STATISTICA MEDICA

Dott.ssa Marta Di Nicola

N.P.D. 3° Blocco 2° piano

0871-3554007

marta.dinicola@unich.it

http://www.biostatistica.unich.it

Cos'è la Statistica Medica-Biostatistica?

La biostatistica è la branca della statistica che enfatizza l'applicazione della statistica nelle scienze biomediche.

Esempi:

Dental researchers conducted a study to evaluate relevant variables that may assist in identifying orthodontic patients with signs and symptoms associated with sleep apnea and to estimate the proportion of potential sleep apnea patients whose ages range from 8 to 15 years.

In dental sciences, gingival recession represents a significant concern for patients and a therapeutic problem for clinicians. A clinical study was conducted to evaluate and compare the effects of a guided tissue regeneration procedure and connective tissue graft in the treatment of gingival recession defects.

Perché ho bisogno della statistica?

Per sviluppare l'abilità alla lettura critica della letteratura specifica.

Periodontal pathogens and gestational diabetes mellitus.

Dasanayake AP, Chhun N, Tanner AC, Craig RG, Lee MJ, Moore AF, Norman RG.

In previous cross-sectional or case-control studies, clinical periodontal disease has been associated with gestational diabetes mellitus. To test the hypothesis that, in comparison with women who do not develop gestational diabetes mellitus, those who do develop it will have had a greater exposure to clinical and other periodontal parameters, we measured clinical, bacteriological (in plaque and cervico-vaginal samples), immunological, and inflammatory mediator parameters 7 weeks before the diagnosis of gestational diabetes mellitus in 265 predominantly Hispanic (83%) women in New York. Twenty-two cases of gestational diabetes mellitus emerged from the cohort (8.3%). When the cases were compared with healthy control individuals, higher pre-pregnancy body mass index (p=0.004), vaginal levels of Tannerella forsythia (p=0.01), serum C-reactive protein (p=0.01), and prior gestational diabetes mellitus (p=0.006) emerged as risk factors, even though the clinical periodontal disease failed to reach statistical significance (50% in those with gestational diabetes mellitus vs. 37.3% in the healthy group; p=0.38).

J Dent Res. 2008 Apr;87(4):328-33.

La Biostatistica è uno strumento essenziale nella ricerca!

Consente di valutare l'effetto di un trattamento, comparare trattamenti, capire le interazioni tra essi, etc

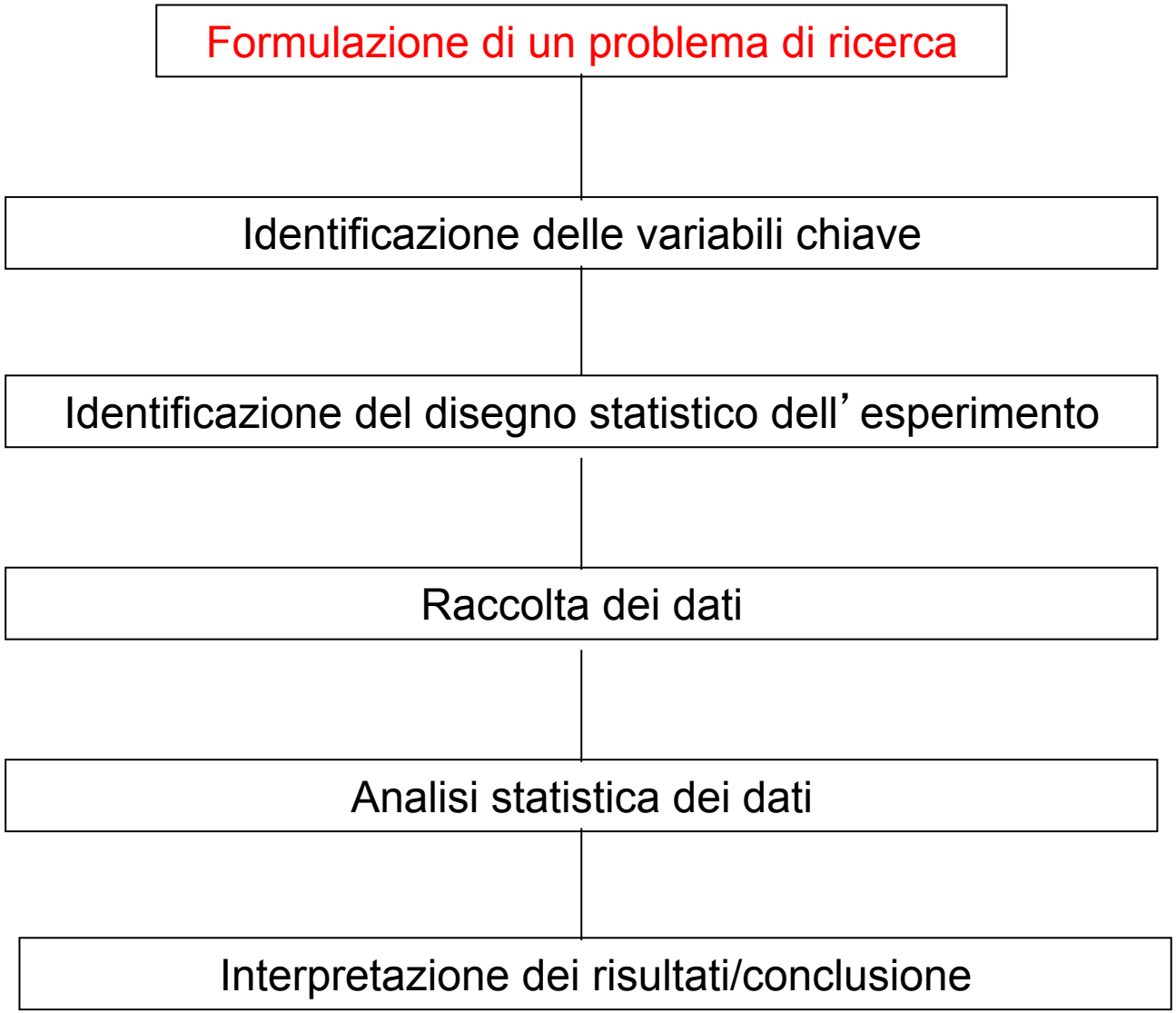
Di quanta matematica ho bisogno?

L'Uso della matematica sarà minimo!!

Come faccio a studiare la statistica?

Ricordando che i concetti e le definizioni presentate sono spesso le basi dei concetti che verranno presentati successivamente.

Una buona idea è allora rivedere frequentemente il materiale per ottenere un'analisi più approfondita e migliorare le tue conoscenze!


Formulazione di un problema di ricerca
Identificazione delle variabili chiave
Identificazione del disegno statistico dell'esperimento
Raccolta dei dati
Analisi statistica dei dati
Interpretazione dei risultati/conclusione

Formulazione di un problema di ricerca

Definire chiaramente la domanda cui vogliamo dare risposta e darne una adeguata giustificazione

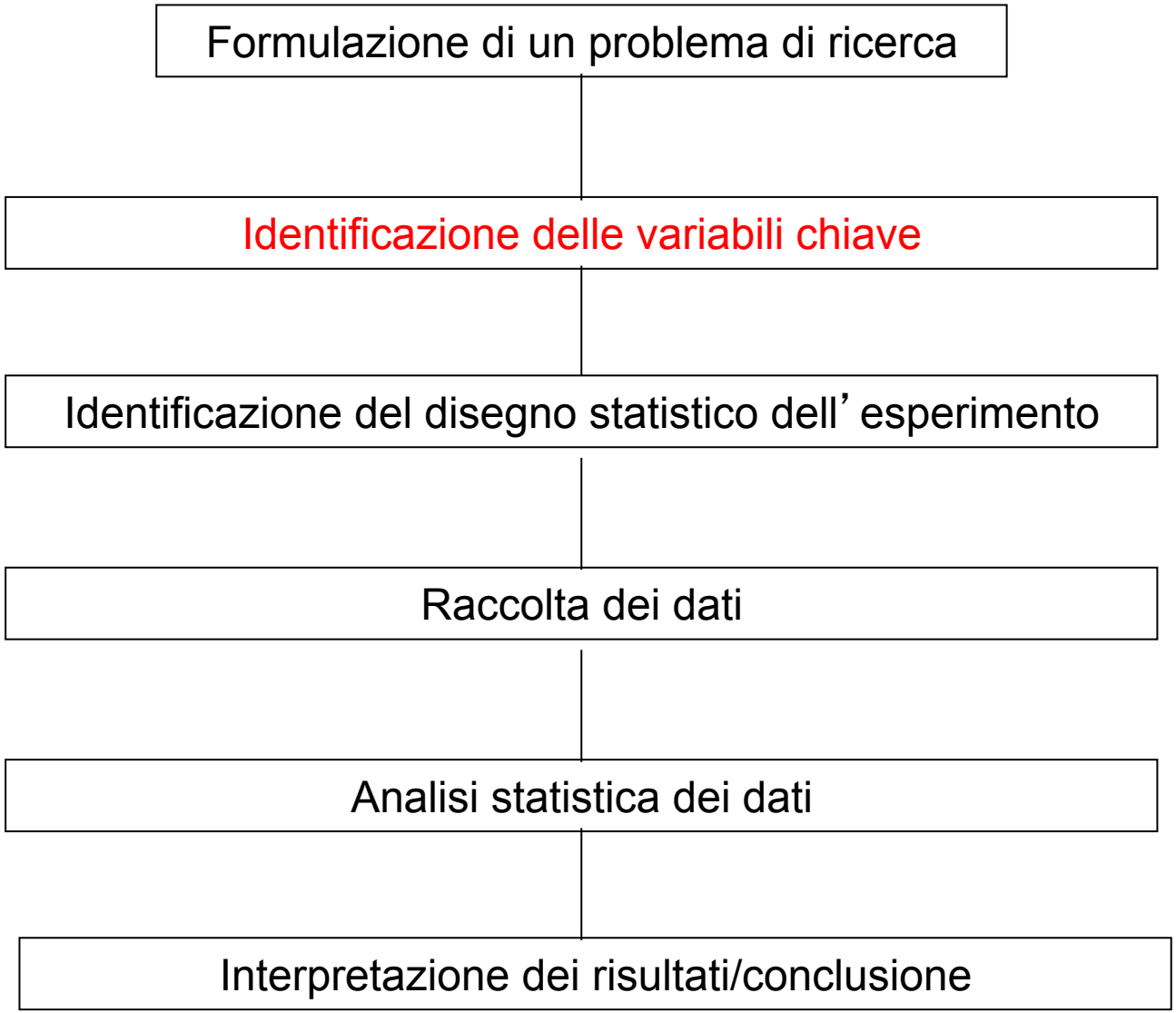
Definire la domanda di studio in modo che possa essere analizzabile in maniera corretta e non ambigua

- Trasformare concetti vaghi (es. fumare di meno, migliorare la prognosi) in definizioni operative che descrivono le misure che si possono fare e che saranno poi valutate
- Tradurre la domanda di studio in affermazioni relative ad attributi (ad es. la media) della popolazione

LA STATISTICA

- La Statistica ha come scopo la conoscenza quantitativa dei fenomeni collettivi. L'analisi statistica mira ad individuare modelli di interpretazione della realtà, attraverso canoni e tecniche che sono astrazioni, semplificazioni di una moltitudine di aspetti e di manifestazioni del reale.

- E' costituita da un insieme dei metodi che consentono di raccogliere, ordinare, riassumere, presentare ed analizzare dati e informazioni, trarne valide conclusioni e prendere decisioni sulla base di tali analisi e risultati.


Formulazione di un problema di ricerca
Identificazione delle variabili chiave
Identificazione del disegno statistico dell'esperimento
Raccolta dei dati
Analisi statistica dei dati
Interpretazione dei risultati/conclusione

GLOSSARIO

POPOLAZIONE: L' insieme di tutti gli elementi (unità statistiche) oggetto dell' osservazione che hanno una o più caratteristiche comuni. Gli elementi possono essere soggetti, oggetti o eventi.

Esistono popolazioni FINITE e popolazioni INFINITE.

- *Tutti i bambini italiani con problemi di carie nella prima infanzia*
- *Tutti gli studenti iscritti ad un corso di odontoiatria nel 2009*
- *Tutte le compresse di vitamine di un lotto di produzione*

CAMPIONE: La parte delle unità statistiche sottoposte all' osservazione.

CARATTERE (O VARIABILE): Ogni caratteristica di un unità statistica che può essere misurata.

Sex	Status of Oral Hygiene	Level of Post-Surgery Pain
F = female	P = poor	N = no pain
M = male	F = fair	M = mild pain
	G = good	S = severe pain
		E = extremely severe pain

Subject No.	Age (yrs.)	BP (mm Hg)	Pocket Depth (mm)	Cholesterol (mg/dl)
1	56	121/76	6.0	167
2	43	142/95	5.5	180
—	—	—	—	—
—	—	—	—	—
115	68	175/124	6.5	243

Note: BP, blood pressure.

Se una variabile può assumere più un di valore con determinate probabilità questa è detta RANDOM o ALEATORIA

MODALITA' : Ogni diversa presentazione del carattere o variabile osservata su ciascuna unità statistica.

Esempi di variabili statistiche

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

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

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

VARIABILI QUALITATIVE

VARIABILI QUANTITATIVE

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

Sesso; professione; diagnosi medica; …

ORDINALI O PER RANGHI
Esiste un criterio predeterminato per ordinare le modalità

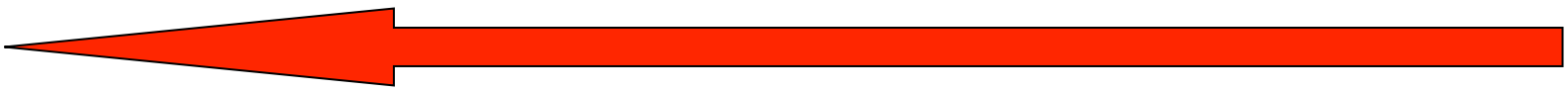
ordine di nascita; giorni della settimana; indice di severità di una malattia;…

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

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

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

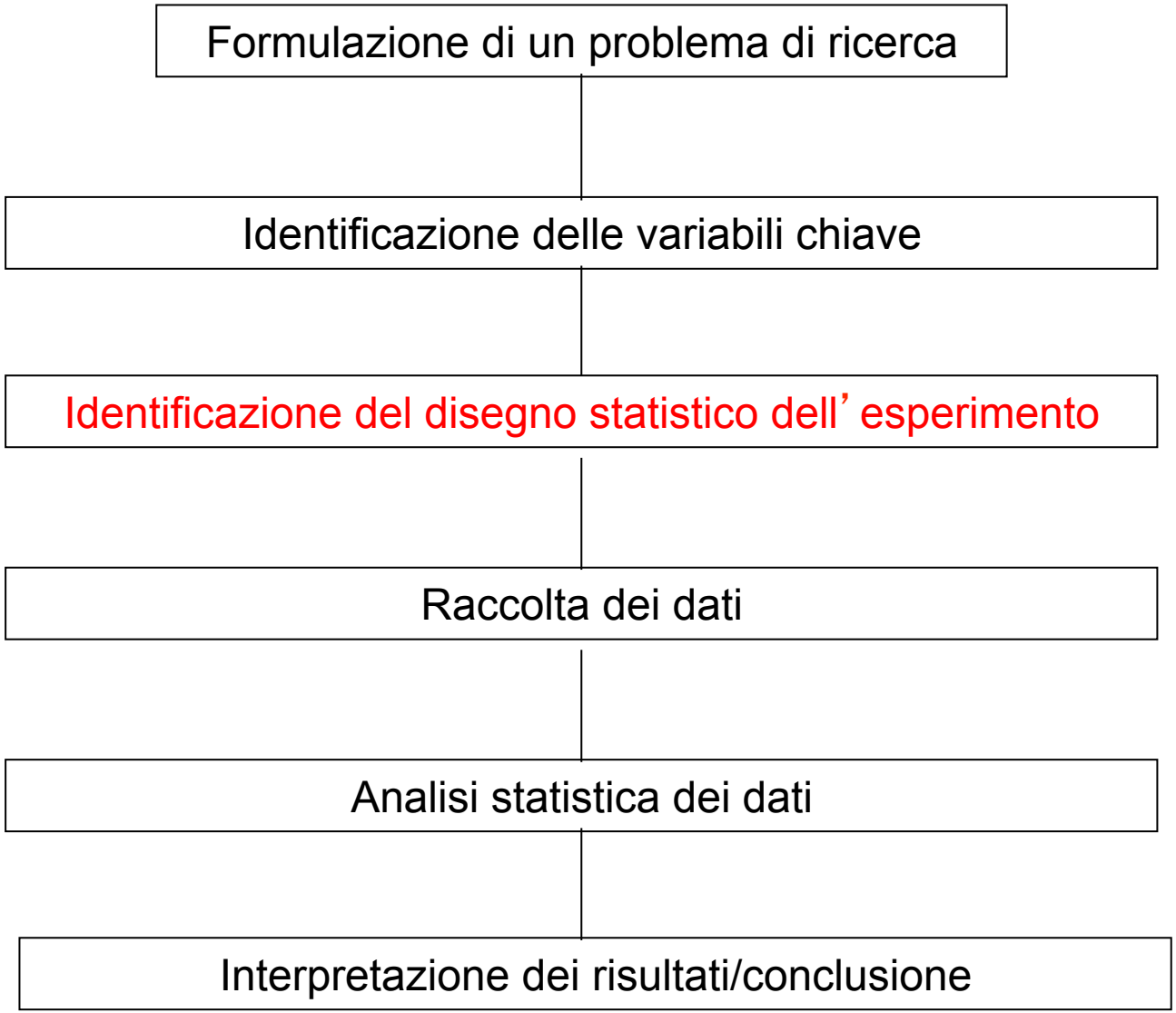
statura; peso; glicemia; PAS;…

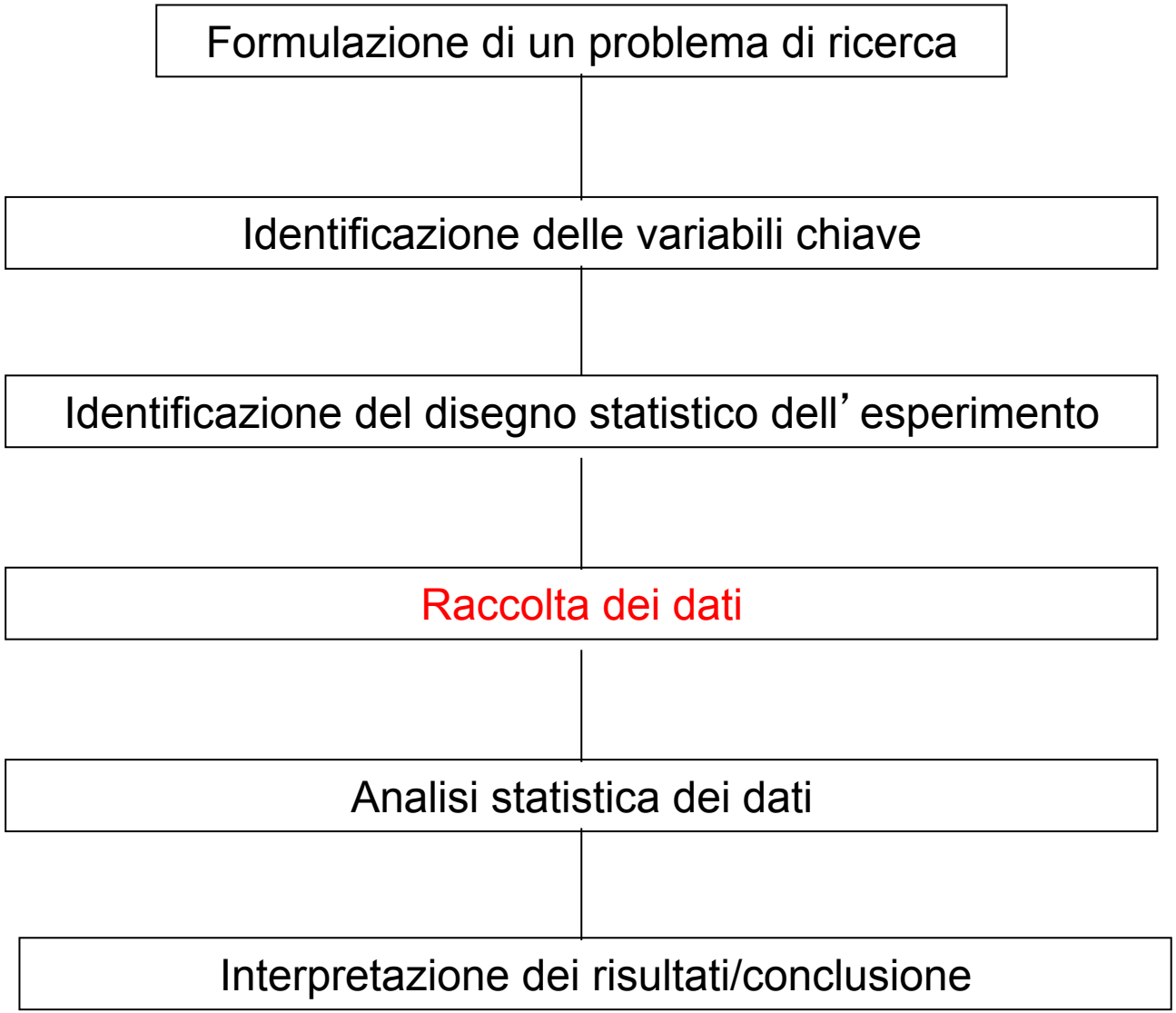


Descrivere e Riassumere i dati

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

nome: Rossi Amerigo	nome: Bianchi Paolo	nome: Valenzi Alberica	nome: Alinori Alfonso
sesso: maschio	sesso: maschio	sesso: femmina	sesso: maschio
età: 32	età: 47	età: 45	età: 27
altezza: 172 cm	altezza: 170 cm	altezza: 168 cm	altezza: 183 cm
peso: 64 Kg	peso: 80 Kg	peso: 51 Kg	peso: 85 Kg
PAS: 140 mm/Hg	PAS: 148 mm/Hg	PAS: 125 mm/Hg	PAS: 138 mm/Hg
glicemia: 190 mg/ 100cc	glicemia: 180 mg/ 100cc	glicemia: 150 mg/ 100cc	glicemia: 170 mg/ 100cc


Formulazione di un problema di ricerca
Identificazione delle variabili chiave
Identificazione del disegno statistico dell'esperimento
Raccolta dei dati
Analisi statistica dei dati
Interpretazione dei risultati/conclusione


Formulazione di un problema di ricerca
Identificazione delle variabili chiave
Identificazione del disegno statistico dell'esperimento
Raccolta dei dati
Analisi statistica dei dati
Interpretazione dei risultati/conclusione

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

Data Base o File Dati.

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

N.	NOME	SESSO	ETA'	ALTEZZA	PESO	PAS	GLIC.
1	Rossi Amerigo	M	32	172	64	140	190
2	Bianchi Paolo	M	47	170	80	148	180
3	ValenziAlberica	F	45	168	51	125	150
4	Alinori Alfonso	M	27	183	85	130	170
5	…	…	…	…	…	…	…
6	…	…	…	…	…	…	…

DISTRIBUZIONI SEMPLICI DI FREQUENZE

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

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

Variabile nominale "Tipologia di corona"

Type of Crown	Number of Crowns
Gold crown	843
Metal ceramic crown	972

Frequenze assolute

Variabile ordinale "Grado di accordo"

Response Category	Number of Individuals
Strongly disagree	24
Disagree	43
Neutral	49
Agree	33
Strongly agree	30

Frequenze assolute

Variabile quantitativa “conta delle colonie batteriche”

120	160	172	172	176	180	184	184	184	185
188	190	192	196	196	200	200	200	206	207
210	213	220	236	250	254	272	272	272	275
275	280	280	282	284	286	287	295	296	299
300	300	300	301	302	304	304	304	304	304
306	306	308	315	320	320	325	330	335	346
346	354	356	358	360	364	365	366	380	380
385	386	390	396	396	396	400	408	410	412
412	416	418	424	438	440	448	476	500	588

Il numero degli intervalli potrebbe essere pari a $\sqrt{n}$

Interval	Frequency	Interval	Frequency
112.5–162.5	2	362.5–412.5	16
162.5–212.5	19	412.5–462.5	6
212.5–262.5	5	462.5–512.5	2
262.5–312.5	27	512.5–562.5	0
312.5–362.5	12	562.5–612.5	1

Frequenze assolute delle classi

Esempio. Distribuzione doppia di frequenze assolute

Carie	Fumatori	Non fumatori	Totale
SI	160	100	260
NO	120	70	190
Totale	280	170	450

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

Se vogliamo confrontare le frequenze le dobbiamo "depurare" dalla numerosità del collettivo; ciò lo si fa dividendo le f.a. per la numerosità (N) della popolazione e moltiplicando per 100

(cioè facendo riferimento ad una ipotetica popolazione di 100 unità).

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

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

Tornando all' esempio precedente

$$\left(\frac{19}{90}\right)\cdot 100 = 21\%$$

Interval	Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
112.5–162.5	2	2	2
162.5–212.5	19	21	23
212.5–262.5	5	6	29
262.5–312.5	27	30	59
312.5–362.5	12	13	72
362.5–412.5	16	18	90
412.5–462.5	6	7	97
462.5–512.5	2	2	99
512.5–562.5	0	0	99
562.5–612.5	1	1	100
Total	90	100	

29=23+6

Le frequenze cumulate indicano quante u.s. si presentano fino a quella modalità. Ha senso calcolare le f.cum solamente per le variabili quantitative o qualitative ordinabili.

I GRAFICI STATISTICI

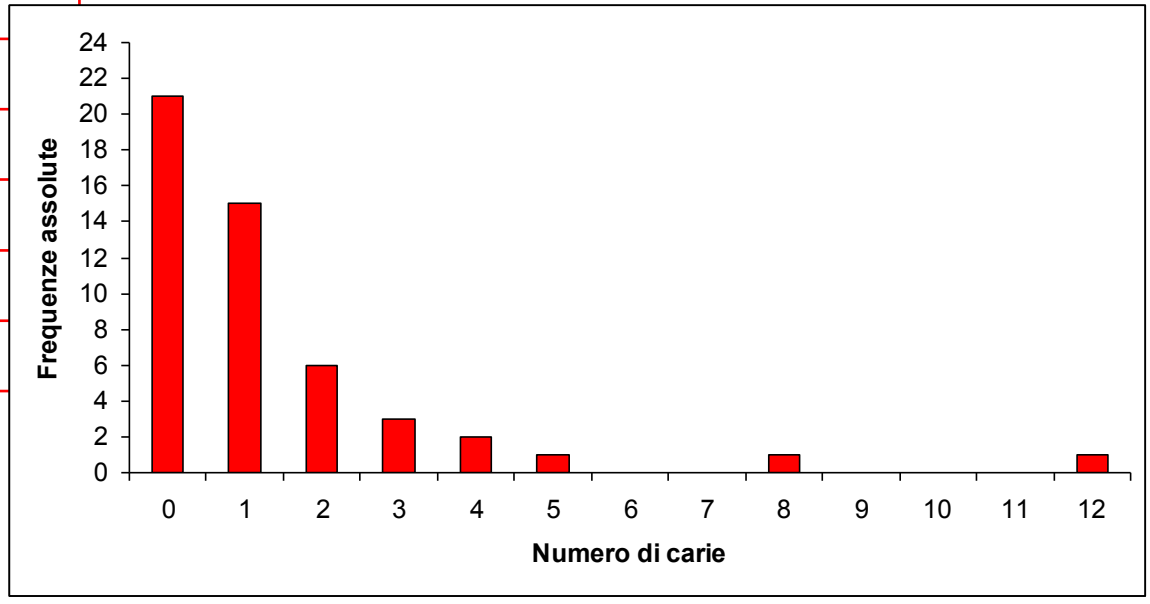
Scopo dei grafici è quello di rendere l'informazione contenuta in una serie di dati:

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

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

ORTOGRAMMA: Usato per variabili qualitative l' altezza delle barre rappresenta frequenze assoluta o percentuale

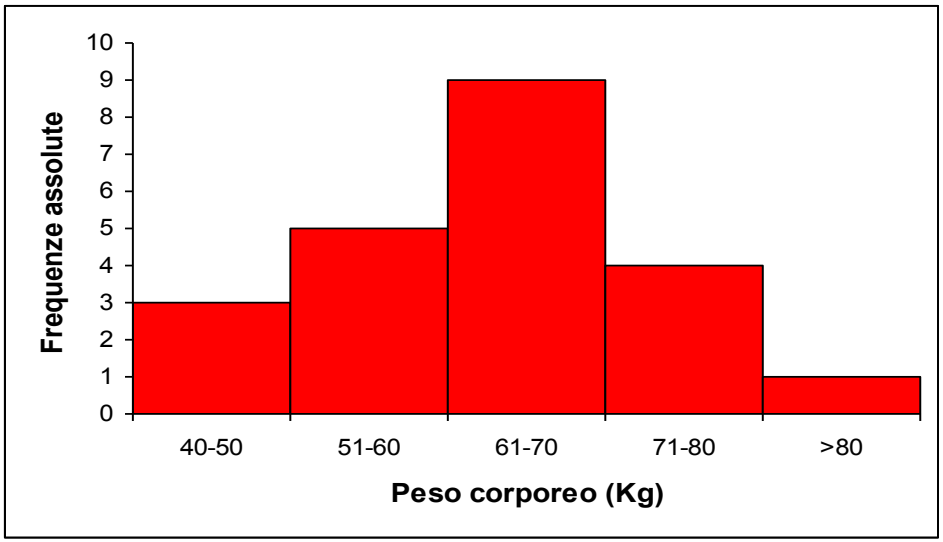
NUMERO DI CARIE	Frequenza assoluta	Frequenza cumulata
0	21	21
1	15	36
2	6	42
3	3	45
4	2	47
5	1	48
8	1	49
12	1	50
Totale	50	



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

PESO CORPOREO	Freq. assoluta	Freq. relativa	Freq. cumulata
40-50	3	0.14	0.14
51-60	5	0.23	0.36
61-70	9	0.41	0.77
71-80	4	0.18	0.95
> 80	1	0.05	1
Totale	22	1	

ISTOGRAMMA Indicato per rappresentare distribuzioni in classi (variabili quantitative continue). Costituiti da una serie di barre rettangolari contigue ognuna in rappresentanza di una classe e con area proporzionata alla rispettiva frequenza.



DIAGRAMMI A SETTORI CIRCOLARI (TORTE) Indicati per variabili qualitative allo scopo di evidenziare le frequenze % delle singole modalità. L'area di un cerchio viene suddivisa in settori proporzionali alle frequenze %

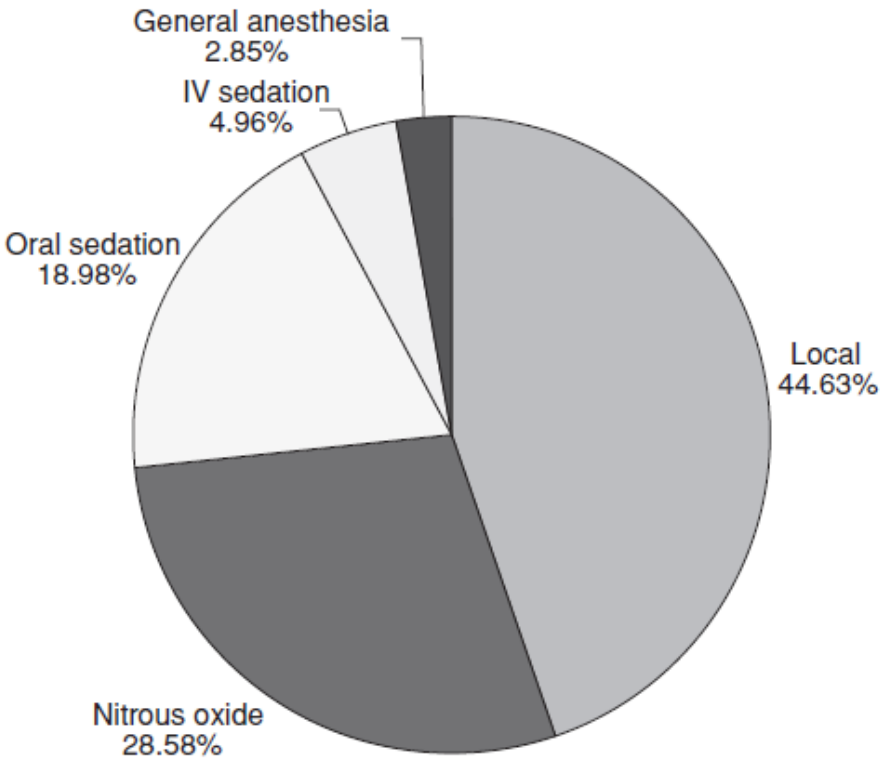


Figure 2.4.3 Type of anesthesia used in dental offices.

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

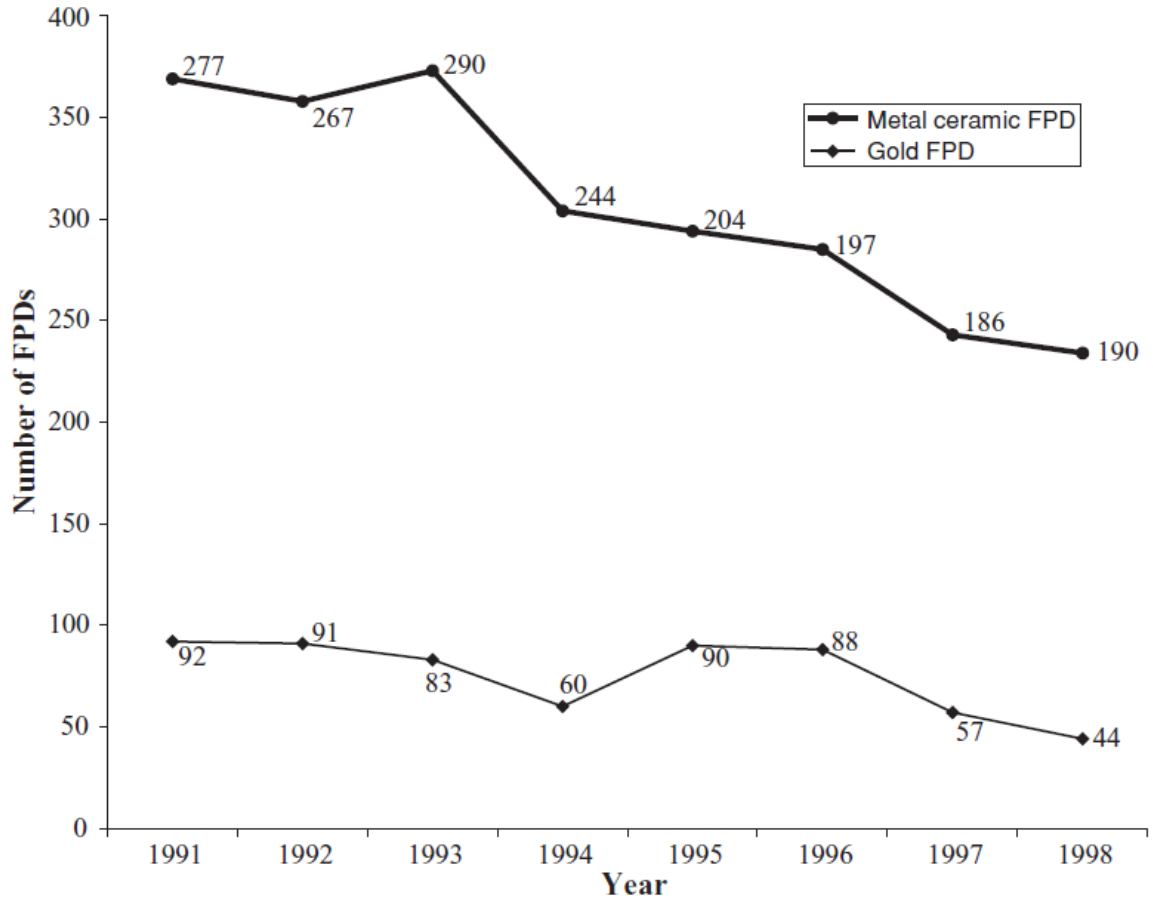
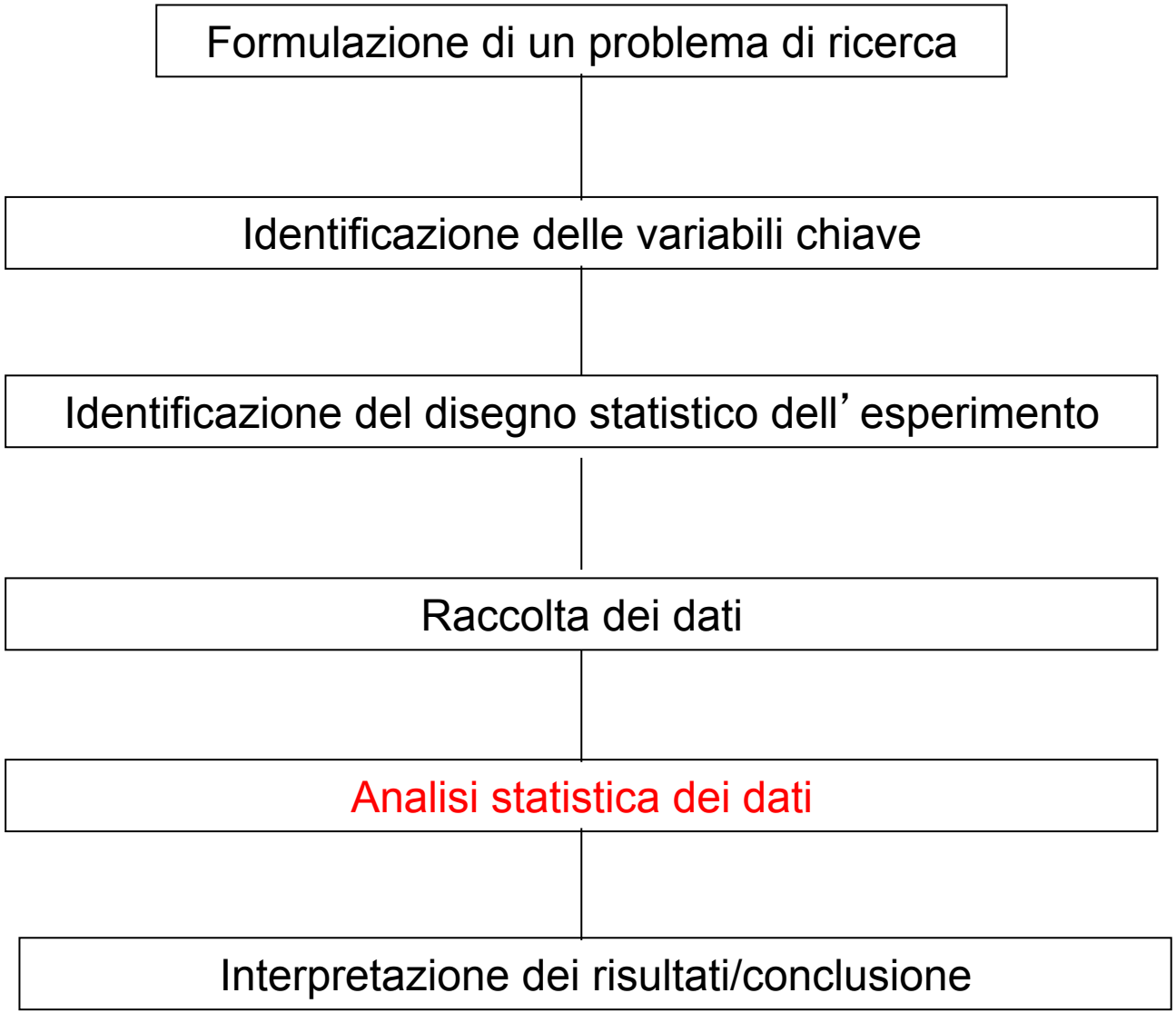


Figure 2.4.5 Gold and metal ceramic fixed partial dentures.


Formulazione di un problema di ricerca
Identificazione delle variabili chiave
Identificazione del disegno statistico dell'esperimento
Raccolta dei dati
Analisi statistica dei dati
Interpretazione dei risultati/conclusione

OBIETTIVO: Individuare un indice che rappresenti significativamente un insieme di dati statistici.

Altezza degli studenti 2004-05 (cm)

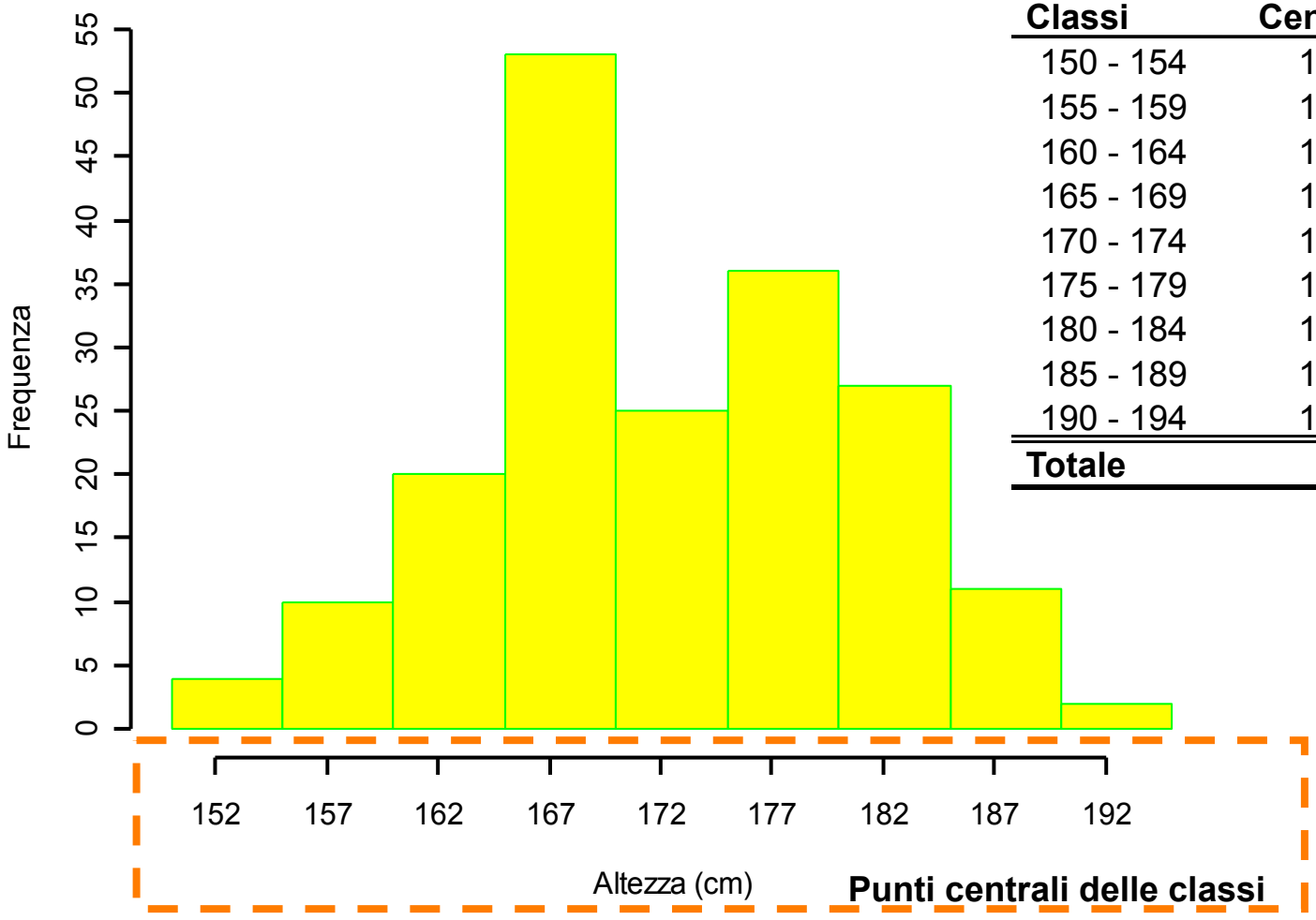


Tabella dei dati

Classi	Punto Centrale	n	%
150 - 154	152	4	2.1
155 - 159	157	10	5.3
160 - 164	162	20	10.6
165 - 169	167	53	28.2
170 - 174	172	25	13.3
175 - 179	177	36	19.1
180 - 184	182	27	14.4
185 - 189	187	11	5.9
190 - 194	192	2	1.1
Totale		**188**	**100.0**

LA MEDIA ARITMETICA

DEFINIZIONE: La media aritmetica è quel valore che avrebbero tutte le osservazioni se non ci fosse la variabilità (casuale o sistematica).

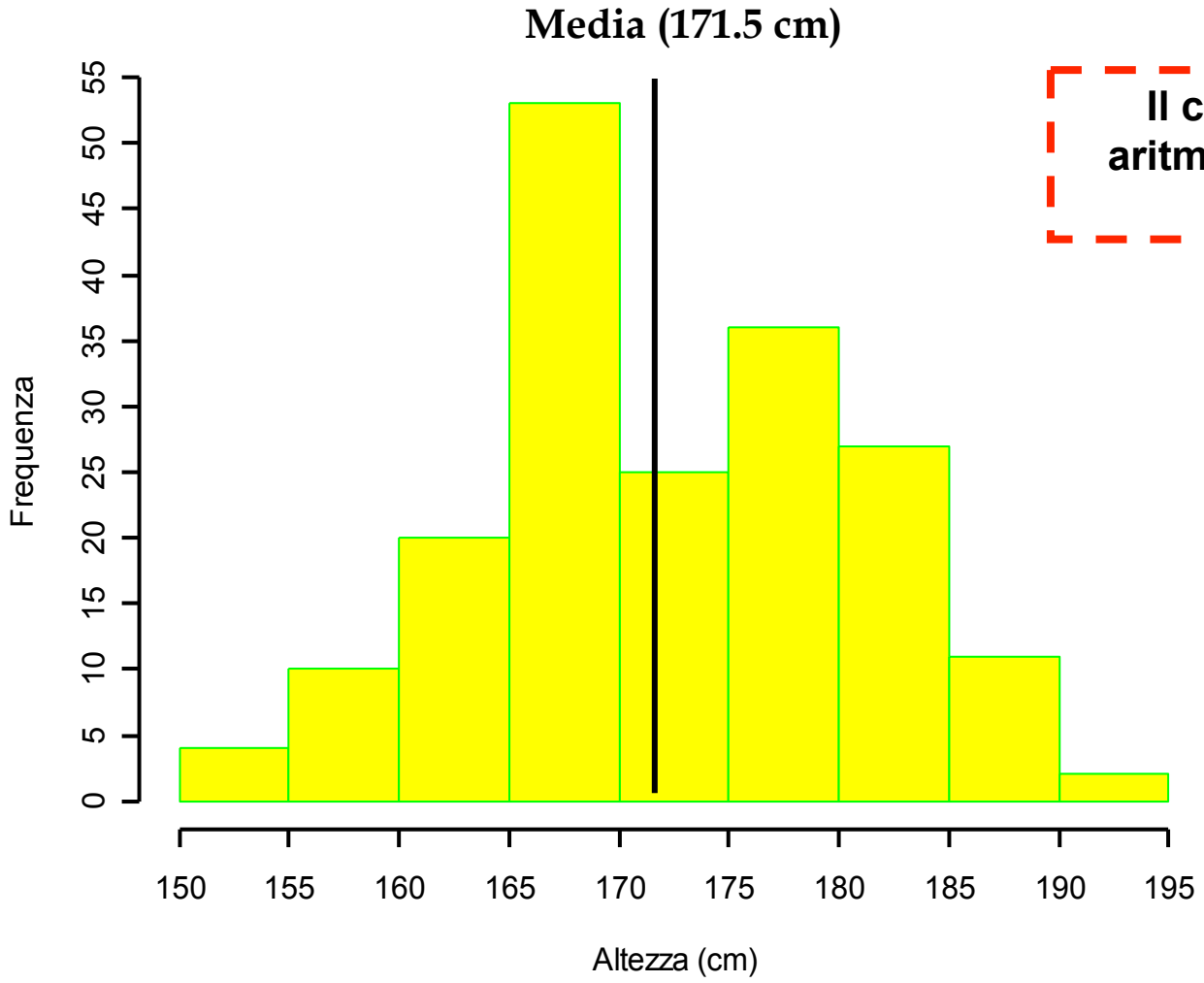
Più precisamente, è quel valore che sostituito a ciascun degli n dati ne fa rimanere costante la somma.

dato un insieme di n elementi {x1, x2, ... xn}

Si dice media aritmetica semplice di n numeri il numero che si ottiene dividendo la loro somma per n.

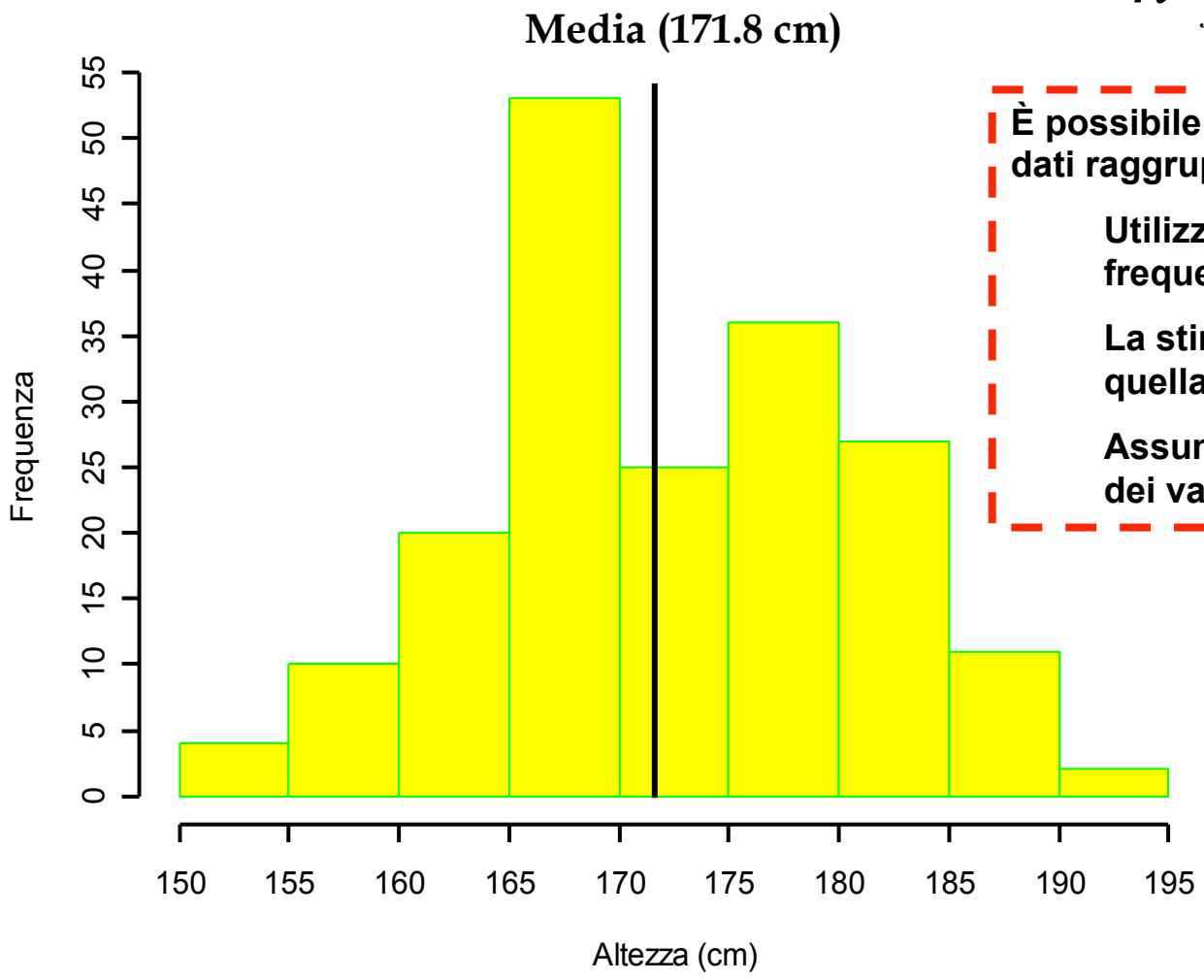
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

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



Il calcolo della media aritmetica utilizza tutte le osservazioni

$$\mu = \frac{1}{n} \sum_{j=1}^{k} f_j \cdot x_{(mid)\,j}$$



È possibile calcolare la media anche dai dati raggruppati in classi:

- **Utilizzo i punti centrali e le frequenze di classe**
- **La stima della media approssima quella ottenuta con i singoli dati**
- **Assume la distribuzione uniforme dei valori all'interno delle classi**

Le proprietà della media aritmetica

1. Compresa tra il minimo dei dati e il massimo dei dati;
2. $\sum_i (x_i - \bar{x}) f_i = 0$ la somma degli scarti dalla media è sempre zero;
3. $\sum_i (x_i - z)^2 f_i$ assume valore minimo per z = media aritmetica;
4. la media dei valori: $k \cdot x_i$ è pari alla media aritmetica · k (dove k è un numero reale qualsiasi);
5. la media dei valori: $x_i \pm h$ è pari a: media aritmetica ±h (dove h è un numero reale qualsiasi).

Limite della media aritmetica

La media aritmetica è la misura di posizione più usata ma. A volte, altre misure come la mediana e la moda si dimostrano utili.

Si consideri un campione di valori di VES (velocità di eritrosedimentazione, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

In questo caso, la media che è = 10 mm/ora non è un valore tipico della distribuzione: soltanto un valore su 7 è superiore alla media!

Il calcolo della media aritmetica utilizzando tutte le osservazioni risulta sensibile ai valori atipici/estremi della distribuzione

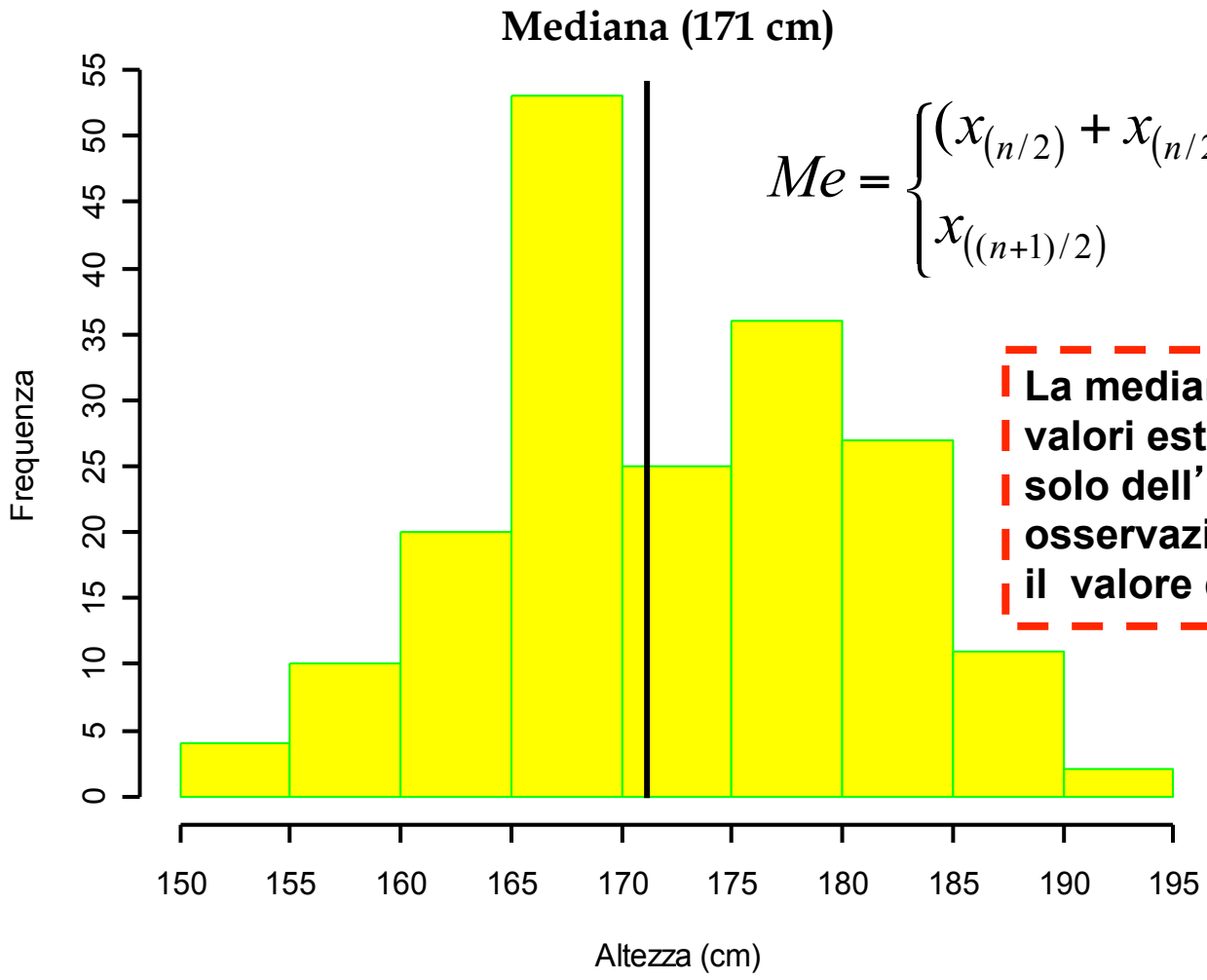
LA MEDIANA

DEFINIZIONE: La mediana (Me) è quell'osservazione che bipartisce la distribuzione in modo tale da lasciare al "di sotto" lo stesso numero di termini che lascia al "di sopra".

L'idea che e alla base della mediana e di cercare un numero che sia più grande di un 50% delle osservazioni e più piccolo del restante 50%.

per dati almeno ordinali

La mediana è quel valore che divide in due la distribuzione ordinata di tutti i valori



$$Me = \begin{cases} (x_{(n/2)} + x_{(n/2+1)})/2 & \textit{se n pari} \\ x_{((n+1)/2)} & \textit{se n dispari} \end{cases}$$

La mediana si dice *resistente* ai valori estremi, perché tiene conto solo dell'ordinamento delle osservazioni, considerando solo il valore centrale

Gingival Index (x_i)	Frequenze assolute	Frequenze Cumulate	Freq.Cum. %
0	4 (22.2)	4	22.2
1	8 (44.4)	4+8 = 12	66.6
2	4 (22.2)	12+4 = 16	88.8
3	2 (11.2)	16+2 = 18	100
Totale	**18**		

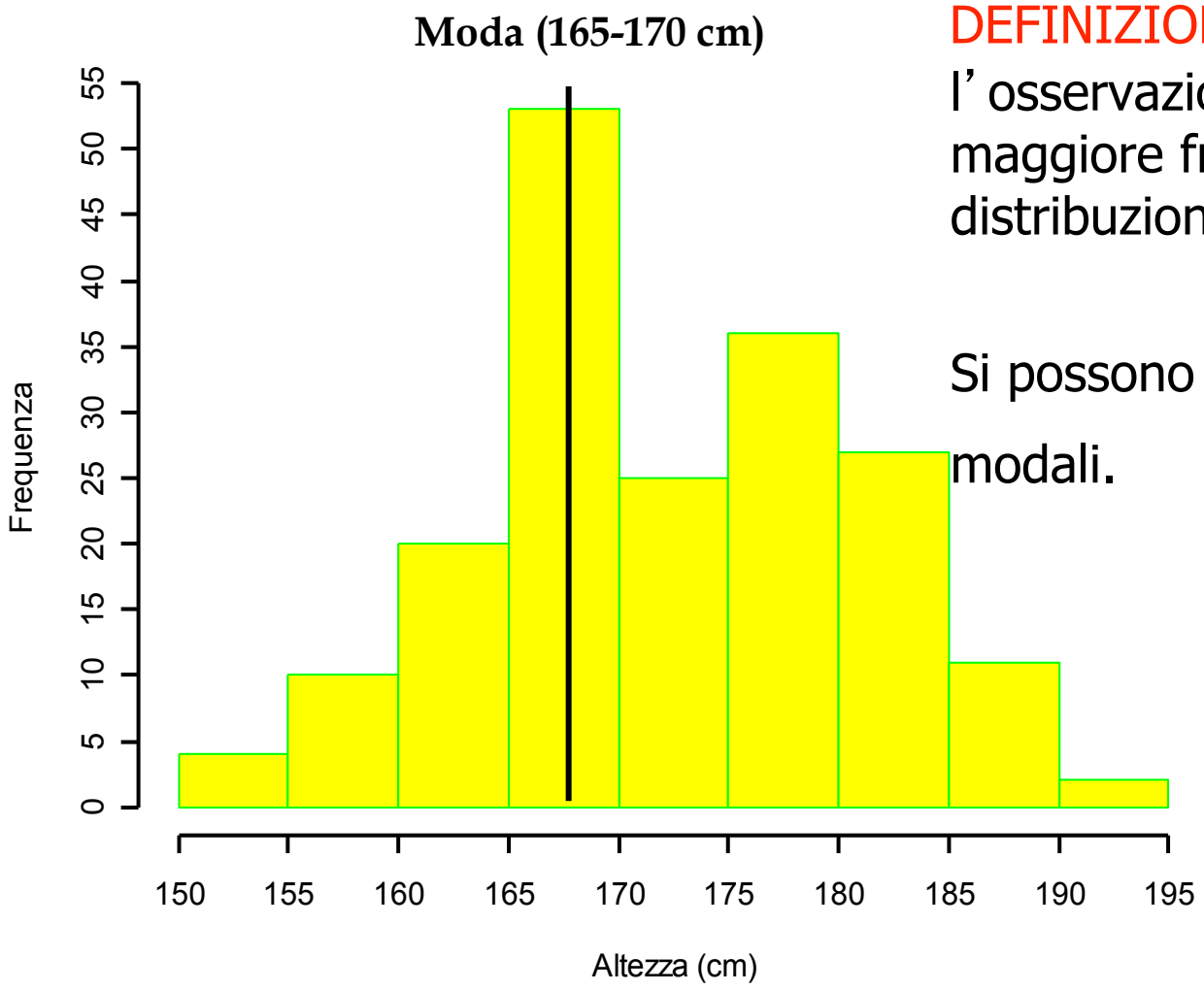
Le fasi operative per il calcolo della mediana sono le seguenti:

1) ordinamento crescente dei dati;

2) se il numero di dati n è dispari, la mediana corrisponde al dato che occupa la (n+1)/2 esima posizione

3) se il numero di dati n è pari, la mediana è data dalla media aritmetica dei due dati che occupano la posizione n/2 e quella n/2+1.

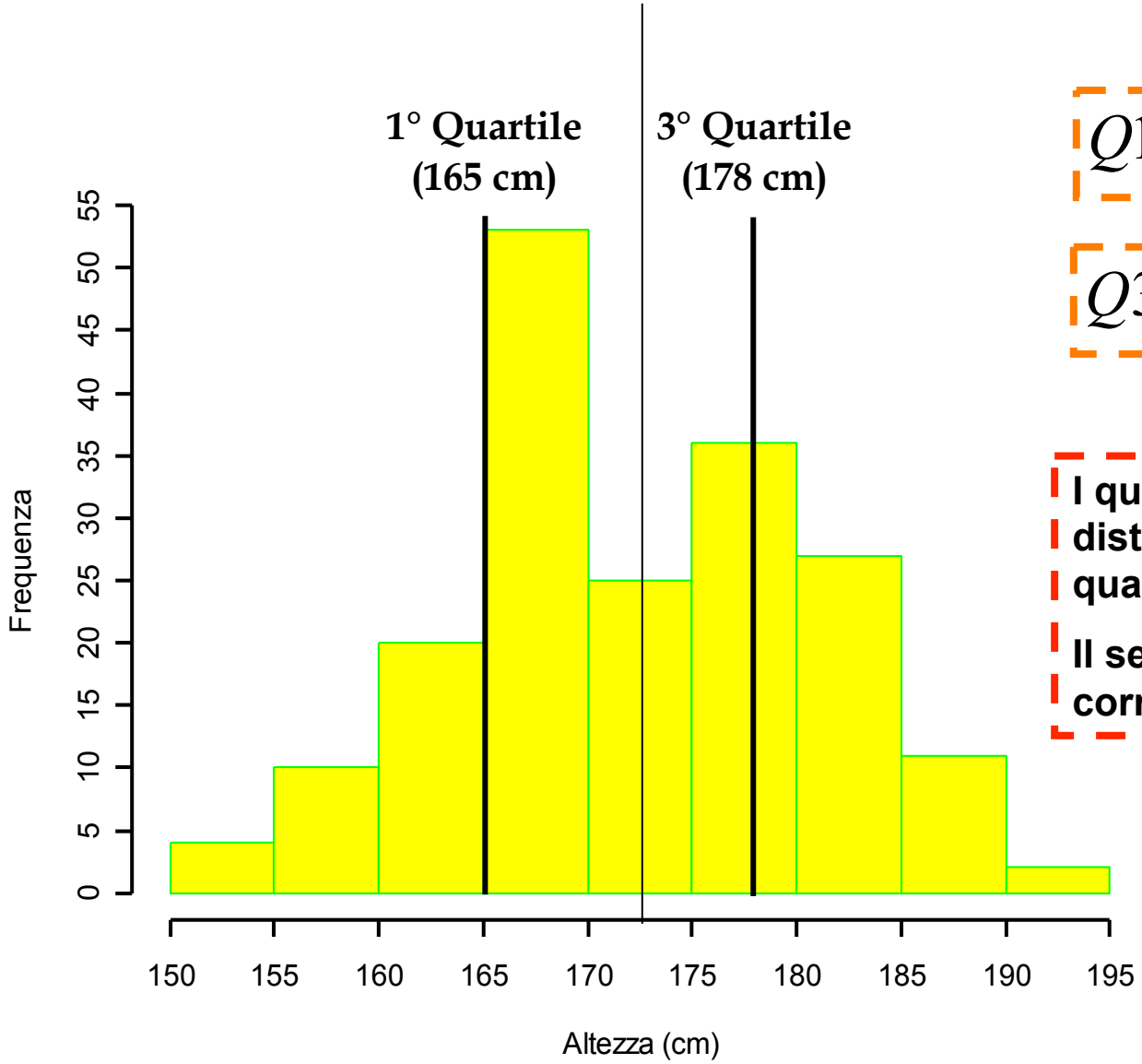
LA MODA



DEFINIZIONE: La Moda (Mo) è l'osservazione che si verifica con maggiore frequenza in una data distribuzione.

Si possono avere anche più valori modali.

Misure di Posizione: i Quartili



$$Q1 = x_{(0.25\cdot(n+1))}$$

$$Q3 = x_{(0.75\cdot(n+1))}$$

I quartili dividono la distribuzione ordinata in quattro parti uguali.

Il secondo quartile (Q2) corrisponde alla *mediana*

Misure di Posizione: Calcolo dei Quartili

Rango	Valore
1	10
2	20
3	40
4	70
5	80
6	90
7	100
8	120

QUARTILE 1

Posizione = 0.25 * (8+1) = 2.25

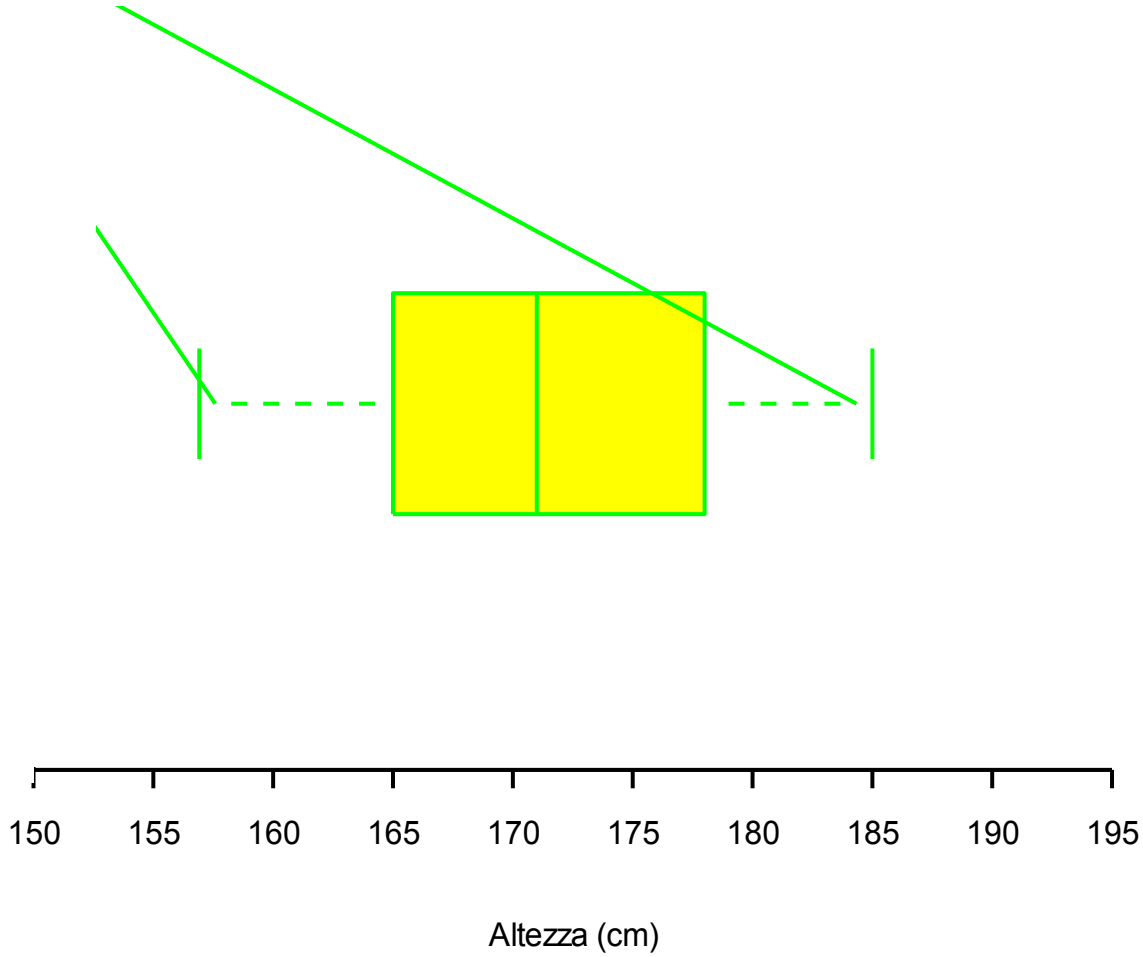
Valore Q1 = 20 + (40-20) * 0.25 = 25

QUARTILE 3

Posizione = 0.75 * (8+1) = 6.75

Valore Q3 = 90 + (100-90) * 0.75 = 97.5

Rango	Valore
1	10
2	20
3	40
4	70
5	80
6	90
7	100
8	120


150
155
160
165
170
175
180
185
190
195
Altezza (cm)

Relazione tra media, mediana e moda

In una distribuzione perfettamente simmetrica, la media, la mediana e la moda hanno lo stesso valore. In una distribuzione asimmetrica, la media si posiziona nella direzione dell'asimmetria. Nelle distribuzioni di dati biologici, l'asimmetria è quasi sempre verso destra (asimmetria positiva, verso i valori più elevati), e quindi la media è > della mediana o della moda

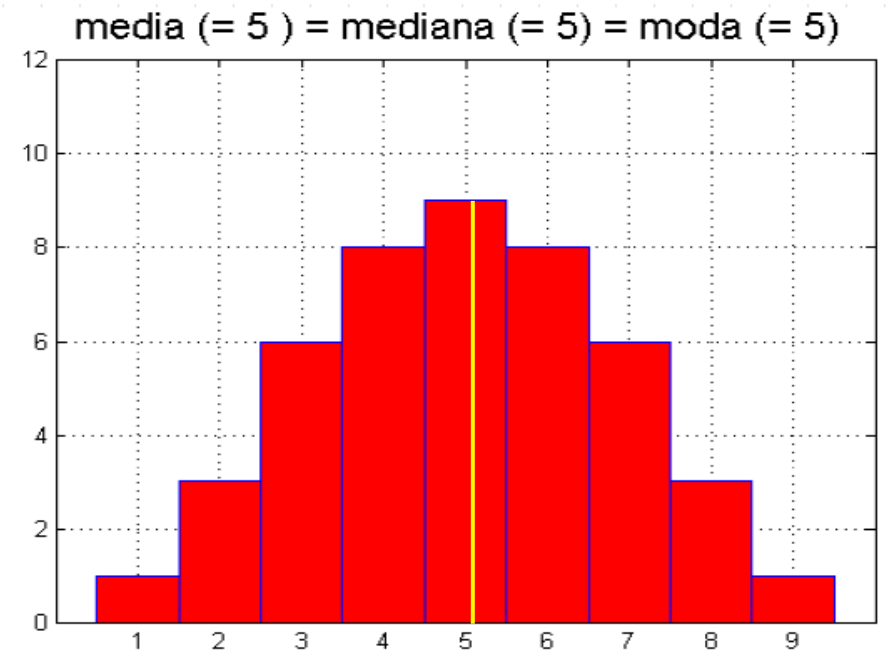
INDICI DI SIMMETRIA

Distribuzione ***simmetrica:***

Le osservazioni equidistanti dalla mediana (coincidente in questo caso col massimo centrale) presentano la stessa frequenza relativa

Un esempio importante è fornito dalla ***distribuzione normale***

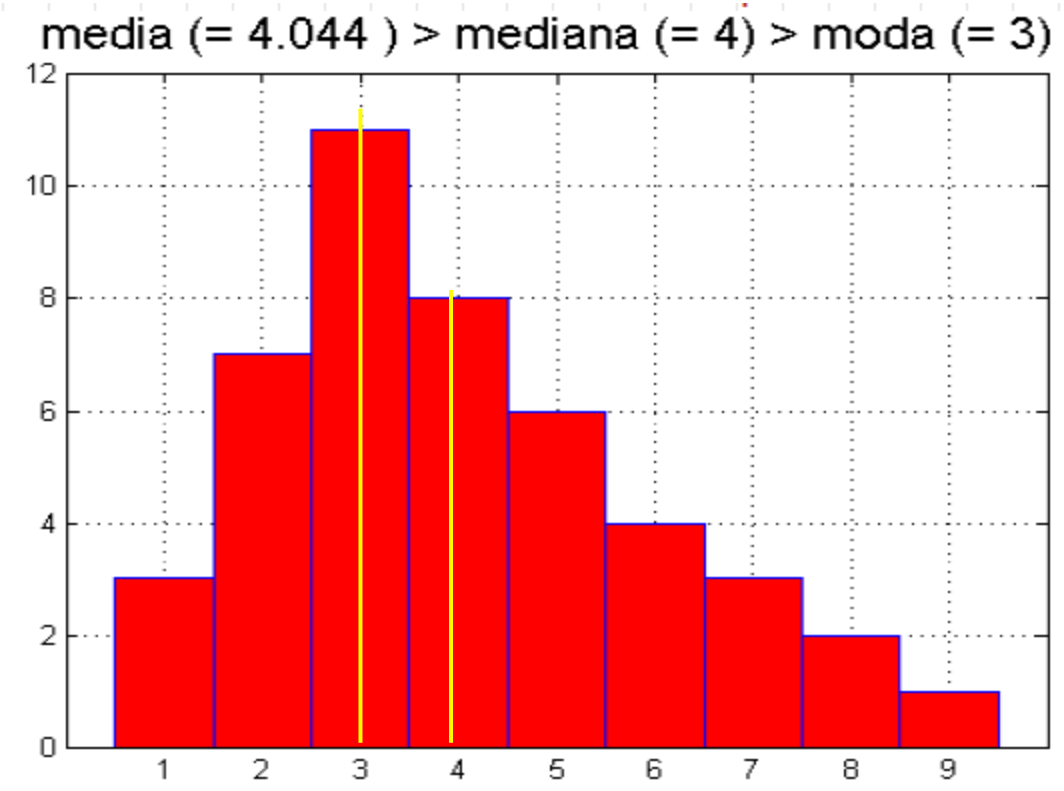
Media = Mediana = Moda



Distribuzione ***asimmetrica positiva***

La curva di frequenza ha una coda più lunga a destra del massimo centrale

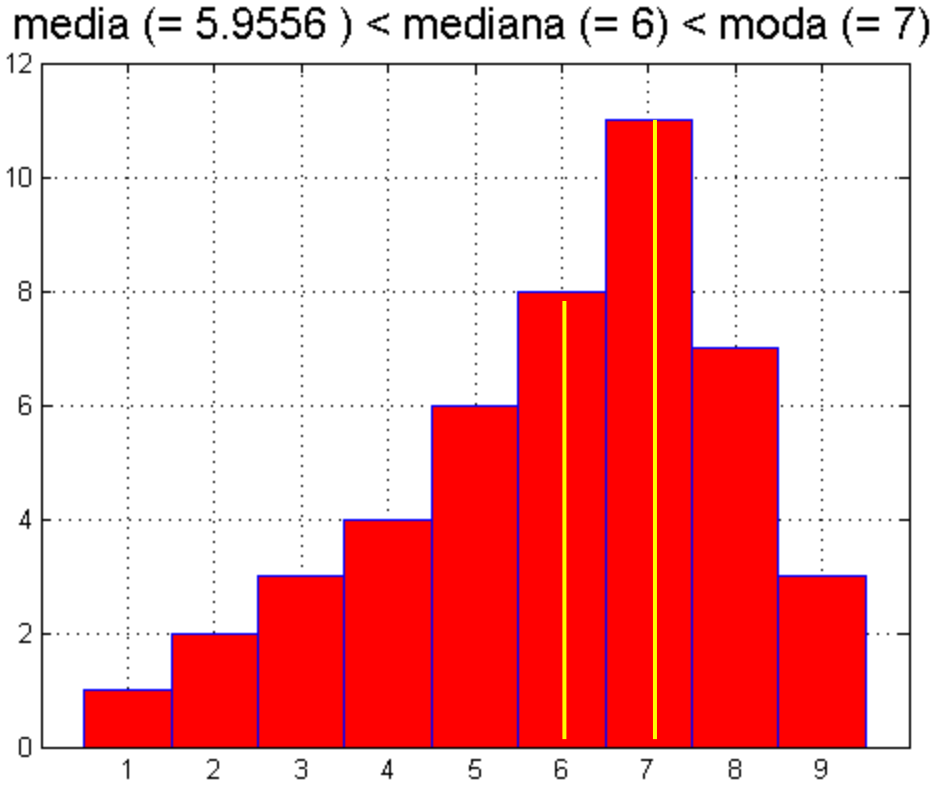
Media > Mediana > Moda



Distribuzione **asimmetrica negativa**

La curva di frequenza ha una coda più lunga a sinistra del massimo centrale

Media < Mediana < Moda



Misure di Dispersione: Il Range

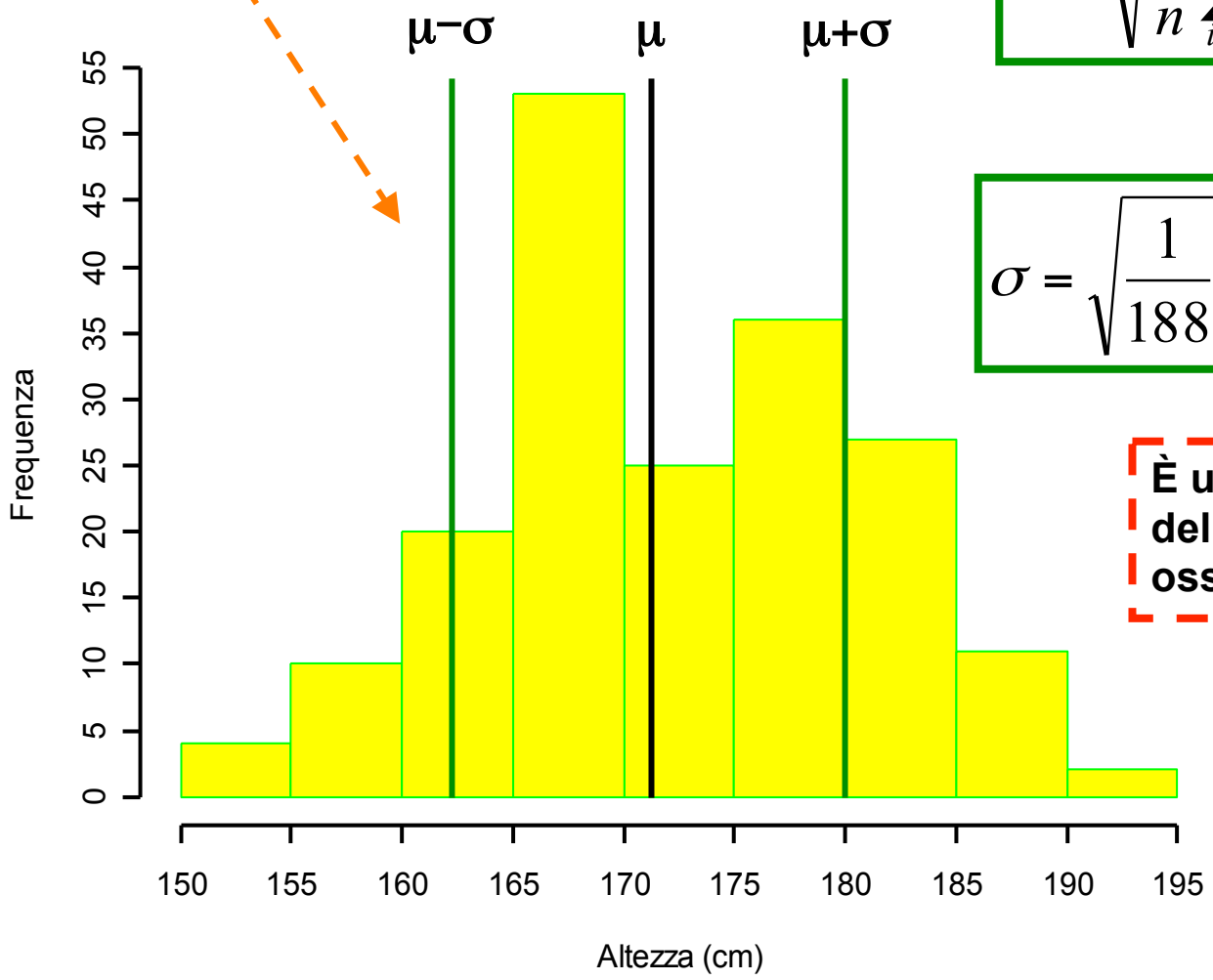
DEFINIZIONE: Il Campo di variazione o Range corrisponde alla differenza fra la modalità più piccola e la modalità più grande della distribuzione

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

- ✓ è troppo influenzato dai valori estremi;
- ✓ tiene conto dei due soli valori estremi, trascurando tutti gli altri;
- ✓ è influenzato dalla numerosità campionaria (sample size).

Misure di Dispersione: La Deviazione Standard

POPOLAZIONE



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

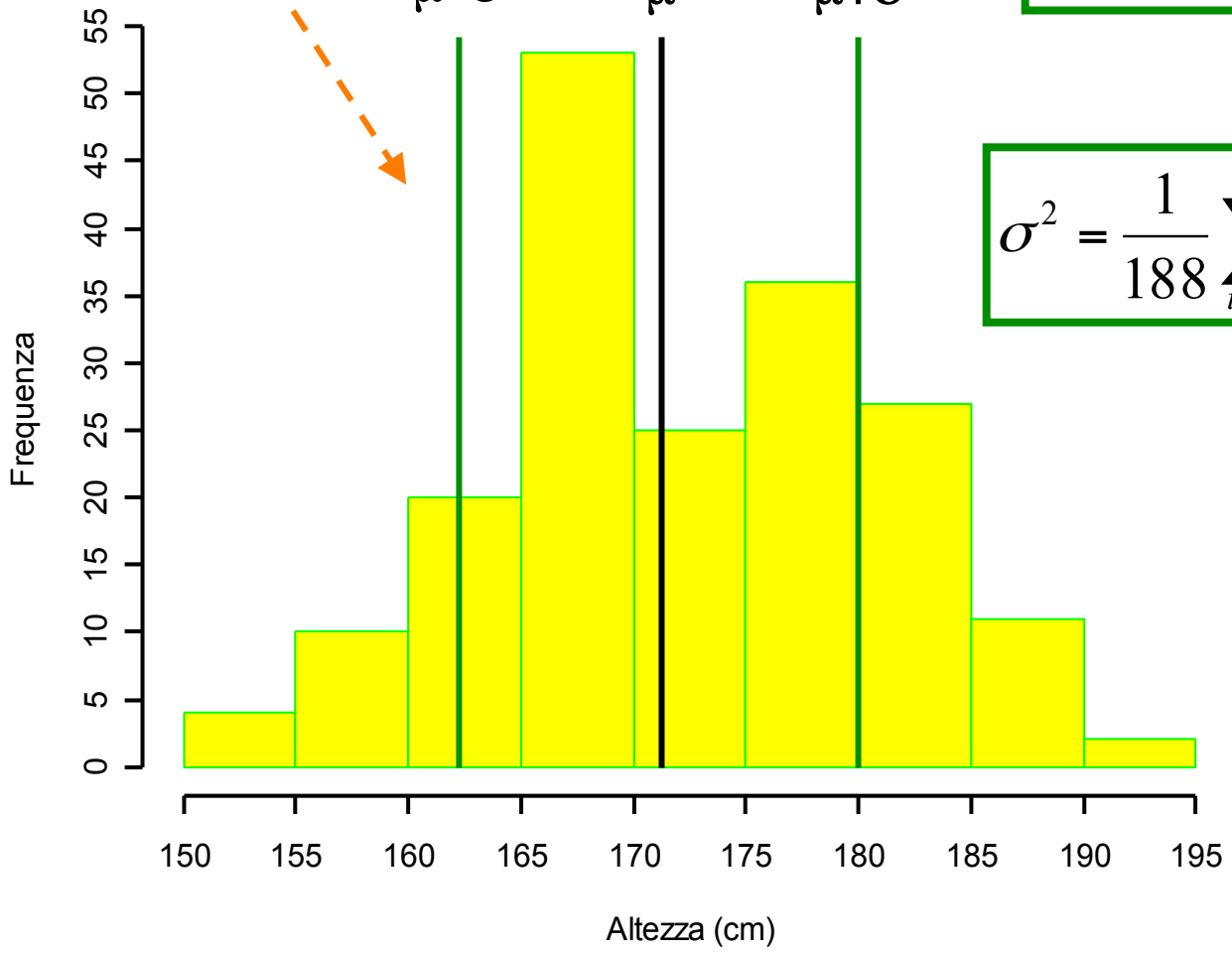
$$\sigma = \sqrt{\frac{1}{188}\sum_{i=1}^{n}(x_i - 171.5)^2} = 8.5$$

È una misura riassuntiva delle differenze di ogni osservazione dalla media

Ha la stessa unità di misura della media aritmetica

Misure di Dispersione: La Varianza

POPOLAZIONE

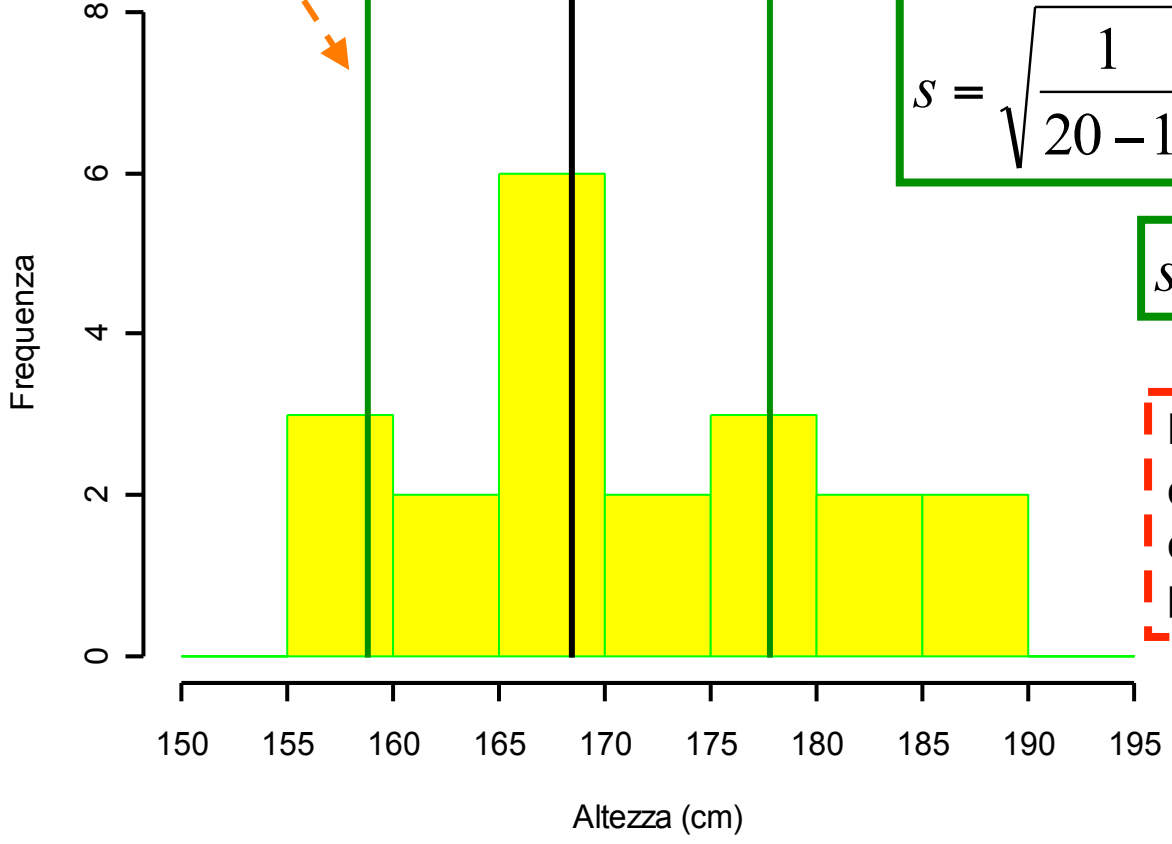


$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{188}\sum_{i=1}^{n}(x_i - 171.5)^2 = 72.5$$

Misure di Dispersione: La Deviazione Standard

CAMPIONE



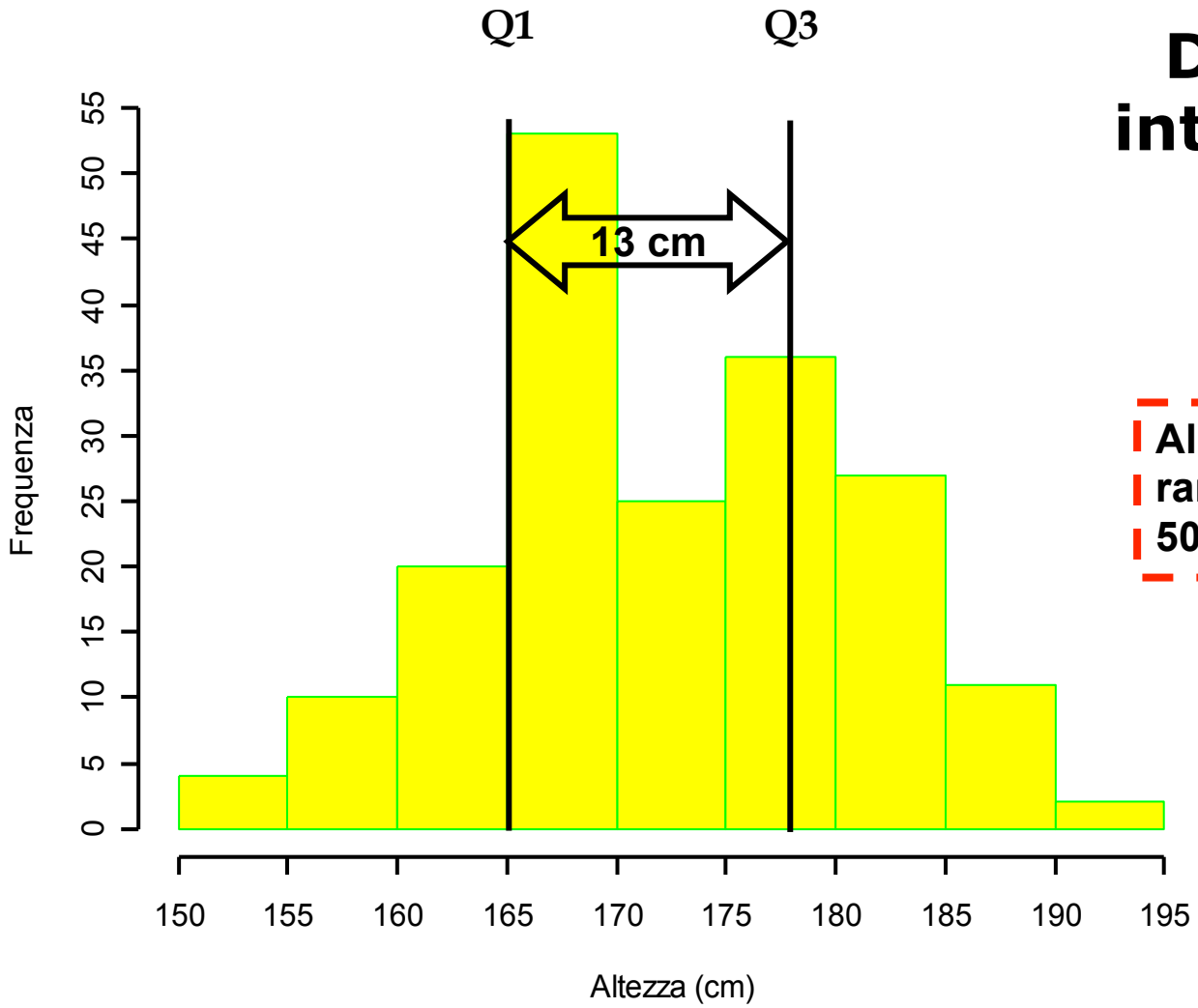
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$s = \sqrt{\frac{1}{20-1}\sum_{i=1}^{n}\left(x_i - 170.15\right)^2} = 9.3$$

$$s^2 = 86{,}87$$

La deviazione standard del campione (s) *stima* la deviazione standard della popolazione (σ)

Misure di Dispersione: La differenza interquartile



Differenza interquartile: Q3 – Q1

All'interno di questo range sono contenute il 50% delle osservazioni

Misure di Dispersione: IL COEFFICIENTE DI VARIAZIONE

$$CV = \frac{\text{Deviazione Standard}}{\text{Media Aritmetica}} \cdot 100$$

La variabilità guarda alle differenze tra le unità sperimentali. E' pero evidente che il significato pratico delle differenze può dipendere dal livello del fenomeno considerato.

Può quindi essere interessante disporre di una qualche misura di variabilità aggiustata in qualche maniera per tenere conto del livello del fenomeno stesso.