PROF. CRESCENZIO GALLO

L'ANALISI STATISTICA DEI DATI NELLA RICERCA SANITARIA

Università di Foggia

Università degli Studi di Foggia

Dipartimenti di Area Medica

LA STATISTICA

COME PRESENTARE IN MANIERA CHIARA I RISULTATI?

- In questa dispensa cercheremo di approfondire le più comuni tecniche statistiche per l'analisi dei dati raccolti nell'ambito della ricerca sanitaria.
- Verranno inoltre forniti i concetti fondamentali per poter interpretare in modo corretto i risultati di uno studio.

RACCOLTA E PRESENTAZIONE DEI DATI

ASPETTI STATISTICI
La Raccolta dei Dati

Le informazioni raccolte nel corso di una sperimentazione clinica controllata (SCC), e abitualmente riportate su un'apposita scheda raccolta dati, riguardano generalmente le caratteristiche socio-demografiche e cliniche dei pazienti arruolati.

Ognuna di queste caratteristiche (ad esempio età, sesso, stadio di malattia, etc.) prende il nome di "**variabile**". Le variabili utilizzate nell'ambito di un qualsiasi studio possono essere di due tipi distinti:

- *numeriche (continue* o *discrete)*;
- *ordinali*;
- *nominali* (o *categoriche*).

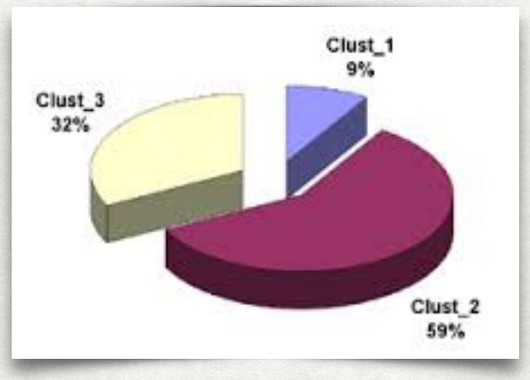
Tale distinzione è fondamentale poiché, come vedremo in seguito, ci guiderà nella scelta delle misure riassuntive e dei test statistici da utilizzare.

ASPETTI STATISTICI
La Raccolta dei Dati

- Le **variabili** numeriche **continue** possono assumere un numero infinito di valori all'interno di un certo ambito.
- Inoltre, la distanza che c'è fra 3 e 4, è la stessa esistente ad es. fra 20 e 21.
- Questo vuol dire che se consideriamo ad esempio il peso, tipica variabile continua, un soggetto che pesa 80 kg avrà un peso che è effettivamente doppio rispetto ad un soggetto che pesi 40 kg.
- Età, pressione arteriosa, glicemia, sono tutti esempi di variabili continue.

ASPETTI STATISTICI
La Raccolta dei Dati

- Le **variabili** numeriche **discrete** si differenziano da quelle continue poiché possono assumere solo un numero finito di valori all'interno di uno specifico intervallo (ad esempio, il numero di figli, nell'intervallo 2-5, può assumere solo il valore di 2, 3, 4, o 5).
- Nelle **variabili ordinali**, pur essendo i valori posti secondo un ordine predeterminato (ad esempio uno scompenso cardiaco di classe IV è più grave di uno di classe III, che a sua volta è più grave di uno scompenso di classe II), non c'è equidistanza fra i valori (non possiamo cioè affermare che uno scompenso di classe IV è il doppio grave di uno di classe II o quattro volte più grave di uno scompenso di classe I).
- Gli stadi di malattia, o le misure di qualità di vita sono tipicamente variabili ordinali.



- Infine, le **variabili nominali** esprimono una qualità del tipo " tutto o nulla", senza nessun ordine prestabilito.
- Ne sono un esempio il sesso, la razza, la presenza/assenza di una complicanza, etc.

ASPETTI STATISTICI

Definizioni di Variabile Numerica e Nominale

- **Variabile (Numerica) Continua**

  Caratterizzata da un infinito numero di valori possibili tra due valori qualsiasi. Si riferisce pertanto ad un insieme *continuo* di valori (ad es. pressione arteriosa, glicemia, etc.).

- **Variabile (Numerica) Discreta**

  Può assumere solo un numero finito di valori all'interno di uno specifico intervallo di numeri interi.

- **Variabile Ordinale**

  Caratterizzata da un ordine predeterminato per classificare le risposte (ad es. stadio di malattia, scala di dolore, etc.).

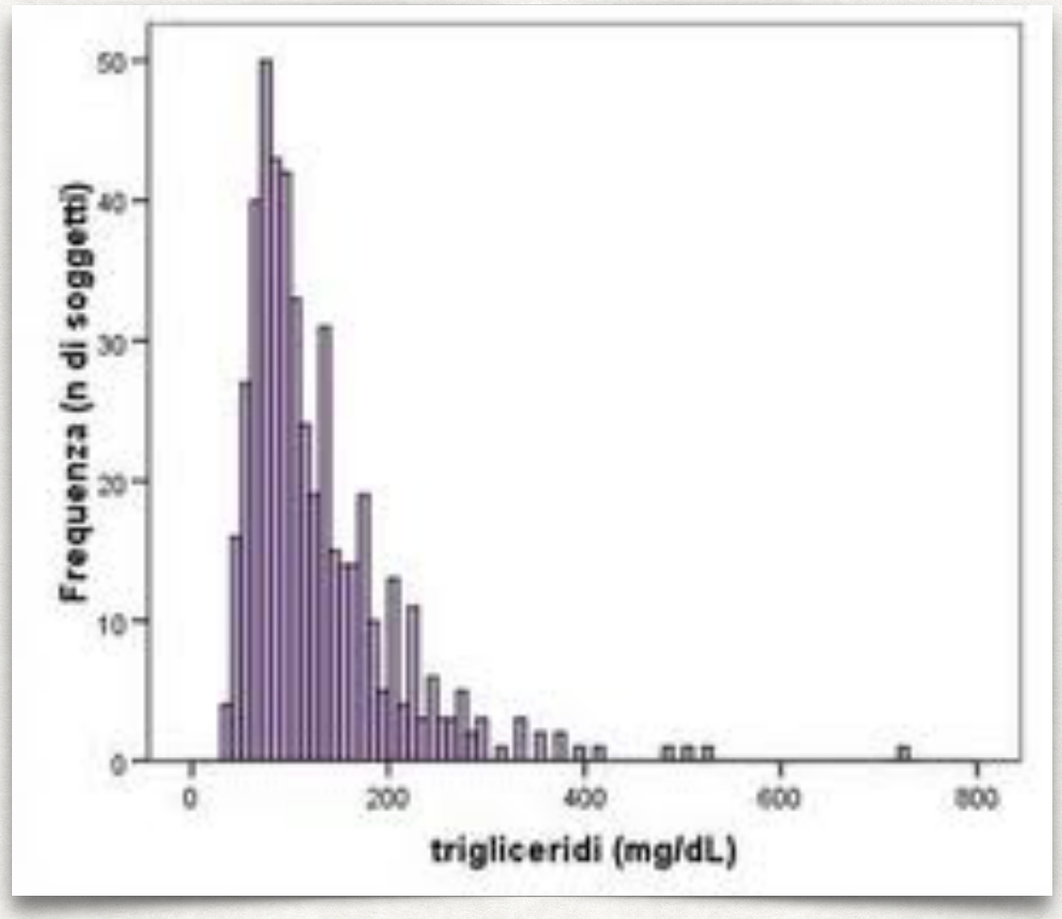
- **Variabile Nominale**

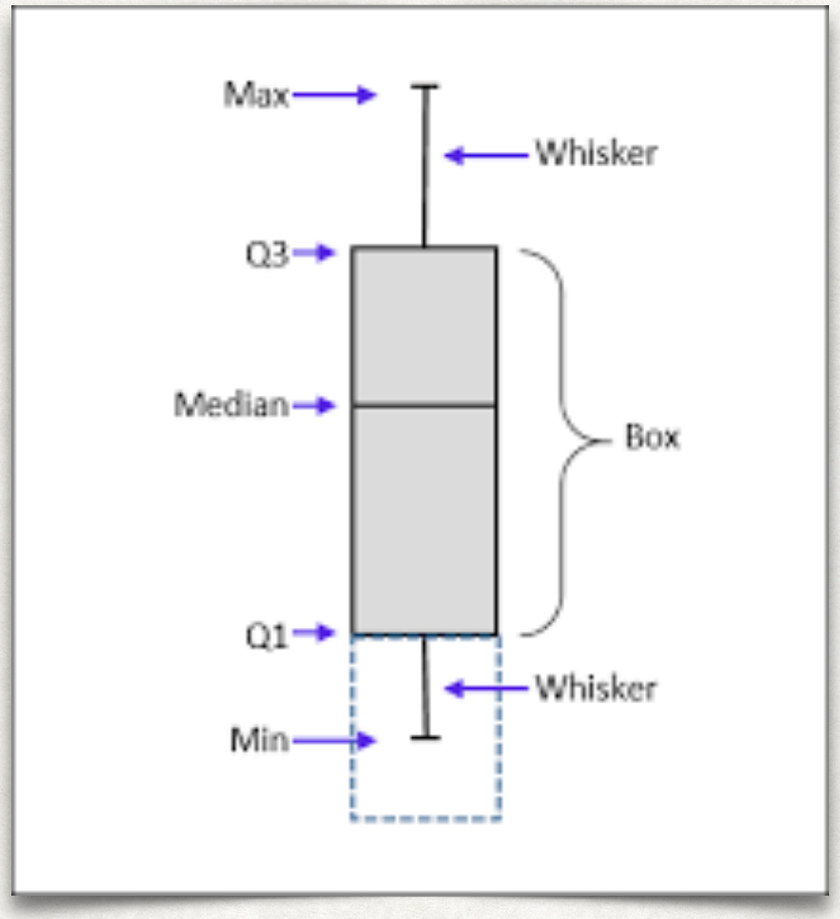
  Esprime una qualità (ad es. sesso, razza, stato vitale). Non implica alcun ordinamento, né valori intermedi tra due valori della variabile (ad es. la variabile "sesso" contempla i soli valori M oppure F, senza alcun ordinamento né valore intermedio).

RAPPRESENTAZIONE GRAFICA DEI DATI

VARIABILE CONTINUA

Una variabile numerica può essere rappresentata graficamente mediante un **istogramma** o un **box-plot**.

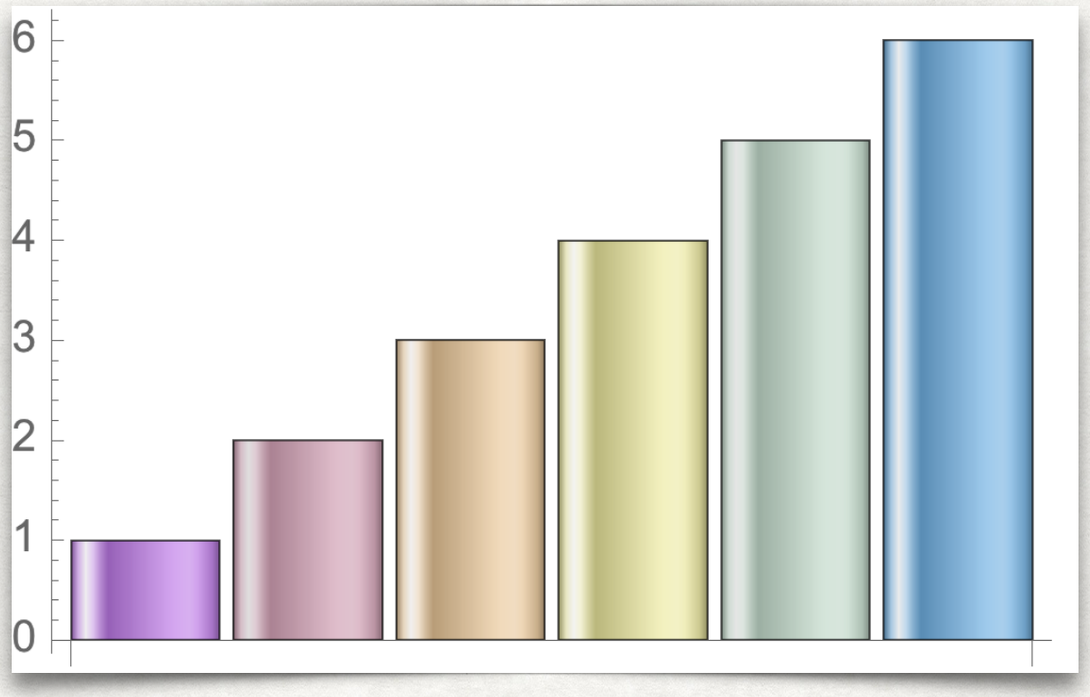


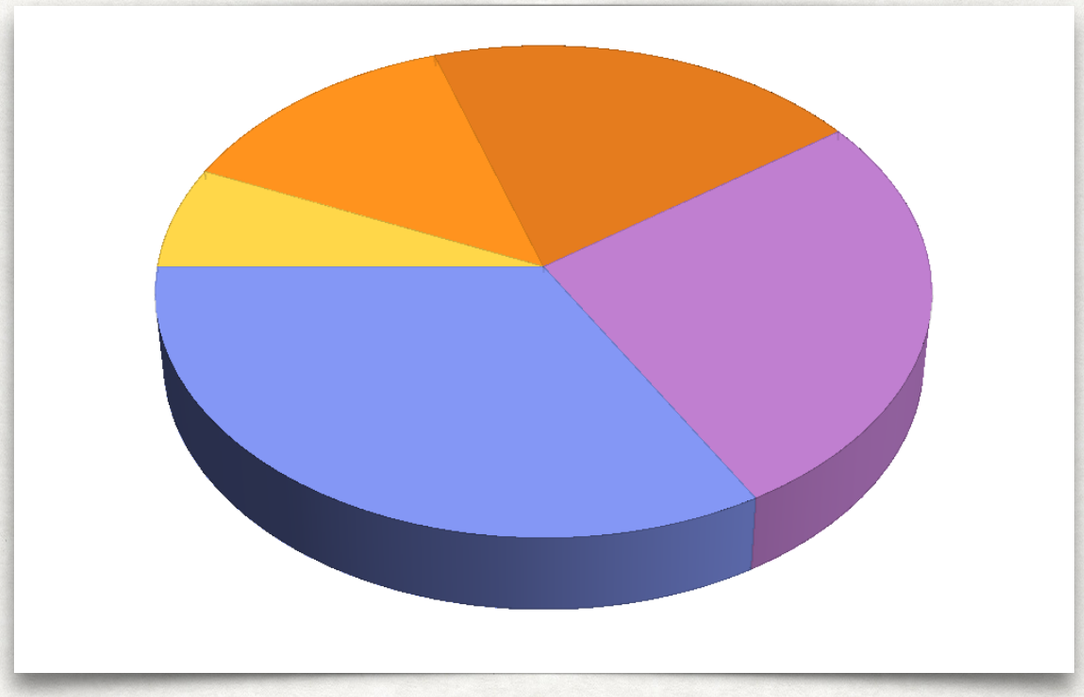


RAPPRESENTAZIONE GRAFICA DEI DATI

VARIABILE ORDINALE / NOMINALE

Una variabile ordinale o nominale viene rappresentata con un **grafico a barre** o **a torta**.





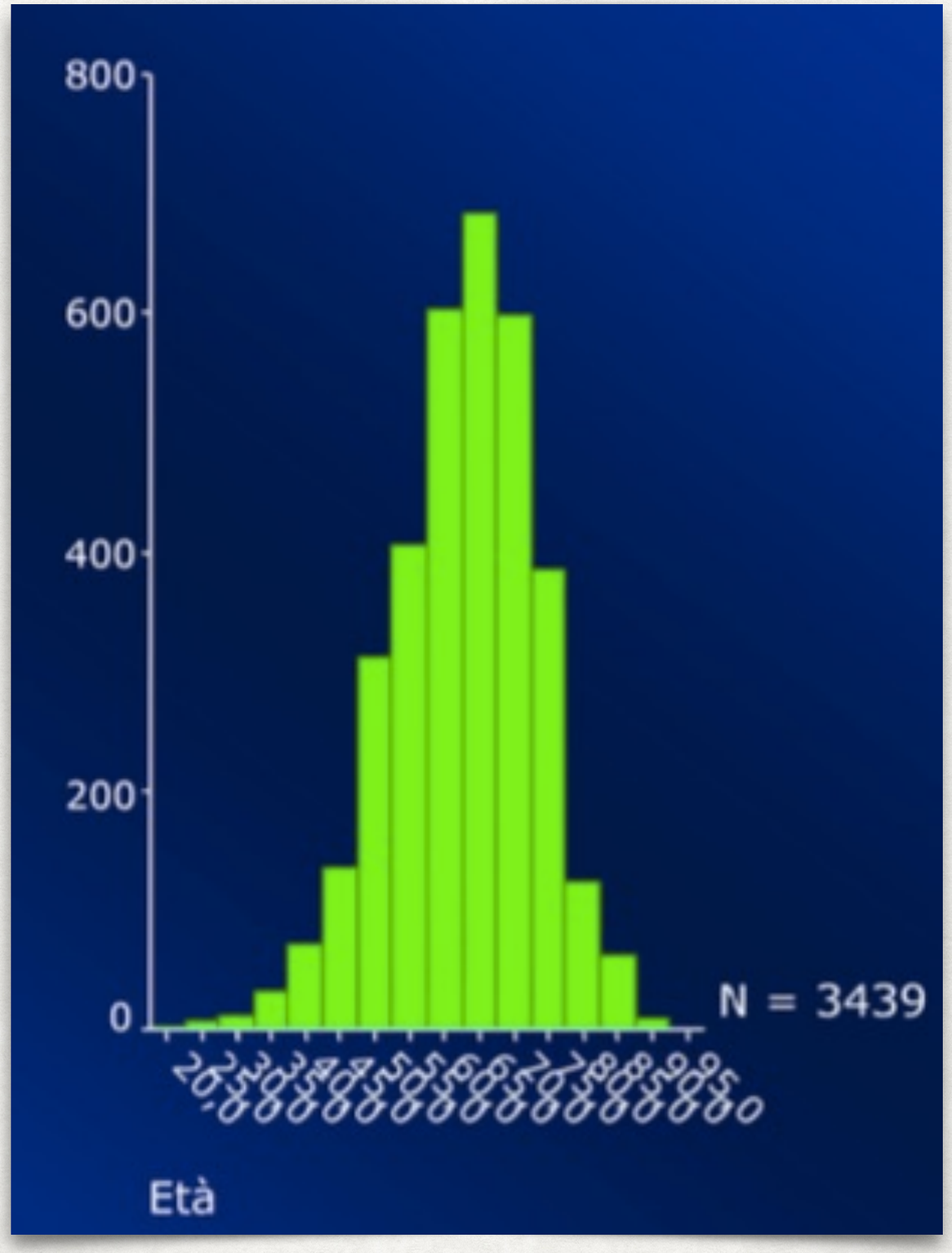
Una variabile discreta/ordinale può anche essere rappresentata con un box-plot.

RAPPRESENTAZIONE GRAFICA DEI DATI

Istogramma per Variabili Numeriche

L'istogramma mostra la distribuzione dei valori dividendo i valori in intervalli equamente distanziati e tracciando sotto forma di barra il conteggio dei casi in ciascun intervallo.

- *Un istogramma viene costruito dividendo i valori della variabile in oggetto in intervalli equamente distanziati, e riportando sotto forma di colonna il numero di soggetti che presentano un valore all'interno dell'intervallo.*
- *L'altezza della colonna è proporzionale al numero di soggetti in quell'intervallo.*
- *Nell'esempio in figura è stato tracciato l'istogramma dell'età in un campione di 3439 soggetti.*
- *L'età è stata divisa in intervalli di 5 anni, ed è stato fatto il conteggio di quanti soggetti avessero un'età che rientrava in ognuno degli intervalli.*
- *Dal grafico si può osservare come la maggior parte dei soggetti avesse un'età fra i 50 e i 70 anni, mentre solo pochi individui avevano un'età al di sotto dei 30 anni o al di sopra degli 80.*

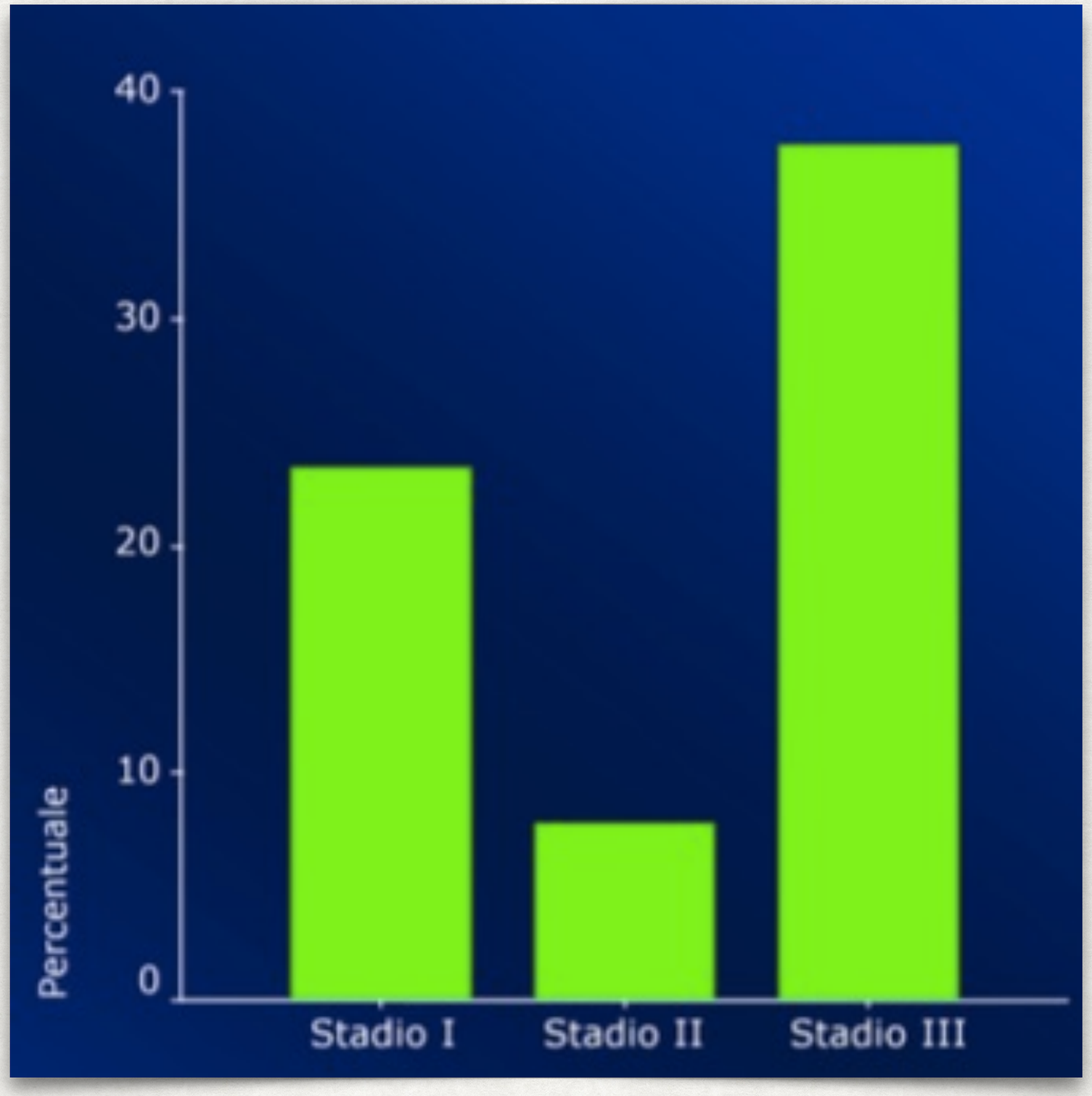


RAPPRESENTAZIONE GRAFICA DEI DATI

Grafico a Barre per Variabili Ordinali o Nominali

Il grafico a barre permette di semplificare il confronto visivo delle categorie di una variabile ordinale.

- *Nel caso di variabili ordinali o nominali, è possibile utilizzare i grafici a barre.*
- *Analogamente all'istogramma, l'altezza di ogni barra dipenderà dal numero di soggetti che rientrano in quella classe.*
- *Di solito si preferisce tuttavia riportare i dati come percentuali, piuttosto che come numeri assoluti.*
- *Nell'esempio riportato in figura, possiamo vedere quale sia la percentuale di soggetti con malattia in stadio I, II e III.*

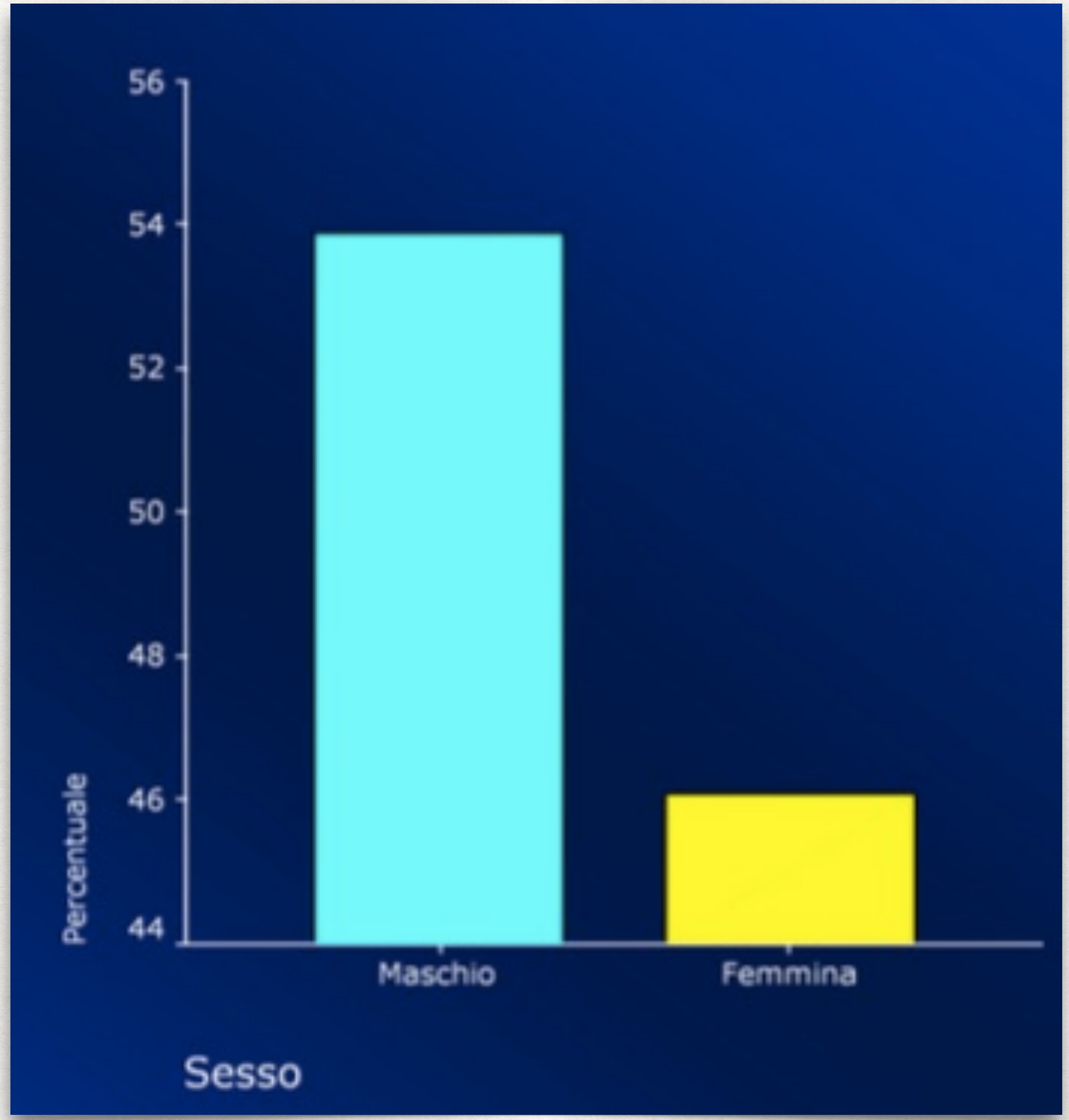


RAPPRESENTAZIONE GRAFICA DEI DATI

Grafico a Barre per Variabili Ordinali o Nominali

Il grafico a barre permette di presentare le caratteristiche riassuntive di una variabile nominale.

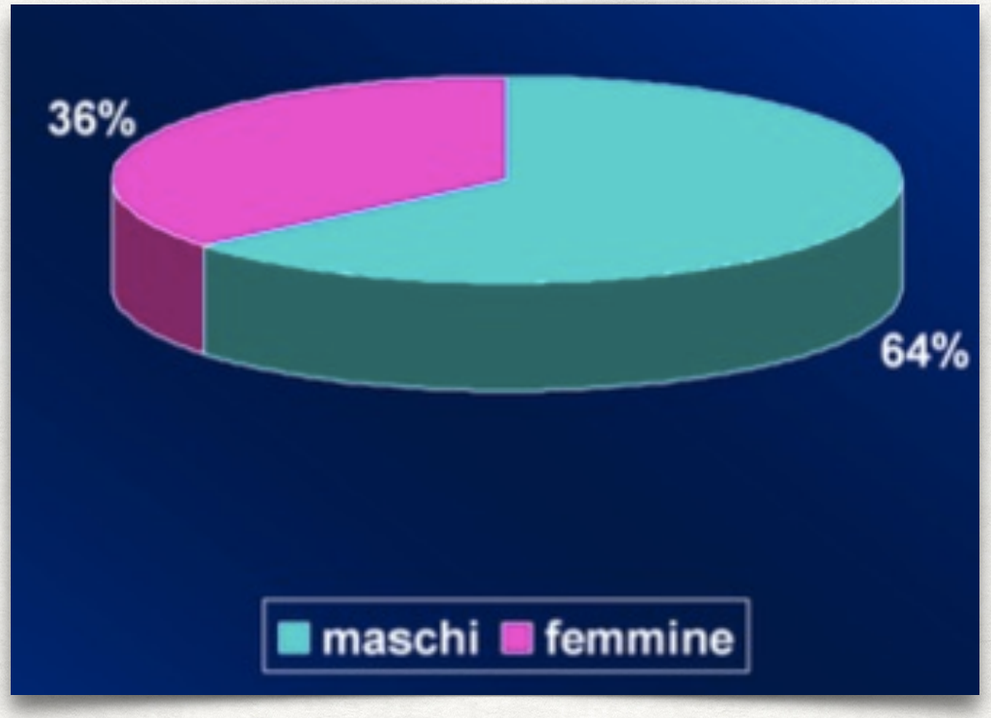
- *Analogamente, per una variabile nominale (in questo caso il sesso), l'altezza delle barre dipende dal numero (o dalla percentuale) di soggetti in ognuna delle classi.*



RAPPRESENTAZIONE GRAFICA DEI DATI

GRAFICO A TORTA

Come il grafico a barre, anche il grafico a torta permette di presentare graficamente la percentuale di soggetti in ogni classe (categoria) della variabile.



- *Il grafico a torta è una modalità alternativa di rappresentare variabili ordinali o nominali.*
- *Il grafico diventa tuttavia di difficile lettura se la variabile è divisa in molte classi.*

MISURE RIASSUNTIVE DEI DATI

MISURE RIASSUNTIVE DEI DATI

MEDIA, MEDIANA, MODA, RANGE, DEVIAZIONE STANDARD

Oltre che graficamente, è necessario riassumere le informazioni a disposizione sotto forma numerica.

A questo proposito, per una adeguata descrizione della distribuzione dei valori di una variabile nella popolazione in studio abbiamo necessità di due **misure**, una di *tendenza centrale*, l'altra di *dispersione*.

- *La misura di tendenza centrale ci dice attorno a quale valore tendono a raggrupparsi le osservazioni, mentre la misura di dispersione ci indicherà di quanto le singole osservazioni si discostano dal valore centrale.*
- *Per una variabile continua, la misura di tendenza centrale è rappresentata dalla* ***media****, mentre la misura di dispersione è rappresentata dalla* ***deviazione standard****.*
- *Per le variabili discrete si utilizzano invece la* ***mediana*** *e il* ***range****.*

MISURE RIASSUNTIVE DEI DATI

MEDIA, MEDIANA, MODA, RANGE, DEVIAZIONE STANDARD

Le variabili numeriche (continue e discrete) possono quindi essere riassunte da una **misura di tendenza centrale** e da una **misura di dispersione**.

Variabile	Misura di tendenza centrale	Misura di dispersione
Continua	Media (**μ**)	Deviazione standard (**σ**)
Discreta	Mediana	Range

- Per qualsiasi tipo di variabile è inoltre possibile calcolare la **moda**, cioè il valore più frequente nella distribuzione.
- Per le variabili ordinali e nominali non è possibile utilizzare misure riassuntive (i valori della variabile sono categorie, ordinate o meno), ma solo la **percentuale di casi** in ogni categoria.

MEDIA aritmetica μ

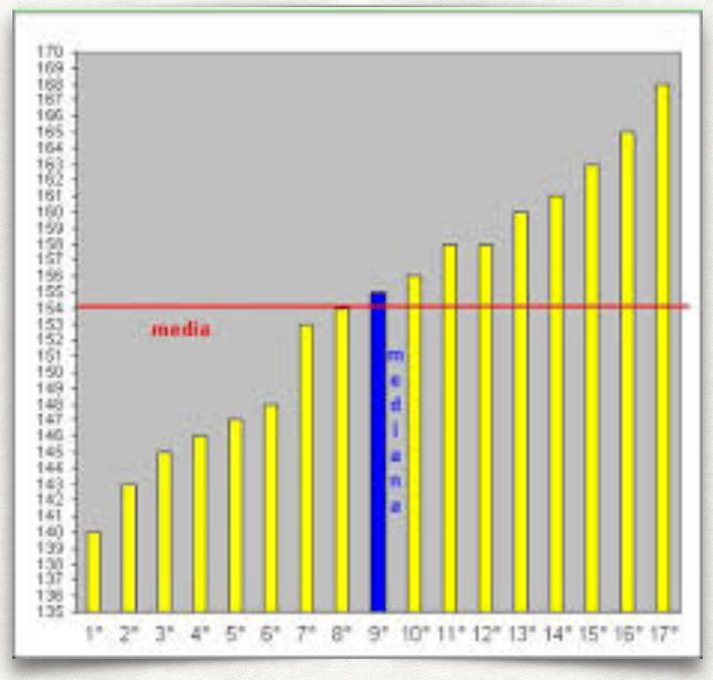
Somma (Σ) dei valori di una variabile (X) per ogni osservazione divisa per il numero di osservazioni (N):

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

- Se ad es. abbiamo le cinque osservazioni 2, 12, 6, 3, 7 la loro media sarà: (2 + 12 + 6 + 3 + 7) / 5 = 6
- Se invece abbiamo le sei misure 1, 2, 4, 5, 10, 20 la media sarà: (1 + 2 + 4 + 5 + 10 + 20) / 6 = 7

MEDIANA

Valore che divide la popolazione a metà.



- È il valore centrale per un numero dispari di osservazioni, mentre corrisponde alla media dei due valori centrali nel caso di un numero pari di osservazioni.
- Per determinarla, occorre prima disporre i valori in ordine crescente.
- Se ad es. abbiamo le 5 osservazioni: 2, 3, 6, 7, 12 la mediana sarà: 6.
- Se abbiamo le 6 osservazioni: 2, 3, 6, 7, 12, 15 la mediana sarà: (6+7)/2 = 6.5
- In altre parole, rappresenta quel valore che divide a metà la popolazione, così che il 50% ha un valore pari o più alto della mediana, e il 50% un valore pari o più basso.

DEVIAZIONE STANDARD σ (e VARIANZA σ^2)

Misura quanto i valori di una distribuzione si allontanano dalla media.

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

- Se ad es. abbiamo le 5 osservazioni: 2, 3, 6, 7, 12 la media è: 6
- La varianza σ^2 è data da: $[(2-6)^2+(3-6)^2+(6-6)^2+(7-6)^2+(12-6)^2] / 5$
  $= [16+9+0+1+36] / 5 = 62 / 5 = 12.4$
- Quindi la deviazione standard σ è data dalla radice quadrata di 12.4 cioè 3.52

RANGE

Si riferisce al valore minimo e al valore massimo riscontrati nel campione in studio.

- Disponendo in ordine crescente i valori, il range è dato dal primo e dall'ultimo valore nella sequenza.
- Se ad es. abbiamo le 5 osservazioni: 2, 3, 6, 7, 12 il range è: 2 — 12
- Nel caso di una variabile ordinale, il range (vale a dire il valore minimo e quello massimo riscontrati) ne misura la dispersione.

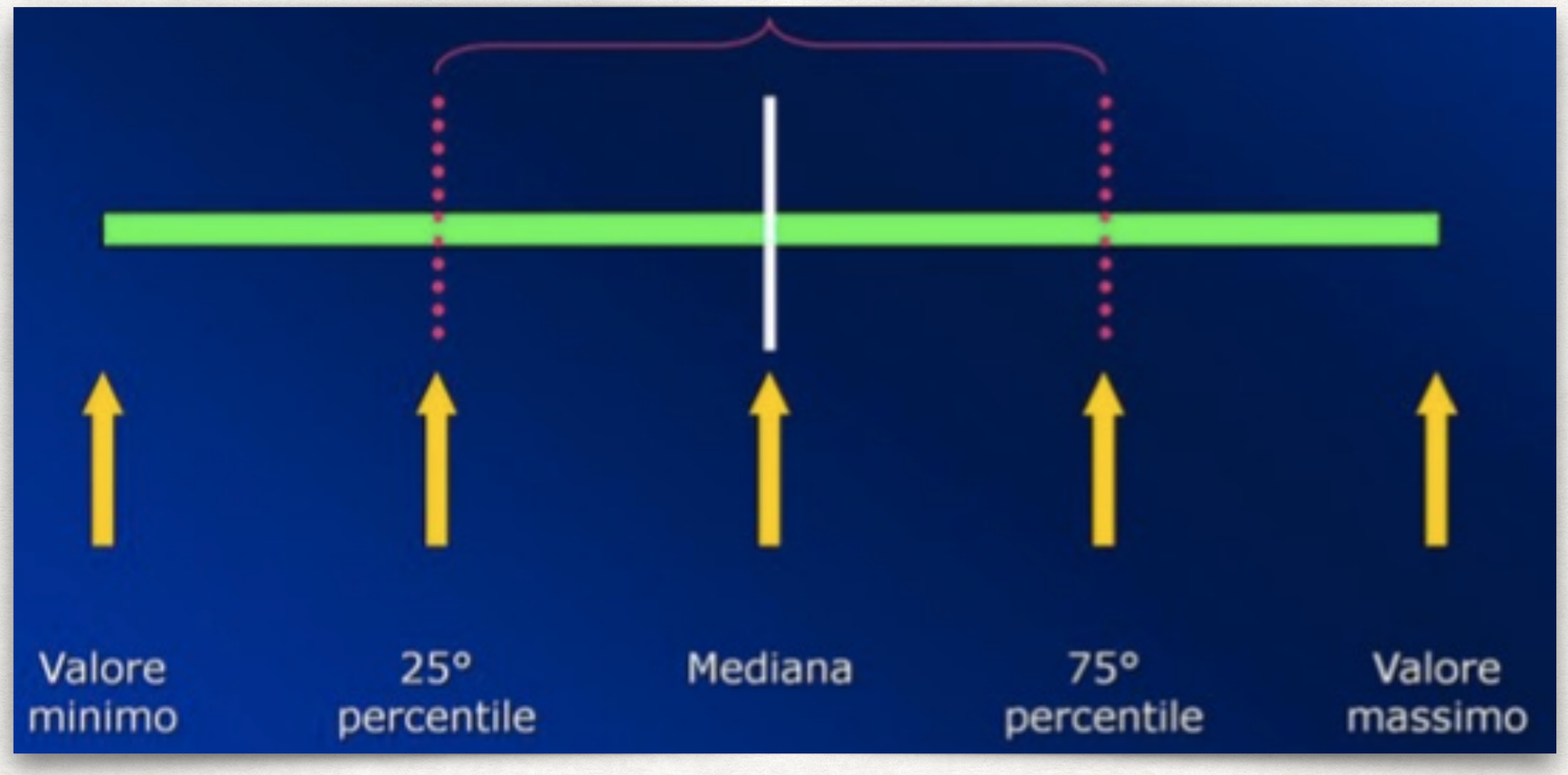
RANGE INTERQUARTILE

Si riferisce ai valori relativi al primo e al terzo quartile (25° e 75° percentile).

- Il range potrebbe dare una informazione imprecisa; infatti, il valore più alto e quello più basso potrebbero rappresentare valori estremi atipici, e non riflettere la reale variabilità della misura nella popolazione in esame.
- Se ad esempio avessimo 100 soggetti, tutti di età pari a 50 anni salvo due soggetti, uno di 10 anni e uno di 90, il range (10–90) potrebbe indurci a pensare che si tratti di una popolazione di età molto più variabile di quanto essa effettivamente sia.
- Per tale motivo, si utilizza di solito il range interquartile.
- Ad esempio i quartili sono i valori che dividono la popolazione in quattro gruppi, ognuno dei quali contiene il 25% del campione.

RANGE INTERQUARTILE

Si riferisce ai valori relativi al primo e al terzo quartile (25° e 75° percentile).

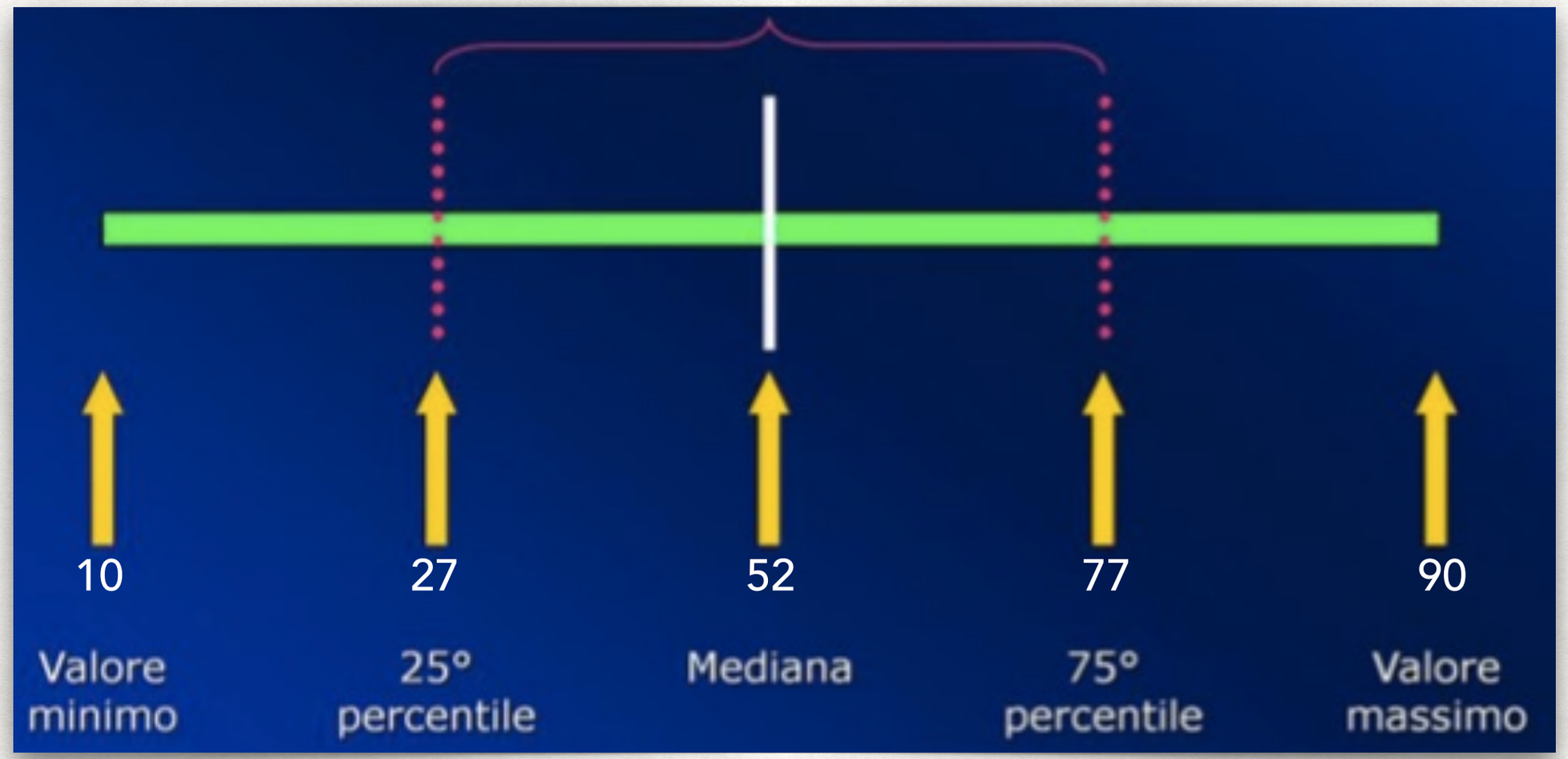


- Più in generale, i *percentili* sono i valori di una variabile che dividono la popolazione in studio in gruppi di uguale numerosità.
- Come è facile intuire, la mediana corrisponde al 50° percentile.
- Il range interquartile è quindi definito dai valori del 25° e del 75° percentile.

MISURE RIASSUNTIVE DEI DATI

ESEMPIO

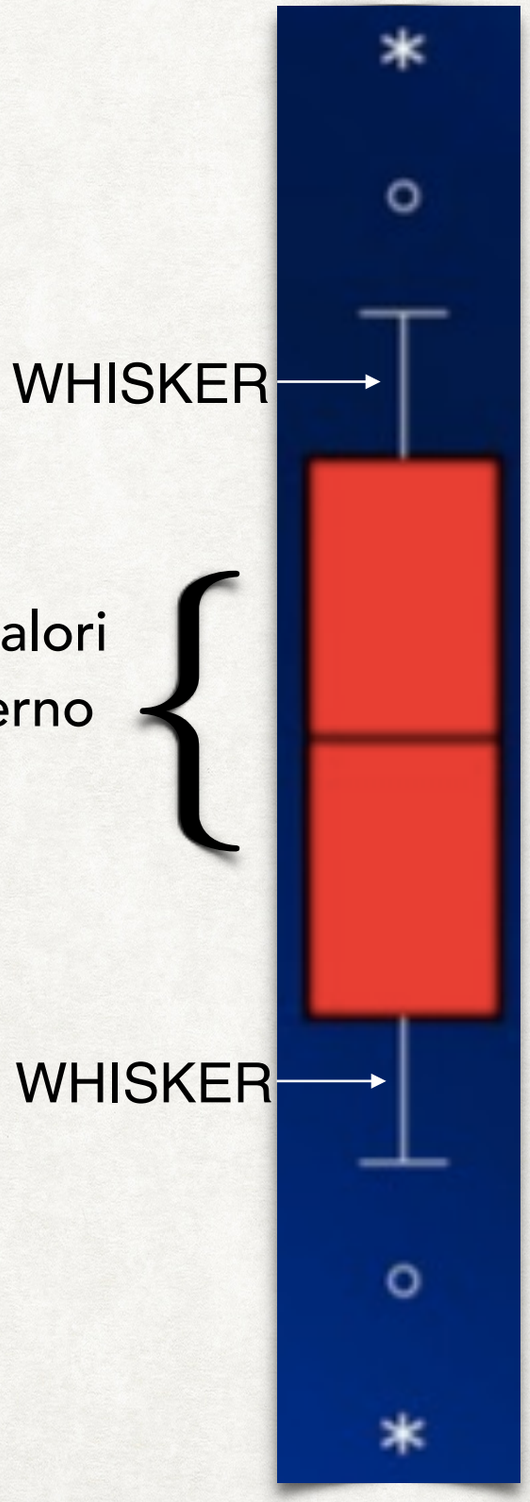
Età mediana di un campione di pazienti = 52 anni.



- Range: 10 — 90 anni.
- Range interquartile: 27 — 77 anni.
- Il 25% ha meno di 27 anni, il 50% ha tra 27 e 77 anni, il 25% ha più di 77 anni.

RAPPRESENTAZIONE DEI DATI

BOX-PLOT



valore estremo (far outlier): valore > Q3 + 3×(Q3–Q1)

outlier: valore > Q3+1.5×(Q3–Q1)

Max: valore non outlier più alto

WHISKER

Q3 (75° PERCENTILE)

Il 50% dei casi ha valori che cadono all'interno del **BOX**

MEDIANA

Q1 (25° PERCENTILE)

WHISKER

Min: valore non outlier più basso

outlier: valore < Q1 – 1.5×(Q3–Q1)

valore estremo (far outlier): valore < Q1 – 3×(Q3–Q1)

RAPPRESENTAZIONE DEI DATI

BOX-PLOT

- La definizione dei percentili e del range interquartile ci permette di descrivere un'altra modalità grafica di rappresentazione dei dati, rappresentata dai **box plot**.
- Il box plot è un grafico a forma di "scatola" che contiene molte informazioni riguardo la nostra variabile.
- La linea centrale che taglia il box rappresenta la mediana, mentre gli estremi del box rappresentano il range interquartile; in altre parole, il 50% delle osservazioni ha un valore compreso all'interno del box.
- Il grafico ci mostra inoltre i cosiddetti "outliers", vale a dire quei valori che sono più bassi del 25° percentile di oltre 1.5 volte la lunghezza del box o che sono più alti del 75° percentile di oltre 1.5 volte la lunghezza del box.

RAPPRESENTAZIONE DEI DATI

BOX-PLOT

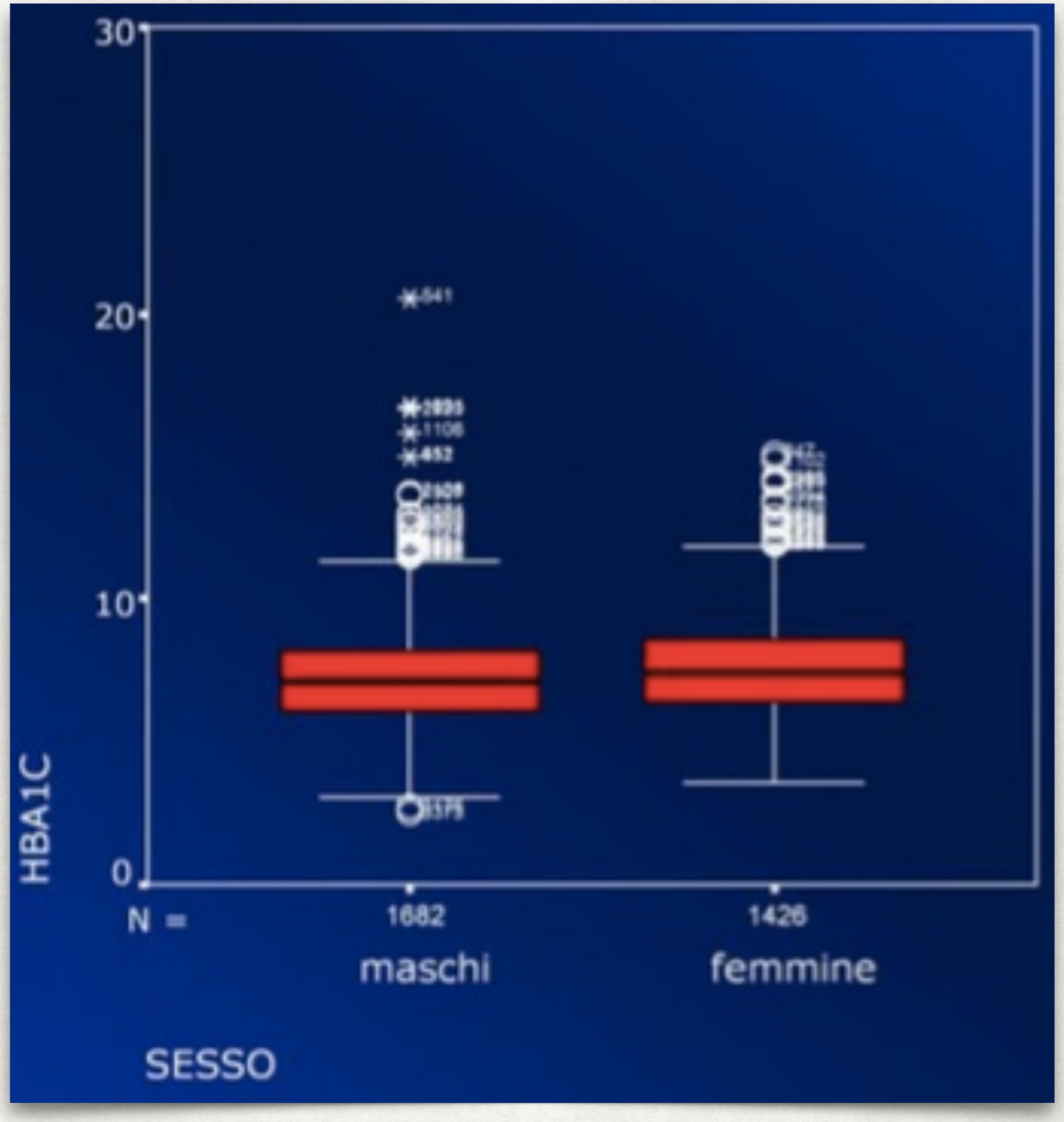
- Infine, vengono rappresentati i valori estremi, cioè quei valori che si discostano dal 25° e dal 75° percentile di oltre 3 volte la lunghezza del box.
- Queste informazioni sono importanti perché ci permettono di identificare quei valori che necessitano di verifica, perché potrebbero trattarsi di errori nell'immissione dei dati.
- Qualora non si trattasse di errori di digitazione, è comunque necessario verificare se si tratti di valori biologicamente plausibili o di errori di misurazione.
- Se i valori sono plausibili, non vanno ovviamente rimossi.

RAPPRESENTAZIONE DEI DATI

BOX-PLOT

- In questo esempio, i box plot sono stati utilizzati per riassumere graficamente i valori di emoglobina glicosilata (HbA_{1c}) in una popolazione di individui affetti da diabete, riassumendo il dato separatamente per maschi e femmine.

- Il grafico mostra che, mentre nelle femmine esistono solo alcuni outliers, contraddistinti dai cerchietti, fra i maschi sono presenti anche alcuni valori estremi, rappresentati dagli asterischi.



RAPPRESENTAZIONE DEI DATI

Forma Tabellare

- I dati, oltre che in forma grafica, vengono generalmente riassunti in **forma tabellare**.
- Le **variabili numeriche** vengono espresse come media e deviazione standard, o come mediana e range (o range interquartile).
- Le **variabili categoriche** vengono espresse come percentuali.

Questa diapositiva riporta un esempio di riassunto in forma tabellare delle caratteristiche di una popolazione, divise per braccio dello studio.

Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial

	Diastolic blood pressure target group		
	≤90 mm Hg (n=6264)	≤85 mm Hg (n=6264)	≤80 mm Hg (n=6262)
Age (years)	61·5 (7·5)	61·5 (7·5)	61·5 (7·5)
Body-mass index (kg/m²)	28·4 (4·7)	28·5 (4·7)	28·4 (4·6)
Diastolic blood pressure (mm Hg)	105 (3·4)	105 (3·4)	105 (3·4)
Systolic blood pressure (mm Hg)	170 (14·4)	170 (14·0)	170 (14·1)
Serum creatinine (μmol/L)	89 (26)	89 (23)	89 (23)
Serum cholesterol (mmol/L)	6·0 (1·1)	6·1 (1·1)	6·1 (1·2)
Men/women (%)	53/47	53/47	53/47
Previous treatment (%)	52·3	52·7	52·6
Smokers (%)	15·9	15·8	15·9
Previous MI (%)	1·6	1·5	1·5
Other previous CHD (%)	5·9	6·0	5·9
Previous stroke (%)	1·2	1·2	1·2
Diabetes mellitus (%)	8·0	8·0	8·0

Data are mean (SD) or % of group. MI=myocardial infarction; CHD=coronary heart disease.

Table 1: **Characteristics at randomisation**

Lancet1998; 351: 1755–62

DISTRIBUZIONE NORMALE (GAUSSIANA) E ASIMMETRICA

DISTRIBUZIONE NORMALE (O GAUSSIANA)

Definizioni

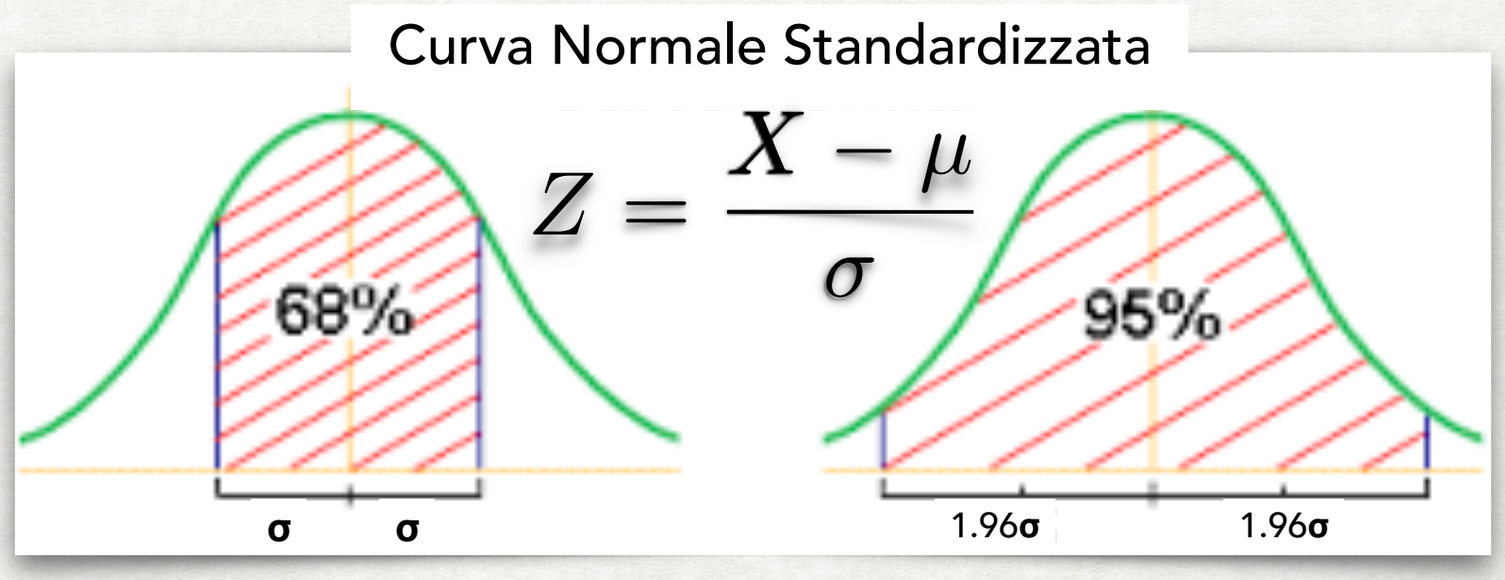
Una variabile continua X è **distribuita normalmente** quando media, moda e mediana coincidono.

Inoltre, una distribuzione normale Z è **standardizzata** se ha media nulla e deviazione standard unitaria (si ottiene sottraendo ad ogni osservazione la media e dividendo per la deviazione standard). In una distribuzione normale:

- circa 2/3 delle osservazioni (68%) si trovano entro 2 deviazioni standard (μ±1σ);
- circa il 95% delle osservazioni si trovano entro quasi 4 deviazioni standard (μ±1.96σ);
- circa il 99% delle osservazioni si trovano entro quasi 6 deviazioni standard (μ±2.58σ).

Inoltre, in una distribuzione normale:

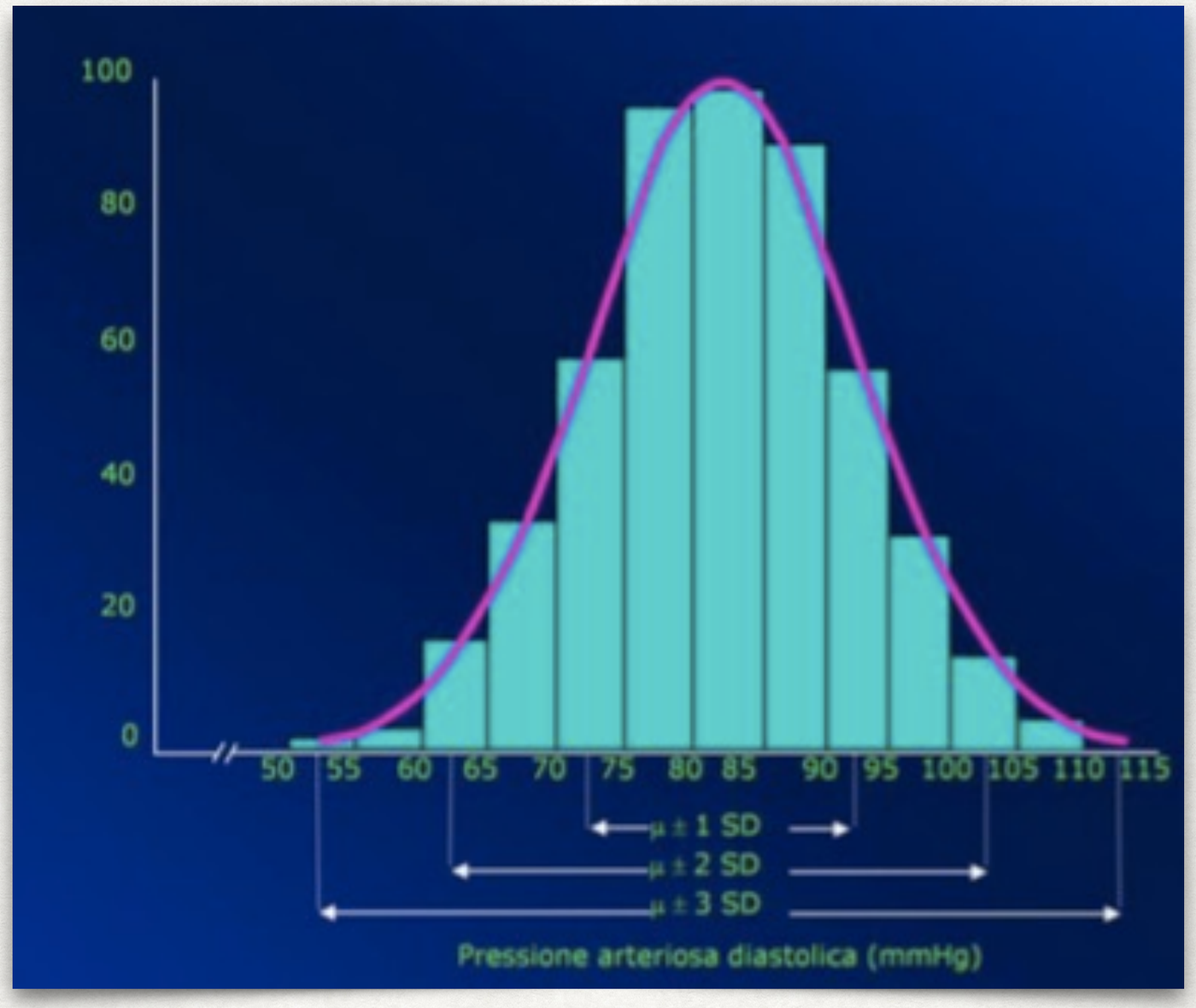
- 2.5° percentile ≃ μ–2σ
- 16° percentile = μ–1σ
- 50° percentile = μ
- 84° percentile = μ+1σ
- 97.5° percentile ≃ μ+2σ



DISTRIBUZIONE NORMALE (O GAUSSIANA)

DEFINIZIONI

- Il grafico mostra una classica distribuzione normale, riferita ai valori di pressione diastolica in una popolazione.
- Dall'immagine si può notare come, in presenza di una distribuzione normale, molti soggetti tendono ad avere un valore vicino alla media, mentre, man mano che ci allontaniamo dal valore medio, troviamo sempre meno soggetti.
- In particolare, solo il 2.5% dei soggetti avrà un valore superiore a $\mu+1.96\sigma$, e analogamente solo il 2.5% dei soggetti avrà un valore inferiore a $\mu-1.96\sigma$.

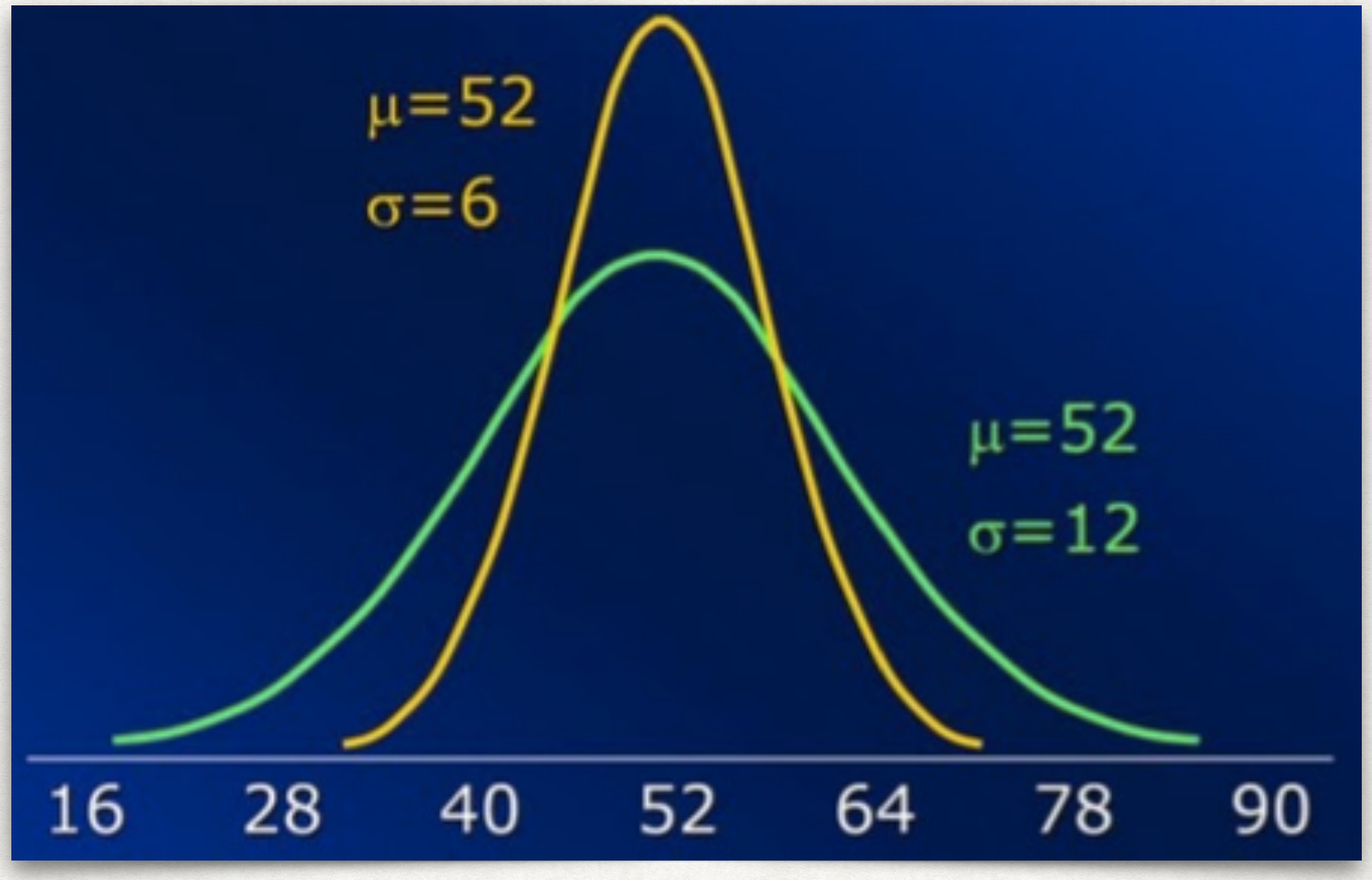


- Inoltre, i soggetti con valori superiori a $\mu+2.58\sigma$ e quelli con valori inferiori a $\mu-2.58\sigma$ saranno soltanto lo 0.5% in entrambi i casi.

DISTRIBUZIONE NORMALE (O GAUSSIANA)

Definizioni

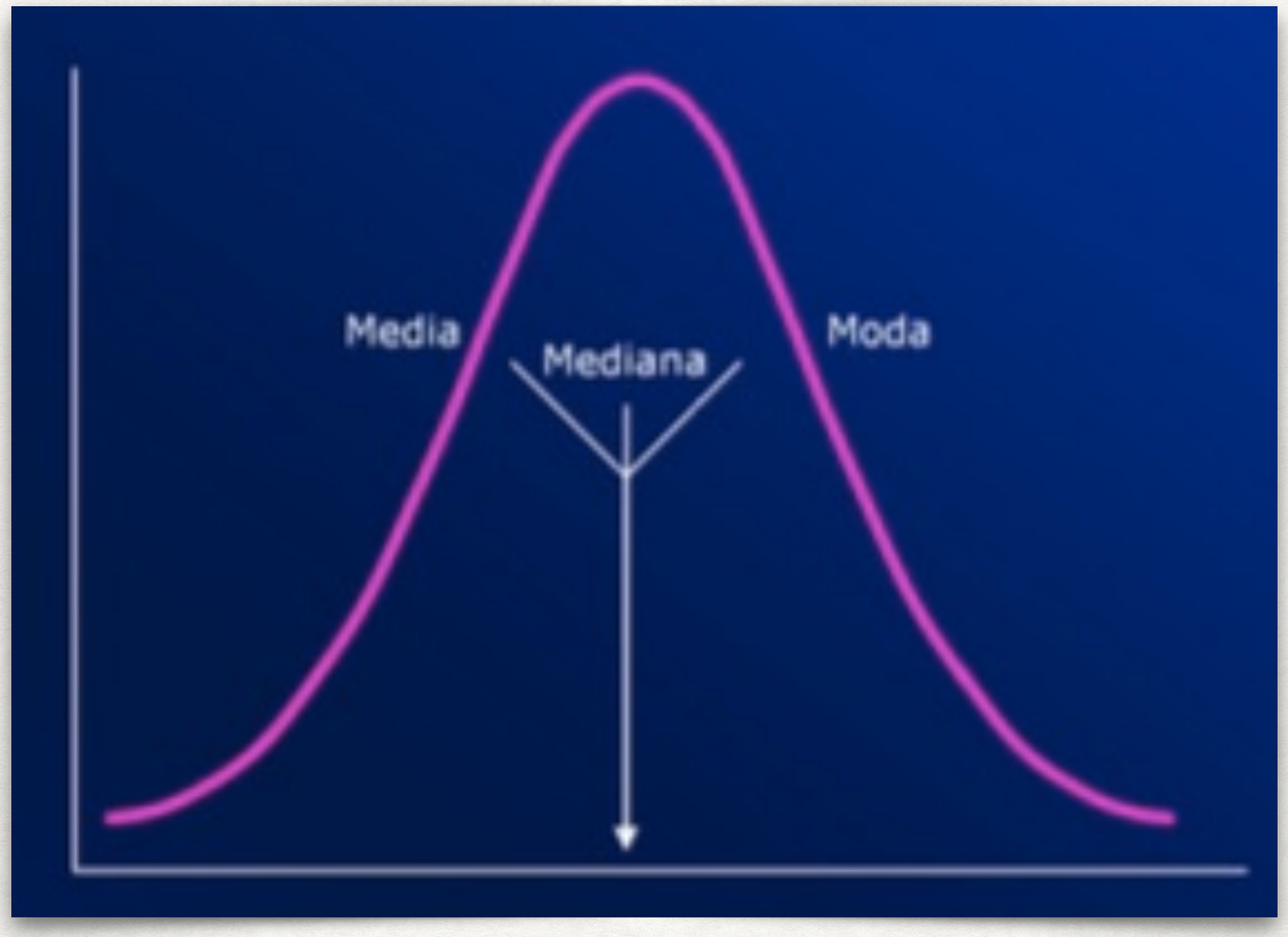
- Le due curve rappresentano la distribuzione dell'età in due campioni.
- Come si può constatare, il valore medio è lo stesso, ma, nel caso della curva verde, la deviazione standard è doppia rispetto a quella gialla.



- Questo vuol dire che i soggetti appartenenti alla popolazione rappresentata in giallo hanno un'età più omogenea e meno variabile di quelli appartenenti alla popolazione in verde.

DISTRIBUZIONE NORMALE (O GAUSSIANA)

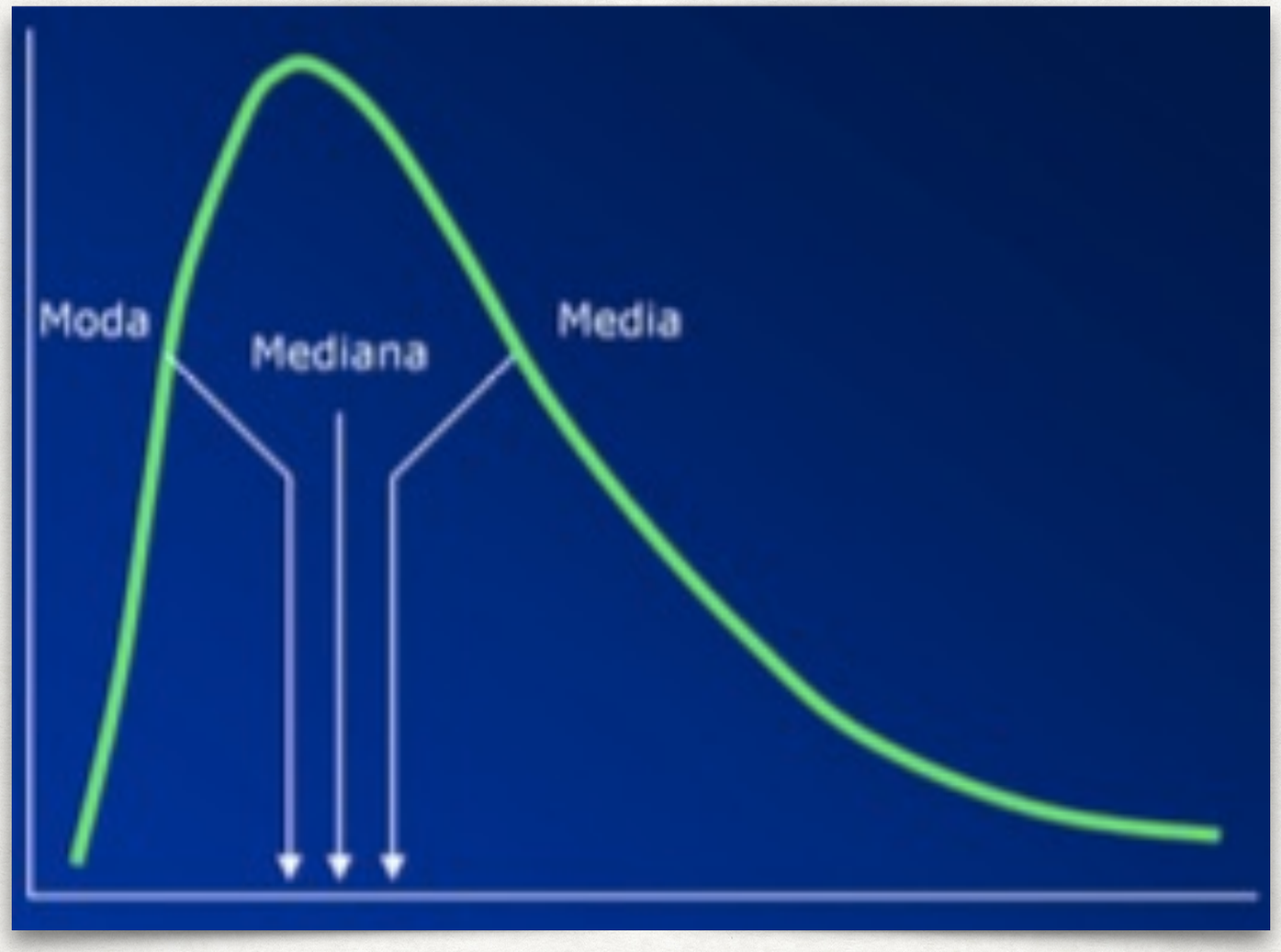
DEFINIZIONI



- Come già precisato, una delle condizioni perché una variabile continua possa essere considerata normalmente distribuita è rappresentata dalla coincidenza fra media, moda e mediana.
- In altri termini, possiamo dire che la curva è **simmetrica** rispetto al valore centrale.

DISTRIBUZIONE ASIMMETRICA A DESTRA

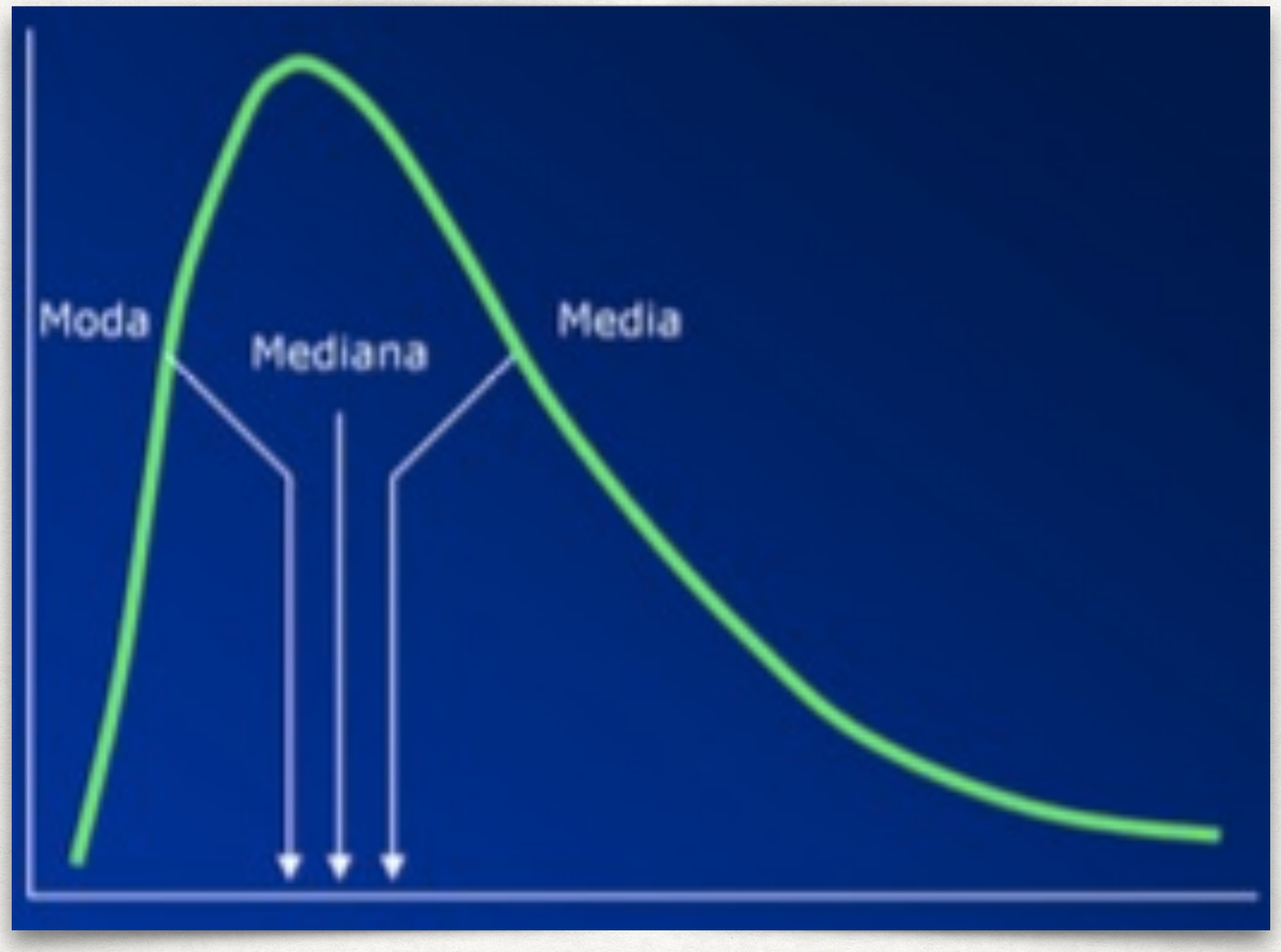
Right Skewness



- Può tuttavia capitare che una variabile continua assuma valori che non sono simmetricamente distribuiti.
- Ad esempio, in alcuni casi può essere presente una certa percentuale di soggetti che hanno un valore più alto rispetto al resto del campione, conferendo alla curva un aspetto asimmetrico con una "coda" a destra.

DISTRIBUZIONE ASIMMETRICA A DESTRA

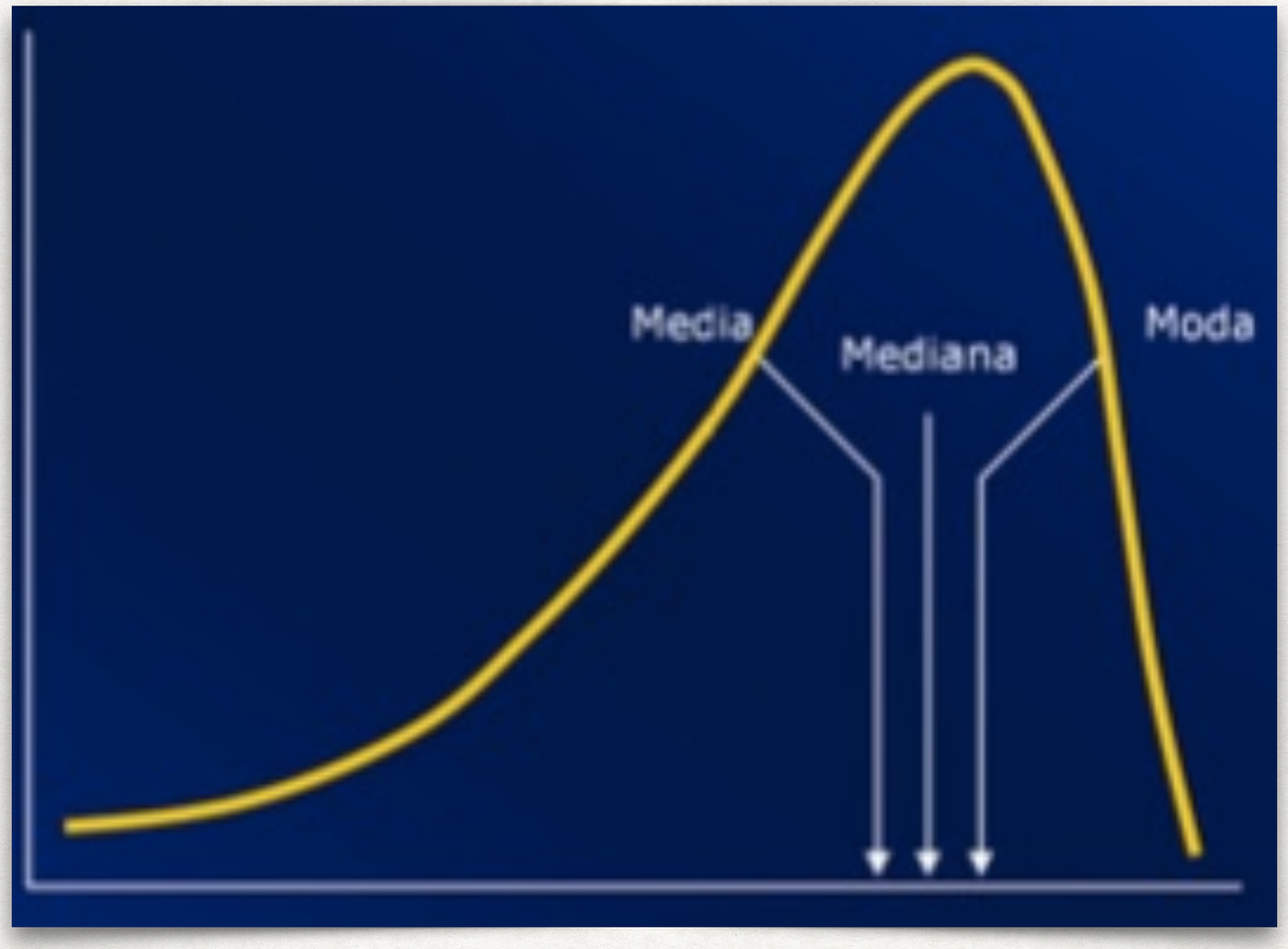
Right Skewness



- Ciò potrebbe verificarsi se ad esempio si misurassero i livelli di glicemia in una popolazione.
- La maggior parte dei soggetti avrà valori nella norma, ma una piccola percentuale potrebbe avere valori consistentemente superiori, perché ad esempio affetti da diabete senza saperlo.
- In questi casi, il valore della media sarà superiore a quello della mediana.

DISTRIBUZIONE ASIMMETRICA A SINISTRA

Left Skewness

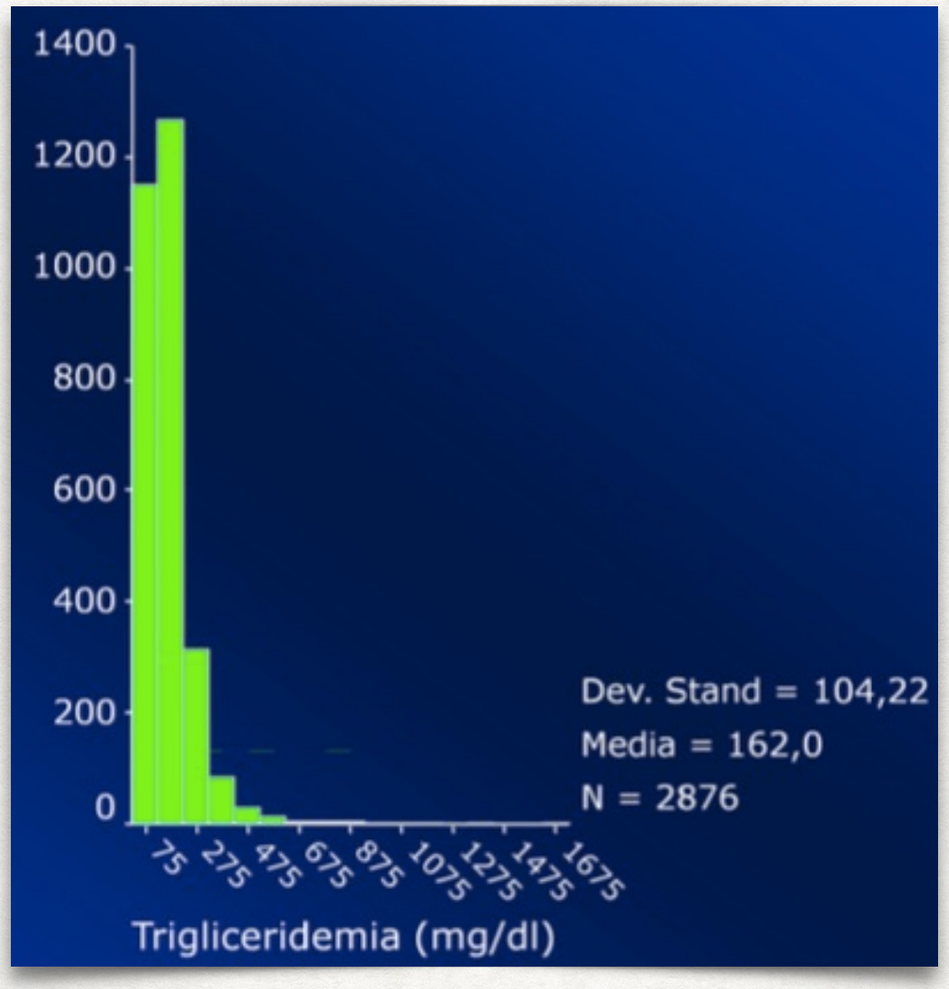


- In altri casi potrebbe verificarsi la situazione opposta, con una certa percentuale di soggetti che presentano valori più bassi rispetto al resto della popolazione.
- In questo caso la curva presenterà una "coda" a sinistra e il valore della media sarà inferiore a quello della mediana.

TRASFORMAZIONE DEI DATI

Da Distribuzione Asimmetrica a Normale

- A volte una variabile con una distribuzione chiaramente asimmetrica può essere matematicamente trasformata così da renderne normale la distribuzione.
- Nell'esempio riportato nel grafico, riguardante i valori dei trigliceridi nel sangue in un campione di soggetti con diabete, si può chiaramente vedere come la distribuzione sia asimmetrica, con una "coda" a destra.
- Tale distribuzione è determinata dalla presenza di una certa proporzione di pazienti con valori di trigliceridemia marcatamente più elevati rispetto al resto del campione.

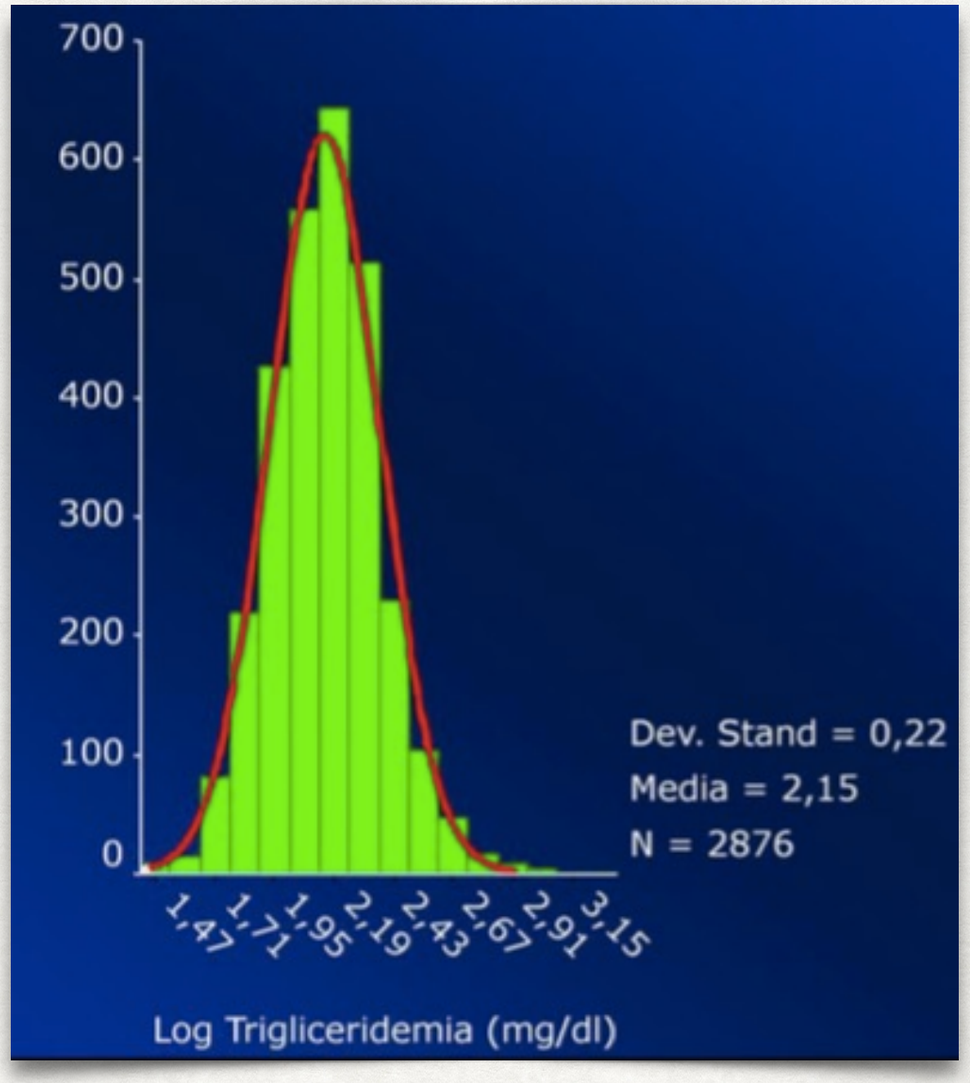


- Che la distribuzione non sia normale lo si evince anche osservando la media e la deviazione standard.
- Come già sottolineato, se la distribuzione fosse normale, il 95% dei valori dovrebbe cadere entro l'intervallo $\mu \pm 1.96\sigma$.
- Se tuttavia sottraiamo al valore medio, pari a 162 mg/dl, il doppio di σ (cioè 208.44) otterremo un numero negativo!
- In questi casi una trasformazione matematica della variabile può talvolta permetterci di ricondurre la distribuzione alla normalità.

TRASFORMAZIONE DEI DATI

DA DISTRIBUZIONE ASIMMETRICA A NORMALE

- Questo grafico riporta la distribuzione del logaritmo della trigliceridemia.
- Confrontando questo istogramma con il precedente si vede chiaramente come la distribuzione sia ora simmetrica e non presenti più la lunga coda a destra, riacquistando il classico aspetto a campana rovesciata.
- L'uso del logaritmo, della radice quadrata, o di altre funzioni matematiche può quindi talvolta essere utile per ottenere la normalità (vedremo in seguito che il requisito di normalità di una variabile continua è indispensabile per l'uso di alcuni test statistici, chiamati parametrici).



TEST DI IPOTESI

IL TEST DI IPOTESI

Una volta riassunti i dati in forma grafica o tabellare, il passo successivo consiste nell'esplorare se esistano particolari associazioni (ad es. "*Esiste un rapporto tra obesità e controllo metabolico?*"; "*I valori di HbA_{1c} sono più alti nei soggetti con basso livello di scolarità?*").

TEST STATISTICI

I test statistici servono ad accettare o a respingere (confutare) un'ipotesi

IL TEST DI IPOTESI

- Ogni test statistico parte dall'**ipotesi nulla** di assenza di differenze.
- Nell'esempio precedente l'ipotesi nulla è rappresentata da assenza di differenze nei valori medi di HbA_{1c} in base al BMI (assenza di associazione).
- Scopo del test statistico è quello di accettare l'ipotesi nulla o di confutarla, accettando quindi l'**ipotesi alternativa** che esistano differenze significative.

IL TEST DI IPOTESI

Tutti i test statistici partono dall'**ipotesi nulla** che non esista una relazione fra le variabili in studio.

Se ad esempio volessimo testare l'ipotesi che, fra i soggetti con diabete, il livello di controllo metabolico dipende dal livello di obesità, espresso dall'indice di massa corporea (o BMI), l'ipotesi nulla di partenza sarebbe rappresentata da una completa assenza di correlazione fra controllo metabolico e BMI.

Scopo del test statistico sarà quello di suggerirci se accettare questa ipotesi nulla, o se invece rifiutarla, accettando quindi l'**ipotesi alternativa** che ci sia un rapporto significativo fra controllo metabolico e livello di obesità.

IL TEST DI IPOTESI

- Poiché i soggetti studiati, per numerosi che siano, rappresentano sempre un piccolo campione di tutti quelli potenzialmente arruolabili e a causa della variabilità biologica, **i risultati di un test statistico saranno sempre in termini di probabilità**.
- Questo perché **il campione da noi studiato potrebbe non essere rappresentativo** dell'universo dei pazienti e le conclusioni a cui arriviamo potrebbero pertanto essere erronee.
- Nell'accettare o confutare l'ipotesi nulla è quindi **sempre possibile commettere un errore**; se, tuttavia, la probabilità di commettere tale errore è molto bassa, allora accetteremo con "sufficiente" fiducia le conclusioni a cui siamo arrivati.

In base a quale regola accettiamo o respingiamo l'ipotesi nulla?

Valore di *p* (*p*-value)

Indica la probabilità di commettere un errore rifiutando l'ipotesi nulla, e cioè la probabilità di sbagliare affermando che ci sia una differenza tra i gruppi messi a confronto.

$p \leq 0.05$ → Risultato statisticamente significativo → Rifiutiamo l'ipotesi nulla

IL TEST DI IPOTESI

ESEMPIO

- I soggetti con BMI compreso tra 25 e 27 hanno una HbA_{1c} di 7.28±1.7
- I soggetti con BMI>30 hanno una HbA_{1c} di 7.51±1.6

Ipotesi nulla: i livelli medi di HbA_{1c} non differiscono in base al BMI.

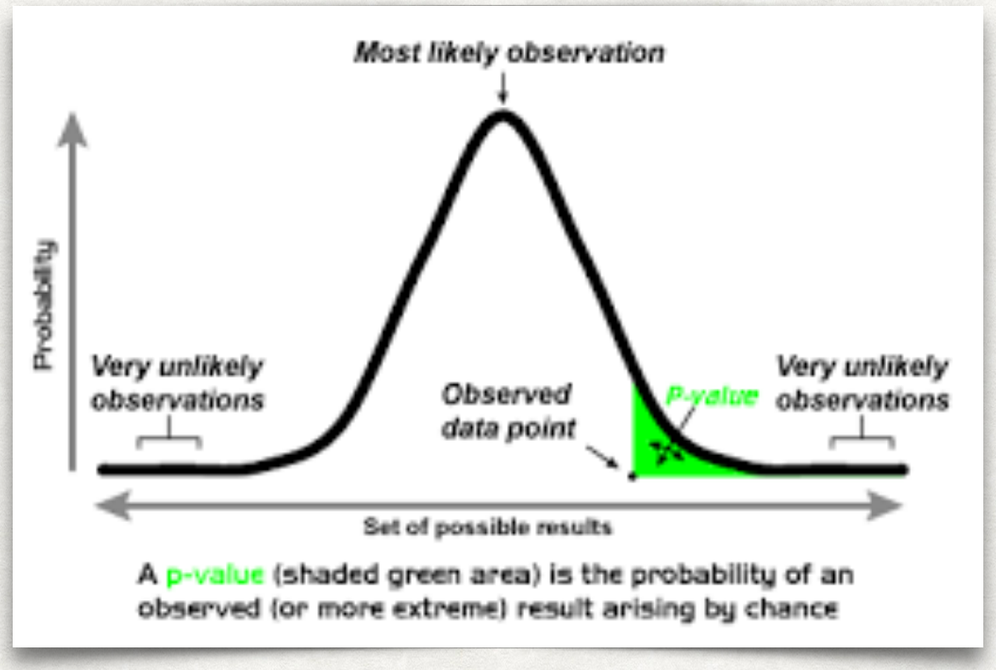
Ipotesi alternativa: i livelli medi di HbA_{1c} sono significativamente più alti in soggetti con BMI>30 rispetto a quelli con BMI tra 25 e 27.

Utilizzando il test statistico appropriato [test *t* di Student per dati non appaiati (*)] il valore di *p* risulta essere di 0.009

(*) *Più avanti illustreremo i criteri per la scelta del test statistico appropriato in relazione alla distribuzione.*

$p = 0.009$

- Rifiutando l'ipotesi nulla e affermando quindi che esiste una differenza in termini di HbA_{1c} in base ai livelli di BMI, avremo una probabilità inferiore all'1% di sbagliare (9 su mille).
- In altre parole, la probabilità che non vi sia associazione è inferiore all'1%.



IL TEST DI IPOTESI

ESEMPIO

Un risultato statisticamente significativo non equivale a dire che esso è clinicamente rilevante.

- Supponiamo che trattando 10000 diabetici con l'ipoglicemizzante A si ottengano valori di HbA_{1c} di 7.1±1.3 e che trattando altrettanti soggetti con l'ipoglicemizzante B si ottengano valori di HbA_{1c} di 7.3±1.4
- Applicando il test statistico appropriato si ottiene un valore di $p<0.001$
- La differenza è statisticamente significativa ed è altamente improbabile che sia dovuta al caso.
- Ma è anche rilevante da un punto di vista clinico?

IL TEST DI IPOTESI

- Attenzione! Molto spesso la **significatività statistica** viene confusa con la **rilevanza clinica** del risultato ottenuto.
- Il fatto che il risultato ottenuto sia statisticamente significativo implica solo che è molto probabile che questo risultato sia vero, e non dovuto al caso.
- **Il valore di *p* non può tuttavia dirci se tale risultato è anche importante dal punto di vista clinico, poiché questo è un giudizio che spetta solo a chi sta valutando i risultati, e prescinde dalla statistica.**

IL TEST DI IPOTESI

- Se ad esempio su un campione di 20 000 pazienti con diabete confrontiamo l'efficacia di due farmaci ipoglicemizzanti ed otteniamo valori medi di HbA_{1c} di 7.1±1.3 con il farmaco A e valori medi di 7.3±1.4 con il farmaco B, tale differenza nei valori medi risulterà altamente significativa.
- È sufficiente questo dato per affermare che bisogna preferire il farmaco A? Probabilmente no.
- Infatti, dai dati epidemiologici è difficile immaginare che una differenza così piccola nei valori medi di HbA_{1c} si possa tradurre in una differenza importante nel rischio di sviluppare le complicanze della malattia.

IL TEST DI IPOTESI

- Certamente, se i farmaci A e B avessero lo stesso profilo di tollerabilità e lo stesso costo, dovremmo sempre preferire quello che si è dimostrato più efficace, anche se di poco.
- Se al contrario il farmaco B fosse meglio tollerato o meno costoso, allora potrebbe in specifiche circostanze essere preferito, nonostante una efficacia lievemente (anche se statisticamente significativa) inferiore.

VALORI DI *P* E INTERVALLI DI CONFIDENZA

p<0.05

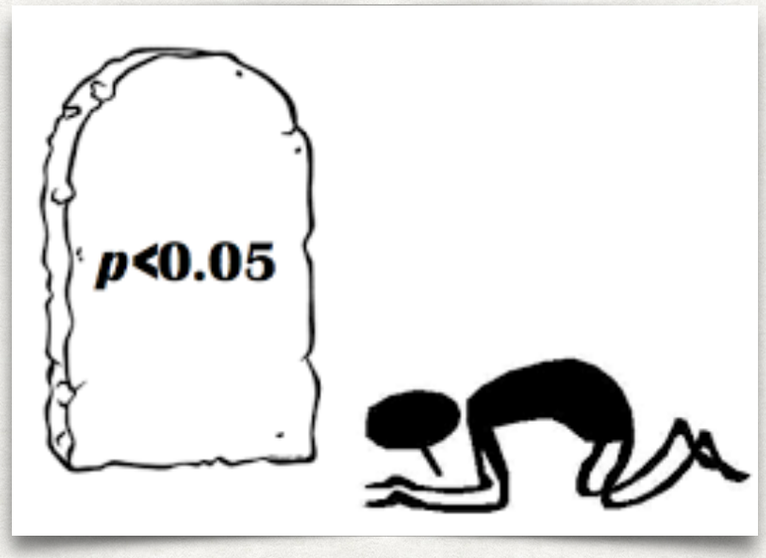
- Se rifiutiamo l'ipotesi nulla (cioè se dichiariamo che un trattamento è più efficace dell'altro) abbiamo una probabilità inferiore al 5% di affermare il falso.

Limiti

- Valore di soglia arbitrario (perché ad es. non 0.03 oppure 0.06?).
- La probabilità di trovare un p<0.05 aumenta con il numero di test eseguiti.
- Il valore di p non fornisce indicazioni sulla rilevanza clinica della differenza trovata.

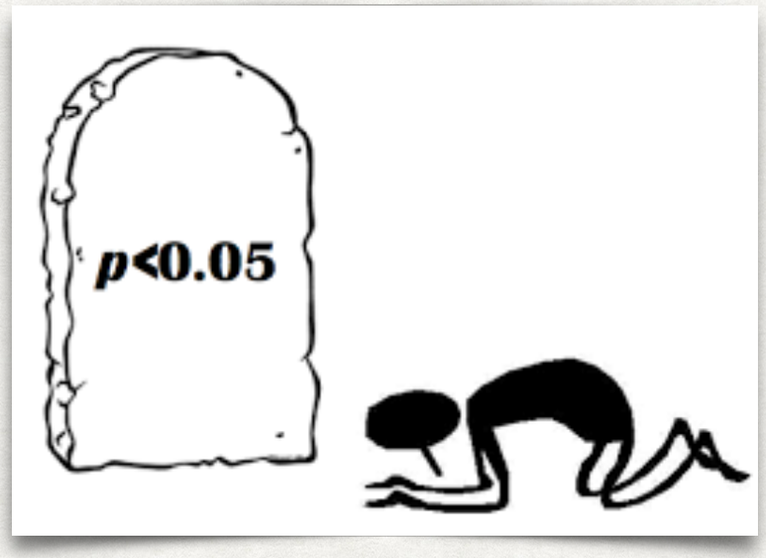
VALORI DI *P* E INTERVALLI DI CONFIDENZA

- Da quanto detto, appare chiaro che il valore di p ha dei limiti.
- Innanzitutto, il valore soglia per definire un risultato come statisticamente significativo è arbitrariamente posto a 0.05.
- Ma possiamo affermare con sicurezza che un p=0.06 o 0.07 non lo sia?
- In effetti, il rischio di accettare erroneamente l'ipotesi alternativa passerebbe dal 5% al 6-7%, ma resterebbe comunque molto basso!



VALORI DI *P* E INTERVALLI DI CONFIDENZA

- Inoltre, quando applichiamo un test statistico, il valore di p ottenuto si riferirà alla probabilità di errore per quello specifico test.
- Se nell'ambito di uno studio — come sempre accade — facciamo tante analisi statistiche sui nostri dati, la probabilità di trovare un valore di $p<0.05$ per puro caso aumenta in modo sostanziale all'aumentare del numero di test eseguiti.
- Infine, abbiamo già sottolineato come il valore di p non fornisca alcuna indicazione riguardo la rilevanza clinica dei risultati ottenuti.



SCELTA DEL TEST STATISTICO

SCELTA DEL TEST STATISTICO

Per decidere il tipo di test più appropriato per un determinato quesito bisogna domandarsi:

1. **Che tipo di variabili devo studiare?**
   - Numeriche (continue o discrete)
   - Categoriche (ordinali o nominali)
2. **Se la variabile è numerica, è distribuita normalmente?**
3. **Quanti sono i gruppi che devo confrontare?**
   - 2 gruppi;
   - 3 o più gruppi.

SCELTA DEL TEST STATISTICO

Metodi parametrici

- Basati su media μ e deviazione standard σ (o varianza σ^2).
- Si utilizzano nel caso di variabili distribuite normalmente.

Metodi non parametrici

- Basati su frequenza e percentili.
- Si utilizzano nel caso di variabili la cui distribuzione non sia normale.

I primi sono definiti "parametrici" in quanto si basano sui parametri *media* e *deviazione standard*. Tali test possono pertanto essere utilizzati solo per variabili normalmente distribuite. In tutti gli altri casi si utilizzano i metodi non parametrici, che prescindono dalla distribuzione dei dati.

I TEST STATISTICI PARAMETRICI

- I metodi statistici **parametrici** sono test statistici basati sui parametri *media* e *deviazione standard*.
- Per tale motivo essi sono utilizzabili solo per variabili numeriche normalmente distribuite.

Come si valuta la normalità di una distribuzione?

Approccio grafico

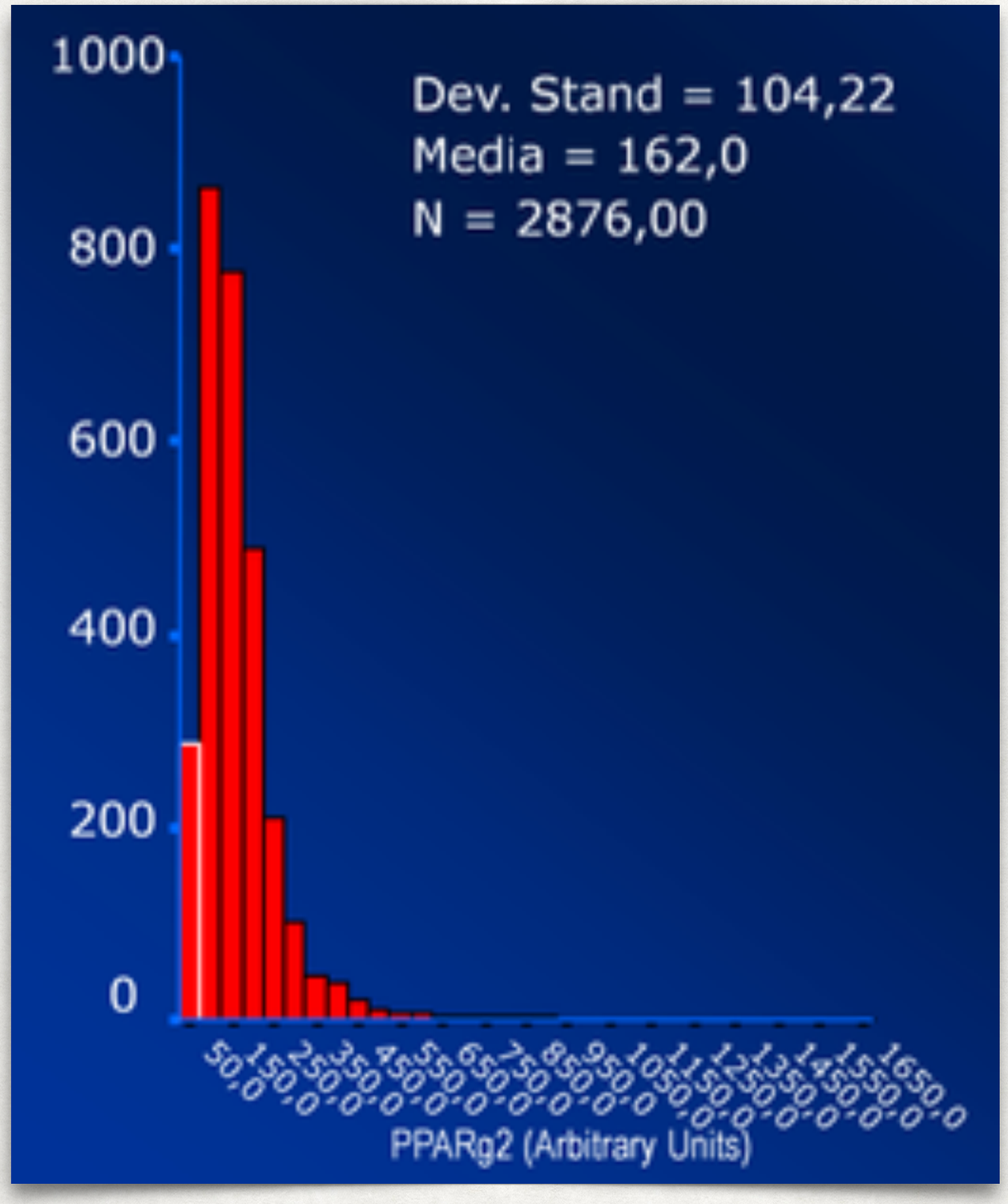
Skewness (asimmetria) e Kurtosis (curtosi)

Test statistici (Kolmogorov-Smirnov, χ^2, Shapiro-Wilk, …)

È possibile utilizzare diversi approcci, più o meno "esatti", basati su una valutazione grafica, che tuttavia potrebbe essere imperfetta, o su parametri numerici.

I TEST STATISTICI PARAMETRICI

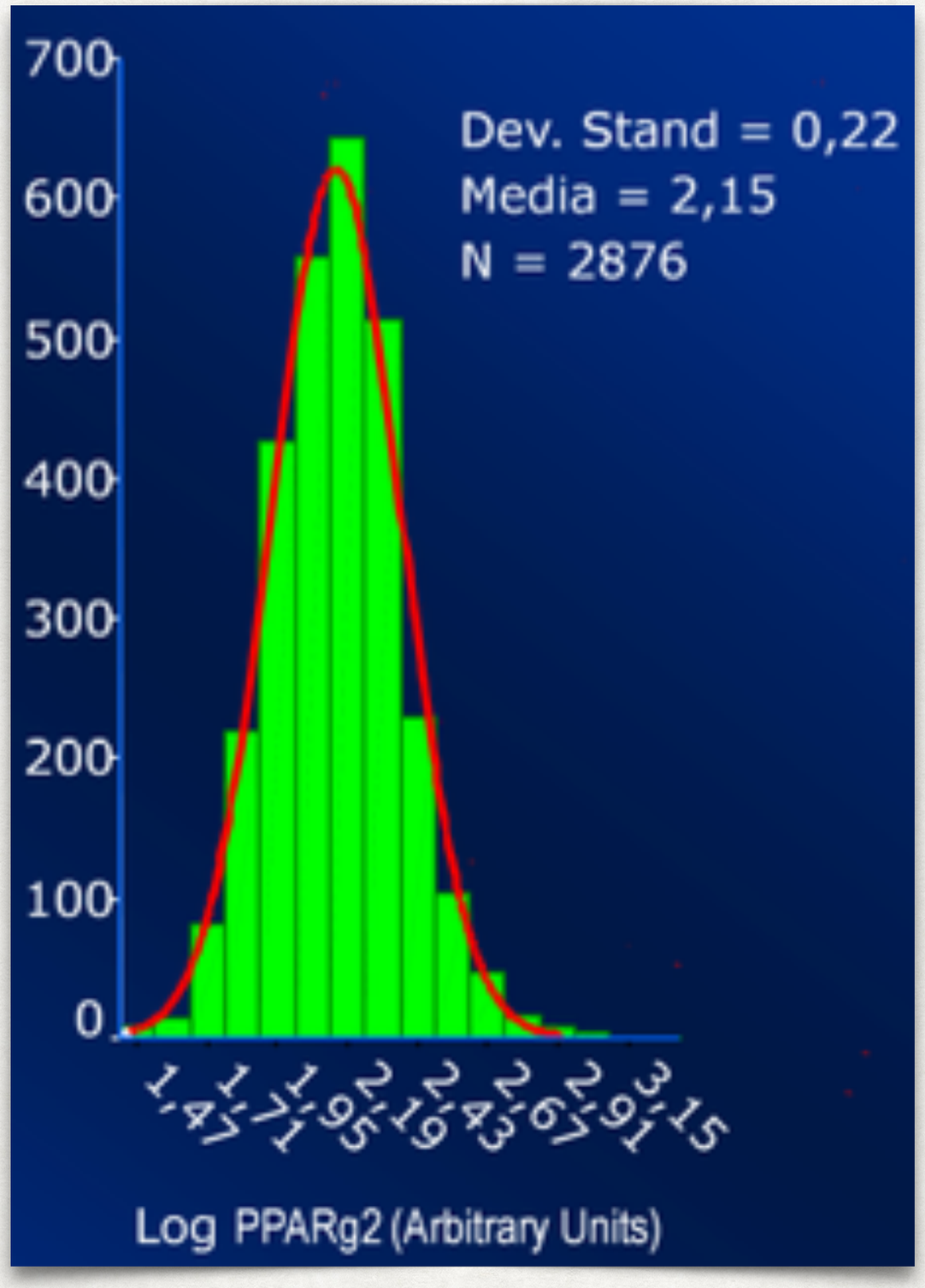
Approccio Grafico



- L'approccio grafico, come già detto, consiste nel disegnare l'istogramma della variabile in oggetto, per verificare se essa presenta o meno una distribuzione "a campana", simmetrica.
- L'esempio mostra come la distribuzione dei livelli di espressione genica di PPARg2 sia chiaramente asimmetrica (grafico a sinistra).

I TEST STATISTICI PARAMETRICI

Approccio Grafico

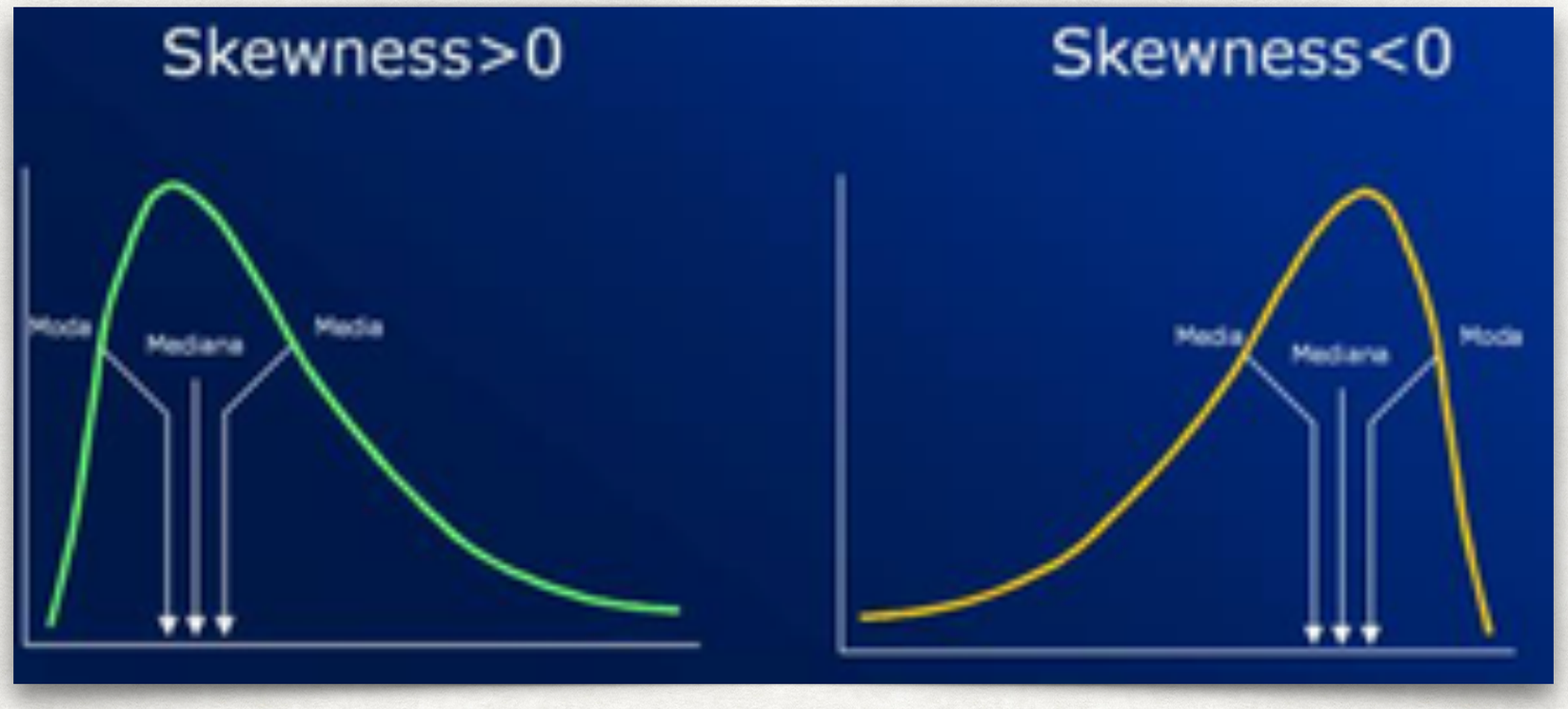


- Quando tuttavia disegniamo il grafico del logaritmo dei livelli di mRNA di PPARg2, la distribuzione della nuova variabile è sicuramente più vicina alla normalità.
- L'impressione grafica non è tuttavia sufficiente.
- Ricordiamo infatti che, oltre all'aspetto a campana, perché una variabile sia normalmente distribuita è necessario che il 95% delle osservazioni cadano entro l'intervallo definito da media ± (circa) 2 deviazioni standard (nell'esempio 2.15±0.44).

I TEST STATISTICI PARAMETRICI

Skewness e Kurtosis

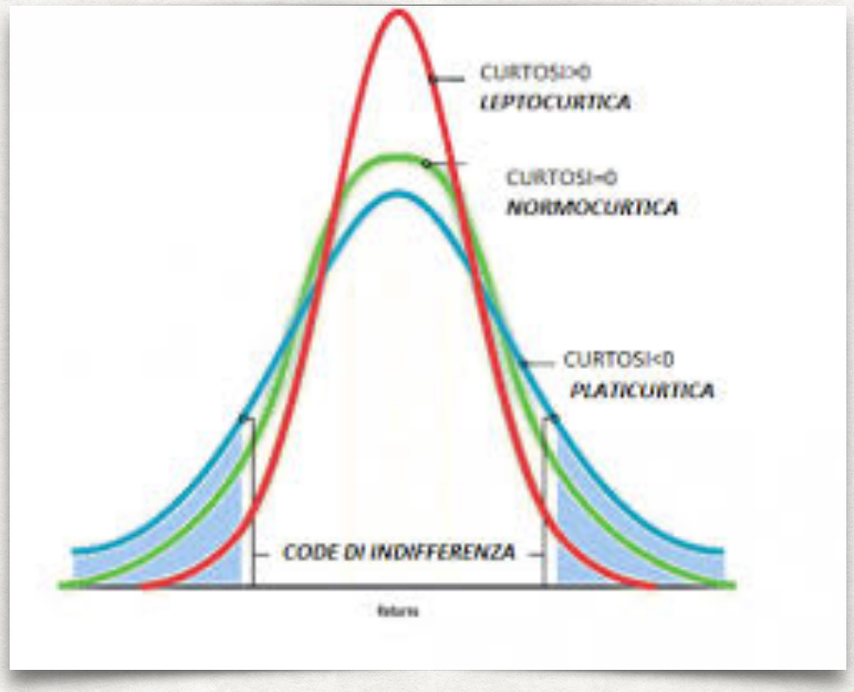
- Due indici numerici abitualmente riportati da tutti i software statistici riassumono in modo efficiente le informazioni che riguardano la distribuzione di una variabile continua: si tratta della *skewness* e della *kurtosis*.
- La **skewness** indica il livello di asimmetria della distribuzione. In caso di simmetria perfetta, il valore sarà pari a zero. Se il valore è positivo, allora la distribuzione sarà asimmetrica verso destra, mentre in caso di valore negativo essa sarà asimmetrica verso sinistra.



I TEST STATISTICI PARAMETRICI

Skewness e Kurtosis

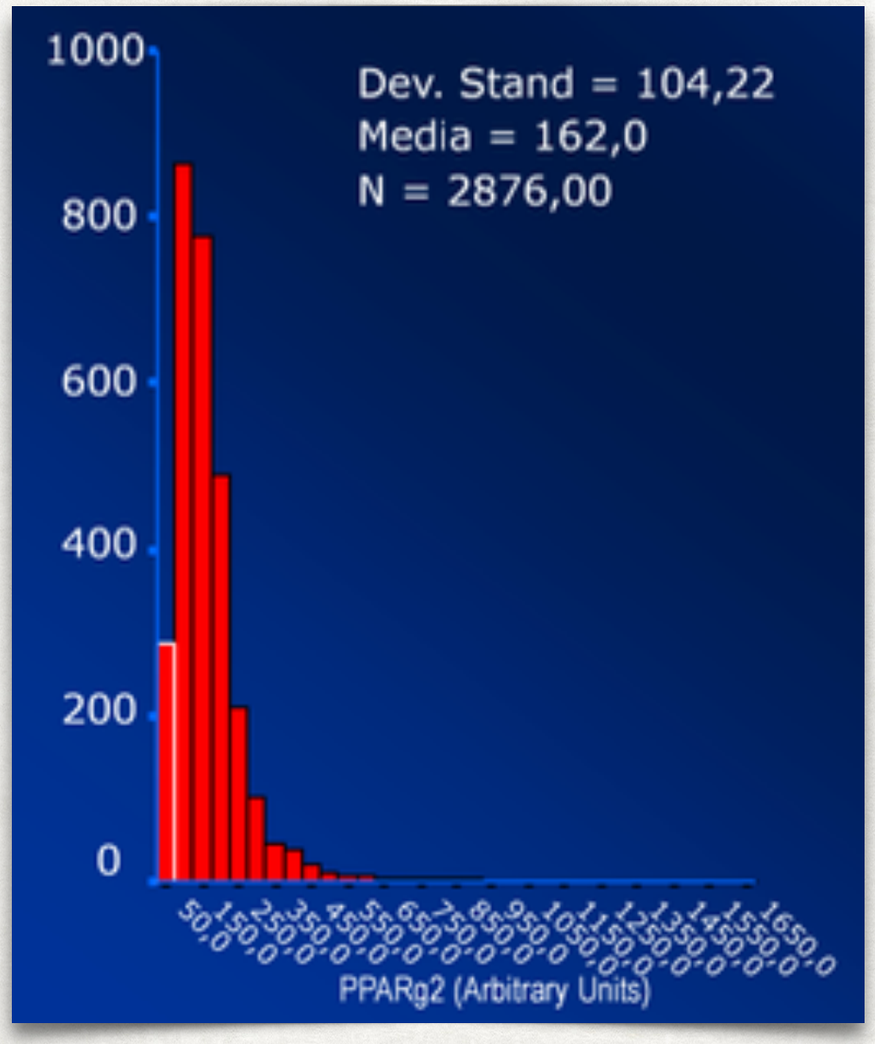
- La **kurtosis** indica invece se la distribuzione a campana è molto "stretta" o slargata.
- Se la distribuzione è normale, la kurtosis sarà uguale a zero (95% delle osservazioni contenute entro $\mu \pm 1.96\sigma$).
- Se la kurtosis è negativa, allora la distribuzione è platicurtica, cioè slargata, il che equivale a dire che meno del 95% delle osservazioni cadono entro $\mu \pm 1.96\sigma$
- Al contrario, se la kurtosis è positiva, allora la distribuzione è leptocurtica, cioè "stretta", cioè oltre il 95% delle osservazioni cade entro $\mu \pm 1.96\sigma$



I TEST STATISTICI PARAMETRICI

Skewness e Kurtosis

						Asimmetria			Curtosi		
	N	Minimo	Massimo	μ	σ	Valore	SE	Valore/SE	Valore	SE	Valore/SE
PPARg2	2876	28	1631	162.03	104.219	4.11	0.046	89.3	33.842	0.091	371.9



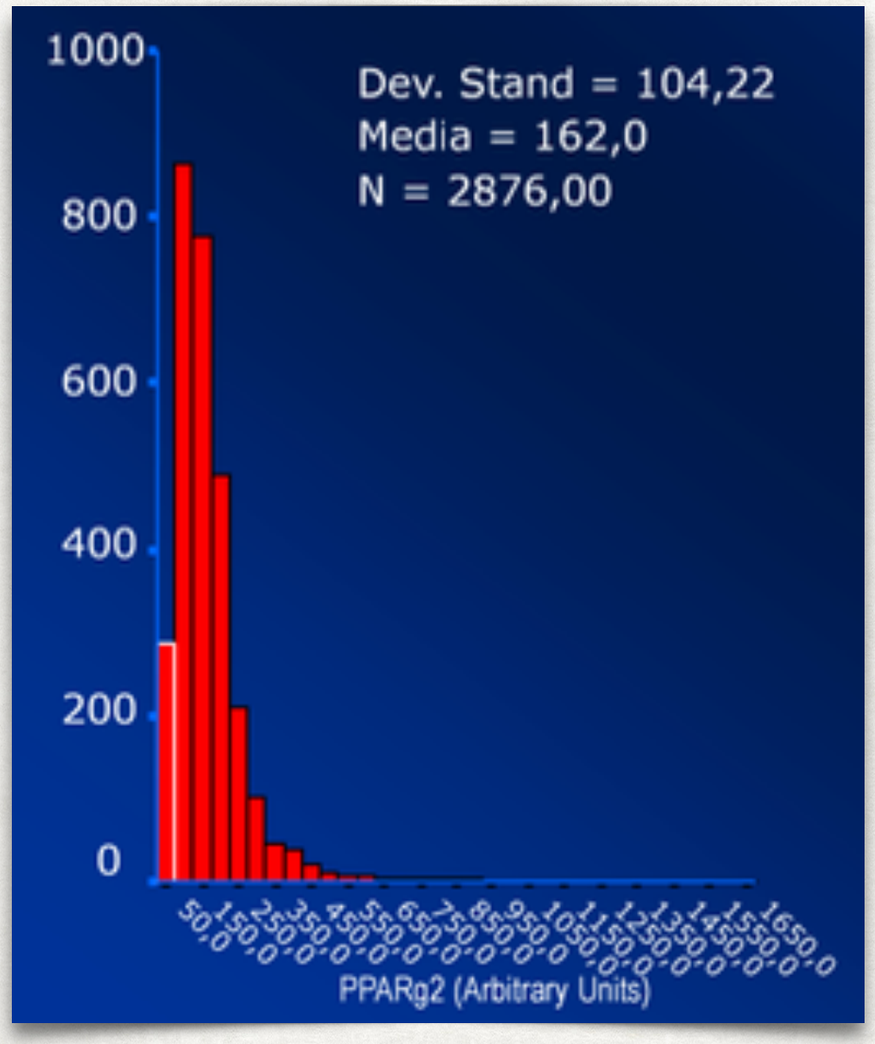
- Se la misura di asimmetria o di curtosi divisa per il proprio errore standard SE(*) dà un valore maggiore di 2 o minore di −2 allora la distribuzione non può essere considerata normale.
- Nell'esempio (livello di espressione genica di PPARg2) si può osservare come il valore di asimmetria sia positivo, pari a 4.11: questo equivale a dire che la distribuzione è asimmetrica verso destra, come chiaramente evidente dall'istogramma.
- Tale valore, diviso per il suo errore standard (pari a 0.046), dà un risultato di 89.3, ben maggiore di 2!

(*) Lo SE per la skewness è approssimativamente dato da $(6/N)^{1/2}$

I TEST STATISTICI PARAMETRICI

Skewness e Kurtosis

						Asimmetria			Curtosi		
	N	Minimo	Massimo	μ	σ	Valore	SE	Valore/SE	Valore	SE	Valore/SE
PPARg2	2876	28	1631	162.03	104.219	4.11	0.046	89.3	33.842	0.091	371.9

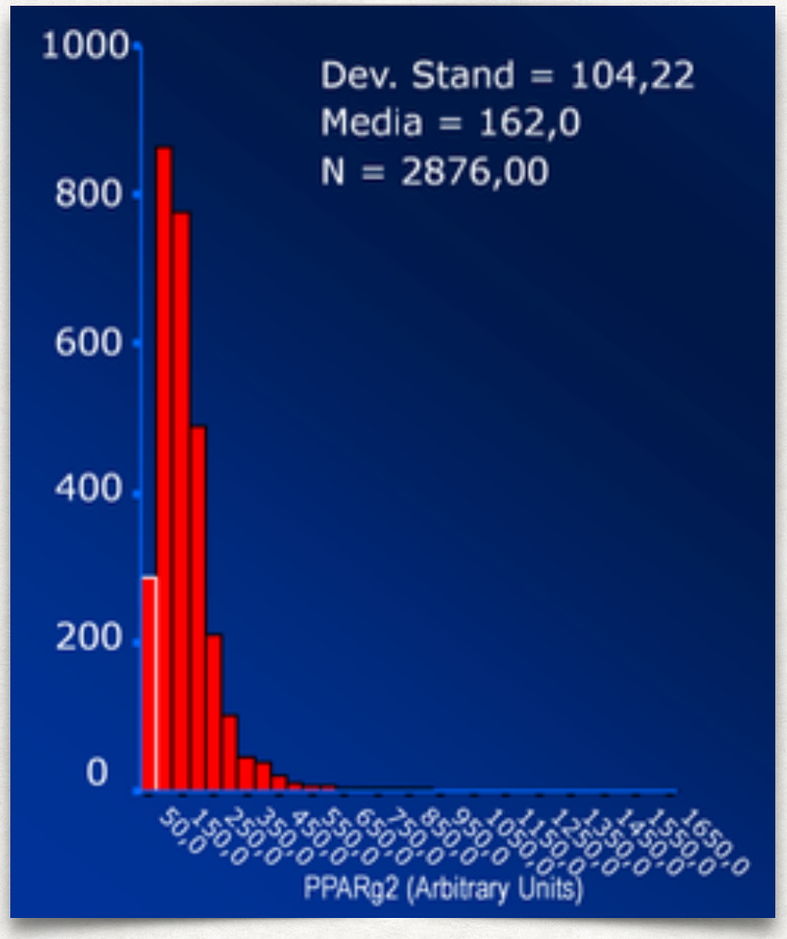


- Analogamente, la curtosi risulta positiva e molto elevata (33.842), ad indicare che la distribuzione è molto "stretta". Il valore ottenuto dividendo la curtosi per il suo SE (*) è pari a 371.9
- Pertanto, la distribuzione in questione si discosta notevolmente da una distribuzione normale.

(*) Lo SE per la kurtosis è approssimativamente dato da $(24/N)^{1/2}$

I TEST STATISTICI PARAMETRICI

TEST DI KOLMOGOROV-SMIRNOV



- Un'ultima modalità per valutare se una distribuzione differisce in maniera significativa dalla normalità è rappresentata dall'applicazione di un test statistico, chiamato di Kolmogorov-Smirnov.
- Tale test parte dall'ipotesi nulla che la distribuzione in esame non differisce da una distribuzione normale.

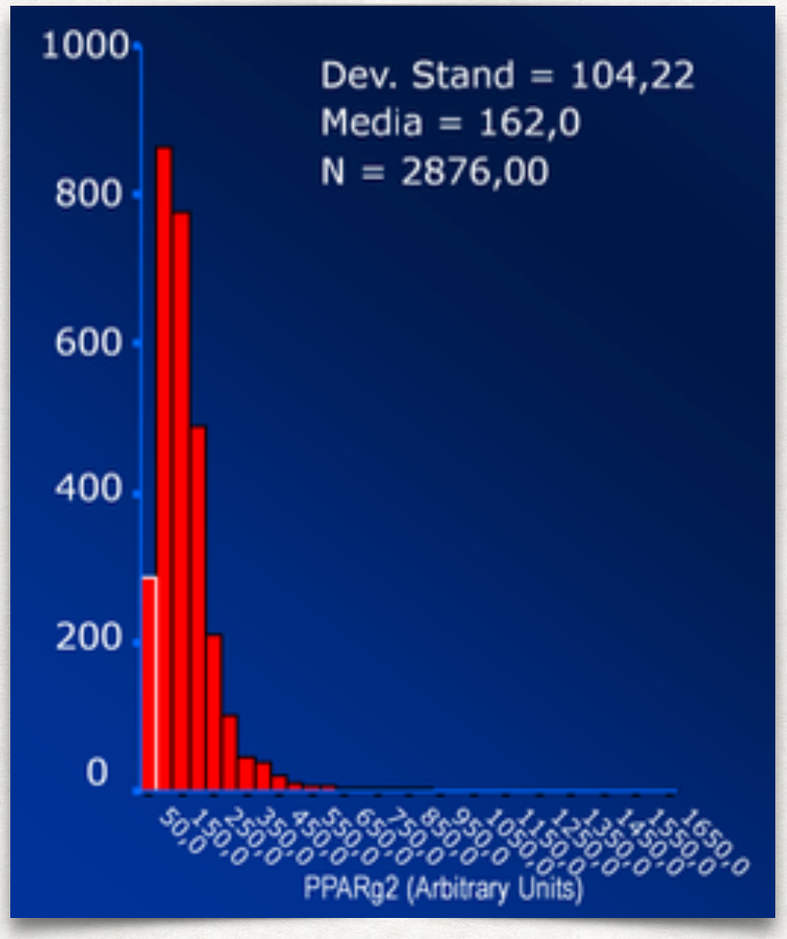
Test di K-S per un campione

PPARg2	
N	2876
μ	162.03
σ	104.22
Z di K-S	7.716
p-value (2 code)	<0.0001

Se $p < 0.05$ la distribuzione differisce significativamente da una distribuzione normale.

I TEST STATISTICI PARAMETRICI

TEST DI KOLMOGOROV-SMIRNOV



Test di K-S per un campione

PPARg2	
N	2876
μ	162.03
σ	104.22
Z di K-S	7.716
p-value (2 code)	<0.0001

Se $p < 0.05$ la distribuzione differisce significativamente da una distribuzione normale.

- Se il test porta ad un valore di $p<0.05$, allora dovremo rifiutare l'ipotesi nulla, ed affermare che la distribuzione in esame differisce significativamente da una distribuzione normale.
- Applicando il test di K-S ai nostri dati sui trigliceridi, otterremo un valore di $p<0.0001$.
- Pertanto, dovremo concludere che la distribuzione differisce in modo significativo da una distribuzione normale.

ALCUNI UTILI INDICATORI STATISTICI

Standard Error (SE), Relative SE (RSE), Coefficient of Variation (CV)

N	μ	σ	**SE**	RSE	CV (RSD)
2876	162.03	104.219	**±1.943**	1.2%	89.3

- Oltre ai normali parametri statistici Media e Deviazione Standard, si utilizza anche l'**errore standard della media** (SEM = Standard Error of the Mean, anche semplicemente SE) che misura la "dispersione" della distribuzione campionaria della media.
- La media campionaria è una stima della media della popolazione. Tuttavia, diversi campioni tratti dalla stessa popolazione hanno in generale valori diversi della media campionaria, per cui vi è una distribuzione di medie campionarie (con propria media e varianza).
- L'errore standard della media è la deviazione standard delle medie campionarie su tutti i campioni possibili (di una data dimensione) estratti dalla popolazione.
- Esso può essere calcolato come il rapporto tra la deviazione standard e la radice quadrata della dimensione del campione:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{N}}$$

Nell'esempio: SE = σ/√N = 1.943;

μ±1.96SE corrisponde all'intervallo di confidenza al 95% della media = [158.22, 165.84]

ALCUNI UTILI INDICATORI STATISTICI

Standard Error (SE), Relative SE (RSE), Coefficient of Variation (CV)

N	μ	σ	SE	**RSE**	CV (RSD)
2876	162.03	104.219	±1.943	**1.2%**	89.3

- L'**errore standard relativo** di una media campionaria (RSE) è l'errore standard diviso per la media ed espresso in percentuale.
- Esso può essere calcolato solo se la media è un valore diverso da zero.
- Una variabile con un errore standard relativo inferiore può dirsi avere una misura più precisa, dal momento che ha proporzionalmente meno variazione campionaria intorno alla media.

$$\mathrm{RSE} = \frac{\mathrm{SE}}{\mu}$$

Nell'esempio: RSE = SE/μ = 1.943/162.03 = 0.012 = 1.2%

ALCUNI UTILI INDICATORI STATISTICI

Standard Error (SE), Relative SE (RSE), Coefficient of Variation (CV)

N	μ	σ	SE	RSE	**CV (RSD)**
2876	162.03	104.219	±1.943	1.2%	**64.32%**

- Il **Coefficiente di Variazione** (CV), noto anche come deviazione standard relativa (RSD), è una misura standardizzata di dispersione di una distribuzione di probabilità o distribuzione di frequenza.
- Il CV o RSD è ampiamente usato in analisi di laboratorio per esprimere la precisione e la ripetibilità di un test.
- Esso mostra il grado di variabilità rispetto alla media della popolazione.
- È spesso espresso in percentuale, ed è definito come il rapporto tra la deviazione standard e la media (o il suo valore assoluto).

$$\mathrm{CV} = \mathrm{RSD} = \frac{\sigma}{\mu}$$

Nell'esempio: CV = σ/μ = 104.219/162.03 = 0.6432 = 64.32%

ALCUNI UTILI INDICATORI STATISTICI

Quartile Coefficient of Dispersion (QCD)

Dataset A = {2, 4, 6, 8, 10, 12, 14}						
N	range	μ	mediana	Q1	Q3	**QCD**
7	12	8	8	4	12	**50%**

Dataset B = {1.8, 2, 2.1, 2.4, 2.6, 2.9, 3}						
N	range	μ	mediana	Q1	Q3	**QCD**
7	1.2	2.4	2.4	2	2.9	**18%**

- Una possibilità più robusta del coefficiente di variazione è data dal **Coefficiente Quartile di Dispersione** (Quartile Coefficient of Dispersion, **QCD**).
- Il coefficiente quartile di dispersione è una statistica descrittiva che misura la dispersione ed è usata per confronti interni e tra distribuzioni.
- Esso è dato dal range interquartile diviso per la somma: (Q3-Q1) / (Q3 + Q1).

Nell'esempio:

$$\text{QCD dataset A} = (Q3-Q1)/(Q3+Q1) = (12-4)/(12+4) = 8/16 = 0.5 = 50\%$$

$$\text{QCD dataset B} = (Q3-Q1)/(Q3+Q1) = (2.9-2)/(2.9+2) = 0.9/4.9 = 0.18 = 18\%$$

Il coefficiente quartile di dispersione del dataset *A* è 2.7 volte maggiore (0.5/0.18) di quello del dataset *B*

I TEST STATISTICI NON PARAMETRICI

- I metodi statistici non parametrici sono test statistici basati sui ranghi delle osservazioni, cioè sul loro numero d'ordine invece che sulle osservazioni in sé.
- Quindi, trasformando dei valori in ranghi si perde l'informazione relativa alla misura, mantenendo solo quella relativa all'ordinamento.
- Tutte le volte in cui una variabile numerica non è normalmente distribuita, o ci troviamo di fronte ad una variabile categorica (ordinale o nominale), sarà necessario utilizzare i **test non parametrici**.
- Tali test si basano sui ranghi delle osservazioni, non sul loro reale valore. In altre parole, le osservazioni vengono messe in ordine crescente, e ad ognuna si attribuisce un numero corrispondente alla posizione che quell'osservazione occupa nella graduatoria (rango, *rank*).
- I test statistici non parametrici vengono quindi basati sul confronto fra le somme dei ranghi.

I TEST STATISTICI NON PARAMETRICI

Esempio di Calcolo dei Ranghi

Siano dati i seguenti valori di una variabile: 11, 25, 3, 26, 20

- Per calcolare i ranghi vanno prima ordinati i valori della variabile: 3, 11, 20, 25, 26
- Quindi si assegnano i ranghi, che assumeranno i seguenti valori: 1, 2, 3, 4, 5

Siano dati i seguenti valori di una variabile: 11, 25, 3, 25, 20

- Per calcolare i ranghi vanno prima ordinati i valori della variabile: 3, 11, 20, 25, 25
- Quindi si assegnano i ranghi, che assumeranno i seguenti valori: 1, 2, 3, **4.5, 4.5**

Per i dati identici viene considerata la media dei ranghi altrimenti attribuibile alle misure (nel caso fossero state diverse).

Pertanto nell'esempio il valore **4.5** deriva dalla media dei valori di ranghi che sarebbero stati assegnati ai due valori (4 e 5).

I TEST STATISTICI NON PARAMETRICI

Applicabilità

I test statistici non parametrici sono giustificati quando:

- le variabili hanno **evidenti scostamenti dalla normalità**, oppure sono **fortemente asimmetriche** o presentano **più di un picco** (distribuzione multimodale);
- il campione è **troppo piccolo** per verificare se la distribuzione dei dati è normale;
- le osservazioni sono rappresentate da **variabili categoriche**.

I test non parametrici vengono quindi utilizzati quando una variabile non è normalmente distribuita, ma anche quando i valori a disposizione sono pochi, e non è quindi possibile capire quale sia la distribuzione.

Se ad esempio misuriamo il peso in 5 soggetti, sarà ben difficile capire se queste misure si distribuiscono in modo normale!

Come regola generale, se il numero di osservazioni è <20, non bisogna mai usare un test parametrico.

SCELTA DEL TEST STATISTICO

I metodi non parametrici sono tanto più "potenti" (nel senso statistico) quanto più la distribuzione si allontana da quella normale, mentre sono meno potenti dei corrispondenti metodi parametrici quando ci si avvicina alla normalità

- Mentre non vi è alcun dubbio sulla necessità di dover utilizzare un test non parametrico per le variabili nominali e ordinali, ci si potrebbe chiedere in quale errore si incorrerebbe se, in presenza di una variabile continua, si utilizzasse sempre un test parametrico, a prescindere dalla distribuzione della variabile o, al contrario, sempre un test non parametrico.
- A questo proposito, è importante sottolineare che i test non parametrici sono tanto più potenti quanto più la distribuzione si discosta dalla normalità, mentre, in presenza di una distribuzione normale, essi tendono ad essere meno potenti dei test parametrici.
- Come conseguenza, se utilizziamo un test parametrico quando la distribuzione è chiaramente non-normale, o di converso utilizziamo un test non parametrico in presenza di una distribuzione normale, potremmo ottenere dei risultati non statisticamente significativi, quando invece lo sarebbero se si fosse utilizzato il test più appropriato.

SCELTA DEL TEST STATISTICO

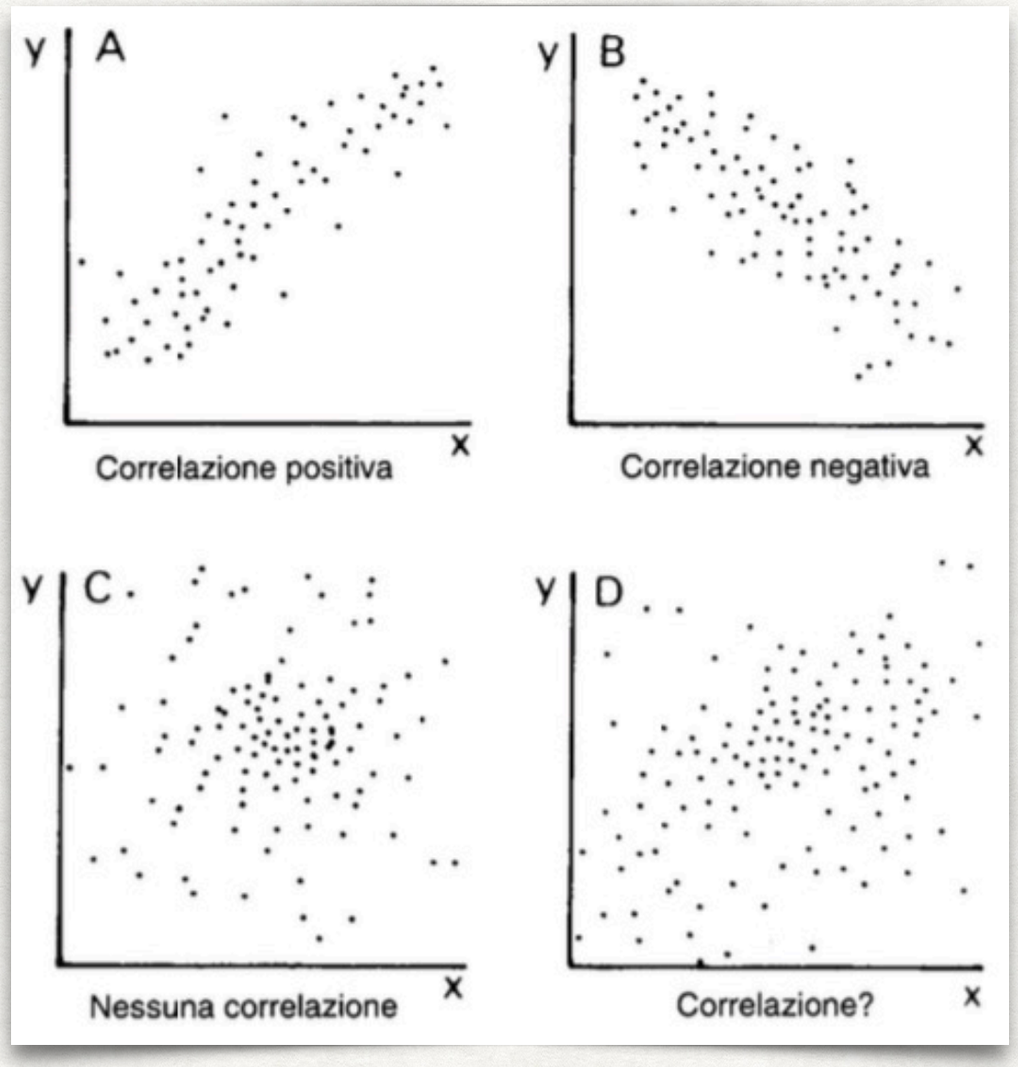
	Test parametrici	Test non parametrici
Misure riassuntive	Media (μ) e deviazione standard (σ)	Mediana e range
Confronto fra 2 variabili	Regressione lineare	Chi-quadrato (χ^2)
Misura di correlazione fra 2 variabili	R di Pearson	ρ di Spearman τ-b di Kendall
Confronto dei valori fra 2 gruppi differenti	Test t di Student "unpaired"	Test di Mann-Whitney (u-test)
Confronto prima-dopo (before-after)	Test t di Student "paired"	Test sign-rank di Wilcoxon
Confronto di valori fra più gruppi	Analisi della varianza (ANOVA)	Test di Kruskal-Wallis
Analisi multivariata	Regressione multipla	Regressione logistica

Questa tabella riassume i test statistici di più comune impiego.
Risulta evidente come, per ogni situazione, esista un test parametrico ed il corrispettivo test non parametrico.
Non c'è quindi alcuna ragione per utilizzare a tutti i costi un test parametrico, anche quando non appropriato.

CORRELAZIONE E REGRESSIONE

CORRELAZIONE E REGRESSIONE

- La **correlazione** analizza se esiste una relazione tra due variabili (come e quanto due variabili variano insieme).
- La **regressione** analizza la forma della relazione funzionale tra variabili (relazione causa-effetto).
- **Regressione semplice** (lineare o non lineare): determinare la forma della relazione tra 2 variabili (una indipendente ed una dipendente).



- **Regressione multipla**: determinare la forma della relazione tra più variabili (più indipendenti ed una dipendente).
- Conoscendo la forma della relazione funzionale tra variabile indipendente e dipendente è possibile stimare il valore della variabile dipendente conoscendo quello della variabile indipendente (solo interpolazione).

CORRELAZIONE FRA VARIABILI NORMALI

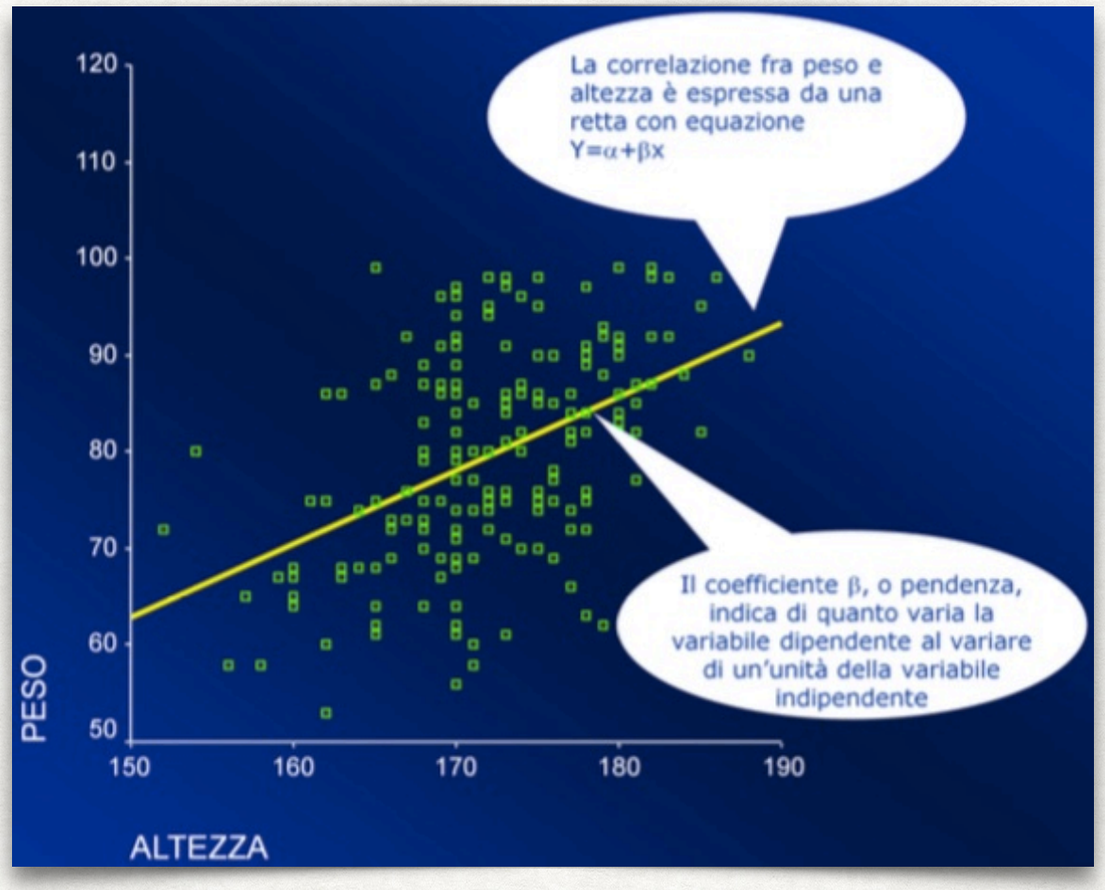
Che relazione intercorre fra peso e altezza?
La pressione arteriosa sistolica è correlata all'età?

A questi quesiti (ricerca di una correlazione fra due variabili normalmente distribuite) è possibile dare una risposta con la **regressione lineare**.

Con tale tecnica è possibile valutare se e con che misura i valori di una variabile (**variabile dipendente**) variano (sono cioè correlati) col variare di una seconda variabile (variabile indipendente).

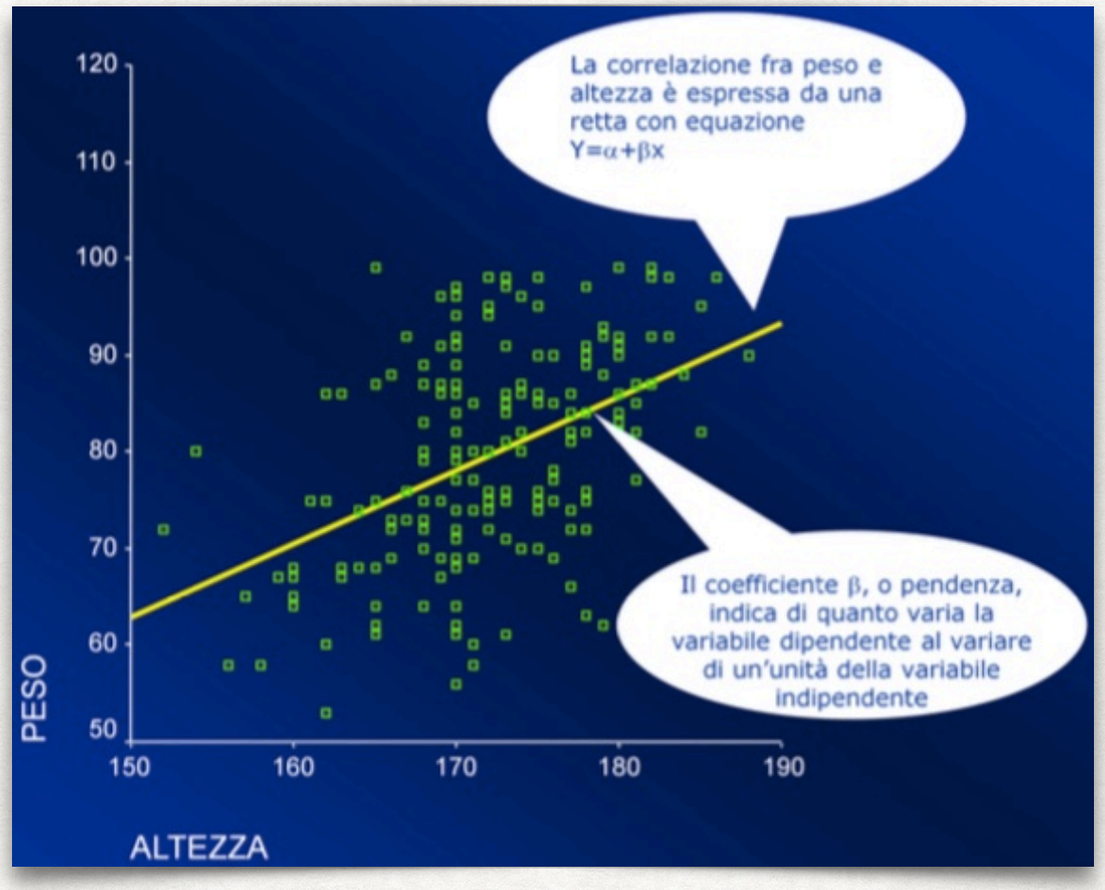
Ad esempio, nel rapporto fra peso e altezza il peso sarà la variabile dipendente (cioè quella che varia al variare dell'altezza), mentre nel rapporto fra pressione sistolica ed età, sarà la prima a variare in funzione della seconda, per cui la nostra variabile dipendente sarà la pressione sistolica, quella indipendente l'età.

CORRELAZIONE FRA VARIABILI NORMALI



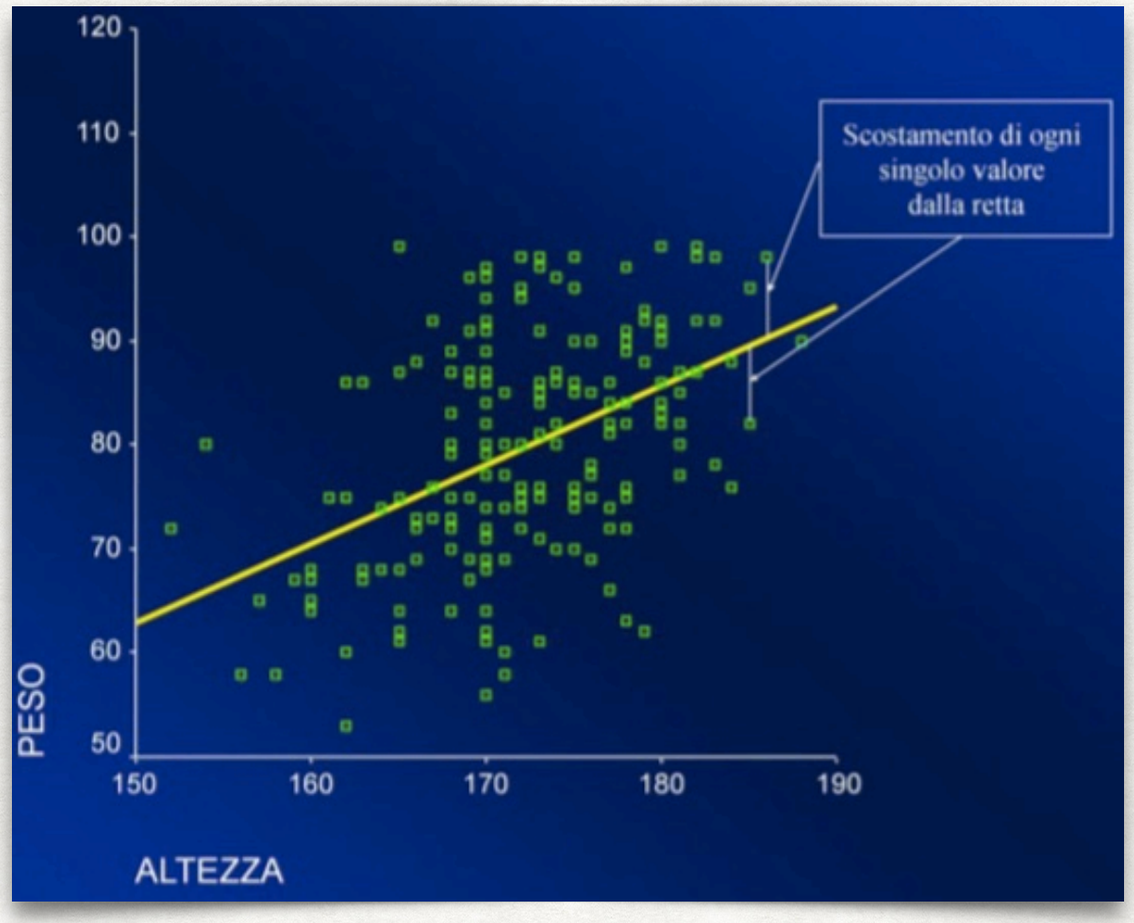
- Nella regressione lineare, il rapporto esistente fra variabile dipendente e variabile indipendente può essere rappresentato in un grafico con assi cartesiani.
- In esso sulle ascisse sono riportati i valori della variabile indipendente, e sulle ordinate i valori della variabile dipendente.
- Quanto più il rapporto fra le due variabili è stretto, tanto più i punti del grafico tenderanno a disporsi lungo una linea retta.

CORRELAZIONE FRA VARIABILI NORMALI



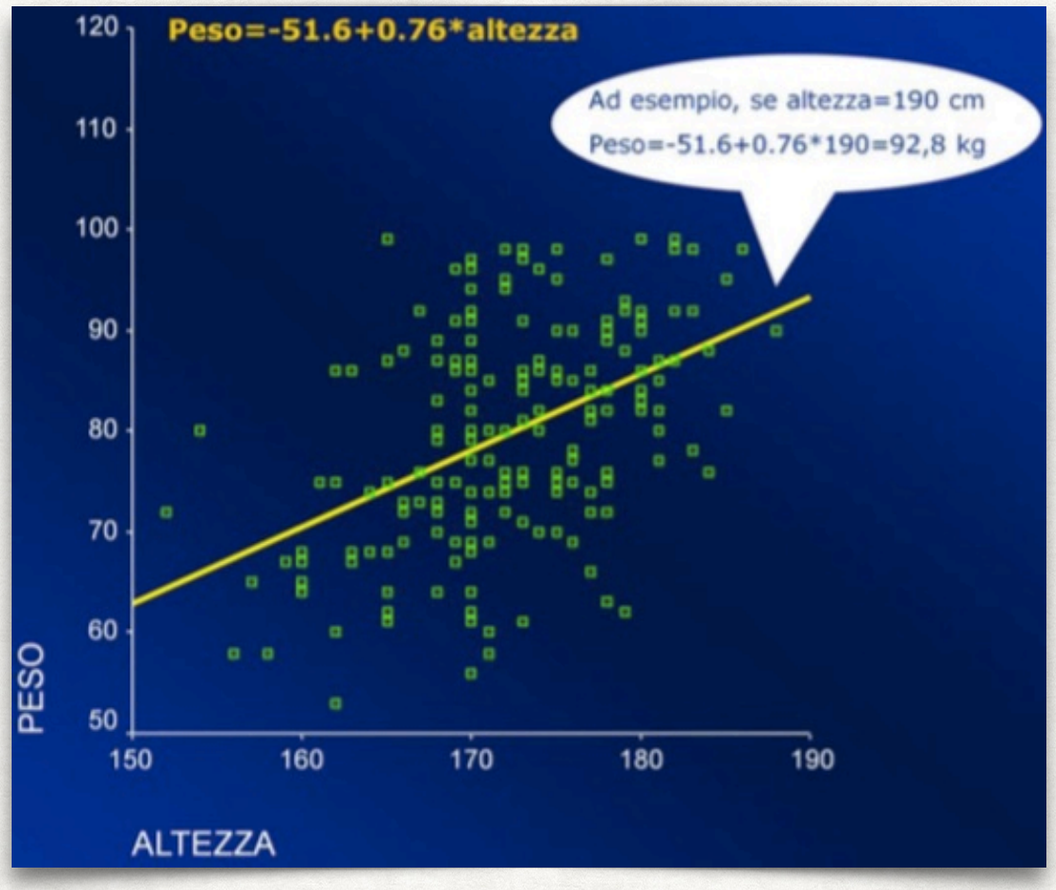
- Da un punto di vista matematico, il rapporto fra le due variabili è espresso proprio dall'equazione di una retta, dove y rappresenta il valore della variabile dipendente e x il valore della variabile indipendente.
- Il coefficiente β, anche chiamato pendenza, indica di quanto varia y per ogni variazione di una unità di x.
- Il valore di α, o intercetta, rappresenta invece il valore che y assume quando x è uguale a zero.

CORRELAZIONE FRA VARIABILI NORMALI



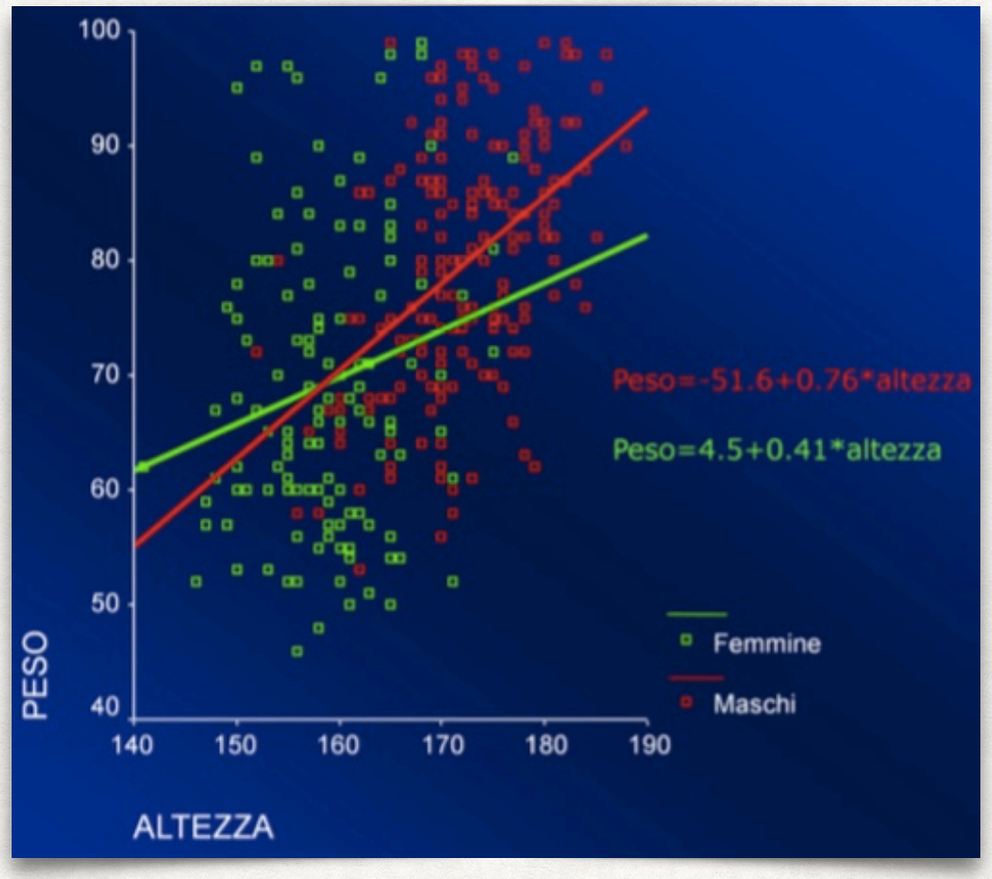
- Si noti che fra tutte le possibili rette che descrivono il rapporto fra le due variabili, verrà scelta quella per la quale sia minima la somma del valore al quadrato di tutti gli scostamenti di ogni singolo punto dalla retta (metodo dei minimi quadrati).
- *N.B. I valori degli scostamenti sono elevati al quadrato in quanto, rispetto al valore stimato dall'equazione, i valori sarebbero a volte positivi e a volte negativi, per cui la loro somma sarebbe uguale a zero.*

CORRELAZIONE FRA VARIABILI NORMALI



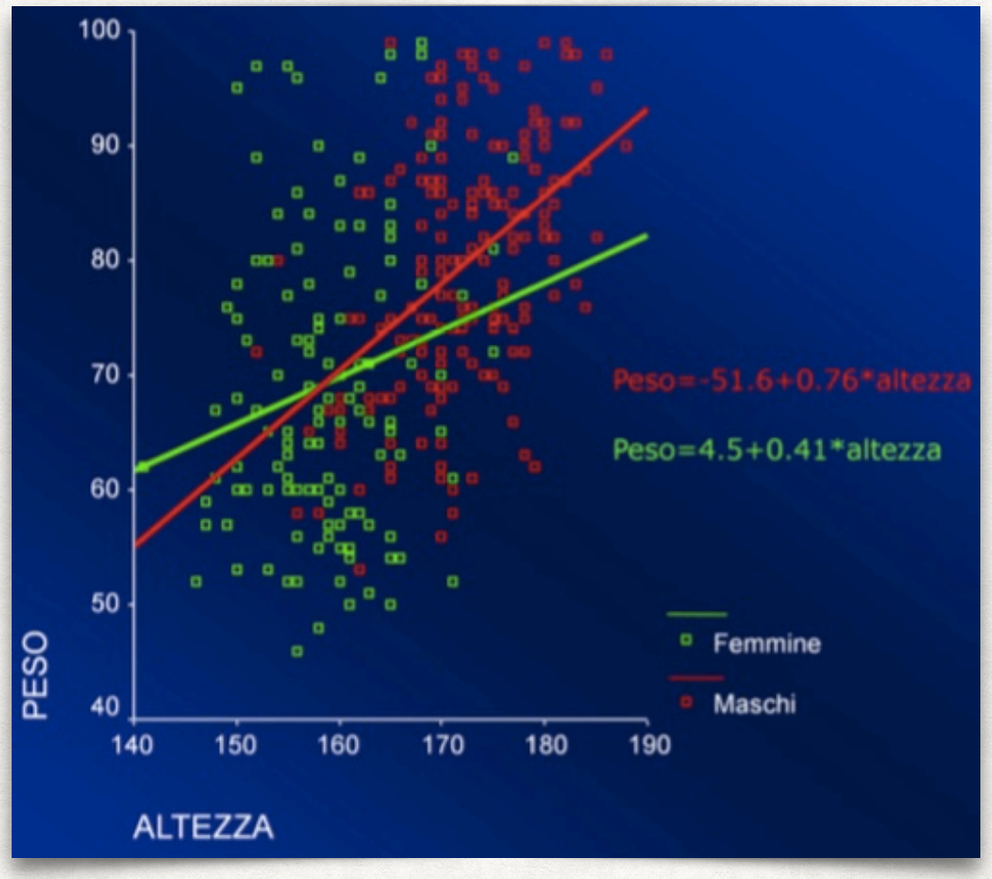
- Nel caso del rapporto fra peso e altezza riportato in figura, l'equazione ci dice che, in media, il peso tende a crescere di 0.76 kg per ogni cm in più di altezza.
- Si noti che il valore dell'intercetta, negativo, è puramente teorico, in quanto rappresenterebbe il peso di un soggetto alto 0 cm!
- Dall'equazione della retta possiamo stimare che, ad esempio, un soggetto alto 1.90 cm ed appartenente alla popolazione in studio (soggetti di sesso maschile affetti da diabete), avrebbe in media un peso di 92.8 kg.

CORRELAZIONE FRA VARIABILI NORMALI



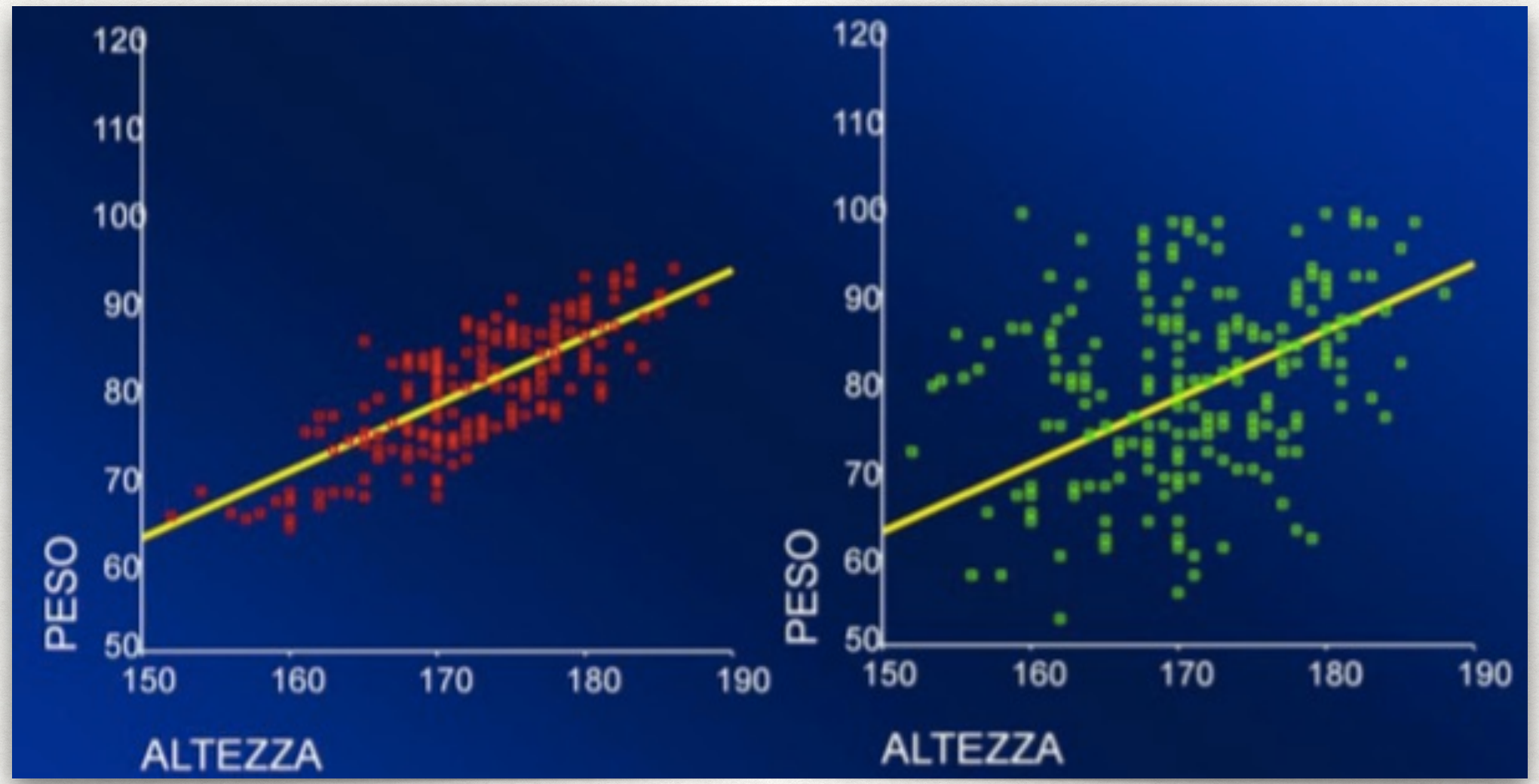
- In questo grafico il rapporto fra peso e altezza è stato stimato separatamente per maschi e femmine.
- Come si può constatare, la retta che descrive il rapporto fra peso e altezza nei maschi (rossa) è più "ripida" di quella delle femmine (verde).
- La pendenza (o coefficiente beta) per i maschi è infatti più elevata (0.76 per i maschi, 0.41 per le femmine).
- In altre parole il peso tenderà ad aumentare in media di 0.76 kg per ogni cm in più di altezza fra i maschi, mentre tenderà ad aumentare in media di 0.41 kg per ogni cm fra le donne.

CORRELAZIONE FRA VARIABILI NORMALI



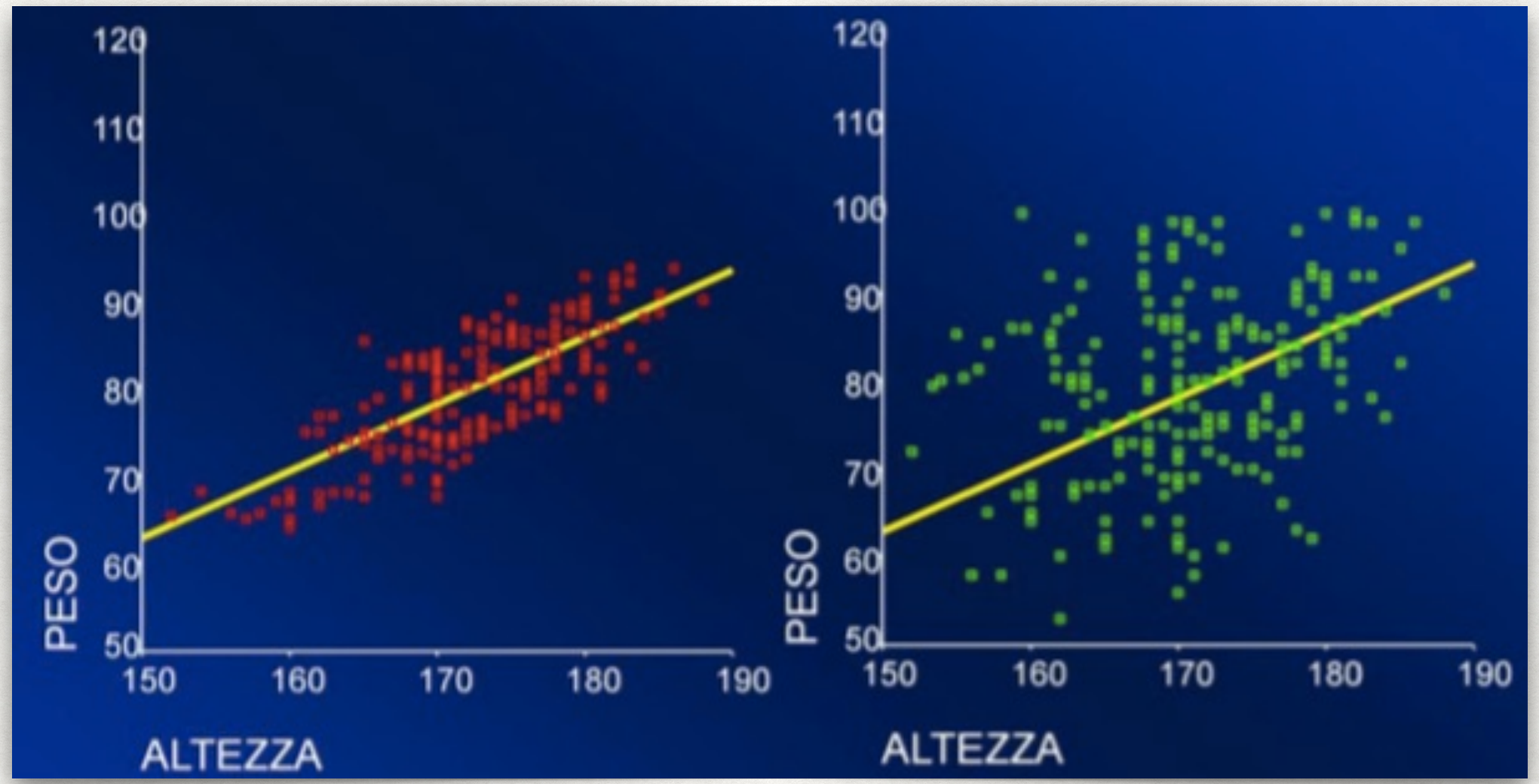
- Quando il coefficiente β è positivo, allora le due variabili sono positivamente correlate (la variabile dipendente tende a crescere al crescere della variabile indipendente, e la retta tende a "salire" da sinistra a destra).
- Quando invece il beta è negativo, allora fra le due variabili ci sarà un rapporto inverso (la variabile dipendente tende a decrescere al crescere della variabile indipendente, e la retta tende a "scendere" da sinistra a destra).
- In assenza di correlazione fra le due variabili, il coefficiente β avrà un valore vicino a zero; graficamente, la retta sarà grosso modo parallela all'asse delle ascisse.

CORRELAZIONE FRA VARIABILI NORMALI



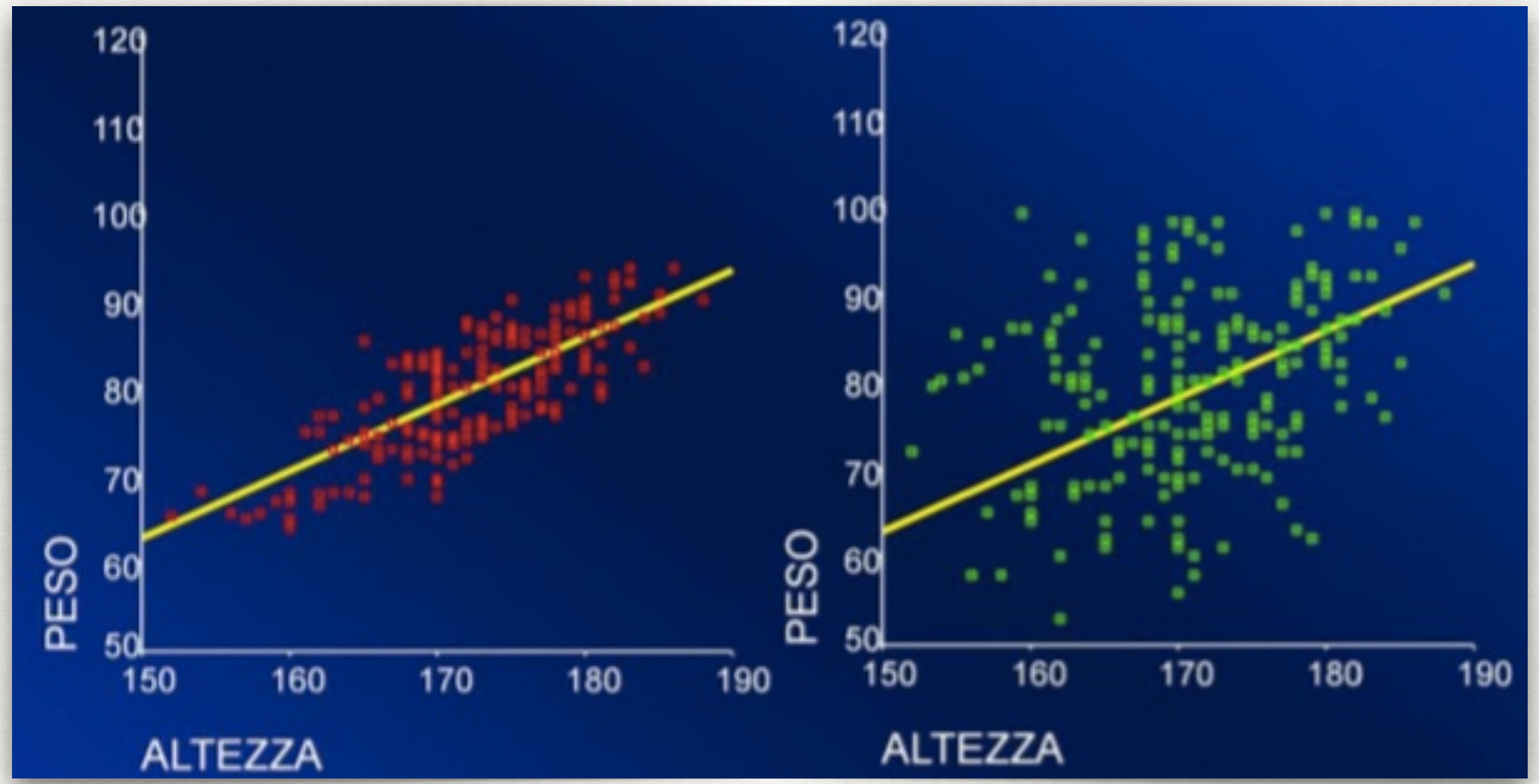
- Sebbene la regressione lineare fornisca una indicazione di quanto tenda a variare la variabile dipendente in funzione della variabile indipendente, essa non ci fornisce alcuna indicazione su quanto sia forte tale associazione.
- Dal punto di vista grafico, una prima indicazione deriverà dalla dispersione dei punti attorno alla retta.

CORRELAZIONE FRA VARIABILI NORMALI



- Se i punti tendono ad essere molto vicini alla retta, assumendo una classica disposizione lineare (come nel grafico a sinistra) allora possiamo arguire che il rapporto lineare fra le due variabili sia molto stretto.
- Se al contrario i punti sono molto dispersi, a formare una nuvola (come nel grafico a destra) allora l'associazione sarà meno forte.

CORRELAZIONE FRA VARIABILI NORMALI



È tuttavia possibile andare oltre l'impressione grafica, ed esprimere la forza dell'associazione in termini quantitativi, utilizzando il coefficiente di correlazione $\boldsymbol{R}$ di Pearson.

CORRELAZIONE FRA VARIABILI NORMALI

Coefficiente di Correlazione R di Pearson

- L'esistenza di una relazione lineare fra due variabili continue X ed Y normalmente distribuite può anche essere espressa dal coefficiente di correlazione R di Pearson.
- Il coefficiente R varia fra -1 e +1, ed è è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili:

$$R = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- La covarianza misura quanto le due variabili X ed Y varino assieme.

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

CORRELAZIONE FRA VARIABILI NORMALI

COEFFICIENTE DI CORRELAZIONE R DI PEARSON

- Se $0 < R \leq 0.3$ si ha correlazione *debole*; se $0.3 < R \leq 0.7$ si ha correlazione *moderata*; se $R > 0.7$ si ha correlazione *forte*.
- Il segno indica la direzione della correlazione:
  - se R è negativo, all'aumentare della variabile indipendente la variabile dipendente diminuisce;
  - se R è positivo, all'aumentare della variabile indipendente la variabile dipendente aumenta;
  - se $R = 0$ non c'è correlazione fra le due variabili.

R^2 indica la percentuale di variabilità della variabile dipendente "spiegata" dalla variabile indipendente (anche noto come *coefficiente di determinazione*).

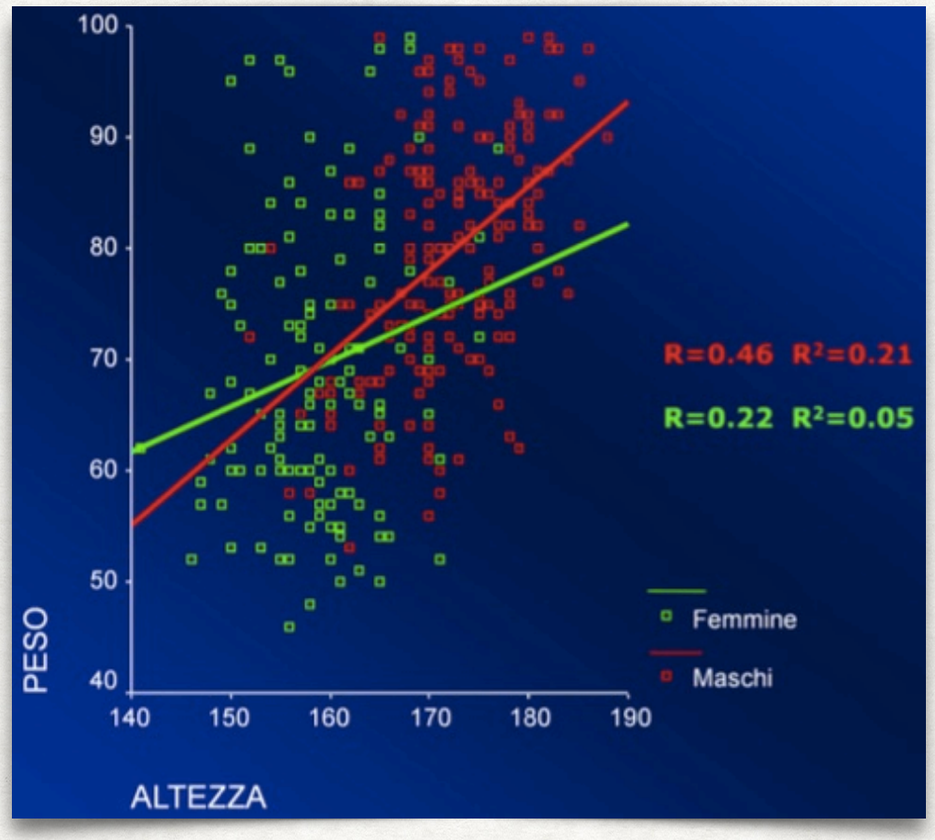
CORRELAZIONE FRA VARIABILI NORMALI

COEFFICIENTE DI CORRELAZIONE R DI PEARSON

- Il coefficiente di correlazione R di Pearson esprime quindi la forza dell'associazione lineare fra due variabili continue normalmente distribuite.
- Se R = 1 oppure –1, allora la correlazione è perfetta (graficamente, tutti i punti cadrebbero esattamente sulla retta).
- Pertanto, quanto più R si avvicina a 1 o a –1, tanto maggiore sarà la forza dell'associazione fra le due variabili.
- Il valore di R elevato al quadrato fornisce un'ulteriore informazione: dice infatti quale sia la percentuale della variabilità della variabile dipendente spiegata dalla variabile indipendente.
- Se R = 1 oppure –1, R^2 sarebbe uguale a 1; in altre parole, il 100% della variabilità della nostra variabile dipendente sarebbe spiegato dalla variabile indipendente.

CORRELAZIONE FRA VARIABILI NORMALI

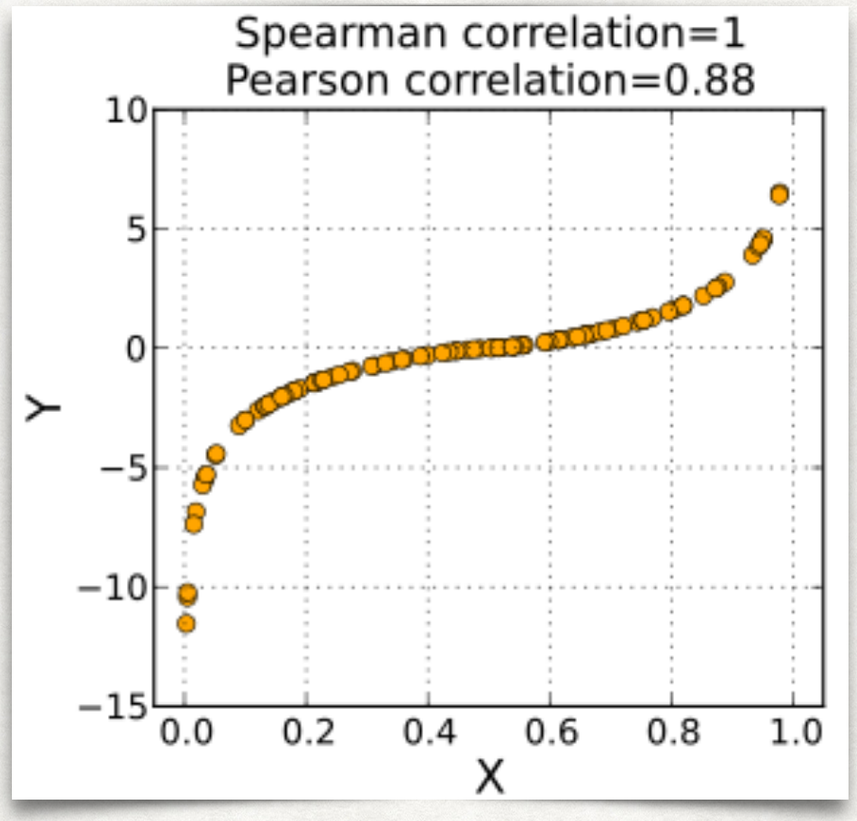
COEFFICIENTE DI CORRELAZIONE R DI PEARSON



- Nel caso del rapporto peso-altezza nei maschi e nelle femmine, si può vedere come il valore di R sia molto più alto nei maschi.
- Questo vuol dire che, mentre nei maschi la variabilità nel peso è legata in modo importante alla variabilità dell'altezza, nelle femmine tale correlazione è meno forte. In altre parole, fra le donne più che fra gli uomini, potremo trovare a parità di altezza persone con peso molto diverso.
- Il valore di R^2 ci indica che fra i maschi la variabilità nell'altezza è responsabile del 21% della variabilità nel peso, mentre fra le donne l'altezza spiega solo il 5% della variabilità del peso.

CORRELAZIONE FRA VARIABILI NON NORMALI

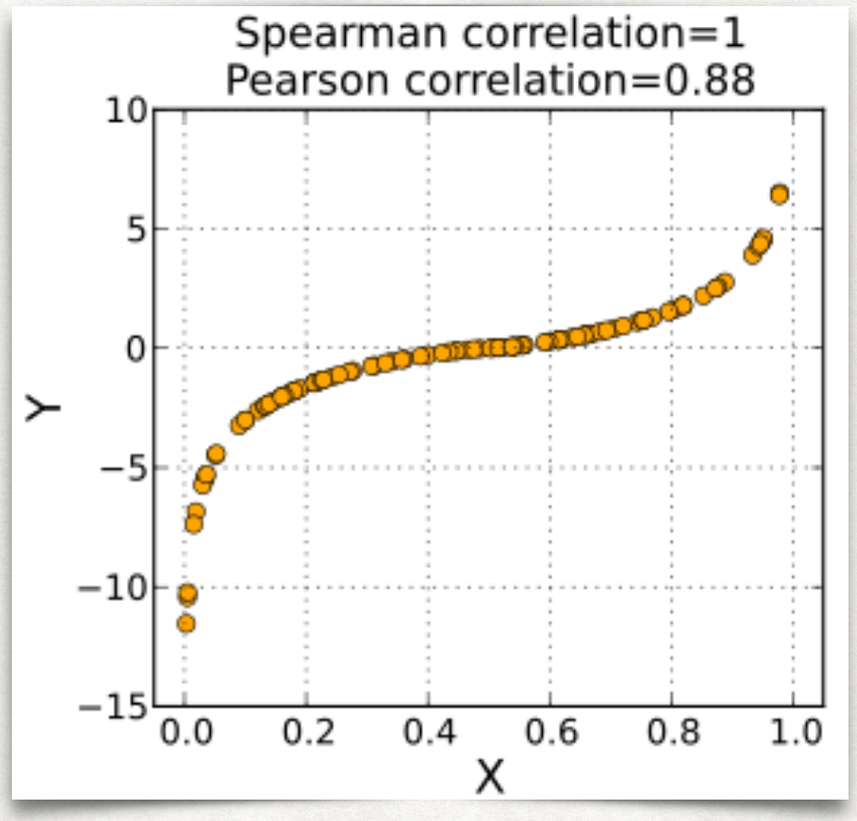
COEFFICIENTE DI CORRELAZIONE PER RANGHI DI SPEARMAN



- L'**indice di correlazione ρ per ranghi di Spearman** è una misura statistica non parametrica di correlazione.
- Essa misura il grado di relazione tra due variabili per le quali non si fa altra ipotesi della misura ordinale, ma possibilmente continua.
- Diversamente dal coefficiente di correlazione lineare di Pearson, il coefficiente di Spearman non misura una relazione lineare anche qualora vengano usate misure intervallari.
- A livello pratico il coefficiente ρ è semplicemente un caso particolare del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente.

CORRELAZIONE FRA VARIABILI NON NORMALI

COEFFICIENTE DI CORRELAZIONE PER RANGHI DI SPEARMAN



- L'**indice di correlazione ρ per ranghi di Spearman** è dato dalla formula:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} D_i^2}{N(N^2 - 1)}$$

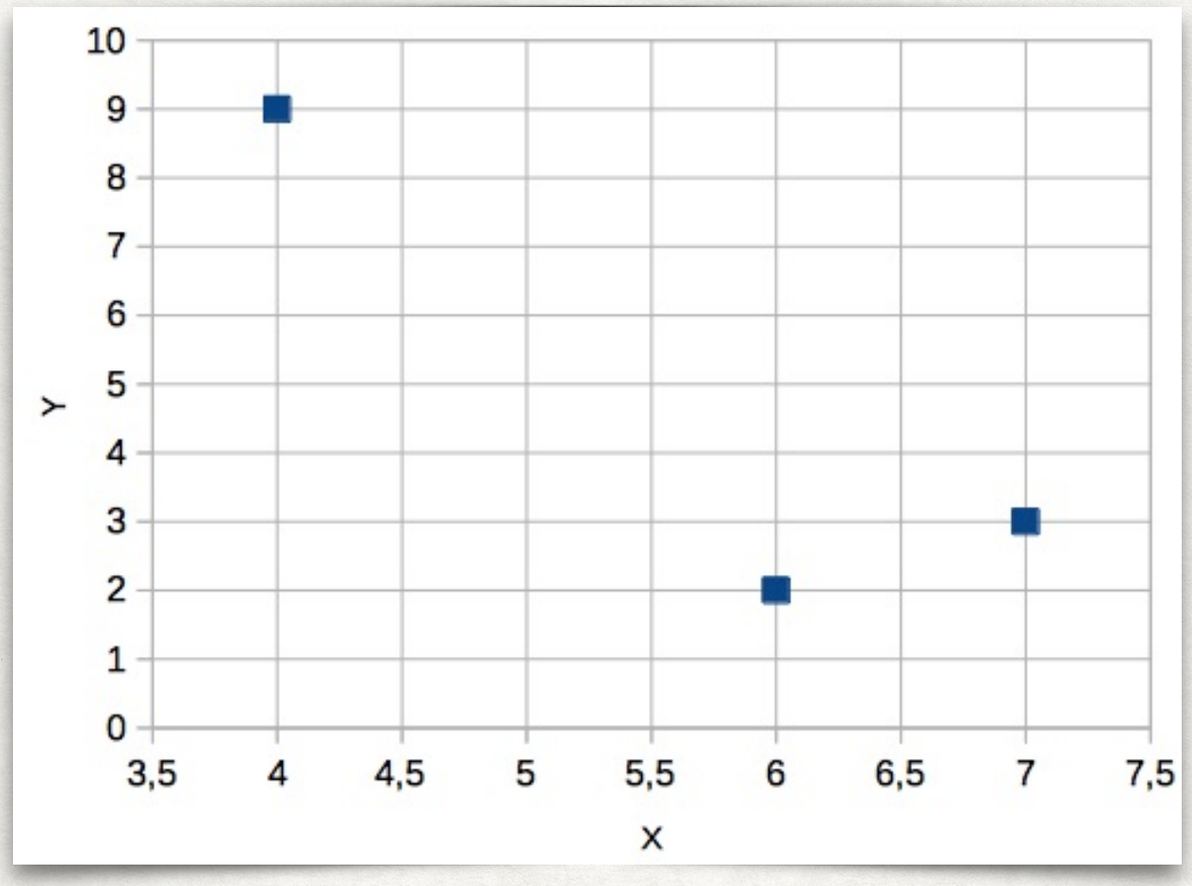
dove $D_i = r_i - s_i$ è la differenza dei ranghi (essendo r_i e s_i rispettivamente il rango della prima variabile e della seconda variabile della i-esima osservazione).

- Nell'esempio il coefficiente di Spearman è pari a 1. Contrariamente il coefficiente di Pearson non è ottimale.

CORRELAZIONE FRA VARIABILI NON NORMALI

COEFFICIENTE DI CORRELAZIONE PER RANGHI DI SPEARMAN

Data 1	Data 2	Rank 1	Rank 2	d	d^2
6	2	2	1	1	1
4	9	1	3	2	4
7	3	3	2	1	1



- Esempio di calcolo.

$$1-\left(\frac{6\Sigma d^2}{n(n^2-1)}\right)=1-\left(\frac{6\times 6}{n(n^2-1)}\right)$$

$$1-\left(\frac{6\times 6}{3(3^2-1)}\right)=-0.5$$

In un foglio elettronico si possono usare le funzioni RANGO e CORRELAZIONE.

CONFRONTO FRA GRUPPI

CONFRONTO FRA VALORI MEDI DI DUE GRUPPI (UNPAIRED)

Che differenza esiste tra i livelli medi pressori tra gruppo di controllo e pazienti trattati?

Per rispondere a questo quesito, si ricorre al **test *t* di Student**:

$$t = \frac{\text{Differenza tra le medie campionarie}}{\text{Errore standard della differenza tra le medie campionarie}}$$

- Un'altra situazione in cui spesso ci si imbatte analizzando i dati di uno studio consiste nel confrontare i valori medi di una variabile fra due gruppi di individui.
- Ad esempio, possiamo chiederci se il farmaco A abbia avuto un effetto maggiore del placebo sui valori di pressione arteriosa.
- Se la variabile in questione è normalmente distribuita, a questo tipo di quesiti si può rispondere utilizzando il test *t* di Student per dati non appaiati.

CONFRONTO FRA VALORI MEDI DI DUE GRUPPI (UNPAIRED)

Che differenza esiste tra i livelli medi pressori tra gruppo di controllo e pazienti trattati?

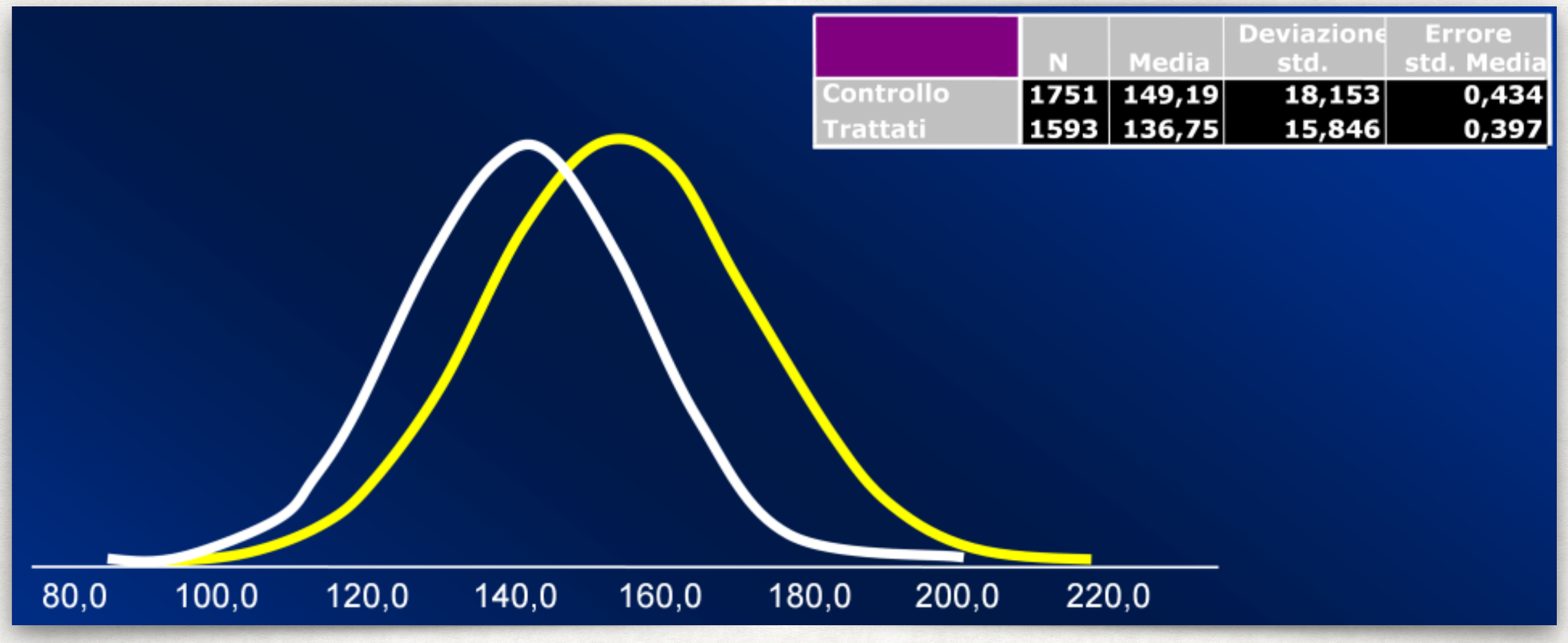
Per rispondere a questo quesito, si ricorre al **test *t* di Student**:

$$t = \frac{\text{Differenza tra le medie campionarie}}{\text{Errore standard della differenza tra le medie campionarie}}$$

- Tale test si basa sulla differenza fra i valori medi nei due gruppi messi a confronto, divisa per l'errore standard di tale differenza.
- L'errore standard fornisce una misura di quanto accuratamente sia stata stimata tale differenza fra i valori medi (più individui studieremo, più la stima sarà accurata).
- Più grande sarà il valore di t, più sicuri saremo nel rifiutare l'ipotesi nulla di non differenza fra i due valori medi.
- In particolare, il risultato sarà considerato statisticamente significativo ($p<0.05$) se il valore di t sarà > 1.96

CONFRONTO FRA VALORI MEDI DI DUE GRUPPI (UNPAIRED)

TEST T DI STUDENT

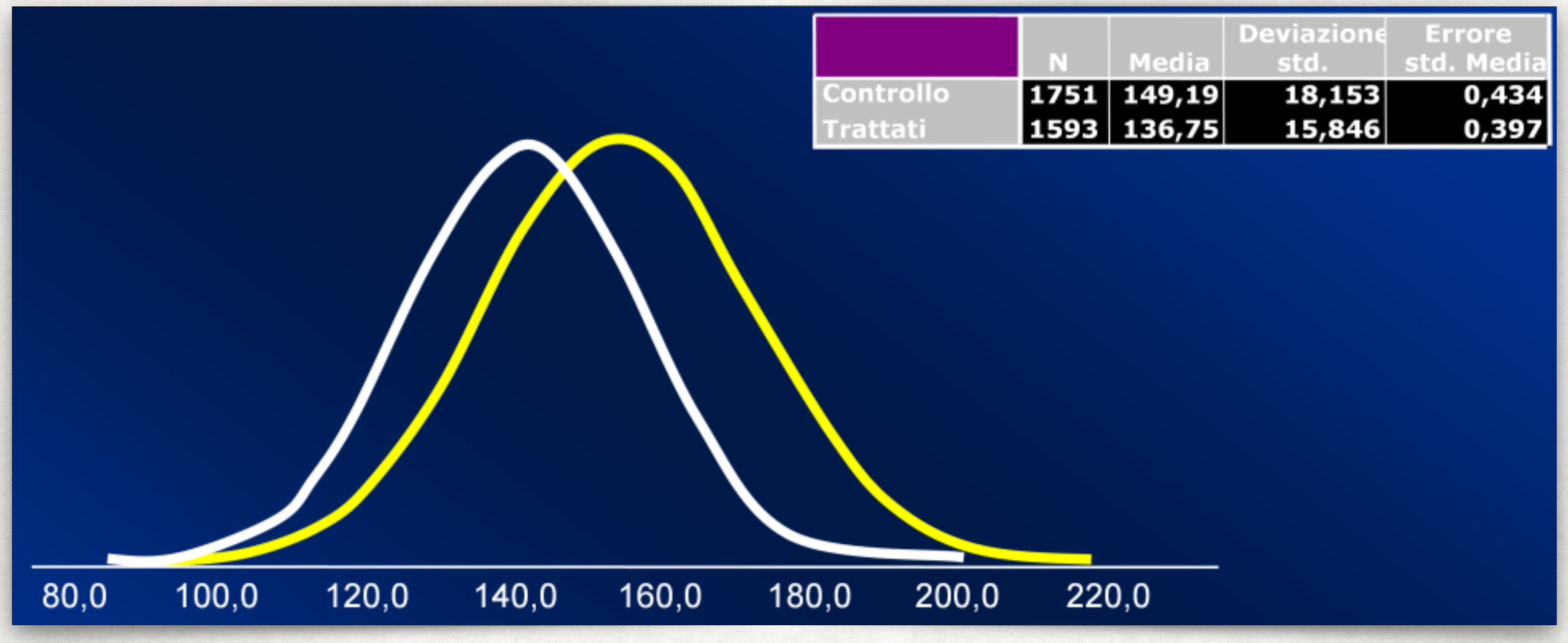


N.B. *L'errore standard della differenza fra le medie si calcola come la radice quadrata della somma degli errori standard delle due medie.*

- Differenza fra le medie: 149.19 − 136.75 = 12.44
- Errore standard della differenza = $\sqrt{(0{,}434^2+0{,}397^2)} = \sqrt{0.346} = 0.588$
- $t = 12.44 / 0.588 = 21.156 \Longrightarrow p < 0.0001$

CONFRONTO FRA VALORI MEDI DI DUE GRUPPI (UNPAIRED)

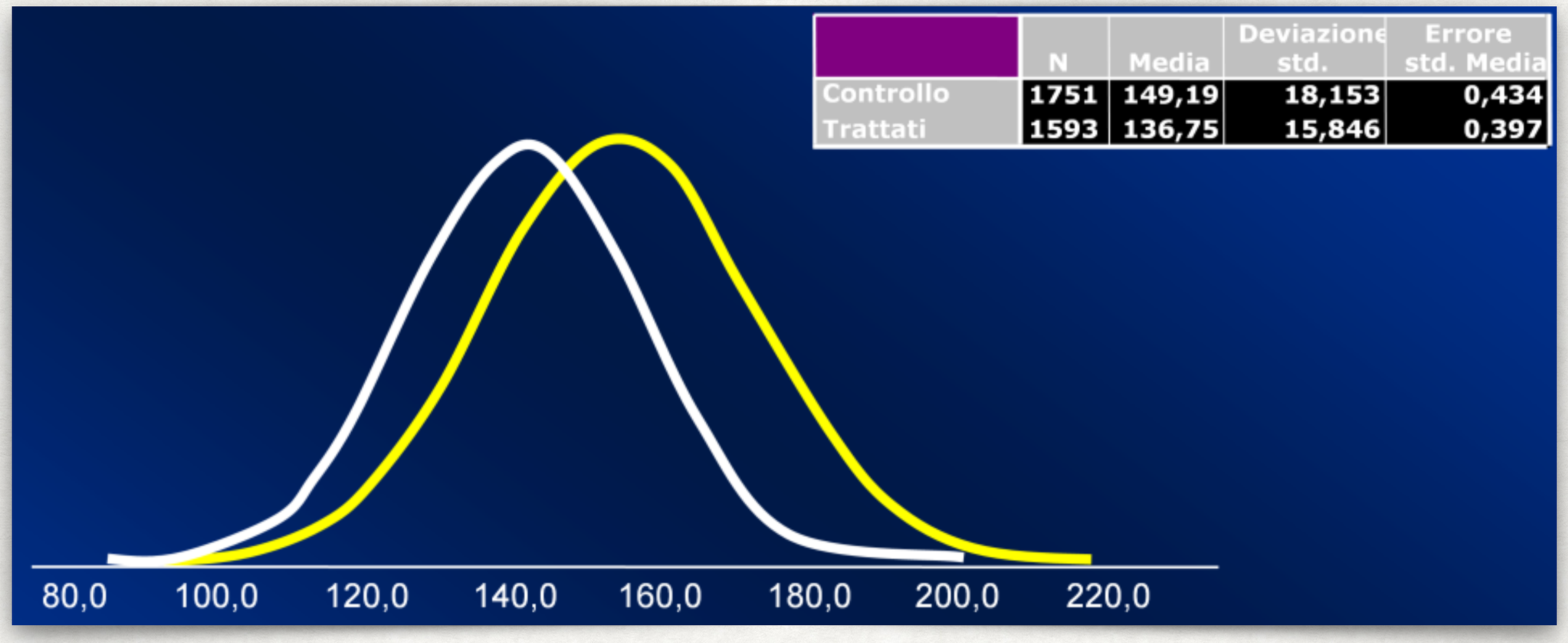
TEST T DI STUDENT



- Questo esempio rappresenta una applicazione pratica del test *t* di student per il confronto dei valori medi di pressione arteriosa sistolica fra un gruppo di 1593 soggetti trattati con un farmaco anti-ipertensivo e 1751 assegnati al gruppo di controllo, trattato con un placebo.
- I valori pressori medi sono di circa 137 mmHg nel primo gruppo e di 149 mmHg nel secondo. Il grafico riporta la distribuzione dei valori pressori nei due gruppi.
- L'ipotesi nulla di partenza è che le due distribuzioni non siano significativamente diverse.

CONFRONTO FRA VALORI MEDI DI DUE GRUPPI (UNPAIRED)

TEST T DI STUDENT



	N	Media	Deviazione std.	Errore std. Media
Controllo	1751	149,19	18,153	0,434
Trattati	1593	136,75	15,846	0,397

- Rifiutando l'ipotesi nulla, si afferma che i valori pressori ottenuti con il farmaco in studio sono invece significativamente più bassi di quelli ottenuti con il placebo.
- L'applicazione del test t di student (differenza fra le medie divisa per il suo errore standard) porterà ad un valore del t pari a oltre 21, cui corrisponderà un valore di $p<0.0001$.
- Possiamo quindi concludere che i valori pressori ottenuti con il farmaco sono significativamente più bassi di quelli ottenuti con il placebo.

CONFRONTO FRA VALORI MEDI DI UN CAMPIONE (PAIRED)

TEST T DI STUDENT (PER DISTRIBUZIONE NORMALE)

- Nel test *t* di Student appaiato (**paired *t*-test**), viene selezionato un campione e vengono prese due misurazioni per ciascun elemento del campione (ad esempio "prima" e "dopo" un trattamento).
- Ogni serie di misurazioni è considerata un campione. A differenza del test di ipotesi "unpaired", i due campioni non sono indipendenti l'uno dall'altro.
- I campioni appaiati sono anche chiamati "campioni abbinati" o "misure ripetute".
- La distribuzione dei dati deve essere normale.

CONFRONTO FRA VALORI MEDI DI UN CAMPIONE (PAIRED)

TEST T DI STUDENT (PER DISTRIBUZIONE NORMALE)

- *Ad esempio, se si vuole determinare se bere un bicchiere di vino o bere un bicchiere di birra ha lo stesso o diverso impatto sulla memoria, un approccio è quello di prendere un campione di — diciamo — 40 persone, facendo bere alla metà di loro un bicchiere di vino e all'altra metà un bicchiere di birra.*
- *Si sottopone poi ciascuna delle 40 persone a un test di memoria e si confrontano i risultati.*
- *Questo è l'approccio a "campioni indipendenti".*

CONFRONTO FRA VALORI MEDI DI UN CAMPIONE (PAIRED)

Test T di Student (per Distribuzione Normale)

- *Un altro approccio potrebbe essere quello di prendere un campione di 20 persone, somministrare ad ogni persona un bicchiere di vino e sottoporla quindi a un test di memoria.*
- *Si fa successivamente bere alle stesse persone un bicchiere di birra, eseguendo nuovamente il test della memoria. Infine si confrontano i risultati.*
- *Questo è l'approccio utilizzato con* ***campioni appaiati****.*

Il vantaggio di questo secondo approccio è il campione può essere più piccolo. Poiché i soggetti campionati sono gli stessi per la birra e il vino, ci sono meno possibilità che qualche fattore esterno (variabile di confondimento) influenzi il risultato.

Il problema con questo approccio è che è possibile che i risultati della seconda prova di memoria siano inferiori semplicemente perché la persona ha assorbito più alcol.

Questo può essere corretto separando le prove, ad esempio effettuando il test con birra un giorno dopo la prova con il vino.

CONFRONTO FRA VALORI MEDI DI UN CAMPIONE (PAIRED)
TEST T DI STUDENT (PER DISTRIBUZIONE NORMALE)

- Nel test *t* di Student appaiato (**paired *t*-test**) lo scopo è quindi quello di confrontare le medie di due campioni non indipendenti (essendo le misure prese per gli stessi soggetti).
- L'analisi è eseguita sulle N **differenze** $d_i = (y_i - x_i)$ tra le singole coppie di osservazioni (y_i=dopo; x_i=prima).
- L'ipotesi nulla è che la media delle differenze sia uguale a zero:
  $\boldsymbol{H_0: \mu_d = 0}$
- L'ipotesi alternativa è che la media delle differenze non sia uguale a zero:
  $\boldsymbol{H_A: \mu_d \neq 0}$
- La statistica che il test calcola è la seguente:

$$t = \frac{\mu_d}{\mathrm{SE}(\mu_d)}$$

dove $\mathrm{SE}(\mu_d) = \sigma_d / \sqrt{N}$

CONFRONTO FRA VALORI MEDI DI UN CAMPIONE (PAIRED)

Test T di Student (per Distribuzione Normale)

- Sotto l'ipotesi nulla, questa statistica segue una distribuzione t di Student con $n-1$ gradi di libertà.
- Si utilizza quindi l'appropriata tabella della distribuzione t di Student (con $n-1$ gradi di libertà) per confrontare il valore della statistica. Questo fornisce il *p*-value per il test.
- ATTENZIONE: *Il test è valido se la distribuzione delle differenze è approssimativamente normale. Perciò non è consigliabile utilizzare il paired t-test se ci sono outlier estremi o la distribuzione è fortemente non normale.*
- Sarebbe utile calcolare un intervallo di confidenza per la differenza media per sapere entro quali limiti la reale differenza della popolazione probabilmente si trova.
- Un intervallo di confidenza del 95% per la differenza media reale è dato da:

$$\mu_d \pm [t^* \times \mathrm{SE}(\mu_d)]$$

  dove t^* è il punto 2.5% della distribuzione t con $n-1$ gradi di libertà.
- N.B. *Nel foglio elettronico è disponibile la funzione TESTT per calcolare sia il test t paired che l'unpaired.*

CONFRONTO FRA VALORI MEDI DI UN CAMPIONE (PAIRED)

TEST T DI STUDENT (PER DISTRIBUZIONE NORMALE)

Esempio. Si supponga di avere le seguenti misure per un gruppo di 20 persone:

Soggetto	Prima	Dopo	Differenza
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

- Calcolando la media e la deviazione standard delle differenze abbiamo: $\mu_d = 2.05$ e $\sigma_d = 2.837$; di conseguenza $SE(\mu_d) = 2.837 / \sqrt{20} = 0.634$
- Quindi $t = 2.05 / 0.634 = 3.231$ con 19 df.
- Dalla tabella della distribuzione t di Student con 19 gradi di libertà abbiamo $p = 0.004$.
- Perciò vi è una forte evidenza che, in media, il trattamento porta ad un miglioramento.
- Inoltre, poiché il punto al 2.5% della distribuzione t con 19 df è 2.093, il CI al 95% sarà: $2.05 \pm (2.093 \times 0.634) = [0.72, 3.38]$
- Possiamo quindi essere sicuri al 95% che il miglioramento della differenza media reale giaccia tra 0.72 e 3.38.

CONFRONTO FRA VALORI MEDI DI PIÙ GRUPPI

ANOVA

Nel caso in cui si vogliano confrontare i valori medi di più gruppi (con distribuzione normale), si ricorre all'**analisi della varianza** (ANOVA).

- Qualora si fosse interessati a confrontare i valori medi di una variabile fra più di due gruppi (ad esempio la pressione arteriosa sistolica confrontando tre trattamenti diversi), sarà necessario utilizzare un test statistico che rappresenta una estensione del test t di Student: l'analisi della varianza.
- Descrivere in dettaglio tale tipo di analisi va oltre i nostri scopi; è sufficiente sapere che si parte dall'ipotesi nulla che i valori medi (tre o più) messi a confronto siano uguali.
- Il test produce un valore (che si chiama F), al quale corrisponde un valore di p; se quest'ultimo è < 0.05, allora rifiutiamo l'ipotesi nulla e concludiamo che i valori medi messi a confronto non sono uguali.

CORRELAZIONE FRA VARIABILI CATEGORICHE

Che relazione intercorre fra trattamento ed evento?

	Evento SI	Evento NO	Totale
Placebo	20	10	30
Farmaco attivo	8	22	30

- Per rispondere a questo quesito (ricerca di una correlazione fra due variabili categoriche), bisogna innanzitutto conoscere il numero di pazienti in ciascuna categoria di risposta (sia in numeri assoluti che in percentuale).
- Si costruisce così la **tabella di contingenza 2x2**, che confronta direttamente la distribuzione degli eventi in base ai trattamenti somministrati.

Fra i test più utilizzati per l'analisi statistica di dati clinici va sicuramente annoverato il test del chi quadrato (χ^2). Tale test viene utilizzato per confrontare percentuali fra due o più classi di una variabile. Se ad esempio stiamo confrontando l'effetto di un farmaco rispetto al placebo riguardo lo sviluppo di un particolare evento (ad esempio insorgenza di infarto del miocardio, mortalità, sviluppo di complicanze, etc.), allora la domanda che ci porremo è la seguente: *la percentuale di eventi che si è verificata nei soggetti trattati con il farmaco è più bassa, uguale, o più alta di quella che si è avuta nei soggetti che hanno assunto il placebo*?

CORRELAZIONE FRA VARIABILI CATEGORICHE

Che relazione intercorre fra trattamento ed evento?

	Evento SI	Evento NO	Totale
Placebo	20	10	30
Farmaco attivo	8	22	30

- Per rispondere a questa domanda, è necessario innanzitutto costruire una **tabella di contingenza**, nella quale riporteremo, in quattro cellette separate, il numero di soggetti trattati con il farmaco o con il placebo, che abbiano sviluppato o meno l'evento di interesse.
- Nell'esempio riportato, sono stati studiati complessivamente 60 pazienti, di cui 30 trattati con il farmaco attivo e 30 con il placebo.
- Fra i soggetti trattati con il farmaco attivo, 8 hanno sviluppato l'evento e 22 no, mentre fra quelli che hanno assunto il placebo 20 soggetti su 30 hanno sviluppato l'evento. Tale differenza è statisticamente significativa?

CORRELAZIONE FRA VARIABILI CATEGORICHE

Tabella di contingenza farmaco/evento osservato

	Evento SI	Evento NO	Totale
Placebo n° (%)	20 (66.7%)	10 (33.3%)	30 (100%)
Farmaco n° (%)	8 (28.6%)	22 (73.3%)	30 (100%)
TOTALE n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% di eventi osservati

Tabella di contingenza farmaco/evento atteso

	Evento SI	Evento NO	Totale
Placebo n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
Farmaco n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
TOTALE n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% di eventi attesi ipotizzando che farmaco o placebo abbiano la stessa efficacia

- Se nella tabella di contingenza riportiamo le percentuali di eventi/non eventi in relazione al trattamento, si evidenzia come la percentuale di pazienti che hanno avuto un evento è del 66.7% fra quelli che hanno assunto il placebo e del 28.6% fra quelli trattati con il farmaco.
- Sommando il numero di pazienti che hanno avuto un evento, a prescindere dal trattamento ricevuto, vediamo come questo sia pari a 28.

CORRELAZIONE FRA VARIABILI CATEGORICHE

Tabella di contingenza farmaco/evento <u>osservato</u>

	Evento SI	Evento NO	Totale
Placebo n° (%)	20 (66.7%)	10 (33.3%)	30 (100%)
Farmaco n° (%)	8 (28.6%)	22 (73.3%)	30 (100%)
TOTALE n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% di eventi osservati

Tabella di contingenza farmaco/evento <u>atteso</u>

	Evento SI	Evento NO	Totale
Placebo n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
Farmaco n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
TOTALE n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% di eventi attesi ipotizzando che farmaco o placebo abbiano la stessa efficacia

- A questo punto, l'ipotesi nulla da testare è che la percentuale di eventi sia la stessa nei due gruppi messi a confronto.
- Poiché i due gruppi sono di uguale numerosità, essendo il numero totale di eventi pari a 28, dovremmo aspettarci, se l'ipotesi nulla fosse vera, che la metà di questi eventi si fosse verificata fra i pazienti che hanno assunto il placebo e l'altra metà fra i pazienti che hanno assunto il trattamento attivo.

CORRELAZIONE FRA VARIABILI CATEGORICHE

	Evento SI	Evento NO	Totale
PLACEBO osservati	20	10	30
attesi	14	16	
FARMACO osservati	8	22	30
attesi	14	16	
TOTALE n°	28	32	60

Test del χ^2 (chi quadrato)

$$\chi^2 = \sum_i \frac{(\text{Osservati}_i - \text{Attesi}_i)^2}{\text{Attesi}_i}$$

$\chi^2 = (20-14)^2 / 14 + (10-16)^2 / 16 + (8-14)^2 / 14$
$+ (22-16)^2 / 16 = 9.643$

- Il test del chi quadrato si basa sul calcolo delle differenze fra valori osservati e valori attesi entro ognuna delle 4 cellette.
- Il valore di ciascuna differenza (Osservati−Attesi) viene elevato al quadrato e diviso per il rispettivo valore atteso. Si sommano quindi i valori ottenuti, come nell'esempio riportato.
- Un valore di chi quadrato pari a 3.84 corrisponde ad un valore di p=0.05.
- Se otteniamo quindi, in una tabella 2x2, un valore $\chi^2 \geq 3.84$ (cioè $p \leq 0.05$), respingeremo l'ipotesi nulla e concluderemo che le percentuali sono significativamente diverse.
- Nel caso dell'esempio, a un chi quadrato di 9.643 corrisponde in valore di p=0.002.

DALL'ANALISI UNIVARIATA

ALL'ANALISI MULTIVARIATA

DALL'ANALISI UNIVARIATA ALL'ANALISI MULTIVARIATA

- Tutte le volte che, nell'ambito di una sperimentazione, si verifica uno sbilanciamento in uno o più fattori prognostici importanti oppure si vuole stimare il ruolo di un fattore in assenza di problemi di confondimento creati da altre variabili **è necessario l'impiego di analisi multivariate**.
- Nelle sperimentazioni cliniche controllate la randomizzazione rappresenta la tecnica più efficace per garantire che i gruppi messi a confronto siano quanto più possibile confrontabili per quanto riguarda tutte le caratteristiche cliniche e socio-demografiche.
- Solo in questo modo abbiamo sufficienti garanzie che, qualora si documentassero differenze di efficacia fra i trattamenti messi a confronto, tali differenze possano essere effettivamente attribuibili al trattamento e non a differenze di fondo nei gruppi messi a confronto.
- Tuttavia, poiché la randomizzazione è per natura un processo di assegnazione del tutto casuale, potrebbe accadere che i gruppi in studio possano differire per quanto riguarda una o più caratteristiche ritenute importanti.
- Nell'analisi dei dati dello studio sarà quindi necessario tenere conto di questi sbilanciamenti, utilizzando appropriate tecniche di analisi multivariata.

DALL'ANALISI UNIVARIATA ALL'ANALISI MULTIVARIATA

	Farmaco A	Farmaco B	*p*
HbA_{1c}	7.1±1.6	7.5±1.7	0.0001
Età	62±10	64±11	0.0001
BMI	27.5±3.9	28.4±5.1	0.0001
Durata	11.2±8.6	11.7±8.5	0.1300

- Supponiamo di avere due gruppi di pazienti affetti da diabete mellito di tipo 2, uno dei quali sia stato randomizzato ad essere trattato con l'ipoglicemizzante A e l'altro con l'ipoglicemizzante B.
- Dalla tabella si desume che il controllo metabolico, espresso come valori medi di emoglobina glicosilata (HbA_{1c}), è significativamente(*) migliore con il farmaco A rispetto a B.
- Tuttavia, analizzando le caratteristiche dei pazienti nei due gruppi, osserviamo che i pazienti assegnati al farmaco B sono significativamente(*) più anziani e presentano un indice di massa corporea (BMI) significativamente(*) più alto.

Il farmaco A è davvero più efficace del farmaco B, o la differenza nel controllo metabolico è legata alla più giovane età e al BMI più basso?

Quale sarebbe la differenza vera nei valori di HbA_{1c} se i due gruppi avessero stessa età e BMI?

(*) Significativamente dal punto di vista statistico…

DALL'ANALISI UNIVARIATA ALL'ANALISI MULTIVARIATA

	Farmaco A	Farmaco B	*p*
HbA_{1c}	7.1±1.6	7.5±1.7	0.0001
Età	62±10	64±11	0.0001
BMI	27.5±3.9	28.4±5.1	0.0001
Durata	11.2±8.6	11.7±8.5	0.1300

- I due gruppi non differiscono invece (significativamente) per quanto riguarda la durata del diabete.
- A questo punto sorgono spontanee le domande:
  - *le differenze nei valori medi di HbA_{1c} sono davvero legate al trattamento, o sono una conseguenza della differenza di età e di peso corporeo*?
  - *quale sarebbe l'effetto vero del trattamento se i due gruppi avessero la stessa età e lo stesso BMI?*
- A queste domande è possibile dare una risposta utilizzando le analisi multivariate, che permettono di valutare il ruolo indipendente di ciascun fattore considerato, a parità di tutti gli altri.

Il farmaco A è davvero più efficace del farmaco B, o la differenza nel controllo metabolico è legata alla più giovane età e al BMI più basso?

Quale sarebbe la differenza vera nei valori di HbA_{1c} se i due gruppi avessero stessa età e BMI?

DALL'ANALISI UNIVARIATA ALL'ANALISI MULTIVARIATA

L'analisi multivariata ci permette di valutare il ruolo indipendente di ogni singola variabile (fattore), a parità di tutte le altre.

$y = \alpha + \beta x$ ⇐ analisi univariata

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots$ ⇐ analisi multivariata

$$\mathrm{HbA}_{1c} = \alpha + \beta_1 \cdot \mathrm{farmaco} + \beta_2 \cdot \mathrm{BMI} + \beta_3 \cdot \mathrm{età} + \beta_4 \cdot \mathrm{durata}$$

- Nel caso di una variabile dipendente continua, come nel nostro esempio, l'analisi multivariata appropriata è rappresentata dalla **regressione lineare multipla**, che rappresenta la normale estensione della regressione lineare semplice, di cui abbiamo già parlato.
- A differenza della regressione lineare semplice, nella quale avevamo una sola variabile indipendente, in questo caso andremo a valutare in che misura varia la variabile dipendente al variare di più variabili indipendenti. In questo caso, quindi, stimeremo più parametri , uno per ogni variabile indipendente inclusa nel modello.
- Nel nostro esempio, andremo a valutare come variano i livelli di HbA_{1c} al variare del farmaco utilizzato, del BMI, dell'età dei pazienti e della durata della malattia.

DALL'ANALISI UNIVARIATA ALL'ANALISI MULTIVARIATA

	β	*p*
HbA_{1c}	0.31	0.0001
Età	-0.0053	0.08
BMI	18	0.007
Durata	14	0.001

Interpretazione

- *a parità di età, BMI e durata del diabete, il farmaco B è associato ad una HbA_{1c} significativamente più alta di 0.31 rispetto al farmaco A;*
- *i valori di HbA_{1c} crescono inoltre in modo significativo all'aumentare del BMI e della durata del diabete, mentre tendono a ridursi all'aumentare dell'età (sebbene quest'ultimo risultato non raggiunga la significatività statistica).*

- I risultati della regressione multipla applicata al nostro esempio sono riportati nella tabella. Essa ci mostra come il tipo di trattamento sia significativamente correlato ai valori di HbA_{1c}, così come il BMI e la durata del diabete.
- L'interpretazione dei risultati è pertanto la seguente: a parità di età, BMI e durata del diabete, i valori di HbA1c sono significativamente più alti nei pazienti trattati con il farmaco B rispetto a quelli trattati con il farmaco A (il valore del β ci dice che in media sono più alti di 0.31).
- La regressione ci fornisce anche altre indicazioni, poiché ora sappiamo che, a parità di trattamento, il controllo metabolico è comunque significativamente peggiore all'aumentare del BMI e della durata del diabete, mentre non cambia sostanzialmente al cambiare dell'età.

REGRESSIONE LOGISTICA

- Nell'esempio precedente la **variabile dipendente** (HbA_{1c}) era una variabile **continua**, **normalmente distribuita**. In questi casi si utilizza la **regressione multipla**.
- Nei casi in cui la **variabile dipendente** è **dicotomica** (Sì/No), si utilizza la **regressione logistica** binaria.
- Analogamente, per variabili continue con distribuzione molto distante da quella normale, variabili ordinali o nominali a più livelli, si utilizza la regressione logistica dopo aver dicotomizzato la variabile dipendente.
- Se quindi l'end-point dello studio in esame è rappresentato da una variabile categorica invece che continua, il modello di analisi multivariata da utilizzare è rappresentato dalla regressione logistica binaria.
- Ricordiamo inoltre che i test parametrici, a cui appartiene anche la regressione lineare multipla, richiedono che la variabile dipendente sia normalmente distribuita. Qualora non lo fosse, è preferibile dicotomizzare la variabile rispetto ad un valore clinicamente rilevante, ed utilizzare la regressione logistica.

REGRESSIONE LOGISTICA

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

- La regressione logistica è un'estensione del test del χ^2, e i risultati vengono espressi come "**odds ratio**" (OR) con gli intervalli di confidenza (CI) al 95%:

$$\mathrm{OR} = e^{\beta_1}$$

- Un OR > 1 indica un eccesso di rischio, mentre un valore di OR < 1 indica un rischio più basso rispetto alla categoria di riferimento.
- Se l'intervallo di confidenza al 95% include il valore nullo di 1, allora il risultato non è statisticamente significativo.

REGRESSIONE LOGISTICA

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots$$

- L'equazione che descrive la regressione logistica non è quindi molto diversa da quella della regressione multipla; anche in questo caso si studia infatti il rapporto lineare fra variabile dipendente ed una serie di variabili indipendenti, ad ognuna delle quali corrisponde un coefficiente β.
- Ciò che contraddistingue la regressione logistica rispetto alla regressione multipla è la grandezza che si trova a sinistra dell'equazione.
- Qui infatti non abbiamo più il valore di y, e cioè della variabile dipendente, ma il logaritmo dell'*odd*, cioè della probabilità di avere l'evento di interesse divisa per la probabilità di non averlo.
- Tale caratteristica fa sì che l'esponenziale di ciascun coefficiente β rappresenterà l'*odds ratio*, e quindi esprimerà il rischio di avere l'evento per una specifica categoria rispetto a quella di riferimento.

ODDS RATIO

Con il termine inglese **odds** si intende il rapporto tra la probabilità *p* di un evento e la probabilità che tale evento non accada (cioè la probabilità 1−*p* dell'evento complementare):

$$\frac{p}{1-p}$$

Esempio (Wikipedia).

- Nella cittadina statunitense di Framingham, su un totale di 656 soggetti appartenenti alla classe di età fra 50 e 60 anni all'epoca della prima rilevazione, 130 di costoro svilupparono una coronaropatia durante un follow-up di 12 anni.
- La probabilità *p* dell'evento "sviluppo di una coronaropatia" è data dalla proporzione 130/656 = 0.20 = 20%.
- L'*odds* di sviluppare una coronaropatia è: *p* / (1−*p*) = 0.20 / (1−0.20) = 0.20 / 0.80 = 0.25; ossia, 1 a 4 (25%).

ODDS RATIO

- Il logaritmo naturale dell'*odds* è detto **logit**.
- Il rapporto tra due *odds* è detto *odds ratio* (OR) e misura l'associazione tra due fattori, per esempio tra un fattore di rischio e una malattia.
- Il calcolo dell'*odds ratio* prevede il confronto tra le frequenze di comparsa dell'evento (ad esempio, malattia) rispettivamente nei soggetti esposti e in quelli non esposti al fattore di rischio in studio.
- Se OR = 1, significa che il fattore di rischio è ininfluente sulla comparsa della malattia.
- Se OR > 1, il fattore di rischio è o può essere implicato nella comparsa della malattia.
- Se OR < 1, il fattore di rischio in realtà è una difesa contro la malattia.
- Esso è utilizzato negli studi retrospettivi (casi-controlli), dove non è necessaria la raccolta dei dati nel tempo; infatti, esso non calcola un andamento ed è, anzi, indipendente dal fattore durata.

RISCHIO RELATIVO

- Negli studi prospettici si utilizza invece, allo stesso scopo, il calcolo del rischio relativo.
- Il **rischio relativo** (*risk rate*, **RR**) è la probabilità che un soggetto, appartenente ad un gruppo esposto a determinati fattori, sviluppi la malattia, rispetto alla probabilità che un soggetto appartenente ad un gruppo non esposto sviluppi la stessa malattia.
- Questo indice è utilizzato negli studi di coorte dove l'esposizione è misurata nel tempo:

$$\mathrm{RR} = \frac{I(\text{esposti})}{I(\text{non esposti})}$$

dove I = incidenza, che si definisce come:

$$\mathrm{I} = \frac{\text{n.ro nuovi ammalati}}{\text{n.ro totale persone - n.ro ammalati}}$$

- Se:
  - RR = 1 il fattore di rischio è ininfluente sulla comparsa della malattia;
  - RR > 1 il fattore di rischio è implicato nel manifestarsi della malattia;
  - RR < 1 il fattore di rischio dipende dalla malattia (fattore di difesa).
- Esempi di applicazione di tale formula sono gli studi riguardanti la correlazione tra il fumo e lo sviluppo di cancro al polmone, nei quali sono stati riscontrati RR > 17.

RISORSE WEB PER IL CALCOLO STATISTICO

RISORSE WEB PER IL CALCOLO STATISTICO

Corso di **Analisi dei Dati**

(docente Prof. Barbaranelli, La Sapienza)

https://elearning2.uniroma1.it/course/view.php?id=1796

Questo corso fornisce le basi necessarie per lo studio delle principali tecniche di analisi multivariata dei dati.

Pacchetto **Real Statistics**

http://www.real-statistics.com

Questo sito è una guida pratica su come fare analisi statistica con il foglio elettronico, senza la necessità di sofisticati (e costosi) software specializzati. Comprende un software gratuito che estende le capacità statistiche del foglio elettronico, in modo da eseguire con facilità un'ampia varietà di analisi statistiche.

RISORSE WEB PER IL CALCOLO STATISTICO

Testo online **Handbook of Biologica Statistics**

(di J.H. McDonald)

http://www.biostathandbook.com

Si tratta di un completo ed esauriente textbook disponibile online, che mostra come effettuare la scelta del test statistico appropriato per un particolare esperimento, e quindi applicarlo ed interpretarne i risultati. È corredato di molti esempi e risorse disponibili su web.

Software statistico e di Data Mining **Tanagra**

http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

TANAGRA è un software di "data mining" per uso accademico e di ricerca. Propone svariati metodi di data mining dall'analisi esplorativa dei dati, all'apprendimento statistico, machine learning e database.

RISORSE WEB PER IL CALCOLO STATISTICO

Software statistico e di Data Mining **Orange**

https://orange.biolab.si/

ORANGE è un software open source di "machine learning" e visualizzazione dati per uso professionale e personale. Possiede un'interfaccia applicativa interattiva orientata al workflow per l'analisi dei dati, con una ricca toolbox di metodi e algoritmi.

Software statistico e di analisi esplorativa dei dati **Jamovi**

https://www.jamovi.org

Jamovi è un foglio di calcolo statistico open source e gratuito, basato sul linguaggio statistico R, di facile e immediato utilizzo. È particolarmente adatto per l'analisi esplorativa rapida di dataset, e contiene efficaci presentazioni tabellari e grafiche dei dati.

Grazie per l'attenzione