STATISTICA DESCRITTIVA
MEDIE

SAPIENZA
UNIVERSITÀ DI ROMA

annarita.vestri@uniroma1.it

STATISTICA DESCRITTIVA

La Sintesi Statistica

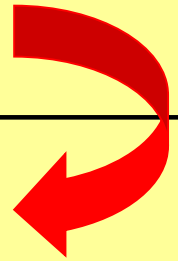
Una serie di dati numerici è compiutamente descritta da tre proprietà principali:

- La **tendenza centrale** o **posizione**
- La **dispersione** o **variabilità**
- La **forma**

Queste misure descrittive sintetiche, riassuntive dei dati tabellari, sono chiamate:

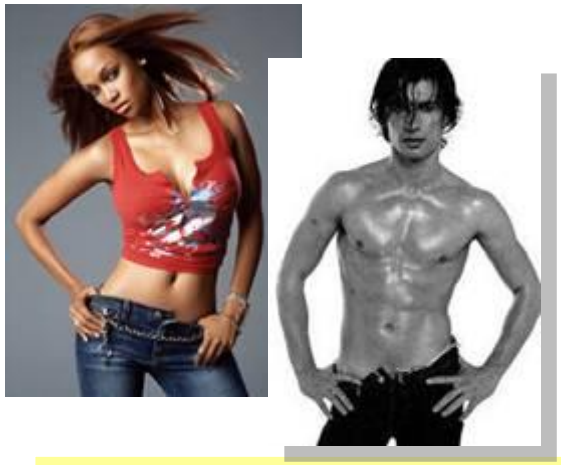
- **statistiche**, quando sono calcolate su un campione di dati (si esprimono con lettere dell'alfabeto latino)
- **parametri**, quando descrivono la popolazione od universo dei dati (si esprimono con lettere dell'alfabeto greco

Medie ferme	Medie lasche
Aritmetica	Moda
Geometrica	Mediana
Armonica	Quantili
Quadratica	



Medie lasche: MODA

- E’ la scelta fatta dalla maggioranza della popolazione
- In statistica è lo stesso
- Si definisce MODA di un insieme dei dati o di una distribuzione di frequenza la modalità della variabile alla quale corrisponde la **massima frequenza**
- E’ l’unica media che si può applicare indifferentemente a serie e a seriazioni



MODA

E' la scelta fatta dalla maggioranza della popolazione, lo stile che "tutti" seguono

Si definisce moda (*classe modale*) di un insieme di dati o di una distribuzione di frequenza la modalità / il valore (*l'intervallo di classe*) della variabile cui corrisponde la massima frequenza.

esempio:
X = tipo di parto
(50 neonati)

modalità	frequenza assoluta	frequenza relativa	frequenza relativa percentuale
normale	35	0.70	70%
forcipe	1	0.02	2%
cesareo	14	0.28	28%
TOTALE	50	1.00	100%

MODA

Esempio 4 (seriazione)

Numero di figli	n_i
0	8
1	14
2	20
3	6
4	4
5	2
Totale	54

Massima frequenza

Esempio 4 (seriazione)

Numero di figli	n_i
0	8
1	14
2	20
3	6
4	4
5	2
Totale	54

Massima frequenza

…quindi

Esempio 4 (seriazione)

Numero di figli	n_i
0	8
1	14
2	20
3	6
4	4
5	2
Totale	54

Massima frequenza

…quindi

Mo = 2

Nel caso di una variabile continua espressa in classi, tutte di uguale ampiezza, si può individuare la classe modale

Esempio 6

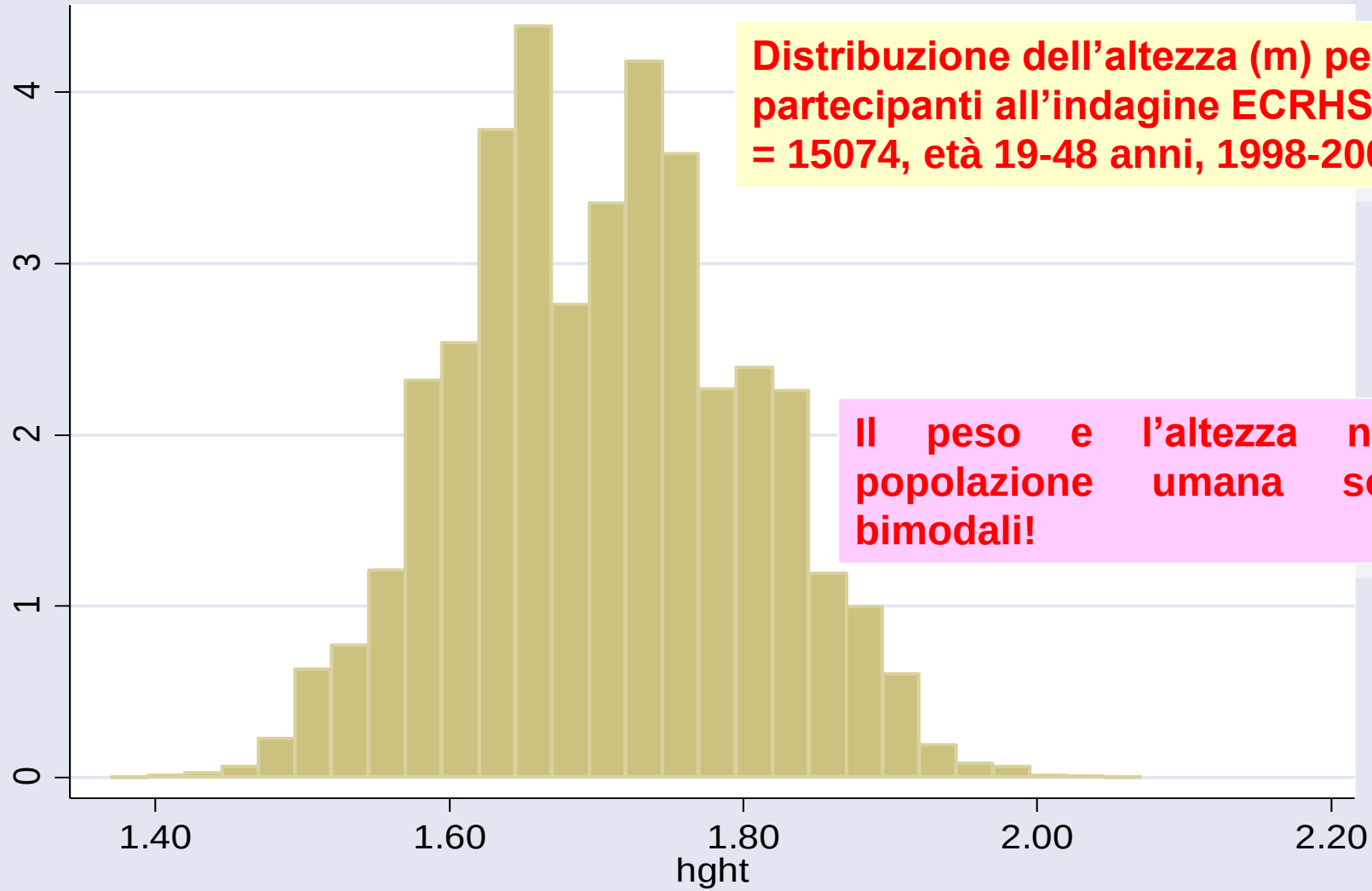
Altezza (cm)	n_i
150 - 159	49
160 – 169	54
170 - 179	61
180 - 189	16
Totale	180

Esempio 6

Classe modale = 170 - 179

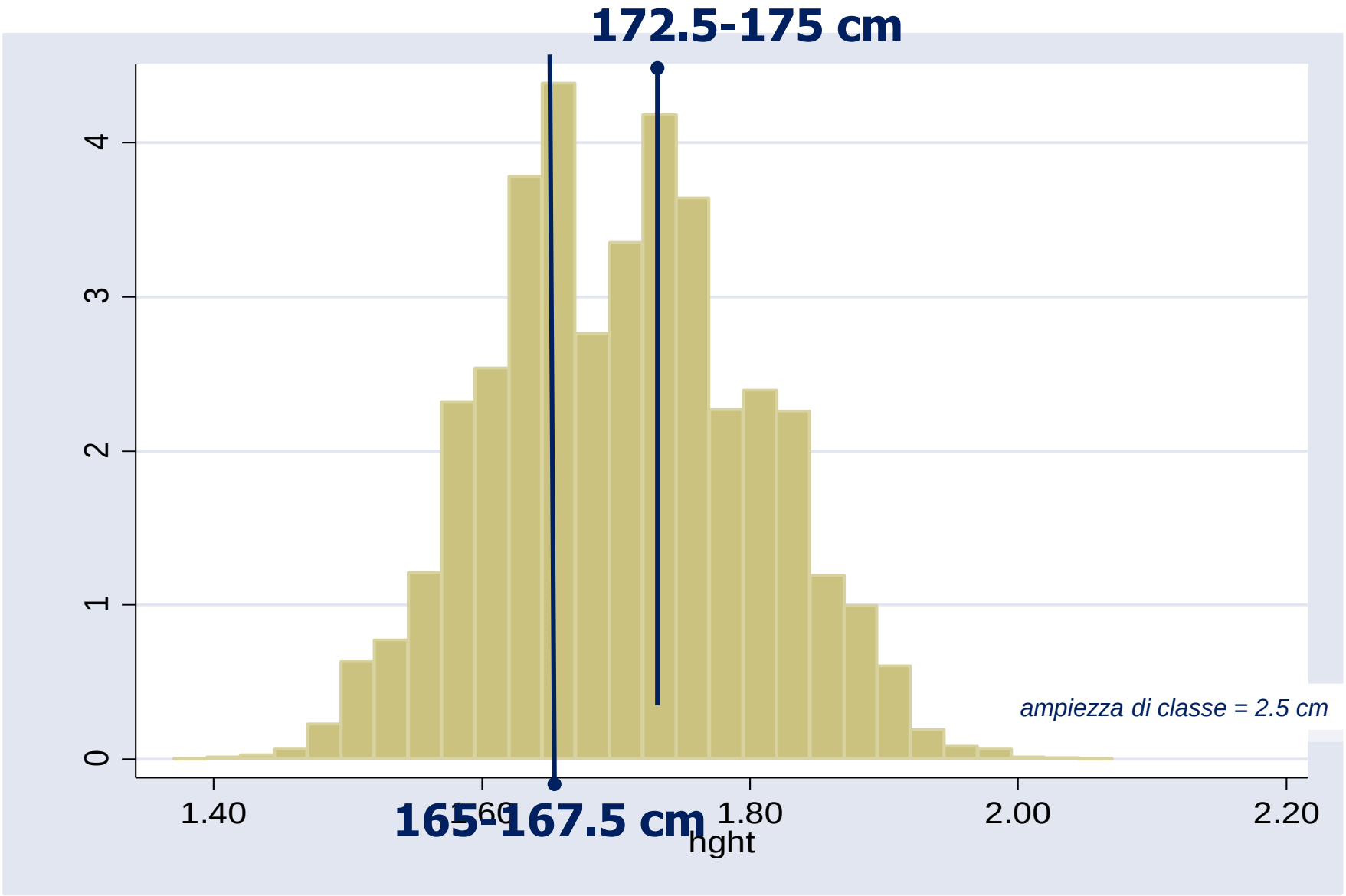
Altezza (cm)	n_i
150 - 159	49
160 – 169	54
170 - 179	61
180 - 189	16
Totale	180

MA LA MODA E' SEMPRE UNA SOLA?

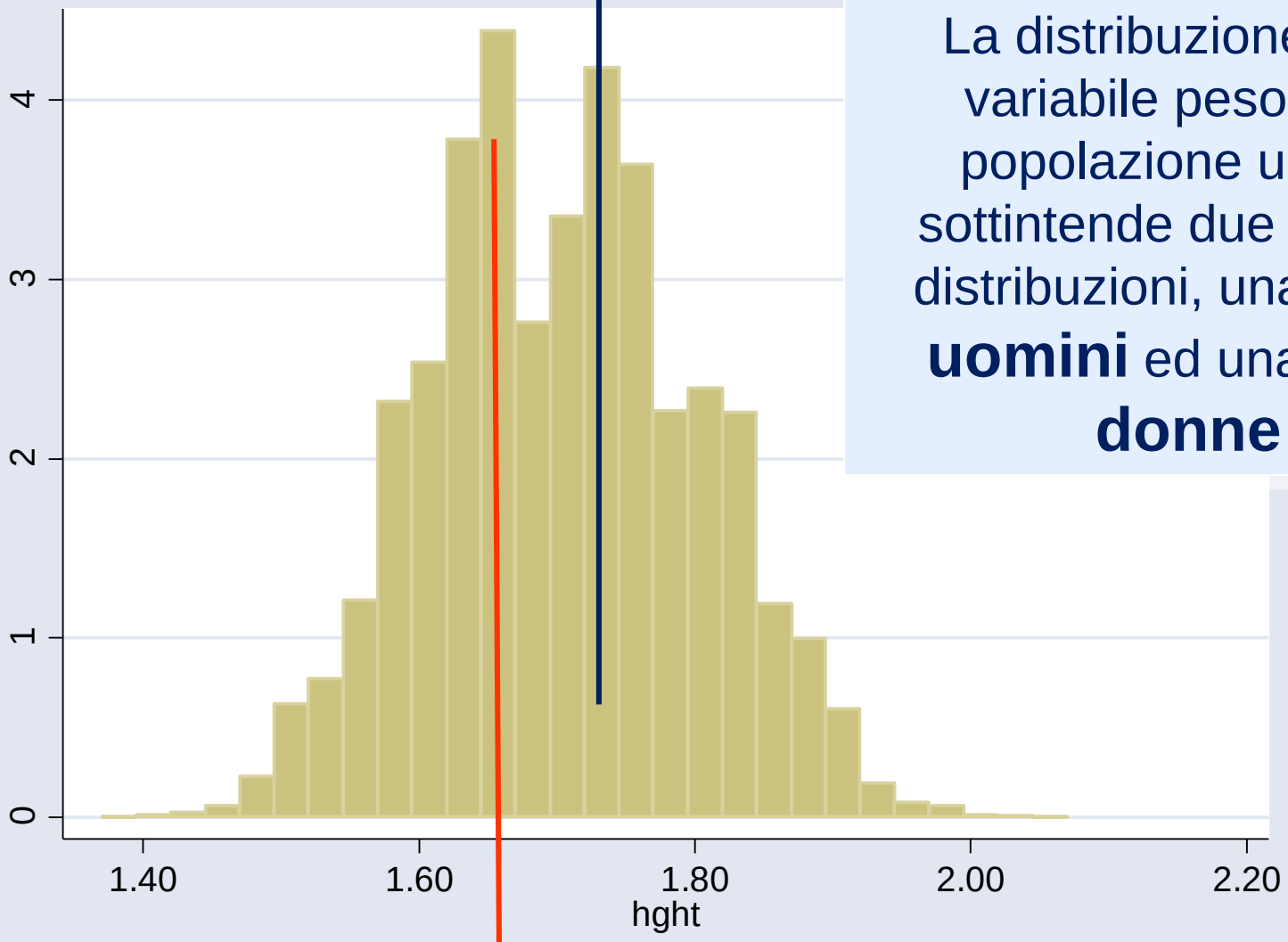


Distribuzione dell'altezza (m) per i partecipanti all'indagine ECRHS II (n = 15074, età 19-48 anni, 1998-2002)

Il peso e l'altezza nella popolazione umana sono bimodali!


172.5-175 cm
0
1
2
3
4
1.40
1.80
2.00
2.20
hght
ampiezza di classe = 2.5 cm
165-167.5 cm

172.5-175 cm



La distribuzione della variabile peso nella popolazione umana sottintende due diverse distribuzioni, una per gli **uomini** ed una per le **donne**

165-167.5 cm

MEDIANA

Si dice **mediana** il valore che occupa il posto centrale in una distribuzione statistica di frequenza i cui valori sono disposti in ordine crescente

- Mediana (Me) è sinonimo di **50-esimo percentile** o di **II quartile**

se n è *dispari* → $\mathbf{Me = x_{[(n+1)/2]}}$

se n è *pari* → $\mathbf{Me = [x_{n/2} + x_{(n/2+1)}] / 2}$

Esempio 1: numero dispari di valori
Abbiamo undici partecipanti a un seminario di formazione ai quali chiediamo l'età; le risposte sono le seguenti:

28, 34, 51, 19, 62, 43, 29, 38, 45, 26, 49

Il primo passo è di mettere le risposte in ordine crescente:

19, 26, 28, 29, 34, 38, 43, 45, 49, 51, 62

1. metto le unità in ORDINE crescente di altezza

es. sulla mediana

19, 26, 28, 29, 34, 38, 43, 45, 49, 51, 62

2. identifico l'unità centrale nella serie ordinata di dati

19, 26, 28, 29, 34, 38, 43, 45, 49, 51, 62

3. la mediana è il **VALORE** che la variabile altezza assume sull'unità che divide il campione in due parti numericamente uguali

formalmente:

$$\rightarrow \quad Me = x_{[(n+1)/2]} = x_{(11+1/2)} = x_6$$

19, 26, 28, 29, 34, **38**, 43, 45, 49, 51, 62

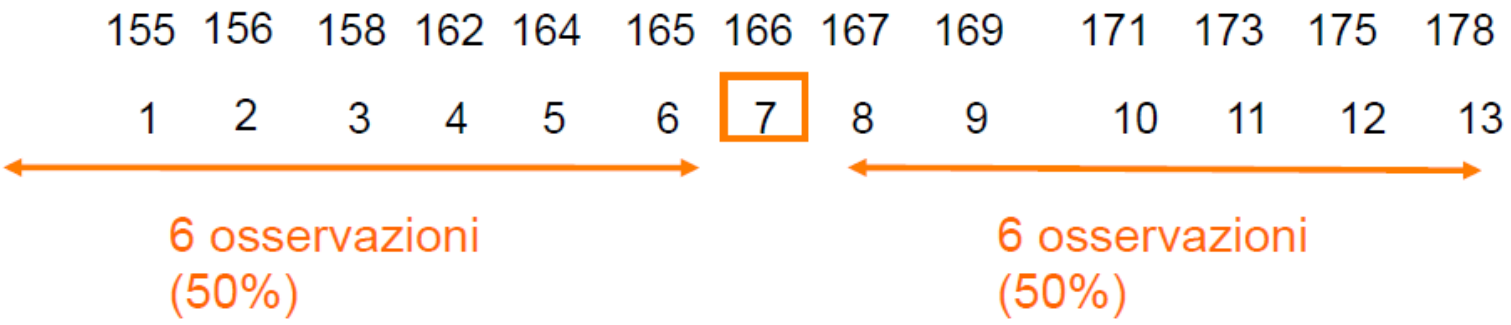
NB: le misure di posizione sono *valori* o *modalità*, NON *frequenze*!

Nel caso di distribuzioni di frequenza, occorre

1) Ordinare la seriazione (nel caso non lo sia)
2) Calcolare le frequenze cumulate
3) Se N dispari il valore centrale è nel posto (N+1)/2
4) Se N pari i valori centrali sono nei posti N/2 e N/2 +1
5) Individuare in quale frequenza cumulata si trova la mediana e quindi a quale modalità corrisponde

Esempio: In un campione di 13 soggetti viene osservato il carattere Altezza (cm):
173 155 162 165 167 175 171 169 164 178 156 158 166

→ Ordiniamo in senso crescente le osservazioni, attribuendogli la pozione in graduatoria (RANGO):



mediana = 166

n pari → mediana = modalità intermedia fra quelle di posto n/2 e n/2+1 (ad esempio, se n=6, è la modalità centrale fra la 3° e la 4°)

n dispari → mediana = modalità di posto (n+1)/2

Esempio 10: N dispari

Età	n_i
20	8
21	11
22	23
23	44
24	52
25	27
26	20
Totale	185

Esempio 10: N dispari

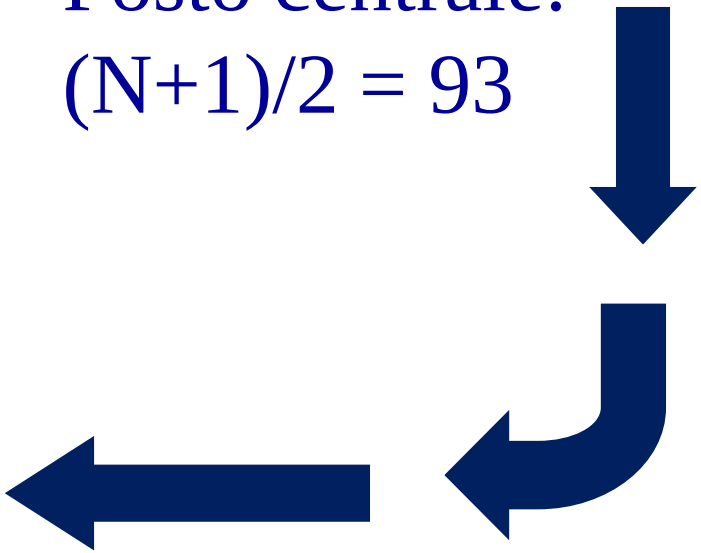
Età	n_i	N_i
20	8	8
21	11	19
22	23	42
23	44	86
24	52	138
25	27	165
26	20	185
Totale	185	

Posto centrale:
$(N+1)/2 = 93$

Esempio 10: N dispari

Età	n_i	N_i
20	8	8
21	11	19
22	23	42
23	44	86
24	52	138
25	27	165
26	20	185
Totale	185	

Posto centrale:
(N+1)/2 = 93



Me = 24

Esempio 11: N pari DATI IN CLASSI

Altezza	n_i
145 - 149	11
150 - 154	12
155 - 159	14
160 - 164	15
165 - 169	22
170 - 174	60
175 e oltre	66
Totale	200

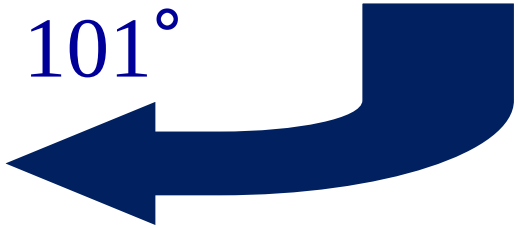
Esempio 11: N pari DATI IN CLASSI

Altezza	n_i	N_i
145 - 149	11	11
150 - 154	12	23
155 - 159	14	37
160 - 164	15	52
165 - 169	22	74
170 - 174	60	134
175 e oltre	66	200
Totale	200	

Posti centrali

N/2 = 100°

N/2 + 1 = 101°



Me = 170 - 174

La mediana è particolarmente utile nella sintesi di **distribuzioni asimmetriche**; in questo caso infatti la media aritmetica, considerando anche i valori estremi anomali, finirebbe col sovrastimare il fenomeno

I PERCENTILI

K-MO PERCENTILE (n_i)

Quel **VALORE** x_i della variabile tale per cui il k% delle osservazioni del campione assume valori $\leq x_i$.

K è noto anche come RANGO PERCENTILICO

I PERCENTILI PIU' NOTI:

25° → 1° QUARTILE

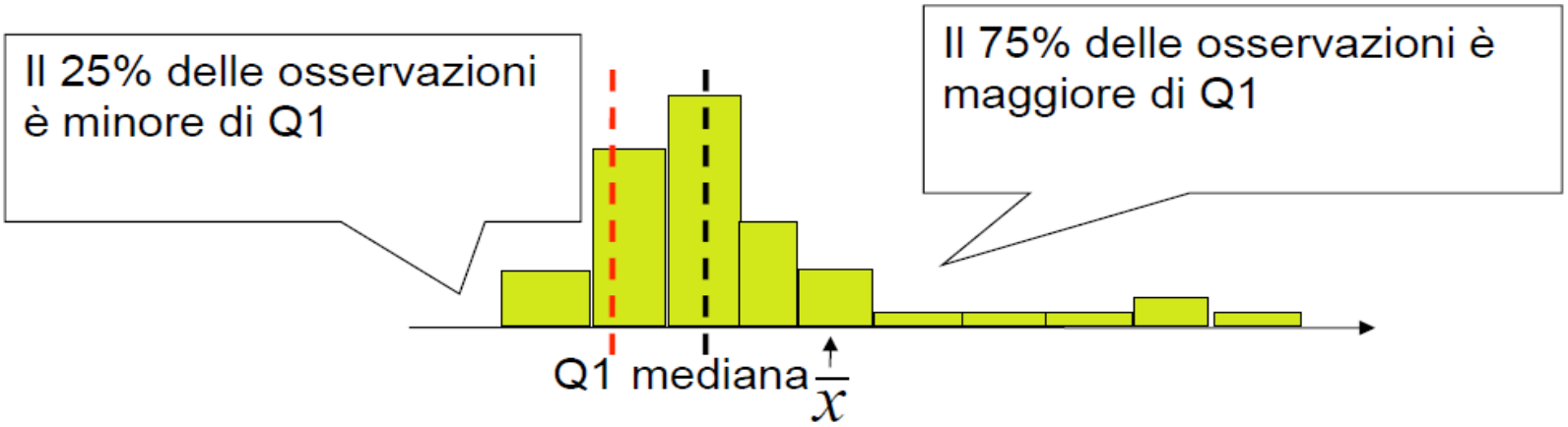
50° → 2° QUARTILE o MEDIANA

75° → 3° QUARTILE

° **QUARTILE – 1° QUARTILE** **DIFFERENZA INTERQUARTILICA**

Generalizzazione della mediana: quantili

- La mediana separa la distribuzione in due parti, ognuna comprendente il 50% delle osservazioni
- I quantili separano la distribuzione ad altre frazioni percentuali, ad esempio:
  - Il 1^0 quartile (Q1) separa il primo 25% dal restante 75%
  - Il 3^0 quartile (Q3) separa il primo 75% dal restante 25%
  - Il 1^0 decile separa il primo 10% dal restante 90%
  - Il 95° percentile è tale che solo il 5% ha un valore superiore a esso
  - etc.



Nota: la mediana e tutti i quantili possono essere calcolati anche per caratteri QUALITATIVI ORDINATI

Calcolo dei quartili (1)

Osservato un collettivo di n unità occorre considerare se $\frac{n}{4}$ è un numero intero o meno

- $\frac{n}{4}$ intero:

$$Q_1 = \frac{1}{2}(x_{\frac{n}{4}} + x_{\frac{n}{4}+1}) \qquad ; \qquad Q_3 = \frac{1}{2}(x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1})$$

- $\frac{n}{4}$ non intero:

  Indichiamo con $\left[\frac{n}{4}\right]$ la parte intera di $\frac{n}{4}$. Allora si ha

$$Q_1 = x_{\left[\frac{n}{4}\right]+1} \qquad ; \qquad Q_3 = x_{\left[\frac{3n}{4}\right]+1}$$

Calcolo dei quartili (2)

Esempio ($\frac{n}{4}$ intero):

$$n = 8: \quad 4, 6, 12, 3, 5, 6, 9, 7 \quad \Rightarrow \quad 3, 4, 5, 6, 6, 7, 9, 12$$

$$x_{[2]} = 4 \text{ e } x_{[3]} = 5 \quad \Rightarrow \quad Q_1 = \frac{4+5}{2} = 4.5$$

$$x_{[6]} = 7 \text{ e } x_{[7]} = 9 \quad \Rightarrow \quad Q_3 = \frac{7+9}{2} = 8.0$$

Esempio ($\frac{n}{4}$ non intero):

$$n = 9: \quad 4, 6, 12, 3, 5, 6, 10, 7, 1 \quad \Rightarrow \quad 1, 3, 4, 5, 6, 6, 7, 10, 12$$

$$Q_1 = x_{[3]} = 4$$

$$Q_3 = x_{[7]} = 7$$

Calcolo dei quartili (3)

Dati sotto forma di **tabella di frequenza**

- Il **primo quartile** (Q_1) coincide con la modalità per cui la frequenza cumulata relativa supera per la prima volta il valore 0.25
- Il **terzo quartile** (Q_3) coincide con la modalità per cui la frequenza cumulata relativa supera per la prima volta il valore 0.75

Numero di Figli	n_i	f_i	F_i
0	8	0.148	0.148
1	14	0.259	0.407
2	20	0.370	0.778
3	6	0.111	0.889
4	4	0.074	0.963
5	2	0.037	1.000
Totale	54	1	

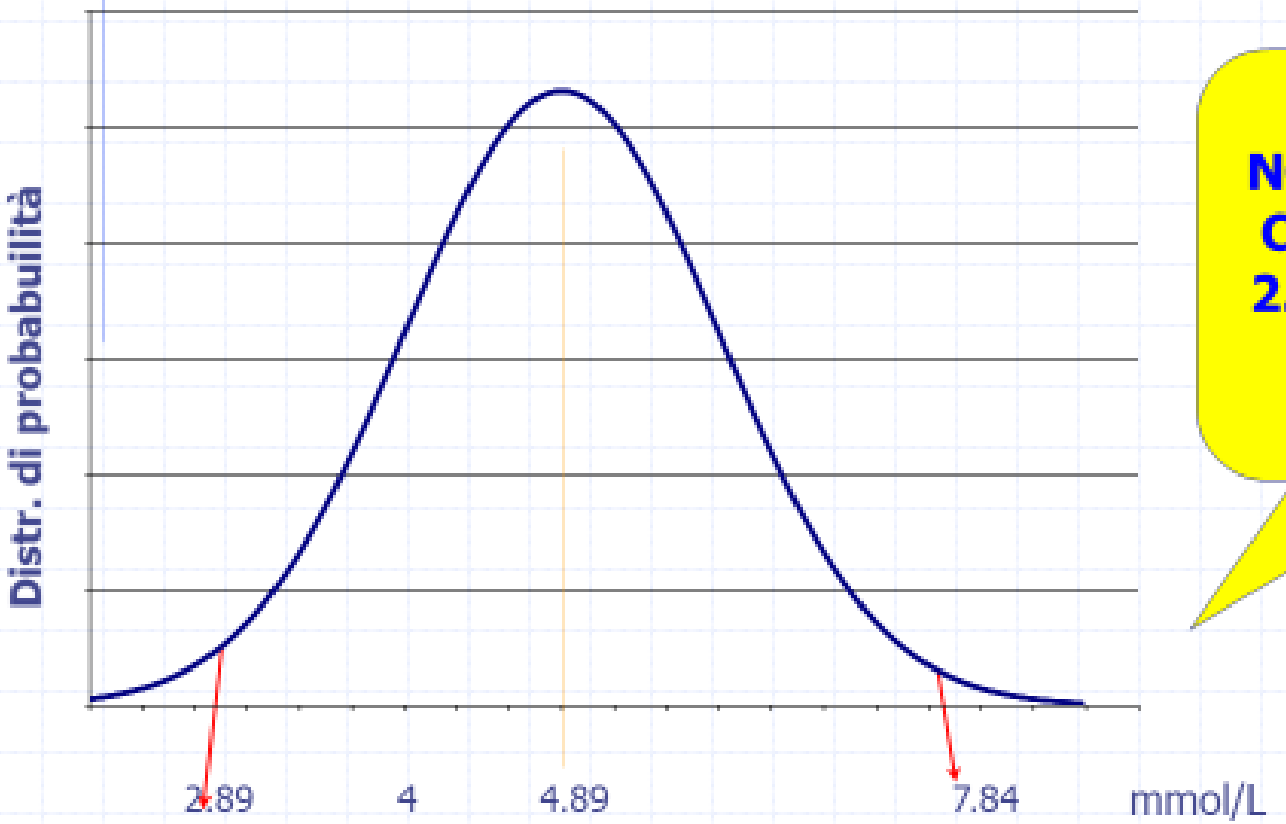
Quindi si ha che $Q_1 = 1$ e $Q_2 = Q_3 = 2$

ESEMPI DELL'UTILITA' DEI PERCENTILI NELLA PRATICA CLINICA

1. intervallo di normalità di un parametro biologico
2. curve auxologiche

PERCENTILI E VALORI DI NORMALITA'

Esempio: i valori di normalità dell'UREA sono compresi tra **2.89** e **7.84** mmol/L. Rappresentiamo la distribuzione di frequenza dei dati relativi ai valori della concentrazione di urea:



I VALORI DI NORMALITA' SONO COMPRESI TRA IL 2.5° PERCENTILE E IL 97.5° PERCENTILE

2.5° PERCENTILE = 2.89

97.5° PERCENTILE = 7.84

INTERVALLI DI RIFERIMENTO

Tabella 1
Limiti di riferimento riferibili per la concentrazione di creatinina plasmatica. Modificato da rif. 20

	Percentile, mg/dL	
Età (sesso)	2,5°	97,5°
Sangue del cordone	0,52	0,97
Neonati pretermine 0–21 giorni	0,32	0,98
Neonati a termine 0–14 giorni	0,31	0,92
2 mesi - <1 anno	0,16	0,39
1 anno - <3 anni	0,17	0,35
3 anni - <5 anni	0,26	0,42
5 anni - <7 anni	0,29	0,48
7 anni - <9 anni	0,34	0,55
9 anni - <11 anni	0,32	0,64
11 anni - <13 anni	0,42	0,71
13 anni - <15 anni	0,46	0,81
Adulti (maschi)	0,72	1,18
Adulti (femmine)	0,55	1,02

INTERVALLI DI RIFERIMENTO

Tabella 2

Limiti di riferimento riferibili per le concentrazioni di attività catalitica di aspartato amminotransferasi (AST) e alanina amminontrasferasi (ALT) nel siero. Modificato da rif. 22

	Femmine		Maschi		Cumulativo	
	2,5° percentile	97,5° percentile	2,5° percentile	97,5° percentile	2,5° percentile	97,5° percentile
AST, U/L	11,0	33,4	13,9	35,1	11,0	34,0
ALT, U/L	7,8	41,0	9.0	59,0	-	-

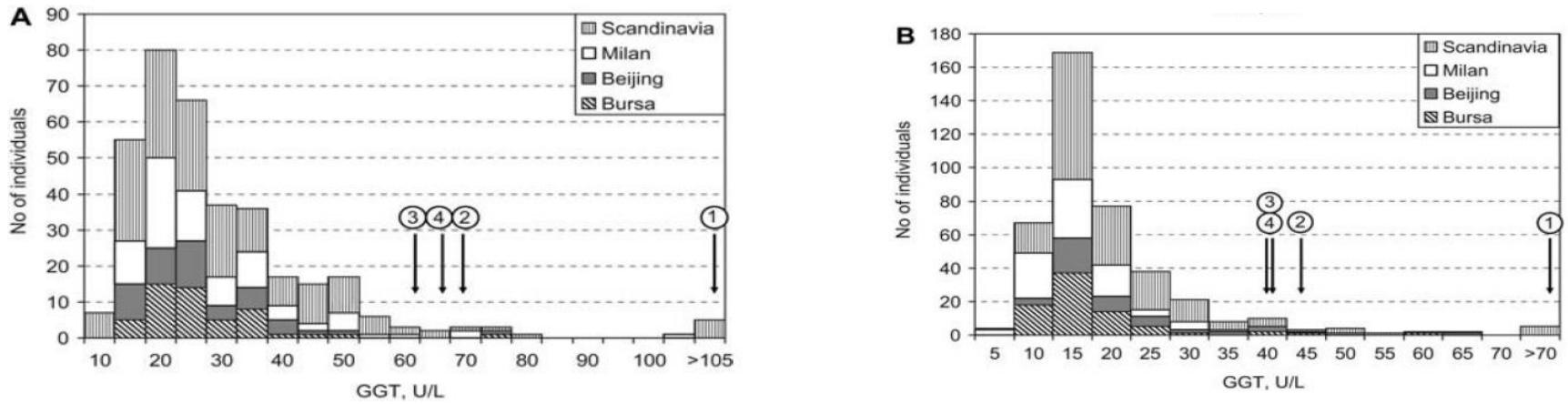


Figura 1

Distribuzione dei risultati relativi ai valori di riferimento riferibili per la γ-glutamiltransferasi (GGT) in 4 diverse regioni. Da rif. 22. A) Distribuzione dei valori della GGT nei maschi. Le frecce indicano il 97,5° percentile, calcolato come segue: 1) Scandinavia: 114 U/L, 2) Milano: 69 U/L, 3) Pechino: 64 U/L e 4) Bursa (Turchia): 66 U/L. B) Distribuzione dei valori della GGT nelle femmine. Le frecce indicano il 97,5° percentile, calcolato come segue: 1) Scandinavia: 69 U/L, 2) Milano: 44 U/L, 3) Pechino: 39 U/L e 4) Bursa: 41 U/L.

LE CURVE DI ACCRESCIMENTO

- Elaborate dai centri auxologici o statistici delle differenti nazioni

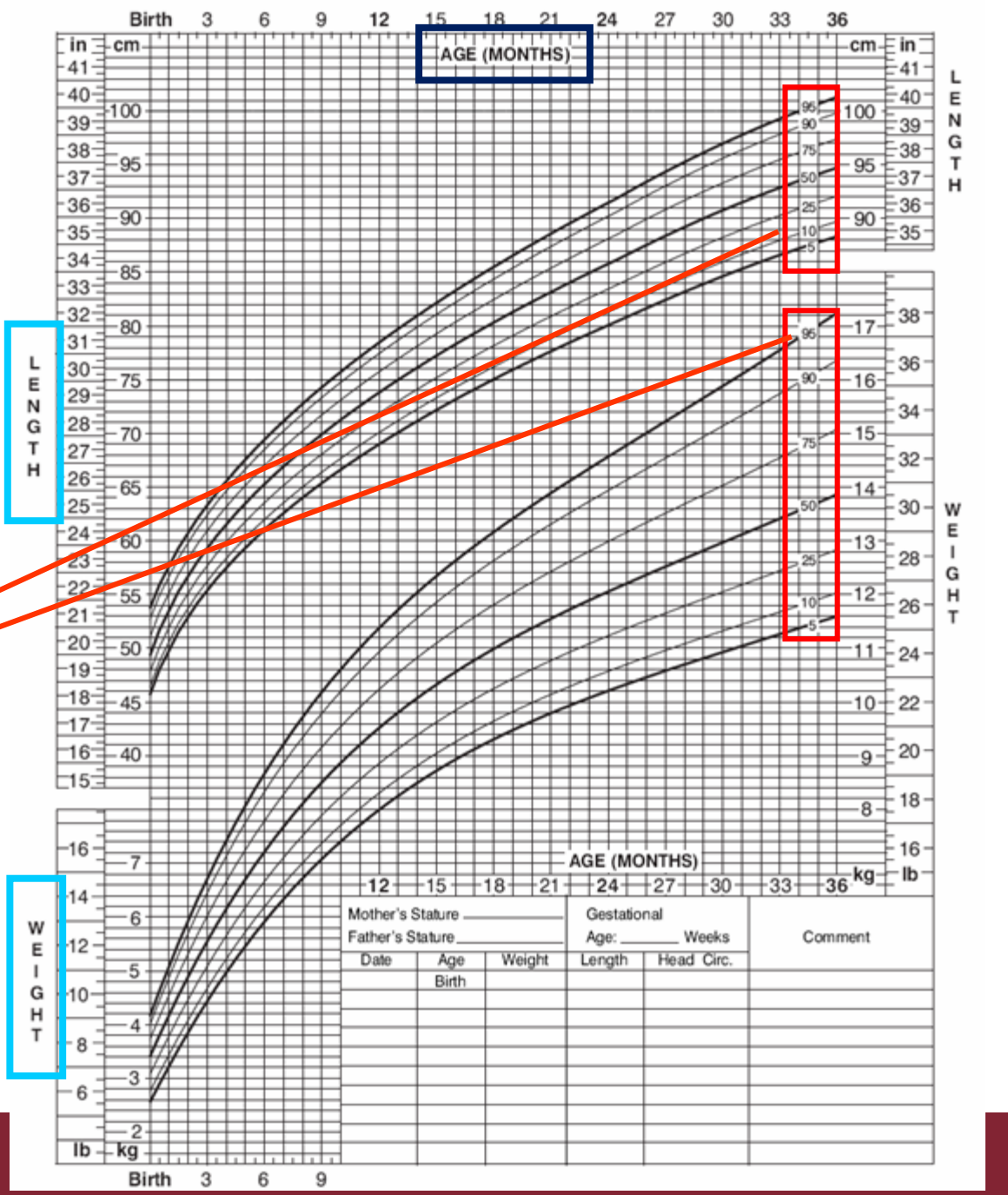
- **Rappresentano il modo in cui la popolazione cresce in funzione dell'età**

- Indicano a quali valori percentilici appartengono le varie stature e pesi che persone di sesso femminile e maschile possono presentare.

BAMBINE DA 0 A 36 MESI

RAPPRESENTAZIONE DEI PERCENTILI DI PESO E STATURA PER ETA'

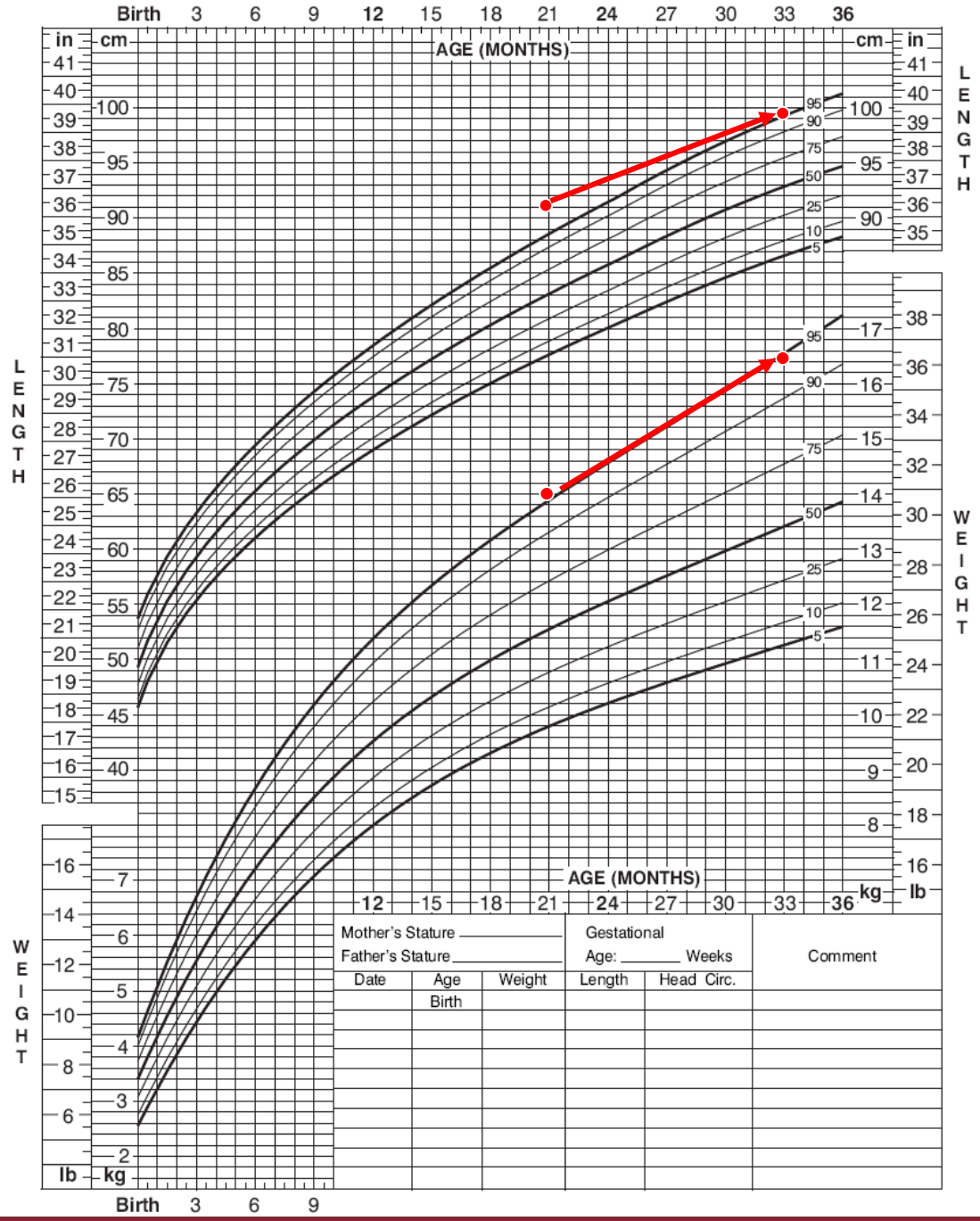
PERCENTILI



Esercizio: Utilizzando la tavola allegata, confrontare l'accrescimento staturo-ponderale di 2 sorelline nell'arco di un anno

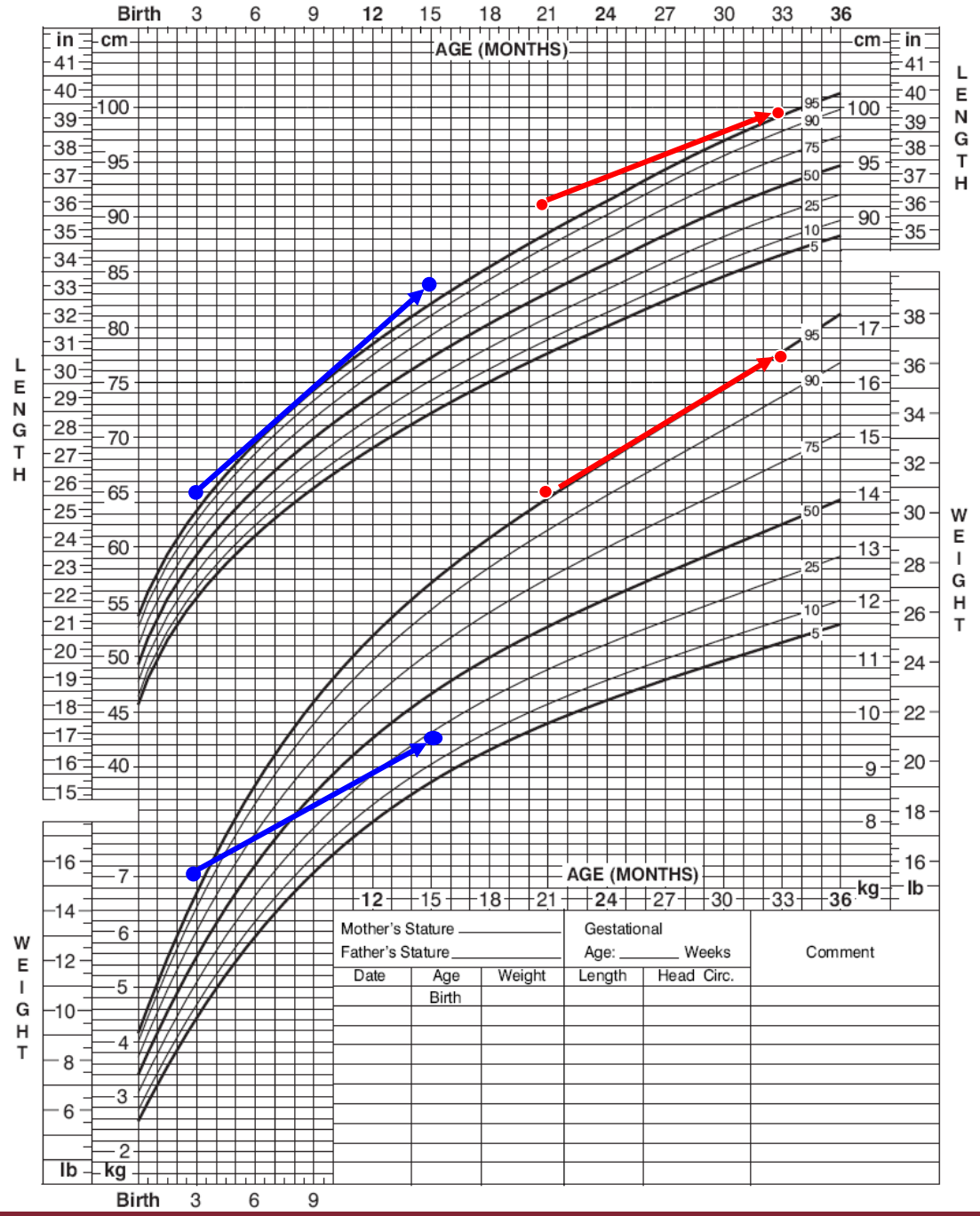
- La prima pesava 14 kg a **21 mesi** e 16.5 kg a **33 mesi**

- Nello stesso periodo la sua statura è aumentata da 91 cm a 100 cm



- Tra i **3** e i **15 mesi** di età il peso della seconda bambina è cresciuto da 7 a 9.5 Kg

- La sua statura è aumentata da 65 a 84 cm



centro di una distribuzione

dato un insieme di n elementi $\{x_1, x_2, \ldots x_N\}$

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

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

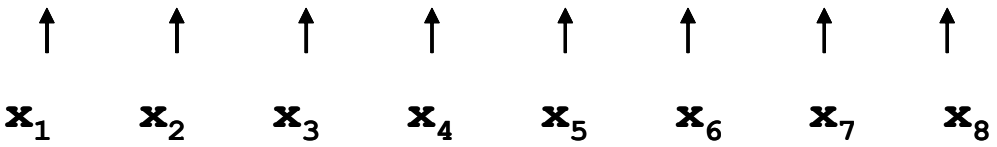
MEDIA ARITMETICA

La media aritmetica di un insieme di osservazioni è pari alla somma dei **valori** diviso il numero totale delle osservazioni

Formalmente: siano (x_1, x_2, … , x_n) le osservazioni della variabile X su un campione di *n* unità statistiche, allora

$$\bar{x} = \sum_{i=1}^{n} x_i \Big/ n = (x_1 + x_2 + .. + x_n)/n$$

esempio:
(8 osservazioni)

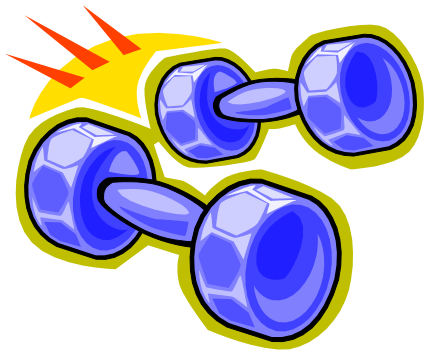


$$\bar{x} = (5+16+13+27+11+5+13+13)/8 = 103/8 = 12.9$$

MEDIA ARITMETICA PONDERATA - I

Se una variabile assume lo stesso valore in più unità statistiche, la media può essere calcolata moltiplicando quel valore per la frequenza con cui compare nella distribuzione

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{x_1 n_1 + x_2 n_2 + .. + x_k n_k}{n}$$



k = numero di valori che la variabile può assumere

x_i = valore assunto dalla variabile nel soggetto i-esimo

n_i = frequenza corrispondente al valore x_i

In formula

$$M(x) = \bar{x} = \frac{\sum_{i=1}^{K} x_i n_i}{\sum_{i=1}^{K} n_i}$$

Per distribuzioni semplici: $n_i = 1$

Esempio 1

Numero di figli	n_i	$x_i n_i$
0	8	0
1	14	14
2	20	40
3	6	18
4	4	16
5	2	10
Totale	54	98

$$\bar{x} = \frac{98}{54} = 1.8148 \cong 1.81$$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{x_1 n_1 + x_2 n_2 + .. + x_k n_k}{n}$$

esempio sulla media aritmetica ponderata:

$x_1 \longrightarrow 5$
$x_2 \longrightarrow 16$
$x_3 \longrightarrow 13$
$x_4 \longrightarrow 27$
$x_5 \longrightarrow 11$
$x_6 \longrightarrow 5$
$x_7 \longrightarrow 13$
$x_8 \longrightarrow 13$

la variabile può assumere 5 valori (k = 5)

x_i	n_i	$x_i n_i$
5	2	10
11	1	11
13	3	39
16	1	16
27	1	27
Totale	8	103

$$\bar{x} = (10 + 11 + 39 + 16 + 27)/8 = 103/8 = 12.9$$

Media per dati in tabella di frequenza: Esempio 1 (1)

Numero di sigarette	Soggetti
5	20
10	30
15	35
20	12
n	97

Interpretiamo le quantità nella tabella: ci sono 20 soggetti che fumano 5 sigarette al giorno, 30 che ne fumano 10, etc.

Per applicare la formula sopra scritta, dovrei riscrivere i dati sotto forma di lista ossia 5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,10,10,10,10,10,10,10,10,....

Media per dati in tabella di frequenza: Esempio 1 (2)

Possiamo calcolare la media a partire dalla tabella di frequenza nel seguente modo:

$$\bar{x} = \frac{\sum_{i=j}^{k} x_j n_j}{n} = \frac{(5 * 20) + (10 * 30) + (15 * 35) + (20 * 12)}{20 + 30 + 35 + 12} = 12.01$$

dove:

- $k = 4$ è il numero di classi o di modalità assunte dal carattere
- x_i è la modalità i-esima del carattere (ossia 5, 10, 15 e 20)
- n_i è la frequenza della modalità i (ossia 20, 30, 35 e 12)

xi	ni	xi*ni
0	18	0
1	27	27
2	31	62
3	19	57
4	5	20
totale	100	166

$$\bar{x} = \frac{166}{100} = 1.66$$

Moda=

Mediana=

Nel caso di variabili continue le cui modalità sono raggruppate in classi, per il calcolo della media si considera usualmente il valore centrale (medio) di ogni classe, ipotizzando l'equidistribuzione dei dati all'interno della classe stessa

Media per dati raggruppati in classi

Supponiamo di avere la distribuzione di frequenze di un carattere quantitativo X suddiviso in K classi.

Si può approssimare la media aritmetica del carattere con la seguente espressione

$$\bar{X}_n \simeq \frac{c_1 \cdot n_1 + \cdots + c_K \cdot n_K}{n} = \frac{1}{n} \sum_{j=1}^{K} c_j \cdot n_j$$

dove c_j è il valore centrale della classe j-esima (calcolato come media dei 2 valori estremi) e n_j è la corrispondente frequenza assoluta.

Ipotesi di **equidistribuzione delle frequenze** all'interno di ogni classe.

Media per dati raggruppati in classi: Esempio 1 (1)

Livello Colesterolo	Soggetti
$[80, 120)$	13
$[120, 160)$	150
$[160, 200)$	442
$[200, 240)$	299
$[240, 280)$	115
$[280, 320)$	34
$[320, 360)$	9
$[360, 400)$	5
Totale	1067

Come calcolare la media ???

Media per dati raggruppati in classi: Esempio 1 (2)

- Assumiamo che tutti i valori che rientrano in un determinato intervallo siano uguali al punto medio di quell'intervallo
- Nel nostro caso, ciò significa che assumiamo che le 13 persone della prima classe abbiano tutte colesterolo pari a $(80 + 120)/2 = 100$.
- Pertanto il problema si riduce a quello illustrato in precedenza, ossia possiamo riscrivere la tabella come

Valore centrale classe di colesterolo	Soggetti
100	13
140	150
180	442
220	299
260	115
300	34
340	9
380	5
Totale	1067

Media per dati raggruppati in classi: Esempio 1 (3)

Nel nostro caso:

$k = 8$ è il numero di classi o di modalità assunte dal carattere (nel nostro esempio 8)

c_i è il valore centrale della classe (calcolato come media dei 2 valori estremi)

n_i è la frequenza della classe i

- Nel nostro caso, allora la media la calcoliamo come

allora la media la calcoliamo come

$$\begin{aligned}
\bar{x} &= \frac{\sum_{j=1}^{k} c_j n_j}{n} \\
&= [(100 * 13) + (140 * 150) + (180 * 442) + (220 * 299) \\
&+ (260 * 115) + (300 * 34) + (340 * 9) \\
&+ (380 * 5)] / (13 + 150 + 442 + 299 + 115 + 34 + 9 + 5) \\
&= 199.3 mg/100 ml
\end{aligned}$$

Media per dati raggruppati in classi: Esempio 2

Distribuzione dei pazienti per classi di età

Classe di età (c_i)	**Frequenza**
[30-40] (35)	432
(40-50] (45)	2840
(50-60] (55)	1630
(60-70] (65)	781
oltre 70 (75)	93
	5776

$$\bar{X}_n = \frac{(35 \cdot 432) + (45 \cdot 2840) + (55 \cdot 1630) + (65 \cdot 781) + (75 \cdot 93)}{5776} = 50.26$$

NOTA: ultima classe è aperta. Abbiamo assunto come estremo superiore della classe un valore ragionevole (80 anni)

Se l’ipotesi di equidistribuzione non fosse accettabile (classi aperte…), come valore medio della classe si può scegliere un valore arbitrario ma plausibile,

oppure è necessario conoscere analiticamente i veri valori delle unità statistiche e tramite questi calcolare il vero valore medio della classe

Età al parto	n_i	Valore classe
Fino a 15	3	?
15 - 20	10	17.5
20 - 25	36	22.5
.....	.....	.....
.....	.....	.....
45 e oltre	8	?

peso	xi	ni	xi*ni
48.5 – 51.5	50	24	1200
51.5 - 54.5	53	36	1908
54.5 - 57.5	56	12	672
57.5 - 60.5	59	6	354
totale		78	4134

$$\bar{x} = \frac{4134}{78} = 53$$

Proprietà della media aritmetica

1 Il valore della media è sempre compreso tra il minimo e il massimo dei valori presi in esame

$$x_1 \leq \overline{x} \leq x_K$$

Data la seguente distribuzione di frequenza relativa alle stature di un collettivo di 82 ragazzi

Classe(cm)	n_i
150-154	2
155-159	6
160-164	11
165-169	18
170-174	25
175-179	13
180-184	7
Totale	82

Indicare quale valore rappresenta la media aritmetica:

a) 155.5
b) 181.3
c) 149.3
d) 174.2
e) 169.62

SCARTO E SCOSTAMENTO

Nota la media aritmetica, si possono calcolare

Gli scarti $s_i = x_i - \bar{x}$

Gli scostamenti $sh_i = x_i - h$

purché $h \neq \bar{x}$

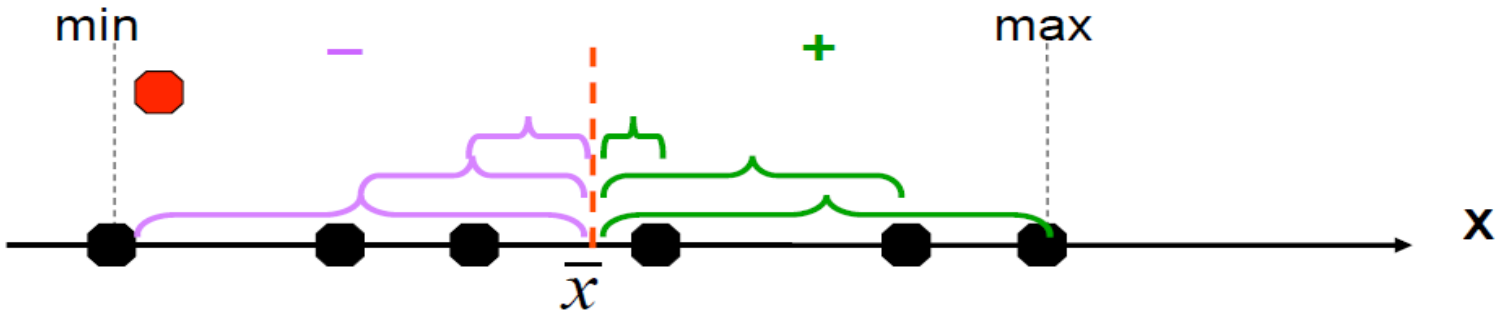
Proprietà della media aritmetica

2 La somma algebrica degli scarti è sempre uguale a 0

$$\sum_{i=1}^{K} s_i n_i = 0$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + .. + (x_n - \bar{x}) = 0$$

Principali proprietà della media



$$\min(x_i) \leq \bar{x} \leq \max(x_i)$$

La media è interna al range, ossia, è sempre compresa fra l'osservazione più bassa e quella più alta

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

La somma degli scarti dalla media è nulla: ossia, la media si colloca "al centro" dei valori osservati, bilanciando scarti positivi e scarti negativi

$$\text{dist} = \sqrt{\sum_{i=1}^{n} (x_i - C)^2}$$

Se misuriamo la distanza delle osservazioni da un valore C secondo questa misura globale, essa assume il minimo se C è la media aritmetica: ossia, la media aritmetica è il punto "globalmente meno distante" dalle osservazioni

(Altre medie (quadratica; geometrica; armonica) godono di altre proprietà, ma sono meno utili: le trascuriamo)

Proprietà della media aritmetica

3 La somma dei quadrati degli scarti è sempre minore della somma dei quadrati degli scostamenti

$$\sum_{i=1}^{K} s_i^{\,2} n_i < \sum_{i=1}^{K} s_{hi}^{\,2} n_i$$

Proprietà della media aritmetica

4 Se un collettivo è diviso in gruppi di cui sono note le medie, la media generale è data dalla media delle medie di gruppo ponderate con la numerosità di gruppo

MEDIA ARITMETICA PONDERATA - II

Date più medie e le singole frequenze con cui sono state calcolate, la media generale può essere calcolata come media ponderata delle medie

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + .. + \bar{x}_k n_k}{n_1 + n_2 + .. + n_k}$$

k = numero di gruppi

$\bar{x}_i$ = media aritmetica nel gruppo i-esimo

n_i = numerosità del gruppo i-esimo

$\bar{x}$ = media aritmetica complessiva

esempio: *valore medio dell'altezza nei maschi e nelle femmine matricole della Facoltà di Medicina (A.A. 2015/2016)*

Sesso	n_i	$\bar{x}_i$
maschi	34	177
femmine	91	166.1
Totale	125	

$$\bar{x} = \frac{177*34+166.1*91}{125} = 169.1$$

QUALE MISURA DI POSIZIONE UTILIZZARE?

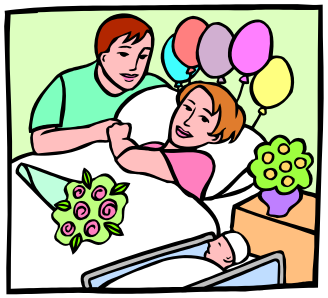
TIPO DI VARIABILE	OPERAZIONI CONSENTITE	MODA	MEDIANA	MEDIA
nominale	= ≠	Sì	No	No
ordinale	= ≠ < >	Sì	Si	No
quantitativa	= ≠ < > − + (/ *)	Sì	Sì	Sì

CONFRONTO TRA LE MISURE DI POSIZIONE PER UNA VARIABILE QUANTITATIVA

	MODA	MEDIANA	MEDIA ARITMETICA
PRO	**Buona misura quando un valore ha una frequenza relativa molto elevata**	**Buona misura con distribuzioni asimmetriche** *(es. tempo di sopravvivenza)* **e in presenza di dati estremi**	**Buona misura con distribuzioni simmetriche** *(es. molti parametri biologici)* **Facile da trattare matematicamente** **Utilizza tutta l'informazione contenuta nei dati**
CONTRO	**Dipende dal raggruppamento arbitrario dei dati** **Varia molto da campione a campione**	**Difficile da trattare matematicamente** **Non tiene conto della grandezza delle singole osservazioni**	**E' inaffidabile in caso di distribuzioni asimmetriche**

CONFRONTO TRA LE MISURE DI POSIZIONE PER UNA VARIABILE QUANTITATIVA

esempio:
Supponiamo di avere le Degenze Ospedaliere di 9 individui *(espresse in giorni)*



CAMPIONE **3 4 4 4 5 6 8 12 95**

Moda = **4**

Mediana = **5**

Media ≈ **16** (senza *outliers* sarebbe circa **6**)

La media aritmetica è poco "robusta" in presenza di **valori anomali** (outliers)!

Robustezza della mediana

La mediana non cambia o cambia di poco (è "robusta") in presenza di alcuni dati molto estremi (ad es. con alcuni valori molto alti rispetto agli altri)

Vediamo per esempio che succede se nel campione precedente i due soggetti più alti sono ancora più alti:

173 155 162 165 167 ~~175~~ 171 169 164 ~~178~~ 156 158 166 $\quad \bar{x} = 166.1$

189 210 $\quad \bar{x} = 169.6$

155 156 158 162 164 165 **166** 167 169 171 173 189 210

1 2 3 4 5 6 **7** 8 9 10 11 12 13

6 osservazioni (50%) — 6 osservazioni (50%)

mediana = 166

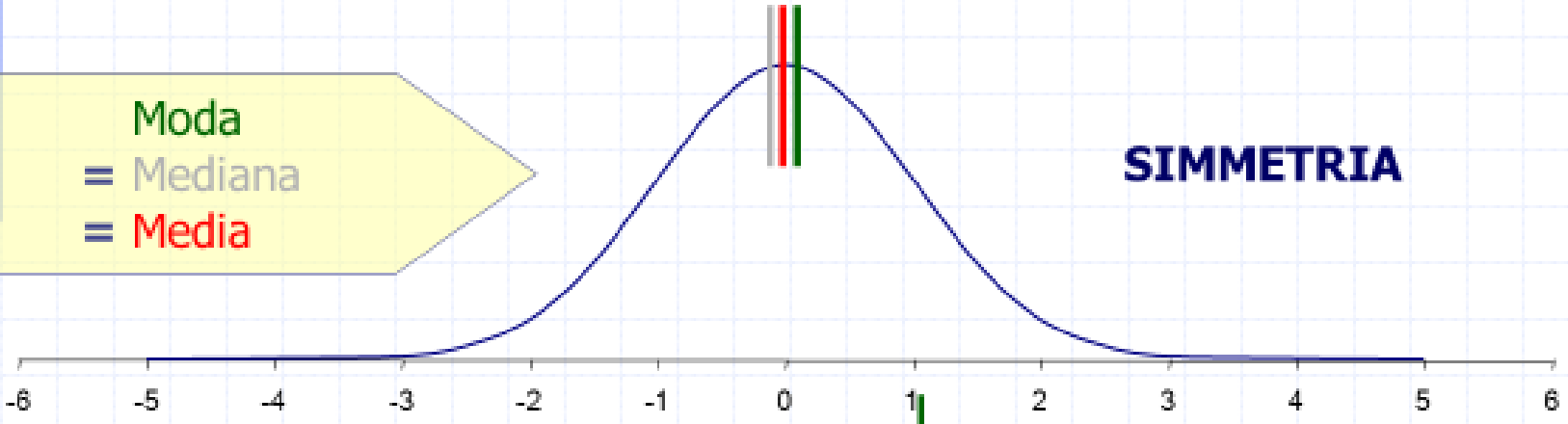
→ La mediana non cambia poichè l'ordinamento delle prime n osservazioni non cambia (invece la media cambia perché l'ammontare totale cambia)

RELAZIONE TRA MODA, MEDIANA E MEDIA ARITMETICA

ASIMMETRIA POSITIVA

Moda
< Mediana
< Media

Moda
= Mediana
= Media



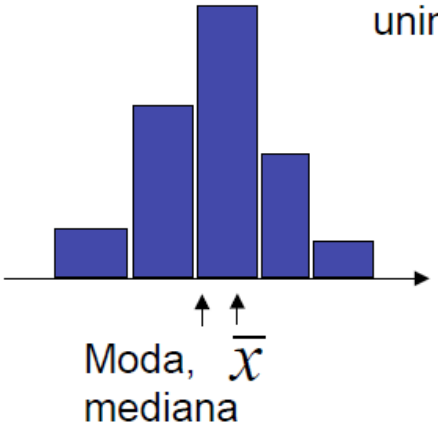
SIMMETRIA

ASIMMETRIA NEGATIVA

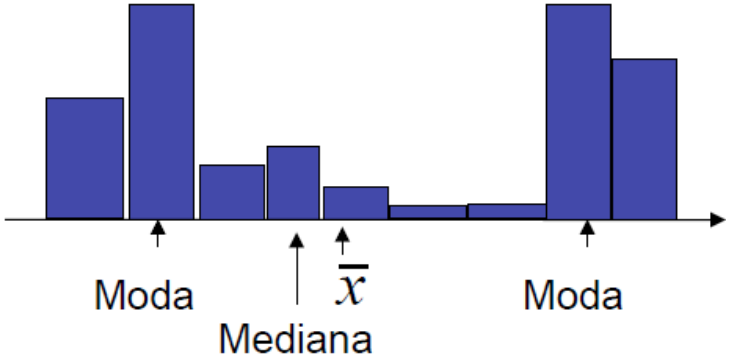
Moda
> Mediana
> Media

Forma della distribuzione e indici

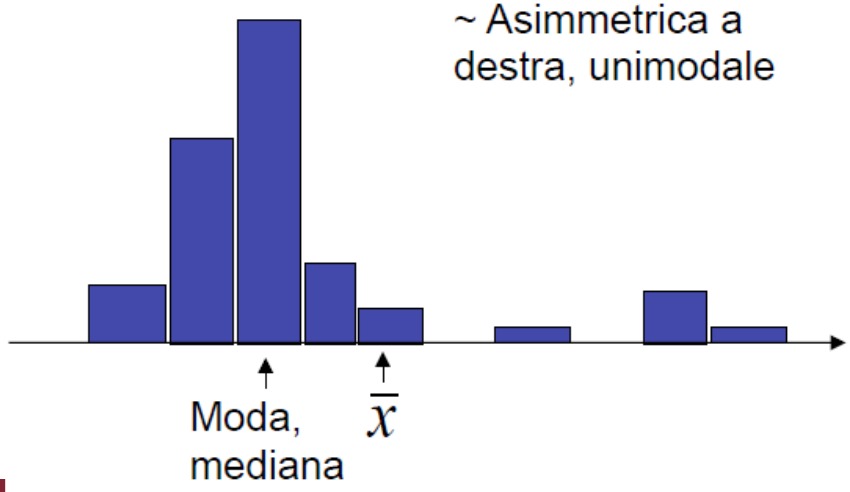
~ Simmetrica, unimodale



~ Simmetrica, bimodale (2 sottopopolazioni?)

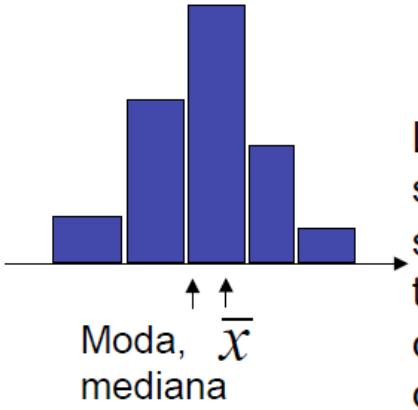


~ Asimmetrica a destra, unimodale

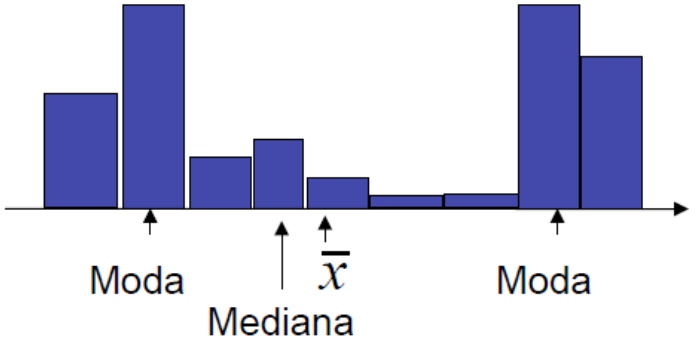


La forma della distribuzione è individuabile (in maniera grossolana) a partire dagli indici sintetici – e viceversa.

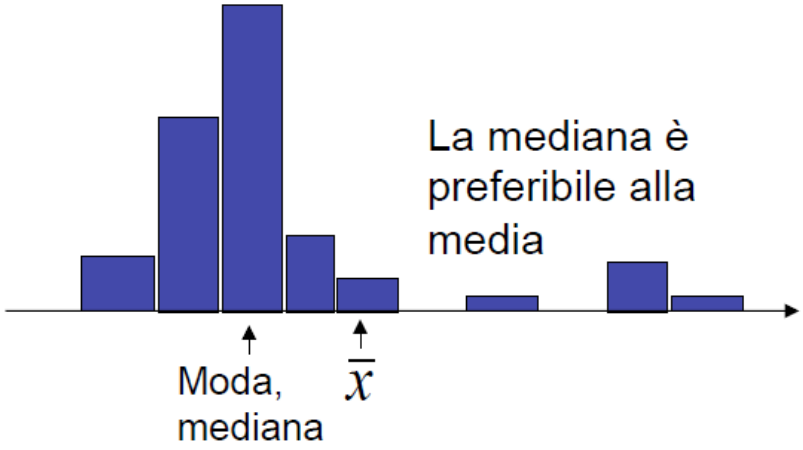
Appropriatezza degli indici



La media è una sintesi soddisfacente, tende a coincidere con la mediana, e con la moda



E' opportuno rimarcare la bimodalità: ne' media ne' mediana sono sintesi soddisfacenti



La mediana è preferibile alla media

quale misura di posizione usare?

- Il proprietario di una ditta afferma "Lo stipendio mensile nella nostra ditta è 2.700 euro"
- Il sindacato dei lavoratori dice che “lo stipendio medio è di 1.700 euro”.
- L'agente delle tasse dice che “lo stipendio medio è stato di 2.200 euro”.

Queste risposte diverse sono state ottenute tutte dai dati della seguente tabella.

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

Media aritmetica= € 2.700

Mediana = € 2.200

Moda = € 1.700

interpretazione delle misure di posizione

- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro

- La **moda** ci dice che la paga mensile più comune è di 1.700 euro

- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valor**i e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.

- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.

- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.