

MORE ABOUT VARIANCE AND STANDARD DEVIATION

Let us define the vector ξ whose n components are the $x_i - \bar{x}$, i.e. the vector

$$\xi = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

We may write the **sample variance** as

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

and the **standard deviation** as

$$s = \sqrt{s^2} = \frac{|\xi|}{\sqrt{n-1}}$$

Note that the standard deviation is measured in the same units as the variables x_i and their mean \bar{x} .

UN ALTRA PROPRIETA' DELLA VARIANZA CAMPIONARIA

Se si somma una costante c a tutti i dati la varianza non cambia.

In formule: data una variabile x di taglia n con valori x_1, \dots, x_n si definisce la variabile y con valori $y_i = x_i + c$, $i = 1, \dots, n$.

Sappiamo che $\bar{y} = \bar{x} + c$ quindi

$$\begin{aligned}s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i + c - \bar{x} - c)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = s_x^2\end{aligned}$$

Utile per fare i calcoli.

ANCORA UNA PROPRIETA' DELLA VARIANZA CAMPIONARIA

Si moltiplicano tutti i valori per un numero $c \neq 0$ e si ottiene la variabile numerica y con valori

$$y_i = cx_i, \quad i = 1, 2, \dots, n$$

La media campionaria $\bar{y} = c\bar{x}$ e la varianza

$$\begin{aligned} s_y^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n (cx_i)^2 - n(c\bar{x})^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n c^2 x_i^2 - n c^2 \bar{x}^2 \right) \\ &= c^2 \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) = c^2 s_x^2 \end{aligned}$$

RIEPILOGO

Data una variabile numerica x di taglia n con valori

$$x_1, x_2, \dots, x_n$$

e due numeri non nulli a e b , la variabile numerica y di taglia n con valori

$$y_1 = ax_1 + b, y_2 = ax_2 + b, \dots, y_n = ax_n + b$$

valgono le seguenti relazioni:

$$\bar{y} = a\bar{x} + b, \quad s_y^2 = a^2 s_x^2$$

che possono essere direttamente verificate come **Esercizio**.

La seconda relazione può anche essere espressa in termini di scarto quadratico come $s_y = |a|s_x$ dove $|a| = \sqrt{a^2}$ è il modulo (o valore assoluto) di a .

Esercizio di ricapitolazione 1

Si consideri il seguente campione di taglia $n = 9$:

1 6 5 6 2 5 2 6 3

Si calcoli la **mediana**, la **moda**, la **media campionaria**, la **varianza** e lo **scarto quadratico**.

La **media campionaria** è :

$$\bar{x} = \frac{1}{9}(1 + 6 + 5 + 6 + 2 + 5 + 2 + 6 + 3) = \frac{36}{9} = 4$$

Ordiniamo i dati in ordine crescente

1 2 2 3 5 5 6 6 6

la **mediana** corrisponde al valore centrale per cui $m = x_5 = 5$ e la **moda** corrisponde al valore della classe più numerosa per cui è uguale a 6.

Calcoliamo ora la **varianza campionaria s^2** :

Usando la formula $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ otteniamo:

$$s^2 = \frac{1}{8} (9 + 4 + 4 + 1 + 1 + 1 + 4 + 4 + 4) = \frac{32}{8} = 4$$

Usando la formula $s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$:

$$s^2 = \frac{1}{8} (1 + 4 + 4 + 9 + 25 + 25 + 36 + 36 + 36 - 144) = 4$$

Lo **scarto quadratico** è semplicemente **s** = 2.

Esercizio di ricapitolazione 2

Le temperature alle ore 6 del primo gennaio sono state rilevate all'Aquila negli ultimi 100 anni. In questo esercizio le temperature sono le variabili x_i con $i = 1, 2, \dots, n$ e $n = 100$.

Dai dati risulta che la temperatura media, calcolata in gradi Celsius, è stata $\bar{x} = -2,4$ con scarto quadratico $s_x = 4,4$ (si noti che a volte si usa dire che la temperatura è stata di $-2,4 \pm 4,4$).

Quale è stata la temperatura media e lo scarto quadratico in gradi Fahrenheit?

Le temperature y_i in gradi Fahrenheit si ottengono dalle temperature x_i in gradi Celsius dalla relazione $y_i = ax_i + c$ con $a = 1,8$ e $c = 32$ (alla pressione di una atmosfera l'acqua congela a 32 gradi e bolle a 212 gradi Fahrenheit).

Si avrà $\bar{y} = a\bar{x} + c = -2,4 \cdot 1,8 + 32 = 27,68$ e $s_y = as_x = 4,4 \cdot 1,8 = 7,92$.

COPPIE DI DATI

Spesso nell'indagine statistica si eseguono analisi di tipo comparativo: si osservano più variabili su un medesimo gruppo di individui.

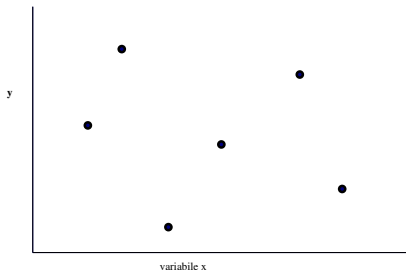
Esempio. Una ragazza vuole acquistare un'automobile usata di un certo modello. Rileva in **5 giorni** diversi gli annunci sul giornale registrando il **numero di anni dell'auto (variabile numerica x)** e il **prezzo in migliaia di euro (variabile numerica y)**.

Anni: x	9	12	8	15	6
Costo: y	18	8	17	7	20

Si possono considerare le due variabili separatamente producendo grafici e indici statistici separatamente per entrambi, ma il problema tipico, in questo caso, è verificare se esiste una dipendenza (**correlazione**) tra le due variabili.

Il primo passo utile per indagare qualitativamente l'eventuale dipendenza delle due variabili consiste nel disegnare un grafico detto **diagramma di dispersione** o **scatterplot**.

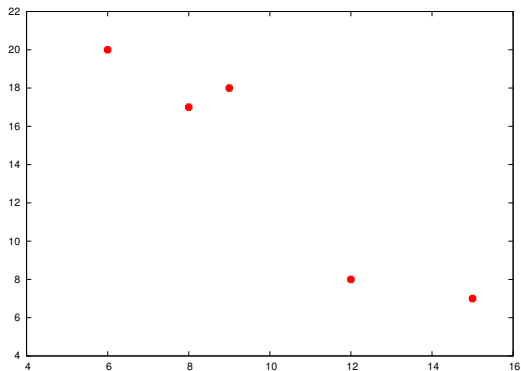
Si pongono in ascissa i dati relativi alla variabile x , in ordinata quelli relativi alla variabile y e si rappresentano con punti le singole osservazioni. Se esiste una relazione semplice tra le due variabili, il diagramma dovrebbe evidenziarla.



In questo esempio (che non è quello dell'auto) sembra che non vi sia una correlazione: i punti **sono sparsi senza apparenti regolarità**.

Invece nel caso dell'auto da acquistare:

Anni (x)	9	12	8	15	6
Costo (y)	18	8	17	7	20



Al crescere degli anni il prezzo diminuisce (non proprio inaspettato!).

COVARIANZA CAMPIONARIA

Date n osservazioni congiunte di due variabili x ed y che indichiamo con

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

il seguente numero si chiama **covarianza campionaria** delle due variabili

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Se si vuole la somma in forma esplicita si può scrivere:

$$s_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n-1}.$$

Given the vectors

$$\boldsymbol{\xi} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

and

$$\boldsymbol{\eta} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$$

we may rewrite the **sample covariance** as

$$s_{xy} = \frac{\boldsymbol{\xi} \cdot \boldsymbol{\eta}}{n - 1}.$$

La **covarianza** è positiva se dati sono ordinati in modo approssimativamente crescente nello scatter plot, negativa se sono ordinati in modo approssimativamente decrescente.

Esempio. Acquisto di un'automobile usata.

Anni (x)	9	12	8	15	6
Costo (y)	18	8	17	7	20

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = \frac{9 + 12 + 8 + 15 + 6}{5} = 10$$

$$\bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = \frac{18 + 8 + 17 + 7 + 20}{5} = 14$$

La covarianza è quindi

$$\begin{aligned}s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\&= \frac{(-1)(4) + (2)(-6) + (-2)(3) + (5)(-7) + (-4)(6)}{4} = \\&= \frac{-4 - 12 - 6 - 35 - 24}{4} = -20,25\end{aligned}$$

La covarianza è negativa perché i dati sono ordinati in modo approssimativamente decrescente.

UNA PROPRIETÀ DELLA COVARIANZA

È facile verificare (stessi passi che abbiamo usato per la varianza) che

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

che implica che

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

ANCORA UNA PROPRIETÀ DELLA COVARIANZA

È facile verificare (stessi passi che abbiamo usato per la varianza) che la **covarianza** di coppie di dati

$$(ax_i + c, by_i + d), \quad i = 1, \dots, n$$

è uguale a

$$ab \cdot s_{xy}$$

Esercizio. Si mostri che la covarianza dei dati che si ottengono da

Anni (x)	9	12	8	15	6
Costo (y)	18	8	17	7	20

con le trasformazioni lineari

$$(3x_i - 18, 2y_i - 40)$$

è uguale a $3 \cdot 2 \cdot (-20, 25) = -121, 5$

Con le trasformazioni lineari si ottengono le seguenti coppie di dati:

	9	18	6	27	0
	-4	-24	-6	-26	0

Le **medie** sono 12 (prima riga) e -12 (seconda riga).

La **covarianza** è

$$\begin{aligned} &= \frac{(-3)(8) + (6)(-12) + (-6)(6) + (15)(-14) + (-12)(12)}{4} = \\ &= \frac{-24 - 72 - 36 - 210 - 144}{4} = \mathbf{-121,5} \end{aligned}$$