Introduzione e statistica descrittiva

Docente: Prof.ssa Paola Borrelli
paola.borrelli@unich.it

La statistica medica è uno strumento per…….

1) Identificare le cause e i fattori di rischio delle malattie

2) Conoscere la storia naturale della patologia

3) Analizzare la distribuzione delle malattie nelle popolazioni

4) Valutare gli interventi sanitari

La statistica medica

metodologia generale per lo studio dei fenomeni

collettivi e quindi della variabilità attraverso:

- ✓ osservazione dei fenomeni
- ✓ traduzione in simboli
- ✓ evidenza delle irregolarità
- ✓ verifica di ipotesi

Osservazione dei fenomeni

1) La statistica medica insegna come occorre osservare la realtà, come raccogliere i dati, con quali strumenti, su chi, su quanti, etc.

Traduzione in simboli

2) La statistica medica insegna come riassumere in modo corretto i dati raccolti, cioè come esprimerli e rappresentarli in modo sintetico e appropriato in funzione del tipo di variabile e della sua distribuzione

Evidenza delle irregolarità

3) Una appropriata analisi statistica di variabili raccolte, consente di evidenziare quei fenomeni che si scostano dalla “normalità” cioè dai valori più frequenti in un set di dati: spesso l’approfondimento su tali irregolarità è molto informativo e può generare nuove ipotesi

Verifica di ipotesi

4) È l’aspetto più interessante della metodologia statistica: la possibilità di verificare se un’ipotesi posta sia o non sia confutata dalla osservazione della realtà

Studio dei fattori di rischio per le patologie cardiovascolari nella popolazione generale

La conoscenza dello stato di salute di una popolazione è un obiettivo prioritario per l'appropriatezza degli interventi sanitari, la razionalizzazione nella produzione di servizi e nel consumo di risorse.

1. Chi?
2. Cosa?
3. Perchè?
4. Come?

1. Chi?

Popolazione di un comune Italiano

2. Cosa?

Analizzare la presenza di fattori di rischio per l'aterosclerosi per valutare l'incidenza di ischemia miocardica e cerebrale nella popolazione generale.

3. Perché?

Monitorare lo stato di salute della popolazione

4. Come?

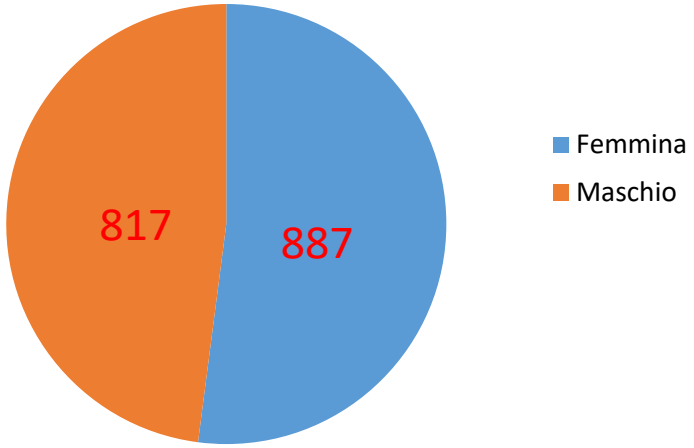
Attraverso fonti di dati idonee a identificare le caratteristiche anagrafiche e cliniche dei soggetti appartenenti alla popolazione

Matrice dei dati

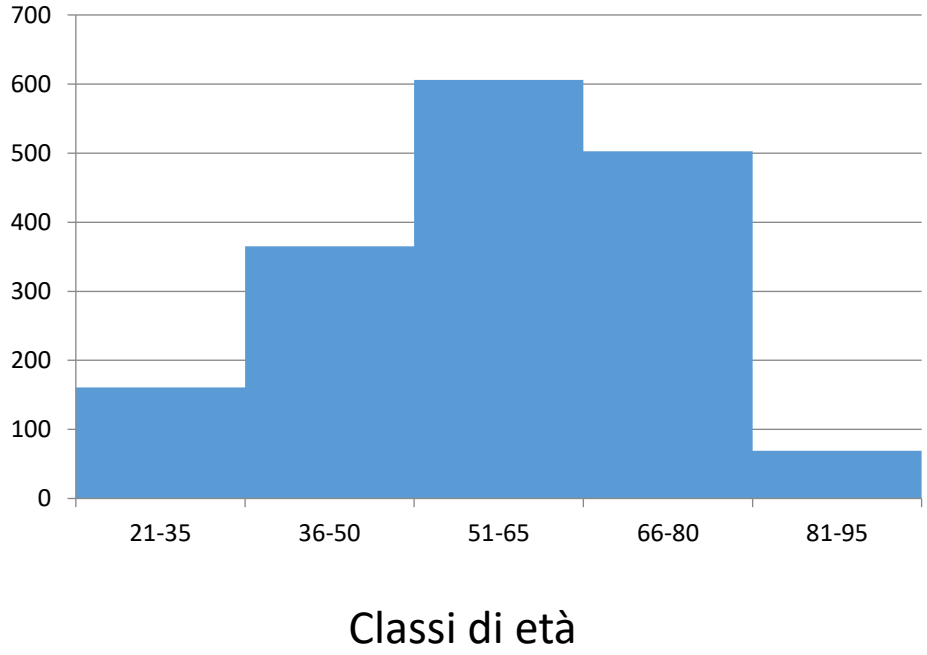
IDNUM	Sesso	ETA96	Maggioredi65	Hpt	ILP	Diabete	Fumo	Obeso	RIC	GD	NRIC_CVD	GD_CVD	RIC_CVD	Decesso
1	Femmina	67	1	0	0	0	0	0	0		0		0	0
2	Femmina	64	0	1	1	0	0	0	1	2	1	2	1	1
3	Femmina	80	1	1	1	0	0	0	0		0		0	1
4	Femmina	66	1	0	1	0	0	0	0		0		0	0
6	Femmina	48	0	0	0	0	1	0	0		0		0	0
7	Femmina	79	1	1	0	0	1	1	0		0		0	0
9	Femmina	63	0	0	1	0	0	0	3	41	0		0	0
10	Femmina	66	1	0	1	0	0	0	0		0		0	0
11	Femmina	85	1	1	0	1	0	1	2	14	1	13	1	1
14	Femmina	63	0	0	1	0	0	0	0		0		0	0
16	Femmina	64	0	0	1	0	0	0	1	2	0		0	0
17	Femmina	71	1	0	0	0	0	0	0		0		0	0
25	Femmina	63	0	1	1	0	0	0	0		0		0	0
28	Femmina	55	0	1	1	0	0	0	0		0		0	1
29	Femmina	50	0	0	1	0	0	0	2	20	0		0	0
32	Femmina	77	1	0	0	0	0	0	0		0		0	0
33	Femmina	55	0	0	0	0	0	0	0		0		0	0
35	Femmina	57	0	0	1	0	0	0	1	9	0		0	0
37	Femmina	72	1	1	1	1	0	1	5	22	2	13	1	1
38	Femmina	69	1	1	1	0	0	1	1	7	0		0	0
49	Femmina	52	0	0	1	0	0	0	0		0		0	0

Caratteristiche dei soggetti

Distribuzione di frequenza delle caratteristiche dei soggetti (n=1704)	
Ipertensione, n (%)	
Si	**427 (25.06%)**
No	**1277 (74.94%)**



Età media: 58±15 anni



Distribuzione di frequenza del diabete per genere (n=1704)			
	Maschi	Femmine	**Totale**
Diabete, n (%)			
Si	22 (2.69%)	19 (2.14%)	**41 (2.41%)**
No	795 (97.31%)	868 (97.86%)	**1663 (97.59%)**
	817 (100%)	**887 (100%)**	**1704 (100%)**

Confronto delle caratteristiche rilevate con la presenza di ricovero per malattie cardiovascolari

	Età>65 aa	Fumo	Obesità	Hpt	Ilp	Diabete
χ^2	54.486	3.536	0.005	45.667	2.288	**3.699**
p-value	**0.0000**	**0.060**	**0.945**	**0.00**	**0.130**	**0.054**

- Età ≥65 anni
- Presenza di ipertensione
- Presenza di diabete

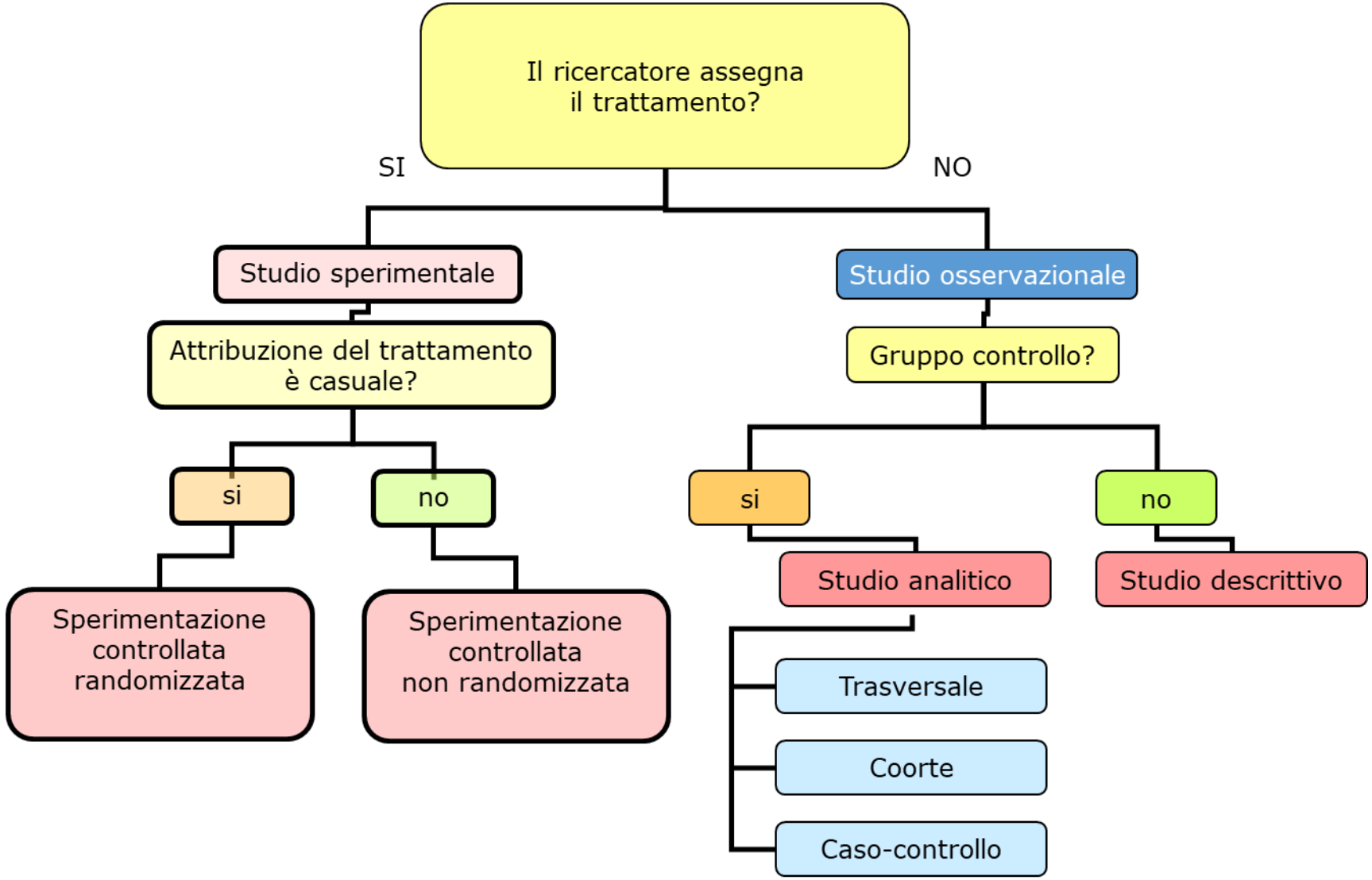
↓

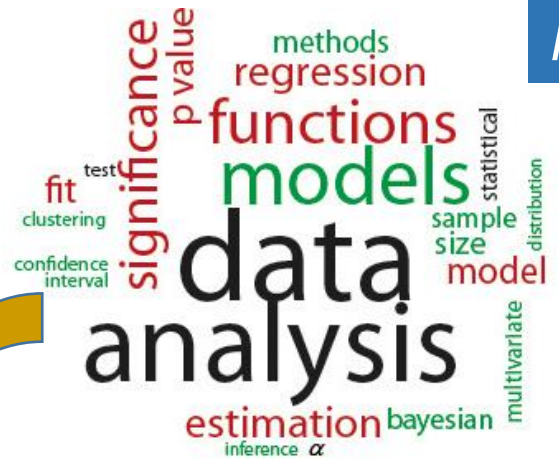
Ricovero per malattie cardiovascolari

Conclusioni: Esiste una relazione tra il profilo di salute e l'incidenza di ricovero per patologia cardiovascolare. I dati suggeriscono una pianificazione di attività preventive volte alla riduzione dei fattori di rischio.

Le tappe di una ricerca

- **Fase preliminare:**
  - chiarire lo scopo dell'indagine
  - formulare l'ipotesi scientifica
- **Pianificare l'indagine**
- **Analizzare i dati**
- **Interpretare i risultati**
- **Produrre evidenza**


Il ricercatore assegna il trattamento?
SI
NO
Studio sperimentale
Attribuzione del trattamento è casuale?
si
no
Sperimentazione controllata randomizzata
Sperimentazione controllata non randomizzata
Studio osservazionale
Gruppo controllo?
si
no
Studio analitico
Studio descrittivo
Trasversale
Coorte
Caso-controllo



STATISTICA INFERENZIALE

STIMARE

DESCRIVERE

STATISTICA DESCRITTIVA

Campione casuale

TRARRE CONCLUSIONI

STATISTICA INFERENZIALE

CAMPIONAMENTO

Popolazione

Popolazione obiettivo

insieme di tutti gli ipotetici elementi oggetto del nostro interesse, legate da una CARATTERISTICA COMUNE che consente di stabilire un criterio di appartenenza alla popolazione stessa

FINITA

se è possibile produrre l'elenco di tutti gli elementi oggetto di interesse

INFINITA

se si tratta di una popolazione ideale, di cui non è possibile produrre un elenco

Popolazione campionata (o base di campionamento)

rappresenta l'aspetto operativo della popolazione obiettivo

Se la popolazione è finita

è possibile ottenere la lista dei soggetti della popolazione stessa.

Se la popolazione è infinita

è necessario scegliere una definizione operativa della popolazione

Campione

sottoinsieme della popolazione di numerosità limitata selezionato attraverso la tecnica del **campionamento**.

- ✓ RAPPRESENTATIVO della popolazione
- ✓ SUFFICIENTEMENTE ampio
- ✓ L'informazione deve essere raccolta su tutti (o quasi tutti) gli elementi del campione

Tipi di Campionamento

I) campionamenti probabilistici: la scelta delle unità statistiche da sottoporre allo studio è regolata dalle leggi della probabilità;
II) campionamenti non probabilistici: la scelta delle unità statistiche da sottoporre allo studio non è di tipo probabilistico

Campionamenti probabilistici

- ✓ **CASUALE SEMPLICE**
- ✓ **CLUSTER o GRAPPOLO**
- ✓ **STRATIFICATO**

Campionamento casuale semplice

- ✓ Ogni elemento della popolazione ha la stessa probabilità di essere incluso nel campione
- ✓ unità di campionamento = individuo
- ✓ Procedura:
  - ➢ preparare una base di campionamento
  - ➢ scegliere l'ampiezza del campione
  - ➢ selezionare gli elementi del campione utilizzando una tavola di numeri casuali

Campionamento a CLUSTER o GRAPPOLO

- ✓ Un campione casuale semplice viene selezionato sulla base di un gruppo di soggetti
- ✓ unità di campionamento = cluster (città, classi di una scuola, famiglie …)

Campionamento STRATIFICATO

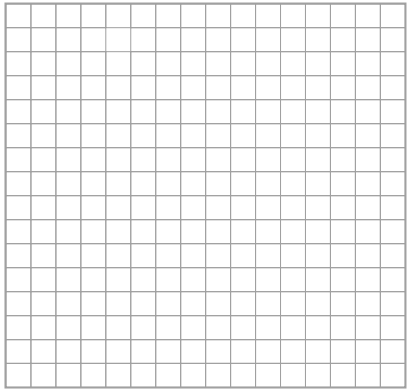
- ✓ La popolazione è suddivisa in strati omogenei al loro interno rispetto alla variabile di stratificazione
- ✓ Un campione casuale semplice viene selezionato ENTRO ogni strato
- ✓ è possibile un campionamento PROPORZIONALE o un campionamneto NON PROPORZIONALE

ESEMPIO:

Si ha una popolazione composta da 256 soggetti di cui 64 femmine e 192 maschi.

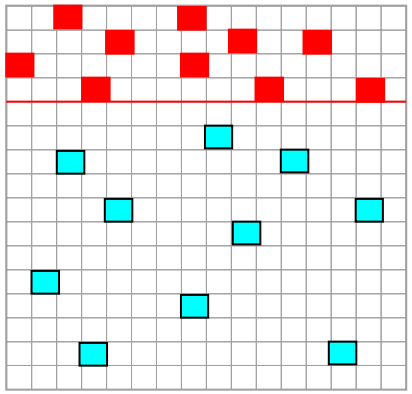
1° STRATO = FEMMINE

2° STRATO = MASCHI



STRATIFICATO NON PROPORZIONALE

campione di 20 soggetti



Femmine

Maschi

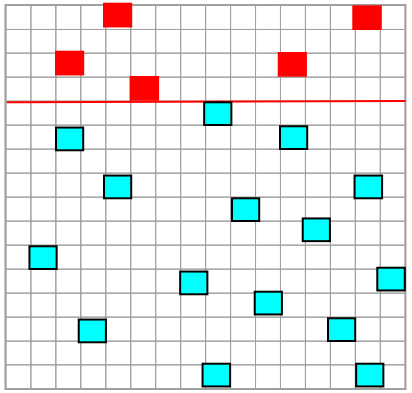
viene estratta la stessa proporzione di individui per ciascuno strato

STRATIFICATO PROPORZIONALE

popolazione

25% femmine

75% maschi



campione di 20 soggetti

5 femmine

15 maschi

Funzioni della Statistica medica

DESCRITTIVA: insieme di tecniche per la raccolta, l'organizzazione, la sintesi di dati. Prima fase di qualsiasi ricerca

INFERENZIALE: insieme di procedimenti utili a trarre conclusioni sulla popolazione obiettivo

Unità statistica Minima unità da cui si raccolgono i dati in una indagine

Variabile Caratteristica che può assumere valori diversi nelle diverse unità statistiche

Osservazione Valore assunto da una variabile in una determinata unità statistica

Variabili

QUALITATIVE

- NOMINALI
  - Dicotomiche
  - Politomiche
- ORDINALI

esiste un ordine tra le modalità

QUANTITATIVE

- DISCRETE
- CONTINUE

Variabili Qualitative

Caratteristiche che non sono misurabili, ma che possono essere classificate in categorie

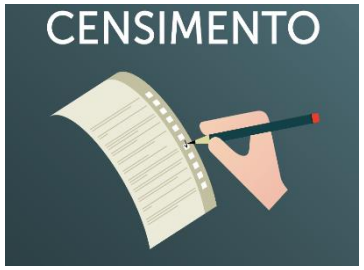
Variabili Quantitative

Derivano da conteggi (variabili discrete) o misure (variabili continue). Sono espresse da numeri.

Come sintetizzare le informazioni raccolte?

- Matrice dei dati
- Tabelle a singola e doppia entrata
- Distribuzioni di frequenza assoluta, percentuale e cumulata
- Rappresentazioni grafiche
- Calcolo delle misure di sintesi

Matrice di dati



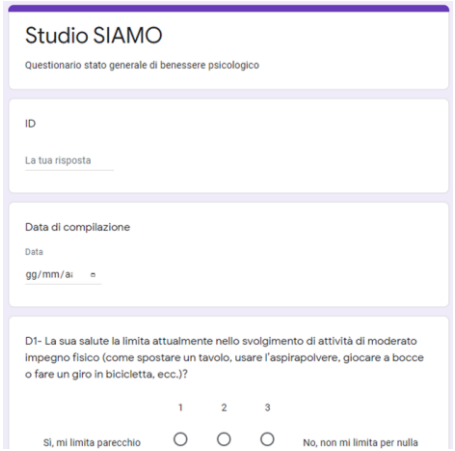
1) Fonti ufficiali

Rilevazioni demografiche
Rilevazioni del sistema sanitario

SDO
Scheda di dimissione ospedaliera

2) Fonti non ufficiali

Indagini ad hoc


Studio SIAMO

Questionario stato generale di benessere psicologico

ID

La tua risposta

Data di compilazione

Data

gg/mm/aa

D1- La sua salute la limita attualmente nello svolgimento di attività di moderato impegno fisico (come spostare un tavolo, usare l'aspirapolvere, giocare a bocce o fare un giro in bicicletta, ecc.)?

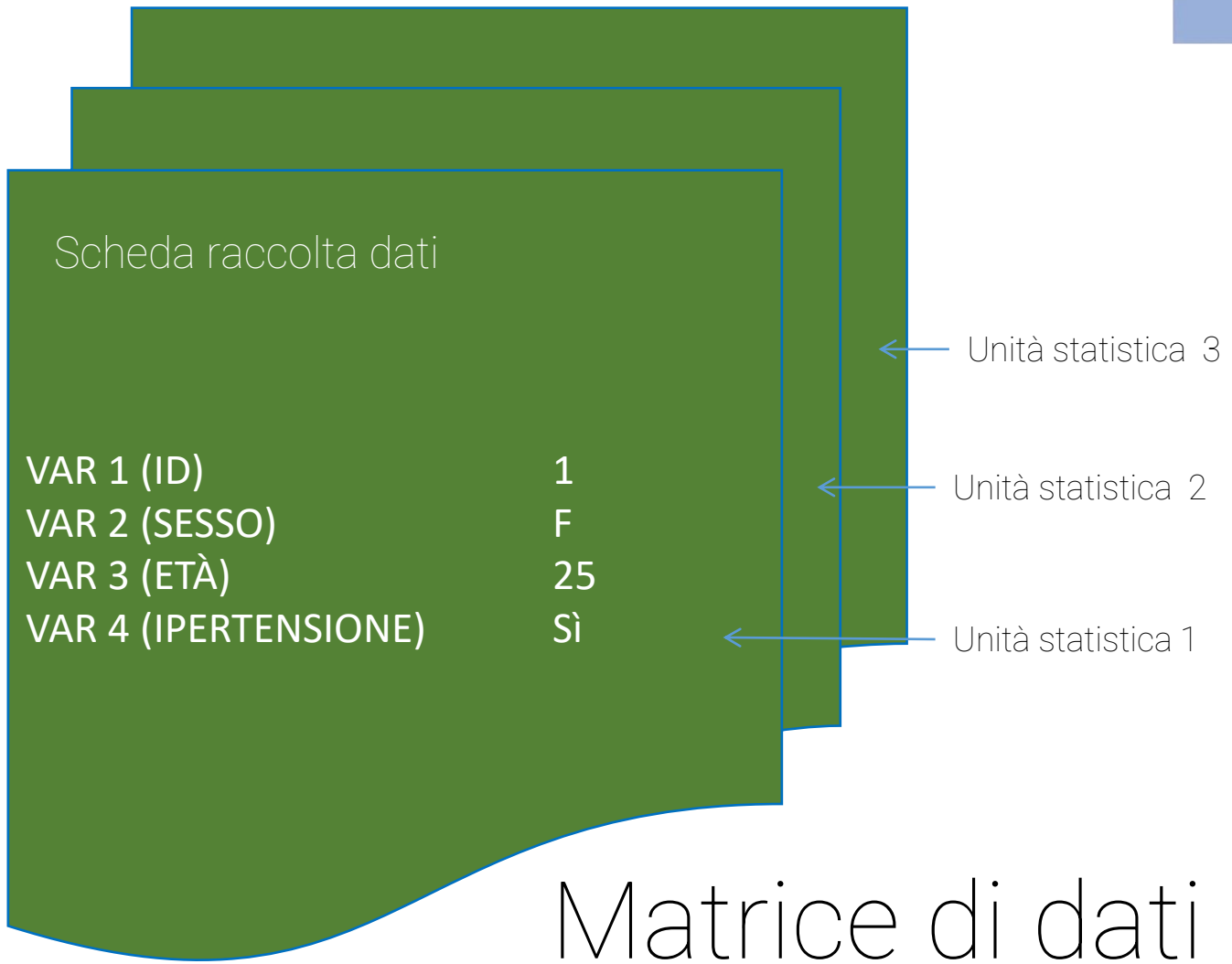
1 2 3

Sì, mi limita parecchio ○ ○ ○ No, non mi limita per nulla

I dati raccolti vengono registrati su una **scheda di raccolta dati**.

Ogni scheda di raccolta dati contiene le informazioni di una singola **unità statistica**.

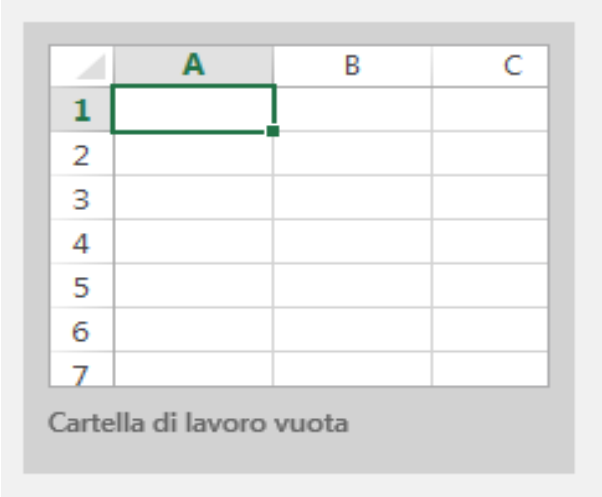
Ogni campo in una scheda di raccolta dati (per esempio ogni domanda in un questionario) corrisponde ad una **variabile**.



Matrice di dati

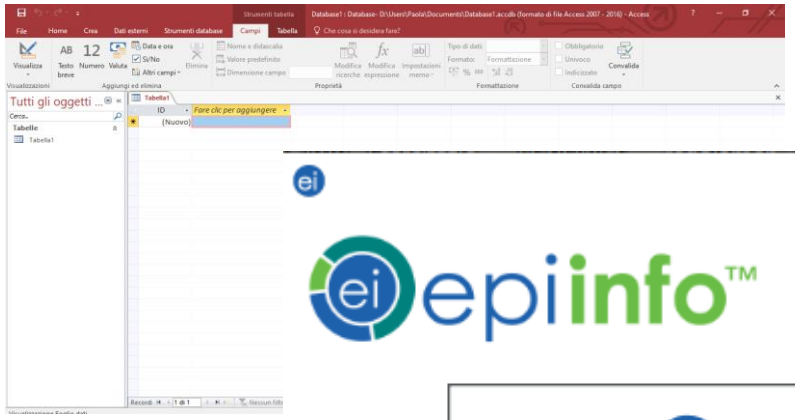
Foglio elettronico

- Lavora per coordinate
- Colonne, righe, celle
- Cartella
- Inserimento sequenziale tramite i fogli disponibili



Database

- Lavora per tabelle
- Record e campi
- Contenitore di oggetti (tabelle, maschere, query, ecc.)
- Raccolta e gestione delle informazioni con la possibilità di creare relazioni con altre tabelle

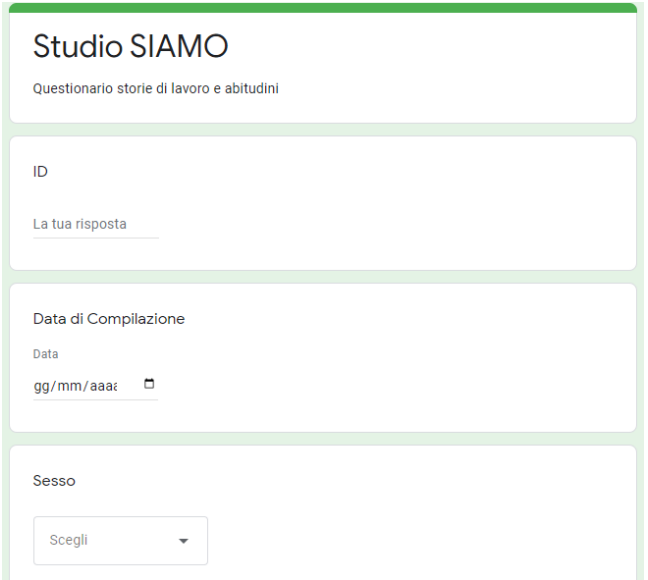


epiinfo™

CREATE FORMS
Create surveys or questionnaires with field validation and skip logic.

Creazione di moduli

Strumento che consente di raccogliere informazioni dagli utenti tramite un sondaggio o un questionario personalizzato. Le informazioni vengono quindi raccolte e automaticamente collegate a un foglio di calcolo. Il foglio di calcolo è poi compilato con le risposte che gli utenti hanno dato ai sondaggi e ai quiz



Questionario
Patologie cardiovascolari

ID ☐☐

sesso ☐ M ☐ F

anno di nascita ☐☐☐☐

età (anni) ☐☐

Indicare lo stato civile

celibe/nubile ☐
coniugato/a ☐
divorziato/a ☐
vedovo/a ☐
non rilevato ☐

Indicare i valori di pressione sistolica e diastolica

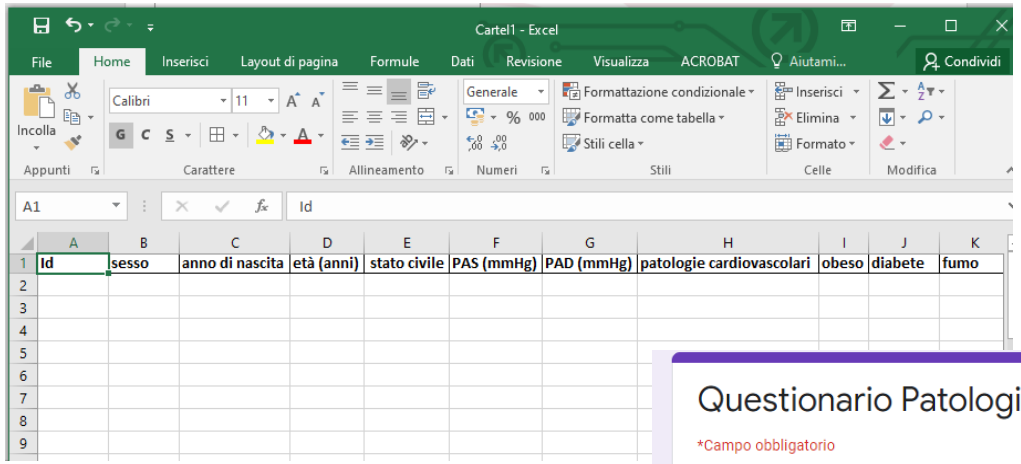
PAS (mmHg) ☐☐☐
PAD (mmHg) ☐☐☐

Indicare la presenza/assenza di patologie cardiovascolari

patologia cardiovascolare ☐ si ☐ no

Indicare la presenza/assenza di fattori di rischio (sono possibili risposte

obeso ☐
diabete ☐
fumo ☐



Id	sesso	anno di nascita	età (anni)	stato civile	PAS (mmHg)	PAD (mmHg)	patologie cardiovascolari	obeso	diabete	fumo

Questionario Patologie Cardiovascolari

Id ☐

sesso ☐ (F, M)

anno di nascita ☐

età ☐

Stato civile ☐

PAS ☐ PAD ☐ Patologie cardiovascolari ☐

Fattori di rischio

☐ Obesità ☐ Diabete ☐ Fumo

Questionario Patologie Cardiovascolari

*Campo obbligatorio

Id *

La tua risposta

sesso *

Scegli

Anno di nascita *

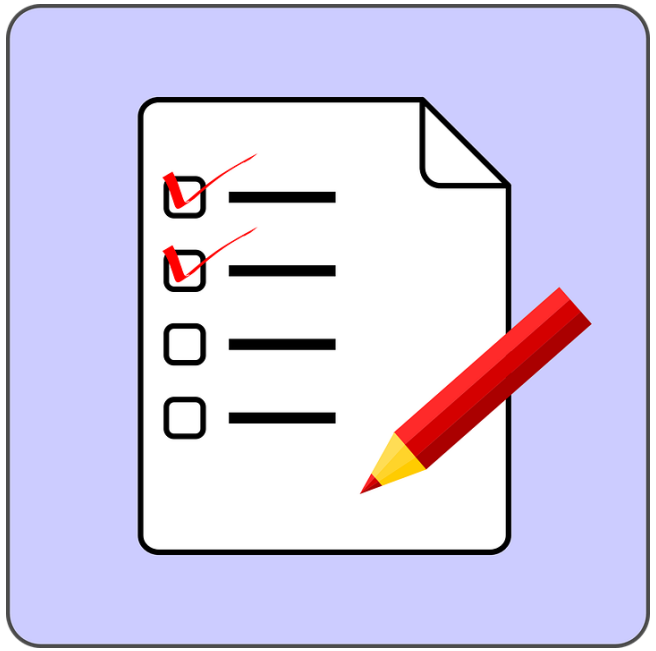
La tua risposta

Età (anni) *

variabili in colonna

id	genere	età (anni)	sindrome metabolica	peso (kg)
1	m	32	no	88
2	f	50	si	94
3	m	45	no	85
4	f	51	no	75
5	m	61	no	70
6	m	39	no	76
7	m	62	no	90
8	f	60	si	84
9	f	38	no	54
10	f	63	si	86

osservazioni in riga per ogni unità statistica



Risposte brevi
Vero/falso
Risposta multipla
Altro

codifica

Legenda

id: codice identificativo
genere: m, f
età: età dei soggetti in anni
sindrome metabolica: si, no
peso: peso dei soggetti in kg

Esempio

malattie di base

0: nessuna/non precisabile
1: idiopatica
2: sofferenza pre-perinatale
3: infezione congenita

Qualità dei dati

Errori di misura

Errori di calcolo

Gli **errori di misura** rappresentano la distanza tra la "risposta reale" e quello che viene registrato sulla scheda di raccolta dati. Gli **errori di calcolo**, invece, sono errori che accadono durante la manipolazione dei dati.

Come tenere gli errori sotto controllo?

- Tutti i dati contengono errori: è opportuno cercarli
- Non usare lo 0 come codice del dato mancante
- Stabilire delle regole comuni
- Verificare la coerenza tra il dato inserito e il dato cartaceo

Tabelle a singola e doppia entrata

SEMPLICI: unità statistiche classificate secondo UNA SOLA variabile

TABELLE

DOPPIA ENTRATA: unità statistiche classificate secondo DUE variabili

DISTRIBUZIONE DI FREQUENZA

Tabella semplice

Sesso	f_x
M	345
F	155
Totale	500

Peso dei soggetti

Peso (in Kg)	f_x
50	2
55	5
65	15
….	
86	2
Totale	500

Tabelle a doppia entrata

Calcolo delle frequenze percentuali nella tabella a doppia entrata

Tipo di attività	Sesso		
	M	F	Totale
Sport individuale	130	20	150
Sport di squadra	70	0	70
Running	40	15	55
Palestra	45	45	90
Gruppi di cammino	12	60	72
Altro	48	15	63
	345	155	500

Distribuzione di due variabili

Raggruppare in classi le osservazioni della variabile

- quante classi?
- di quale ampiezza?

Prima regola: il numero delle classi nella tabella di frequenza deve essere approssimativamente uguale alla radice quadrata del valore della dimensione n del campione, $\sqrt{n}$

Seconda regola: determinare l'ampiezza di ogni classe

valore massimo-valore minimo/numero delle classi

Esempio

- Campione di 30 soggetti (n)
- Valore massimo della distribuzione è 42
- Valore minimo della distribuzione è 5

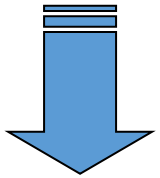
Calcolo

Numero delle classi = $\sqrt{30}$ **, 5 classi**

Ampiezza delle classi = 42-5/5 =**7.4**

CLASSI MUTUAMENTE ESCLUSIVE

CLASSI ESAUSTIVE



Devono comprendere tutti i valori dell'insieme di dati

Avere la stessa ampiezza

Non devono essere sovrapposte

7 classi, ampiezza 5

Peso (in Kg)	f_x
55-59	20
60-64	55
65-69	110
70-74	150
75-79	98
80-84	52
85-89	15
	500

Distribuzioni di frequenza assoluta, percentuale e cumulata

- **ASSOLUTE** = n° di volte che si osserva ciascuna modalità (o osservazione) di una variabile (max= n° totale delle unità stat)

- **RELATIVE** = freq. assoluta/ n° totale unità stat

- **PERCENTUALI** = freq. relativa x 100

- **CUMULATE** = somma delle frequenze (assolute, relative, percentuali) delle osservazioni precedenti all'osservazione data più la frequenza dell'osservazione stessa)

	fx	fr	frX%
Maschi	345	345/500=0.7*100=70	
Femmine	155	155/500=0.3*100=30	

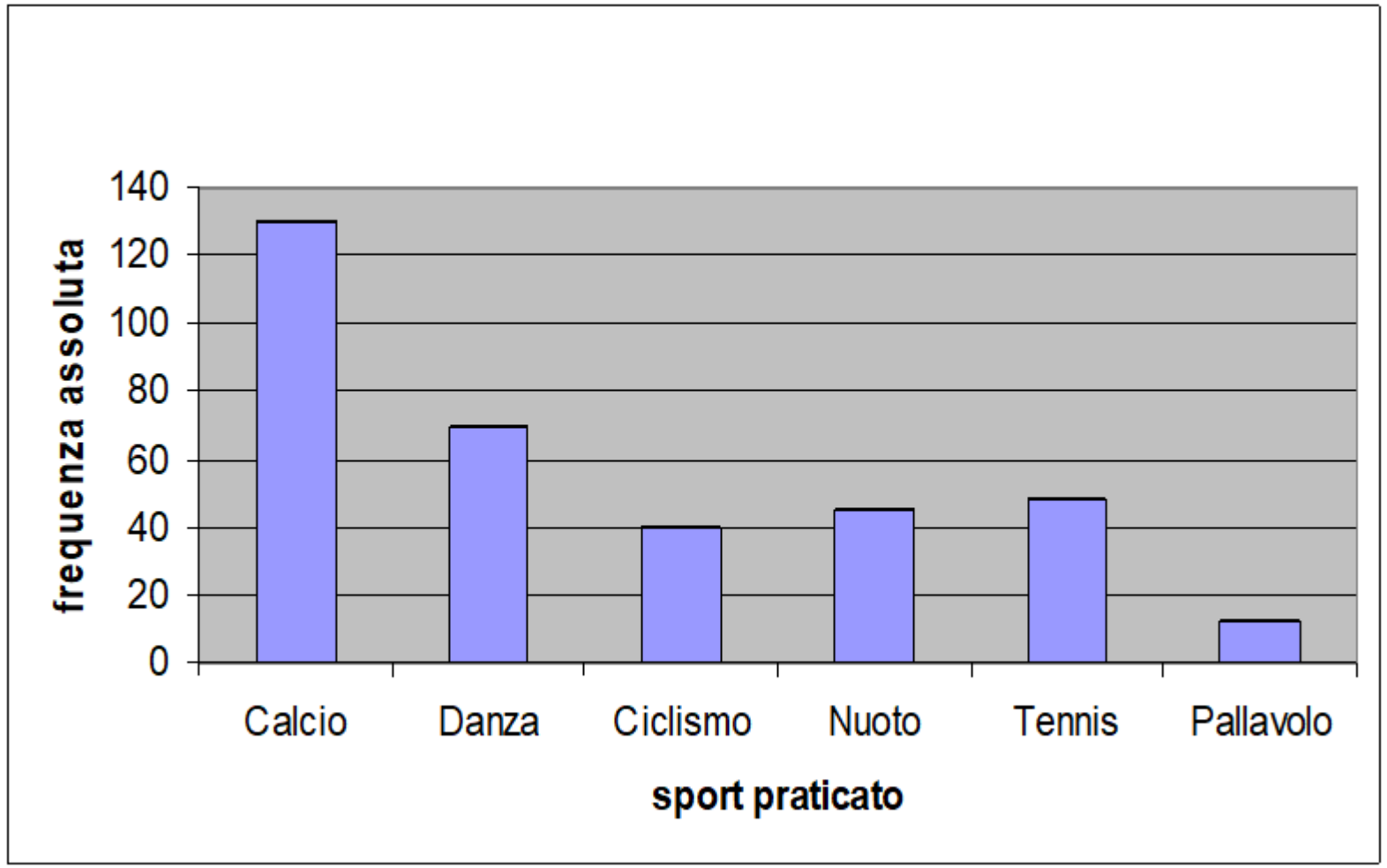
Classi di peso	f_x	$f_r = f_x / n$	$f_{\%} = f_r X100$	f_{ca}
55-59	20	0.04	4	20
60-64	55	0.11	11	20+55=75
65-69	110	0.22	22	20+55+110=185
70-74	150	0.30	30	20+55+110+150=335
75-79	98	0.196	20	20+…….+98=433
80-84	52	0.104	10	20+…….+52=485
85-89	15	0.03	3	20+……+15=500
Totale	500 (n)	1.00	100	

Rappresentazioni grafiche

VARIABILI QUALITATIVE

- DIAGRAMMA A BARRE O COLONNE (la base delle colonne è uguale, l'altezza è proporzionale alla frequenza)

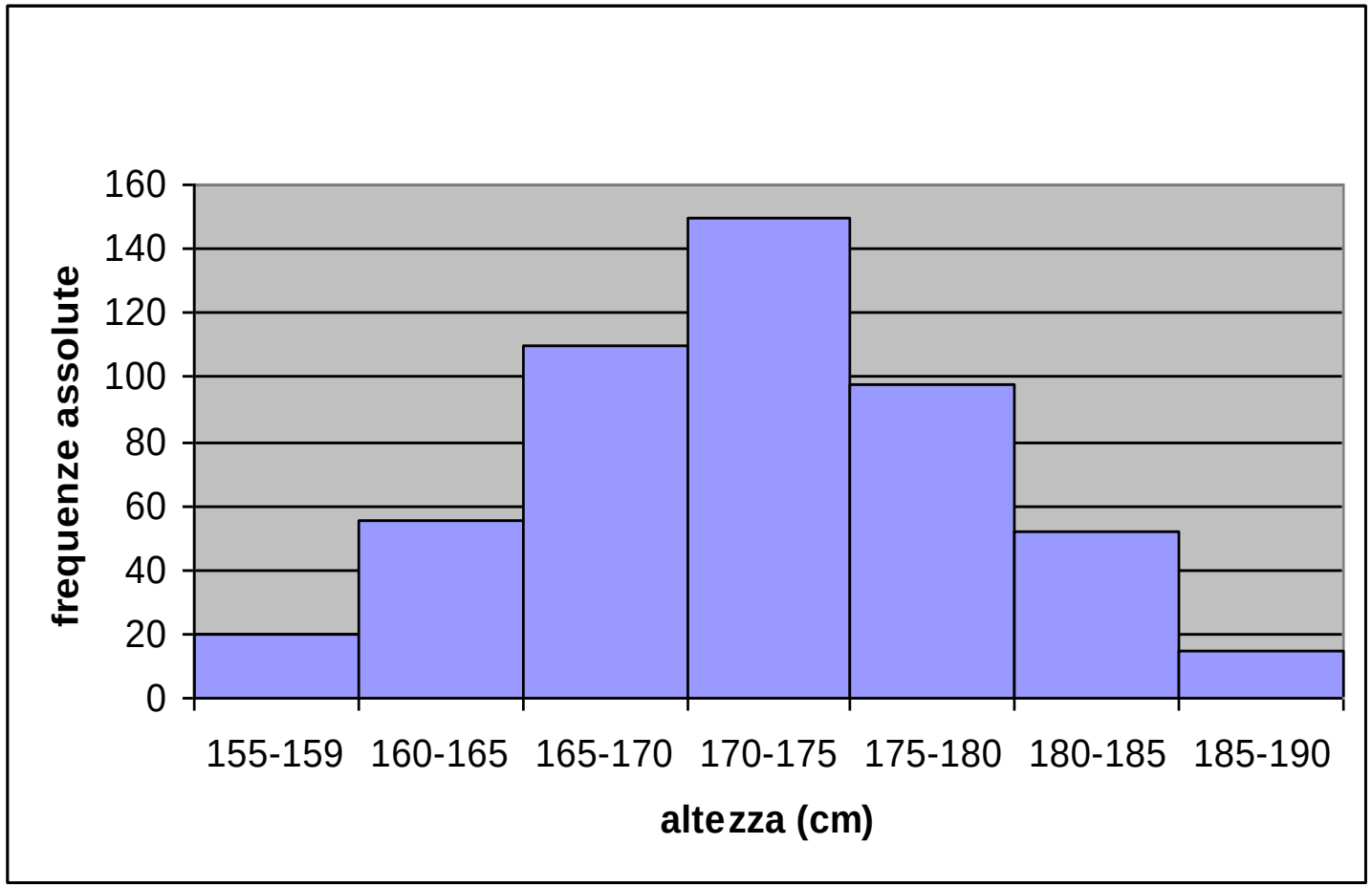
- DIAGRAMMA CIRCOLARE o AREOGRAMMA (ogni settore circolare è proporzionale alla frequenza)

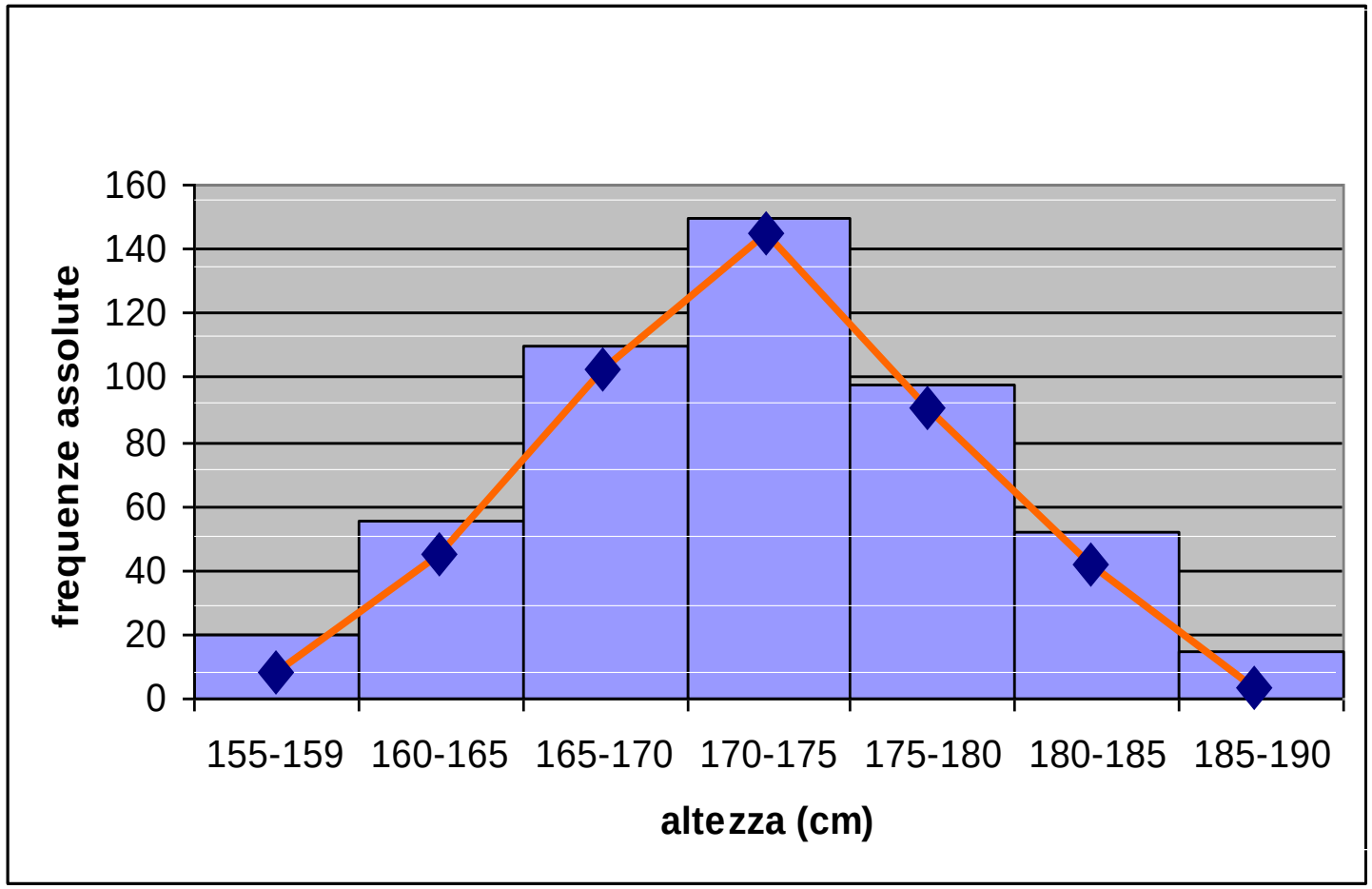

frequenza assoluta
140
120
100
80
60
40
20
0
Calcio
Danza
Ciclismo
Nuoto
Tennis
Pallavolo
sport praticato

VARIABILI QUANTITATIVE

- ISTOGRAMMA (colonne non distanziate di area proporzionale alla frequenza)

- POLIGONO di FREQUENZA (ogni punto ha coordinate date dal valore centrale di classe e dalla frequenza di classe)


160
140
120
100
80
60
40
20
0
frequenze assolute
155-159
160-165
165-170
170-175
175-180
180-185
185-190
altezza (cm)


160
140
120
100
80
60
40
20
0
frequenze assolute
155-159
160-165
165-170
170-175
175-180
180-185
185-190
altezza (cm)

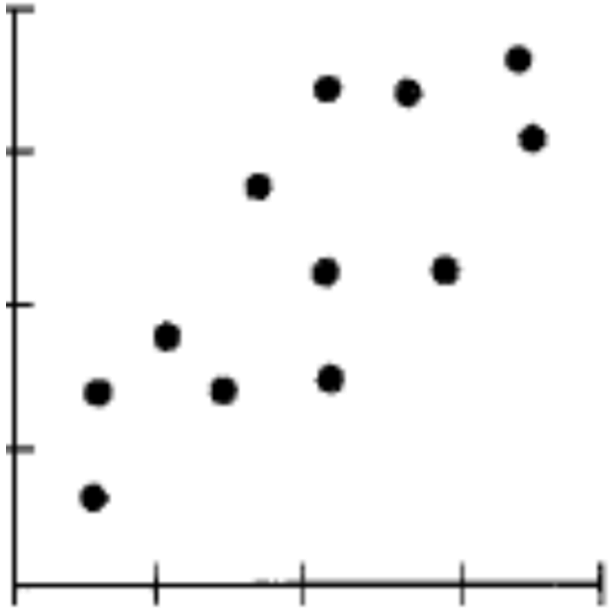
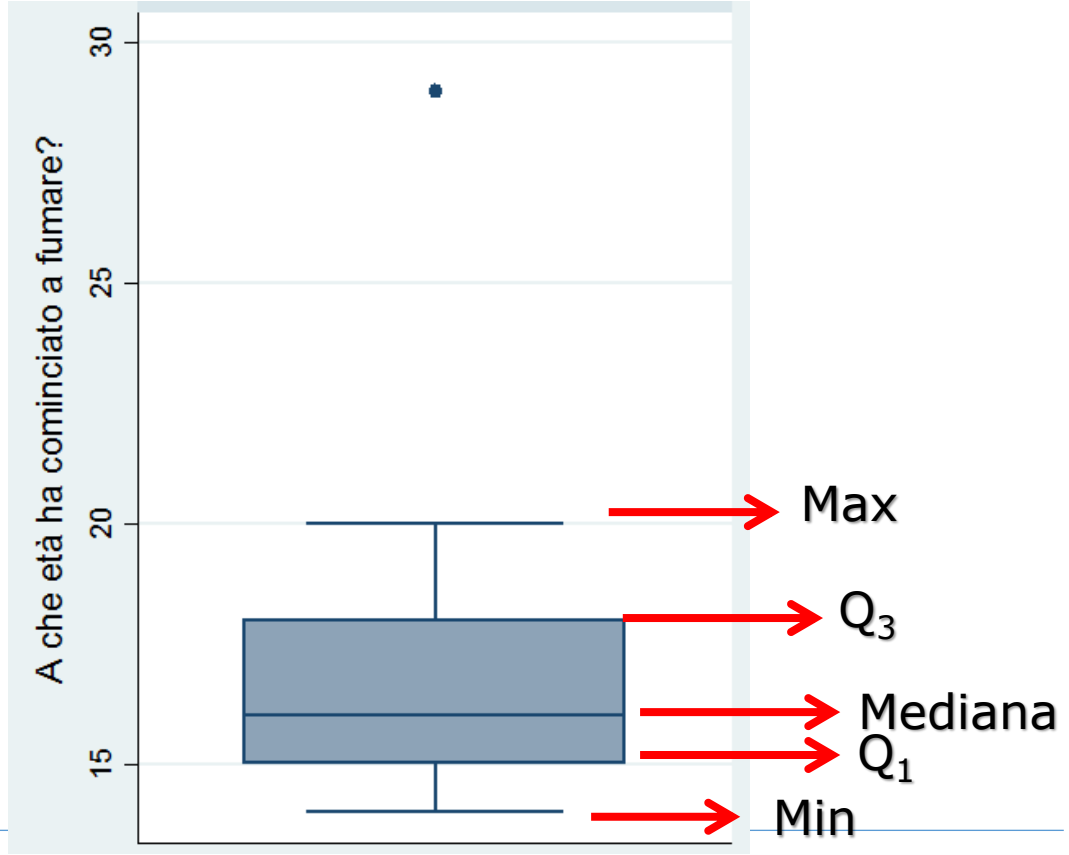


Diagramma a dispersione (x, y)
Relazione tra le variabili quantitative

Box plot o diagramma scatola e baffi
Rappresenta le misure di sintesi



Misure di sintesi sul campione

- Misure di POSIZIONE (media, moda, mediana, percentili)
- le osservazioni si raggruppano attorno a uno o più valori di forte densità

- Misure di DISPERSIONE (range, varianza, dev. standard e coeff. di variazione)
- le osservazioni sono più o meno distanti tra di loro, cioè hanno una certa variazione

Confronto tra media, mediana e moda

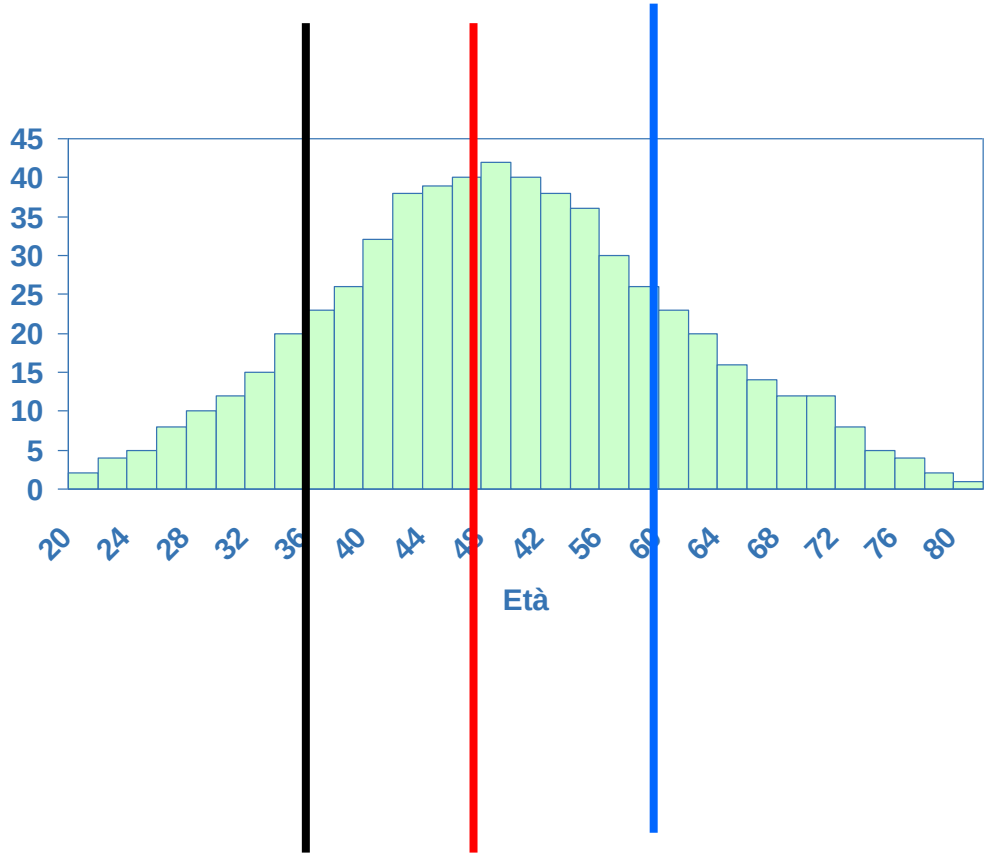
Misura	Definizione	Esistenza	Uso	Tiene conto di tutti i valori?	È condizionata dai valori estremi?
Media	SOMMA delle osservazioni di una variabile divisa per il numero totale di unità statistiche	sempre	Var. quantit.	si	si
Mediana	osservazione centrale che divide a metà la serie ordinata delle osservazioni	sempre	Var. quantit.	no	no
Moda	osservazione con frequenza più elevata	può non esistere, può non essere unica	Utile anche per dati nominali	no	no

Quartili

- **I quartile**: separa il 25% inferiore delle osservazioni dal 75% superiore delle osservazioni

- **II quartile**: coincide con la mediana, separa il 50%inferiore delle osservazioni dal 50% superiore delle osservazioni

- **III quartile**: separa il 75% inferiore delle osservazioni dal 25% superiore delle osservazioni



I quartile Mediana III quartile

25^ 50^ 75^

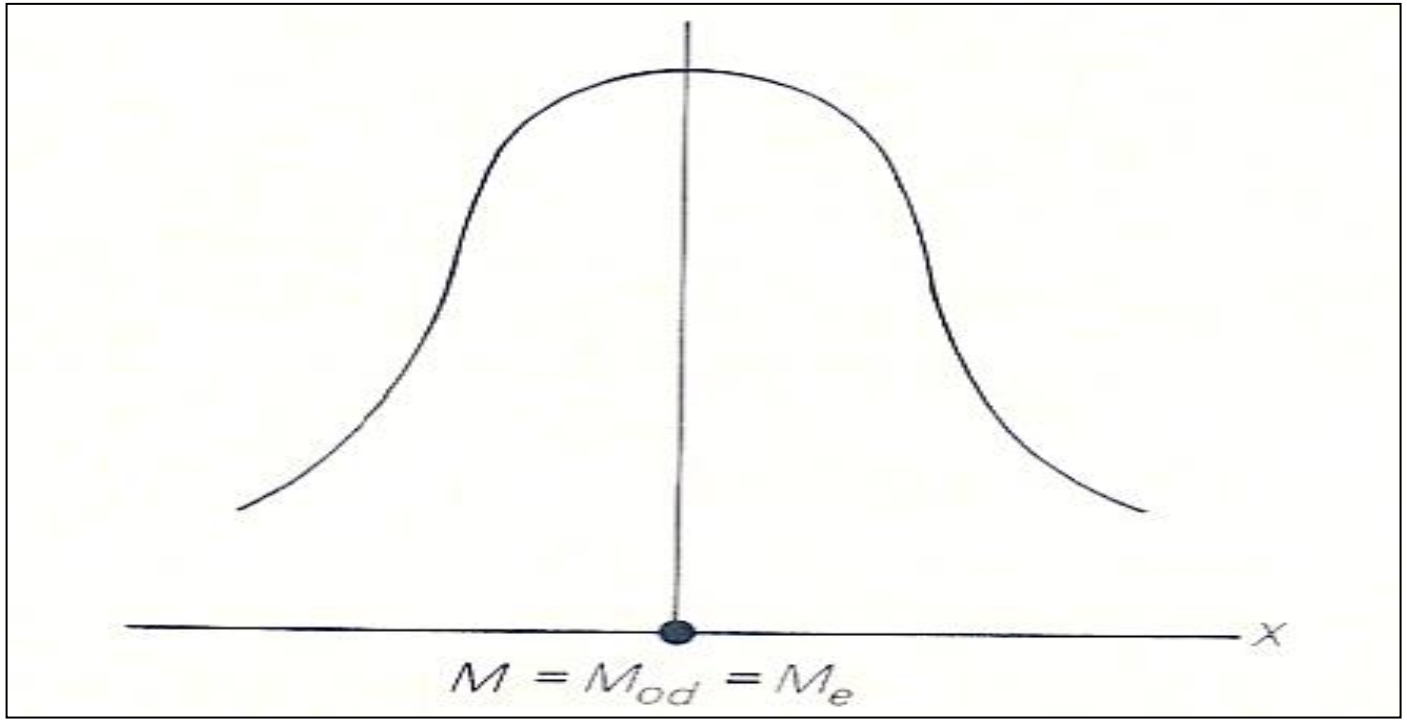
INTERVALLO INTERQUARTILE

IQR= è la differenza tra il 75^ e il 25^ percentile

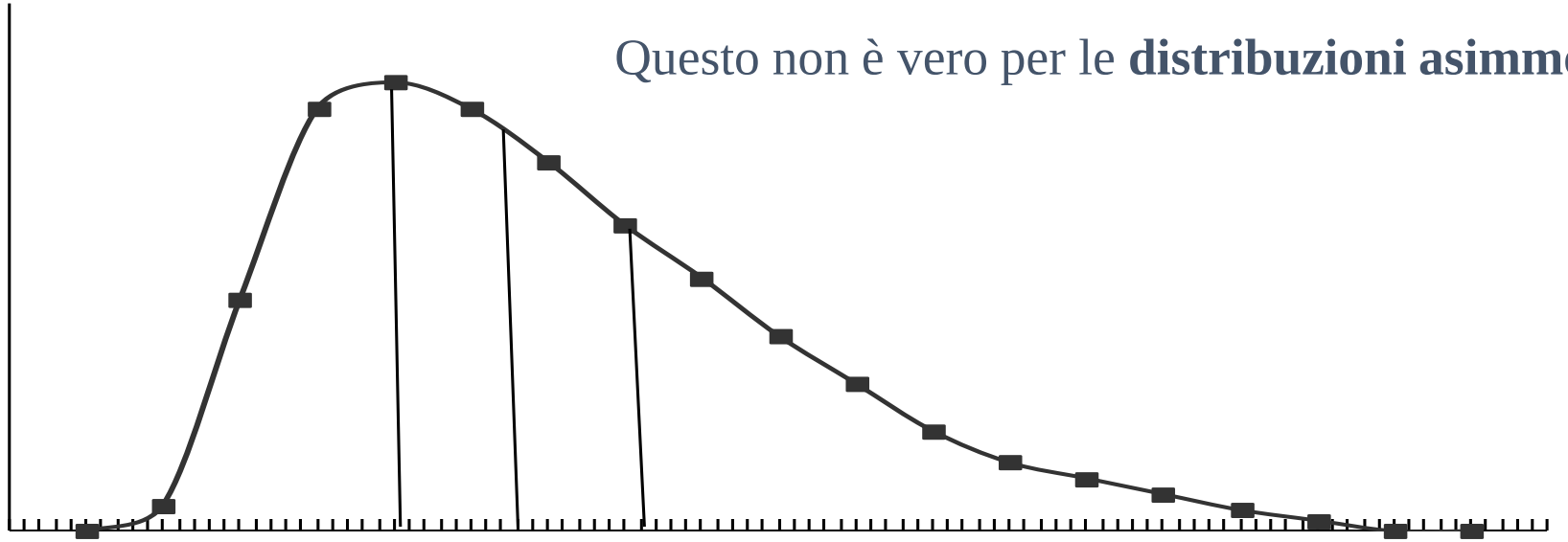
Misura di dispersione associata alla mediana

Relazione tra indici e distribuzione dei dati

In una distribuzione unimodale simmetrica
media, moda e mediana COINCIDONO



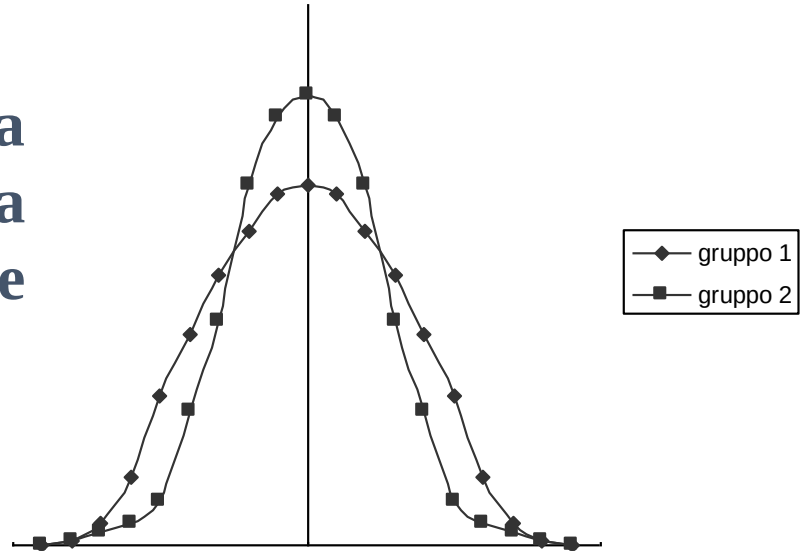
Questo non è vero per le **distribuzioni asimmetriche**



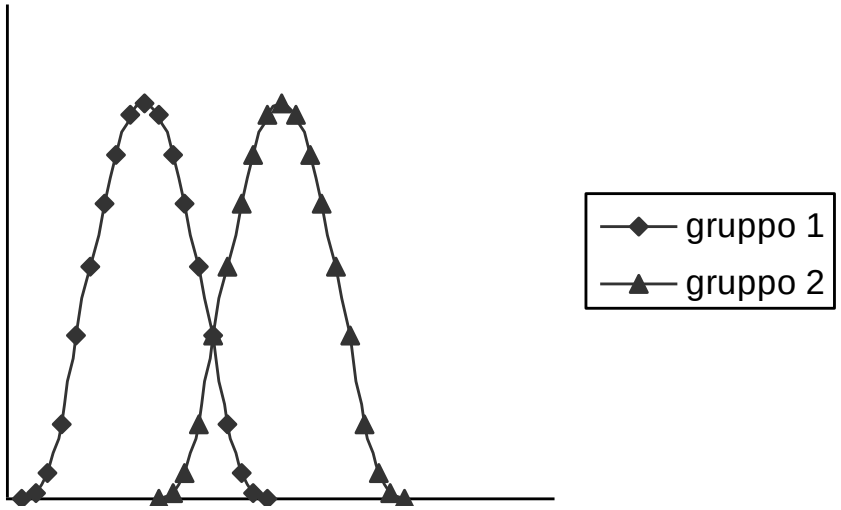
moda<mediana<media

- La distribuzione è asimmetrica a destra (asimmetria positiva).
- La mediana è spostata a destra rispetto alla moda e la media è ancora più a destra della mediana.
- Più è alto il grado di asimmetria e più media, moda e mediana saranno lontane l'una dall'altra.
- L'aggiunta di un dato è sufficiente a far variare di molto la media mentre la mediana e la moda rimangono pressoché inalterate.
- Per dati fortemente asimmetrici, la mediana tiene conto in maniera più accurata di dove si trova la maggior parte dei dati, a differenza della media.

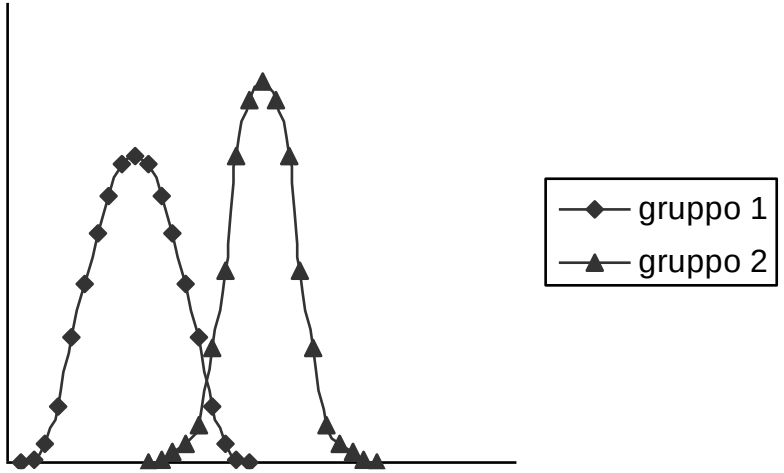
Per descrivere correttamente una distribuzione di frequenza (probabilità), alla misura di posizione bisogna sempre abbinare una misura di dispersione



Stessa media,
diversa deviazione standard



Media diversa,
stessa deviazione standard



Media diversa,
deviazione standard diversa

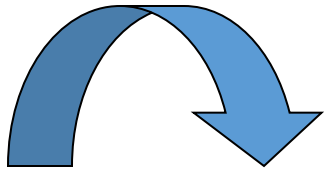
Quale misura di DISPERSIONE usare?

DEVIAZIONE STANDARD (*S*)⇒se si deve **descrivere** la **variabilità** di una variabile quantitativa

COEFF. di VARIAZIONE (*CV* o *CV%*)⇒se si deve **stabilire** una **differenza** nella **variabilità**

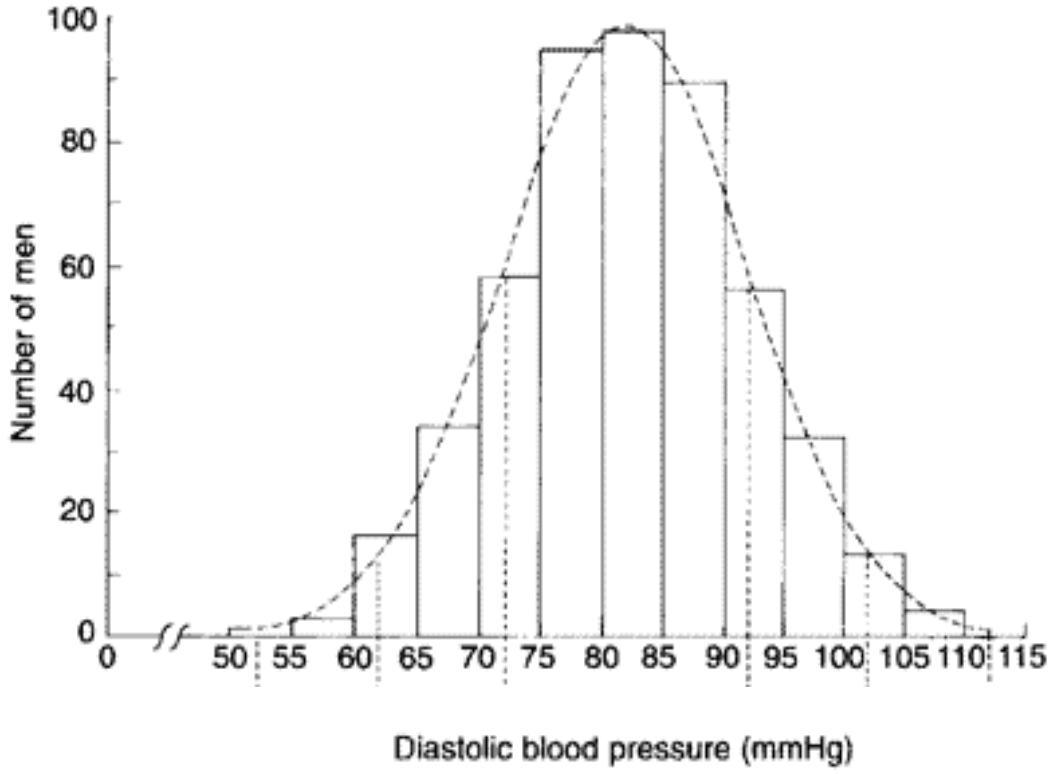
RANGE⇒se si vuole dare una descrizione in più

Se i dati provengono da una popolazione distribuita in modo NORMALE (o *Gaussiano*)



La Deviazione Standard fornisce un’utile base per interpretare i dati in termini di probabilità

LA DISTRIBUZIONE NORMALE (o distribuzione Gaussiana)

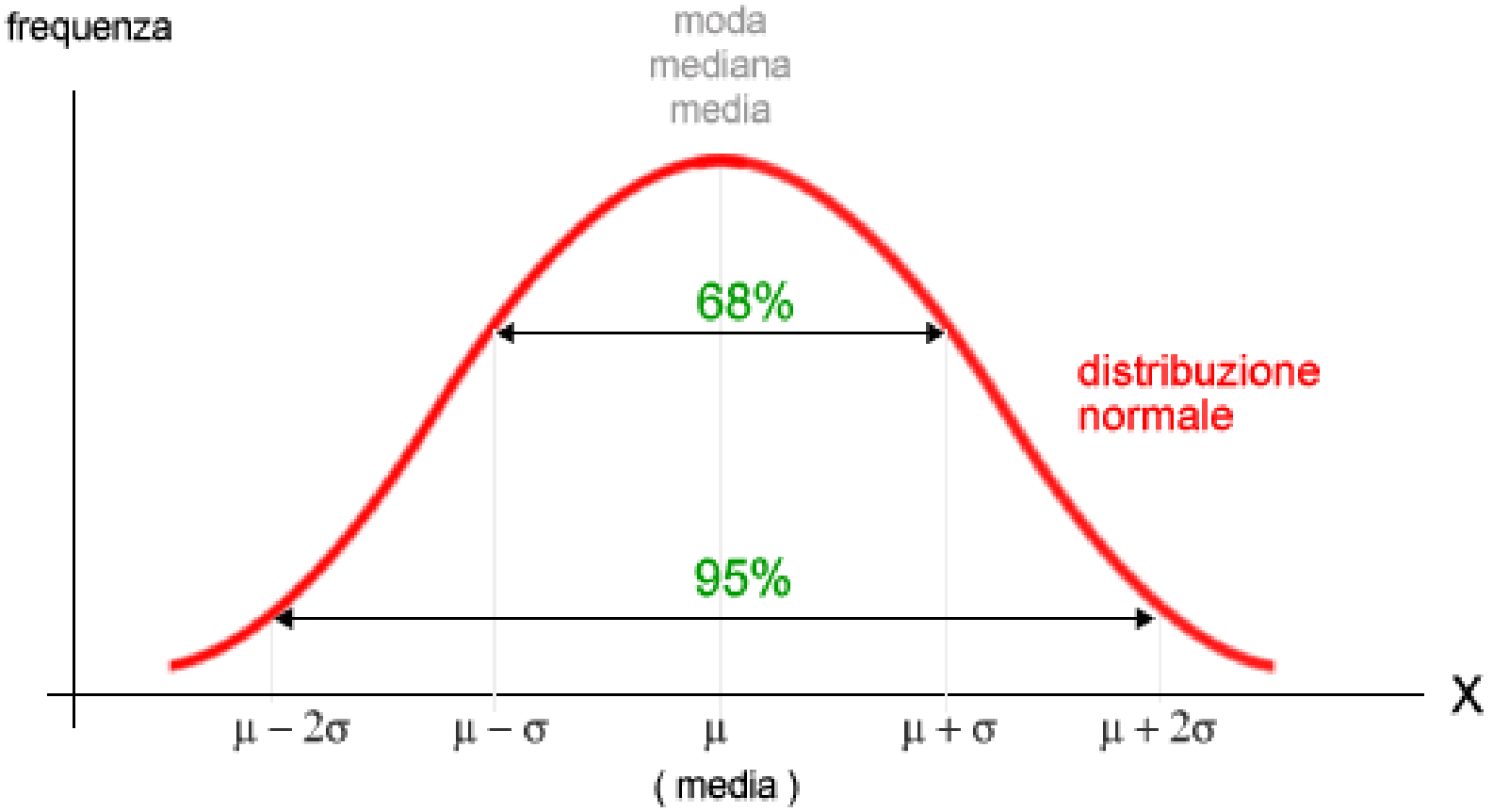


Distribuzione di valori di pressione diastolica di 500 uomini, media=82 mmHg, ds=10 mmHg

La maggior parte delle variabili biologiche seguono una distribuzione normale (ex: altezza di uomini e donne adulti, pressione di una popolazione di individui sani…)

Caratteristiche:

- Le misure di tendenza centrale coincidono
- E' simmetrica intorno alla media
- Famiglia di curve definite unicamente da 2 parametri: MEDIA (μ), DEVIAZIONE STANDARD (σ)
- La variabile aleatoria (X) con distribuzione normale assume valori compresi tra $-\infty$ e $+\infty$
- L'area sottesa dalla curva normale vale 1
- A destra e a sinistra della perpendicolare alzata dalla media (asse di simmetria) si trova il 50% dell'area
- Presenta una diminuzione dell'addensamento (frequenza) delle osservazioni man mano che ci si allontana dal valore medio


frequenza
moda
mediana
media
68%
distribuzione
normale
95%
X
μ – 2σ
μ – σ
μ
μ + σ
μ + 2σ
(media)

Regola Empirica

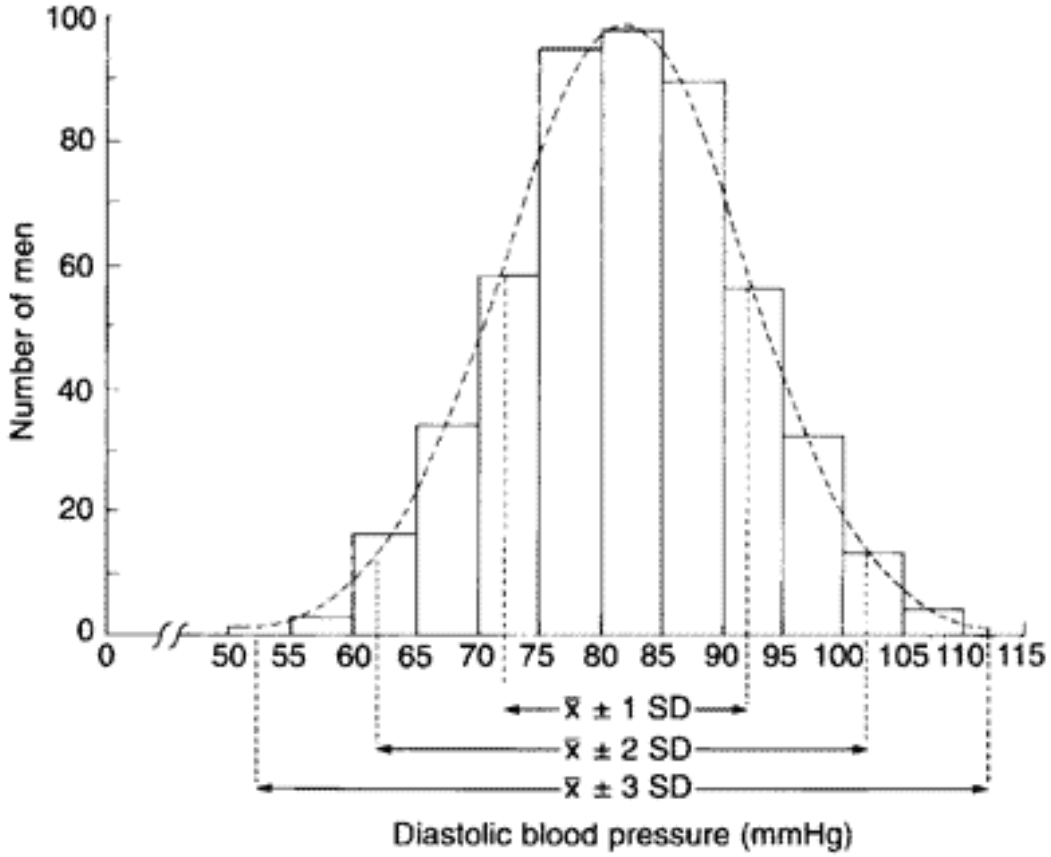
Se le osservazioni seguono una distribuzione normale:

$\bar{x} \pm$ 1 DS include il 68% delle osservazioni

$\bar{x} \pm$ 2 DS include il 95% delle osservazioni

$\bar{x} \pm$ 3 DS include il 99.7% delle osservazioni

Curva normale calcolata dai valori di pressione diastolica di 500 uomini, media=82 mmHg, ds=10 mmHg



il 68% degli uomini ha una pressione diastolica compresa tra 72 e 92 mmHg

il 95% degli uomini ha una pressione diastolica compresa tra 62 e 102 mmHg