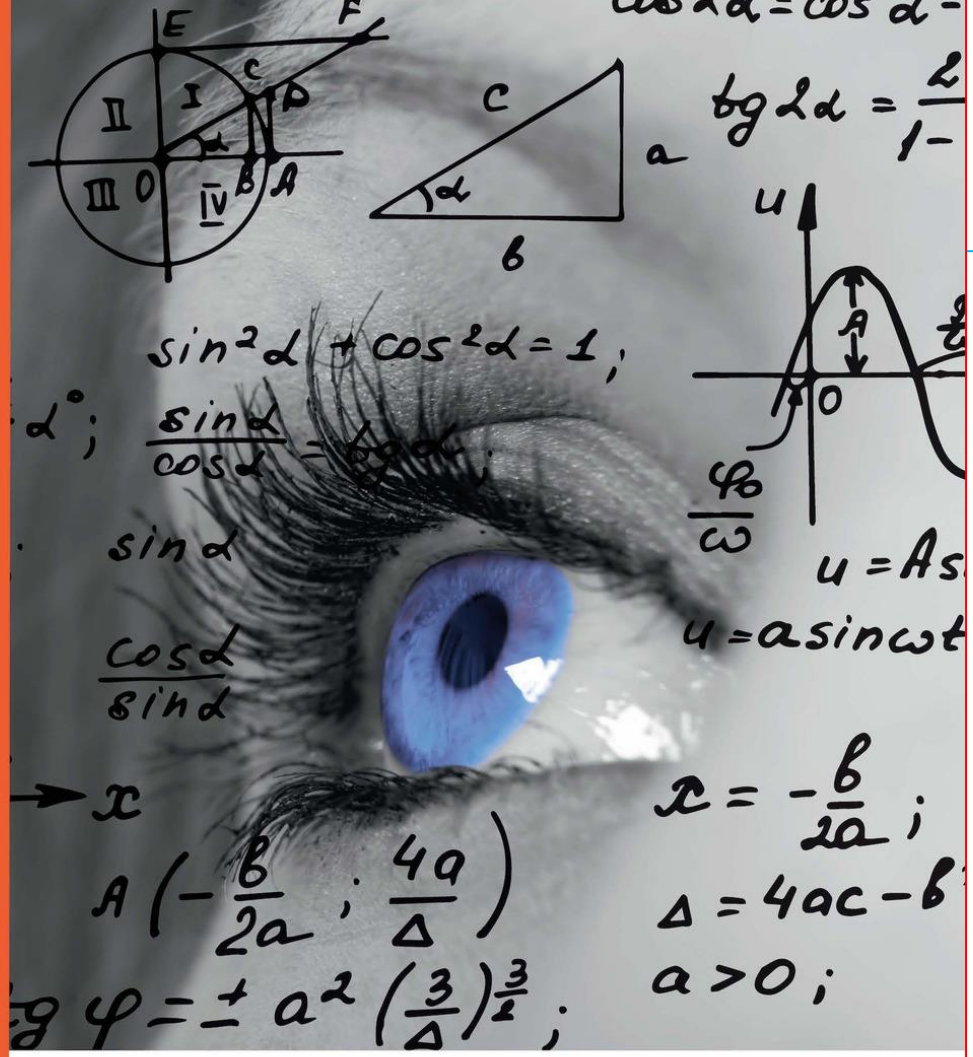


Statistica per le scienze sociali

Enrica Amaturo, Biagio Aragona,
Maria Gabriella Grassia, Carlo Natale Lauro,
Marina Marino

UTET UNIVERSITÀ

4 Sintetizzare e confrontare le distribuzioni

I valori di disuguaglianza

Omogeneità (equilibrio) ed eterogeneità (squilibrio)

Dispersione

Variabilità rispetto a un centro

Altri indici di variabilità

Rappresentare graficamente la variabilità: il box plot

La forma di una distribuzione: asimmetria e curtosi

Concentrazione di una variabile trasferibile

La standardizzazione

Confronti basati sulle differenze

I numeri indice

Capitolo a cura di B. Aragona

I valori di disuguaglianza

Un valore caratteristico di tendenza centrale da solo non consente una rappresentazione soddisfacente di una distribuzione.

Perché si possa attuare una sintesi davvero efficace è necessario tenere conto anche del modo in cui le unità si distribuiscono tra le diverse modalità della variabile, cioè della loro **disuguaglianza**.

La Variabilità

Contemporaneamente ai valori caratteristici di tendenza centrale, bisogna, quindi, studiare la **variabilità**.

La **variabilità** esprime la tendenza delle unità di un collettivo ad assumere diverse modalità (valori) del carattere

I valori di disuguaglianza

I **valori (o indici) di disuguaglianza** indicano come sono diverse le unità statistiche rispetto alla variabile considerata.

Ci aiutano a determinare se i nostri dati sono lontani dal valore centrale e quindi se il valore centrale scelto è adeguato per rappresentare la popolazione dello studio

Maggiore è la variabilità, meno rappresentativo è il valore centrale

Categorie di valori di disuguaglianza

E’ possibile distinguere quattro categorie di indici di disuguaglianza:

- indici di mutabilità

  misurano **l’omogeneità/eterogeneità** tra le modalità di una distribuzione di frequenza

- intervalli di variazione

- indici di dispersione rispetto ad un valore di tendenza centrale

  definiti anche indici di variabilità misurano la disuguaglianza delle unità rispetto ad un valore centrale

- indici di disuguaglianza a coppie

  definiti anche indici di mutua variabilità o variabilità reciproca, misurano la disuguaglianza tra le unità prese a coppia

Proprietà degli indici di disuguaglianza

Un indice di disuguaglianza deve soddisfare almeno due requisiti:

- deve assumere il **valore minimo (nullo)** se e solo se tutte le unità della distribuzione presentano uguale modalità del carattere

- deve **aumentare** all'aumentare della "diversità" tra le modalità assunte dalle varie unità

Ulteriore differenze tra valori di disuguaglianza

Un'ulteriore distinzione tra gli indici è la seguente

a. **Assoluti:** utilizzano la stessa unità di misura della modalità della distribuzione, ma non consentono di fare confronti fra distribuzioni statistiche espresse in unità di misure diverse

b. **Relativi:** depurano la distribuzione dall'unità di misura, per questo motivo sono particolarmente adatti per operare confronti tra distribuzioni. Si ottengono rapportando un indice assoluto al suo massimo o ad un valore centrale (di solito la media)

Categorie di valori di disuguaglianza

E’ possibile distinguere tre categorie di indici di disuguaglianza:

- indici di mutabilità

  misurano **l’omogeneità/eterogeneità** tra le modalità di una distribuzione di frequenza

- **Intervalli di variazione**

- **indici di dispersione rispetto ad un valore di tendenza centrale**

  definiti anche indici di variabilità misurano la disuguaglianza delle unità rispetto ad un valore centrale

Indici di mutabilità

Un indice di mutabilità è **l'Indice di eterogeneità del Gini** che misura il numero di cambiamenti necessari per raggiungere l'omogeneità:

$$E = 1 - \sum_{i=1}^{k} (f_i)^2$$

Questo indice ha come **minimo 0** nel caso una modalità abbia una frequenza relativa pari a 1 e le altre modalità pari a 0.

Il massimo dell'indice di eterogeneità si ha invece quando tutte le frequenze relative sono uguali e cioè quando **ogni frequenza relativa è uguale** a $\frac{1}{k}$ $\boldsymbol{con\ k}$ pari al numero di modalità assunte dal carattere.

Il valore massimo dell'indice è, quindi, pari a:

$$E = 1 - \sum_{i=1}^{k} \left(\frac{1}{k}\right)^2 = 1 - \sum_{i=1}^{k} \left(\frac{1}{k}\right)^2 = 1 - k\left(\frac{1}{k^2}\right) = \mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{k}} = \frac{\boldsymbol{k-1}}{\boldsymbol{k}}$$

Trasformando l'indice da assoluto a relativo (dividendo per il valore massimo) avremo l'indice relativo di eterogeneità del Gini:

$$\boldsymbol{e} = \frac{\boldsymbol{E}}{\frac{\boldsymbol{k-1}}{\boldsymbol{k}}} = \frac{\boldsymbol{Ek}}{\boldsymbol{k-1}}$$

Indice di mutabilità

Gli indici di mutabilità sono gli unici indici calcolabili per le variabili qualitative sconnesse

Categorie di valori di disuguaglianza

E’ possibile distinguere tre categorie di indici di disuguaglianza:

- **indici di mutabilità**

  misurano **l’omogeneità/eterogeneità** tra le modalità di una distribuzione di frequenza

- intervalli di variazione

- **indici di dispersione rispetto ad un valore di tendenza centrale**

  definiti anche indici di variabilità misurano la disuguaglianza delle unità rispetto ad un valore centrale

Intervalli di variazione

Il **range** (campo di variazione) è la differenza tra il valore massimo ed il valore minimo della distribuzione

$$\text{Range} = Max(x) - Min(x)$$

La **differenza interquartile** è la differenza tra il terzo e il primo quartile e può essere utilizzata come misura di dispersione di una distribuzione ordinata di frequenze.

$$IQR = Q_3 - Q_1$$

NB: *La differenza interquartile rappresenta il campo di variazione per il 50% delle unità più vicine alla mediana*

Intervalli di variazione

Facili da calcolare

Possono essere anche determinati per le variabili qualitative con categorie ordinate

Non forniscono alcuna informazione su ciò che succede all'interno degli estremi considerati

Categorie di valori di disuguaglianza

E’ possibile distinguere tre categorie di indici di disuguaglianza:

- **indici di mutabilità**

  misurano l’omogeneità/eterogeneità tra le modalità di una distribuzione di frequenza

- **indici di dispersione rispetto a determinati valori**

- indici di dispersione rispetto ad un valore di tendenza centrale

  definiti anche indici di variabilità misurano la disuguaglianza delle unità rispetto ad un valore centrale

Gli indici di dispersione rispetto ad un valore di tendenza centrale

Gli indici di dispersione rispetto alla:

Media $M(X)$

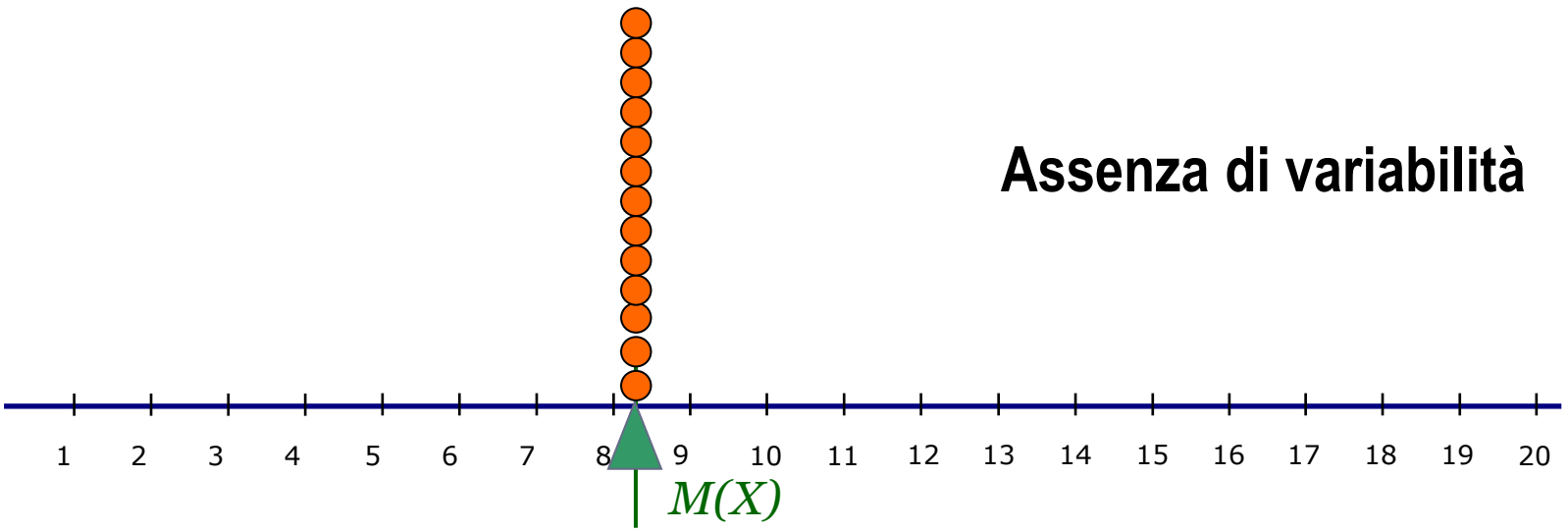
Mediana $Me(X)$

I più diffusi sono:

- Devianza
- Varianza
- Scarto quadratico medio
- Scostamento semplice medio assoluto dalla media
- Scostamento semplice medio assoluto dalla mediana

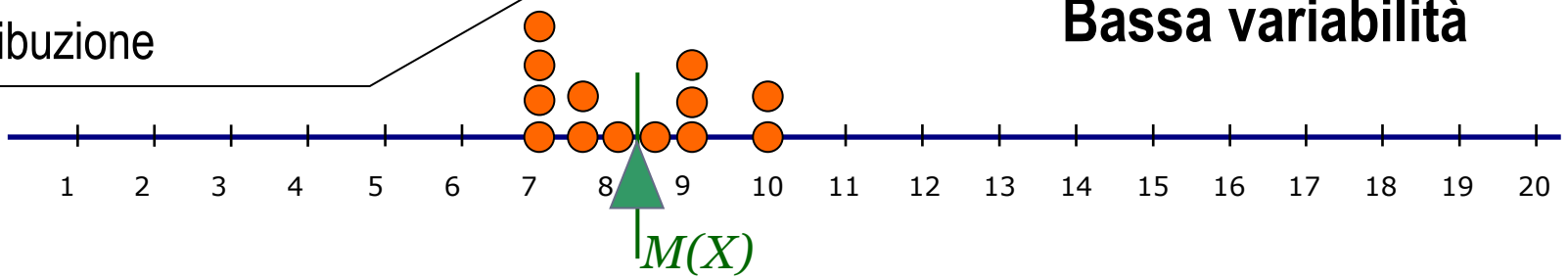
Coinvolgono tutti i valori del collettivo

Gli indici di dispersione rispetto alla media

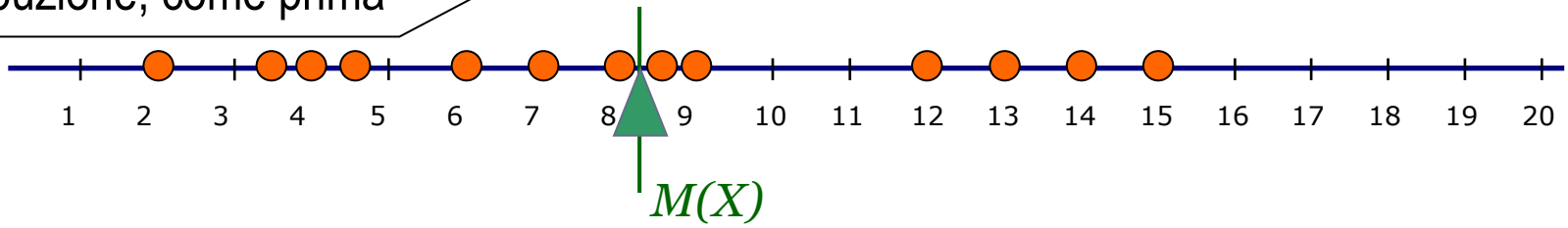


Gli indici di dispersione rispetto alla media

Il valor medio da' una buona rappresentazione dei valori della distribuzione



Lo stesso valor medio non da' una buona rappresentazione dei valori della distribuzione, come prima



Gli indici di dispersione rispetto alla media: La **Devianza**

Definiamo **devianza** di un insieme di N osservazioni $x_1, x_2, \ldots, x_n$ con media $M(x)$ la **somma degli scarti al quadrato dalla media aritmetica**

$$Devianza = \sum_{i=1}^{N} (x_i - M(X))^2$$

Gli indici di dispersione rispetto alla media: La **Devianza**

Definiamo **devianza** di un insieme di N osservazioni $x_1, x_2, \ldots, x_n$ con media $M(x)$ la **somma degli scarti al quadrato dalla media aritmetica**

$$Devianza = \sum_{i=1}^{N} (x_i - M(X))^2$$

- Questo indice confronta le modalità osservate con la media aritmetica della distribuzione
- Riflette tutti i valori del collettivo (o della distribuzione)
- Considera il quadrato delle differenze di ogni modalità dalla media

Gli indici di dispersione rispetto alla media: La **Devianza**

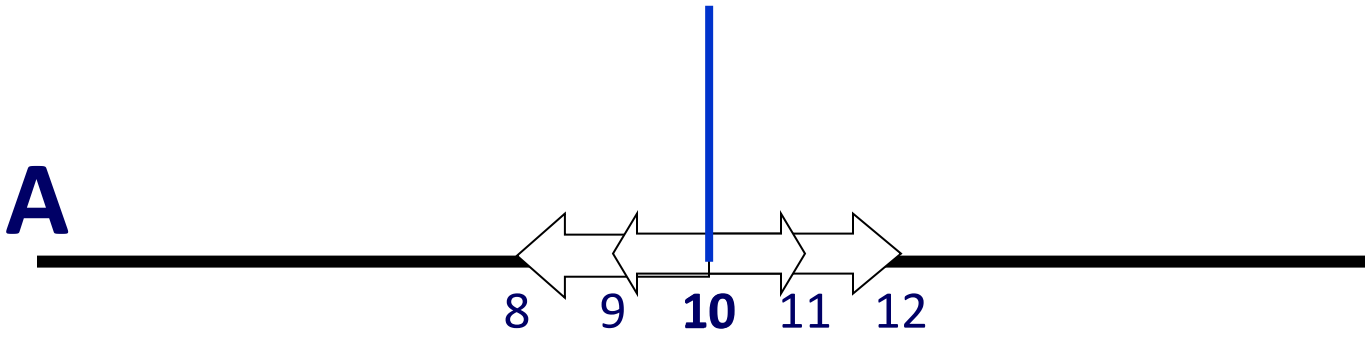
$$Devianza = Dev = \sum_{i=1}^{N} (x_i - M(X))^2$$

1) Perché ogni scarto dalla media è al quadrato e non utilizziamo solo la somma degli scarti?

Guardiamo l'esempio seguente

Gli indici di dispersione rispetto alla media: La Devianza Esempio

Consideriamo il collettivo **A**: 8, 9, 10, 11, 12



Come misurare la variabilità del collettivo rispetto alla media?

Considero gli scarti dalla media

8-10= -2
9-10= -1
10-10= 0
11-10= +1
12-10= +2

Gli indici di dispersione rispetto alla media: La Devianza Esempio

Calcoliamo la varianza del collettivo **A**: 8, 9, 10, 11, 12

scarti dalla media

8-10= -2	+
9-10= -1	+
10-10= 0	+
11-10= +1	+
12-10= +2	=0

Ricordiamo che la somma degli scarti dalla media è uguale a 0

$$Dev_A = (8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2 = 10$$

Gli indici di dispersione rispetto alla media: La **Devianza**

Riflette tutti i valori del collettivo (o della distribuzione)

Cresce all’aumentare della numerosità del collettivo

Gli indici di dispersione rispetto alla media: La Varianza

Definiamo **varianza** di un insieme di n osservazioni x_1 , x_2 ,…,x_n con media $M(x)$ la **media degli scarti al quadrato dalla media aritmetica**

$$Var(X) = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - M(X))^2}{N} = \frac{Dev\ (X)}{N}$$

Gli indici di dispersione rispetto alla media: La Varianza

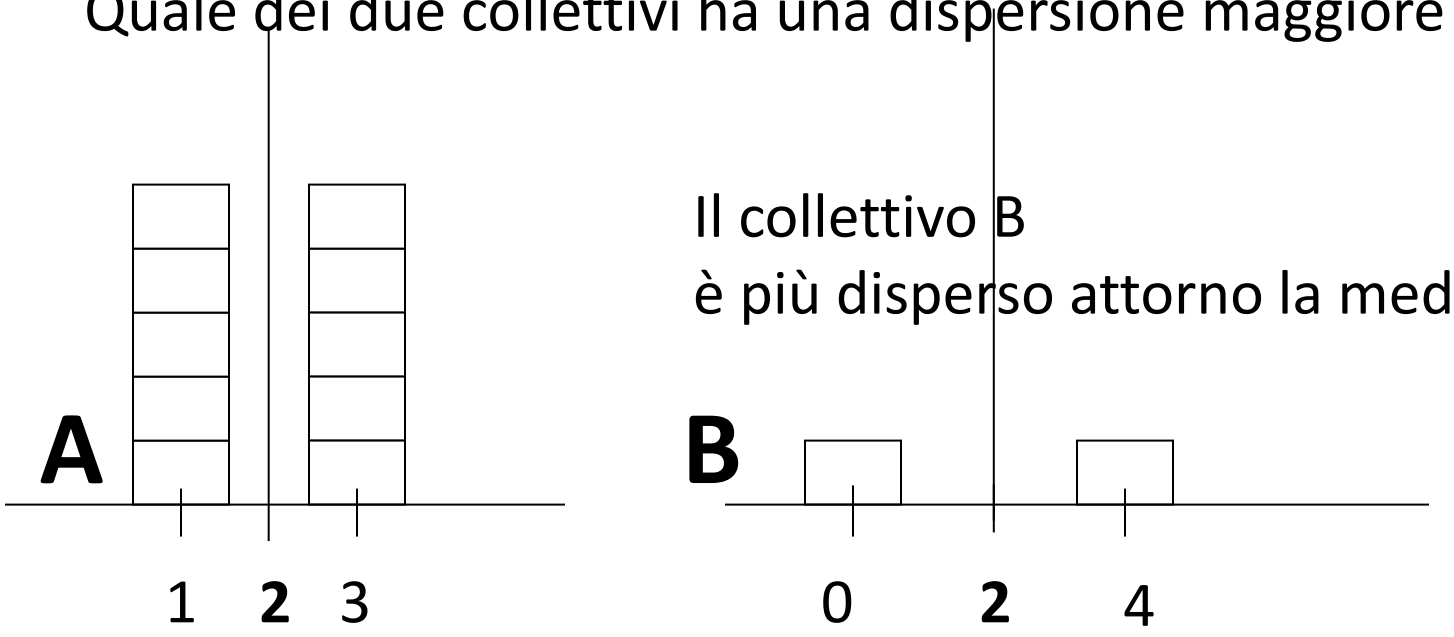
$$Var(X) = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - M(X))^2}{N} = \frac{Dev\,(X)}{N}$$

2) Perché calcolare la varianza come media degli scarti al quadrato?

Guardiamo l'esempio seguente

Gli indici di dispersione rispetto alla media: La Varianza

Quale dei due collettivi ha una dispersione maggiore?



Il collettivo B
è più disperso attorno la media

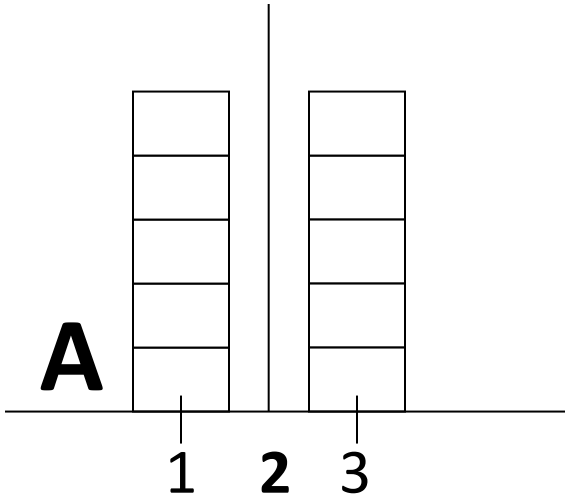
Calcoliamo la somma dei quadrati degli scarti per entrambe le distribuzioni

$\text{Somma}_A = \underbrace{(1-2)^2 + \ldots + (1-2)^2}_{5 \text{ volte}} + \underbrace{(3-2)^2 + \ldots + (3-2)^2}_{5 \text{ volte}} = 10$

$\text{Somma}_B = (0-2)^2 + (4-2)^2 = 8$ ⟶ !

Gli indici di dispersione rispetto alla media: La Varianza
Esempio

Quale dei due collettivi ha una dispersione maggiore?



Il collettivo B
è più disperso attorno la media

Calcoliamo la somma dei quadrati degli scarti per entrambe le distribuzioni

$Somma_A = (1-2)^2 + \ldots + (1-2)^2 + (3-2)^2 + \ldots + (3-2)^2 = 10$

$Somma_B = (1-3)^2 + (5-3)^2 = 8$

Quando dividiamo per il numero di osservazioni, la dispersione delle due distribuzioni è correttamente calcolata

$Var_A = Somma_A/N = 10/10 = 1$

$Var_B = Somma_B/N = 8/2 = 4$

Gli indici di dispersione rispetto alla media: La Varianza

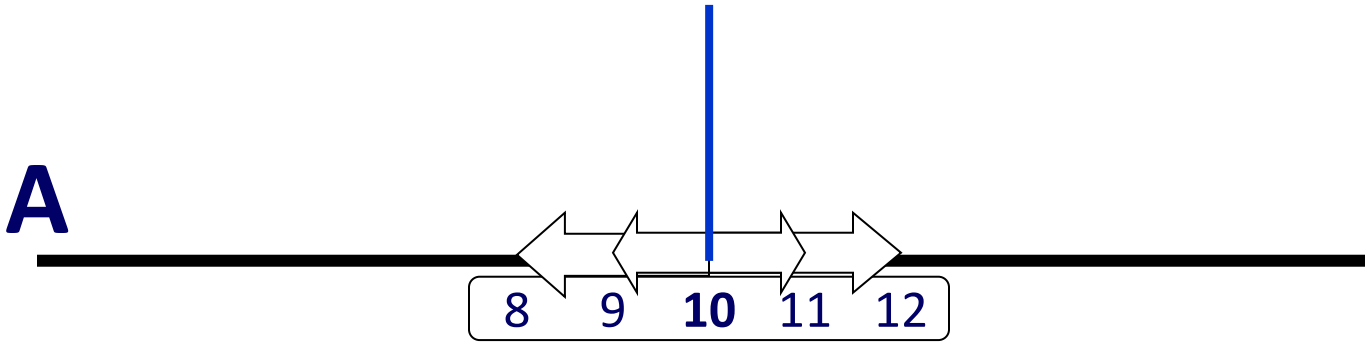
$$Var(X) = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - M(X))^2}{N} = \frac{Dev\,(X)}{N}$$

3) La varianza è un indice che soddisfa i due requisiti richiesti per un indice di disuguaglianza? Aumenta all'aumentare della variabilità?

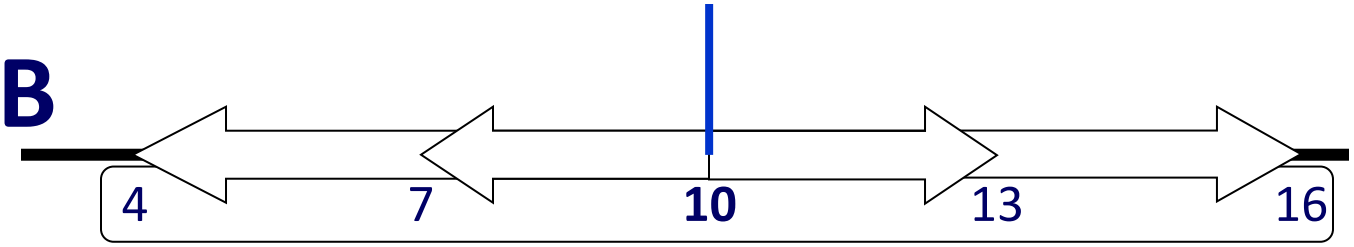
Guardiamo l'esempio seguente

Gli indici di dispersione rispetto alla media: La Varianza
Esempio

Consideriamo due piccoli collettivi:
Collettivo **A**: 8, 9, 10, 11, 12
Collettivo **B**: 4, 7, 10, 13, 16



…la dispersione in B
è molto maggiore rispetto ad A



Gli indici di dispersione rispetto alla media: La Varianza
Esempio

Calcoliamo la varianza dei due collettivi

$$\sigma_A^2 = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} = 2$$

$$\sigma_B^2 = \frac{(4-10)^2 + (7-10)^2 + (10-10)^2 + (13-10)^2 + (16-10)^2}{5} = 18$$

la varianza è maggiore nel collettivo B perché il collettivo B è più disperso intorno alla media rispetto ad A

Gli indici di dispersione rispetto alla media: La Varianza

- **Calcolo semplificato della varianza**

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \big(M(X)\big)^2 =$$

$$= M(X^2) - \big(M(X)\big)^2$$

Gli indici di dispersione rispetto alla media: La Varianza
Esempio

Calcoliamo la varianza della popolazione **A**: 8, 9, 10, 11, 12

$$\sigma_A^2 = M(X^2) - \left(M(X)\right)^2$$

$$\sigma_A^2 = \frac{8^2 + 9^2 + 10^2 + 11^2 + 12^2}{5} - 10^2 = 2$$

Gli indici di dispersione rispetto alla media: La Varianza

Riflette tutti i valori del collettivo (o della distribuzione)

NON Cresce all'aumentare della numerosità del collettivo

Cresce all'aumentare della dispersione

L'unità di misura della varianza è quella della variabile al **quadrato**

X = Altezza (m)

Var (X) = Altezza (m^2)

Gli indici di dispersione rispetto alla media: Lo Scarto quadratico medio

La varianza non possiede la stessa unità di misura dei valori della distribuzione

Si può utilizzare perciò come indice di variabilità la **deviazione standard** o **scarto quadratico medio** che è espresso nella stessa unità di misura del carattere

La radice quadrata della varianza è lo *scarto quadratico medio* (definita anche deviazione standard)

$$\sigma = \sqrt{Var(X)}$$

Gli indici di dispersione rispetto alla media: Lo Scarto quadratico medio

Età	Voto di diploma (in 100)
18	63
18	75
22	68
26	81
21	80
18	97

N=6

Quale delle due variabili ha maggiore variabilità?

$$\sigma = \sqrt{Var(X)}$$

Gli indici di dispersione rispetto alla media: Lo Scarto quadratico medio

Riflette tutti i valori del collettivo (o della distribuzione)

NON Cresce all'aumentare della numerosità del collettivo

Cresce all'aumentare della dispersione

L'unità di misura è quella della variabile

Non consente di confrontare variabili con differenti unità di misura o stessa unità di misura ma medie differenti, o numerosità molto differenti
(E' un indice assoluto)

Gli indici di dispersione rispetto alla media: Coefficiente di variazione

Osservazione: la varianza e la deviazione standard sono indici che risentono dell'unità di misura e dell'ordine di grandezza dei dati. Pertanto il confronto della variabilità tra collettivi diversi o variabili diverse risulta compromesso

Per confrontare la variabilità di due differenti caratteri, può essere utilizzato il coefficiente di variazione

$$cv = \frac{\sigma}{M} \quad \Rightarrow \quad cv = \frac{\sigma}{M} * 100$$

Esempio

Esempio 9 industrie con dispositivo anti-inquinante di tipo A e 9 di tipo B

Tipo	Quantità di pulviscolo								
A	69	80	44	52	54	54	86	77	66
B	35	62	43	23	30	28	22	40	25

$M_A = 64{,}67$

$M_B = 34{,}22$

$\sigma_A = 13{,}65$

$\sigma_B = 12{,}02$

$CV_A = 21\%$

$CV_B = 35\%$

La distribuzione B è più variabile della distribuzione A

Gli indici di dispersione rispetto alla media: Coefficiente di variazione

E’ un indice relativo

Ha significato solo se la media è positiva

Non è calcolabile quando la media è nulla o prossima allo zero

Gli indici di dispersione rispetto alla media
Calcolo per Distribuzioni di frequenze con k modalità

Formule

$$Dev = \sum_{i=1}^{k} (x_i - M)^2 n_i$$

$$Var(X) = \sigma^2 = \frac{1}{N} \sum_{i=1}^{k} (x_i - M)^2 n_i = \sum_{i=1}^{k} (x_i - M)^2 f_i$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{k} (x_i - M)^2 n_i} = \sqrt{\sum_{i=1}^{k} (x_i - M)^2 f_i}$$

Gli indici di dispersione rispetto alla media
Calcolo per Distribuzioni di frequenze con k modalità

Esempio

Età studenti del Corso	Frequenze assolute (n_i)
18	2
19	44
20	66
21	32
22	18
23	13
24	9
25	6
Totale	*190*

$$Var(X) = \frac{\sum_i (x_i - M)^2 \cdot n_i}{n}$$

Gli indici di dispersione rispetto alla media
Calcolo per Distribuzioni di frequenze con k modalità

Esempio

Età studenti del Corso	Frequenze assolute (n_i)
18	2
19	44
20	66
21	32
22	18
23	13
24	9
25	6
Totale	*190*
Età media	*20,6579*

$$Var(X) = \frac{\sum_i (x_i - M)^2 \cdot n_i}{n}$$

Gli indici di dispersione rispetto alla media
Calcolo per Distribuzioni di frequenze con k modalità

Esempio

Età studenti del Corso	Frequenze assolute (n_i)	$(x_i - m)$	$(x_i - m)^2 n_i$
18	2	-2,66	14,13
19	44	-1,66	120,94
20	66	-0,66	28,57
21	32	0,34	3,75
22	18	1,34	32,42
23	13	2,34	71,31
24	9	3,34	100,53
25	6	4,34	113,12
Totale	*190*		*484,76*

Età media 20,6579

$$Var(X) = \frac{\sum_i (x_i - M)^2 \cdot n_i}{n}$$

$$Var(X) = \frac{484,76}{190} = 2,55 \quad \Rightarrow \quad Sqm(X) = \sqrt{2,55} = 1,60 \text{ anni}$$

Gli indici di dispersione rispetto alla media

Calcolo per Distribuzioni di frequenze con modalità in classi

Formule

$$Dev \cong \sum_{i=1}^{k} (c_i - M)^2 n_i$$

$$\mathrm{Var(x)} \cong \frac{1}{N} \sum_{i=1}^{k} (c_i - M)^2 n_i = \sum_{i=1}^{k} (c_i - M)^2 f_i$$

$$\sigma \cong \sqrt{\frac{1}{N} \sum_{i=1}^{k} (c_i - M)^2 n_i} = \sqrt{\sum_{i=1}^{k} (c_i - M)^2 f_i}$$

Altri indici

- **Lo scostamento semplice medio dalla media aritmetica:**

$$S_M = \frac{1}{N}\sum_{i=1}^{N}|x_i - M|$$

N.B. $S_M \leq \sigma$

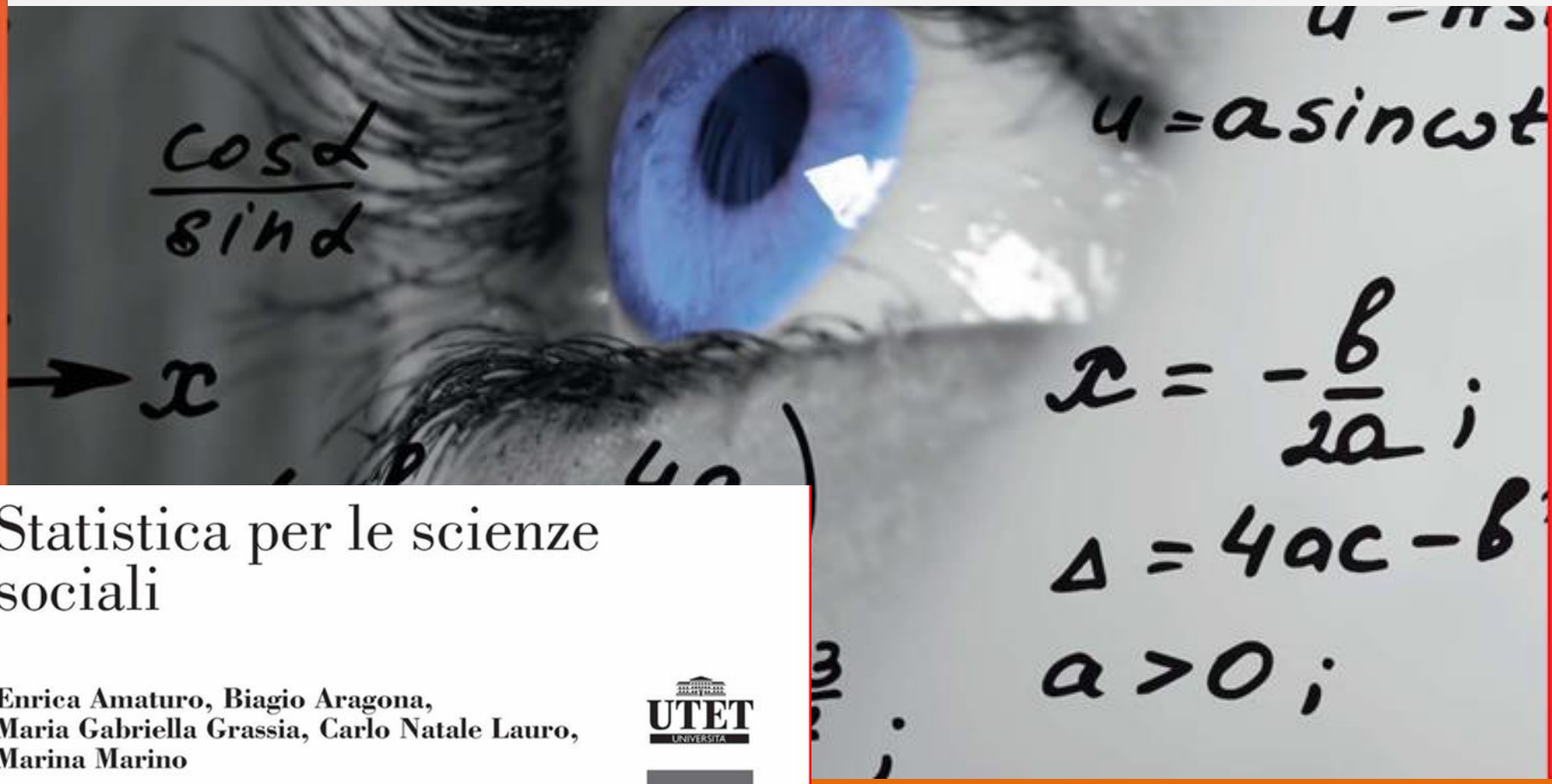
- **Lo scostamento semplice medio dalla mediana:**

$$S_{M_e} = \frac{1}{N}\sum_{i=1}^{N}|x_i - M_e|$$

Si usa in presenza di valori anomali: la mediana è un indice più robusto

Video da vedere
https://www.youtube.com/watch?v=4zKpqhh_cUg
https://www.youtube.com/watch?v=NXi7O1p8xqo
https://www.youtube.com/watch?v=JjILOUvRzpc

Altri link utili:
https://www.youtube.com/watch?v=FnbEug4V0fU



Statistica per le scienze sociali

Enrica Amaturo, Biagio Aragona,
Maria Gabriella Grassia, Carlo Natale Lauro,
Marina Marino

UTET UNIVERSITÀ