"Anticipare la crescita con le nuove competenze sui Big Data" Operazione Rif. PA 2023-19167/RER approvata con DGR n° 843 del 29 maggio 2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027 Regione Emilia-Romagna

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Operazione Rif. PA 2023-19167/RER/10/1, “ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA”, approvata dalla Regione Emilia-Romagna con DGR n° 843 del 29/05/2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027

DATA LAB

LE MISURE DI VARIABILITA'

In assenza di variabilità in una popolazione la statistica non sarebbe necessaria: un singolo *elemento* o unità campionaria sarebbe sufficiente a determinare tutto ciò che occorre sapere su una popolazione. Ne consegue, perciò, che nel presentare informazioni su un campione non è sufficiente fornire semplicemente una misura della *media ma servono informazioni sulla variabilità.*

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

Soggetti	I gruppo	II gruppo
1 2 3 4 5	20 30 40 50 60	10 25 40 55 70
Tot	200	200
Media Aritmetica	200/5=40	200/5=40

LE MISURE DI VARIABILITÀ

- ✓ Campo di variazione (range);
- ✓ Devianza;
- ✓ Varianza;
- ✓ Deviazione Standard;
- ✓ Coefficiente di variazione (variabilità relativa).

IL CAMPO DI VARIAZIONE O RANGE

DEFINIZIONE: Il *Campo di variazione o Range* corrisponde alla differenza fra la modalità più piccola e la modalità più grande della distribuzione.

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

Limiti del campo di variazione:

- ✓ è troppo influenzato dai valori estremi;
- ✓tiene conto dei due soli valori estremi, trascurando tutti gli altri.

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

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

2ª idea  $\sum_{i=1}^{n} |x_i - \bar{x}|$

3ª idea  $\sum_{i=1}^{n} (x_i - \bar{x})^2$

LA DEVIANZA

DEFINIZIONE: La somma dei quadrati degli scarti dalla media aritmetica

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

Esempio 9 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
$\bar{x} = 95$	94	1596

La quantità 1596 esprime la *Devianza*

LA VARIANZA

DEFINIZIONE: La somma dei quadrati degli scarti dalla media aritmetica divisi per la numerosità campionaria

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

LA DEVIAZIONE STANDARD

DEFINIZIONE: La radice quadrata della varianza

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

Calcolare la deviazione standard (DV) delle seguenti 10 osservazioni (mm):

81 79 82 83 80 78 80 87 82 82

1. Si calcoli la media, $\bar{x}$:

$$\bar{x} = \frac{\sum x}{n} = \frac{814}{10} = 81.40$$

2.Si calcolino gli scarti dalla media sottraendo da ciascun valore la media; si elevi al quadrato tale quantità (il quadrato elide il segno -):

$(81-81.4)^2 = 0.16$ $(78-81.4)^2 = 11.56$

$(79-81.4)^2 = 5.76$ $(80-81.4)^2 = 1.96$

$(82-81.4)^2 = 0.36$ $(87-81.4)^2 = 31.36$

$(83-81.4)^2 = 2.56$ $(82-81.4)^2 = 0.36$

$(80-81.4)^2 = 1.96$ $(82-81.4)^2 = 0.36$

3. Si sommino tali quantità: la somma è pari a 56.4. La somma $\sum (x - \overline{x})^2$ è detta somma dei quadrati degli scarti o, più semplicemente, somma dei quadrati.

4. Si divida tale quantità per il numero di osservazioni meno 1:

$$\frac{\text{somma dei quadrati}}{(n-1)} = \frac{\sum\left(x-\bar{x}\right)^2}{n-1} = \frac{56.4}{9} = 6.27$$

5. La deviazione standard è la radice quadrata di tale valore:

$$DS = \sqrt{6.27} = 2.50\,\text{mm}$$

Quindi la deviazione standard del campione di 10 unità estratto dalla popolazione è pari a 2.50 mm.

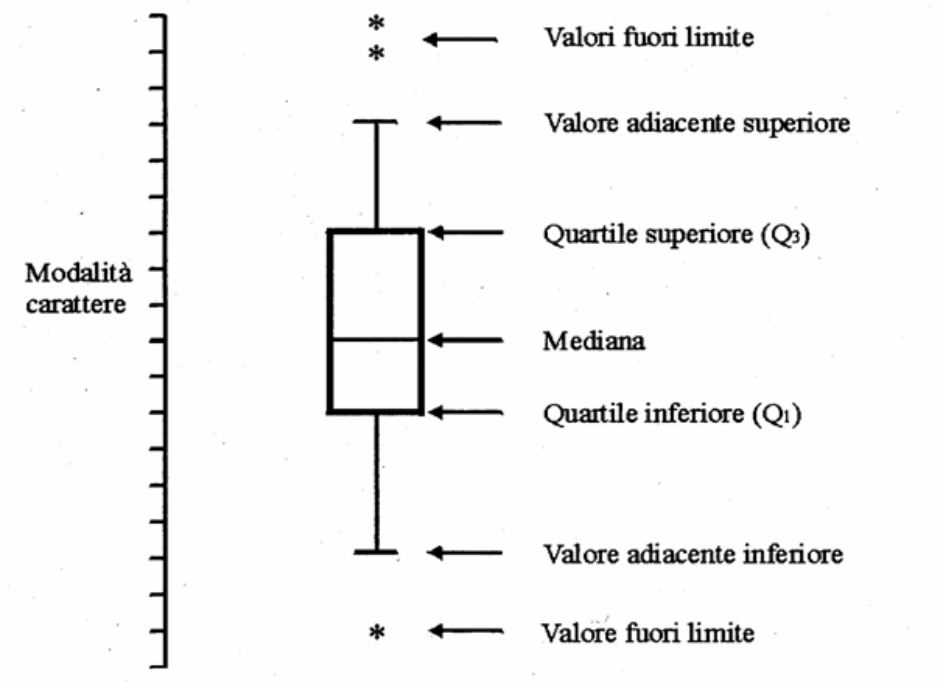
SCARTO INTERQUARTILE (IQR)

Scarto interquartile = (3°quartile)-(1°quartile)

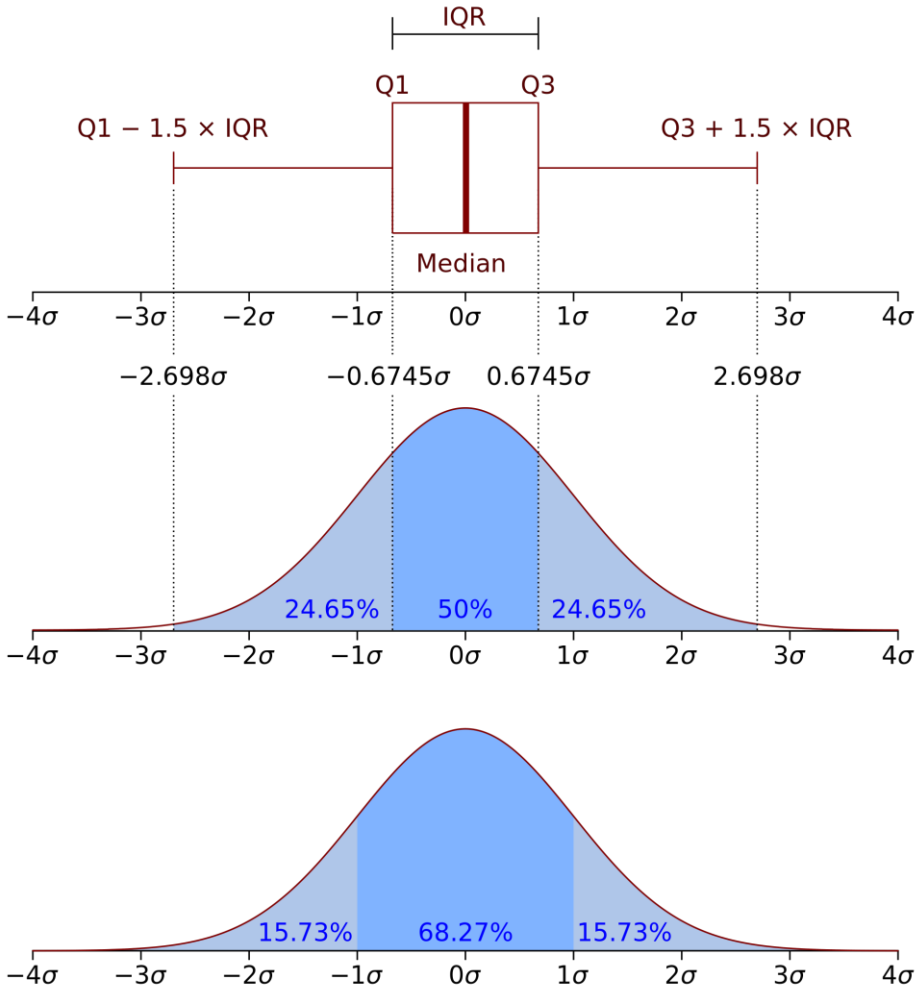
E' molto più *resistente* della varianza in presenza di poche osservazioni estreme. Per questo motivo e usato soprattutto nelle situazioni in cui si sospetta la possibile presenza di osservazioni anomale.

BOX-PLOT

Il nome deriva dall'inglese (*box and whiskers plot* spesso, anche in italiano, abbreviato in *boxplot*).



Scarto interquartile (o **differenza interquartile** o **ampiezza interquartile**, o *IQR*) è la differenza tra il terzo e il primo quartile, ovvero l'ampiezza della fascia di valori che contiene la metà "centrale" dei valori osservati.
Lo scarto interquartile è un indice di dispersione, cioè una misura di quanto i valori si allontanino da un valore centrale. Viene utilizzato nel disegno del diagramma box-plot.



Particolari percentili sono i **quartili**:

Q_1: **primo quartile**: 25-esimo percentile
Q_2: **secondo quartile**: 50-esimo percentile, ovvero la Mediana
Q_3: **terzo quartile**: 75-esimo percentile

Si può dare una definizione quantitativa in termini dei quartili:

Def.: Un valore x del campione si definisce **outlier** se

$$x \leq Q_1 - 1.5(Q_3 - Q_1) \quad \text{oppure} \quad x \geq Q_3 + 1.5(Q_3 - Q_1);$$

in particolare e' detto **outlier debole** se

$$Q_1 - 3(Q_3 - Q_1) < x \leq Q_1 - 1.5(Q_3 - Q_1)$$

oppure

$$Q_3 + 1.5(Q_3 - Q_1) \leq x < Q_3 + 3(Q_3 - Q_1),$$

outlier forte se

$$x \leq Q_1 - 3(Q_3 - Q_1) \text{ oppure} \quad x \geq Q_3 + 3(Q_3 - Q_1)$$

COESIONE ITALIA 21-27 EMILIA-ROMAGNA

Cofinanziato dall'Unione europea

REPUBBLICA ITALIANA

Regione Emilia-Romagna

DATA LAB

GUARDA AVANTI

Big Data, nuove competenze per nuove professioni.

www.bigdata-lab.it

FORM INDUSTRIA EMILIA ROMAGNA

ifoa Sapere utile

REISS ROMOLI la passione della conoscenza

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Università degli Studi di Ferrara

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITÀ DI PARMA

POLITECNICO MILANO 1863 POLO TERRITORIALE DI PIACENZA

UNIVERSITÀ CATTOLICA del Sacro Cuore