FISICA APPLICATA E BIOSTATISTICA

MODULO MAT7 — Statistica: descrivere e confrontare i dati biomedici

Variabilità Biologica, Misure di Sintesi, Distribuzione Normale e Inferenza

Prof.ssa Valentina Notarstefano | Università degli Studi di Teramo

Perché la Statistica è importante per un biotecnologo

La biologia è variabile: ripetendo una misura si ottiene una distribuzione di valori, non un unico numero.

La statistica descrittiva riassume questa variabilità — media, mediana, deviazione standard, quartili.

La distribuzione normale trasforma una variabilità in un intervallo di riferimento (media ± 2 DS = 95% centrale).

La statistica inferenziale chiede se una differenza è reale: trattati vs controllo (il test t).

E ci dice quanto è buono un test diagnostico: la sua sensibilità e specificità.

Dati Numerici

- Se gli oggetti dei dati hanno lo stesso insieme fisso di attributi numerici, allora tali oggetti possono essere pensati come punti in uno spazio multidimensionale, in cui ogni dimensione rappresenta un attributo distinto

- Un tale insieme di dati può essere rappresentato da una matrice dei dati n-per-d, con n righe, una per ciascun oggetto, e d colonne, una per ciascun attributo

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Dati Categorici

- Dati costituiti da una raccolta di record, ciascuno formato da un insieme fisso di attributi categorici

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

Popolazione, Unità Statistica, Campione Statistico

- **Popolazione statistica**: insieme degli elementi a cui si riferisce l'indagine statistica:
  - opinioni degli americani riguardo a una nuova elezione presidenziale: tutti i cittadini statunitensi
  - geni sovraespressi negli individui affetti da obesità: tutti gli individui obesi
  - …

- **Unità statistica**: ciascun elemento della popolazione statistica, la più piccola unità da cui vengono raccolti i dati:
  - Un cittadino, un individuo….

- **Campione statistico (campione)**: un qualsiasi insieme di unità statistiche estratte dall'intera popolazione. Un campione è quindi un sottoinsieme di misurazioni selezionato dalla popolazione
  - 50 individui con condizioni legate all'obesità (selezionati casualmente).

Variabile Quantitativa e Variabile Qualitativa

- **Variabile quantitativa**: quando assume valori numerici:
  - Continua: assume valori continui all'interno di un intervallo (il peso e l'altezza di una persona, i livelli di intensità dei campioni su microarray, il livello di espressione genica, ecc.)
  - Discreta: assume valori discreti come il numero di campioni, il numero di geni sovraespressi, il numero di pazienti, ecc.

- **Variabile qualitativa**: quando assume valori non numerici
  - Ordinale: i dati sono ordinati (buono–medio–scarso, freddo–tiepido–caldo…)
  - Categoriale: maschio/femmina, fenotipo, gruppi di pazienti malati/sani, ecc.

Frequenze Assolute e Percentuali

- Quando il campione di cui si vogliono descrivere le variabili statistiche è molto ampio, anziché considerare tutti i valori si possono riportare solo i valori distinti, indicando per ciascuno **quante volte compare**.
- Sia definita una variabile statistica discreta. Definiamo **modalità** i valori distinti e **frequenza assoluta di una modalità** il numero di volte in cui essa è osservata nell'espressione della variabile statistica. $Y = (y_1, y_2, \dots, y_N)$ $y_1, \dots, y_N$

Y=(60, 80, 92, 100, 83, 96, 74, 63, 80, 100, 90, 75, 74, 92)

Modalità	60	80	92	100	83	96	74	63	90
Frequenza assoluta	1	2	2	2	1	1	2	1	1

Frequenze Assolute e Percentuali

Sia Y una variabile statistica e f la frequenza assoluta della modalità z. Definiamo **frequenza relativa** della modalità z il rapporto f/N, dove N è il numero di elementi del campione.
La **frequenza percentuale** si ottiene moltiplicando la frequenza relativa per 100.

Media

- Media di una popolazione: somma di tutti i valori delle variabili della popolazione divisa per il numero di unità della popolazione (N)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Dove N è il numero di elementi della popolazione, X_i è la i-esima osservazione della variabile X_i

- Media di un campione: somma di tutti i valori delle variabili di un sottoinsieme della popolazione divisa per il numero di unità di quel campione (n)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Dove n è il numero di elementi del campione, X_i è la i-esima osservazione della variabile X_i

Dato il seguente insieme di misurazioni dei livelli di espressione genica:

55.20	18.06	28.16	44.14	61.61	4.88	180.29	399.11	97.47	56.89	271.95	365.29	807.80

media della popolazione:

$$\mu = \frac{\sum_{i=1}^{13} 55.20 + 18.06 + 28.16 + 44.14 + 61.61 + \ldots + \ldots + 807.80}{13} = \frac{2390{,}85}{13} = 183.9115$$

media del campione: (55.20; 18.06; 28.16; 44.14):

$$\overline{X} = \frac{55.20 + 18.06 + 28.16 + 44.14}{4} = \frac{145.56}{4} = 36.39$$

La media di un qualsiasi campione $\bar{X}$ può differire sensibilmente da quella dell'intera popolazione μ. Più grande è il campione, più la media campionaria sarà vicina alla media della popolazione

Media

- Media ponderata di una popolazione: a ciascuna variabile è assegnato un peso; tutti i valori delle variabili vengono sommati, ciascuno moltiplicato per il proprio peso, e il totale ottenuto è diviso per la somma dei pesi

$$\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$$

Esercizio 1 – Media aritmetica e ponderata

a) Livello di espressione di un gene (unità relative) in n = 8 campioni:

12, 15, 11, 14, 18, 13, 16, 13

Calcolare la media campionaria.

b) L'HbA1c (%) è misurata in tre gruppi di pazienti di numerosità diversa:

Gruppo	Pazienti (n)	HbA1c media (%)
A	10	5,2
B	30	6,0
C	20	5,5

Calcolare la media complessiva dei 60 pazienti. Perché non basta fare la media delle tre medie?

Richiami

media campionaria: $\bar{x} = \Sigma\, x_i / n$

media ponderata: $\Sigma\,(p_i \cdot x_i) / \Sigma\, p_i$

(pesi p_i = numerosità dei gruppi)

Esercizio 1 – Media aritmetica e ponderata: svolgimento

a) Media campionaria

$\bar{x} = (12 + 15 + 11 + 14 + 18 + 13 + 16 + 13) / 8 = 112 / 8 =$ **14**

b) Media ponderata (il peso di ciascun gruppo è il suo numero di pazienti)

- Somma dei pesi: $\Sigma p_i = 10 + 30 + 20 = 60$
- Somma dei prodotti: $\Sigma p_i \cdot x_i = 10 \cdot 5{,}2 + 30 \cdot 6{,}0 + 20 \cdot 5{,}5 = 52 + 180 + 110 = 342$
- Media ponderata = 342 / 60 = **5,7 %**

Perché non la media delle medie? (5,2 + 6,0 + 5,5) / 3 = 5,57 %: darebbe lo stesso peso a un gruppo di 10 pazienti e a uno di 30. Il gruppo B, più numeroso, deve contare di più.

Moda

- La moda è il valore più frequente di una distribuzione, o più precisamente, la categoria più ricorrente della variabile (cioè quella associata alla frequenza più alta).

962	1005	1003	768	980	965	1030	1005	975	989	955	783	1005

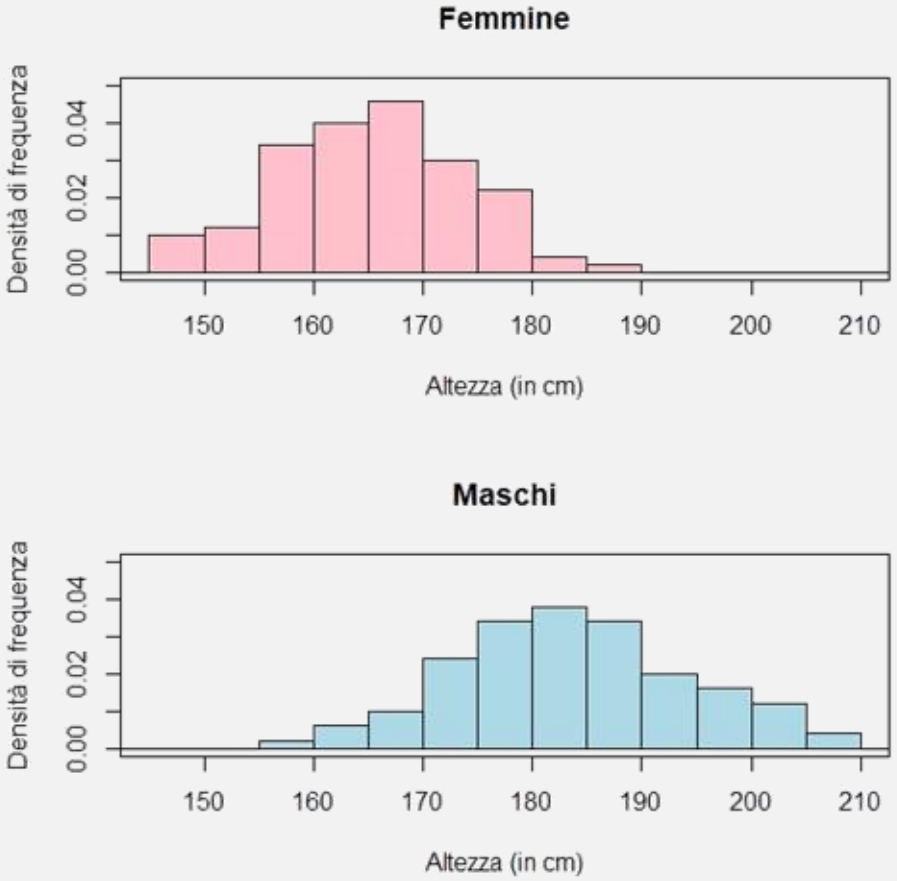
- La moda di questo campione è 1005, poiché ricorre 3 volte.

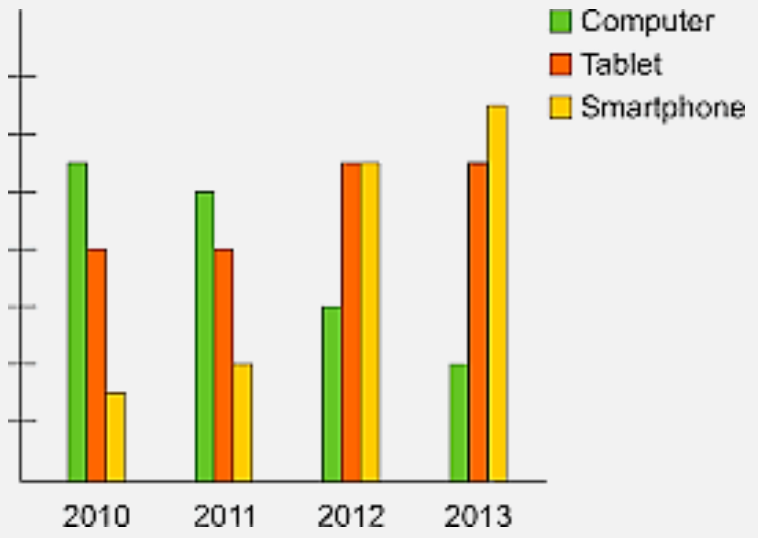
Caratteristiche della moda:

- è utilizzata esclusivamente a scopo descrittivo, perché è meno stabile e meno oggettiva rispetto ad altre misure di tendenza centrale
- per individuare la moda di una distribuzione si possono utilizzare metodi grafici, come gli istogrammi
- può differire all'interno della stessa serie di dati quando si formano classi di distribuzione (intervalli) di ampiezza diversa
- per individuare la moda all'interno di una classe di frequenza, senza conoscere come sono distribuiti i dati, si adotta l'ipotesi di una distribuzione uniforme.

Istogrammi

- Un istogramma descrive la frequenza relativa dei dati all'interno di un intervallo (a, b) ed è utilizzato per visualizzare la distribuzione dei dati.





MODULO MAT7 — BIOSTATISTICA

Distribuzioni Unimodali/Bimodali

- Una distribuzione può presentare più mode:
  - Distribuzioni unimodali: distribuzioni di frequenza che presentano una sola moda, cioè un solo punto di massimo (che rappresenta sia il massimo relativo sia quello assoluto)
  - Distribuzioni bimodali o k-modali: distribuzioni di frequenza che presentano due o più mode, cioè due (o k) massimi relativi
    - Esempio: misurando le altezze di un gruppo di adolescenti in cui la parte maggiore è femminile e quella minore maschile, si ottiene una distribuzione bimodale, con una moda primaria e una secondaria.

MODULO MAT7 — BIOSTATISTICA

Distribuzione Zero-Modale

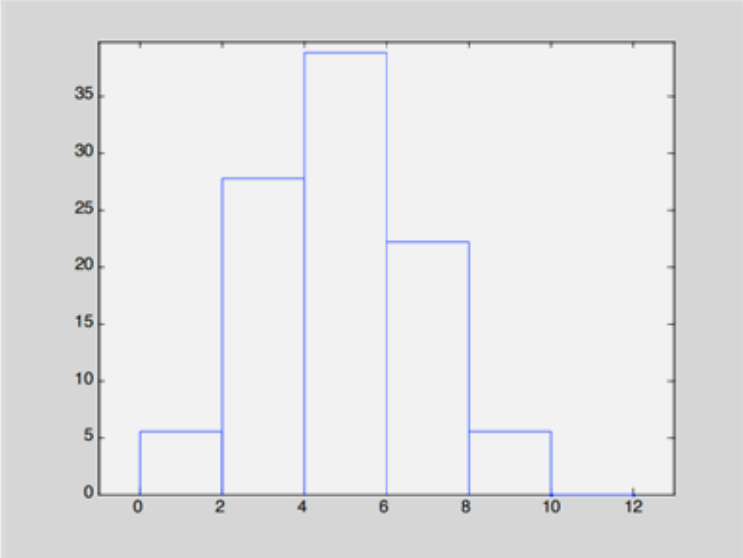
- Nessun valore ha una frequenza maggiore degli altri:

$$A = \{1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6\}$$

Distribuzione unimodale

C'è un solo valore con una frequenza più elevata degli altri

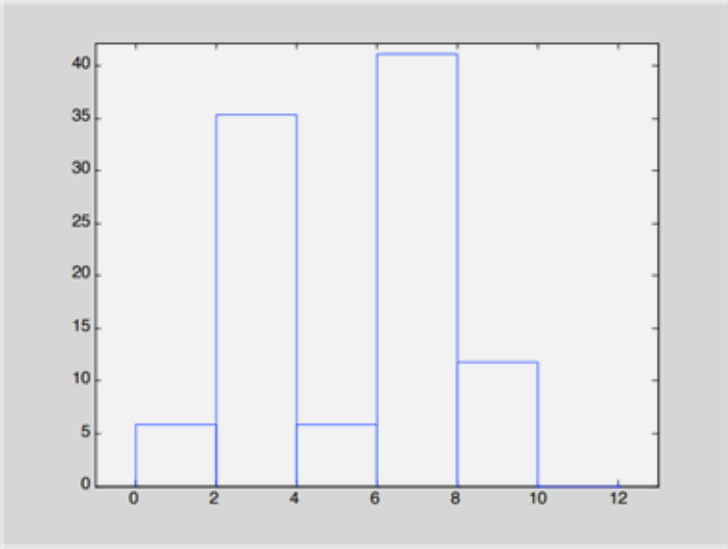
A = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8}

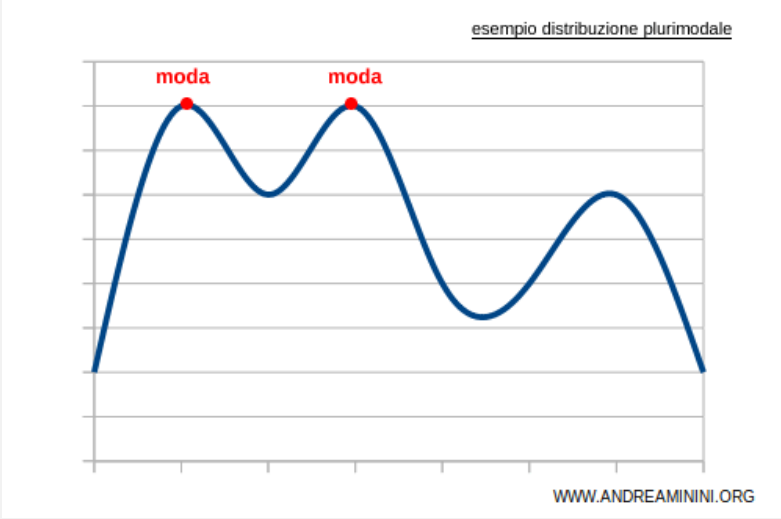


Distribuzione bimodale

Ci sono due valori con una frequenza più elevata degli altri.

A = {1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8}





Esercizio 2 – Frequenze e moda

In un ambulatorio si registra il numero di accessi in un anno di N = 20 pazienti:

3, 2, 2, 3, 1, 1, 4, 1, 2, 0, 1, 1, 2, 1, 2, 0, 0, 2, 1, 1

Si chiede di:

- costruire la tabella con modalità, frequenza assoluta, relativa e percentuale;
- individuare la moda;
- calcolare la percentuale di pazienti con almeno 2 accessi.

Richiami

frequenza relativa = f / N

frequenza percentuale = f / N × 100

moda = modalità con la frequenza assoluta più alta

Esercizio 2 – Frequenze e moda: svolgimento

Passo 1. Contare quante volte compare ogni modalità: è la frequenza assoluta f. Verifica: Σ f = N = 20.

Passo 2. Dividere ogni f per N = 20 (frequenza relativa) e moltiplicare per 100 (percentuale). Esempio, modalità 1: 8 / 20 = 0,40 → 40%.

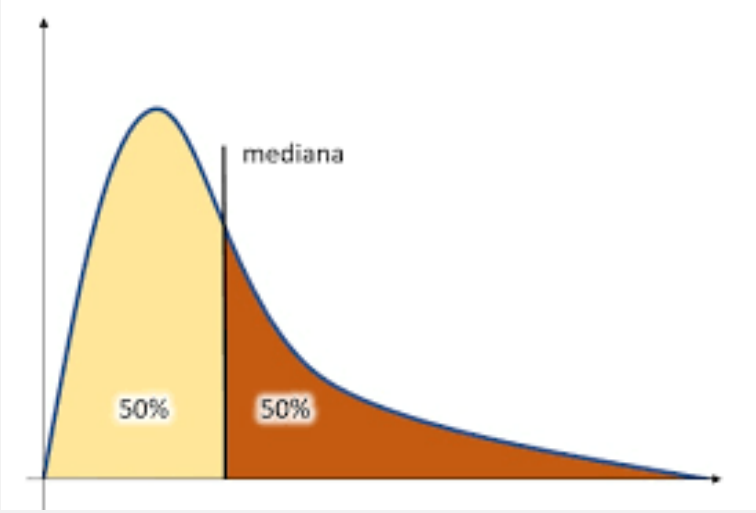
b) Moda: la frequenza più alta è 8, quindi la moda è **1 accesso**.

c) Almeno 2 accessi: modalità 2, 3 e 4 → 6 + 2 + 1 = 9 pazienti → 9 / 20 = 0,45 → **45%**.

Modalità	f	f / N	%
0	3	0,15	15%
1	**8**	**0,40**	**40%**
2	6	0,30	30%
3	2	0,10	10%
4	1	0,05	5%
Totale	**20**	**1,00**	**100%**

Mediana

- La mediana è il valore che occupa la posizione centrale in un insieme di dati ordinato.
- È una **misura robusta**, poiché è minimamente influenzata dalla presenza di valori anomali.
- Caratteristiche:
  - si utilizza quando si vuole attenuare l'effetto dei valori estremi;
  - in una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.



Calcolo della Mediana

Per calcolare la mediana di un insieme di dati occorre:

1. ordinare i valori in senso crescente o decrescente e contare il numero totale n di dati;
2. se il numero (n) di dati è dispari, la mediana corrisponde al valore numerico del dato centrale, cioè quello che occupa la posizione $(n+1)/2$;
3. se il numero (n) di dati è pari, la mediana si stima utilizzando i due valori centrali che occupano le posizioni $n/2$ e $n/2+1$:
   a. con un numero ridotto di osservazioni, la mediana è la media aritmetica di queste due osservazioni intermedie;
   b. con un numero elevato di osservazioni raggruppate in classi, si ricorre talvolta alle proporzioni.

Esempio.
Si consideri il seguente campione:

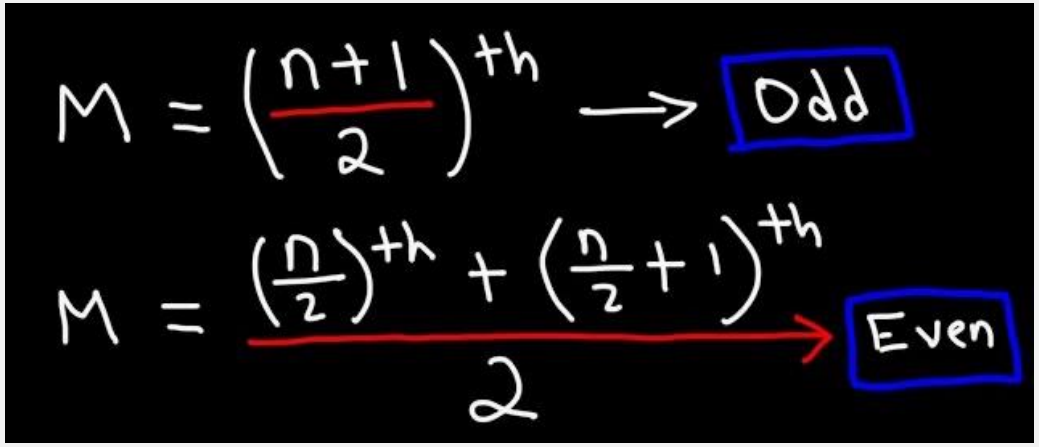
96	78	90	62	73	89	92	84	76	86

Ordinare i campioni in senso crescente:

62	73	76	78	**84**	**86**	89	90	92	95

1. Poiché il numero di campioni è pari ($n = 10$), la mediana si calcola come media dei due elementi centrali:

$$mediana = 84 + 862 = 85$$



Esercizio 3 – Mediana e robustezza

Concentrazione di proteina C reattiva (mg/L) in 7 pazienti:

6, 35, 4, 7, 5, 5, 8

- Calcolare mediana e media.
- Si aggiunge un ottavo paziente con valore 6: ricalcolare la mediana.
- Se il valore 35 fosse in realtà 350, come cambierebbero media e mediana? Quale indice è più robusto?

Richiami

mediana: ordinare i dati;

n dispari → posizione $(n + 1)/2$

n pari → media delle posizioni $n/2$ e $n/2 + 1$

Esercizio 3 – Mediana e robustezza: svolgimento

a) n = 7. Ordino i dati: 4 · 5 · 5 · **6** · 7 · 8 · 35

- n è dispari → posizione (7 + 1)/2 = 4 → mediana = **6 mg/L**
- Media = (4 + 5 + 5 + 6 + 7 + 8 + 35) / 7 = 70 / 7 = **10 mg/L**

b) n = 8. Ordino: 4 · 5 · 5 · **6** · **6** · 7 · 8 · 35

- n è pari → posizioni n/2 = 4 e n/2 + 1 = 5 → mediana = (6 + 6) / 2 = **6 mg/L**

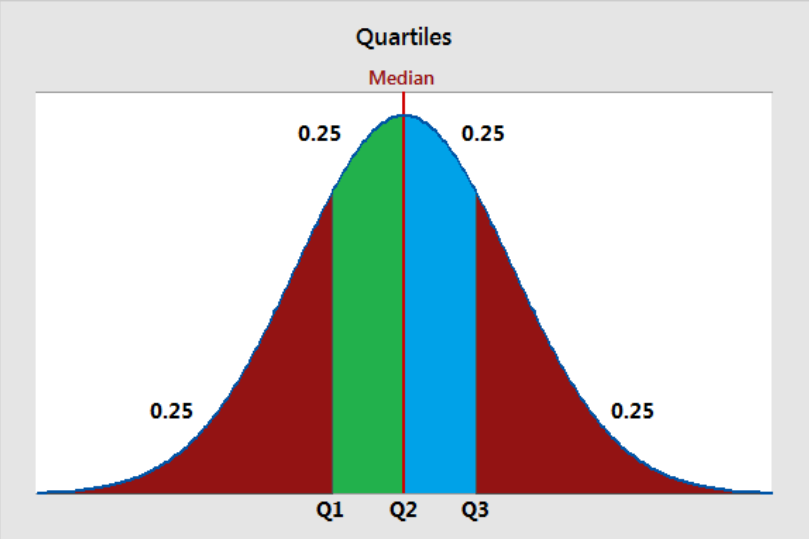
c) Valore anomalo 350 invece di 35 (n = 7)

- Somma = 70 − 35 + 350 = 385 → media = 385 / 7 = **55 mg/L**
- Dati ordinati: 4 · 5 · 5 · 6 · 7 · 8 · 350 → il 4° valore è ancora 6 → mediana = **6 mg/L**

La mediana non cambia, la media sì: la mediana è robusta rispetto ai valori anomali.

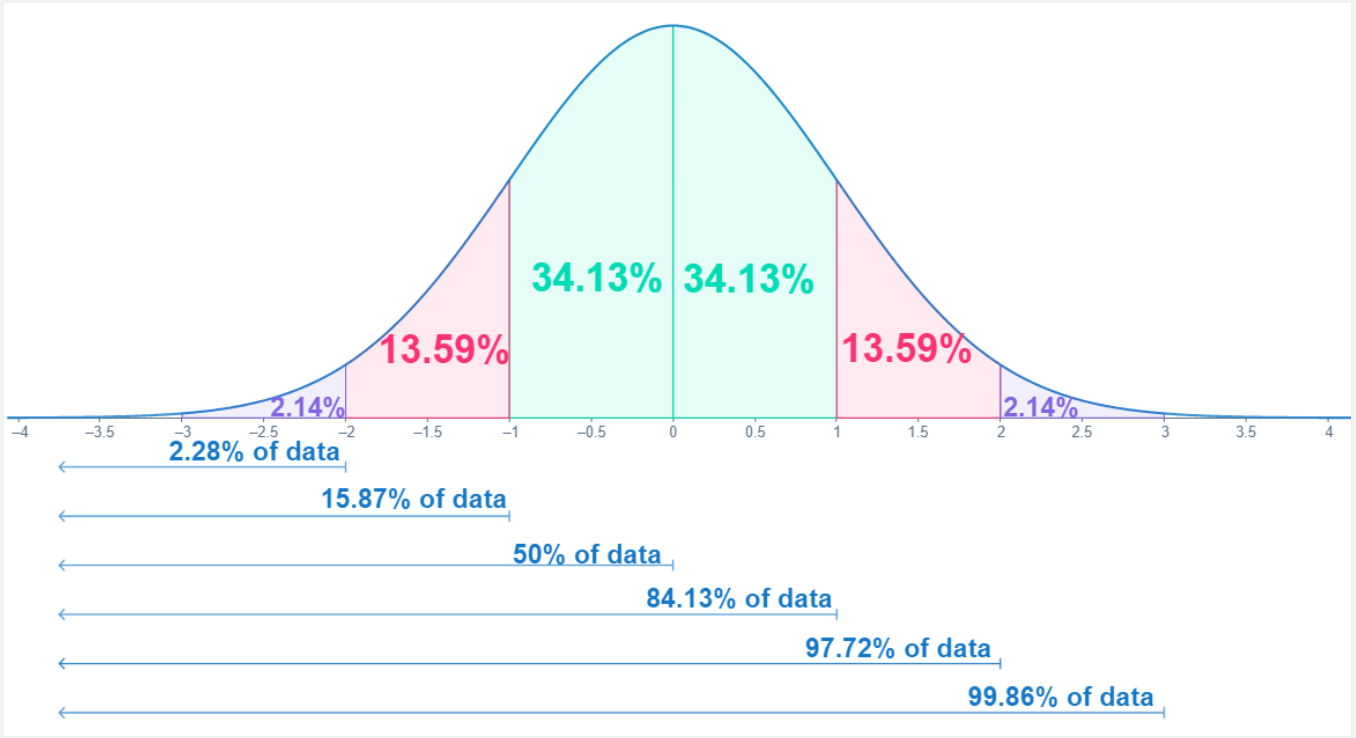
Quantili

- I quantili sono una famiglia di misure, cui appartiene anche la mediana, che si distinguono in base al numero di parti uguali in cui dividono una distribuzione.
- I quartili suddividono la distribuzione in 4 parti di uguale frequenza, dove ogni parte contiene la stessa frazione di osservazioni:
  - Il primo quartile è definito come il numero $q1$ tale che il 25% dei dati statistici sia minore o uguale a $q1$.
  - Il secondo quartile è definito come il numero $q2$ tale che il 50% dei dati statistici sia minore o uguale a $q2$. Il secondo quartile corrisponde alla mediana.
  - Il terzo quartile è definito come un numero $q3$ tale che il 75% dei dati statistici sia minore o uguale a $q3$.



Decili e Percentili

- Analogamente, si definiscono:
  - Decili: 9 punti che dividono la distribuzione ordinata in 10 parti uguali
  - Percentili: 99 punti che dividono la distribuzione ordinata in 100 parti uguali



Campo di Variazione

- Il campo di variazione di una distribuzione è la differenza tra il valore più grande e il valore più piccolo della distribuzione:

$$C = x_{max} - x_{min}$$

Questo indice è relativamente grossolano, poiché non fornisce alcuna informazione sulla variabilità dei valori intermedi.

- Esempio: il campo di variazione della seguente distribuzione:

$$25 - 26 - 28 - 29 - 30 - 32 \rightarrow C = 32 - 25 = 7$$

Varianza

- Varianza della popolazione: misura ottimale per caratterizzare la variabilità di una popolazione. Indicatore della dispersione di una variabile o di una distribuzione statistica che si ottiene calcolando la media dei quadrati degli scarti dalla media aritmetica (μ).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

dove N è il numero di osservazioni dell'intera popolazione; μ è la media della popolazione; x_i è la i-esima osservazione

- Varianza di un campione:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

dove n è il numero di osservazioni del campione; $\bar{X}$ è la media del campione; x_i è la i-esima osservazione

Quando n è sufficientemente grande, la differenza tra le due formule è minima.

Deviazione Standard, o Scarto Quadratico Medio

- Deviazione standard della popolazione:

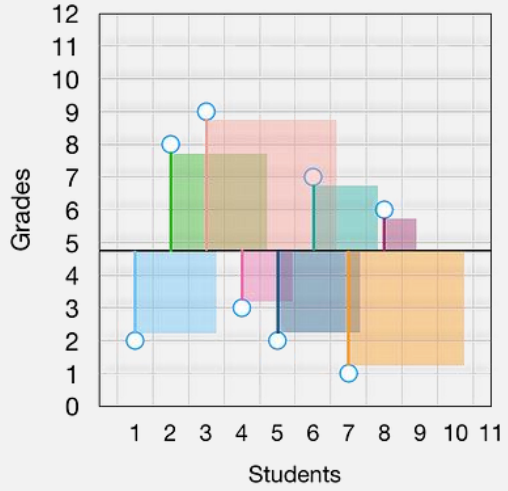
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

dove N è il numero di osservazioni dell'intera popolazione; μ è la media della popolazione; X_i è la i-esima osservazione.

- Deviazione standard del campione:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

dove n è il numero di osservazioni del campione; $\bar{X}$ è la media del campione; x_i è la i-esima osservazione.



$$\sum (x_n - \bar{x})^2 =$$

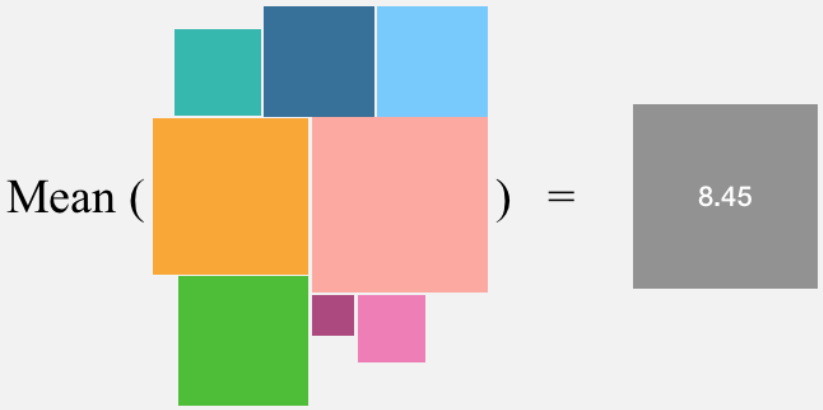
$$7.5625 + 10.5625 + 18.0625 + 3.0625 + 7.5625 + 5.0625 + 14.0625 + 1.5625 = 67.5$$

Varianza

$$\frac{\sum (x_n - \bar{x})^2}{N} = \frac{67.5}{8} = 8.45 \text{ points}^2$$

Deviazione standard

$$\frac{\sum (x_n - \bar{x})^2}{N}$$





Esempio.
Consideriamo i voti di due studenti:

- Anna: *30, 30, 28, 27, 26*
- Stefano: *21, 30, 30, 30, 30*

Entrambi hanno la stessa media dei voti (28,2)

Calcoliamo la deviazione standard:

- $\sigma_{Anna} = 1.78$
- $\sigma_{Stefano} = 4.02$

Che cosa significa? I voti di Anna sono più concentrati (più vicini tra loro) rispetto a quelli di Stefano

Esercizio 4 – Quartili

Tempo di guarigione (giorni) in N = 11 pazienti:

8, 3, 12, 5, 4, 15, 6, 9, 4, 10, 7

- Calcolare la mediana (Q2), il primo quartile (Q1) e il terzo quartile (Q3).
- Calcolare lo scarto interquartile e il campo di variazione.
- Tra quali valori guarisce il 50% centrale dei pazienti?

Richiami

Q1: 25% dei dati ≤ Q1

Q2 = mediana

Q3: 75% dei dati ≤ Q3

Metodo: Q1 e Q3 sono le mediane della metà inferiore e della metà superiore dei dati ordinati (la mediana è esclusa se n è dispari).

scarto interquartile = Q3 – Q1

Esercizio 4 – Quartili: svolgimento

Posizione	1	2	3	4	5	6	7	8	9	10	11
Dato ordinato	3	4	**4**	5	6	**7**	8	9	**10**	12	15

Passo 1. Ordino i dati (tabella sopra).

Passo 2. Mediana. N = 11 (dispari) → posizione (11 + 1)/2 = 6 → Q2 = **7 giorni**

Passo 3. Q1. Metà inferiore (5 valori sotto la mediana): 3 · 4 · **4** · 5 · 6 → il 3° valore → Q1 = **4 giorni**

Passo 4. Q3. Metà superiore (5 valori sopra la mediana): 8 · 9 · **10** · 12 · 15 → il 3° valore → Q3 = **10 giorni**

Passo 5. Scarto interquartile = Q3 − Q1 = 10 − 4 = **6 giorni**

Passo 6. Campo di variazione = $x_{max} - x_{min}$ = 15 − 3 = **12 giorni**

Interpretazione. Il 50% centrale dei pazienti guarisce tra 4 e 10 giorni; il 25% guarisce in 4 giorni o meno e il 25% in 10 giorni o più.

Esercizio 5 – Variabilità: varianza e deviazione standard

Frequenza cardiaca a riposo (bpm) di n = 5 volontari:

62, 70, 66, 74, 68

- Calcolare il campo di variazione e la media.
- Calcolare varianza e deviazione standard del campione.
- Se questi 5 valori fossero l'intera popolazione di interesse, quanto varrebbero varianza e deviazione standard?

Richiami

$C = x_{max} - x_{min}$

$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$

$s = \sqrt{s^2}$

$\sigma^2 = \Sigma (x_i - \mu)^2 / N$

$\sigma = \sqrt{\sigma^2}$

Esercizio 5 – Variabilità: svolgimento

Campo di variazione. $C = 74 - 62 =$ **12 bpm**

Media. $\bar{x} = (62 + 70 + 66 + 74 + 68) / 5 = 340 / 5 =$ **68 bpm**

Scarti dalla media e loro quadrati (tabella): la somma degli scarti è 0 (controllo!), la somma dei quadrati è 80.

Campione ($n - 1 = 4$): $s^2 = 80 / 4 =$ **20 bpm²**; $s = \sqrt{20} =$ **4,47 bpm**

Popolazione ($N = 5$): $\sigma^2 = 80 / 5 =$ **16 bpm²**; $\sigma = \sqrt{16} =$ **4 bpm**

Dividendo per $n - 1$ invece che per N si ottiene un valore un po' più grande; con n grande la differenza è minima.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
62	−6	36
70	+2	4
66	−2	4
74	+6	36
68	0	0
Σ	**0**	**80**

Correlazione

- È un approccio più rigoroso che consente di studiare il grado di intensità della relazione lineare tra coppie di variabili

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

Coefficiente di correlazione di Pearson

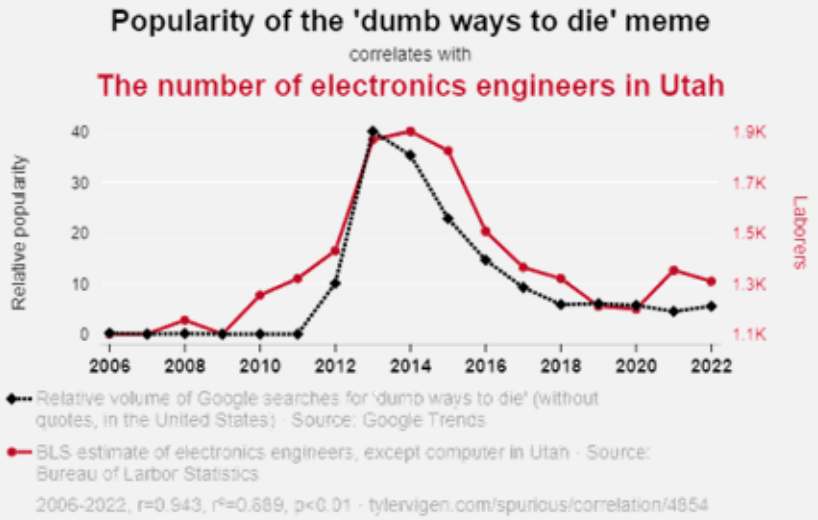
Il coefficiente di correlazione consente di:

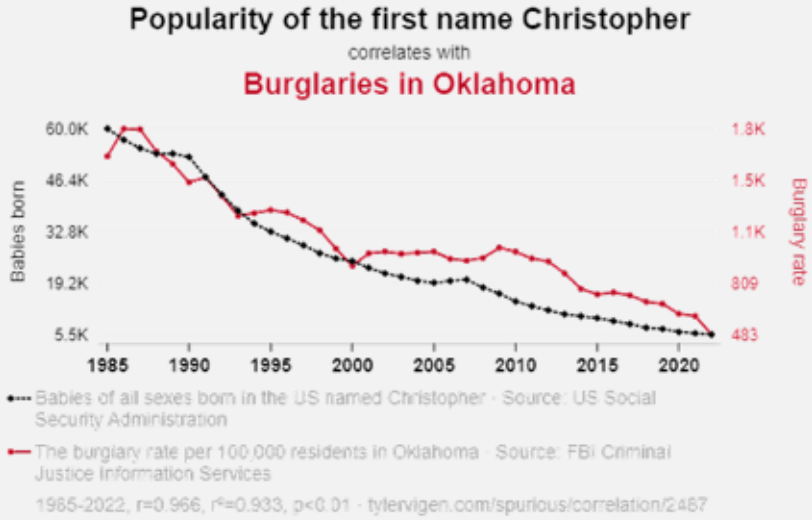
- riassumere la forza della relazione lineare tra le variabili
- valutare l'associazione apparente tra le variabili

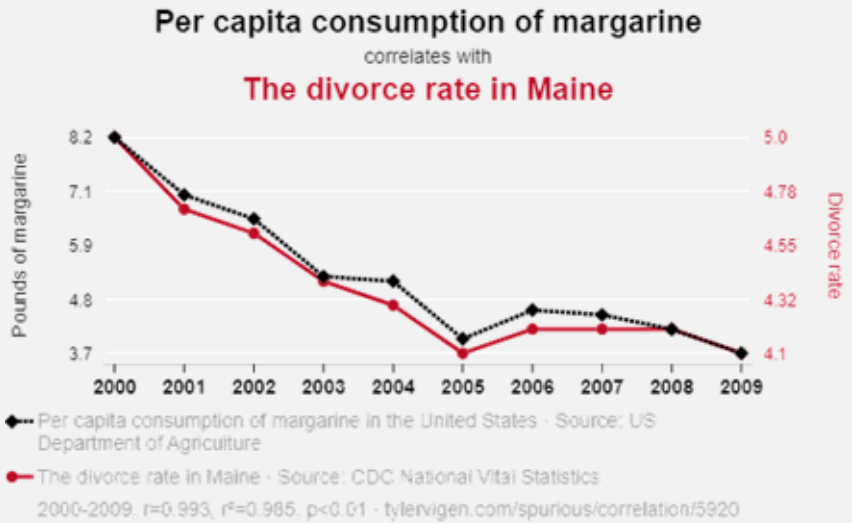
Il coefficiente di correlazione:

- varia da −1 a 1 (se uguale a 1 o −1: correlazione perfetta)
- è positivo quando i valori delle variabili crescono insieme
- è negativo quando i valori di una variabile crescono al diminuire dei valori dell'altra
- non è influenzato dalle unità di misura

CORRELAZIONE NON SIGNIFICA CAUSALITÀ







Esercizio 6 – Correlazione di Pearson

In 5 pazienti ipertesi si somministrano dosi crescenti di un farmaco (x) e si misura la riduzione della pressione sistolica (y):

Paziente	1	2	3	4	5
Dose x (mg)	10	20	30	40	50
Riduzione y (mmHg)	4	8	9	10	14

Calcolare il coefficiente di correlazione di Pearson r e commentarlo.

Richiami

$r = \Sigma(dx \cdot dy) / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)}$

con $dx = x - \bar{x}$ e $dy = y - \bar{y}$

$-1 \leq r \leq 1$

Esercizio 6 – Correlazione di Pearson: svolgimento

Passo 1. Medie: $\bar{x} = 150 / 5 =$ **30 mg**; $\bar{y} = 45 / 5 =$ **9 mmHg**

Passo 2. Scarti dalle medie, prodotti e quadrati:

x	y	dx = x − 30	dy = y − 9	dx · dy	dx²	dy²
10	4	−20	−5	100	400	25
20	8	−10	−1	10	100	1
30	9	0	0	0	0	0
40	10	+10	+1	10	100	1
50	14	+20	+5	100	400	25
Σ		**0**	**0**	**220**	**1000**	**52**

Passo 3. $r = 220 / \sqrt{(1000 \cdot 52)} = 220 / \sqrt{52000} = 220 / 228{,}04 =$ **0,96**

Passo 4. Commento. r è vicino a +1: correlazione lineare positiva molto forte (a dosi maggiori corrispondono riduzioni maggiori). r non dipende dalle unità di misura. **Attenzione:** correlazione non significa causalità.

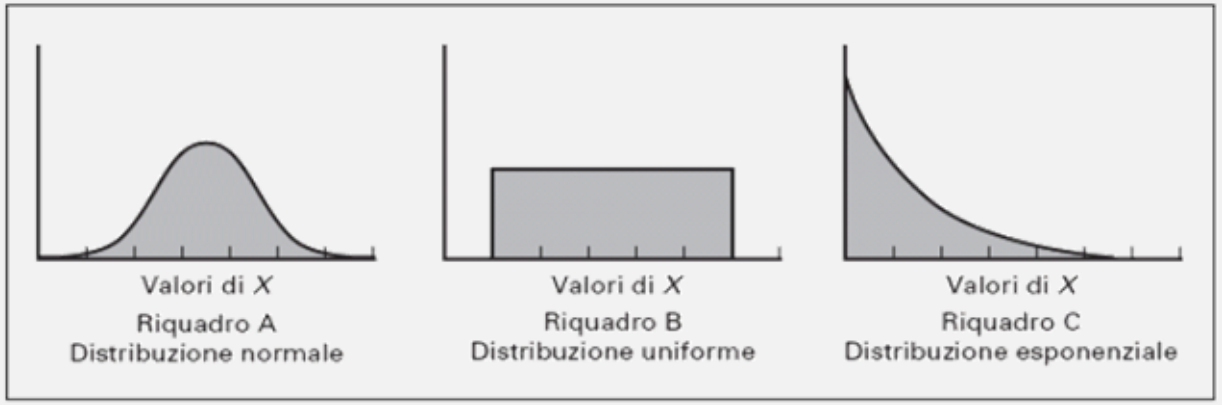
$$r = \Sigma(dx \cdot dy) / \sqrt{(\ \Sigma dx^2 \cdot \Sigma dy^2\ )}$$

Distribuzione di Probabilità Continue

- Una funzione di densità di probabilità continua è un modello che definisce analiticamente come sono distribuiti i valori assunti da una variabile casuale continua

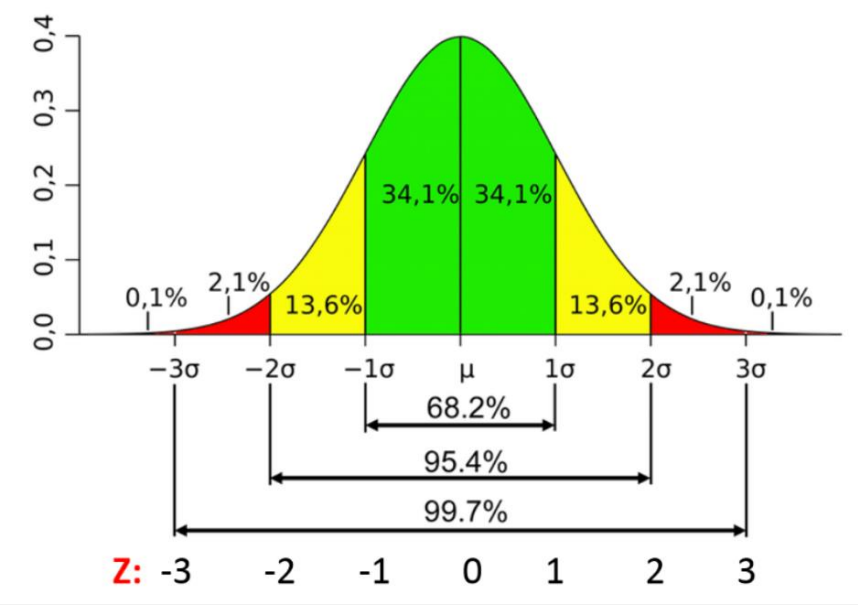
- Quando si dispone di un'espressione matematica adatta a rappresentare un fenomeno continuo, siamo in grado di calcolare la probabilità che la variabile casuale assuma valori compresi in determinati intervalli

- La figura rappresenta graficamente tre funzioni di densità di probabilità: normale, uniforme ed esponenziale

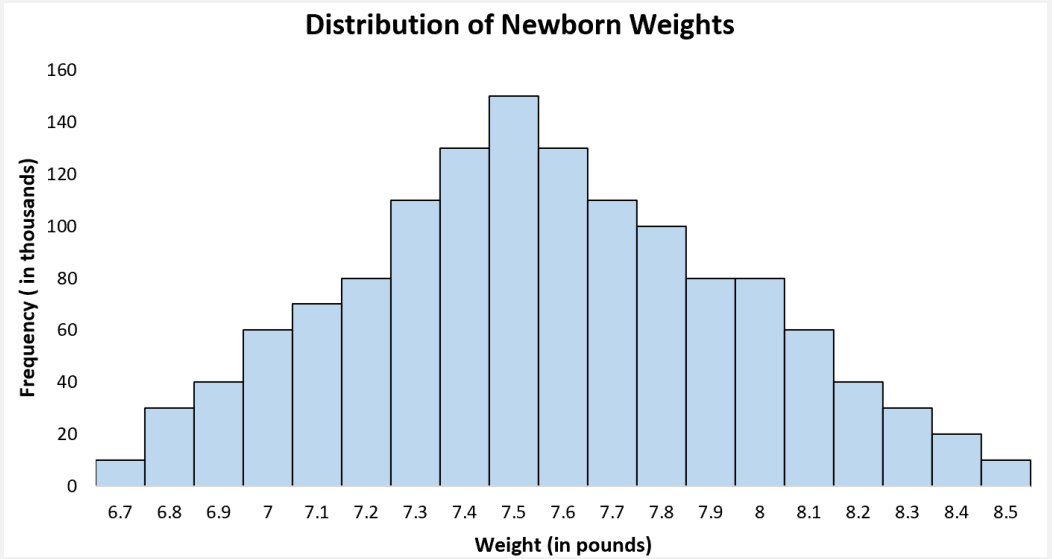


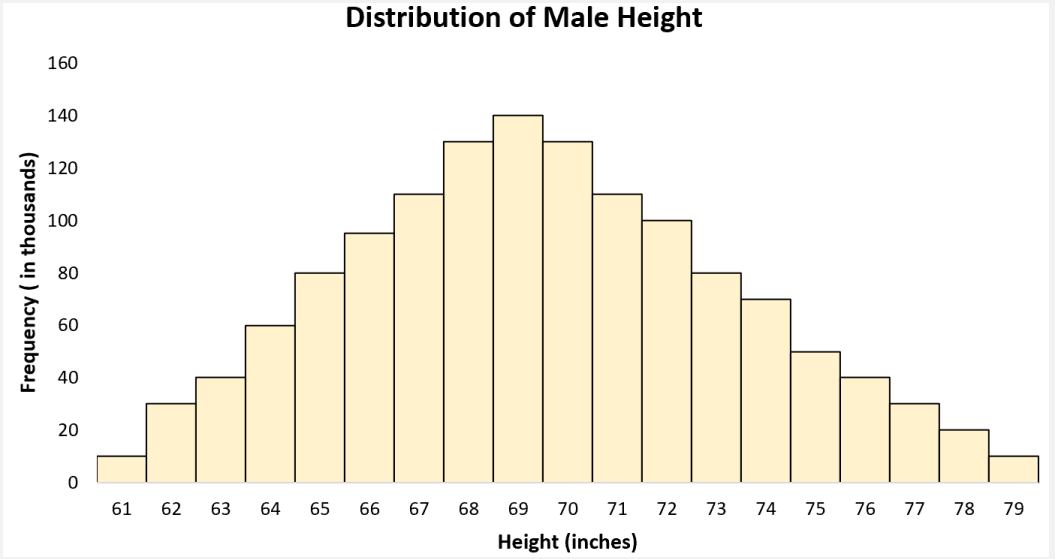
Distribuzione Normale

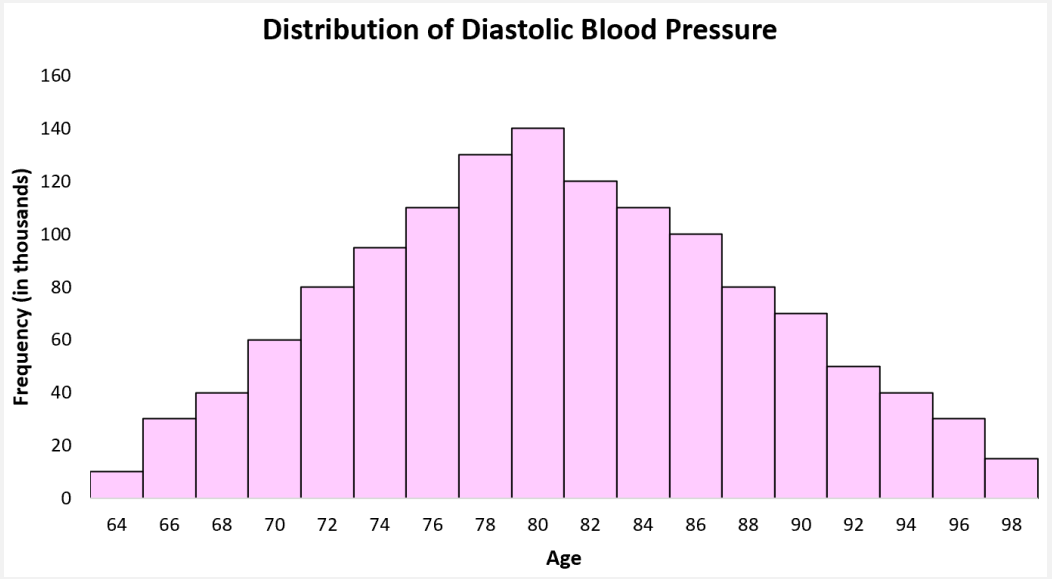
- La distribuzione normale (o distribuzione gaussiana) è la distribuzione continua più utilizzata in statistica.
- La distribuzione normale è importante in statistica per tre motivi fondamentali:
  1. Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale.
  2. La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.

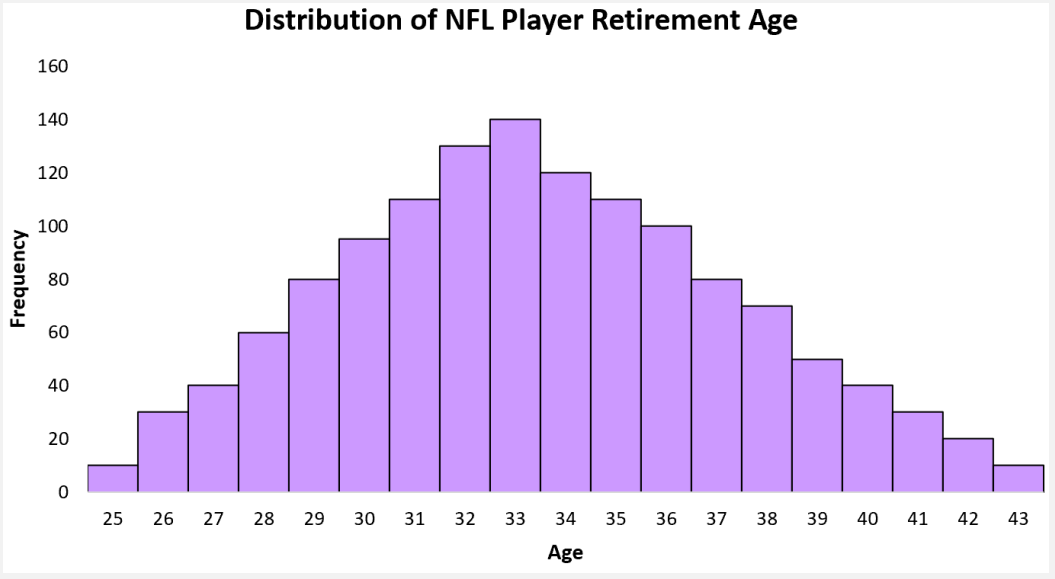





Distribution of Newborn Weights
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
6.7
6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
8.3
8.4
8.5
Weight (in pounds)


Distribution of Male Height
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
Height (inches)


Distribution of Diastolic Blood Pressure
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
98
Age


Distribution of NFL Player Retirement Age
Frequency
0
20
40
60
80
100
120
140
160
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
Age

Distribuzione Normale

La distribuzione normale presenta diverse caratteristiche importanti:

- La distribuzione normale ha una forma campanulare e simmetrica
- Le sue misure di tendenza centrale (valore atteso, mediana) coincidono
- La sua differenza interquartile è pari a 1,33 volte la deviazione standard; copre cioè un intervallo compreso tra $\mu - 2/3\ \sigma$ e $\mu + 2/3\ \sigma$
- La variabile casuale con distribuzione normale assume valori tra $-\infty$ e $+\infty$

Distribuzione Normale

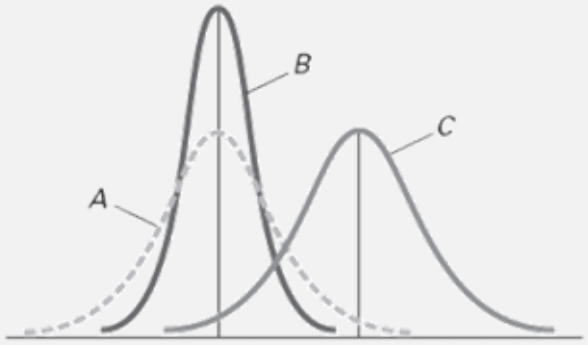
Utilizzeremo il simbolo $f(X)$ per indicare l'espressione matematica di una funzione di densità di probabilità. Nel caso della distribuzione normale, la funzione di densità di probabilità normale è data dalla seguente espressione:

Funzione di densità di probabilità normale

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{1}{2}\right)\left[\frac{X-\mu}{\sigma}\right]^2}$$

Dove μ è il valore atteso della popolazione; σ è la deviazione standard della popolazione; X rappresenta i valori assunti dalla variabile casuale, $-\infty < X < +\infty$

Si noti che, poiché e e π sono costanti matematiche, le probabilità di una distribuzione normale dipendono esclusivamente dai valori assunti dai due parametri μ e σ. Specificando particolari combinazioni di μ e σ, otteniamo diverse distribuzioni di probabilità normali.



Esempio – intervalli di riferimento nella Distribuzione Normale

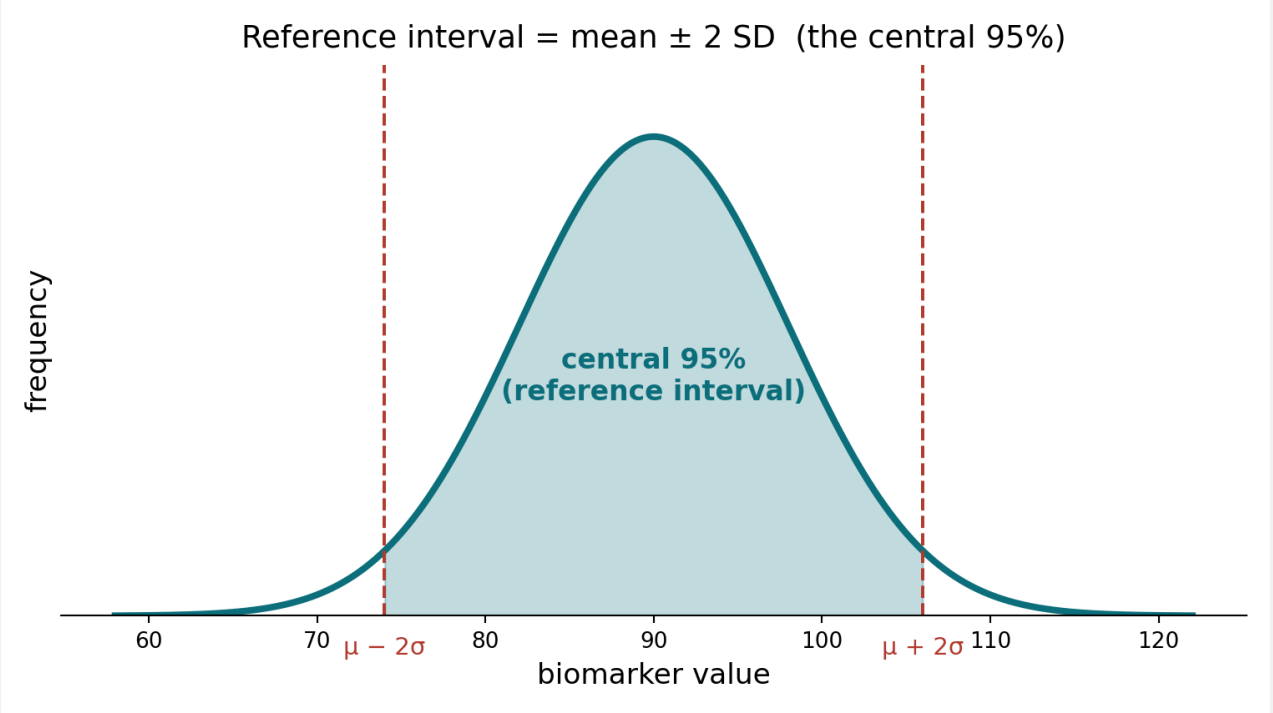
Molte variabili biomediche sono approssimativamente distribuite normalmente attorno a una media.

Il 95% centrale della distribuzione definisce l'intervallo di riferimento (normale):

intervallo di riferimento ≈ media ± 2 DS

Un valore al di fuori viene segnalato come 'anomalo' — eppure circa il 5% delle persone sane cade fuori per puro caso.

È esattamente così che vengono costruiti gli intervalli di riferimento di laboratorio.



Esercizio 7 – Distribuzione normale

L’emoglobina nelle donne sane è distribuita approssimativamente in modo normale, con media μ = 13,5 g/dL e deviazione standard σ = 1,0 g/dL.

- Calcolare l’intervallo di riferimento (95% centrale).

- Tre pazienti hanno emoglobina A = 10,9; B = 14,7; C = 15,8 g/dL. Calcolare lo z-score di ciascuna e dire se rientra nell’intervallo di riferimento.

- Su 400 donne sane, quante ci aspettiamo fuori dall’intervallo?

- Quale intervallo contiene circa il 68% delle donne sane?

Richiami

intervallo di riferimento ≈ μ ± 2σ

μ ± σ ≈ 68%; μ ± 2σ ≈ 95%; μ ± 3σ ≈ 99,7%

z = (x − μ) / σ (numero di DS dalla media)

Esercizio 7 – Distribuzione normale: svolgimento

a) $\mu \pm 2\sigma = 13{,}5 \pm 2 \cdot 1{,}0 \rightarrow$ **11,5 – 15,5 g/dL**

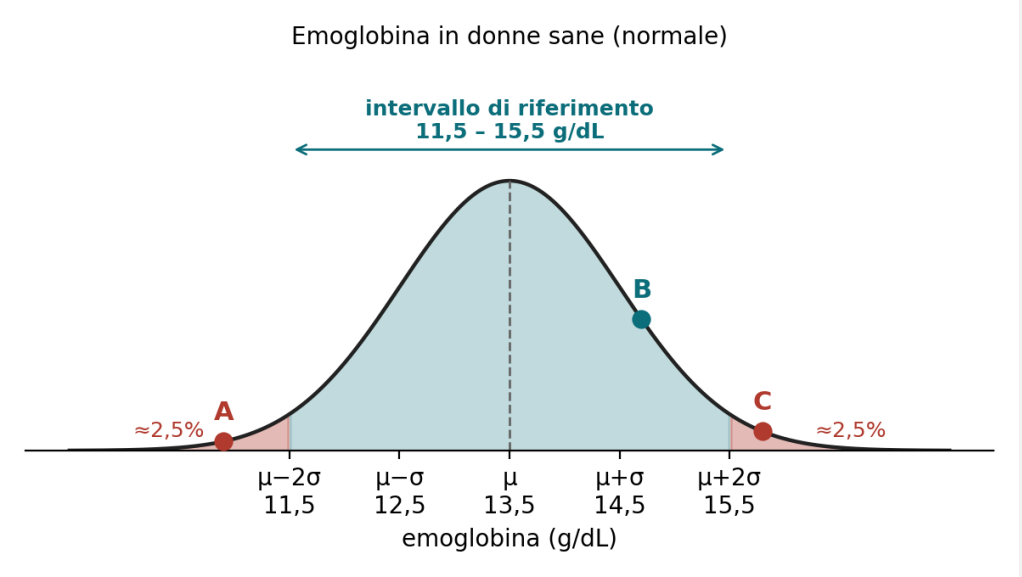
b) $z = (x - \mu) / \sigma$

- A: $z = (10{,}9 - 13{,}5) / 1{,}0 =$ **−2,6** → $|z| > 2$: **fuori** (sotto)
- B: $z = (14{,}7 - 13{,}5) / 1{,}0 =$ **+1,2** → **dentro**
- C: $z = (15{,}8 - 13{,}5) / 1{,}0 =$ **+2,3** → **fuori** (sopra)

c) circa il 5% sta fuori: $0{,}05 \cdot 400 =$ **20 donne** (circa 10 per coda).

d) $\mu \pm \sigma = 13{,}5 \pm 1{,}0 \rightarrow$ **12,5 – 14,5 g/dL**

Essere fuori dall'intervallo non significa essere malati: circa il 5% delle persone sane cade fuori per puro caso.



intervallo di riferimento ≈ $\mu \pm 2\sigma$

$\mu \pm \sigma \approx 68\%$; $\mu \pm 2\sigma \approx 95\%$; $\mu \pm 3\sigma \approx 99{,}7\%$

$z = (x - \mu) / \sigma$ (numero di DS dalla media)

Dal campione alla popolazione: l'Errore Standard

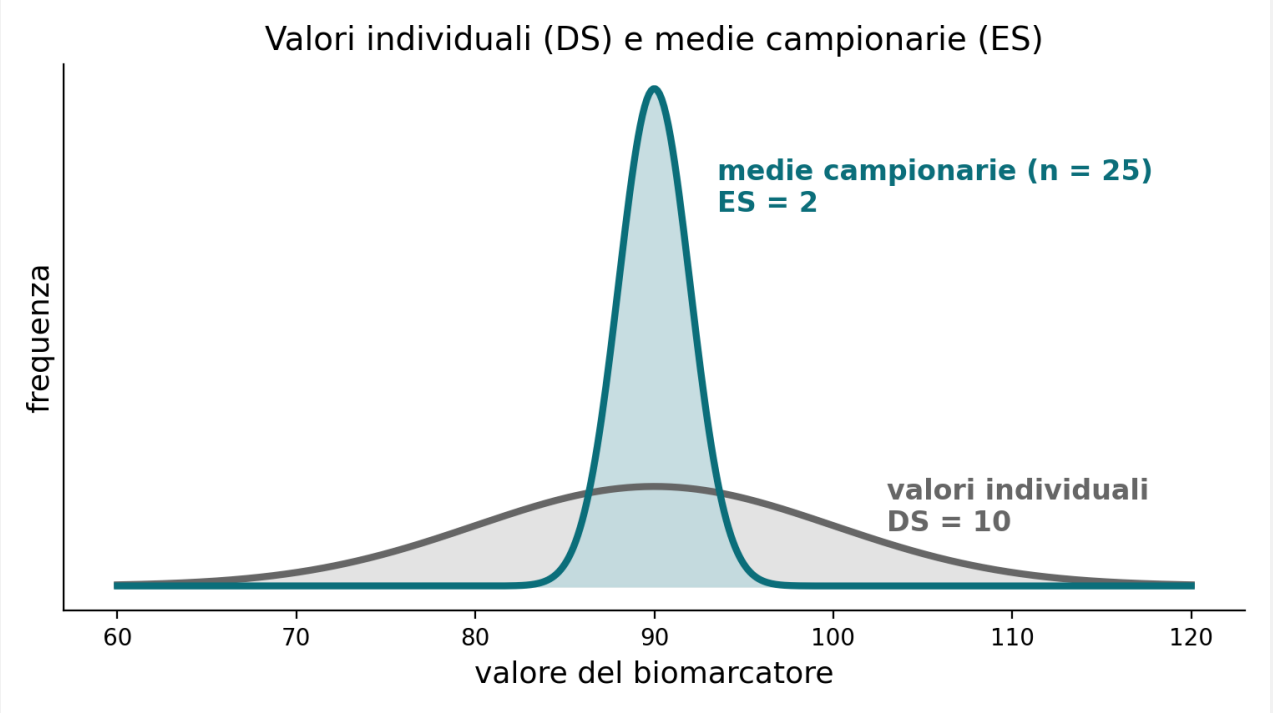
La media di un campione è solo una stima della vera media (della popolazione): un altro campione darebbe una media leggermente diversa.

La dispersione delle medie campionarie attorno alla vera media è l'errore standard della media (ES):

$$ES = DS/\sqrt{n}$$

La DS descrive la variabilità dei singoli valori; l'ES descrive con quale precisione è stimata la media.

La DS non diminuisce al crescere del campione; l'ES sì. Per dimezzare l'ES servono quattro volte i soggetti.



Esempio – L’errore standard di una misura di glicemia

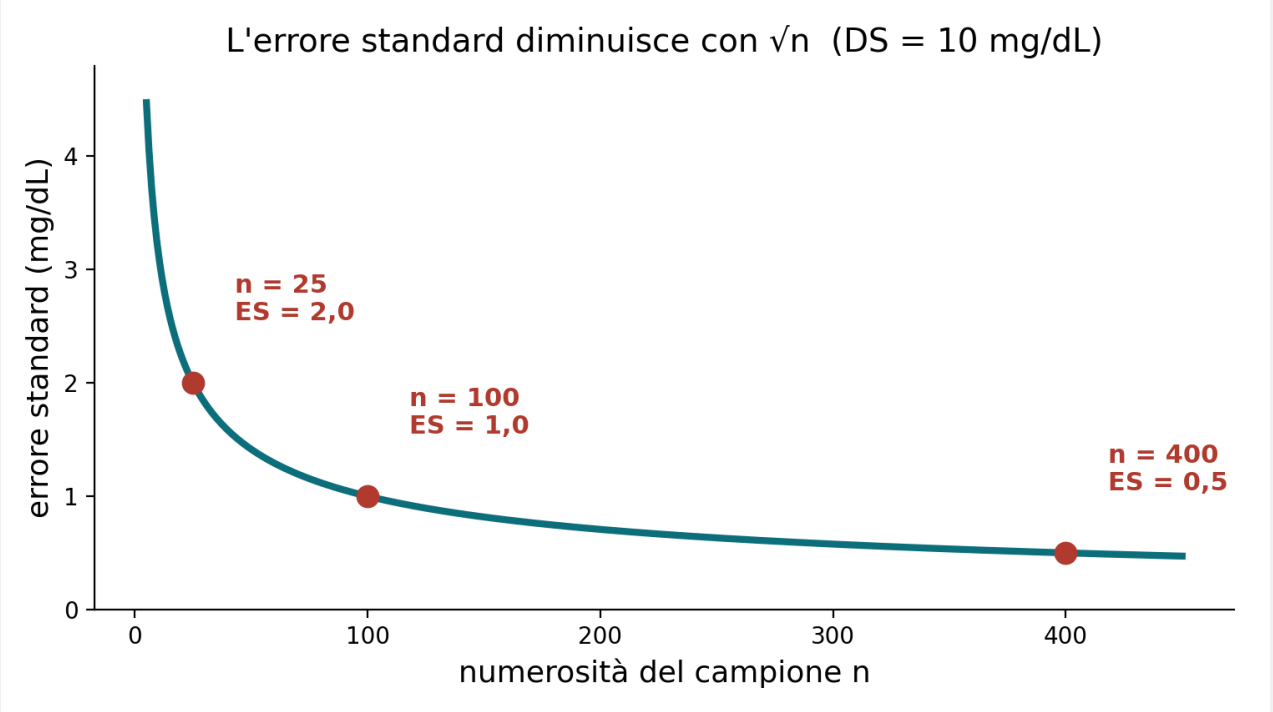
Glicemia a digiuno misurata in n = 25 volontari: media = 92 mg/dL, DS = 10 mg/dL.

$$ES = 10/\sqrt{25} = 2 mg/dL$$

Con n = 100 soggetti: ES = 10 / √100 = 1 mg/dL. Con n = 400: ES = 0,5 mg/dL.

Quadruplicare la numerosità dimezza soltanto l'ES: la precisione cresce con √n.

L'ES è il denominatore del test t: t = (media₁ − media₂) / errore standard.



L’Intervallo di Confidenza della Media

Invece di un singolo numero, si riporta un intervallo di valori plausibili per la vera media.

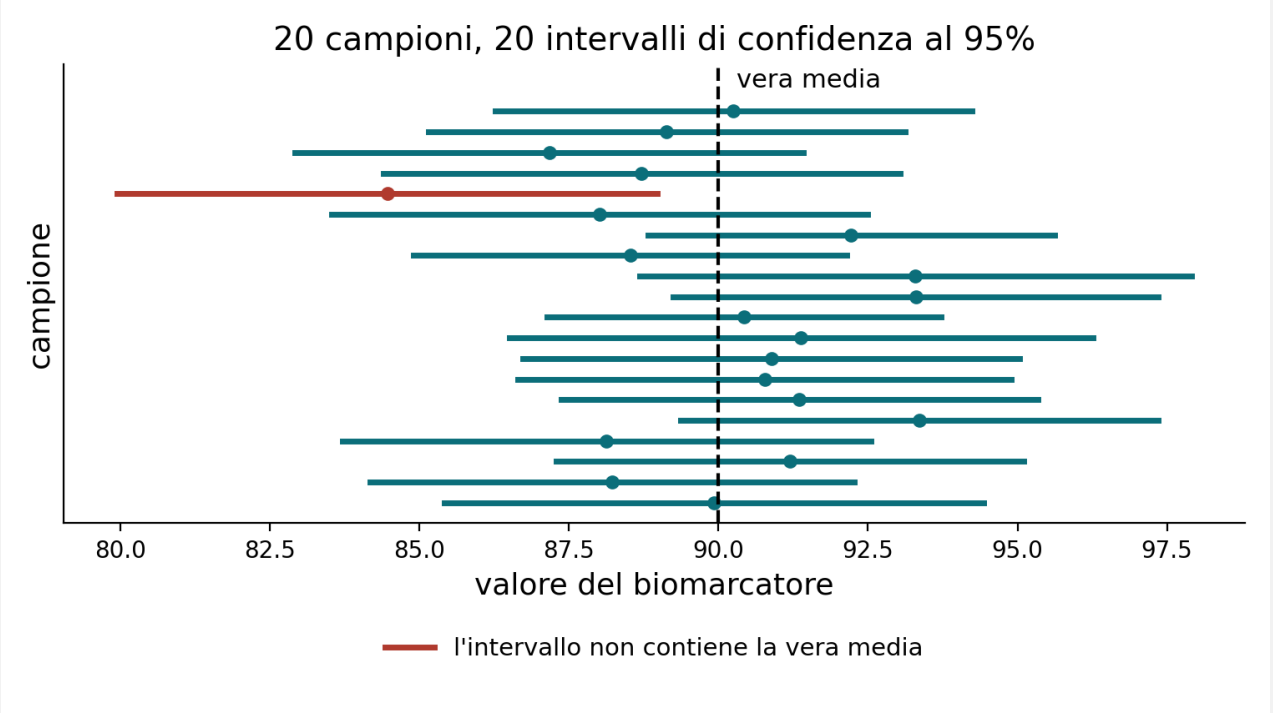
Intervallo di confidenza al 95%:

IC 95% ≈ media ± 2 · ES

(esattamente media ± t · ES, con t ≈ 1,96 per campioni grandi e maggiore per quelli piccoli)

Se ripetessimo lo studio molte volte, il 95% degli intervalli costruiti così conterrebbe la vera media.

Intervallo stretto = stima precisa. Non confonderlo con l'intervallo di riferimento (media ± 2 DS), che descrive i valori individuali.



Esempio – Intervallo di Confidenza al 95% della glicemia

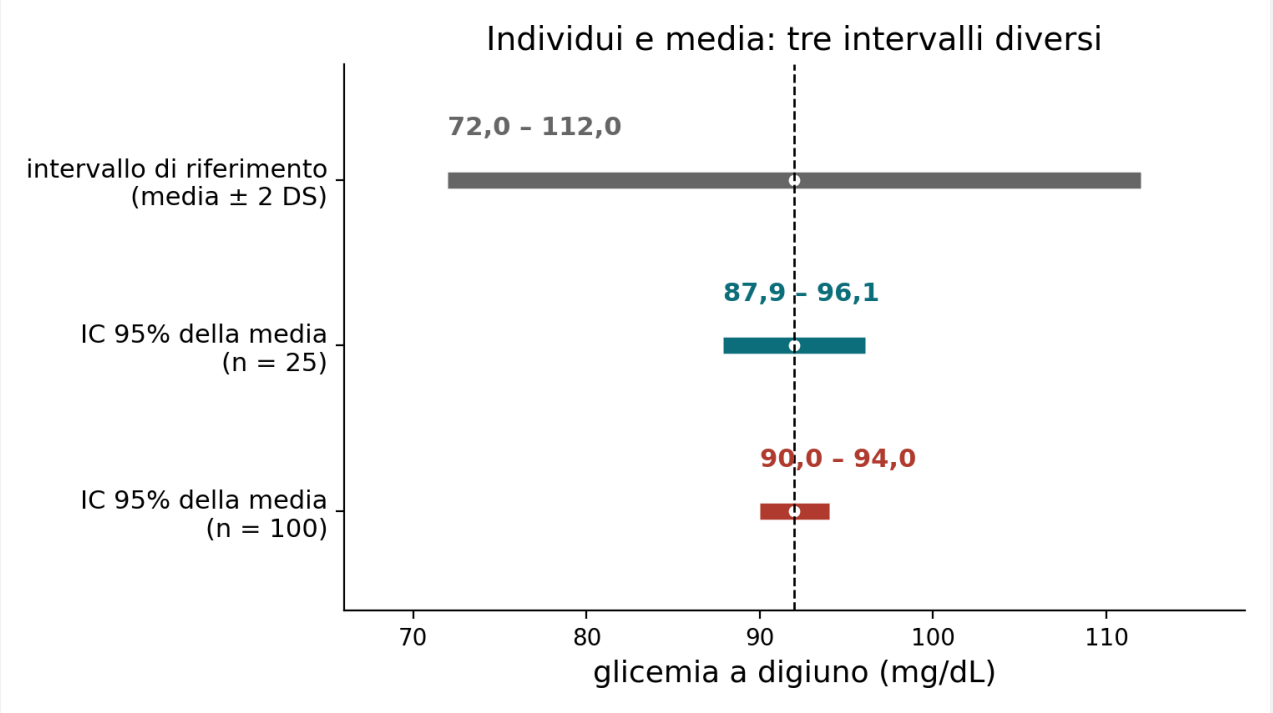
n = 25, media = 92 mg/dL, DS = 10 mg/dL → ES = 2 mg/dL. Per n = 25, t = 2,06.

IC 95% = 92 ± 2,06 · 2 = 87,9 – 96,1 mg/dL

Siamo confidenti al 95% che la vera glicemia media della popolazione sia compresa tra circa 88 e 96 mg/dL.

Con n = 100 (ES = 1 mg/dL) l'intervallo si restringe a 90,0 – 94,0 mg/dL.

L'intervallo di riferimento (media ± 2 DS = 72 – 112 mg/dL) è molto più ampio: descrive gli individui, non la media.



Il Valore *p*

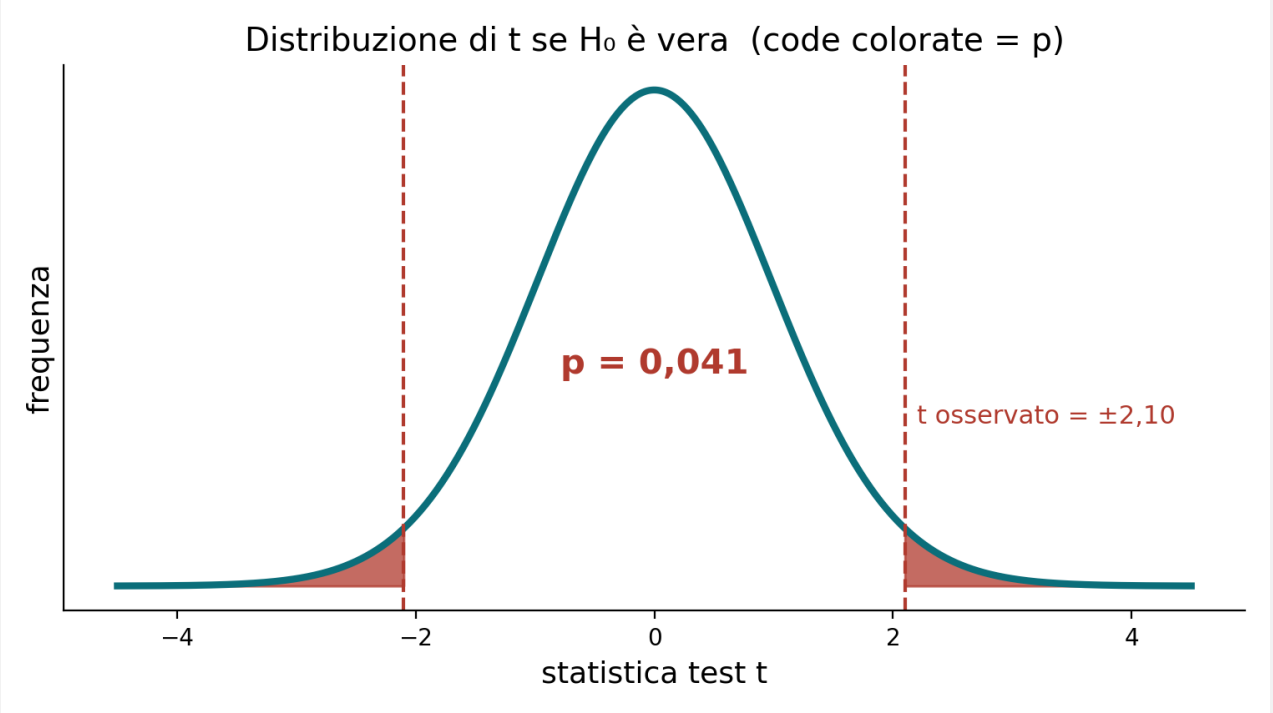
Si parte dall'ipotesi zero H_0: non c'è alcun effetto reale (ad es. trattati e controlli hanno la stessa media).

Il valore p è la probabilità di osservare una differenza grande almeno quanto la nostra se H_0 fosse vera.

$p < 0{,}05 \rightarrow$ statisticamente significativo

p piccolo (per convenzione $p < 0{,}05$): dati difficilmente spiegabili dal solo caso, quindi si rifiuta H_0.

p grande: dati compatibili con il caso; non si può rifiutare H_0 (ma questo non prova che l'effetto non esista).



Leggere correttamente il valore *p*

p NON è la probabilità che H_0 sia vera, e 0,05 è una convenzione, non una legge di natura: p = 0,04 e p = 0,06 raccontano quasi la stessa storia.

Significativo ≠ importante.

Con campioni molto grandi anche una differenza minima può dare p < 0,001: riporta sempre la dimensione dell'effetto e il suo intervallo di confidenza.

p e IC vanno insieme: se l'IC 95% della differenza non contiene 0, allora p < 0,05.

Due modi di sbagliare: errore di I tipo (falso positivo, probabilità α = 0,05) ed errore di II tipo (falso negativo, probabilità β).

I due modi di sbagliare

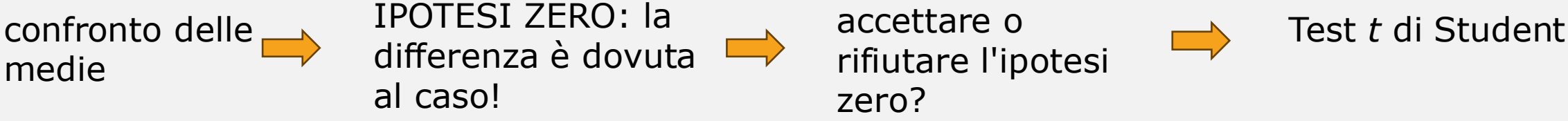
	H_0 vera (nessun effetto reale)	H_0 falsa (effetto reale)
Si rifiuta H_0 (significativo)	Errore di I tipo falso positivo (α)	corretto potenza = 1 − β
Non si rifiuta H_0 (non significativo)	corretto (1 − α)	Errore di II tipo falso negativo (β)

Confrontare due medie: il test *t* di Student

la differenza tra le medie dei due campioni è significativa?

Puoi affermare che la differenza osservata non è dovuta al caso ma che, invece, esiste realmente una differenza tra le medie delle due popolazioni?



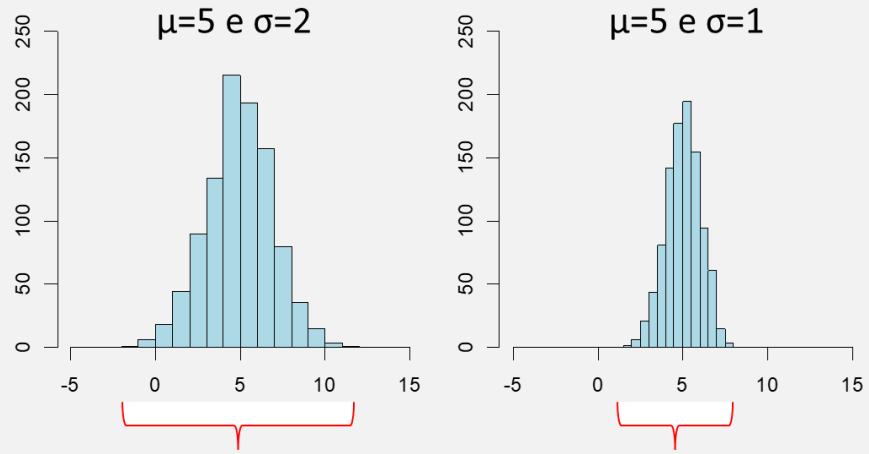


Confrontare due medie: il test *t* di Student

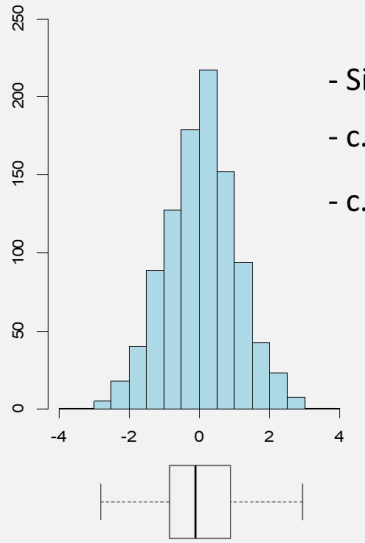
confronto delle medie → IPOTESI ZERO: la differenza è dovuta al caso! → accettare o rifiutare l'ipotesi zero? → Test *t* di Student

Condizioni:

1. Indipendenza delle osservazioni
2. Normalità delle popolazioni a confronto
3. Omogeneità della varianza



Analisi dell'istogramma



- Simmetria (media ≈mediana)
- c. 2/3 dei dati in un intervallo μ±σ
- c. 95% dei dati in un intervallo μ±2σ

Confrontare due medie: il test *t* di Student

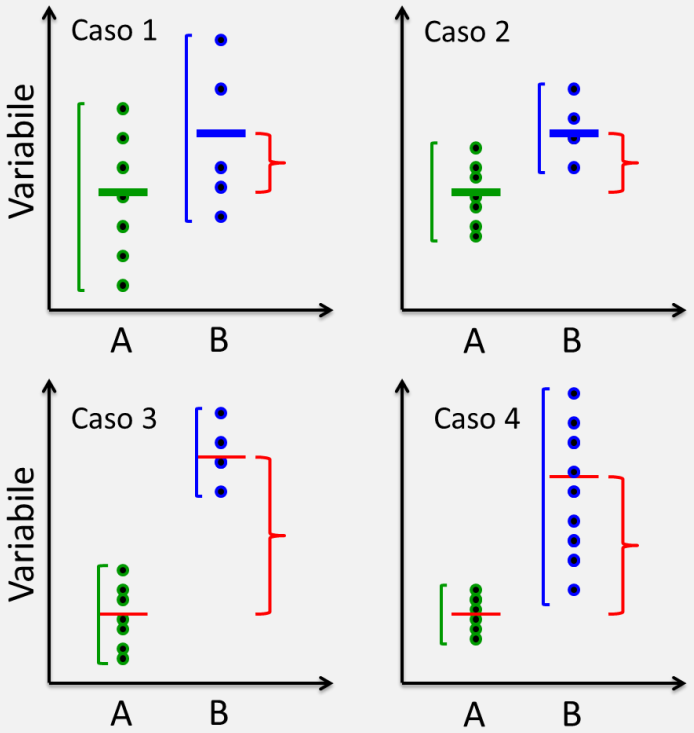
differenza tra le medie

fattore dimensionale

$$t = \frac{m_a - m_b}{S} \sqrt{\frac{n_a\, n_b}{n_a + n_b}}$$

errore standard

t = (media₁ – media₂) / errore standard



differenza tra le medie

variabilità A

variabilità B

Esempio – Confrontare due gruppi con il test *t* di Student

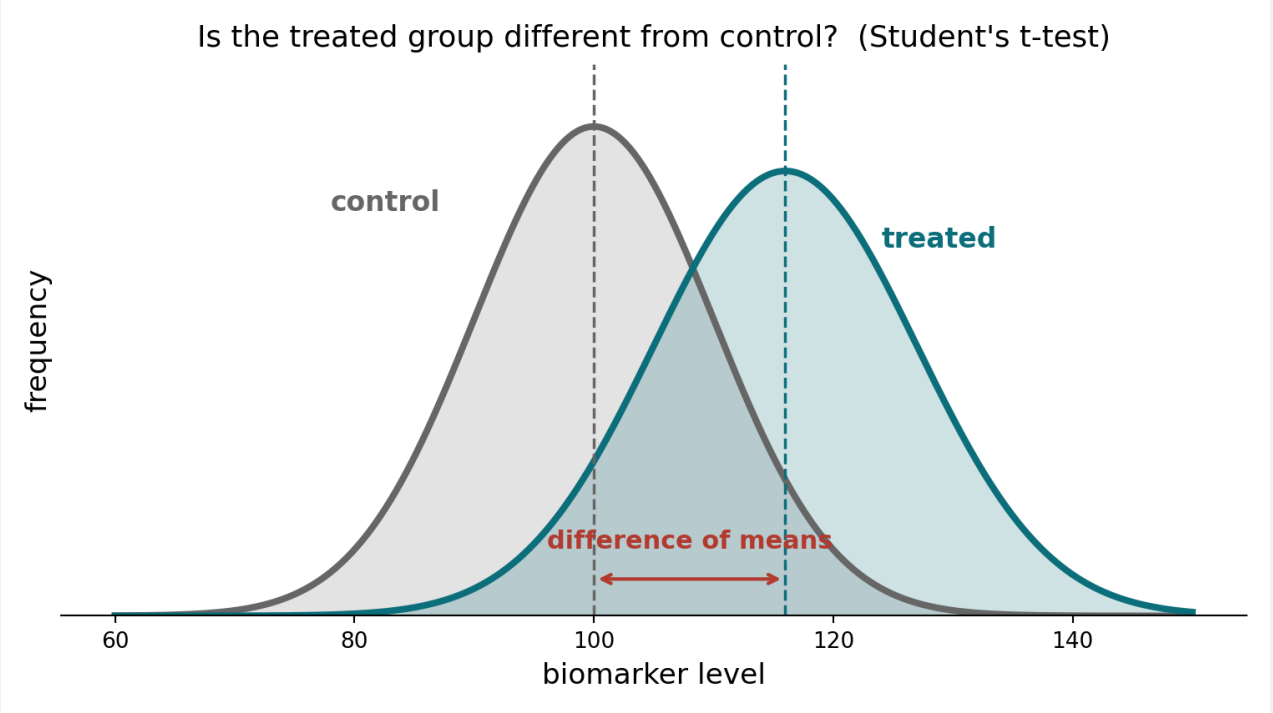
Un trattamento modifica un biomarcatore? Confronta il gruppo trattato con un gruppo di controllo.

Il test t confronta la differenza tra le due medie con la variabilità interna ai gruppi:

$$t = (\text{media}_1 - \text{media}_2) / \text{errore standard}$$

Un valore p piccolo indica che la differenza osservata è improbabile che sia dovuta al solo caso.

Questa è la logica alla base della maggior parte degli esperimenti 'trattati vs controllo'.



Accuratezza diagnostica — sensibilità e specificità

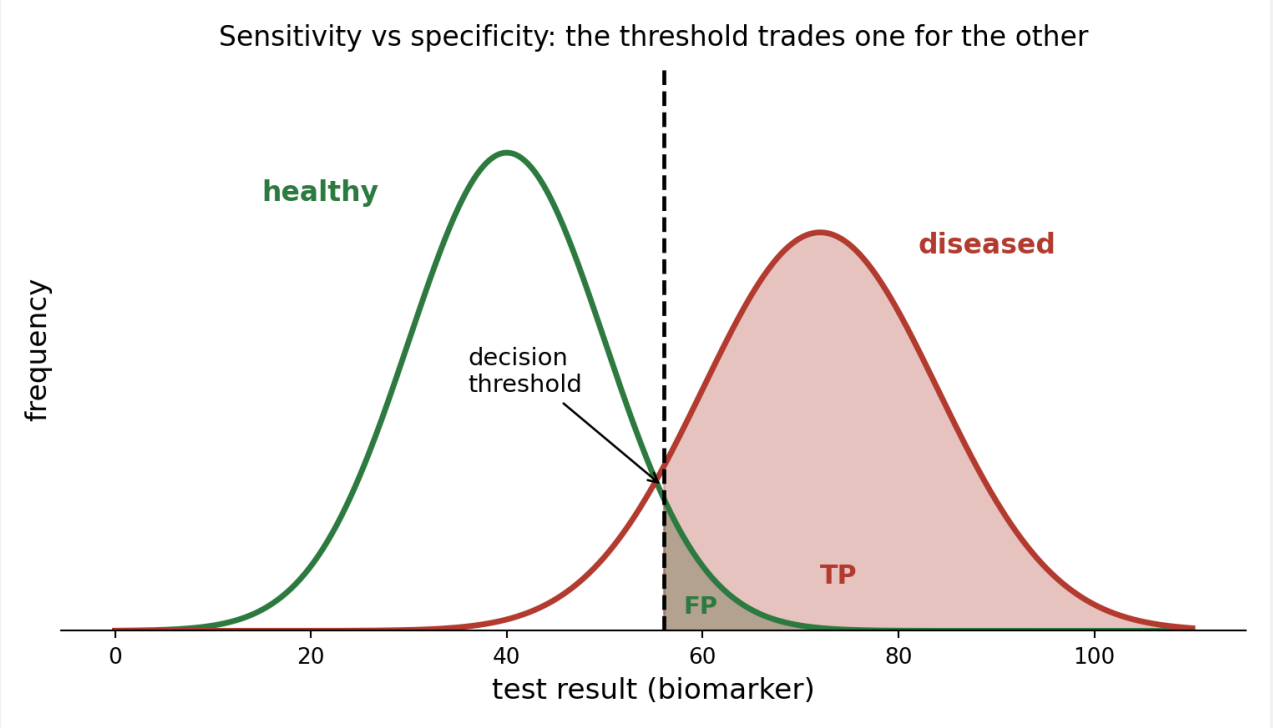
Un test diagnostico deve separare i malati dai sani, ma i loro valori si sovrappongono.

Sensibilità = VP / (VP + FN): la frazione di malati correttamente individuati.

Specificità = VN / (VN + FP): la frazione di sani correttamente esclusi.

Spostare la soglia decisionale scambia sensibilità con specificità.

Anche i valori predittivi (**VPP, VPN**) dipendono da quanto è comune la malattia.



Esercizio 10 – Test diagnostici

Un test rapido è applicato a 1000 persone. La prevalenza della malattia è del 10%: 100 malati e 900 sani. Il test è positivo in 90 dei 100 malati e in 90 dei 900 sani.

- Costruire la tabella 2 × 2 (VP, FN, FP, VN).
- Calcolare sensibilità e specificità.

Richiami

Se = VP / (VP + FN)

Sp = VN / (VN + FP)

VPP = VP / (VP + FP)

VPN = VN / (VN + FN)

Esercizio 10 – Test diagnostici: svolgimento (a – c)

	Malati	Sani	Totale
Test positivo	**VP = 90**	FP = 90	180
Test negativo	FN = 10	**VN = 810**	820
Totale	100	900	1000

a) FN = 100 − 90 = 10 (malati con test negativo); VN = 900 − 90 = 810 (sani con test negativo).

b) Se = VP / (VP + FN) = 90 / 100 = **0,90 = 90%**

Sp = VN / (VN + FP) = 810 / 900 = **0,90 = 90%**

MODULO MAT7 — BIOSTATISTICA

Esercizi di Statistica

Dieci esercizi con calcoli svolti passo per passo (basta una calcolatrice). Ogni esercizio ha una slide con il testo e una con lo svolgimento.

N.	Argomento	Indici utilizzati
1	Media aritmetica e ponderata	media campionaria; media ponderata
2	Frequenze e moda	frequenza assoluta, relativa, percentuale; moda
3	Mediana e robustezza	mediana (n dispari e pari); effetto dei valori anomali
4	Quartili	Q1, Q2, Q3; scarto interquartile; campo di variazione
5	Variabilità	varianza e deviazione standard (campione e popolazione)
6	Correlazione	coefficiente di correlazione di Pearson
7	Distribuzione normale	z-score; intervallo di riferimento media ± 2 DS
8	Errore standard e intervallo di confidenza	ES; IC 95% della media; effetto di n
9	Confronto tra due medie	test t di Student; valore p; IC della differenza
10	Test diagnostici	sensibilità, specificità, VPP, VPN; effetto della prevalenza

MODULO MAT7 — BIOSTATISTICA

Formulario

Indice	Formula
Frequenze relativa e %	f / N $f / N \times 100$ (f = frequenza assoluta, N = numero di dati)
Media campionaria	$\bar{x} = \Sigma x_i / n$
Media ponderata	$\Sigma (p_i \cdot x_i) / \Sigma p_i$ (p_i = pesi)
Mediana	dati ordinati: n dispari → posizione $(n + 1)/2$; n pari → media delle posizioni $n/2$ e $n/2 + 1$
Quartili	Q1, Q2 (= mediana), Q3; scarto interquartile = $Q3 - Q1$
Campo di variazione	$C = x_{max} - x_{min}$
Varianza e DS campione	$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$ $s = \sqrt{s^2}$
Varianza e DS popolazione	$\sigma^2 = \Sigma (x_i - \mu)^2 / N$ $\sigma = \sqrt{\sigma^2}$
Correlazione di Pearson	$r = \Sigma(dx \cdot dy) / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)}$, con $dx = x - \bar{x}$, $dy = y - \bar{y}$
z-score; int. di riferimento	$z = (x - \mu) / \sigma$ intervallo di riferimento $\approx \mu \pm 2\sigma$ (95% centrale)
Errore standard; IC 95%	$ES = s / \sqrt{n}$ $IC\ 95\% = \bar{x} \pm t \cdot ES$ ($gdl = n - 1$)
Test t (due campioni)	$t = (\bar{x}_1 - \bar{x}_2) / ES_{diff}$, $ES_{diff} = \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$, $gdl = n_1 + n_2 - 2$
Test diagnostici	Se = VP/(VP+FN) Sp = VN/(VN+FP) VPP = VP/(VP+FP) VPN = VN/(VN+FN)