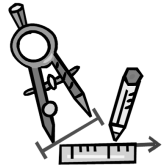
LEZIONI DI STATISTICA MEDICA

Indici di posizione

Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

" un qualsiasi insieme di dati porta in sè una certa quantità di informazione "

OBIETTIVO:
riassumere tutta l'informazione possibile in modo SINTETICO ed EFFICACE



diversi strumenti e possibilità offerti dalla statistica

STATISTICHE DI BASE

Una serie di dati numerici è compiutamente descritta da tre **PROPRIETÀ PRINCIPALI:**

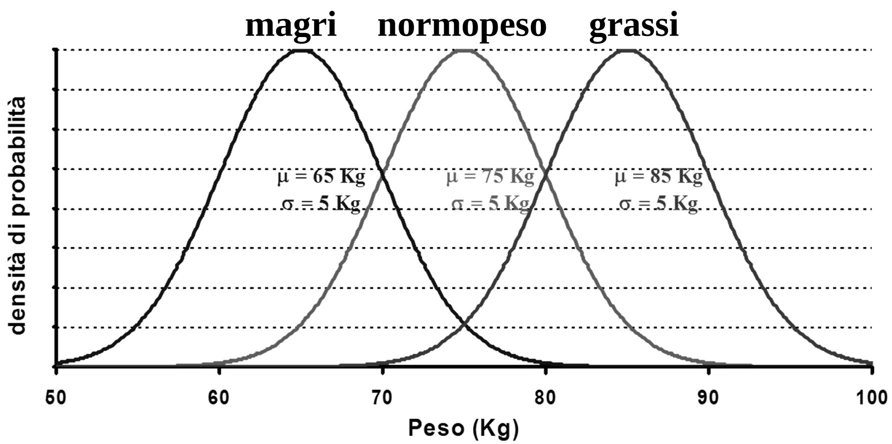
1. **La tendenza centrale o posizione**
2. **La dispersione o variabilità**
3. **La forma**

Queste **misure descrittive sintetiche** sono chiamate:

STATISTICHE ($\bar{x}$, s, p) → quando sono calcolate su un CAMPIONE di dati (si esprimono con lettere dell'alfabeto latino)

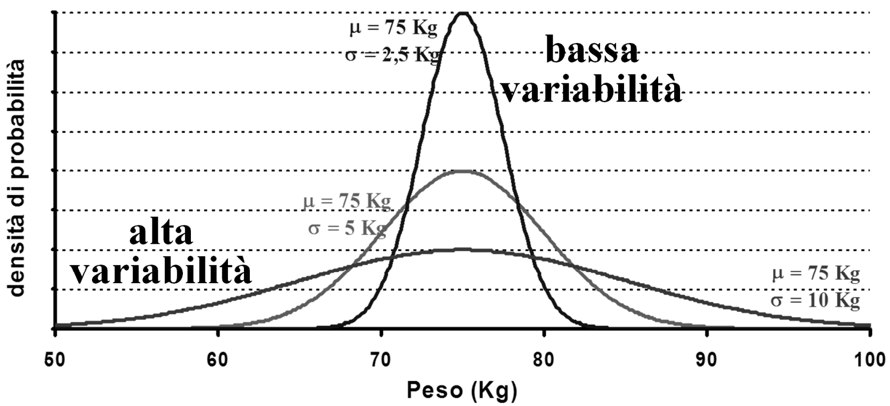
PARAMETRI (μ, σ, π) → quando descrivono la POPOLAZIONE (si esprimono con lettere dell'alfabeto greco)

Posizione differente, uguale dispersione



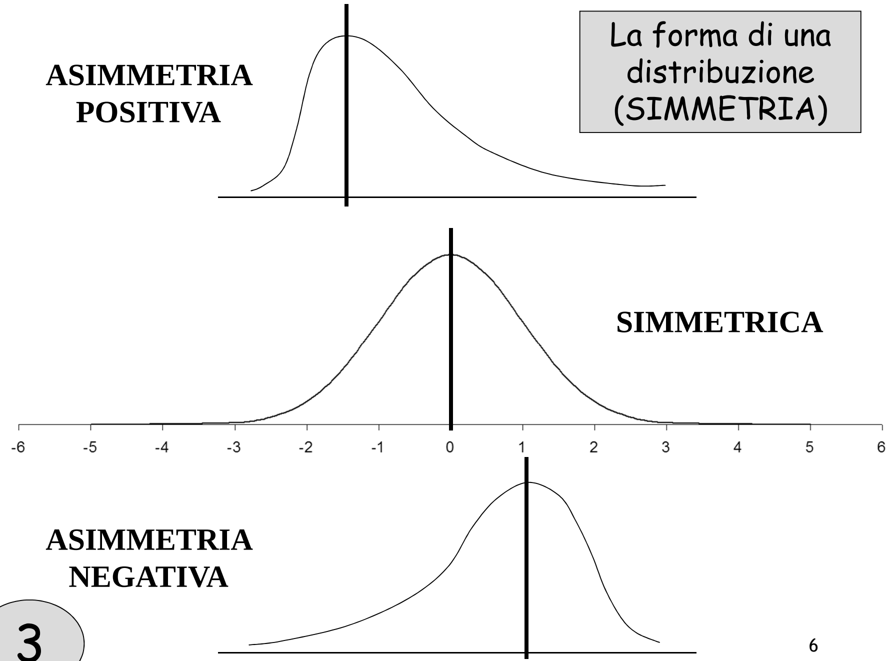
1

Dispersione differente, uguale posizione



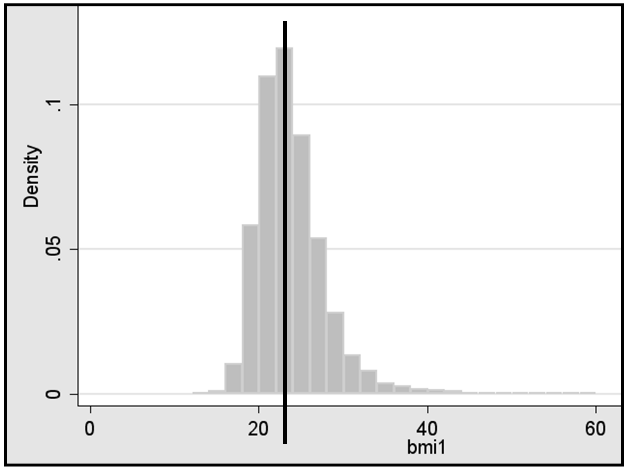
2

La forma di una distribuzione (SIMMETRIA)



3

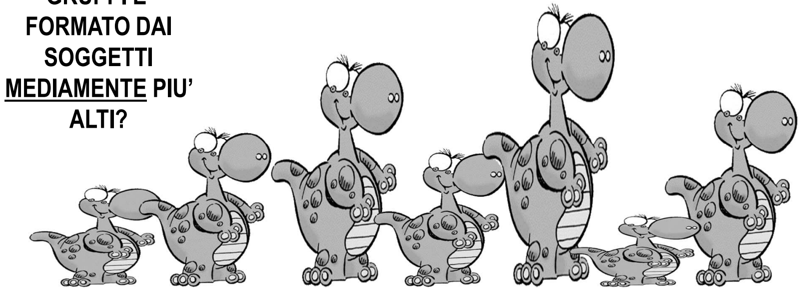
Il BMI in ECRHS
(European Community Respiratory Health Survey)

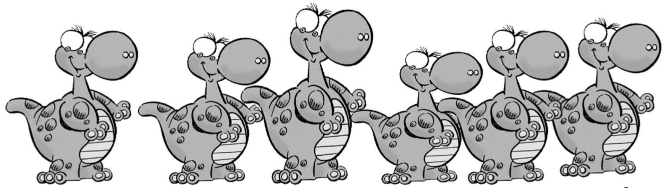


INDICI DI POSIZIONE

QUALE DEI DUE GRUPPI E' FORMATO DAI SOGGETTI <u>MEDIAMENTE</u> PIU' ALTI?

la variabile d'interesse è l'ALTEZZA





INDICI DI POSIZIONE

(measures of location or *central tendency)*

1. MODA
2. MEDIA → aritmetica / armonica / geometrica
3. MEDIANA

MODA

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

in statistica non è diverso

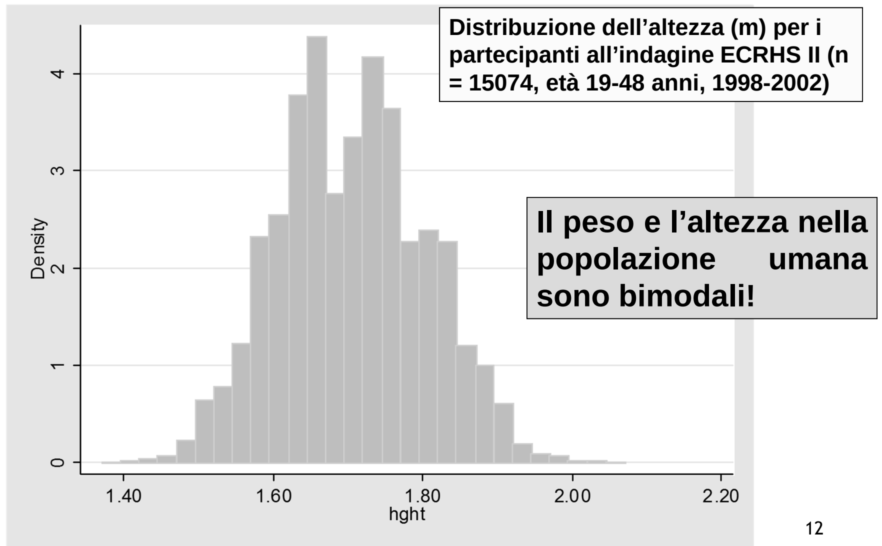
Si definisce moda (*classe modale*) di un insieme di dati o di una distribuzione di frequenza la <u>modalità / il valore</u> (*l'intervallo di classe*) della variabile <u>cui corrisponde la massima frequenza</u>.

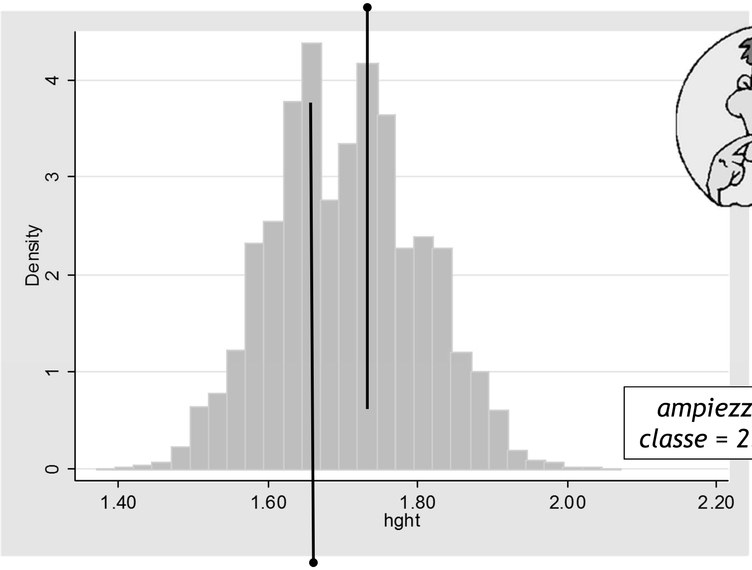
esempio:
X = tipo di parto
(50 neonati)

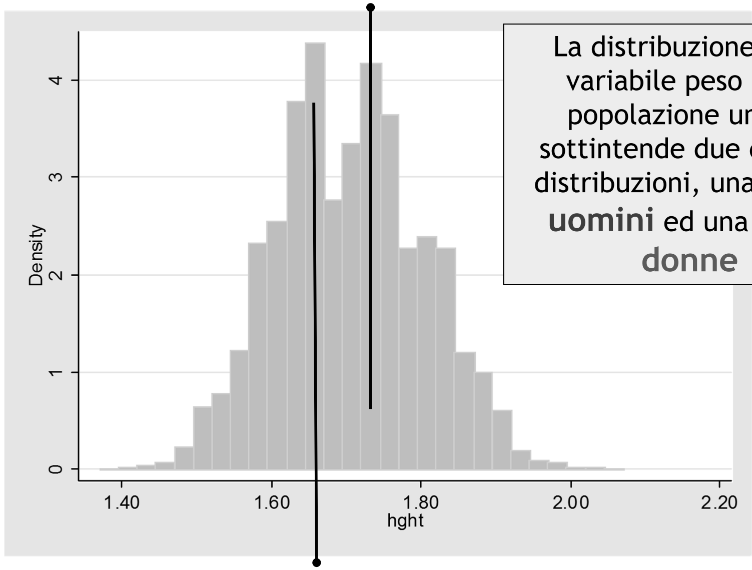
modalità x_i	frequenza assoluta n_i	frequenza relativa p_i	frequenza relativa percentuale p_i (%)
normale	35	0.70	70%
forcipe	1	0.02	2%
cesareo	14	0.28	28%
TOTALE	50	1.00	100%

MODA

MA LA MODA E' SEMPRE UNA SOLA?







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

MEDIANA

- Il valore centrale di una serie **ORDINATA** di dati
- Le osservazioni vengono separate dal valore mediano in due parti numericamente uguali
- Mediana (Me) è sinonimo di **50-esimo percentile** o di **II quartile**

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

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

es. sulla mediana

campione di 5 unità

variabile d'interesse = altezza


165 cm


150 cm


180 cm


50 cm

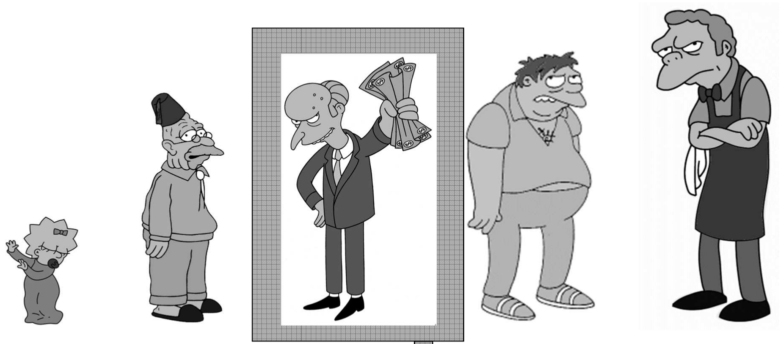

155 cm

1. metto le unità in ORDINE crescente di altezza

es. sulla mediana

campione di **5** unità

variabile d'interesse = altezza



50 cm | 150 cm | 155 cm | 165 cm | 180 cm

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

50 cm | 150 cm | 155 cm | 165 cm | 180 cm

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

formalmente:

n è *dispari* ⇒ $Me = x_{[(n+1)/2]} = x_{(5+1/2)} = x_3$

50 cm | 150 cm | 155 cm | 165 cm | 180 cm

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

esercizio
(moda e mediana)

1. VARIABILE QUALITATIVA

2. VARIABILE QUANTITATIVA

ESERCIZIO - I

I dati seguenti si riferiscono all'abitudine al fumo in un campione di 168 soggetti senza bronchite cronica di età 20-44 anni:

X = abitudine al fumo **x_i = non fumatore**
ex fumatore
moderato fumatore
forte fumatore

	frequenza assoluta
non fumatore	74
ex fumatore	37
moderato fumatore	34
forte fumatore	23
Totale	168

E' possibile determinare la moda e la mediana di questa distribuzione di frequenza? Se la risposta è affermativa calcolatele.

SOLUZIONE - I

	frequenza assoluta	assoluta cumulata
non fumatore	74	74
ex fumatore	37	111
moderato fumatore	34	145
forte fumatore	23	168
Totale	168	

CALCOLO DELLA MODA

massima frequenza assoluta

MODA:
non fumatore

SOLUZIONE - I

	frequenza assoluta	assoluta cumulata
non fumatore	74	74
ex fumatore	37	111
moderato fumatore	34	145
forte fumatore	23	168
Totale	168	

CALCOLO DELLA MEDIANA

per le osservazioni con rango 84 e 85 la variabile assume questa modalità

MEDIANA:
ex fumatore

pari

le osservazioni centrali hanno rango:

- **n/2 = 168/2 = 84**
- **n/2 + 1 = 84 + 1 = 85**

ESERCIZIO - II

I dati seguenti si riferiscono al livello di emoglobina (X) in g/100 ml misurato in un campione di 70 donne:

1. Determinate la mediana della distribuzione (*dati individuali*).

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	15
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

SOLUZIONE - II

n = 70 $\rightarrow$ pari

le osservazioni centrali hanno rango:

- n/2 = 70/2 = 35
- n/2 + 1 = 70/2 + 1 = 36

osservazione con rango 35

osservazione con rango 36

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	15
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

MEDIANA = (11.8 + 11.9) / 2 =
= 11.85 g/100mL

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

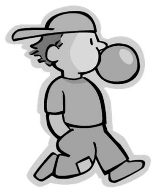
I PERCENTILI PIU' NOTI:

- 25° → 1° QUARTILE
- 50° → 2° QUARTILE o MEDIANA
- 75° → 3° QUARTILE

3° QUARTILE – 1°QUARTILE = **DIFFERENZA INTERQUARTILICA**

ESEMPIO:

Un ragazzo ha la glicemia a 90 mg/dl. Nella sua scuola ci sono **700** ragazzi. Se ordiamo la glicemia in ordine crescente questo ragazzo occupa la posizione **500** (**rango assoluto**).



PIERINO

Qual è il rango percentilico?

RangoPercentilico = RangoAssoluto/ (n+1) * 100

500/(700+1) * 100 = 0,713 * 100 = 71,3 ~ **71**

Il valore di 90 mg/dl corrisponde al 71-esimo percentile

CALCOLO DEL K-ESIMO PERCENTILE - 1

(1. Dati **individuali** disponibili)

1. Si individua il rango assoluto corrispondente al ***k*-esimo percentile**

Rango Assoluto = (n + 1) * *k* / 100

2. Si riporta il VALORE dell'osservazione cui corrisponde quel determinato rango

Esempio

la mediana (*50-esimo percentile, k* = ***50***) di un campione di 89 individui ha **rango**:

(89 + 1) * ***50*** / 100 = 45

il *50*-esimo percentile sarà <u>il valore</u> osservato per la variabile di interesse nell'unità statistica con rango 45

CALCOLO DEL K-ESIMO PERCENTILE - 2

(2. Dati disponibili **in classi** sotto forma di tabella di frequenza)

<u>Nel caso **ad esempio** del 50-esimo percentile (mediana):</u>

- Si individua la CLASSE MEDIANA, ovvero la classe in cui la frequenza relativa cumulativa supera o coincide con il 50%
- esiste una formula per ottenere a questo punto una stima accurata del valore della mediana (si procede per *interpolazione lineare*).

Per gli scopi del corso è sufficiente, partendo dai dati raggruppati in classi, costruire la curva delle frequenze cumulate (ogiva) e identificare graficamente il valore della mediana (si veda diapositiva successiva)

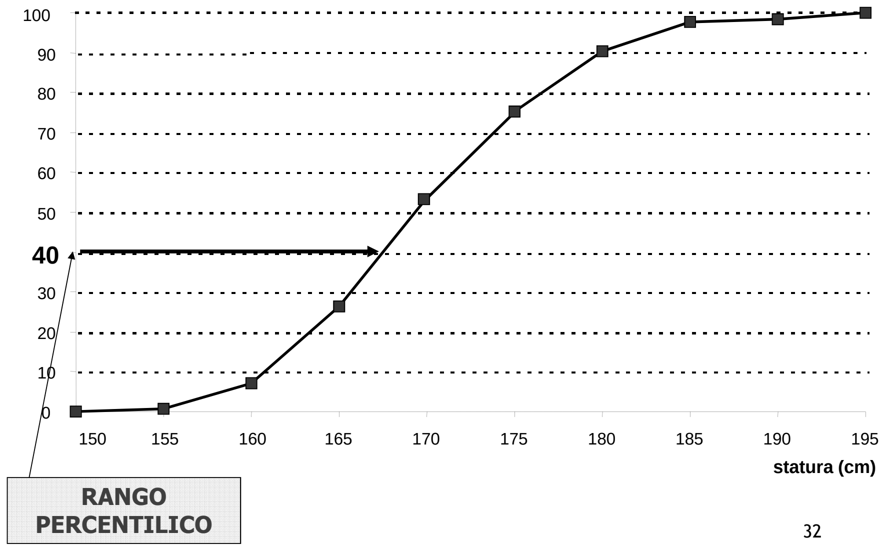
CALCOLO DEL K-ESIMO PERCENTILE - 3

(3. Dati non disponibili, si dispone solamente della **rappresentazione grafica** della frequenza relativa cumulata)

1. Sull'asse delle **ordinate (Y)**, dove è rappresentata la *frequenza relativa cumulata*, si individua il punto corrispondente al **rango percentilico (k)**

2. da qui si traccia una <u>linea orizzontale</u>, che intersechi la linea spezzata (ogiva), che rappresenta l'andamento della frequenza relativa cumulata

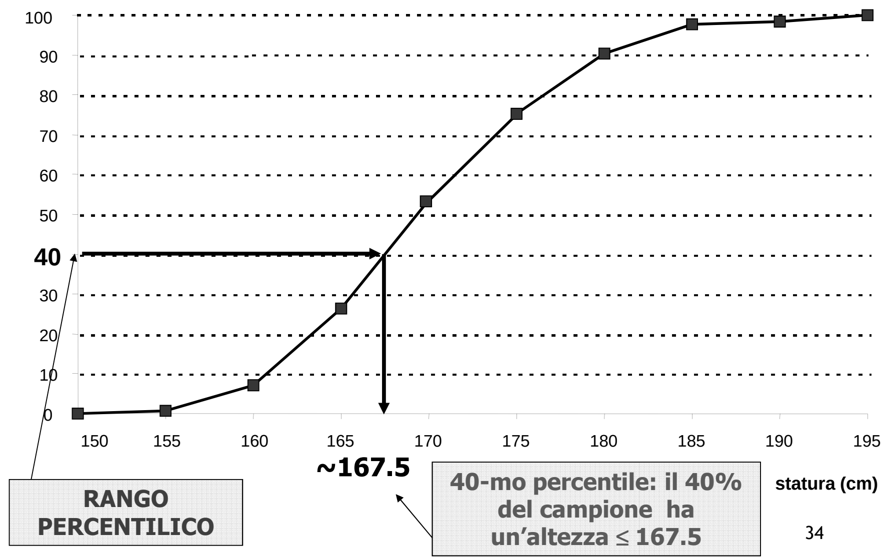
Esempio: calcolo del **40-esimo percentile**



3. dal punto d'intersezione così individuato, si traccia una <u>linea verticale</u> fino all'intersezione con **l'asse delle ascisse (X)**, che rappresenta i valori della variabile oggetto dello studio

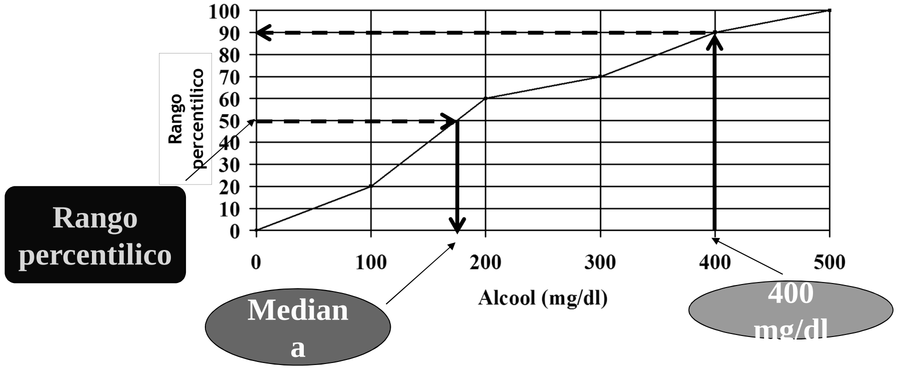
4. **il valore della variabile in corrispondenza del punto d'intersezione con le X rappresenta il k-esimo percentile**

Esempio: calcolo del **40-esimo percentile**



<u>Un'applicazione pratica:</u>

1. si vuole determinare i livelli ematici di alcool, in un gruppo di 250 soggetti (la cui distribuzione di frequenza cumulata è rappresentata in figura), in corrispondenza del **50-esimo percentile**.
2. si vuole conoscere a che percentile corrisponde un livello ematico di alcol di **400 mg/dl** in questo campione



1. La **mediana** è circa pari a **175 mg/dl**
2. **400 mg/dl** corrisponde circa al **90-esimo percentile**

ESERCIZIO

I dati seguenti si riferiscono al livello di emoglobina (X) in g/100 ml misurato in un campione di 70 donne:

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	15
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

2. Raggruppate i dati in intervalli di ampiezza 1 g/100 ml.
3. Determinate la moda e la mediana della distribuzione (*dati raggruppati in intervalli di classe*).

SOLUZIONE

CLASSE	FREQUENZA ASSOLUTA	FREQUENZA CUMULATA ASSOLUTA	FREQUENZA CUMULATA RELATIVA %
[9-10)	4	4	(4/70)*100 = 5.7
[10-11)	14	4+14 = 18	(18/70)*100 = 25.7
[11-12) ←	$\boxed{19}$	18+19 = 37	(37/70)*100 = 52.8
[12-13)	14	37+14 = 51	(51/70)*100 = 72.8
[13-14)	13	64	91.4
[14-15]	6	70	100.0
TOTALE	70		

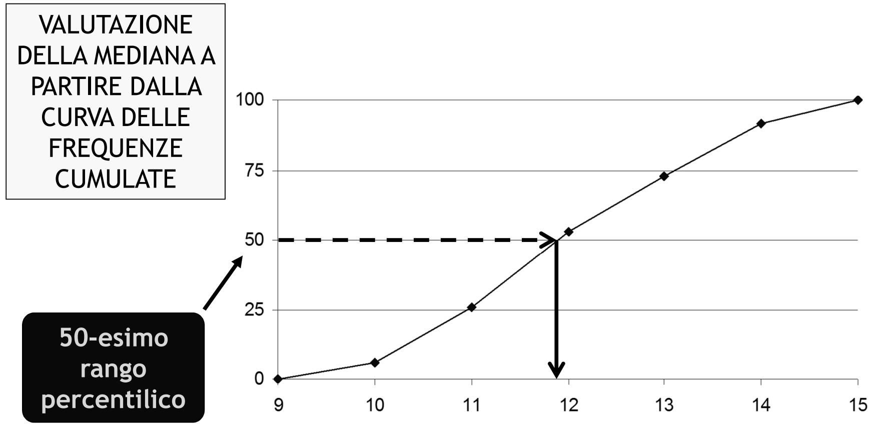
CLASSE MODALE = [11-12) g/100 mL

CLASSE MEDIANA:

1. **le osservazioni centrali hanno rango: 35 e 36**
2. **classe mediana: [11-12)**

SESM

SOLUZIONE



La mediana (50-esimo percentile) è pari a circa 11.9 g/100 mL

SESM

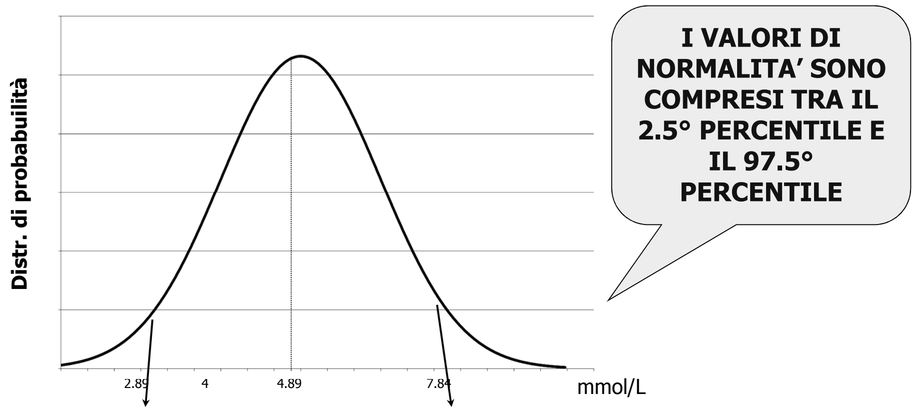
ESEMPI DELL'UTILITA' DEI PERCENTILI NELLA PRATICA CLINICA

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

SESM

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:



2.5° PERCENTILE = 2.89 97.5° PERCENTILE = 7.84

SESM

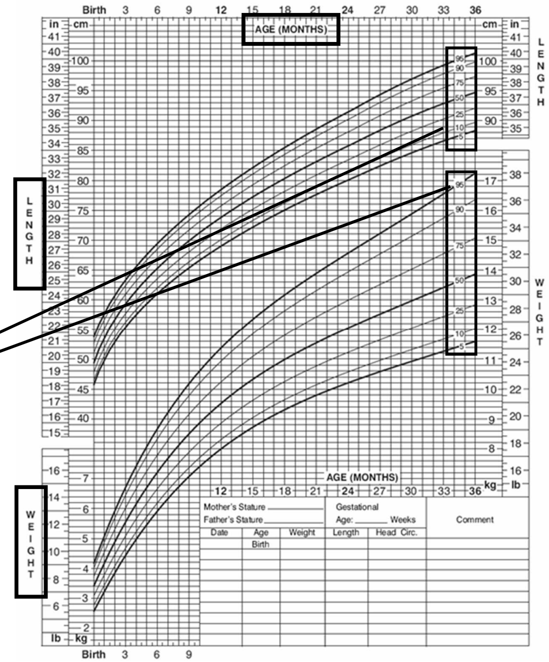
LE CURVE DI ACCRESCIMENTO

- 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

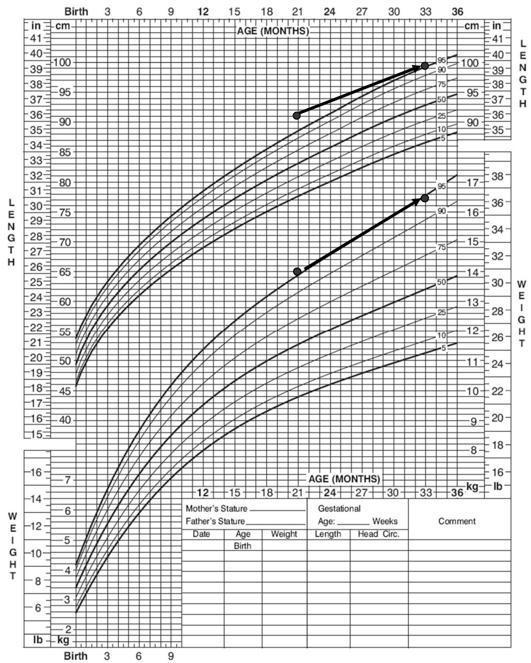
RAPPRESENTAZIONE DEI PERCENTILI DI PESO E STATURA PER ETA'

PERCENTILI

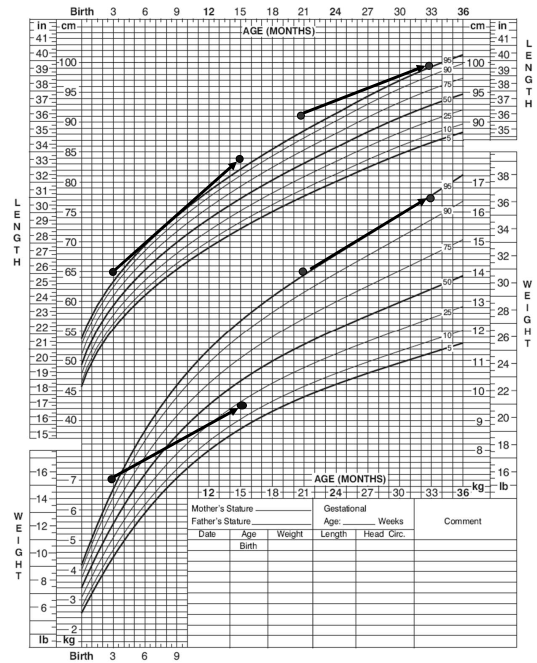


Esercizio: Utilizzando la tavola allegata, confrontare l'accrescimento staturo-ponderale di 2 sorelline nell'arco di un anno

- 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



MEDIA ARITMETICA

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

<u>Formalmente</u>: 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)

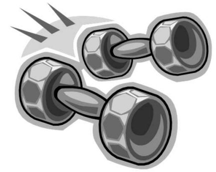
5	16	13	27	11	5	13	13
↑	↑	↑	↑	↑	↑	↑	↑
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

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

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

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



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

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

k = numero di valori che la variabile può assumere

x_i = valore assunto dalla variabile nel sogg. i-esimo

n_i = frequenza corrispondente al valore x_i

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

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

Nel caso di variabili quantitative continue, i dati sono spesso organizzati in classi per una migliore capacità descrittiva

Distribuzione del FEV_1 (cl/s) in 54 soggetti maschi di età 20-44 anni (indagine *European Community Respiratory Health Survey* - ECRHS)

	n_i
[223 – 270.25)	0
[270.25 – 317.5)	3
[317.5 – 364.75)	9
[364.75 – 412)	10
[412 – 459.25)	14
[459.25 – 506.5)	8
[506.5 – 553.75)	7
[553.75 – 601]	3
TOTALE	54

Distribuzione del FEV_1 (cl/s) in 54 soggetti maschi di età 20-44 anni (ECRHS)

Per il calcolo della media ponderata le osservazioni di una classe si considerano coincidenti con il suo valore centrale

punto centrale della 1ª classe: (223+270.25)/2 = 246.625

	x_i	n_i	$x_i n_i$
[223 – 270.25)	246.625	0	0
[270.25 – 317.5)	293.875	3	881.625
[317.5 – 364.75)	341.125	9	3070.125
[364.75 – 412)	388.375	10	3883.750
[412 – 459.25)	435.625	14	6098.750
[459.25 – 506.5)	482.875	8	3863
[506.5 – 553.75)	530.125	7	3710.875
[553.75 – 601]	577.375	3	1732.125
TOTALE		54	23240.25

$$\bar{x} = (x_1 n_1 + x_2 n_2 + .. + x_8 n_8)/n = 23240.25/54 = 430.375 \text{ cl/s}$$

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. 95/96)*

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

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

La media aritmetica gode di diverse proprietà, le due principali dal punto di vista applicativo sono legate al concetto di scarto:

PRIMA PROPRIETA' DELLA MEDIA ARITMETICA

la somma algebrica degli scarti delle osservazioni dalla loro media aritmetica è pari a zero

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \underbrace{(x_1 - \bar{x})}_{\text{scarto (distanza) della prima osservazione dalla media}} + (x_2 - \bar{x}) + .. + (x_n - \bar{x}) = 0$$

Verifica empirica della prima proprietà:

FEV_1	x_i	n_i	$(x_i - \bar{x})n_i$
[223 – 270.25)	246.625	0	0
[270.25 – 317.5)	293.875	3	- 409.5
[317.5 – 364.75)	341.125	9	- 803.25
[364.75 – 412)	388.375	10	- 420
[412 – 459.25)	435.625	14	73.5
[459.25 – 506.5)	482.875	8	420
[506.5 – 553.75)	530.125	7	698.25
[553.75 – 601]	577.375	3	441
TOTALE		54	0

scarti negativi

scarti positivi

$$-409.5 = (293.875 - 430.375) \times 3$$

$\bar{x}$ = 430.375 cl/s

media ponderata

la somma algebrica degli scarti delle osservazioni dalla loro media aritmetica è pari a zero

ESERCIZIO

I dati seguenti si riferiscono al livello di emoglobina (X) in g/100 ml misurato in un campione di 70 donne:

CLASSE	FREQUENZA ASSOLUTA (n_i)
[9-10)	4
[10-11)	14
[11-12)	19
[12-13)	14
[13-14)	13
[14-15]	6
TOTALE	70

3. Determinate la media della distribuzione.

SESM

SOLUZIONE-III

CLASSE	FREQUENZA ASSOLUTA (n_i)
[9-10)	4
[10-11)	14
[11-12)	19
[12-13)	14
[13-14)	13
[14-15]	6
TOTALE	70

$$\bar{x} = (x_1 n_1 + x_2 n_2 + .. + x_6 n_6)/n = 841/70 = 12.0 \text{ g/100 mL}$$

QUALE MISURA DI POSIZIONE UTILIZZARE?

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

SESM

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

SESM

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)!

RELAZIONE TRA MODA, MEDIANA E MEDIA ARITMETICA