

COEFFICIENTE DI CORRELAZIONE

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 **coefficiente di correlazione**

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

dove s_{xy} è la covarianza campionaria

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

mentre s_x ed s_y sono le deviazioni standard di x e di y :

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}$$

ed

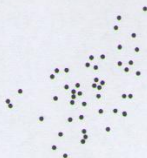
$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})}$$



Correlazione $r = 0$



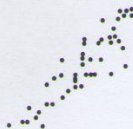
Correlazione $r = -0.3$



Correlazione $r = 0.5$



Correlazione $r = -0.7$



Correlazione $r = 0.9$



Correlazione $r = -0.99$

Si noti che la correlazione è positiva se dati sono ordinati in modo approssimativamente crescente nello scatterplot, negativa se sono ordinati in modo approssimativamente decrescente e approssimativamente nulla quando sono disordinati.

Si noti inoltre che il coefficiente di correlazione è sempre compreso tra -1 e 1. Il valore -1 corrisponde a dati disposti su una retta decrescente mentre il valore 1 a dati disposti su una retta crescente.

UNA PROPRIETÀ DEL COEFFICIENTE DI CORRELAZIONE

- Se r è il coefficiente di correlazione delle coppie di dati (x_i, y_i) , $i = 1, \dots, n$ allora il coefficiente di correlazione delle coppie di dati

$$(ax_i + c, by_i + d), \quad a \neq 0, b \neq 0$$

è

$$\text{sign}(ab) \cdot r$$

dove $\text{sign}(ab)$ indica il segno del prodotto ab . In altre parole, la correlazione rimane la stessa in modulo e non cambia di segno se il prodotto ab è positivo mentre cambia di segno se il prodotto ab è negativo.

Questa proprietà implica che r **non dipende dalla scelta dell'unità di misura**. Per esempio il coefficiente di correlazione tra l'altezza ed il peso di una persona non cambia se l'altezza è misurata in pollici o in centimetri, né se il peso è misurato in chilogrammi o libbre.

Dimostrare questa proprietà è semplice, abbiamo infatti già visto che per le variabili

$$(ax_i + c, by_i + d), \quad a \neq 0, b \neq 0$$

varianze e covarianza si trasformano nel modo seguente

$$s_x \rightarrow |a| \cdot s_x, \quad s_y \rightarrow |b| \cdot s_y, \quad s_{xy} \rightarrow ab \cdot s_{xy}$$

e dato che $r = s_{xy} / (s_x \cdot s_y)$ si avrà:

$$r \rightarrow \frac{ab \cdot s_{xy}}{|a| \cdot s_x \cdot |b| \cdot s_y} = \frac{ab}{|a| \cdot |b|} r = \text{sign}(ab) \cdot r$$

UNA SECONDA PROPRIETÀ DEL COEFFICIENTE DI CORRELAZIONE

- Il coefficiente di correlazione $r=1$ se esistono un numero $b > 0$ ed un numero a qualsiasi tali che

$$y_i = bx_i + a, \quad i = 1, 2, \dots, n$$

Questa proprietà ci dice che c'è una **correlazione lineare** positiva tra le coppie di dati (i dati sono su una retta con coefficiente positivo).

- Il coefficiente di correlazione $r=-1$ se esistono un numero $b < 0$ ed un numero a qualsiasi tali che

$$y_i = bx_i + a, \quad i = 1, 2, \dots, n$$

Cioè la **correlazione è lineare** ma negativa (i dati sono su una retta con coefficiente negativo).

Anche in questo caso la dimostrazione di questa proprietà è semplice. Notiamo preliminarmente che la correlazione tra coppie di variabili identiche ($x_i = y_i$ per ogni i) è 1. Infatti in tal caso avremo $s_x = s_y$ e $s_{xy} = s_x^2$ pertanto

$$r = \frac{s_x^2}{s_x \cdot s_x} = 1$$

Se consideriamo le coppie di variabili

$$(x_i, bx_i + a), \quad b \neq 0$$

per la proprietà precedentemente dimostrata si avrà che il coefficiente di correlazione è

$$r = \frac{b \cdot s_x^2}{|b|s_x \cdot s_x} = \text{sign}(b)$$

che è appunto quello che volevamo dimostrare.

A THIRD PROPERTY OF THE CORRELATION

- The correlation coefficient satisfies $r \in [-1, 1]$.

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

$$\xi = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}), \quad \eta = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$$

we have

$$s_x = \frac{|\xi|}{\sqrt{n-1}}, \quad s_y = \frac{|\eta|}{\sqrt{n-1}}, \quad s_{xy} = \frac{\xi \cdot \eta}{n-1},$$

and, therefore,

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\xi \cdot \eta}{|\xi| |\eta|} = \cos \theta$$

where θ is the angle between ξ and η so that $|r| = |\cos \theta| \leq 1$.

Ultimately, for linear correlations $|r| = 1$, otherwise $r < 1$.

ANCORA UNA PROPRIETÀ

Abbiamo già visto che

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

Da cui si deduce che (si ricordi la definizione $r = s_{xy}/(s_x \cdot s_y)$)

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}$$

Esempio. Acquisto di un'automobile usata.

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

Quale è il coefficiente di correlazione? (abbiamo già visto che la covarianza è negativa e per definizione la correlazione ha lo stesso segno della varianza)

Possiamo sottrarre a tutti i dati x_i una costante e altrettanto fare con i dati y_i , abbiamo infatti visto che questo non modifica il coefficiente di correlazione.

Sottraiamo 9 ai dati di x e sottraiamo 18 ai dati di y , otteniamo:

Anni (x)	0	3	-1	6	-3
Costo (y)	0	-10	-1	-11	2

Per comodità chiamiamo le nuove variabili ancora x_i e y_i .

$$\bar{x} = \frac{3 - 1 + 6 - 3}{5} = 1, \quad \bar{y} = \frac{-10 - 1 - 11 + 2}{5} = -4$$

$$\sum_{i=1}^5 x_i y_i = -30 + 1 - 66 - 6 = -101$$

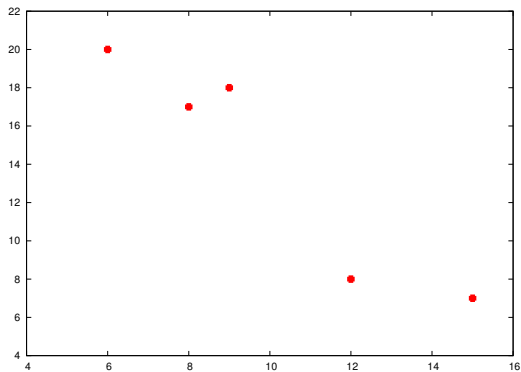
$$\sum_{i=1}^5 x_i^2 = 9 + 1 + 36 + 9 = 55, \quad \sum_{i=1}^5 y_i^2 = 100 + 1 + 121 + 4 = 226$$

$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2\right) \left(\sum_{i=1}^n y_i^2 - n \bar{y}^2\right)}} =$$

$$= \frac{-101 - 5 \cdot 1 \cdot (-4)}{\sqrt{(55 - 5 \cdot 1^2)(226 - 5 \cdot (-4)^2)}} = -\frac{81}{\sqrt{50 \cdot 146}} =$$

si ha quindi:

$$r = -\frac{81}{84,58} = -0,948$$



Al crescere degli anni il prezzo diminuisce. Il coefficiente di correlazione è **$r = -0.948$** molto vicino a -1 quindi **forte correlazione negativa**. I dati sono quasi disposti su una retta (si ricordi che se $r = -1$ allora $y_i = bx_i + a$ con $b < 0$).