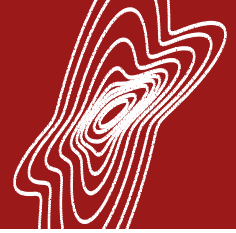


# METODI STATISTICI PER LA BIOINGEGNERIA (B)

## PARTE 1: STATISTICA DESCRITTIVA

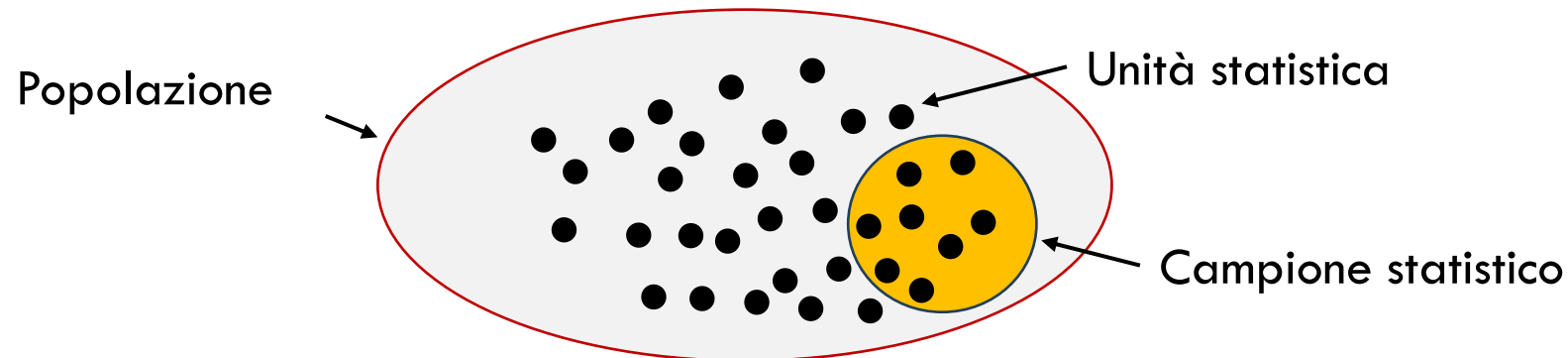
A.A. 2026-2027  
Prof. Enrico Longato

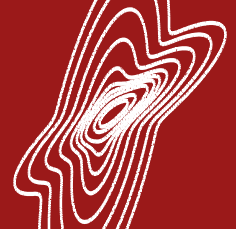


# STATISTICA

**Statistica:** disciplina che si occupa dello studio dei fenomeni collettivi, attraverso la raccolta dati, la loro descrizione ed elaborazione, al fine di trarre conoscenza. È anche definita come l'arte di apprendere dai dati.

- **Fenomeni collettivi:** fenomeni il cui studio richiede una pluralità di osservazioni.
- **Collettivo o popolazione:** insieme di elementi omogenei secondo una o più **caratteristiche** in modo tale che sia possibile stabilire se un elemento appartiene o no alla popolazione.
- **Unità statistica:** singolo elemento che compone la popolazione.
- **Campione statistico:** sottoinsieme di unità estratto da una popolazione.

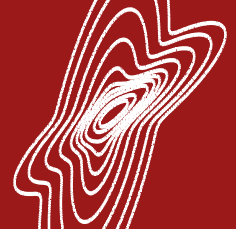




# CARATTERE O VARIABILE

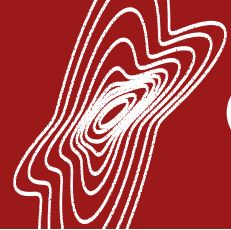


- **Carattere o variabile:** un attributo o una proprietà osservabile delle unità di un collettivo.
- **Variabile quantitativa:** assume valori numerici
  - **Variabile quantitativa discreta:** può assumere delle quantità numeriche distinte, come ad esempio dei numeri interi, preventivamente individuabili ed elencabili.
  - **Variabile quantitativa continua:** può assumere tutti i valori in un certo intervallo di numeri reali.
- **Variabile qualitativa:** assume valori non numerici che possono essere ordinabili (variabile qualitativa **ordinale** o carattere qualitativo rettilineo) o non ordinabili (variabile qualitativa **nominale** o carattere qualitativo sconnesso).



# ESEMPIO

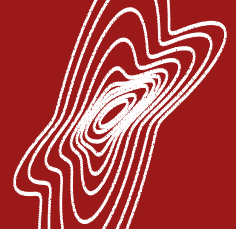
- **Popolazione:** individui residenti in Veneto affetti da diabete.
- **Unità:** un individuo residente in Veneto affetto da diabete.
- **Campione:** soggetti affetti da diabete in cura presso l'azienda ospedaliera di Padova (Azienda Ospedaliera Università Padova, AOUP); pazienti con diabete di AOUP registrati nel sistema dal 23 dicembre 2021 ad oggi.
- **Variabili quantitative continue:** altezza, peso, pressione sanguigna, glicemia a digiuno, concentrazione di colesterolo nel sangue ...
- **Variabili quantitative discrete:** età in anni, numero di componenti del nucleo familiare, anni dalla diagnosi di diabete ...
- **Variabili qualitative ordinali:** livello di istruzione (elementare, media, secondaria, ...), grado di ipertensione (bassa, normale, alta), ...
- **Variabili qualitative nominali:** colore degli occhi (marrone, azzurro, verde, ...), stato coniugale (celibe/nubile, sposato/a, vedovo/a, divorziato/a,...), ...



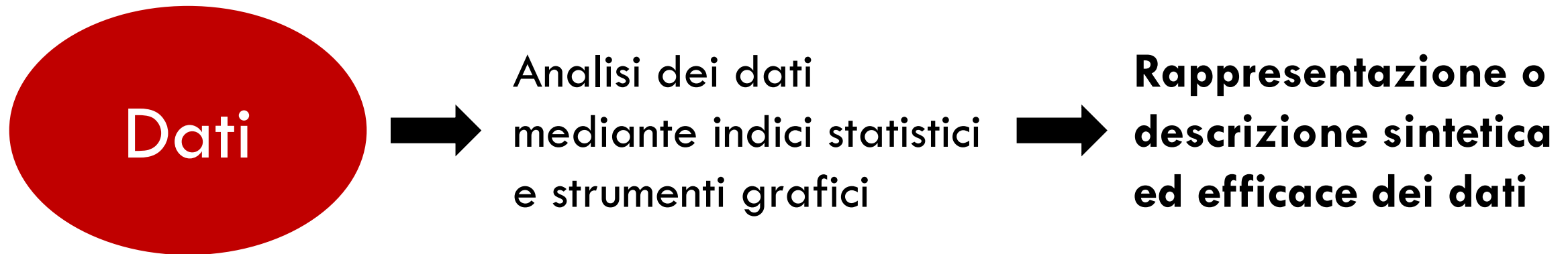
# GLI STEP PRINCIPALI DI UN'INDAGINE STATISTICA



1. Definizione degli **obiettivi** dell'indagine statistica
2. Rilevazione: **raccolta dei dati** relativi alle unità statistiche
  - Rilevazione completa: coinvolge tutte le unità statistiche → **indagine completa**
  - Rilevazione parziale: coinvolge un campione di unità statistiche → **indagine campionaria**
3. **Elaborazione dei dati** per ottenere risultati fruibili
4. **Presentazione dei dati** in forma tabulare o grafica attraverso metriche sintetiche
5. **Interpretazione dei dati** alla luce degli obiettivi prefissati



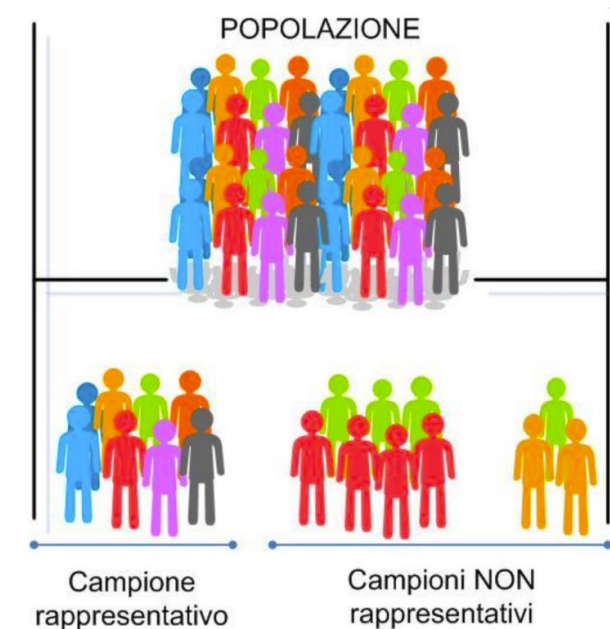
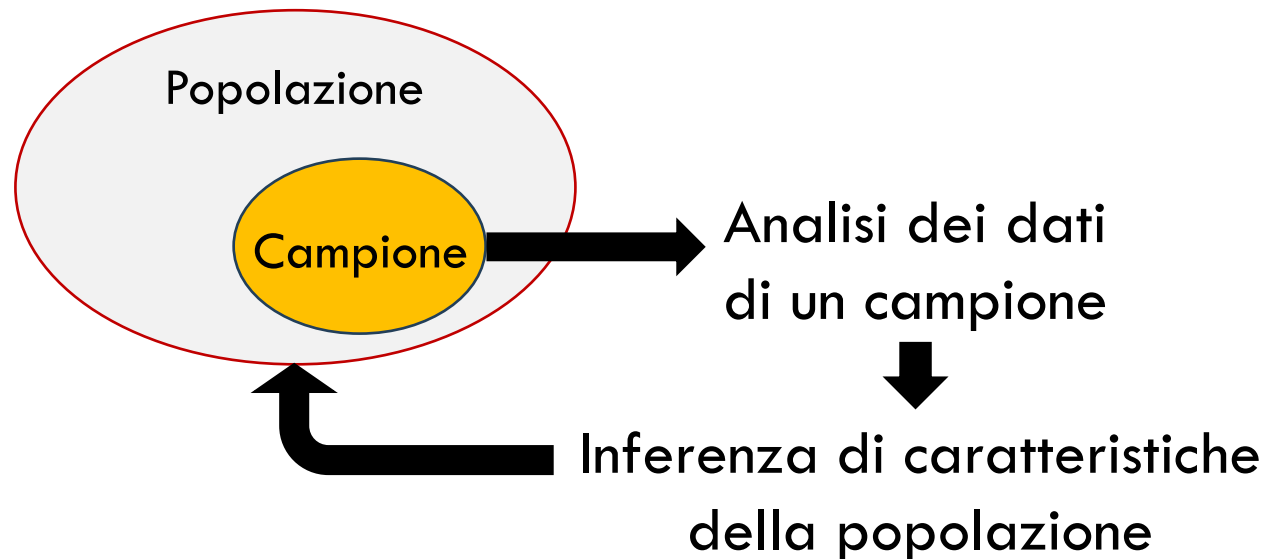
**Statistica descrittiva:** i dati vengono analizzati senza fare assunzioni esterne all'insieme di dati considerato.

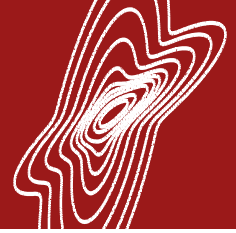


- **Indici statistici:** indicatori di posizione (es. media), di dispersione (es. varianza), di correlazione, di forma (es. curtosi e skewness)
- **Strumenti grafici:** diagrammi a barre, a torta, istogrammi, boxplot

**Statistica inferenziale o inferenza statistica:** i dati di un **campione** statistico vengono analizzati al fine di trarre, o **inferire**, delle **caratteristiche della popolazione** di appartenenza.

Il **campione** deve essere **rappresentativo** della **popolazione** di interesse, altrimenti non è detto che le conclusioni tratte sulla base del campione siano applicabili alla popolazione.

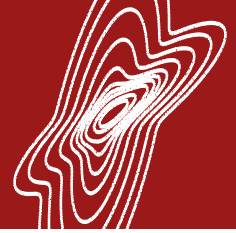




# STATISTICA INFERENZIALE (2/2)



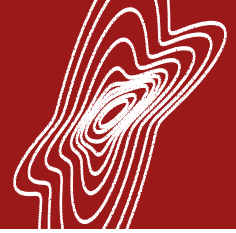
- Per poter inferire delle caratteristiche della popolazione a partire dal campione osservato, si deve tener conto dell'**influenza del caso** sui dati osservati → alcune differenze osservate nel valore di una variabile tra due campioni potrebbero essere semplicemente casuali.
- La statistica inferenziale:
  - si basa su un **modello probabilistico** del fenomeno studiato → modello che descrive la probabilità che i dati osservati assumano certi valori.
  - utilizza quindi i dati per fare inferenze su queste probabilità.
- Nelle prossime lezioni ripasseremo le basi di probabilità.



# STRUMENTI DELLA STATISTICA DESCRITTIVA



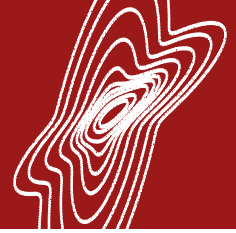
- Tabelle di frequenza
- Indici di posizione
- Indici di dispersione
- Percentili
- Coefficienti di correlazione
- Rappresentazioni grafiche



# TABELLE DI FREQUENZA



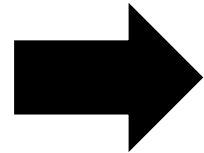
- Dati che possono assumere un numero contenuto di valori distinti possono essere efficacemente rappresentati tramite tabelle di frequenza.
  - Frequenze assolute
  - Frequenze relative o percentuali



# FREQUENZA ASSOLUTA

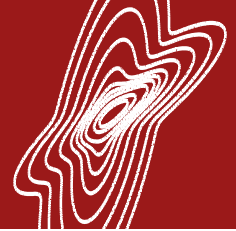
- **Frequenza assoluta:** il numero di unità statistiche che presentano un determinato valore per una data variabile.
- **Esempio:**
  - Variabile 'voto' che può assumere i valori insufficiente, sufficiente, buono, ottimo.
  - Valori osservati per la variabile voto:

- Ottimo
- Buono
- Buono
- Insufficiente
- Ottimo
- Sufficiente
- Sufficiente
- Sufficiente
- Buono
- Buono



**Tabella di frequenza**

| Valore        | Frequenza assoluta |
|---------------|--------------------|
| Insufficiente | 1                  |
| Sufficiente   | 3                  |
| Buono         | 4                  |
| Ottimo        | 2                  |



# FREQUENZA RELATIVA

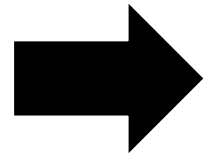
➤ **Frequenza relativa:** la frazione di unità statistiche che presentano un determinato valore per una data variabile. Equivale alla frequenza assoluta divisa per il numero di unità statistiche osservate. Può essere espressa in percentuale.

➤ **Esempio:**

▪ Variabile 'voto' che può assumere i valori insufficiente, sufficiente, buono, ottimo.

▪ Valori osservati per la variabile voto:

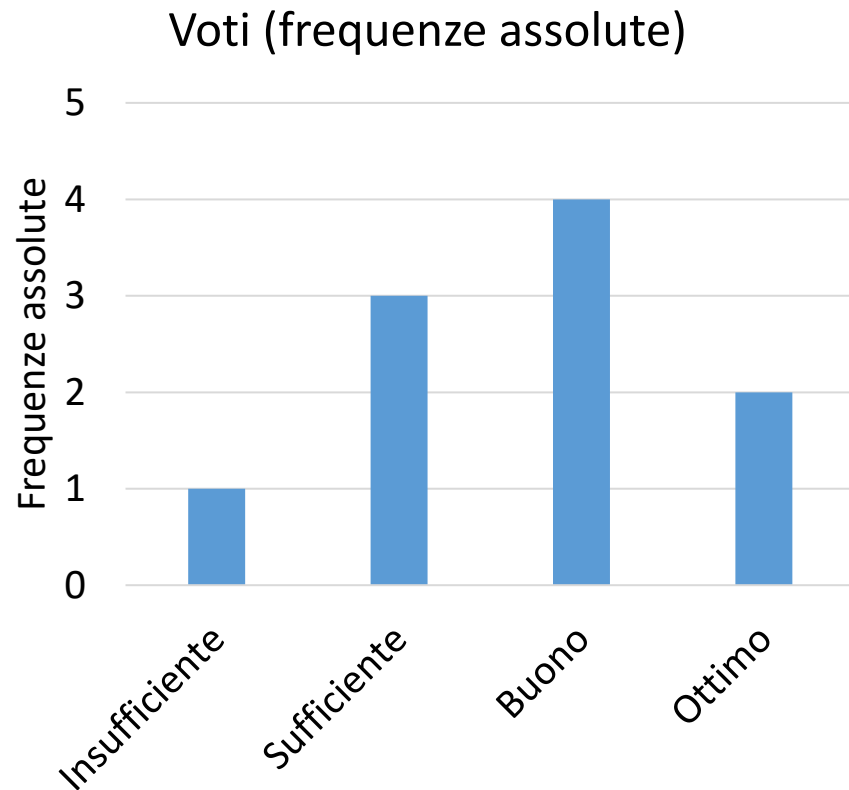
- Ottimo
- Buono
- Buono
- Insufficiente
- Ottimo
- Sufficiente
- Sufficiente
- Sufficiente
- Buono
- Buono



**Tabella di frequenza**

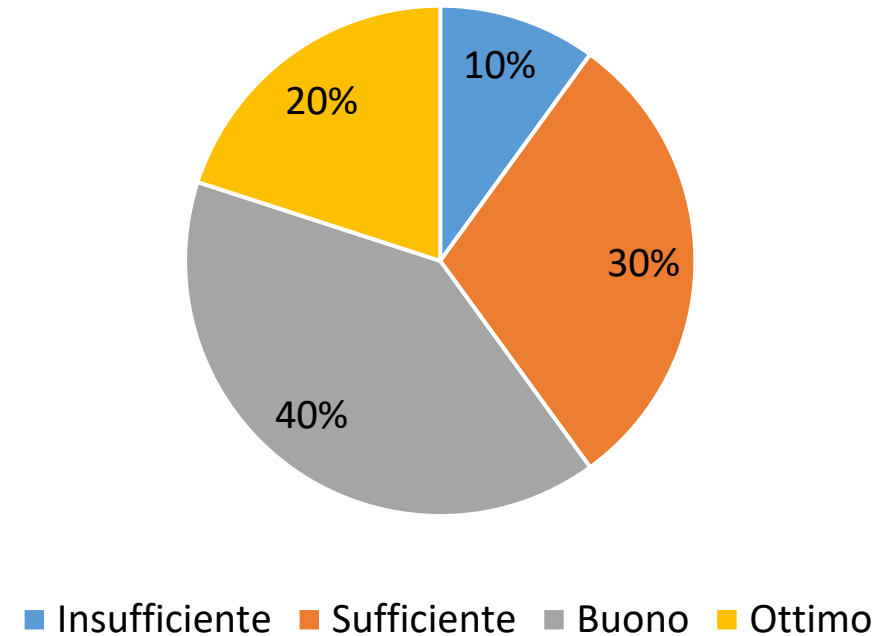
| Valore        | Frequenza<br>relativa | Frequenza<br>relativa [%] |
|---------------|-----------------------|---------------------------|
| Insufficiente | 1/10                  | 10%                       |
| Sufficiente   | 3/10                  | 30%                       |
| Buono         | 4/10                  | 40%                       |
| Ottimo        | 2/10                  | 20%                       |

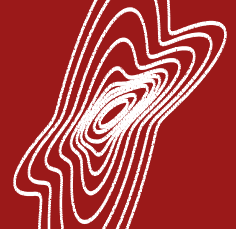
## ➤ Grafico a barre



## ➤ Diagramma a torta

Voti (frequenze relative in %)





# CLASSI DI FREQUENZA

- Quando si vogliono analizzare le frequenze di una variabile quantitativa continua, è utile definire delle classi di frequenza.
- Il range di valori assunti dalla variabile viene diviso in intervalli, detti classi, e per ciascun intervallo si calcolano le frequenze.

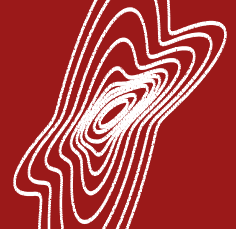
- Esempio:

- Indice di massa corporea:

- 26,0
    - 28,5
    - 35,0
    - 18,7
    - 16,9
    - 19,3
    - 24,2
    - 23,0
    - 27,1
    - 20,2

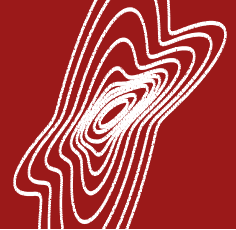
**Tabella di frequenza**

| Valore                      | Frequenza assoluta | Frequenza relativa | Frequenza relativa [%] |
|-----------------------------|--------------------|--------------------|------------------------|
| <18,5<br>(sottopeso)        | 1                  | 1/10               | 10%                    |
| 18,5 - 24,9<br>(normopeso)  | 5                  | 5/10               | 50%                    |
| 25,0 - 29,9<br>(sovrappeso) | 3                  | 3/10               | 30%                    |
| ≥30,0<br>(obesità)          | 2                  | 2/10               | 10%                    |



# SCELTA DELLE CLASSI

- Sulla base di criteri di classificazione esistenti (come nell'esempio dell'indice di massa corporea).
- Oppure, suddividendo il range osservato in un certo numero di intervalli di ampiezza costante.
  - Quanti intervalli? Scegliere in base ai dati osservati.
    - Troppi intervalli: pochi dati per ogni intervallo, non riusciremo a rappresentare in modo efficace la distribuzione dei dati.
    - Troppo pochi intervalli: si perde l'informazione sulla posizione che i dati avevano all'interno degli intervalli di classe.

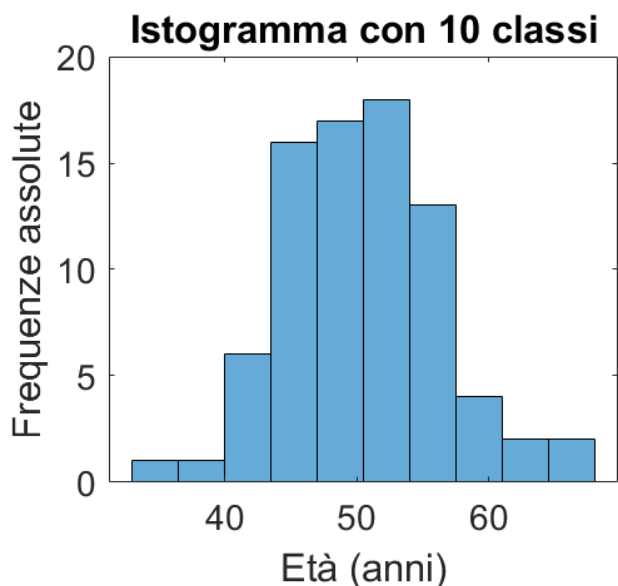


# ISTOGRAMMA

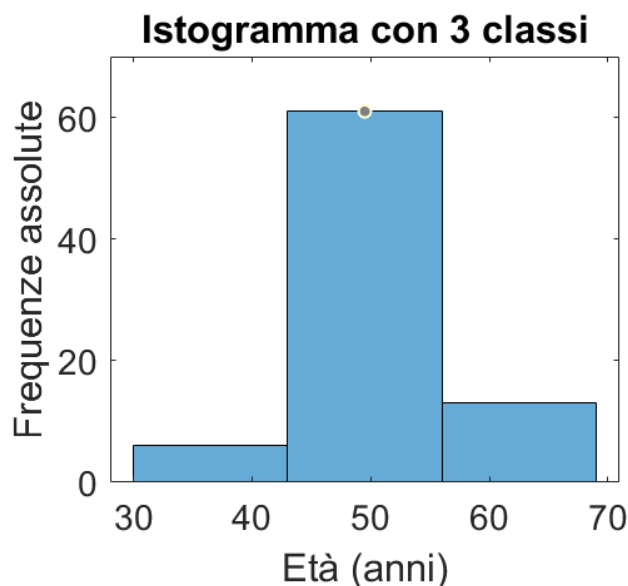
- **Istogramma:** il grafico a barre delle classi di frequenza. Le classi, o intervalli, in cui è diviso l'asse delle ascisse sono anche dette bin.

Esempio:

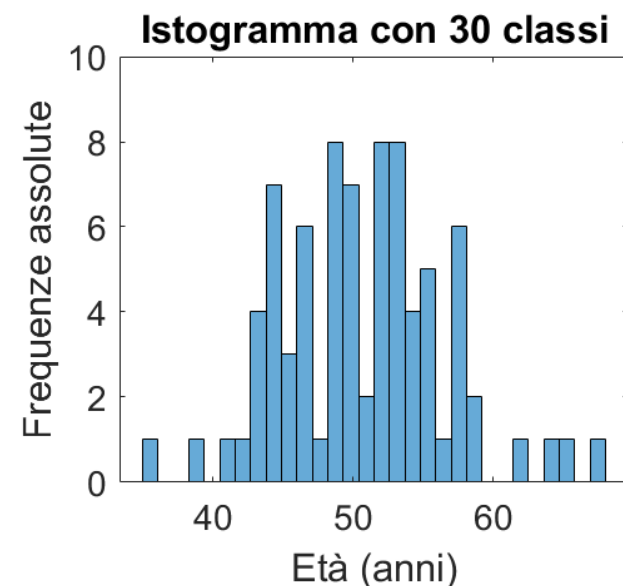
80 valori osservati per la variabile età: 53, 59, 39, 54, 52, 43, 48, 52, 68, 64, 43, 65, 54, 50, 54, 49, 49, 57, 57, 57, 53, 44, 54, 58, 52, 55, 54, 48, 51, 46, 54, 44, 45, 46, 35, 57, 52, 46, 57, 41, 49, 49, 52, 52, 46, 50, 49, 53, 55, 56, 46, 50, 44, 44, 50, 58, 46, 52, 49, 56, 45, 50, 53, 56, 58, 50, 43, 46, 45, 62, 47, 54, 49, 54, 46, 43, 43, 52, 49, 49.



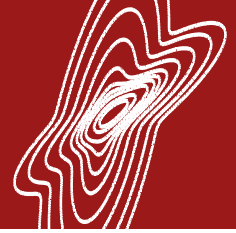
OK



TROPPO POCHE CLASSI



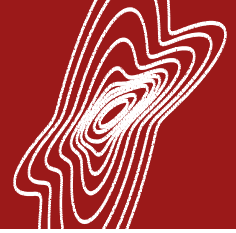
TROPPE CLASSI



# INDICI DI POSIZIONE



- A partire dai dati possiamo calcolare delle **statistiche sintetiche**, ovvero grandezze che rappresentano alcune caratteristiche dei dati.
- Gli **indici di posizione** rappresentano il «centro» dell'insieme di dati, ovvero un valore attorno al quale i dati si distribuiscono.
  - Media campionaria
  - Mediana campionaria
  - Moda campionaria



# MEDIA CAMPIONARIA - DEFINIZIONE



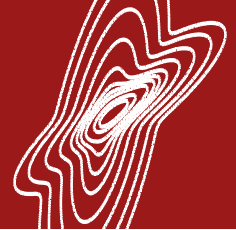
- Supponiamo di avere un insieme di  $n$  dati  $x_1, x_2, \dots, x_n$
- Si dice media campionaria e si denota  $\bar{x}$  la quantità, la media aritmetica dei dati:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$

Esempio:

Voti degli esami sostenuti da Mario Rossi: 24, 26, 30, 24, 28, 30, 26, 24, 27, 26

Media campionaria:  $\bar{x} = \frac{1}{10} \cdot 265 = 26.5$



# MEDIA CAMPIONARIA – CALCOLO BASATO SULLE FREQUENZE



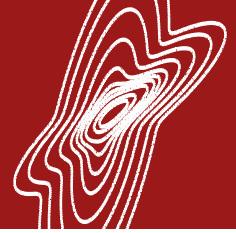
- Dati:  $k$  valori distinti  $v_1, v_2, \dots, v_k$ , con frequenze assolute  $F_1, F_2, \dots, F_k$  e corrispondenti frequenze relative  $f_1, f_2, \dots, f_k$
- Media campionaria: media pesata o ponderata dei valori assunti dai dati, usando come pesi le frequenze relative (cf. formula del valore atteso).

$$\bar{x} := \frac{1}{n} \sum_{i=1}^k F_i \cdot v_i = \sum_{i=1}^k \frac{F_i}{n} \cdot v_i = \sum_{i=1}^k f_i \cdot v_i$$

Esempio:

Voti degli esami sostenuti da Mario Rossi: 24, 26, 30, 24, 28, 30, 26, 24, 27, 26

Media campionaria:  $\bar{x} = \frac{1}{10} \cdot (24 \cdot 3 + 26 \cdot 3 + 27 \cdot 1 + 28 \cdot 1 + 30 \cdot 2) = 26.5$



# MEDIA ARITMETICA VS. MEDIA PONDERATA

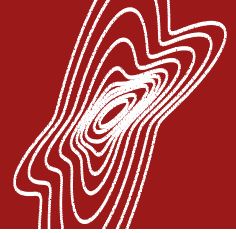


➤ **Media aritmetica** di  $n$  valori  $x_1, x_2, \dots, x_n$ :

$$M_a = \frac{\sum_{i=1}^n x_i}{n}$$

➤ **Media ponderata** di  $k$  valori  $v_1, v_2, \dots, v_k$ :

$$M_p = \frac{\sum_{i=1}^k w_i \cdot v_i}{\sum_{i=1}^k w_i}$$



# PROPRIETA' DELLA MEDIA CAMPIONARIA

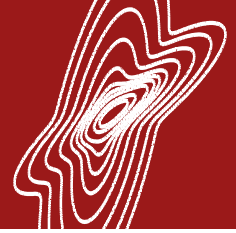
- La media campionaria di una trasformazione lineare è la trasformazione lineare della media campionaria (cf. linearità del valore atteso).

$$y_i := a \cdot x_i + b, \quad i = 1, \dots, n$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n (a \cdot x_i + b) = \frac{1}{n} \sum_{i=1}^n (a \cdot x_i) + \frac{1}{n} \sum_{i=1}^n b = a \cdot \bar{x} + b$$

- La somma degli scarti dalla media (con il loro segno) è 0:

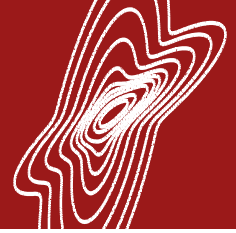
$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$



# MEDIANA CAMPIONARIA



- **Mediana:** il valore centrale una volta che i dati sono stati ordinati dal più piccolo al più grande.
- Definizione formale:  
Dato un insieme di dati di numerosità  $\underline{n}$ , lo si ordina dal minore al maggiore.  
La **mediana campionaria** è:
  - Se  $n$  è dispari, il valore del dato in posizione  $(\underline{n}+1)/2$  (quello in centro);
  - se  $n$  è pari, la media aritmetica tra i valori dei dati in posizione  $\underline{n}/2$  e  $\underline{n}/2+1$  (i due valori centrali).



# MEDIANA CAMPIONARIA - ESEMPI

## ➤ Esempio 1:

Voti degli esami: 24, 26, 30, 24, 28, 30, 26, 24, 27, 27, 29

Voti ordinati: 24, 24, 24, 26, 26, 27, 27, 28, 29, 30, 30

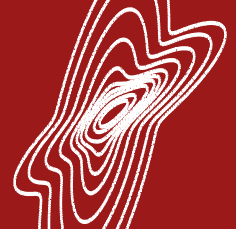
$n = 11 \rightarrow n$  è dispari  $\rightarrow$  la mediana è il valore in posizione 6  $\rightarrow$  mediana = 27

## ➤ Esempio 2:

Voti degli esami: 24, 26, 30, 24, 28, 30, 26, 24, 27, 27

Voti ordinati: 24, 24, 24, 26, 26, 27, 27, 28, 30, 30

$n = 10 \rightarrow n$  è pari  $\rightarrow$  la mediana è la media aritmetica tra i voti in posizione 5 e 6  
 $\rightarrow$  mediana =  $(26+27)/2 = 26.5$



# MEDIA VS. MEDIANA

- La media è calcolata usando i valori di tutti i dati → è influenzata in maniera sensibile da valori eccezionalmente alti o bassi, i cosiddetti *outlier*.
- La mediana fa leva sull'ordine delle osservazioni (interessano quelle che, in ordine, sono intorno alla metà) e non sul loro valore (il valore che entra il gioco è solo quello di uno o due campioni) → non risente particolarmente dei valori estremi.

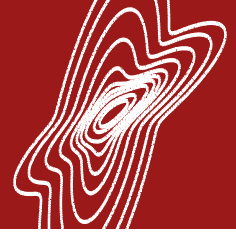
- Esempio:

- Valori di pressione sanguigna sistolica:

105, 108, 109, 110, 110, 111, 112, 113, 115, 150, 160 mmHg

media = 118.45 mmHg

mediana = 111 mmHg



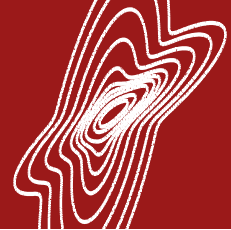
# MODA CAMPIONARIA

- La **moda campionaria** di un insieme di dati, se esiste, è l'unico valore che assume frequenza massima.
- Se i valori con frequenza massima sono più di uno, essi vengono detti **valori modali**.
- La moda, a differenza di media e mediana, ha senso solo per variabili discrete che assumono solo un numero limitato di valori.
- Esempio:

Voti degli esami: 24, 26, 30, 24, 28, 30, 26, 24, 27, 27, 29

Moda = 24 ←

| Voto | Frequenza assoluta |
|------|--------------------|
| 24   | 3                  |
| 26   | 2                  |
| 27   | 2                  |
| 28   | 1                  |
| 29   | 1                  |
| 30   | 2                  |

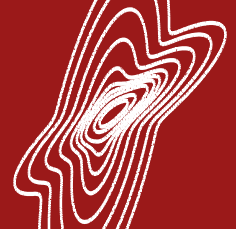


## Esempio:

Stipendio netto mensile dei dipendenti dell'azienda XYZ

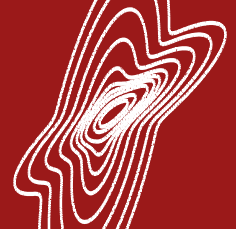
| Stipendio<br>[euro] | Frequenza assoluta<br>[n. dipendenti] |
|---------------------|---------------------------------------|
| 1300                | 2                                     |
| 1700                | 22                                    |
| 2200                | 19                                    |
| 2600                | 3                                     |
| 6500                | 2                                     |
| 9400                | 1                                     |
| 23000               | 1                                     |

- **Media = 2700 euro**
  - Se tutti i dipendenti avessero lo stesso stipendio netto, questo sarebbe di 2700 euro.
- **Mediana = 2200 euro**
  - Circa metà dei dipendenti percepisce meno di 2200 euro, e circa metà più di 2200 euro.
- **Moda = 1700 euro**
  - Lo stipendio netto più comune è 1700 euro.



# INDICI DI DISPERSIONE

- **Gli indici di dispersione** rappresentano quanto i dati sono concentrati o dispersi attorno ai valori tipici (rappresentati dagli indici di posizione).
  - Campo di variazione o range
  - Devianza (*sum of squares*)
  - Varianza campionaria
  - Deviazione standard campionaria
  - Coefficiente di variazione

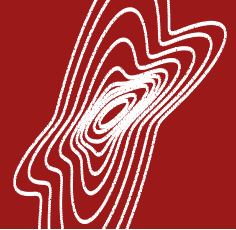


# CAMPO DI VARIAZIONE O RANGE

- Campo di variazione o range: la differenza tra il massimo ed il minimo valore nei dati.

$$\begin{array}{l} x_{\min} = \min(x_1, x_2, \dots, x_n) \\ x_{\max} = \max(x_1, x_2, \dots, x_n) \end{array} \quad \longrightarrow \quad \text{range} = x_{\max} - x_{\min}$$

- Nota 1: Il range è influenzato solo dai valori estremi dei dati, non tiene in considerazione come sono distribuiti gli altri dati.
- Nota 2: a volte, con abuso di notazione, si indica come range l'intervallo  $[x_{\min}, x_{\max}]$ .



# DEVIANZA, VARIANZA E DEVIAZIONE STANDARD CAMPIONARIA



Consideriamo un insieme di  $n$  dati  $x_1, x_2, \dots, x_n$ . I seguenti indici di variabilità ci dicono quanto i dati sono dispersi attorno alla media campionaria  $\bar{x}$ .

➤ **Devianza (*sum of squares*):**

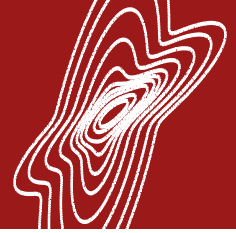
$$D := \sum_{i=1}^n (x_i - \bar{x})^2$$

➤ **Varianza campionaria:**

$$s^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

➤ **Deviazione standard campionaria:**

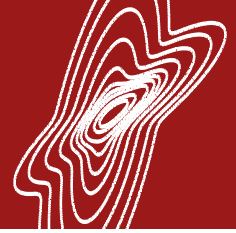
$$s := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$



# DEVIANZA, VARIANZA E DEVIAZIONE STANDARD CAMPIONARIA - NOTE



- La devianza in inglese si chiama *sum of squares*. Il termine inglese *deviance* viene utilizzato in statistica per indicare un'altra quantità.
- Gli stimatori di varianza e deviazione standard campionaria riportati nella slide precedente si dicono stimatori corretti.
- Esistono anche degli stimatori della varianza e della deviazione standard in cui la divisione viene fatta per  $n$  anziché per  $n - 1$ . Questi stimatori forniscono tuttavia delle stime distorte.



# DEVIANZA, VARIANZA E DEVIAZIONE STANDARD CAMPIONARIA - ESEMPIO



Esempio: Consideriamo i seguenti due insiemi di dati aventi la stessa media.

A: 3, 4, 6, 7, 10

B: -20, 5, 15, 24

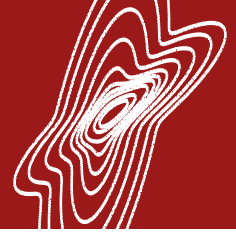
I dati in A e in B presentano la stessa media pari a 6. Tuttavia, i dati in B risultano molto più dispersi attorno alla media, rispetto ai dati in A. Cosa ci dicono gli indici di variabilità?

Insieme A:

- $D = (3-6)^2 + (4-6)^2 + (6-6)^2 + (7-6)^2 + (10-6)^2 = 9 + 4 + 0 + 1 + 16 = 30$
- $S^2 = D/(5-1) = 30/4 = 7.5$
- $S = \sqrt{7.5} = 2.74$

Insieme B:

- $D = (-20-6)^2 + (5-6)^2 + (15-6)^2 + (24-6)^2 = 676 + 1 + 81 + 324 = 1082$
- $S^2 = D/(4-1) = 1082/3 = 360.67$
- $S = \sqrt{360.67} = 18.99$



# DEVIANZA, VARIANZA E DEVIAZIONE STANDARD CAMPIONARIA – PROPRIETA'



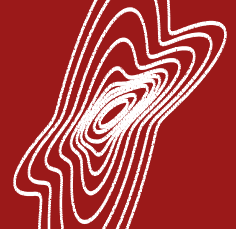
- Devianza, varianza e deviazione standard campionaria sono indici non negativi ed valgono 0 solo quando tutti i valori sono uguali.

- La devianza può essere calcolata anche come:

$$D = \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2$$

- Se  $y_i = a \cdot x_i + b$  con  $i=1, \dots, n$ , e  $D$ ,  $s^2$  e  $s$  sono rispettivamente la devianza, la varianza campionaria e la deviazione standard campionaria dei dati  $x_i$ ,  $i=1, \dots, n$ , allora:

- Devianza di  $y_i$   $i=1, \dots, n \rightarrow a^2 \cdot D$
- Varianza campionaria di  $y_i$   $i=1, \dots, n \rightarrow a^2 \cdot s^2$
- Deviazione standard campionaria di  $y_i$   $i=1, \dots, n \rightarrow |a| \cdot s$



# COEFFICIENTE DI VARIAZIONE

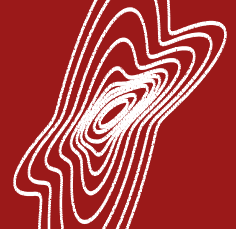
- Dato un insieme di  $n$  dati  $x_1, x_2, \dots, x_n$  con media  $\bar{x} \neq 0$  e deviazione standard  $s$ , si dice coefficiente di variazione, CV:

$$CV = \frac{s}{|\bar{x}|}$$

- Il CV è normalmente è espresso in %:

$$CV\% = \frac{s}{|\bar{x}|} \cdot 100 \%$$

- Esso quantifica quanto è grande la deviazione standard rispetto alla media.



# COEFFICIENTE DI VARIAZIONE - ESEMPIO

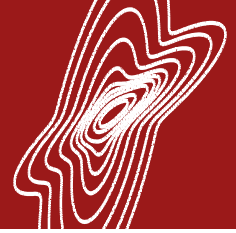
## Esempio:

Consideriamo 3 neonati con peso 3, 4 e 5 Kg (media: 4 Kg, deviazione standard: 1 Kg) e 3 bambini con peso 10, 11 e 12 Kg (media: 11 Kg, deviazione standard: 1 Kg).

E' maggiore la variabilità del peso dei neonati o quella dei bambini?

La deviazione standard è la stessa nei due casi, pari a 1 Kg. Tuttavia, il gruppo dei neonati presenta maggiore variabilità del peso in rapporto alla sua media. Questo risulta evidente calcolando il coefficiente di variazione.

- Neonati:  $CV = 1/4 = 0.25 \rightarrow 25\%$
- Bambini:  $CV = 1/11 = 0.09 \rightarrow 9\%$

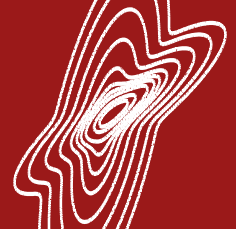


# PERCENTILI CAMPIONARI



- Sia  $k$  un numero intero tra 0 e 100 inclusi. Il  **$k$ -esimo percentile** di un insieme di dati è quel valore che risulta contemporaneamente maggiore o uguale ad almeno una percentuale  $k$  dei dati e minore o uguale ad almeno una percentuale  $100-k$  dei dati.
- Più nello specifico, le due condizioni possono essere verificate contemporaneamente da un solo valore distinto presente nei dati oppure da due:
  - Se un unico valore presente nei dati soddisfa le due condizioni, il  $k$ -esimo percentile coincide con quel valore.
  - Se due valori presenti nei dati soddisfano le condizioni, il  $k$ -esimo percentile è la loro media aritmetica.

NB: nel caso di valori ripetuti nell'insieme dei dati, le regole rimangono invariate con la precisazione che le percentuali  $k$  e  $100-k$  vengono calcolate considerando il totale dei dati, cioè contando tutte le ripetizioni.

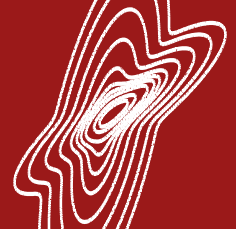


# CALCOLO DEI PERCENTILI CAMPIONARI



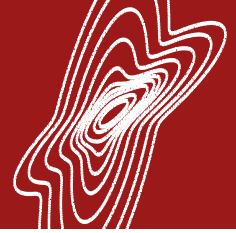
Per calcolare il k-esimo percentile di un insieme di  $n$  dati procediamo come segue.

1. Ordiniamo i dati in senso crescente.
2. Calcoliamo  $p = k/100$  e il prodotto  $n \cdot p$ .
3. Se  $n \cdot p$  non è intero  $\rightarrow$  arrotondiamo  $n \cdot p$  all'intero successivo e il percentile k-esimo è il dato in quella posizione.
4. Se  $n \cdot p$  è intero  $\rightarrow$  il percentile k-esimo è la media tra i dati in posizione  $n \cdot p$  e  $n \cdot p + 1$ .



# QUARTILI

- Il **25-esimo percentile** di un campione di dati viene detto **primo quartile (Q1)**.
- Il **50-esimo percentile** di un campione di dati è la sua **mediana campionaria** e viene anche detto **secondo quartile (Q2)**.
- Il **75-esimo percentile** di un campione di dati viene detto **terzo quartile (Q3)**.
- I quartili dividono i dati in quattro parti, ciascuna contenente circa il 25% dei dati.
- La differenza tra il terzo e il primo quartile viene detta **range interquartile (IQR)**.
- A volte si indica come range interquartile l'intervallo  $[Q1, Q3]$ .



# PERCENTILI CAMPIONARI - ESEMPIO

Esempio:

Consideriamo l'insieme di 10 dati seguente:

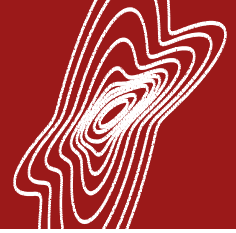
3, 5, 6, 11, 15, 16, 18, 20, 21, 25

Troviamo il 25-esimo percentile.

➤  $p=0.25$ ,  $n \cdot p = 2.5 \rightarrow$  il 25-esimo percentile è il dato in posizione 3  $\rightarrow 6$

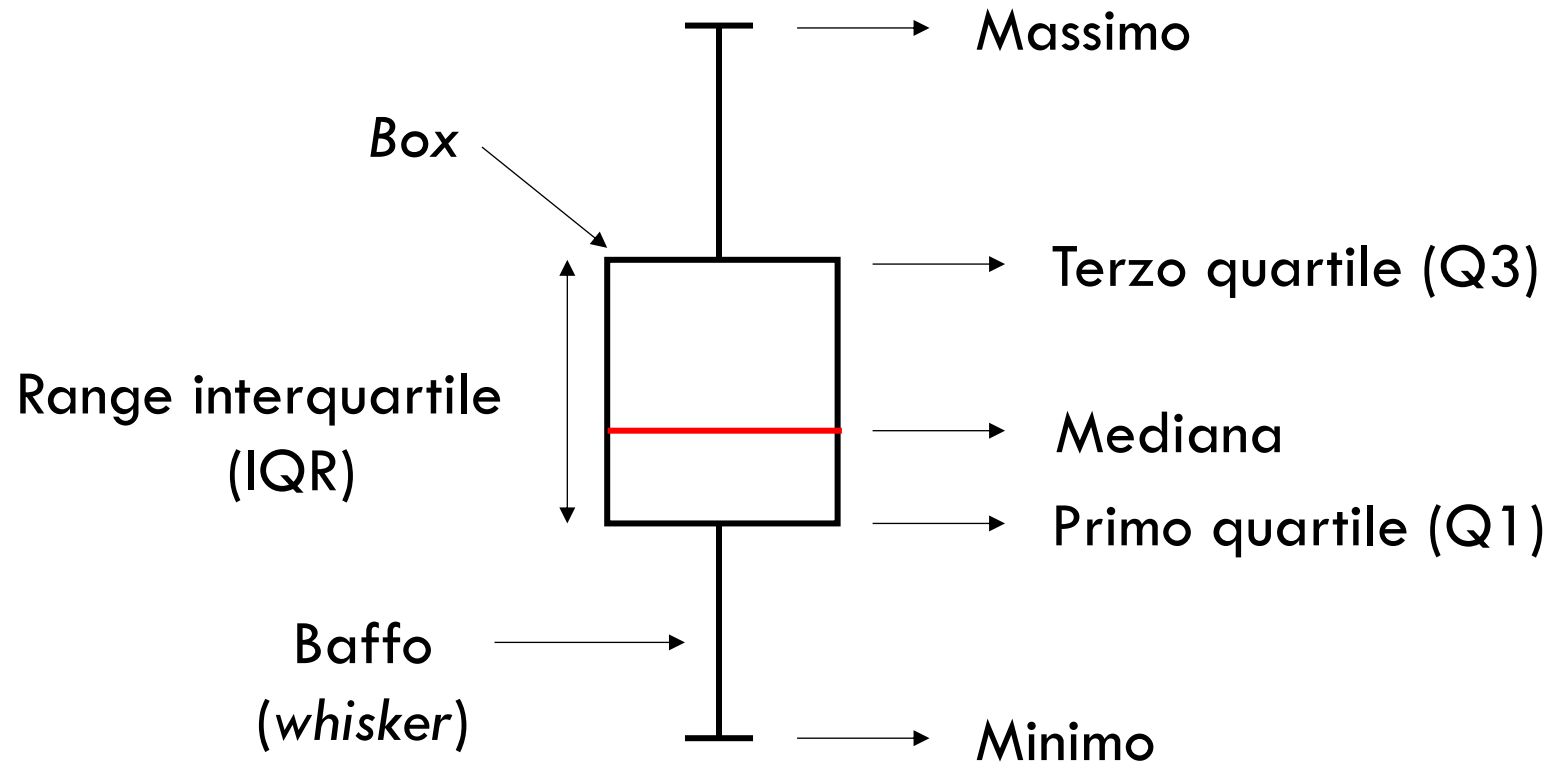
Troviamo il 40-esimo percentile.

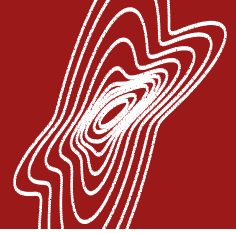
➤  $p=0.40$ ,  $n \cdot p = 4 \rightarrow$  il 40-esimo percentile è la media tra i dati in posizione 4 e 5  $\rightarrow (11+15)/2=13$



# BOXPLOT

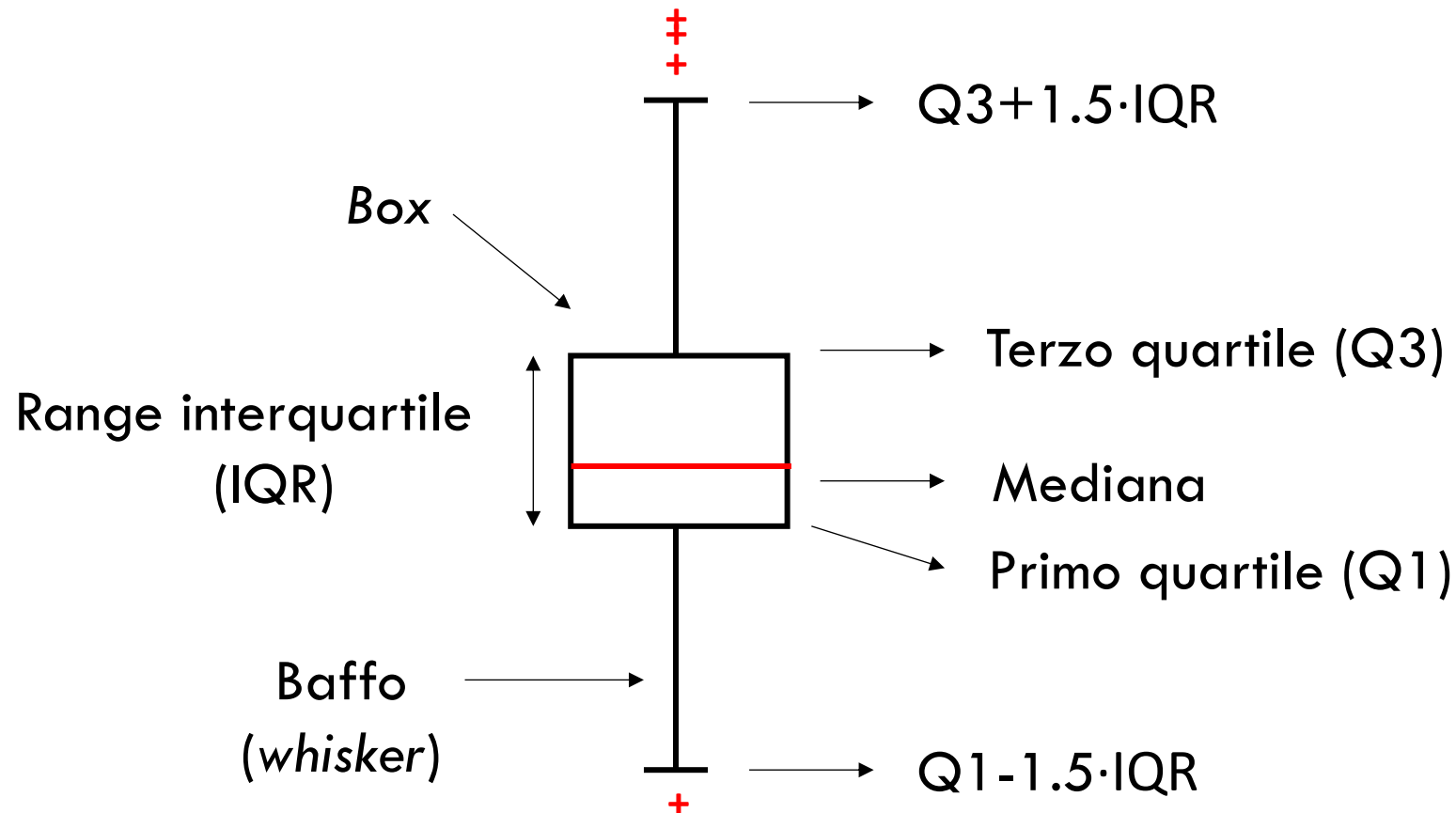
- Il boxplot è un tipo di grafico che consente di visualizzare alcune delle statistiche rappresentative della distribuzione di un insieme di dati.

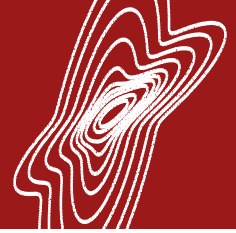




# BOXPLOT CON OUTLIER

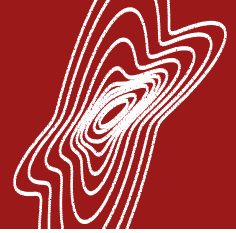
- Eventuali *outlier*, ovvero valori estremi nella distribuzione dei dati, vengono rappresentati oltre il limite dei baffi.





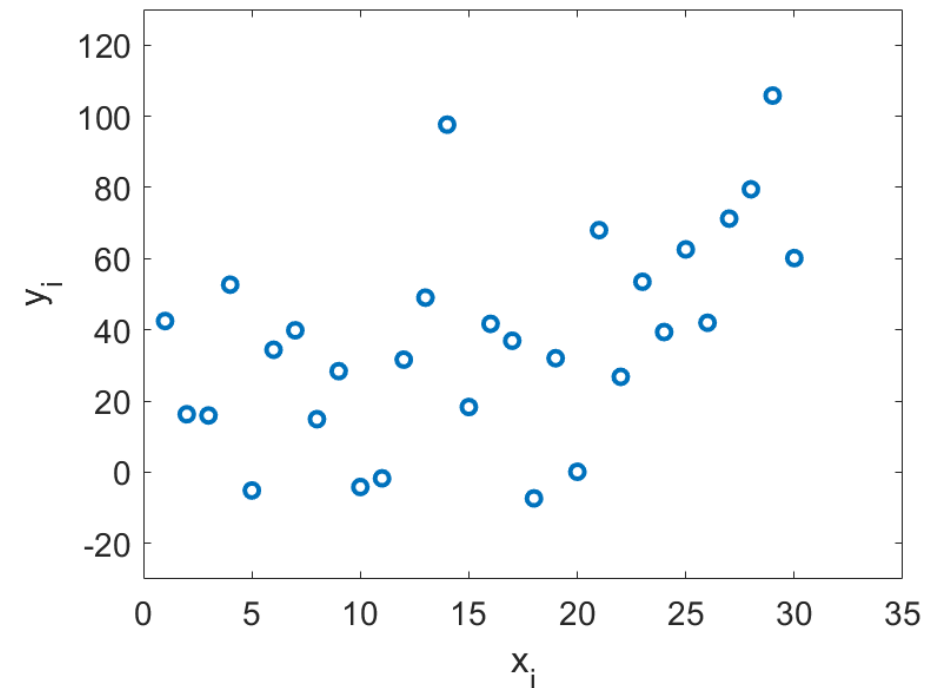
# NOTA SUI BOXPLOT

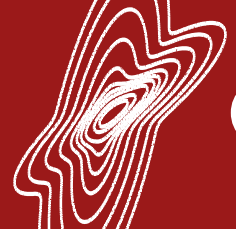
- Esistono anche rappresentazioni alternative meno diffuse dei boxplot (ad esempio, in cui viene rappresentata la media al posto della mediana, l'intervallo  $\text{media} \pm \text{deviazione standard}$  al posto del range interquartile, ecc.).
- Fate sempre attenzione alla definizione di boxplot che state considerando o che state osservando.



# CAMPIONI BIVARIATI

- Spesso vengono analizzate coppie di dati tra i quali esiste una qualche relazione.
- Esempio: vogliamo analizzare congiuntamente la temperatura ambientale e il numero di accessi al pronto soccorso e siamo interessati a capire se esiste una qualche relazione tra le due quantità (al crescere di una quantità aumenta anche l'altra? Il contrario?)
- **Campione bivariato:** un insieme di  $n$  coppie di dati  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .
- Un campione bivariato può essere rappresentato attraverso un **diagramma di dispersione** (*scatterplot*).



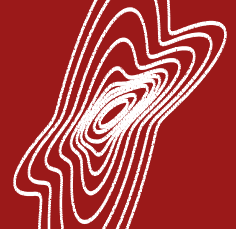


Dato un campione bivariato  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , si dice **coefficiente di correlazione campionaria** (di Pearson) la quantità:

$$r := \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{x,y}}{s_x s_y}$$

dove:

- $s_x$  è la deviazione standard campionaria di  $x_i$ ,  $i = 1, \dots, n$ ;
- $s_y$  è la deviazione standard campionaria di  $y_i$ ,  $i = 1, \dots, n$ ;
- $s_{xy}$  è la cosiddetta **covarianza campionaria** di  $x_i$  e  $y_i$ ,  $i = 1, \dots, n$ .



# PROPRIETA' DEL COEFFICIENTE DI CORRELAZIONE CAMPIONARIA (1 / 2)

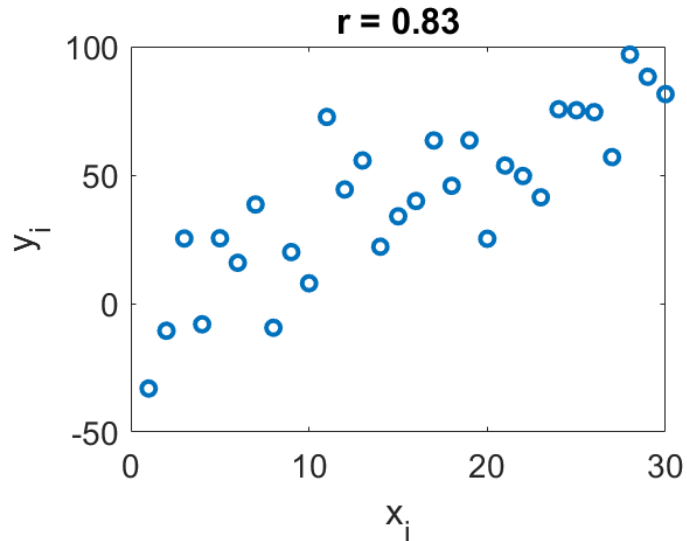


- $-1 \leq r \leq 1$
- Se  $r > 0 \rightarrow$  i dati  $x_i$  e  $y_i$  sono correlati positivamente, ovvero al crescere (decrescere) di  $x_i$ , cresce (decresce)  $y_i$  (e viceversa).
- Se  $r < 0 \rightarrow$  i dati  $x_i$  e  $y_i$  sono correlati negativamente, ovvero al crescere (decrescere) di  $x_i$ , decresce (cresce)  $y_i$  (e viceversa).
- Se  $r = 0 \rightarrow$  i dati non sono correlati, ovvero al crescere (decrescere) di  $x_i$ ,  $y_i$  varia in modo casuale (e viceversa).

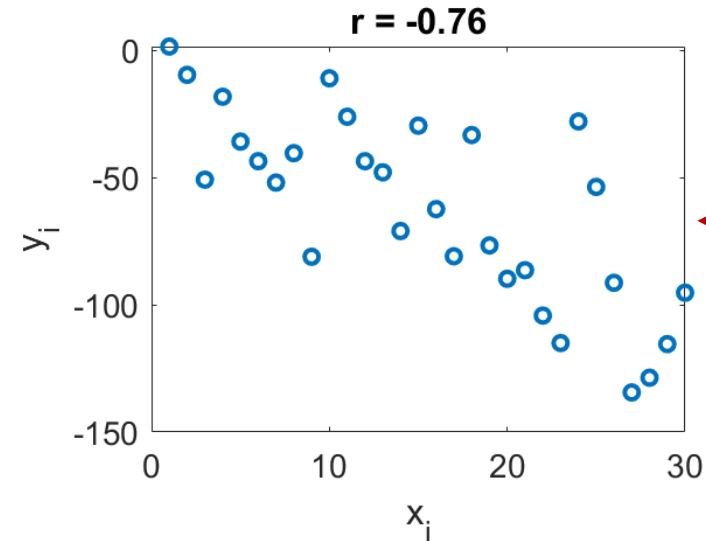


# ESEMPI DI DATI CORRELATI

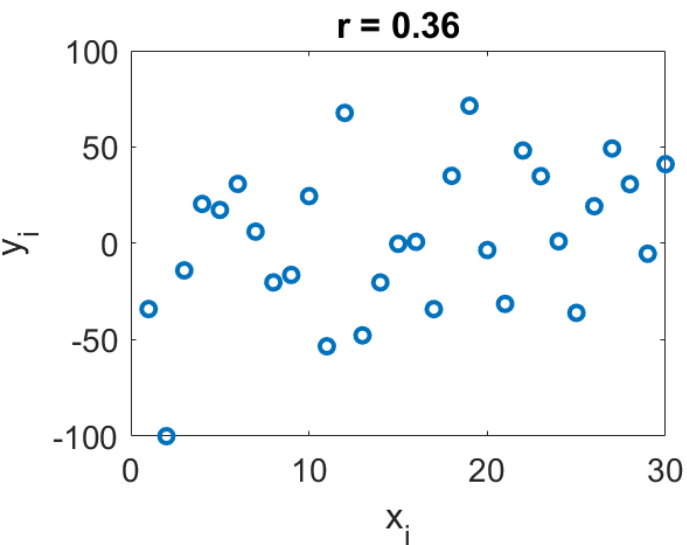
Forte correlazione  
positiva



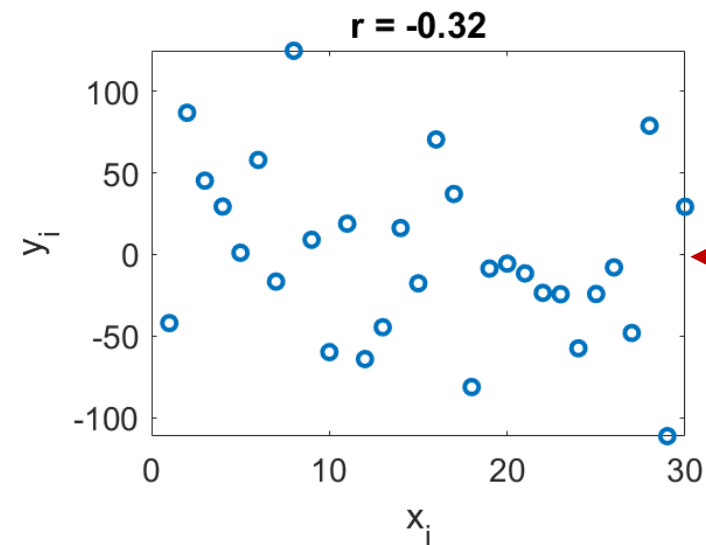
Forte correlazione  
negativa

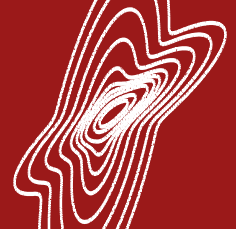


Debole correlazione  
positiva

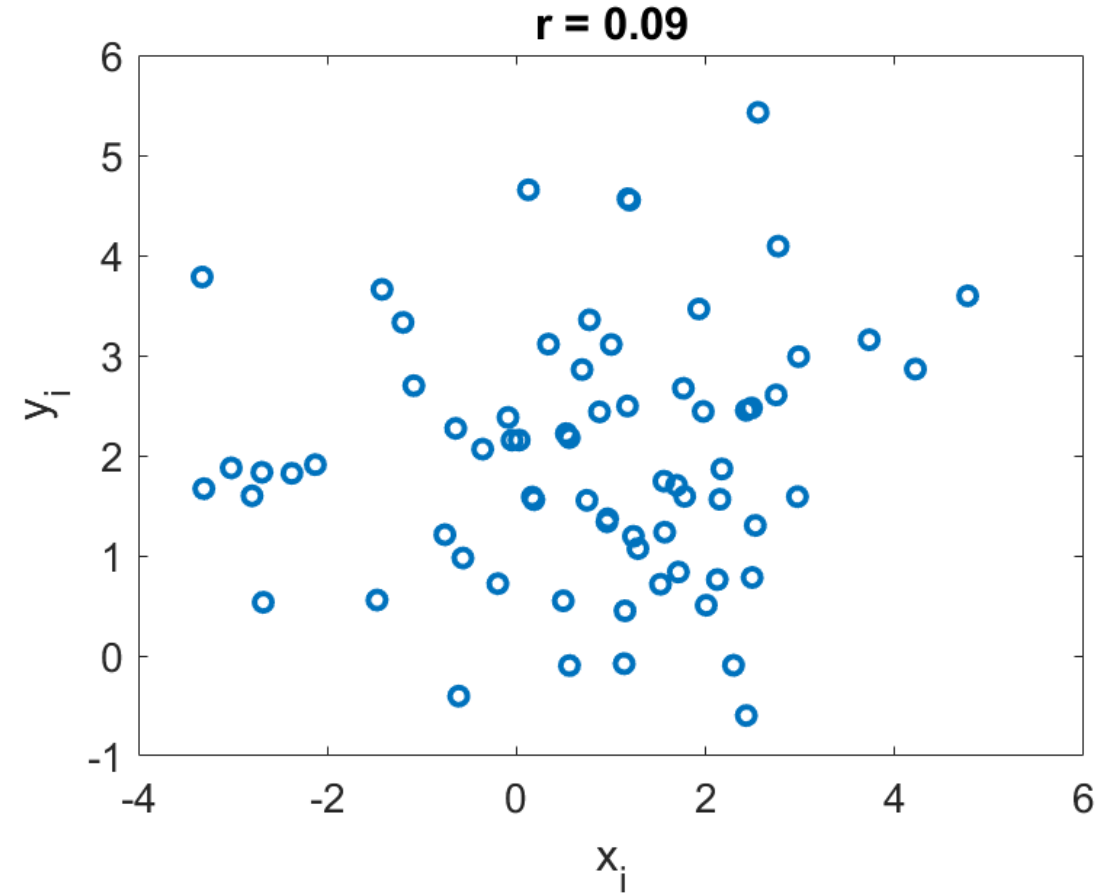


Debole correlazione  
negativa

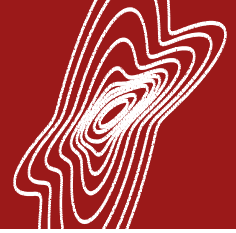




# ESEMPIO DI DATI SCORRELATI



Dati scorrelati



# PROPRIETA' DEL COEFFICIENTE DI CORRELAZIONE CAMPIONARIA (2/2)

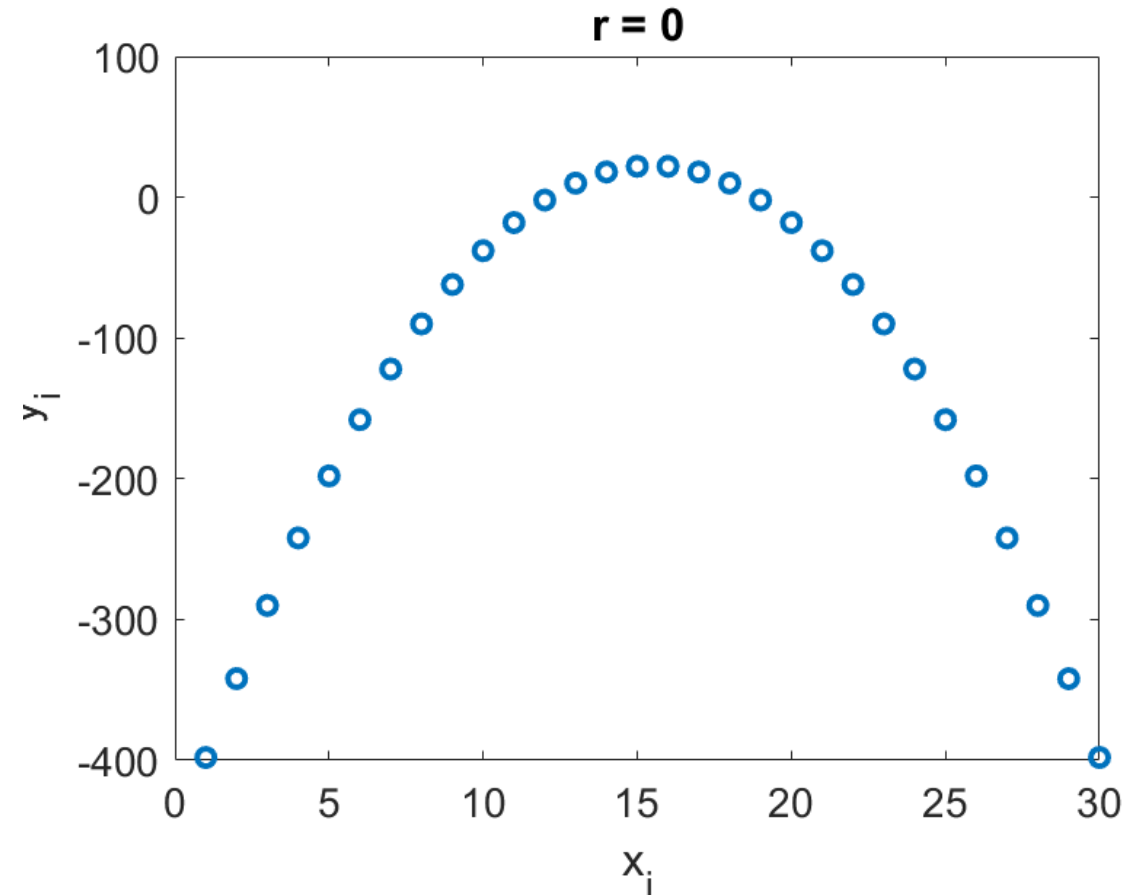
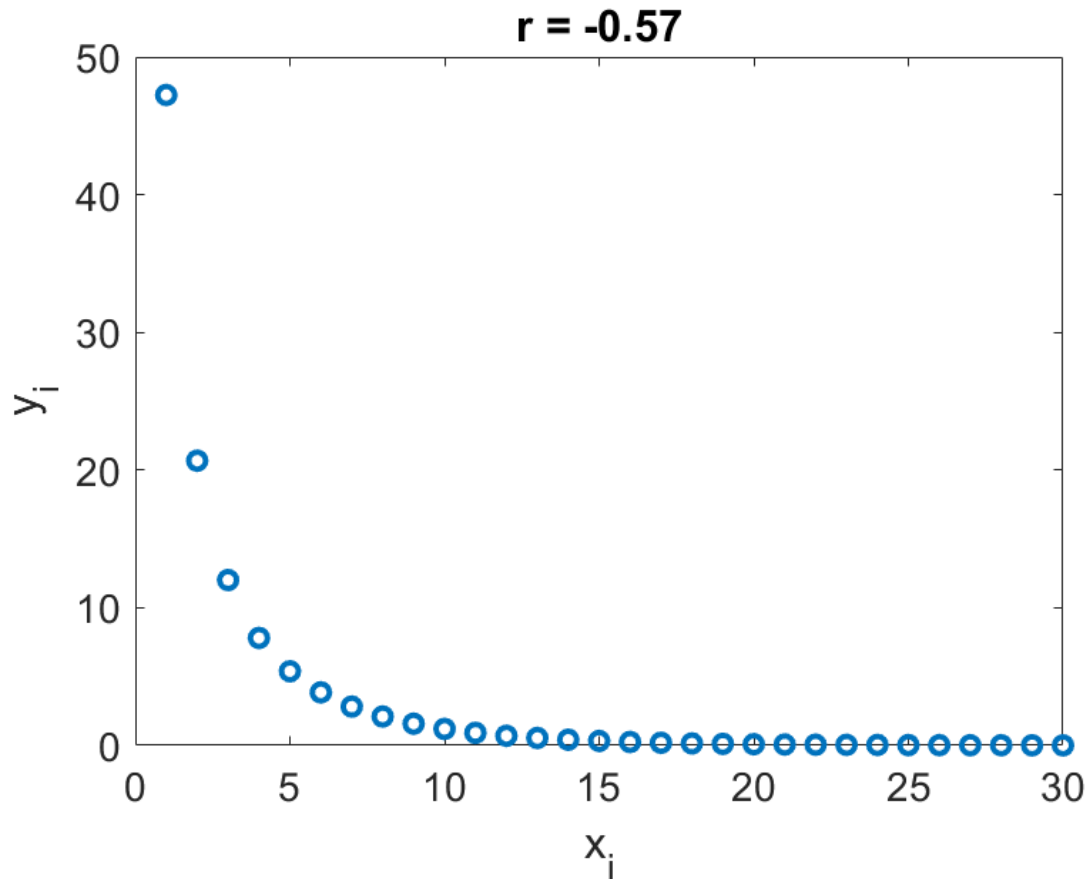


Il coefficiente  $r$  misura il grado di **correlazione lineare**, ovvero quanto i dati tendono a distribuirsi attorno ad una retta.

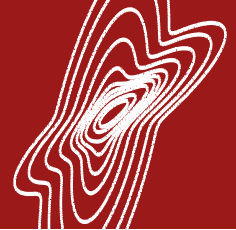
- $r=1$  quando tra  $x_i$  e  $y_i$  sussiste una relazione perfettamente lineare del tipo  $y_i = a \cdot x_i + b$  con  $a > 0$  (e viceversa).
- $r=-1$  quando tra  $x_i$  e  $y_i$  sussiste una relazione perfettamente lineare del tipo  $y_i = a \cdot x_i + b$  con  $a < 0$  (e viceversa).
- Maggiore è il valore assoluto di  $r$ , maggiore è l'intensità della relazione tra  $x_i$ ,  $y_i$ , ovvero il grado di correlazione.

**Nota:** il coefficiente di correlazione campionaria misura l'associazione tra due variabili, ma non denota un rapporto di causa-effetto.

# DATI NON LINEARMENTE CORRELATI



Il coefficiente di correlazione campionaria di Pearson non cattura relazioni di tipo non lineare tra i dati.



# COEFFICIENTE DI CORREAZIONE DI SPEARMAN

- Per catturare **relazioni di tipo non lineare** si può ricorrere al coefficiente di correlazione di Spearman → correlazione di Pearson calcolata tra i **ranghi** delle due grandezze in esame
- Esempio:

| x    | y    | Ranghi di x | Ranghi di y |
|------|------|-------------|-------------|
| 3.6  | 13.1 | 1           | 1           |
| 6.5  | 14.3 | 2           | 2           |
| 10.8 | 14.9 | 3           | 3           |
| 15.1 | 15.1 | 4           | 4           |
| 20.4 | 15.3 | 5           | 5           |
| 35.1 | 15.4 | 6           | 6           |
| 50.3 | 15.5 | 7           | 7           |
| 60.2 | 15.6 | 8           | 8           |

