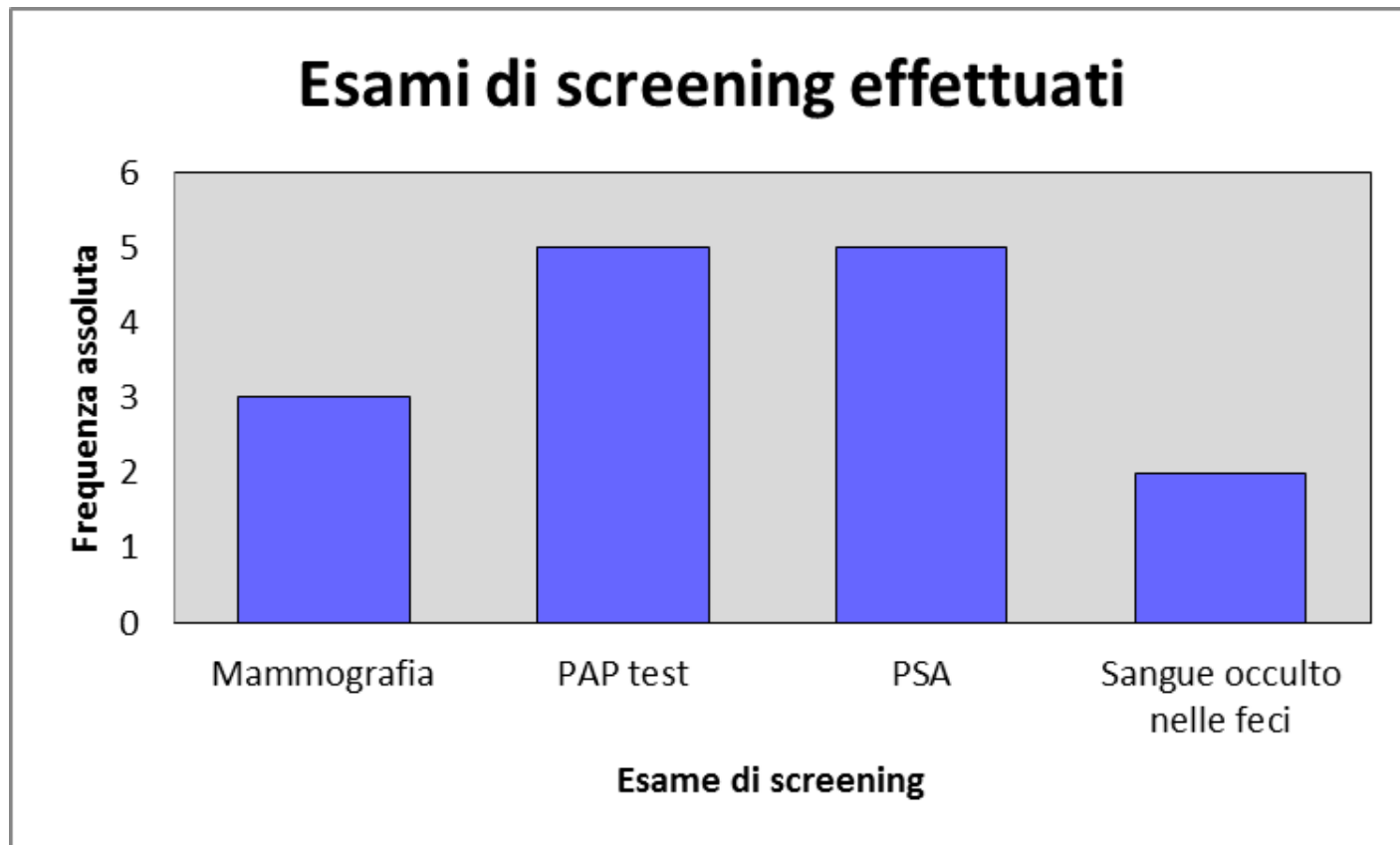
Scopo dei grafici è quello di rendere l'informazione contenuta in una serie di dati:

- ✓ di più facile comprensione;
- ✓ di più diretta lettura.

Pertanto un grafico deve fornire al lettore una informazione sintetica e facile da interpretarsi.

BAR CHART

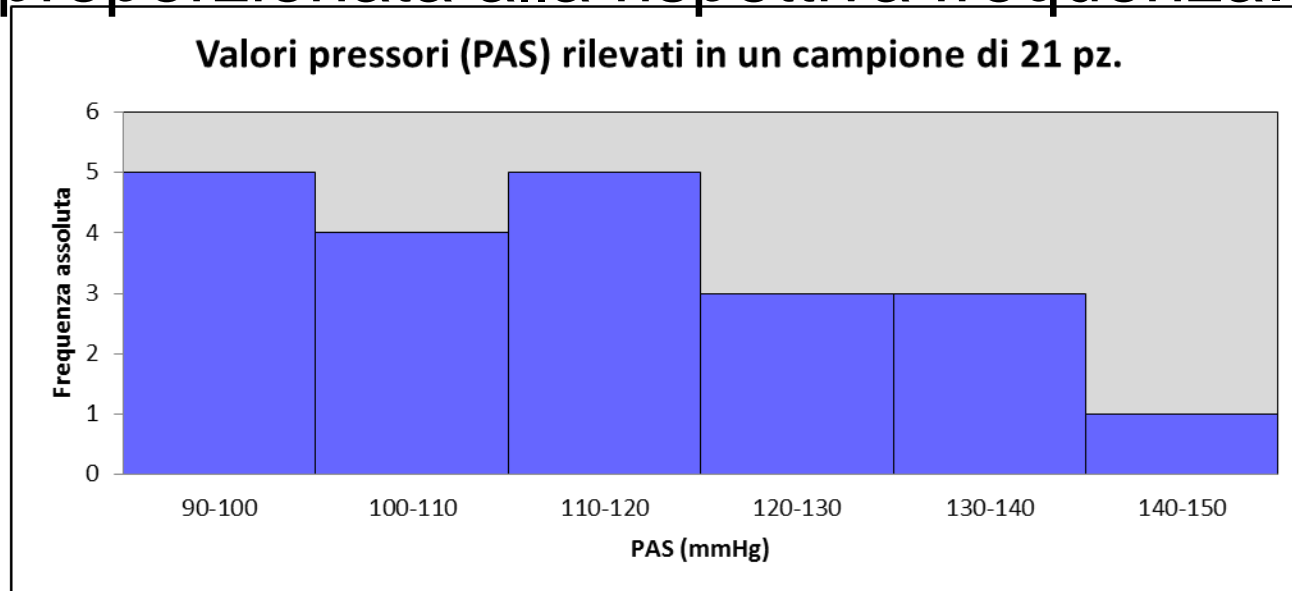
Usati per variabili qualitative, l'altezza delle barre rappresenta freq. assoluta o percentuale



ISTOGRAMMI

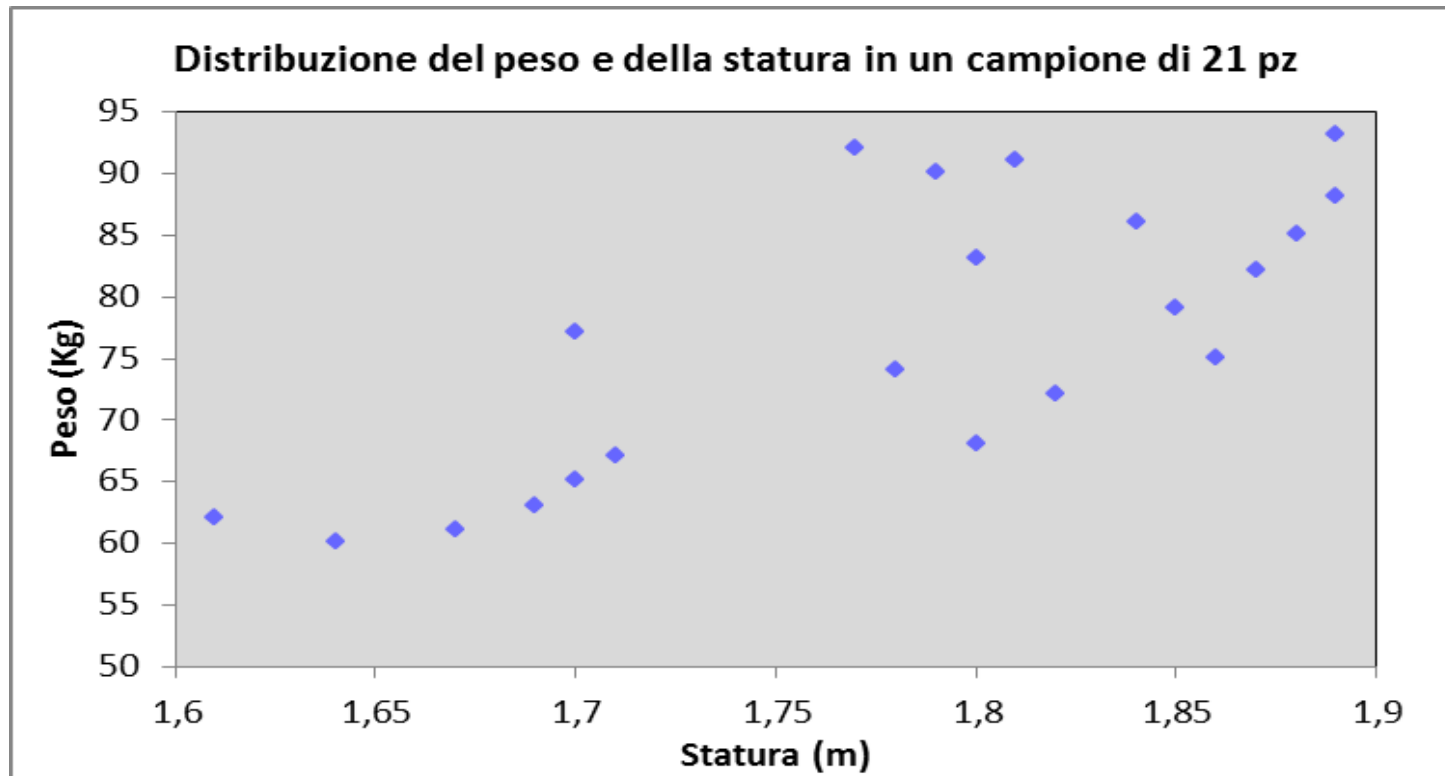
Indicati per rappresentare distribuzioni in classi (variabili quantitative continue).

Costituiti da una serie di barre rettangolari contigue ognuna in rappresentanza di una classe e con area proporzionata alla rispettiva frequenza.



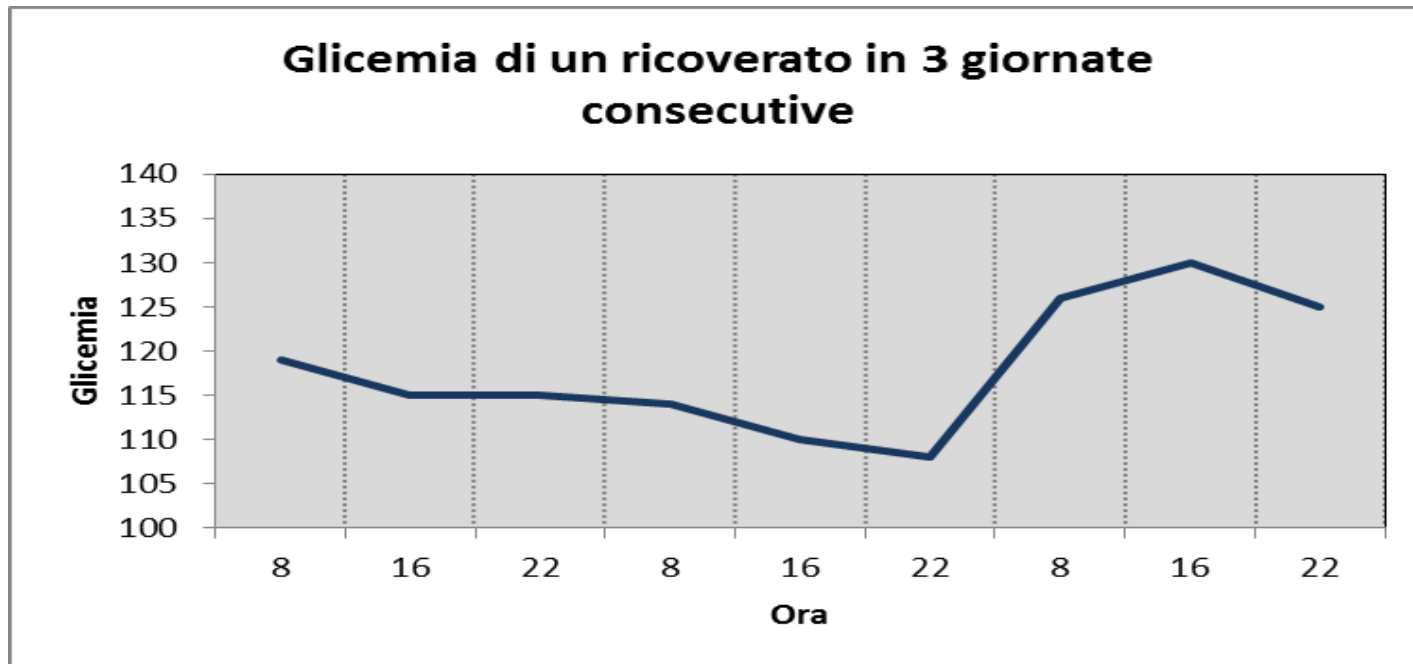
GRAFICI PER PUNTI (Scatter Plot)

Costituito dai punti corrispondenti alle diverse coppie di valori rilevati. Indicati per evidenziare le associazioni tra variabili quantitative.



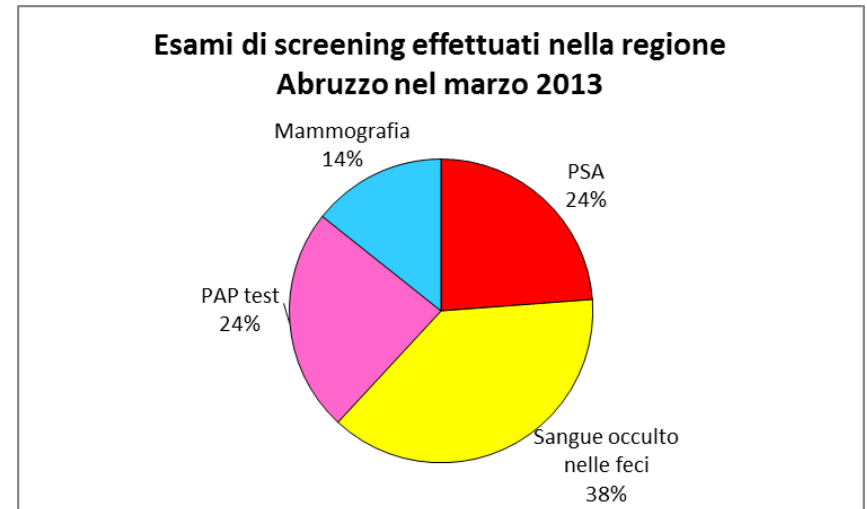
GRAFICI PER SPEZZATE

Si ottengono dai grafici per punti congiungendo i vari punti. Indicati per evidenziare una continuità tra valori come ad es. nella rappresentazione delle serie temporali.



DIAGRAMMI A SETTORI CIRCOLARI (TORTA)

Indicati per variabili qualitative allo scopo di evidenziare le frequenze % delle singole modalità. L'area di un cerchio viene suddivisa in settori proporzionali alle frequenze %



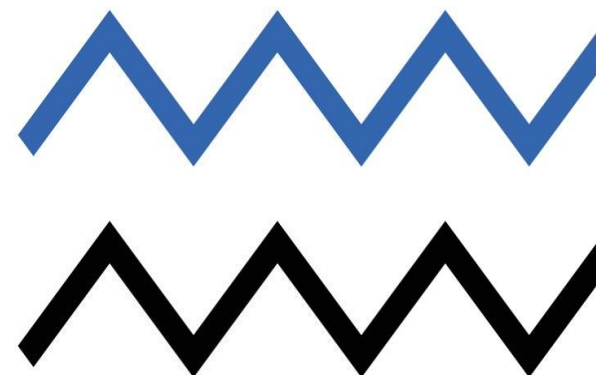


GUARDA AVANTI

Big Data, nuove competenze

per nuove professioni.

www.bigdata-lab.it



Università
degli Studi
di Ferrara



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



UNIVERSITÀ
DI PARMA



POLITECNICO
MILANO 1863
POLO TERRITORIALE DI
PARMA



UNIVERSITÀ
CATTOLICA
del Sacro Cuore



Sapere utile

