

Varianza

Un altro indice molto noto è la **varianza**, che corrisponde allo **scarto quadratico medio elevato al quadrato**.

Ad esempio, la varianza della distribuzione dei prezzi dell'esempio precedente è pari a

$$\sigma^2 = \frac{1}{N} \sum (x_i - M)^2 n_i = \frac{3103.491}{283} = 10.9664$$

Varianza

Esiste una formula indiretta per il calcolo della varianza.

$$\sigma^2 = M(X^2) - [M(X)]^2 = M_2 - M_1^2$$

media
aritmetica
dei
quadrati

media
aritmetica
al
quadrato

Varianza

Applichiamo questa formula all'esempio appena visto. Prima di tutto calcoliamo la **media aritmetica dei quadrati**.

Prezzo	vc	vc ²	ni	vc ² ·ni
1 - 3 €	2	4	37	148
3 - 5 €	4	16	100	1600
5 - 7 €	6	36	62	2232
7 - 9 €	8	64	39	2496
9 - 11 €	10	100	15	1500
11 - 13 €	12	144	14	2016
13 - 15 €	14	196	12	2352
15 - 17 €	16	256	3	768
17 - 19 €	18	324	1	324
Totale			283	13436

$$\begin{aligned} M(X^2) &= \\ &= \frac{13436}{283} = \\ &= 47.4770 \end{aligned}$$

Varianza

Poi calcoliamo il **quadrato della media aritmetica**. La media aritmetica (calcolata prima) era pari a $M=6.0424$.

$$\sigma^2 = M(X^2) - [M(X)]^2 = M_2 - M_1^2$$

$$\sigma^2 = 47.4770 - 6.0424^2 = 47.4770 - 36.5106$$

$$\sigma^2 = 10.9664$$

Decomposizione della varianza

Supponiamo che una popolazione sia divisa in k gruppi di numerosità note, e che siano state calcolate la media e la varianza di un carattere X all'interno di ciascun gruppo.

M_1	M_2	$\dots$	M_k
σ_1^2	σ_2^2	$\dots$	σ_k^2
n_1	n_2	$\dots$	n_k

Decomposizione della varianza

Immaginiamo, partendo da questi dati, di voler calcolare la media e la varianza di X nell'intera popolazione.

M_1	M_2	...	M_k
σ_1^2	σ_2^2	...	σ_k^2
n_1	n_2	...	n_k

Decomposizione della varianza

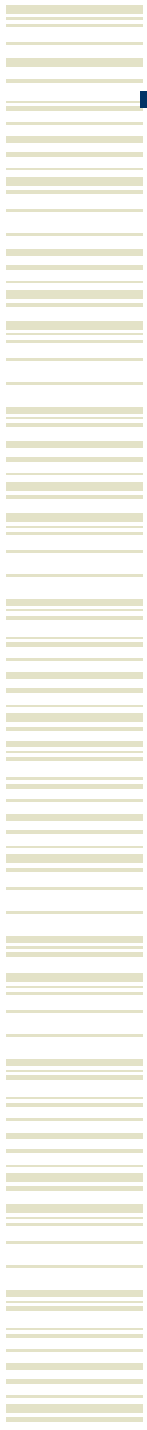
Per la media il problema ha una soluzione molto semplice ed intuitiva: la media generale è data dalla media delle medie parziali, ponderata per la numerosità dei gruppi.

$$M = \frac{M_1 n_1 + M_2 n_2 + \dots + M_k n_k}{n_1 + n_2 + \dots + n_k} = \frac{1}{N} \sum_{i=1}^k M_i n_i$$

Decomposizione della varianza

Per la varianza, invece, vale una regola più complessa.

Devono essere calcolate due grandezze, dette **varianza nei** (VAR_{NEI}) e **varianza fra** (VAR_{FRA}).



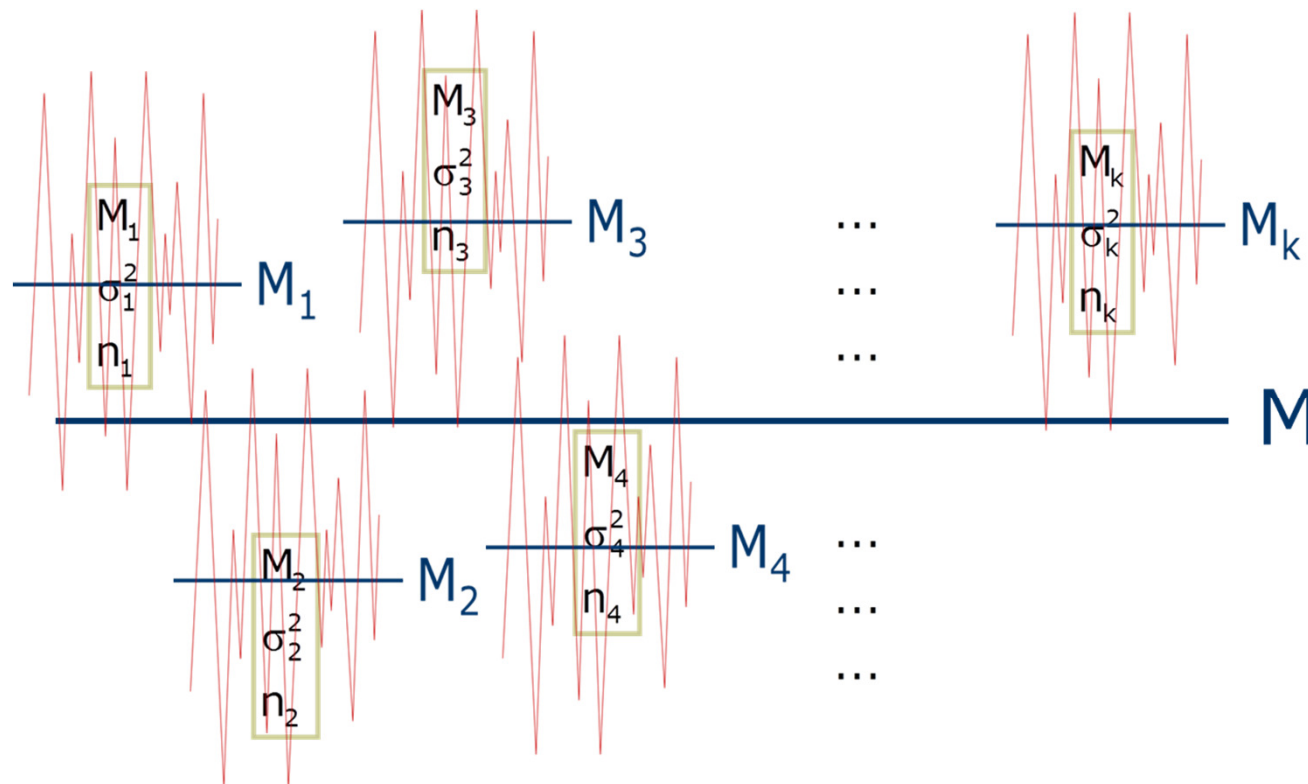
Decomposizione della varianza



Infatti, nel problema che stiamo esaminando, con i soggetti suddivisi in k gruppi, la variabilità totale risente di due tipi di variabilità.

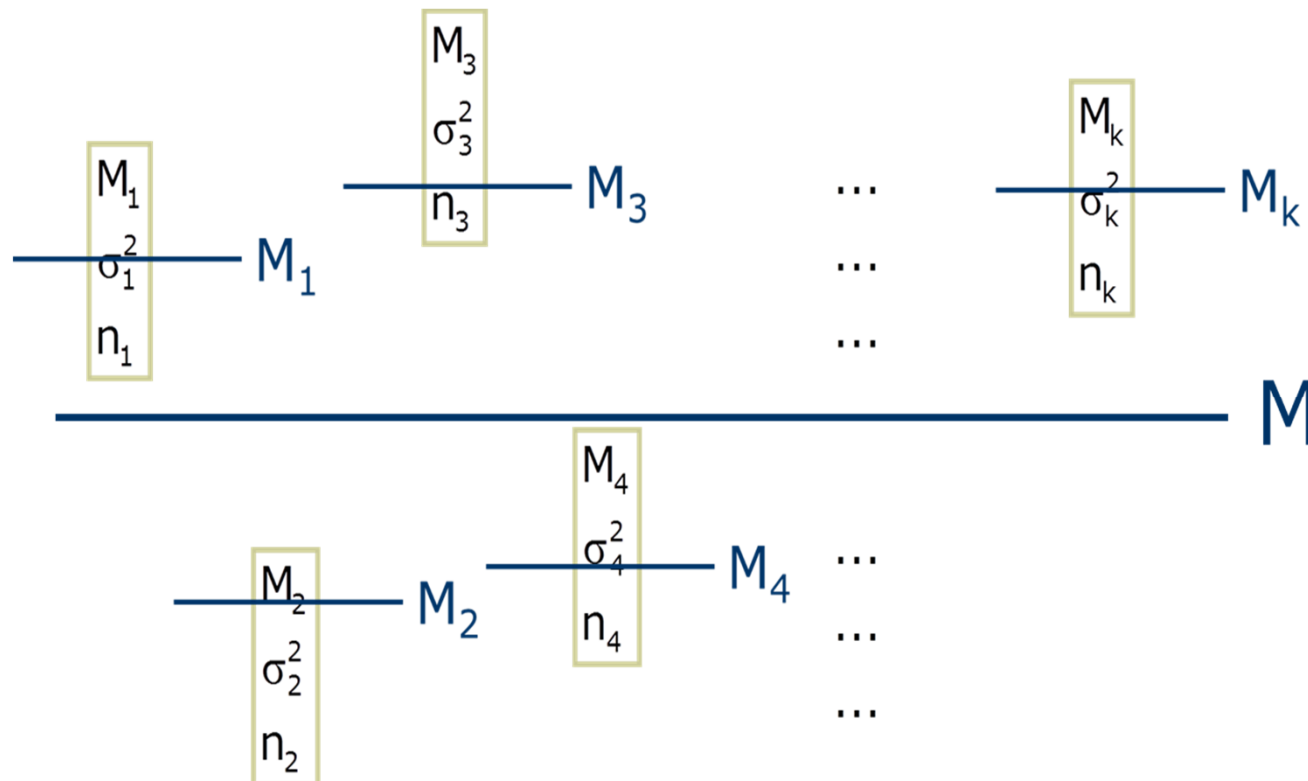
Decomposizione della varianza

C'è una variabilità interna ai gruppi....



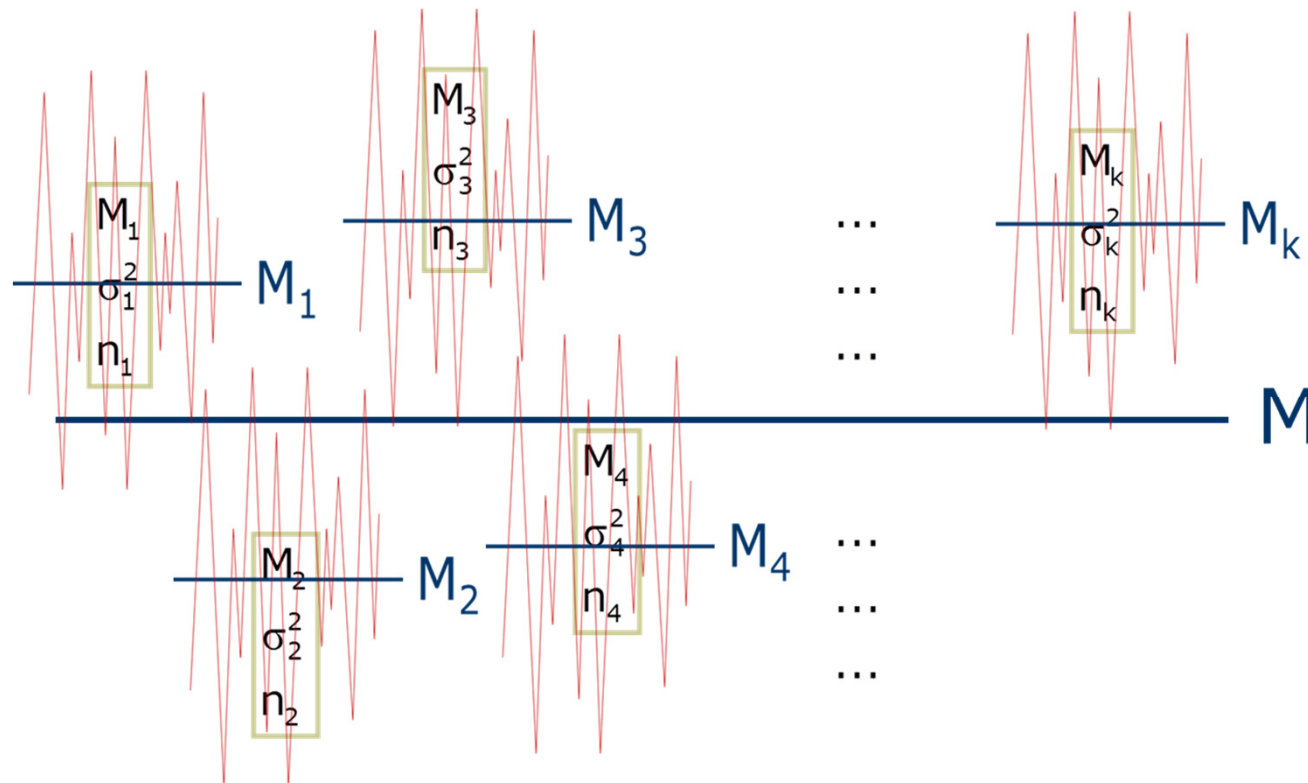
Decomposizione della varianza

... e una variabilità delle medie di gruppo attorno alla media generale



Decomposizione della varianza

La variabilità interna ai gruppi è misurata dalle varianze $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$



Decomposizione della varianza

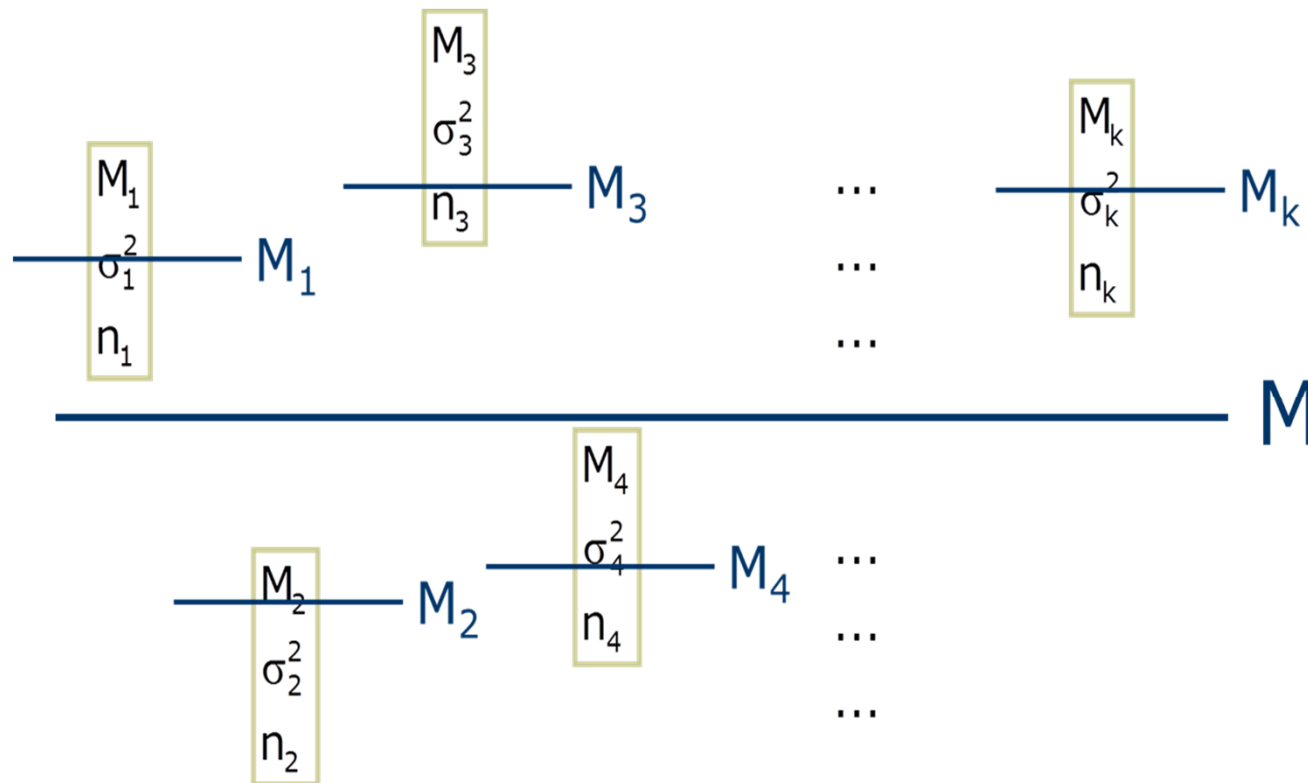
La variabilità interna ai gruppi è misurata dalle varianze $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$

La sua misura di sintesi è una media ponderata delle varianze:

$$VAR_{NEI} = \frac{\sigma_1^2 n_1 + \sigma_2^2 n_2 + \dots + \sigma_k^2 n_k}{n_1 + n_2 + \dots + n_k} = \frac{1}{N} \sum_{i=1}^k \sigma_i^2 n_i$$

Decomposizione della varianza

La variabilità delle medie di gruppo è misurata dalle differenze $(M_1 - M)^2$, $(M_2 - M)^2$, ..., $(M_k - M)^2$



Decomposizione della varianza

La variabilità delle medie di gruppo è misurata dalle differenze $(M_1 - M)^2, (M_2 - M)^2, \dots, (M_k - M)^2$

La sua misura di sintesi è una media ponderata di tali differenze:

$$\begin{aligned} VAR_{FRA} &= \frac{(M_1 - M)^2 n_1 + (M_2 - M)^2 n_2 + \dots + (M_k - M)^2 n_k}{n_1 + n_2 + \dots + n_k} = \\ &= \frac{1}{N} \sum_{i=1}^k (M_i - M)^2 n_i \end{aligned}$$

Decomposizione della varianza

$$VAR_{NEI} = \frac{\sigma_1^2 n_1 + \sigma_2^2 n_2 + \dots + \sigma_k^2 n_k}{n_1 + n_2 + \dots + n_k} = \frac{1}{N} \sum_{i=1}^k \sigma_i^2 n_i$$

$$VAR_{FRA} = \frac{(M_1 - M)^2 n_1 + (M_2 - M)^2 n_2 + \dots + (M_k - M)^2 n_k}{n_1 + n_2 + \dots + n_k} =$$
$$= \frac{1}{N} \sum_{i=1}^k (M_i - M)^2 n_i$$

Decomposizione della varianza

La varianza dell'intera popolazione (VAR_{TOT}) è data dalla somma della varianza nei (VAR_{NEI}) e della varianza fra (VAR_{FRA}).

$$\sigma^2 = VAR_{TOT} = VAR_{NEI} + VAR_{FRA}$$

Decomposizione della varianza

Questa regola è nota come **decomposizione della varianza** e ci sarà utile più avanti, quando inizieremo ad occuparci di statistica bivariata.

$$\sigma^2 = \text{VAR}_{\text{TOT}} = \text{VAR}_{\text{NEI}} + \text{VAR}_{\text{FRA}}$$

Decomposizione della varianza

N=100 vini sono stati sottoposti a un'analisi sensoriale, in base alla quale sono stati classificati in tre categorie (A, B, e C) caratterizzate da apprezzamento complessivo crescente da parte dei giudici. Per ogni bottiglia era noto il prezzo di vendita ed è stata creata una tabella riportante medie e varianze dei prezzi delle tre categorie.

Decomposizione della varianza

	n_i Numero di vini	M_i Prezzi Medi	σ_i^2 Varianze dei prezzi
A	40	17.5	0.5
B	40	18	0.9
C	20	18	0.8

Calcoliamo prezzo medio e varianza dei 100 vini.

Decomposizione della varianza

$$M = \frac{17.5 \times 40 + 18 \times 40 + 18 \times 20}{100} = 17.8$$

$$VAR_{NEI} = \frac{0.5 \times 40 + 0.9 \times 40 + 0.8 \times 20}{100} = 0.72$$

$$VAR_{FRA} = \frac{(17.5 - 17.8)^2 \times 40 + (18 - 17.8)^2 \times 40 + (18 - 17.8)^2 \times 20}{100} = 0.06$$

$$VAR_{TOT} = 0.72 + 0.06 = 0.178$$

Indici di variabilità relativi

Gli intervalli di variazione e gli scostamenti medi che abbiamo visto fin qui risentono di un problema: sono **sensibili all'ordine di grandezza e all'unità di misura** in cui è rilevato il carattere.

Pertanto non si possono utilizzare per confrontare la variabilità di caratteri con ordini di grandezza e/o unità di misura differenti.

Indici di variabilità relativi

Le due tabelle che seguono riportano la produzione di uva di due aziende diverse in 4 anni consecutivi, la prima misurata in quintali, la seconda in kg.

Azienda 1 (q)	
anno 1	400
anno 2	420
anno 3	380
anno 4	410

Azienda 2 (Kg)	
anno 1	40000
anno 2	42000
anno 3	38000
anno 4	41000

Indici di variabilità relativi

Si nota immediatamente che le due aziende hanno prodotto ogni anno esattamente le stesse quantità di uva.

Azienda 1 (q)	
anno 1	400
anno 2	420
anno 3	380
anno 4	410

Azienda 2 (Kg)	
anno 1	40000
anno 2	42000
anno 3	38000
anno 4	41000

Indici di variabilità relativi

Valutiamo ora la variabilità della produzione utilizzando lo scarto quadratico medio.

Azienda 1 (q)	
anno 1	400
anno 2	420
anno 3	380
anno 4	410

Azienda 2 (Kg)	
anno 1	40000
anno 2	42000
anno 3	38000
anno 4	41000

Indici di variabilità relativi

Azienda 1 (q)	
anno 1	400
anno 2	420
anno 3	380
anno 4	410

$$M(X) = \frac{400 + 420 + 380 + 410}{4} = 402.5$$

$$\sigma = \sqrt{\frac{(400 - 402.5)^2 + (420 - 402.5)^2 + (380 - 402.5)^2 + (410 - 402.5)^2}{4}} = 14.7902$$

Più precisamente, lo scarto quadratico medio è pari a **14.7902 q**.

Indici di variabilità relativi

Azienda 2 (Kg)	
anno 1	40000
anno 2	42000
anno 3	38000
anno 4	41000

$$M(X) = \frac{40000 + 42000 + 38000 + 41000}{4} = 40250$$

$$\sigma = \sqrt{\frac{(40000 - 40250)^2 + \dots + (41000 - 40250)^2}{4}} = 1479.02$$

Lo scarto quadratico medio è pari a
1479.02 Kg.

Indici di variabilità relativi

Se ci limitassimo a confrontare la variabilità delle due produzioni semplicemente facendo riferimento allo scarto quadratico medio, dovremmo concludere che la seconda azienda ha una produzione molto più variabile.

Azienda 1 (q)
$\sigma = 14.7902$

Azienda 2 (Kg)
$\sigma = 1479.02$

Indici di variabilità relativi

Quand'anche l'unità di misura fosse la stessa, anche l'ordine di grandezza può comportare distorsioni. Vediamo un altro esempio.

Azienda 1 (q)	
anno 1	400
anno 2	420
anno 3	380
anno 4	410

Azienda 3 (q)	
anno 1	5000
anno 2	5020
anno 3	4980
anno 4	5010

Indici di variabilità relativi

Osservando i dati notiamo che le variazioni assolute sono le stesse per le due aziende.

Ma, **relativamente all'ordine di grandezza**, le variazioni per l'Azienda 1 sono più consistenti....

Azienda 1 (q)		
anno 1	400	-
anno 2	420	+20
anno 3	380	-40
anno 4	410	+30

Azienda 3 (q)		
anno 1	5000	-
anno 2	5020	+20
anno 3	4980	-40
anno 4	5010	+30

Indici di variabilità relativi

Calcoliamo lo scarto quadratico medio della produzione dell'Azienda 3 (quello dell'Azienda 1 lo abbiamo già calcolato).

Azienda 1 (q)		
anno 1	400	-
anno 2	420	+20
anno 3	380	-40
anno 4	410	+30

Azienda 3 (q)		
anno 1	5000	-
anno 2	5020	+20
anno 3	4980	-40
anno 4	5010	+30

Indici di variabilità relativi

Azienda 3 (q)	
anno 1	5000
anno 2	5020
anno 3	4980
anno 4	5010

$$M(X) = \frac{5000 + 5020 + 4980 + 5010}{4} = 5002.5$$

$$\sigma = \sqrt{\frac{(5000 - 5002.5)^2 + \dots + (5010 - 5002.5)^2}{4}} = 14.7902$$

Lo scarto quadratico medio è pari a
14.7902 q.

Indici di variabilità relativi

Ancora una volta, se ci limitassimo a confrontare la variabilità delle due produzioni semplicemente facendo riferimento allo scarto quadratico medio, avremmo un'idea non del tutto corretta, nonostante le due misure siano entrambe espresse in q.

Azienda 1 (q)
$\sigma = 14.7902$

Azienda 3 (q)
$\sigma = 14.7902$

Indici di variabilità relativi

Il confronto potrebbe divenire addirittura assurdo se volessimo confrontare due caratteri completamente differenti.

	prezzo in €	Acidità volatile (g/l)
Den	PM	Chim3
Dolcetto1	5.00	0.33
Dolcetto2	6.45	0.53
Dolcetto3	6.70	0.37
Dolcetto4	8.50	0.74
Dolcetto5	3.95	0.31
Dolcetto6	7.18	0.31
Dolcetto7	4.70	0.3
Dolcetto8	3.90	0.37
Dolcetto9	3.75	0.39
Dolcetto10	3.80	0.34
Dolcetto11	12.00	0.47

Indici di variabilità relativi

Calcoliamo gli scarti quadratici medi di **PM** e **Chim3**

$$\begin{aligned}\sigma(\text{PM}) &= 2.4407 \\ \sigma(\text{Chim3}) &= 0.1257\end{aligned}$$

o, più precisamente

$$\begin{aligned}\sigma(\text{PM}) &= 2.4407 \text{ €} \\ \sigma(\text{Chim3}) &= 0.1257 \text{ g/l}\end{aligned}$$

Indici di variabilità relativi

Possiamo affermare che il prezzo sia più variabile del carattere chimico Acidità volatile?

No: abbiamo sia un problema di unità di misura diverse, sia un problema di ordini di grandezza diversi !!

Indici di variabilità relativi

In tutti questi esempi, la soluzione è utilizzare un **indice di variabilità relativo**.

Ci sono vari modi di rendere relativo un indice di variabilità, quello più semplice consiste nel **rapportarlo alla media del carattere**.

Indici di variabilità relativi

Questo accorgimento si può utilizzare sia con gli intervalli di variazione, sia con gli scostamenti medi.

Tuttavia l'indice più utilizzato è il **coefficiente di variazione φ** , ottenuto dal rapporto tra scarto quadratico medio e media.

$$\varphi = \frac{\sigma}{M}$$

Indici di variabilità relativi

Rapportando lo scarto quadratico medio alla media:

- si ottiene un **indice puro**, cioè senza unità di misura (l'unità di misura è presente sia a numeratore, sia a denominatore, perciò si semplifica)
- **si tiene conto dei diversi ordini di grandezza.**

Indici di variabilità relativi

Vediamo come si sistemano le cose nei tre esempi visti prima. Il **primo esempio**:

Azienda 1 (q)
$\sigma = 14.7902 \text{ q}$
$M = 402.5 \text{ q}$
$\varphi = \mathbf{0.0367}$

Azienda 2 (Kg)
$\sigma = 1479.02 \text{ Kg}$
$M = 40250 \text{ Kg}$
$\varphi = \mathbf{0.0367}$

Indici di variabilità relativi

Vediamo come si sistemano le cose nei tre esempi visti prima. Il **secondo esempio**:

Azienda 1 (q)
$\sigma = 14.7902 \text{ q}$
$M = 402.5 \text{ q}$
$\varphi = \mathbf{0.0367}$

Azienda 3 (q)
$\sigma = 14.7902 \text{ q}$
$M = 5002.5 \text{ q}$
$\varphi = \mathbf{0.0030}$

Indici di variabilità relativi

Vediamo come si sistemano le cose nei tre esempi visti prima. Il **terzo esempio**:

Prezzo
$\sigma = 2.4407 \text{ €}$
$M = 5.99 \text{ €}$
$\varphi = \mathbf{0.4072}$

Acidità volatile
$\sigma = 0.1257 \text{ g/l}$
$M = 0.41 \text{ g/l}$
$\varphi = \mathbf{0.3099}$