Misure di variabilità o dispersione

Intervallo di variazione, distanza interquartile, varianza campionaria e deviazione standard.

Misure di variabilità o dispersione

La media a le mediana sono indici di posizione centrale, misurano cioè il "centro" di un insieme di dati, ma non danno informazione sulla **dispersione** dei dati.

Esempio. Due insiemi di dati, A e B

A: 1, 2, 5, 6, 6

Media=(1+2+5+6+6)/5=4

med= 5

B: -40, 0, 5, 20, 35

Media=(-40+5+20+35)/5=4

med=5

Come misuriamo la dispersione?

Intervallo di variazione

L'intervallo di variazione (range): è l'intervallo che ha come estremi il valore più piccolo dei dati (estremo sinistro) e il valore più grande (estremo destro).

Esempio. A: 1, 2, 5, 6, 6 B: -40, 0, 5, 20, 35

[1, 6] [-40, 35]

La lunghezza di questo intervallo è la differenza tra il dato più grande e il dato più piccolo

A: 6-1=5 B: 35-(-40)=75

Intervallo di variazione

L'intervallo di variazione influenzato dalle osservazioni estreme.

Esempio.

	B	C
	-40, 0, 5, 20, 35	-140, 0, 5, 20, 135
	media=4	media=4
	med=5	med=5
	[-40, 35]	[-140, 135]
	Lungh. = 75	Lungh.= 135-(-140)=275

Distanza interquartile

Distanza interquartile o scarto intequartile: è la differenza tra il terzo quartile Q_3 e il primo quartile Q_1.

Misura la dispersione del 50% dei valori dei dati (valori centrali)

Esempio. Dati ordinati (ordine crescente); n=8

11 13 15 16 19 21 22 25

$Q_1 = 14$ $Q_3 = 21{,}5$

Dist. Inter. $\Delta = Q_3 - Q_1 = 21.5 - 14 = 7{,}5$

Distanza interquartile

Esempio. Tre insiemi di dati A, B, C

A: 1, 2, 5, 6, 6
$Q_1 = 2$, $Q_3 = 6$
$\Delta = Q_3 - Q_1 = 6 - 2 = 4$

B: -40, 0, 5, 20, 35
$Q_1 = 0$, $Q_3 = 20$
$\Delta = 20$

C: -140, 0, 5, 20, 135
$Q_1 = 0$, $Q_3 = 20$
$\Delta = 20$

Il 50% dei valori degli insiemi B e C (valori centrali) sono dislocati su un intervallo più grande rispetto quelli dell'insieme A.

Varianza campionaria

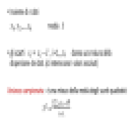
- Insieme di n dati

  $x_1, x_2 ,\ldots, x_n$ media $\bar{x}$

- gli scarti $s_i = x_i - \bar{x}$, i=1,…,n, danno una misura della dispersione dei dat (ci interessano I valori assoluti).

Varianza campionaria: è una misura della media degli scarti quadratici

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$



Varianza campionaria

- Insieme di n dati

  $x_1, x_2, \ldots, x_n$ media $\bar{x}$

- gli scarti $s_i = x_i - \bar{x}$, i=1,…,n, danno una misura della dispersione dei dati. (ci interessano I valori assoluti)

Varianza campionaria: è una misura della media degli scarti quadratici

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

Varianza campionaria

- Come misura di dispersione della popolazione (e non di un campione) si usa la varianza σ^2 (media quadratica), così come definita in probabilità.

- La varianza campionaria s^2 è uno stimatore non distorto della varianza della popolazione σ^2.

Esempio

Insieme di dati 1, 2, 5, 6, 6 n=5

- $\bar{x}$ =4
- Scarti -3, -2, 1, 2, 2
- Quadrato degli scarti 9, 4, 1 4, 4
- Varianza campionaria s^2 =(9+4+1+4+4)/4 = 22/4=5,5

Identità algebrica

- Identità algebrica:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 \; + n\,\bar{x}^2 \text{ - } 2\,\bar{x}\; \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} {x_i}^2 \; - n\,\bar{x}^2$$

  da cui

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} {x_i}^2 \; - n\,\bar{x}^2}{n-1}$$

Esempio.

Insieme di dati 1, 2, 5, 5, 6 n=5

- $\bar{x}$ = 4,
- s^2 = $\frac{1+4+25+25+36 - 5(16)}{4}$ = 5,5

Proprietà 1

Se I valori dei dati vengono incrementati di una quantità c, la varianza campionaria non cambia

- $x_1, x_2, \ldots, x_n$ media $\bar{x}$
- $y_1, y_2, \ldots, y_n$ $y_i = x_i + c$, i=1,..,n $\bar{y} = \bar{x} + c$

$(y_i - \bar{y})^2 = (x_i + c - (\bar{x} + c))^2 = (x_i - \bar{x})^2$ da cui $S_y^2 = S_x^2$

Deviazione standard campionaria

La deviazione standard campionaria è la radice quadrata della varianza

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

E' espressa della stessa unità di misura dei dati

Esempio. Lunghezze (cm) di 6 vermi

8,3 9,2 10,0 10,2 11,5 12,1

$\bar{x}$ =10,2 cm, s^2 =2 cm^2, s=1,4 cm

Proprietà 2

Se moltiplichiamo ciascun valore dei dati per una costante c, la varianza risulta moltiplicata per un fattore c^2, la deviazione standard per un fattore |c|

- $x_1, x_2, \ldots, x_n$ media $\bar{x}$
- $y_1, y_2, \ldots, y_n$ $y_i = cx_i$, i=1,..,n $\bar{y}$ = c $\bar{x}$

$(y_i - \bar{y})^2 = (cx_i - (c\bar{x}))^2 = c^2 (x_i - \bar{x})^2$ da cui

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = c^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$ e quindi $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$

Esempio

Temperatura corporea misurata in gradi Farenheit di 65 soggetti sani

- $\bar{x}$ = 98,10 (° F)
- S=0,70 (°F)

Temperature corporee in gradi Farenheit			
96,30	97,60	98,20	98,70
96,70	97,70	98,20	98,70
96,90	97,80	98,20	98,80
97,00	97,80	98,30	98,80
97,10	97,80	98,30	98,80
97,10	97,80	98,40	98,90
97,10	97,90	98,40	99,00
97,20	97,90	98,40	99,00
97,30	98,00	98,40	99,00
97,40	98,00	98,50	99,10
97,40	98,00	98,50	99,20
97,40	98,00	98,60	99,30
97,40	98,00	98,60	99,40
97,50	98,00	98,60	99,50
97,50	98,10	98,60	
97,60	98,10	98,60	**Media**
97,60	98,20	98,60	**98,10**

Esempio

Temperatura corporea degli stessi 65 soggetti sani espressa in gradi Celsius

- $\bar{x}$ = 36,73(° C)
- S =0,39 (° C)

Temperature corporee in gradi Celsius			
35,73	36,45	36,78	37,06
35,95	36,50	36,78	37,06
36,06	36,56	36,78	37,12
36,12	36,56	36,84	37,12
36,17	36,56	36,84	37,12
36,17	36,56	36,89	37,17
36,17	36,62	36,89	37,23
36,23	36,62	36,89	37,23
36,28	36,67	36,89	37,23
36,34	36,67	36,95	37,28
36,34	36,67	36,95	37,34
36,34	36,67	37,00	37,39
36,34	36,67	37,00	37,45
36,39	36,67	37,00	37,50
36,39	36,73	37,00	
36,45	36,73	37,00	**Media**
36,45	36,78	37,00	**36,73**

Trasformazione delle scale di misura

Le due scale di misura sono legate dalla trasformazione lineare

$$T_c = \frac{5}{9}(T_f - 32)$$

T_c temperature in °C

T_f temperature in °F

La seconda tabella dei valori si ottiene dalla prima sottraendo a ciascun dato la quantità 32 e moltiplicando poi per 5/9. Per le proprietà della media

$$\bar{x}_c = \frac{5}{9}(\bar{x}_f - 32) \qquad 36{,}73 = \frac{5}{9}(98{,}10 - 32)$$

Trasformazione delle scale di misura

Le due scale di misura sono legate dalla trasformazione lineare

$$T_c = \frac{5}{9}(T_f - 32)$$

T_c temperature in °C

T_f temperature in °F

La seconda tabella dei valori si ottiene dalla prima sottraendo a ciascun dato la quantità 32 e moltiplicando poi per 5/9. Per le proprietà della deviazione standard

$$S_c = \frac{5}{9} S_f \qquad 0{,}39 = \frac{5}{9}\ (0{,}70\ )$$

Coefficiente di variazione

La deviazione standard risente dell'unità di misura e dell'ordine di grandezza dei dati.

Esempio. Due campioni di 5 individui maschi (25 e 11 anni)

	Campione 1	Campione 2
Età	**25 anni**	**11 anni**
Peso medio	**70 kg**	**36 kg**
Dev.st.	**3 kg**	**3 kg**

Coefficiente di variazione

- Due distribuzioni con deviazioni standard molto vicine non hanno necessariamente un'analoga dispersione; infatti la dev. st. è "grande" o "piccola" rispetto all'ordine di grandezza delle misure a cui si riferisce, ossia ad un indice di posizione come la media.
- Per confrontare le dispersioni di due diverse distribuzioni occorre confrontare indici indipendenti dall'unità di misura. L'indice più utilizzato è il Coefficiente di Variazione

Coefficiente di variazione

Il coefficiente di variazione è dato dal rapporto tra la deviazione standard campionaria e la media campionaria

$$C.V. = \frac{s}{\bar{x}}$$

Campione 1: C. V. = 3/70 = 0,042 (4.2 %)

Campione 2: C. V. = 3/36 = 0,08 (8 %)

Il coefficiente di variazione esprime la deviazione standard come frazione (percentuale) della media ed è indipendente dall'unità di misura.

Esempio

Campione di cinque individui maschi di 25 anni

- Peso medio=70 kg   dev. St.= 3 Kg
- Altezza media=179cm   dev. St. =4.1cm

Coeff. di variazione

- Peso   c.v.=3/70=0,042   (4,2%)
- Altezza   c.v.=4,1/179=0,022 (2,2 %)

I valori del peso sono più sparsi rispetto a quelli delle altezze (il c.v. è quasi il doppio)