Lezione 4ª- Misure di dispersione o di variabilità

Abbiamo visto che la media è una misura della localizzazione centrale della distribuzione (il centro di gravità). Popolazioni con la stessa media possono avere un grado molto diverso di variazione dei dati. Una maniera per esprimere questa variazione è quello di utilizzare la media come punto di riferimento di ciascun valore, cioè di calcolare la deviazione di ciascun dato dalla media (il suo "scarto" dalla media). Le deviazioni saranno numeri positivi per tutti i valori al di sopra della media e numeri negativi per tutti i valori al di sotto della media. Se noi sommassimo queste deviazioni il risultato sarebbe 0 (i valori positivi sarebbero elisi dai valori negativi). Quest'approccio non ci consentirebbe pertanto di ottenere una misura della variabilità dei dati. Il problema si risolve elevando al quadrato le deviazioni dalla media (il quadrato di un numero negativo è un numero positivo). Se sommiamo i quadrati delle deviazioni (o "scarti") dalla media e dividiamo questa somma per il numero delle osservazioni otteniamo la **deviazione quadratica media** (o scarto quadratico medio) o **varianza**. Per riportare i valori all'unità di misura di partenza possiamo estrarre la radice quadrata della varianza. La radice quadrata della varianza è la misura di distribuzione più usata ed è definita **deviazione standard**.

Un altro modo di esprimere la variabilità di una distribuzione è quella di riferirsi al range di una distribuzione (il valore minimo e il valore massimo). Il **range** dipende esclusivamente dai valori estremi, perciò se il campione di dati è piccolo esso può dare una stima erronea del range della popolazione (questo perché i valori estremi sono rari e possono non essere rappresentati in un piccolo campione).

Esempio 10. Si considerino inizialmente, le seguenti due distribuzioni di valori riferiti all'età di 10 individui

I gruppo	II gruppo	III gruppo
20aa	10aa	35aa
30aa	25aa	37aa
40aa	40aa	40aa
50aa	55aa	43aa
60aa	70aa	45aa
R=40aa	R=60aa	R=10aa

L'età media (media aritmetica) è pari a 40 anni per tutti gruppi, ma nel secondo i dati sono più "dispersi" attorno alla media.

Pertanto accanto ai valori medi vanno introdotti anche indici di misura della VARIABIITA' (O DISPERSIONE) dei dati.

Le misure di dispersione più usate sono:

1. campo di variazione (range);
2. devianza;
3. varianza;
4. deviazione standard;
5. coefficiente di variazione (indice di variabilità relativa);
6. differenza interquartile.

Campo di Variazione o Range

$$R = X_{max} - X_{min}.$$

Limiti del campo di variazione

- è troppo influenzato dai valori estremi;
- tiene conto dei due soli valori estremi, trascurando tutti gli altri.
- tende ad aumentare con l'aumento del numero di osservazioni.

Occorre allora un indice di dispersione che consideri tutti i dati (e non solo quelli estremi), confrontando questi con il loro valor medio.

Tuttavia va ricordato che: $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

Si potrebbe calcolare la somma dei valori assoluti: $\sum_{i=1}^{n} |x_i - \bar{x}|$, ma tale quantità è difficile da trattare matematicamente.

Un indice alternativo è quello di considerare la **somma dei quadrati degli scarti dalla media aritmetica = DEVIANZA =** $\sum_{i=1}^{n} (x_i - \bar{x})^2$

Esempio 5'. Valori del tasso glicemico in 10 soggetti

x_i (glicemia mg/100cc)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
103	+8	64
97	+2	4
90	-5	25
119	+24	576
107	+12	144
71	-24	576
94	-1	1
81	-14	196
92	-3	9
96	+1	1
$\bar{x} = 95$	$\sum_{i=1}^{10} \lvert x_i - \bar{x} \rvert = 94$	$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1596$

La quantità 1596 esprime la ***Devianza della distribuzione*** (**Dev**).

Il limite della Devianza come misura di dispersione è quello di aumentare con il numero di osservazioni. Per ottenere una misura che non dipenda dalla numerosità si può dividere la devianza per il numero n. di dati, ottenendo la ***Varianza***:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{1596}{10} = 159.60\ (mg/100\ cc)^2$$

In pratica il denominatore n è quasi sempre sostituito da (n-1) in modo da ottenere una stima corretta della dispersione della variabile nella popolazione da cui il campione in esame è stato estratto.

$$\mathbf{s^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n - 1} = \frac{1596}{9} = 177.3 (mg/100cc)^2$$

Il limite della Varianza come misura di dispersione è quella di avere una unità di misura espressa al quadrato rispetto all'unità di misura originale, per cui si utilizza la ***Deviazione Standard*** (D.S. o S.D.):

$$s = D.S. = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n - 1}} = \sqrt{177.3} = 13.32 (mg/100\ cc).$$

Essa indica quanto, in media, ciascun elemento si discosta dalla media aritmetica.

La Deviazione Standard è l'indice di variabilità più "reale" e, quindi, più utilizzato

La Deviazione Standard per distribuzioni di frequenza: assume la seguente forma:

$$D.S. = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n - 1}},$$

dove **k** è il numero di modalità della variabile statistica X o il numero di classi in cui i valori di X sono stati raggruppati. In tal caso le x_i sono i valori centrali delle classi.

Esempio 11. Valori pressori massimi rilevati su 5 pazienti ipertesi

PAS (mmHg)	$\mathbf{f_i}$	$\mathbf{x_i \cdot f_i}$	$x_i - \bar{x}$	$\mathbf{(x_i - \bar{x})^2 \cdot f_i}$
170	1	170	-23	529
185	1	185	-8	64
200	1	200	7	49
205	2	410	12	288
Somma	5	965		930

- **Media Aritmetica**: $\bar{x} = 965/5 \text{ mmHg} = 193 \text{ mmHg}$;
- **Range**: R = 205 – 170 = 35 mmHg;
- **Devianza**: Dev = 930 $(\text{mmHg})^2$;
- **Varianza**: $s^2 = 930 / 4 \ (\text{mmHg})^2 = 232{,}5 \ (\text{mmHg})^2$;
- **Dev. St.**: $s = \sqrt{232{,}5} \text{ mmHg} = 15{,}25 \text{ mmHg}$;
- **Coeff. Variaz.**: CV% = (15,25 / 193)·100 = 7,9 %.

Indici di variabiltà relativi

(non dipendono dall'unità di misura)

Coefficiente di Variazione

$$CV = \frac{s}{\bar{x}} 100 = \frac{\text{Deviazione Standard}}{\text{media aritmetica}} 100$$

Per l'Esempio 5' si ottiene : $CV = \frac{13.32 mg / 100 cc}{95 mg / 100 cc} 100 = 14.02\%$

E' interessante anche il confronto tra i coefficienti di variazione delle due serie di dati dell'Esempio 10:

per il gruppo I si ha: $CV_I = \frac{15.8 aa}{40 aa} 100 = 39.50\%$,

per il gruppo II si ha: $CV_{II} = \frac{23.7 aa}{40 aa} 100 = 59.25\%$

risultati che confermano la maggiore variabilità dei dati della seconda serie rispetto alla prima.

Il Coefficiente di Variazione è un numero "puro", in quanto rapporto di due grandezze omogenee, e perciò consente il confronto anche tra variabili eterogenee.

L’uso del C.V. si rende necessario ogni qualvolta si vogliono confrontare le misure di variabilità relative a distribuzioni le cui modalità sono espresse in unità di misure diverse (confronto tra variabilità dell’altezza e del peso) oppure sono espresse nella stessa unità di misura ma il loro valore medio risulta molto diverso (confronto delle variabilità dei pesi fra un campione di neonati ed uno di adulti).

Per il calcolo della mediana (Me) e della Moda (Mo) della distribuzione della pressione si procede come nella tabella:

PAS (mmHg)	f_i	frequ. cumulate
170	1	1
185	1	2
200	1	3
205	2	5

$5 / 2 = 2{,}5 \longrightarrow$ Me = 200

Mo= 205

I due esempi che seguono illustrano il calcolo di indici medi e di variabilità nel caso di dati raggruppati in classi di frequenze.

Esempio 12. Azoto ureico (mg %) in un gruppo di 50 adolescenti

Azoto	val. centr. **(x_i)**	frequenze **(f_i)**	frequ. cum.	**$x_i * f_i$**	**$(x_i - \bar{x})^2 * f_i$**
17.1 – 19	18.05	3	3	54.15	82.3728
19.1 – 21	20.05	6	9	120.30	62.9856
21.1 – 23	22.05	11	20	242.55	16.9136
23.1 – 25	24.05	20	40	481.00	11.5520
25.1 – 27	26.05	8	48	208.40	60.9408
27.1 – 29	28.05	1	49	28.05	22.6576
29.1 – 31	30.05	1	50	30.05	45.6976
Somma		50		1164.50	303.1270

$\bar{x}$ = 1164.50/50 = 23.29 mg %;

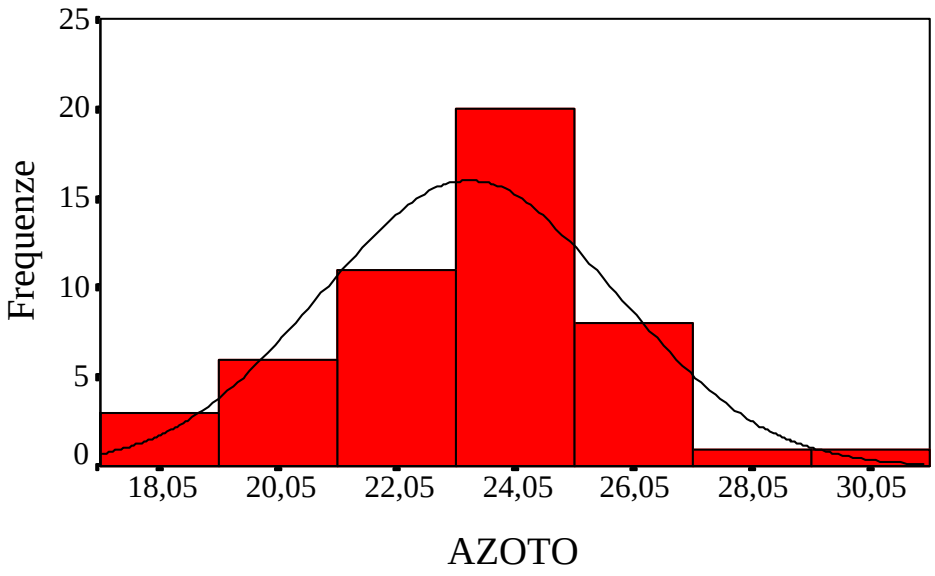
D.S. = $\sqrt{303.12/49} = 2.49$ mg %;

C.V.= 2.49/23.29*100 = 11%

calcolo mediana: N/2 = 50 / 2 = 25 $\longrightarrow$ la <u>classe mediana</u> (classe che comprende la mediana) è data da: 23.1 - 25, ovvero 23.1 < Me < 25;

calcolo moda: la frequenza più elevata si ha per la classe 23.1 - 25, dunque: 23.1 < Mo < 25.

Il grafico seguente mostra l'ISTOGRAMMA della distribuzione dell'azoto e, sovrapposta a questo, la curva della <u>distribuzione normale</u> (per lo studio di tale curva si veda in appunti successivi).


25
20
15
10
5
0
Frequenze
18,05
20,05
22,05
24,05
26,05
28,05
30,05
AZOTO

Esempio 13. Dosaggio della Fosfatasi Alcalina (UA) in 20 studenti

Fosfatasi Alcalina	Valore centrale **(x_i)**	frequenze **(f_i)**	Frequ. cumul.	$x_i * f_i$	$(x_i - \bar{x})^2 * f_i$
30.1 - 60	45.05	1	1	45.05	7056
60.1 - 90	75.05	3	4	225.15	8748
90.1 - 120	105.05	3	7	315.15	1728
120.1 - 150	135.05	7	14	945.35	252
150.1 - 180	165.05	5	19	825.25	6480
180.1 - 210	195.05	0	19	0	0
210.1 - 240	225.05	1	20	225.05	9216
		20		2581	33480

$\bar{x}$ = 2581 / 20 = 129;

D.S. = $\sqrt{33480/19} = 41.98$;

C.V. =41.98/129*100 = 32%

calcolo mediana: N/2 = 20 / 2 = 10 $\rightarrow$ la classe mediana è 120.1 - 150, ovvero 120.1< Me <150;

calcolo moda: la frequenza più elevata si ha per la classe 120.1 - 150, dunque: 120.1< Mo <150.

La misura della variabilità che è usata quando la localizzazione centrale dei dati è espressa dalla mediana è il **range interquartile**. Come abbiamo visto la mediana è usata quando la distribuzione include valori estremi che tenderebbero a influenzare in maniera eccessiva la media. Questi valori estremi tenderebbero a dare anche una stima erronea della variabilità (il range sarebbe troppo elevato). Abbiamo visto che la mediana è una misura centrale che divide in due una distribuzione. Il range interquartile si calcola dividendo in due ciascuna di queste due metà: la

distribuzione è così suddivisa in quattro parti e il range interquartile identifica i valori compresi tra il 1° e il 3° quartile. Il range interquartile ha la proprietà di eliminare l'influenza dei valori estremi e (a differenza del range) di essere relativamente indipendente dalla numerosità del campione. Il range interquartile riunisce il 50% dei valori di una distribuzione, quindi è un'espressione più "raggruppata" della media ±1DS che raccoglie il 66% dei valori di una distribuzione ±1DS.

I QUANTILI

Me — 50% | 50% — Me

Q_i — 25% | 50% | 75% | 100% — Q_1 Q_2 Q_3 Q_4

QUINTILI — 20% | 40 | 60 | 80 | 100%

DECILI — 10% | 50% | 90%

PERCENTILI — 1% | ... | 100 parti

N.B.: $Q_2 = Me$; 4° QUINTILE = 80° PERCENTILE

RANGE = $x_5 - x_1$ non efficiente, poco sensibile

$Q_3 - Q_1$ = distanza interquartile, + attendibile, 50% casi

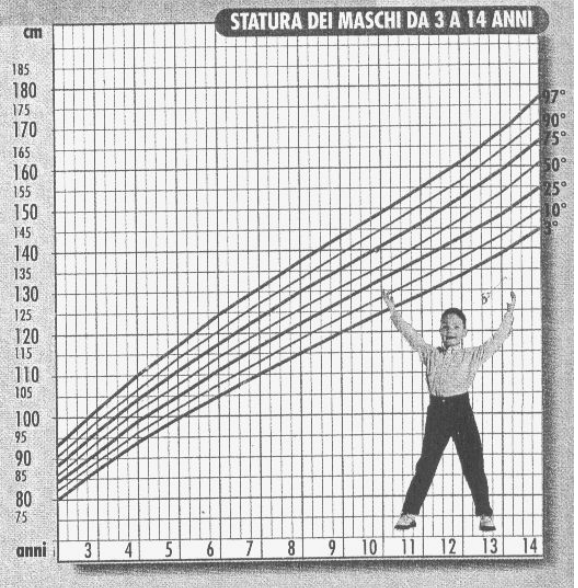
90° - 10° percentile comprende l'80% dei casi.

2.5° ⟷ 97.5°

LEGGI QUI SE TUO FIGLIO È IN SOVRAPPESO

Tabelle dei percentili

ALTEZZA



Per sapere se il peso del tuo bambino è giusto, impara a usare le tabelle dei percentili. Hai mai giocato a battaglia navale? Il procedimento è più o meno lo stesso.

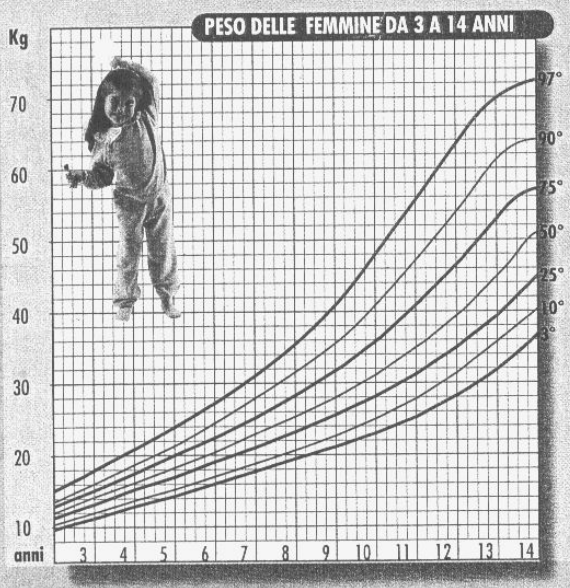
COME SI CALCOLA IL PERCENTILE. Prendi per prima cosa in considerazione la griglia dell'altezza (ce ne sono due, una per i maschi e una per le femmine). Trova, sulla linea orizzontale in basso, il punto corrispondente all'età di tuo figlio e, sulla linea verticale a sinistra, quello corrispondente all'altezza. Adesso, come nella battaglia navale, vedi dove si incontrano le linee che partono dai punti dell'età e dell'altezza. Individuerai così un nuovo punto che dovrebbe cadere all'interno delle curve dei percentili, ognuna delle quali è contrassegnata da un numero, dal 3° al 97°.

A quale percentile corrisponde l'altezza del tuo bambino? Se cade su una delle linee curve non ci sono problemi, basta leggere il numero corrispondente (che si trova sulla destra). Se invece cade nel mezzo di due linee, devi prendere per buono il percentile sottostante.

Adesso, ripeti lo stesso procedimento usando la griglia del peso.

COME SI INTERPRETANO I RISULTATI. Va precisato per prima cosa che tutti i valori compresi tra il 3° e il 97° percentile vanno considerati normali. L'importante è il confronto tra il percentile del peso e quello dell'altezza.

1. Se il percentile del peso corrisponde a quello dell'altezza o se la differenza è minima (per esempio, 50° percentile per il peso e 50°, 25° o 75° per l'altezza), significa che il bambino cresce bene e non occorre modificare il suo menu settimanale.
2. Se il percentile del peso è superiore a quello dell'altezza (per esempio, 97° percentile per il peso e 50° per l'altezza) vuol dire che il bambino è in sovrappeso e va messo a dieta, insieme a tutta la famiglia, per evitare che si senta penalizzato ingiustamente.
3. Infine, se il percentile del peso è inferiore a quello dell'altezza (per esempio 25° percentile per il peso e 75° per l'altezza) il bambino è troppo magro e bisogna aiutarlo a mangiare di più modificando la sua alimentazione e proponendogli piatti stuzzicanti.



PESO