

Statistica descrittiva bivariata


STATISTICA NUMERICA

A.Y. 2024-25

<https://www.andreaminini.org/statistica/covarianza>

https://tommasorigon.github.io/Stat/slides/sl_J.pdf

<https://www.webtutordimatematica.it/materie/statistica-e-probabilita/variabili-aleatorie-e-distribuzioni-di-probabilita/covarianza>



Covarianza fra due variabili

La covarianza indica la tendenza che hanno due variabili (X e Y) a *variare insieme*, ovvero, a *covariare*.

Ad esempio, si può supporre che vi sia una relazione tra l'insoddisfazione della madre e l'aggressività del bambino, nel senso che all'aumentare dell'una aumenta anche l'altra.

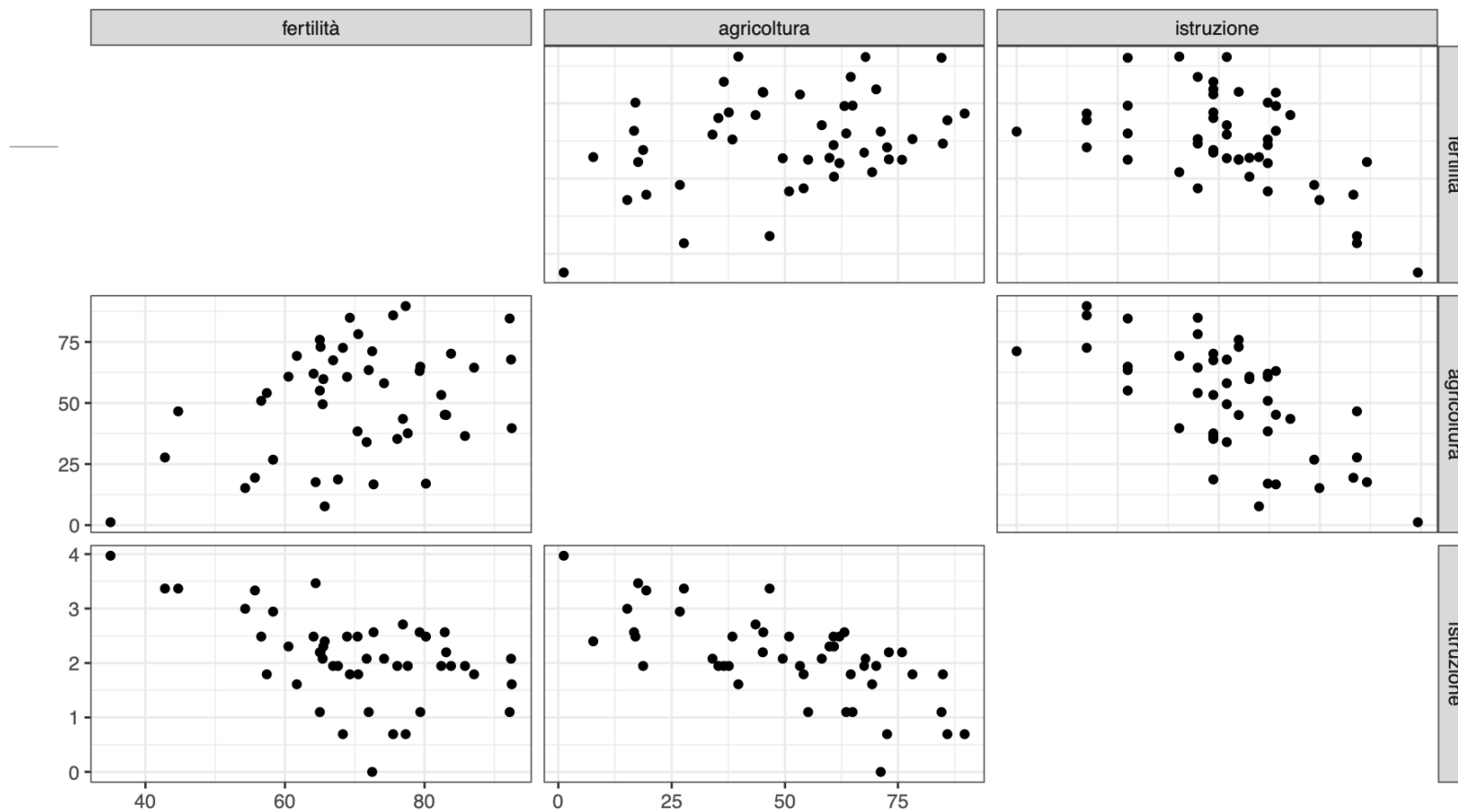
Quando si parla di correlazione bisogna prendere in considerazione due aspetti:
il tipo di relazione esistente tra due variabili e la forma della relazione.

Covarianza fra due variabili: esempio

Consideriamo **tre indicatori socio-economici** disponibili per $n = 47$ province svizzere di lingua francese. I dati sono storici e si riferiscono al 1888. Consideriamo:

- Una **misura di fertilità** (nati per donna), standardizzata in maniera tale che vari tra 0 e 100.
- Percentuale degli **occupati in agricoltura** sul totale degli occupati, interpretabile come un indicatore di urbanizzazione della provincia.
- Il logaritmo della percentuale della popolazione con un'**istruzione** superiore alla scuola primaria.

Il problema che ci poniamo è di cercare di descrivere le **relazioni** esistenti tra i tre indicatori.



Covarianza fra due variabili: esempio

- La percentuale di occupati in agricoltura e fertilità sono **positivamente associati**.
- Province con una alta percentuale di occupati in agricoltura hanno anche una alta fertilità.
Viceversa, basse percentuali di occupati in agricoltura si osservano in province con bassi livelli di fertilità.
- Esiste una **associazione negativa** tra istruzione e fertilità.
Province con un alto livello di istruzione hanno una fertilità più bassa delle province con un basso livello di istruzione.
- Simili considerazioni possono essere fatte per la relazione tra le variabili agricoltura e istruzione, in cui si osserva una **associazione negativa**.

Covarianza fra due variabili: esempio

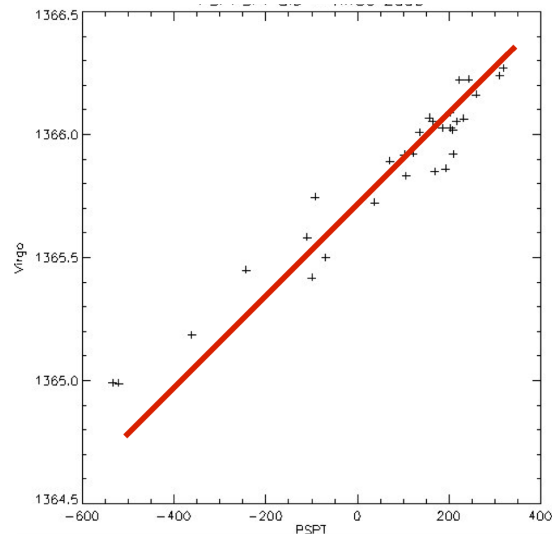
- La relazione tra agricoltura e fertilità sembra **più debole** della relazione esistente tra agricoltura ed istruzione.
- Meno facile è valutare l'intensità delle relazioni intercorrenti tra istruzione e, rispettivamente, agricoltura e fertilità.
- La prima relazione (istruzione — agricoltura) sembra però in una qualche misura **più forte** della seconda (istruzione — fertilità).

Per quantificare queste relazioni, abbiamo pertanto bisogno di un **indice** che sia in grado di identificare forza e direzioni delle associazioni tra variabili.

Relazione lineare fra due variabili

Per quanto riguarda il **tipo di relazione**, essa può essere *lineare* o *non lineare*.

La relazione è di tipo *lineare* se, rappresentata su assi cartesiane, si avvicina alla forma di una retta.

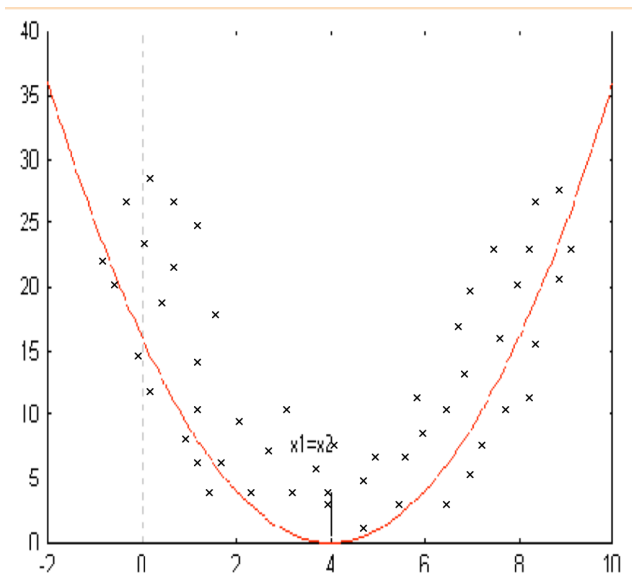


In questo caso, all'aumentare (o al diminuire) di X aumenta (diminuisce) Y.

Ad esempio, all'aumentare dell'altezza di una persona aumenta anche il suo peso.

Relazione non lineare fra due variabili

La relazione è di tipo *non lineare*, se rappresentata su assi cartesiane, ha un andamento curvilineo (parabola o iperbole).



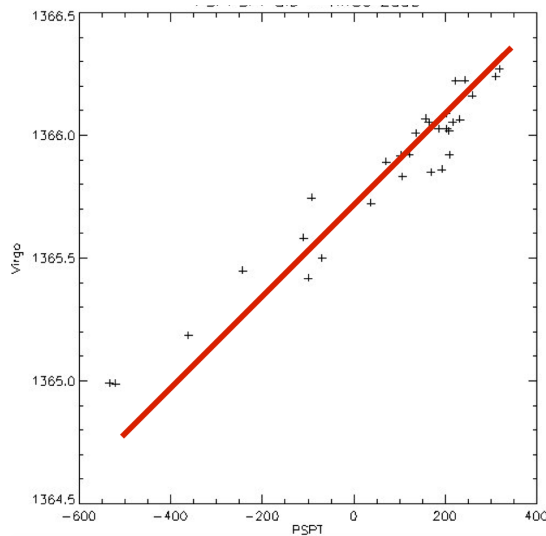
In questo caso a livelli bassi e alti di X corrispondono livelli bassi di Y; mentre a livelli intermedi di X corrispondono livelli alti di Y.

Ad esempio, il tempo impiegato per risolvere un problema è alto quando l'ansia è bassa o alta, è elevato quando l'ansia ha livelli medi.

Forma della relazione fra due variabili

Per quanto riguarda la **forma della relazione**, si distinguono l'*entità* e la *direzione*.

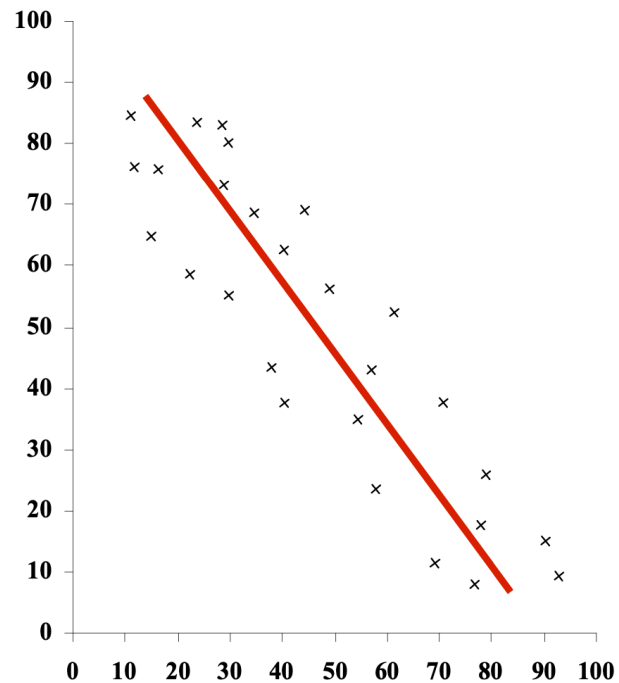
La **direzione** può essere: *positiva*, se all'aumentare di una variabile aumenta anche l'altra.



La **direzione** può essere: *positiva*, se all'aumentare di una variabile aumenta anche l'altra.

Ad esempio, all'aumentare dell'altezza di una persona aumenta anche il suo peso.

Forma della relazione fra due variabili

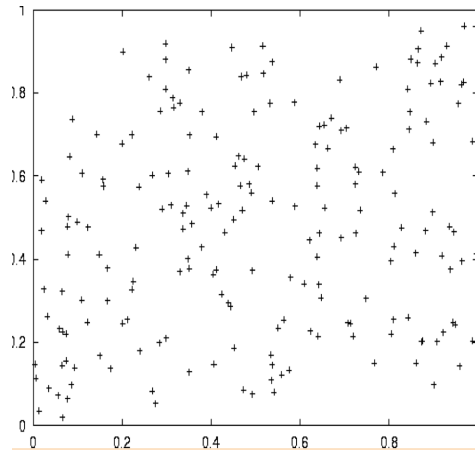


La direzione è *negativa* se all'aumentare di una variabile diminuisce l'altra.

Forma della relazione fra due variabili

L'**entità** si riferisce alla forza della relazione esistente tra due variabili.

Quanto più i punteggi sono raggruppati attorno ad una retta, tanto *più forte* è la relazione tra due variabili.



Se i valori sono dispersi in maniera uniforme, invece, tra le due variabili *non esiste* alcuna relazione.

Covarianza fra due variabili: definizione

Un indicatore che misura la forza della relazione tra due variabili è la **covarianza**.
Si noti che la covarianza è simmetrica, ovvero: $\text{cov}(x,y) = \text{cov}(y,x)$.

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

La covarianza pertanto assume valori **positivi** se la maggior parte dei termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono concordi, ovvero se hanno lo stesso segno.

La covarianza assume invece valori **negativi** se la maggior parte dei termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono discordi, ovvero se hanno segni diversi.

Infine, la covarianza assume valori **prossimi a zero** se i termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono in ugual misura concordi e discordi.

Covarianza fra due variabili: proprietà

- La covarianza tra la variabile x e x stessa è pari alla varianza di x , ovvero

$$\text{cov}(x, x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \text{var}(x) \geq 0.$$

Poiché i termini $(x_i - \bar{x})$ e $(x_i - \bar{x})$ sono necessariamente sempre concordi, in questo caso la covarianza è grande e positiva.

- La covarianza tra la variabile x e $-x$ stessa è pari alla varianza di x cambiata di segno, ovvero

$$\text{cov}(x, -x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(-x_i + \bar{x}) = -\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = -\text{var}(x) \leq 0.$$

Poiché i termini $(x_i - \bar{x})$ e $-(x_i - \bar{x})$ sono necessariamente sempre discordi, in questo caso la covarianza è grande e negativa.

Matrici di covarianza

Nel caso in esame, troviamo che:

$\text{cov}(\text{fertilità}, \text{agricoltura}) = 98.0$,
 $\text{cov}(\text{fertilità}, \text{istruzione}) = -5.1$,
 $\text{cov}(\text{agricoltura}, \text{istruzione}) = -11.9$.

Tipicamente, le varianze e le covarianze di tutte le coppie di variabili vengono organizzate in una matrice, chiamata **matrice delle varianze e covarianze**.

	fertilità	agricoltura	istruzione
fertilità	152.7	98.0	-5.1
agricoltura	98.0	504.8	-11.9
istruzione	-5.1	-11.9	0.6

Matrici di covarianza

- In tale matrice, l'elemento in posizione (i,j) rappresenta la covarianza tra la variabile i -esima e la variabile j -esima.
- Nella diagonale ci sono le varianze, poiché $\text{cov}(x,x) = \text{var}(x)$.
- Inoltre, poiché $\text{cov}(x,y) = \text{cov}(y,x)$, la matrice è **simmetrica**.

Correlazione fra due variabili

Per affermare se la covarianza è piccola o grande dobbiamo confrontarla con il prodotto degli scarti quadratici medi.

Di conseguenza, solitamente la covarianza viene presentata direttamente nella sua forma **normalizzata**, chiamata correlazione.

Tale coefficiente è standardizzato e può assumere valori che vanno da **-1.00** (correlazione perfetta negativa) e **+1.00** (correlazione perfetta positiva). Una correlazione uguale a **0** indica che tra le due variabili non vi è alcuna relazione.

Nota. La correlazione non include il concetto di causa-effetto, ma solo quello di rapporto tra variabili. La correlazione ci permette di affermare che tra due variabili c'è una *relazione sistematica*, ma non che una causa l'altra.

Il coefficiente di correlazione r di Pearson

Tale coefficiente serve a misurare la correlazione tra variabili a intervalli o a rapporti equivalenti. Dette X e Y le due variabili e \bar{X} , \bar{Y} le loro medie, il coefficiente di correlazione Pearson si definisce come (denominato r oppure *cor*).

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Tale coefficiente può assumere valori che vanno da -1.00 (tra le due variabili vi è una correlazione perfetta negativa) e $+1.00$ (tra le due variabili vi è una correlazione perfetta positiva). Una correlazione uguale a 0 indica che tra le due variabili non vi è alcuna relazione.

Il coefficiente di correlazione r di Pearson

Nel caso precedente, troviamo che:

$\text{cor}(\text{fertilità}, \text{agricoltura}) = 0.35$,

$\text{cor}(\text{fertilità}, \text{istruzione}) = -0.52$,

$\text{cor}(\text{agricoltura}, \text{istruzione}) = -0.68$.

Anche le correlazioni vengono tipicamente organizzate in una matrice, chiamata **matrice di correlazione**.

	fertilità	agricoltura	istruzione
fertilità	1	0.35	-0.52
agricoltura	0.35	1	-0.68
istruzione	-0.52	-0.68	1

Il coefficiente di correlazione r di Pearson

- La covarianza e la correlazione misurano esclusivamente **relazioni lineari**. Questo ha importanti conseguenze.
- Se la relazione tra x ed y è monotona ma non lineare, allora $\text{cor}(x,y) < 1$.

Esempio. Si considerino i dati x_1, \dots, x_5 pari a $-2, -1, \dots, 2$ e si consideri

$$y_i = e^{x_i}, \quad i = 1, \dots, 5.$$

Nonostante la relazione tra le variabili x ed y sia monotona, $\text{cor}(x,y) = 0.89 < 1$.

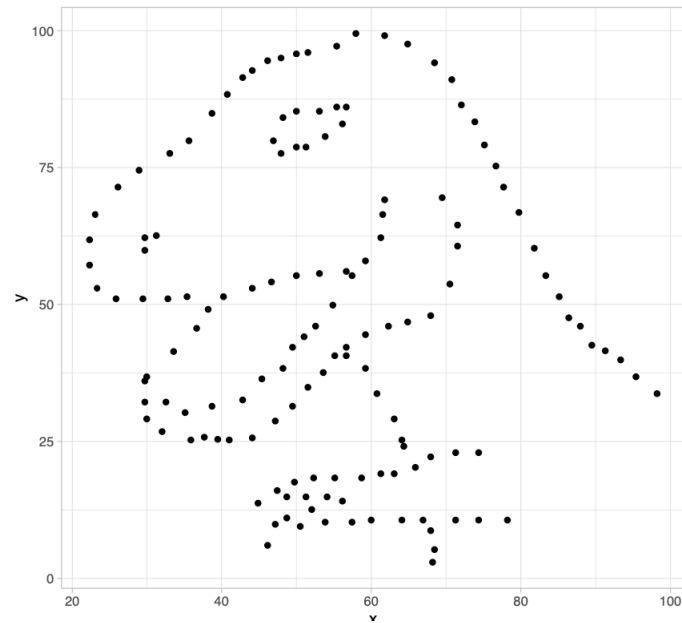
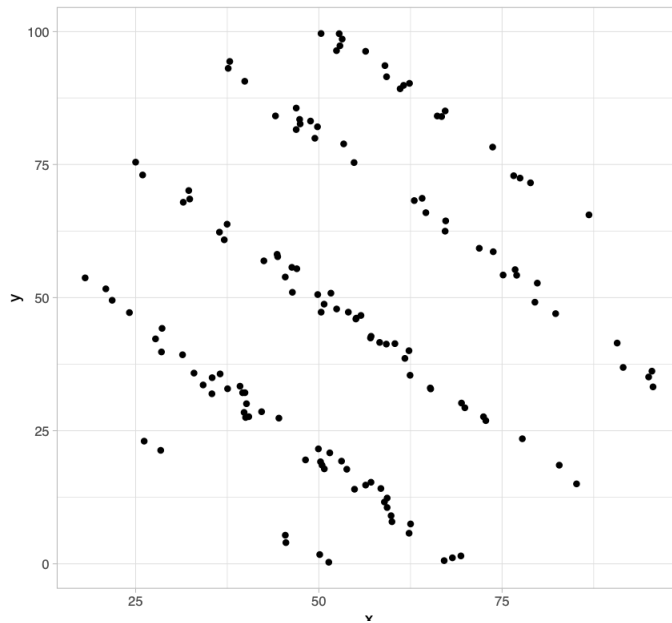
- Il fatto che $\text{cor}(x,y) = 0$ non permette di escludere la presenza di relazioni non-monotone nei dati.

Esempio. Si considerino i dati x_1, \dots, x_5 pari a $-2, -1, \dots, 2$ e si consideri

$$y_i = x_i^2, \quad i = 1, \dots, 5.$$

Nonostante ci sia una relazione ben precisa tra le variabili x ed y , $\text{cor}(x,y) = 0$.

Il coefficiente di correlazione r di Pearson



Questi insiemi di dati hanno **correlazione nulla**