# Leggiamo la realtà attraverso la statistica


## VARIABILITÀ

# Indice

# Che cos’è la VARIABILITA’?

**È l’attitudine di un fenomeno ad assumere modalità diverse nelle unità statistiche del collettivo**

Facciamo un esempio:

. consideriamo le altezze di ogni ragazzo nella nostra classe

. calcoliamo la media, la mediana o la moda

questi indici di posizione non ci informano su come le unità tendono ad assumere misure diverse l’una dall’altra,

noi non abbiamo tutti la stessa altezza… **siamo variabili**

# Basta la media per rappresentare i dati?

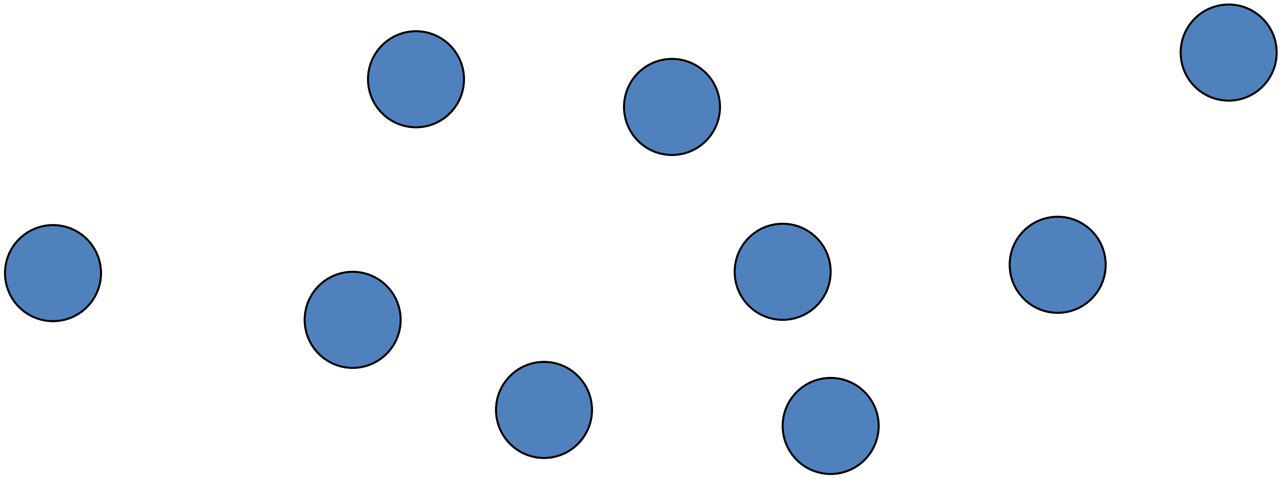

Siamo alla fine del quadrimestre.
Per valutare l'andamento calcoliamo le medie dei voti in matematica.
Ecco i voti ottenuti da tre studenti:

| | | | | | |
|---|---|---|---|---|---|
| MARCO | 6 | 5 | 6 | 7 | MEDIA=6 |
| LUCA | 6 | 6 | 6 | 6 | MEDIA=6 |
| LUIGI | 4 | 4 | 8 | 8 | MEDIA=6 |

Gli studenti hanno la stessa media
ma i voti sono molto... **differenti**

# Ma come si misura la variabilità?

La media non è informativa della variabilità dei dati, dobbiamo COSTRUIRE un INDICE DI VARIABILITA'!

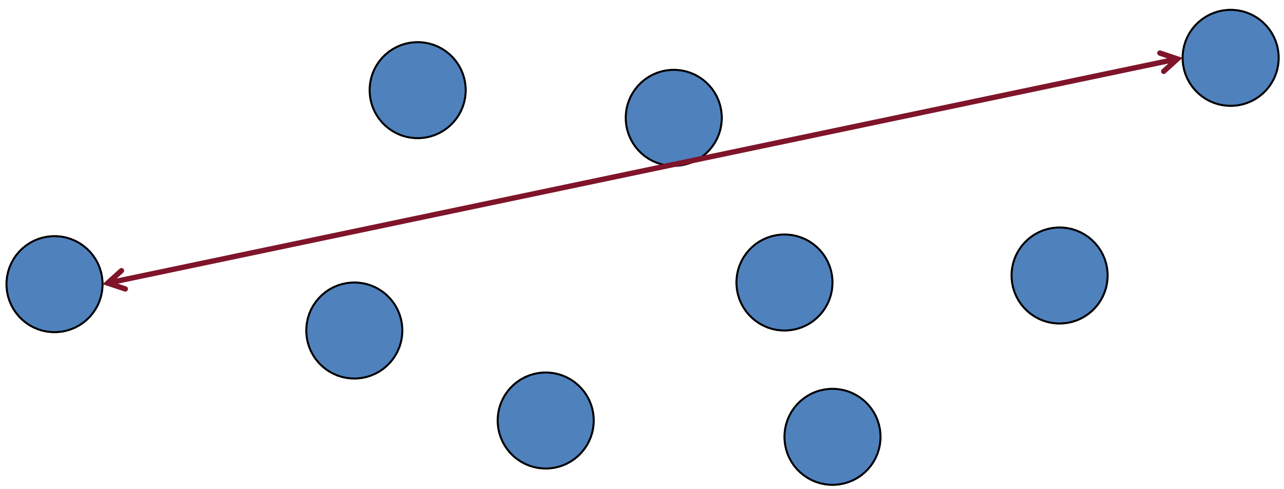


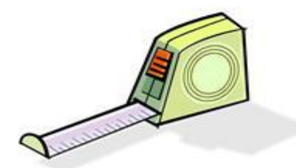
Immaginiamo che ogni pallino sia un unità statistica…

# La variabilità come misura di distanza

La variabilità aumenta all'aumentare della distanza tra le osservazioni

**RANGE o CAMPO DI VARIAZIONE=MAX-min**





È lo strumento più **semplice perché considera la posizione solamente di due unità statistiche**
è la differenza tra il massimo e il minimo osservato!

# Tra i nostri studenti chi è il più variabile?

Proviamo a calcolare subito il **range** prendendo il massimo e sottraendo ad esso il minimo osservato

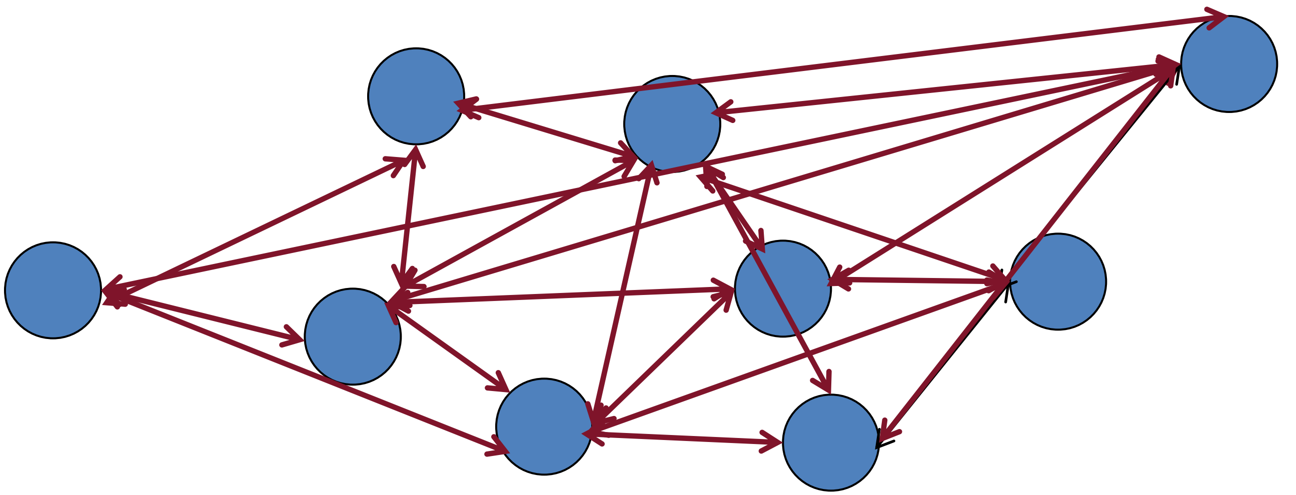
| | | | | | |
|---|---|---|---|---|---|
| MARCO | 6 | 5 | 6 | 7 | RANGE=7-5=2 |
| LUCA | 6 | 6 | 6 | 6 | RANGE=6-6=0 |
| LUIGI | 4 | 4 | 8 | 8 | RANGE=8-4=4 |

Luigi è il più variabile, mentre Luca non varia!

# Il RANGE è una misura immediata ma…

…è molto sensibile
tiene conto solo dei valori estremi

Una misura più adeguata dovrebbe TENER CONTO DI TUTTE LE UNITA' STATISTICHE!



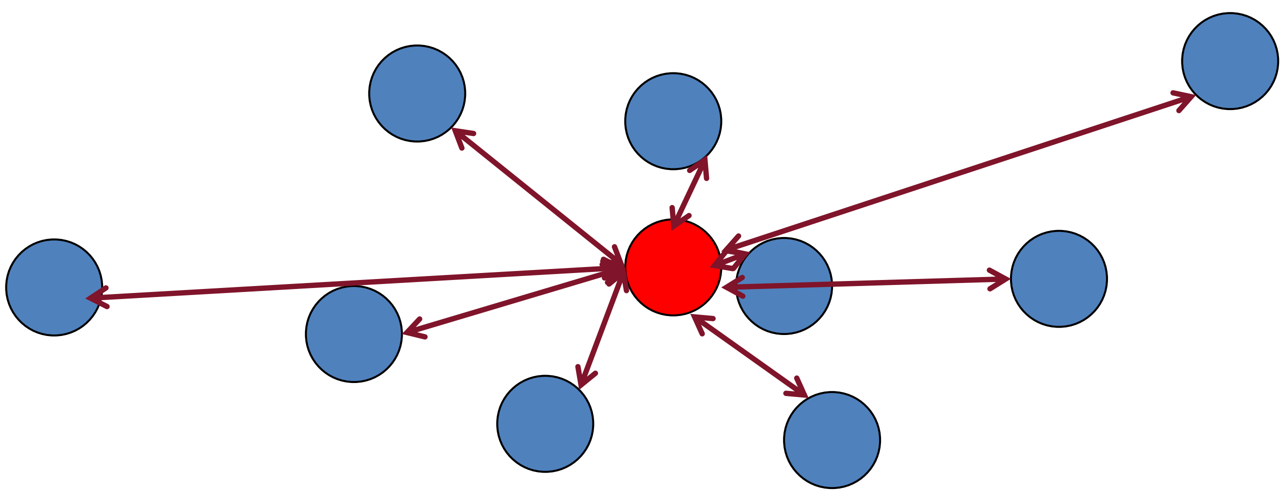
Ma considerare tutte le possibili differenze tra tutte le unità statistiche… è troppo laborioso!

Una soluzione è…

# La variabilità come misura di distanza da un centro

Un indice di variabilità più adeguato può essere una misura della distanza di ogni osservazione dalla media dei dati

La variabilità aumenta all'aumentare della distanza tra le osservazioni e la media.



# Proviamo a costruire un indice di variabilità!

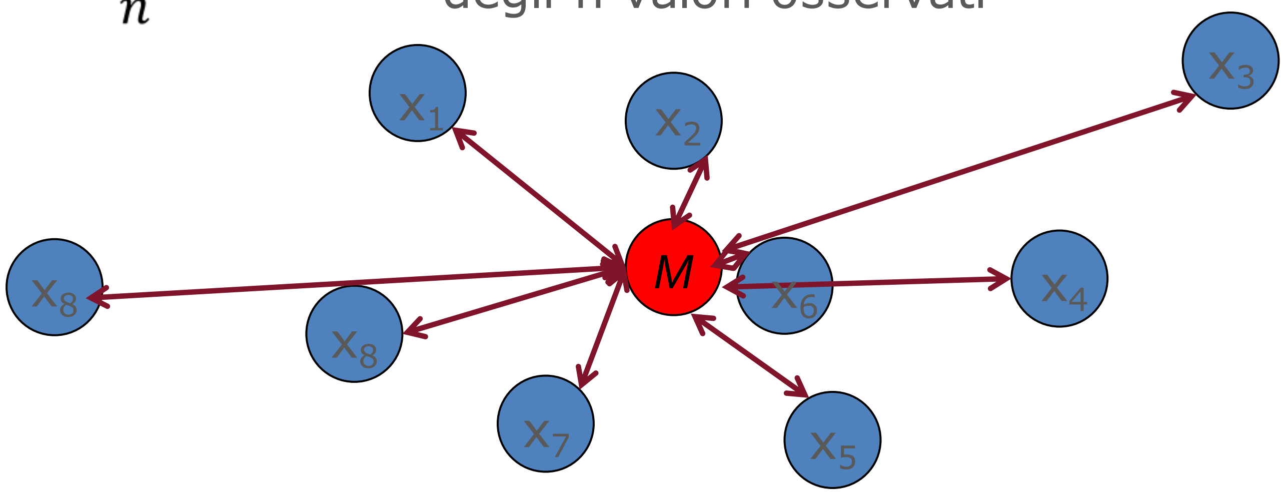
Formalizziamo i nostri dati

X variabile statistica sul fenomeno studiato

$x_1, x_2, x_3, \ldots, x_i, \ldots, x_n$ valori osservati nelle n unità statistiche

$$M = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

è la media aritmetica degli n valori osservati



# Scarti o differenze dalla media

Quale operazione matematica mi permette di misurare una distanza tra due valori osservati? La differenza!

Se prendiamo tutte le differenze **scarti**
e ne facciamo la somma otteniamo però…

$$\left(x_1 - M\right) + \left(x_2 - M\right) + ... + \left(x_n - M\right) = 0$$


La media bilancia i valori più bassi con quelli più alti→
**la somma delle differenze dalla media è NULLA**

# E se eleviamo al quadrato queste differenze?

$$(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2$$

✓ è sempre maggiore di zero

✓ è uguale a zero solo se la variabile non varia

✓ aumenta all'aumentare della distanza delle osservazioni dalla media

✓ aumenta anche all'aumentare del numero delle osservazioni (n)

✓ è espressa nel quadrato dell'unità di misura della variabile

# Varianza

Dividendo per n la somma delle differenze al quadrato

$$\frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n} = \sigma^2$$

si ottiene la **VARIANZA**
viene generalmente indicata con $\sigma^2$ “sigma quadrato”

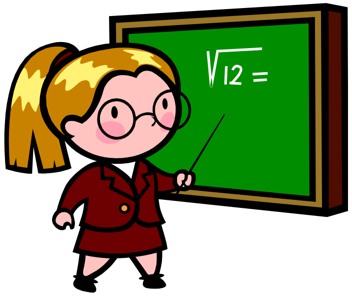
# Scarto quadratico medio *(deviazione standard)*

La varianza è espressa nel quadrato dell'unità di misura della variabile osservata

Si usa più spesso lo **scarto quadratico medio** generalmente indicato con σ "sigma"

Per **n osservazioni** $x_1, x_2, \ldots, x_n$ la formula è

$$\sigma = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}}$$



# Proviamo a fare i calcoli per Marco

Nel caso di una serie di dati

| | **Voti di MARCO** | **Differenze dalla media** | **Quadrati delle differenze** |
|---|---|---|---|
| | 6 | (6-6)=0 | 0 |
| | 5 | (5-6)=-1 | 1 |
| | 6 | (6-6)=0 | 0 |
| | 7 | (7-6)=1 | 1 |
| Somma | 24 | 0 | 2 |
| | | | |
| n=4 | | *Varianza* | *0,5* |
| | | *Scarto quadratico medio* | *0,71* |

# Adesso anche per Luca e Luigi!

| Voti di LUCA | Differenze dalla media | Quadrati delle differenze | Voti di LUIGI | Differenze dalla media | Quadrati delle differenze |
|---|---|---|---|---|---|
| 6 | 0 | 0 | 4 | (4-6)=-2 | 4 |
| 6 | 0 | 0 | 4 | (4-6)=-2 | 4 |
| 6 | 0 | 0 | 8 | (8-6)=2 | 4 |
| 6 | 0 | 0 | 8 | (8-6)=2 | 4 |
| **24** | **0** | **0** | **24** | **0** | **16** |
| | *Varianza* | *0* | | *Varianza* | *4* |
| | *Scarto quadratico medio* | *0* | | *Scarto quadratico medio* | *2* |

Luigi è il più variabile, confermiamo quanto già evidenziato con il RANGE!

# Scarto quadratico medio
# *(deviazione standard)*

Se i dati sono in **tabelle di frequenza**
ogni differenza va pesata con la rispettiva frequenza assoluta

$$\sigma = \sqrt{\frac{(x_1 - M)^2 \cdot n_1 + (x_2 - M)^2 \cdot n_2 + \ldots + (x_k - M)^2 \cdot n_k}{n}}$$

Con i dati dei tre studenti costruiamo una tabella di frequenza e calcoliamo media e scarto quadratico medio, in totale avremo n=12 osservazioni

# Esercitiamoci con i dati in tabella!

| $x_j$ | $n_j$ | $x_j \cdot n_j$ | $(x_j - M)^2$ | $(x_j - M)^2 \cdot n_j$ |
|---|---|---|---|---|
| 4 | 2 | 8 | 4 | 8 |
| 5 | 1 | 5 | 1 | 1 |
| 6 | 6 | 36 | 0 | 0 |
| 7 | 1 | 7 | 1 | 1 |
| 8 | 2 | 16 | 4 | 8 |
| Somma | 12 | 72 | 10 | 18 |

$$M = \frac{72}{12} = 6$$

$$\sigma = \sqrt{\frac{(x_1 - M)^2 \cdot n_1 + (x_2 - M)^2 \cdot n_2 + \ldots + (x_k - M)^2 \cdot n_k}{n}} = \sqrt{\frac{18}{12}} = 1{,}22$$

# Ma è più variabile il peso o l'altezza?

Possiamo confrontare Kg con cm ?

No!

Dobbiamo usare un indice di variabilità come il **coefficiente di variazione**

$$\frac{\sigma}{|M|} \times 100$$

**non ha unità di misura è un numero puro!**

# Marco si confronta con suo fratello!

Suo fratello frequenta l'università e nell'ultimo semestre ha preso i seguenti voti espressi in trentesimi.

| Voti del fratello di Marco | Differenze dalla media | Quadrati delle differenze |
|---|---|---|
| 25 | (25-22)=3 | 9 |
| 20 | (20-22)=-2 | 4 |
| 22 | (22-22)=0 | 0 |
| 21 | (21-22)=-1 | 1 |
| 88 | 0 | 14 |

$$M= \frac{25+20+22+21}{4} = \frac{88}{4} = 22$$

$$\sigma^2 = \frac{(25-22)^2 + (20-22)^2 + (22-22)^2 + (21-22)^2}{4} = \frac{14}{4} = 3{,}5$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{3{,}5} = 1{,}9$$

# Riassumiamo tutto in una tabella!

Chi presenta maggiore variabilità ?

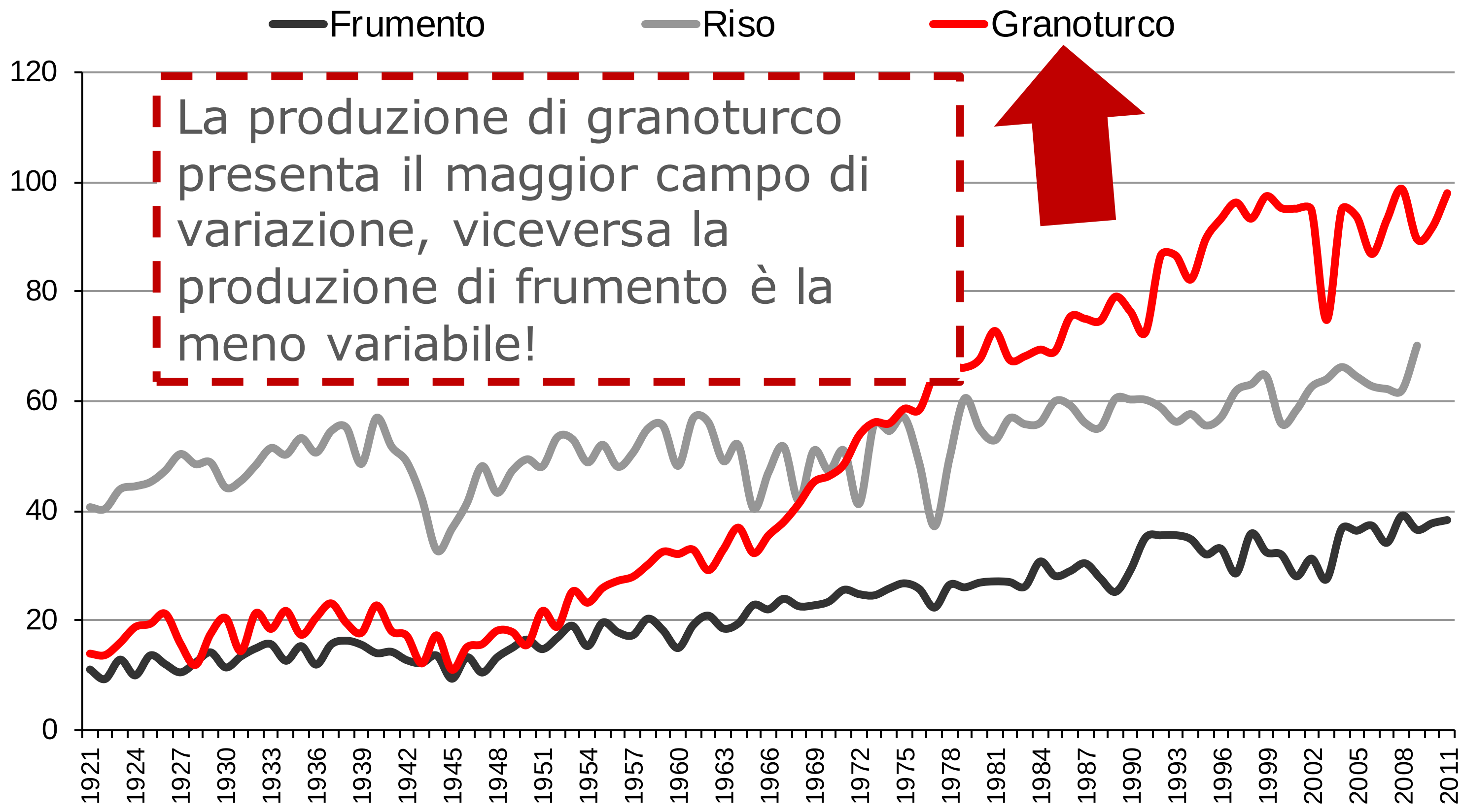
| Indice | Marco | Fratello di Marco |
|---|---|---|
| $M$ | 6 | 22 |
| $\sigma^2$ | 0,5 | 3,5 |
| $\sigma$ | 0,7 | 1,9 |
| $\sigma/M$ | 11,8% | 8,6% |

Chi ha un valore più elevato del coefficiente di variazione: Marco!

# Proviamo con dati reali!

Quale delle tre variabili produzione di frumento, riso,
o granoturco è più variabile?

Produzione di alcuni tipi di cereali - Anni 1921-2011 (quintali per ettaro)




Fonte: Ministero di agricoltura, industria e commercio (fino al 1923); Istituto di economia e statistica agraria (anni 1924-1926); Istat, Stima delle superfici e produzioni delle coltivazioni agrarie, floricole e delle piante intere da vaso (dal 1927)

# …e adesso…
# buon lavoro!