Statistica della Formazione

Slides 6

A.A. 2020-2021

Docente: ANNA LINA SARRA

Modulo 1: elementi di statistica descrittiva

- **Analisi delle distribuzioni doppie: analisi della correlazione**

Dai dati univariati ai dati bivariati

- In molte situazioni interessa studiare se esiste una relazione tra due variabili misurate sulle stesse unità.

Esempio:

– "*il voto di maturità è in relazione con la performance universitaria?*"

- Oppure si desidera prevedere il valore di una variabile conoscendo il valore di un'altra.

Esempio:

– "*conoscendo l'età del paziente, è possibile prevedere la sua pressione arteriosa?*"

- La statistica permette di rispondere a questo tipo di domande, con strumenti adatti alla natura delle variabili in esame. A tale scopo, per variabili quantitative, si tratteranno:
  - La CORRELAZIONE, che misura la dipendenza lineare tra due variabili;
  - La REGRESSIONE, che valuta la relazione lineare tra due variabili.

Correlazione

La correlazione misura l'associazione tra due variabili quantitative.

È lo strumento che si utilizza quando si hanno a disposizione coppie di valori di variabili. Permette di valutare come variano i valori di una variabile al variare dell'altra e viceversa.

- Esempi:
  - Numero di sigarette fumate in gravidanza e tasso di crescita del feto ⇒ all'aumentare del numero di sigarette fumate diminuisce il tasso di crescita (correlazione negativa).
  - Livello di colesterolo e BMI (Body Mass Index = peso (kg)/altezza^2 (m^2)) ⇒ tanto è maggiore il BMI quanto è maggiore il livello di colesterolo (correlazione positiva).
  - Il valor medio della temperatura (ambiente) e il BMI ⇒ non c'è motivo di pensare che la temperatura influenzi il BMI delle persone (assenza di correlazione).
- La relazione può essere valutata tramite:
  - Un grafico (grafico di dispersione)
  - Un indice che quantifica il grado di correlazione (coefficiente di correlazione)

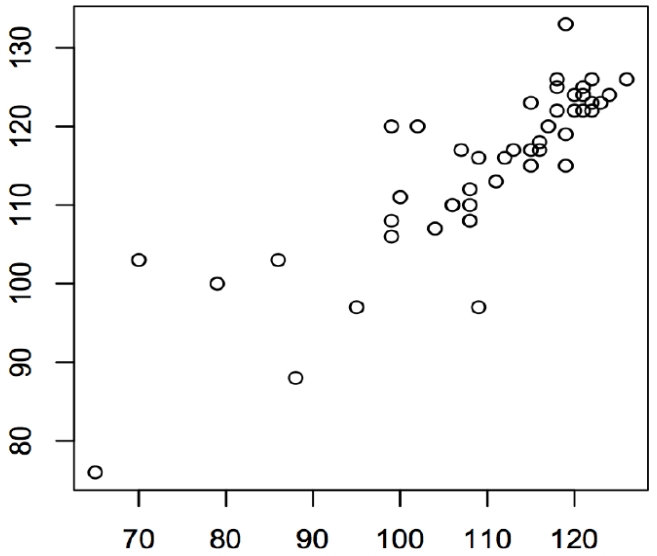
Diagramma a dispersione

Nello studio dell'associazione tra due variabili quantitative misurate sulle stesse unità statistiche, indicate con X e Y , è molto utile disegnare un grafico, il diagramma di dispersione, prima di procedere con altre analisi formali.

Nel grafico di dispersione le coppie

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

di valori di due variabili quantitative misurate sulle n unità sono rappresentati come punti di un piano cartesiano, i cui assi corrispondono alle due variabili.

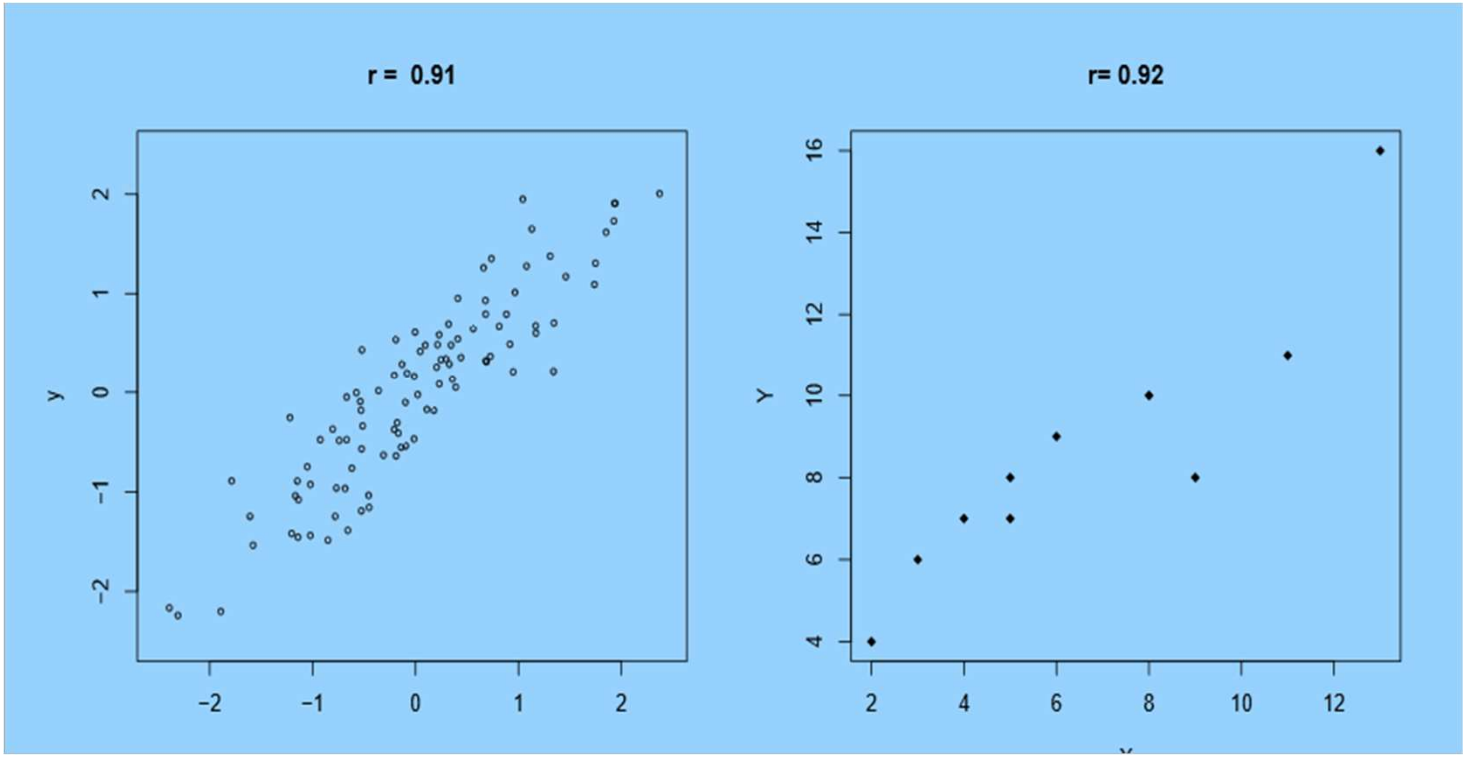


Correlazione

Data una **distribuzione doppia in forma disaggregata**, si dice che tra le due variabili *X* e *Y*

- vi è **correlazione positiva** o concordanza quando esse tendono a crescere (decrescere) insieme
- vi è **correlazione negativa** o discordanza quando al crescere di una variabile l'altra tende a decrescere.

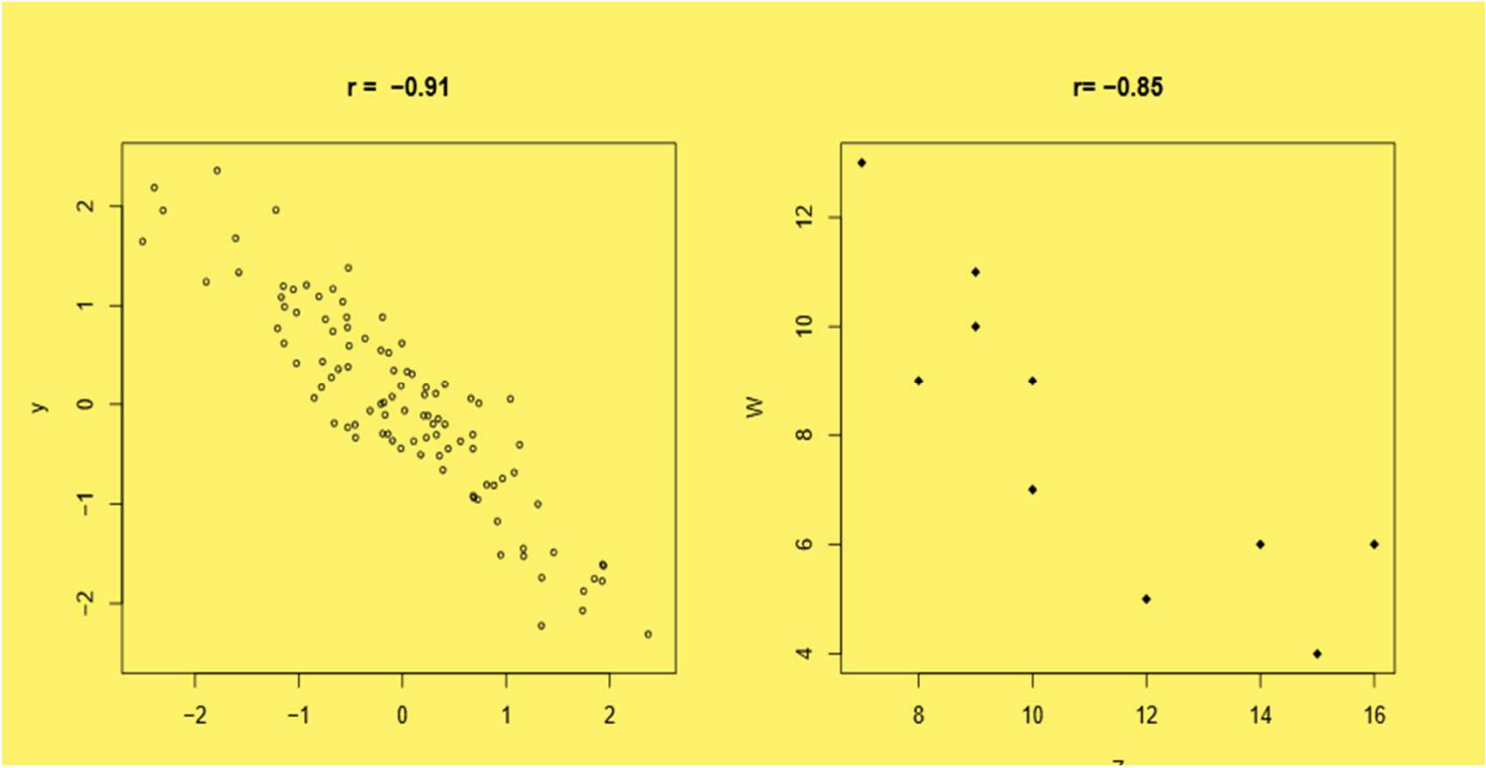
Esempio di correlazione positiva



All'aumentare di X aumenta anche Y, ciascuna variabile a modo suo e viceversa.

È una relazione lineare proporzionale.

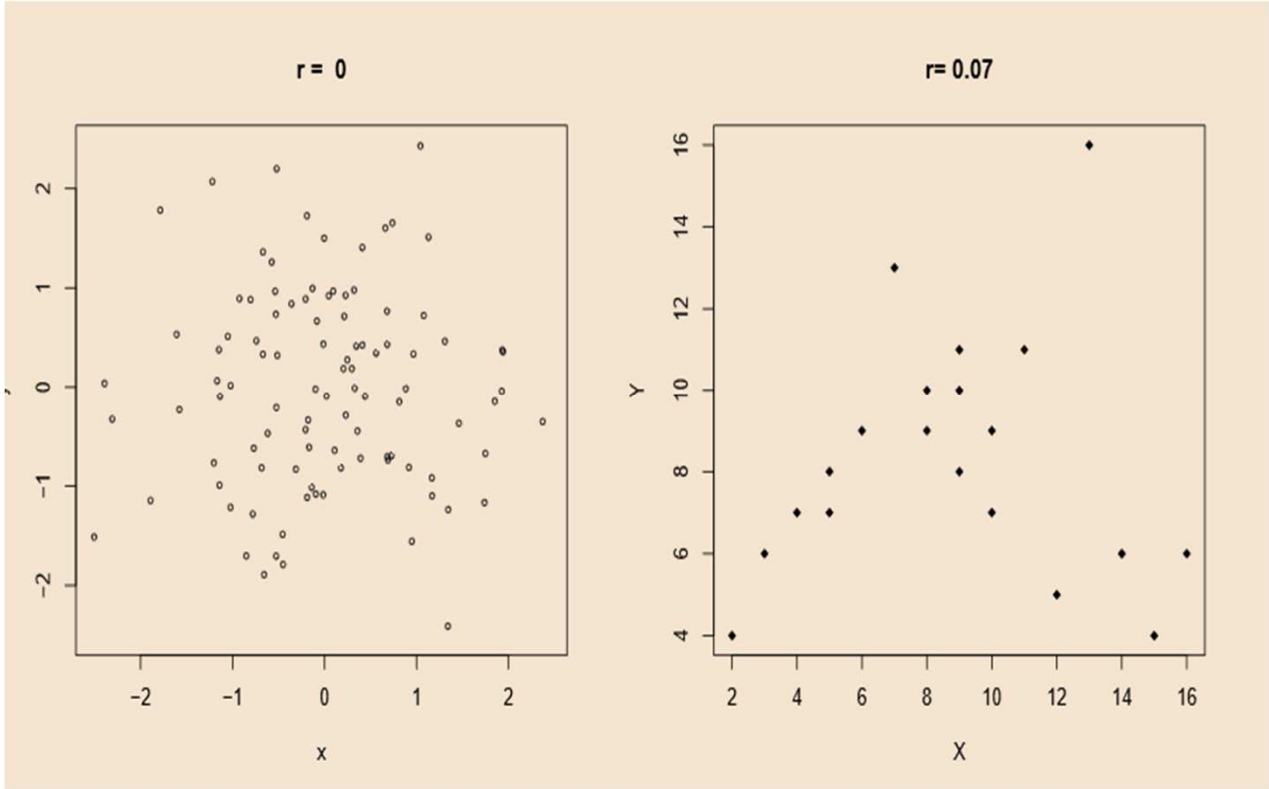
Esempio di correlazione negativa



All'aumentare di X diminuisce Y, ciascuna variabile a modo suo. E viceversa.

È una relazione lineare proporzionale.

Esempio di correlazione nulla



Non c'è alcun legame lineare fra X e Y.

Ciascuna varia indipendentemente dall'altra

COVARIANZA

Per avere una valutazione analitica del grado di associazione tra due variabili quantitative, esiste un indice che misura la dispersione nel piano dei punti dal proprio centro : la COVARIANZA

La covarianza è un indice che esprime la quantità di varianza che due variabili hanno in comune.

La formula deriva da quella della varianza.

In formula

La covarianza è

$$cov(X,Y) = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{N}$$

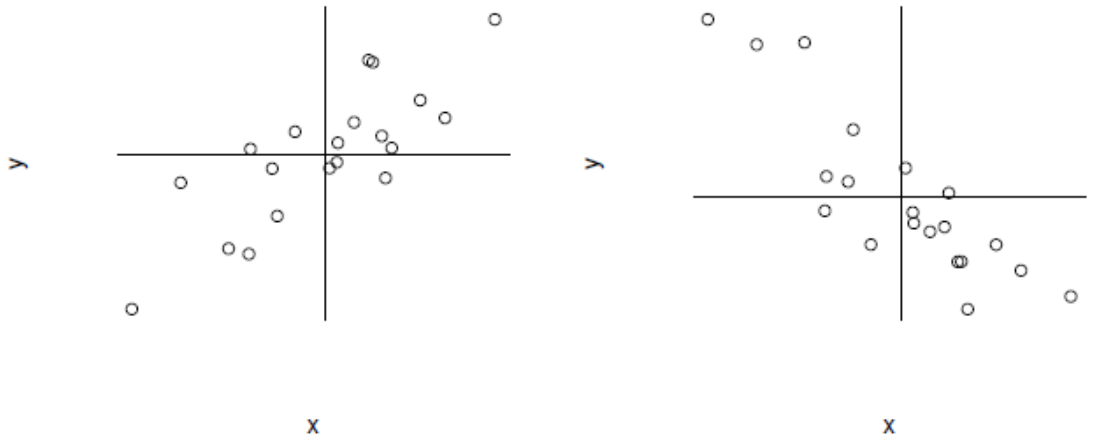
La varianza è

$$var(X) = \frac{\sum(X-\overline{X})^2}{N} = \frac{\sum(X-\overline{X})(X-\overline{X})}{N}$$

Notare la somiglianza tra le due formule

La covarianza

La covarianza, a differenza della varianza che è sempre positiva, misura l'eventuale direzione del legame, ovvero se le due variabili si muovono nella stessa direzione o in direzioni opposte. Il segno della covarianza riflette il senso crescente o decrescente dell'allineamento tendenziale.

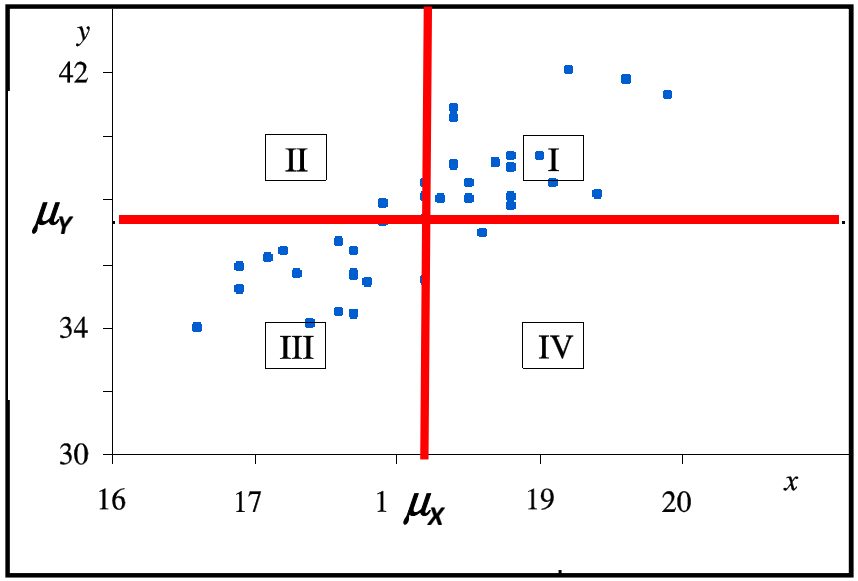


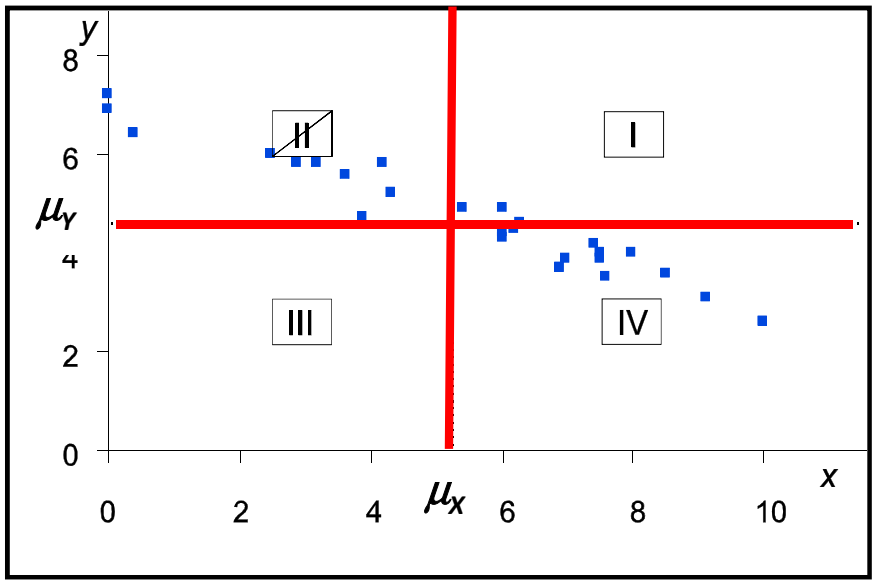
Interpretazione geometrica

$$cov(X,Y) = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{N}$$

Concordanza: i punti osservati sono collocati in prevalenza nel I e nel III quadrante dei nuovi assi cartesiani aventi origine nel punto (μ_X, μ_Y).

Discordanza: i punti osservati sono collocati in prevalenza nel secondo e nel quarto quadrante dei nuovi assi cartesiani aventi origine nel punto (μ_X, μ_Y).





Osservazioni

- Se X e Y sono concordi, allora la covarianza assume segno positivo;
- Se X e Y sono discordi, allora la covarianza assume segno negativo;
- Se la covarianza è nulla, X e Y sono indifferenti (incorrelati).

COVARIANZA **misura assoluta di concordanza** tra due caratteri x e y ed è espressa in un'unità di misura pari al prodotto delle unità di misura dei due caratteri.

Campo di variazione della covarianza

La covarianza può assumere sia valori positivi che negativi.
In particolare vale che

$$-\sigma_x \sigma_y \leq COV(XY) \leq \sigma_x \sigma_y$$

Per ricercare un indice relativo che permetta di effettuare confronti significativi occorrerà **dividere la covarianza per il prodotto degli scarti quadratici medi di X e Y**
L'indice così ottenuto prende valori in [-1,1] e viene detto **coefficiente di correlazione.**

Il coefficiente di correlazione lineare di Bravais-Pearson

In una **distribuzione doppia disaggregata**, il coefficiente di correlazione lineare di Bravais-Pearson è

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}.\frac{y_i - \mu_y}{\sigma_y}\right).$$

Scarti standardizzati della X e della Y

$$z_{x_i} = \frac{x_i - \mu_X}{\sigma_X}; \quad z_{y_i} = \frac{y_i - \mu_y}{\sigma_y}$$

Si può ottenere anche come

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}} = \frac{C_{xy}}{\sqrt{D_x D_y}}$$

dove il segno è determinato dalla codevianza

$$C_{xy} = \sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)$$

Interpretazione geometrica della formula

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right).$$

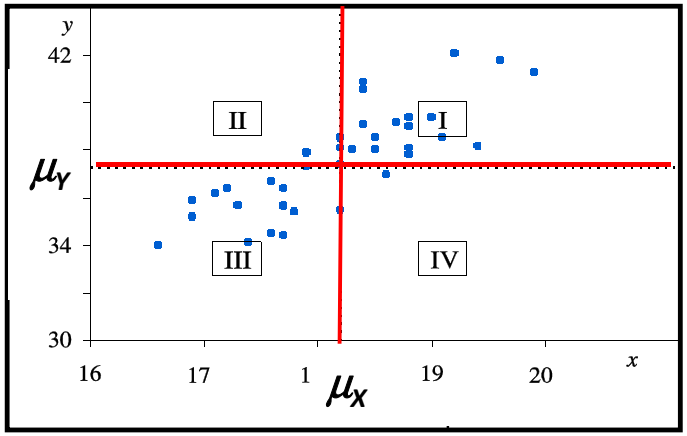
Ne segue che i prodotti

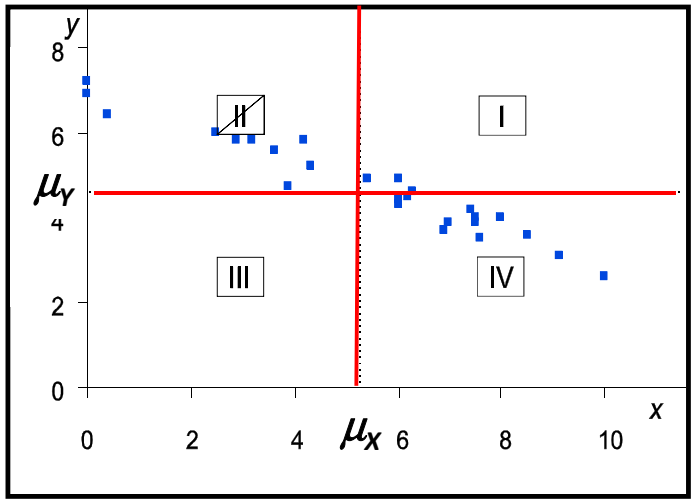
$$\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}$$

sono in prevalenza positivi nel caso di concordanza e prevalentemente negativi nel caso di discordanza. Cosicché la quantità

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right),$$

media di tali prodotti, è positiva in caso di concordanza e negativa nel caso di discordanza.





Coefficiente di correlazione

È un indice statistico che misura l'associazione (relazione) fra due variabili.

Misura come le due variabili si muovono assieme, ossia come co-relano.

Viene espresso come un valore che oscilla fra -1 e 1.

Coefficiente di correlazione

A. Riassunto numerico della forza della relazione fra due variabili

B. Permette di sostituire un diagramma a dispersione con un semplice indice.

È costituito da due parti:

Un segno che indica la direzione della relazione
Un numero fra 0 e 1 che indica la forza della relazione

1 indica una relazione perfetta, esprimibile tramite una formula matematica precisa

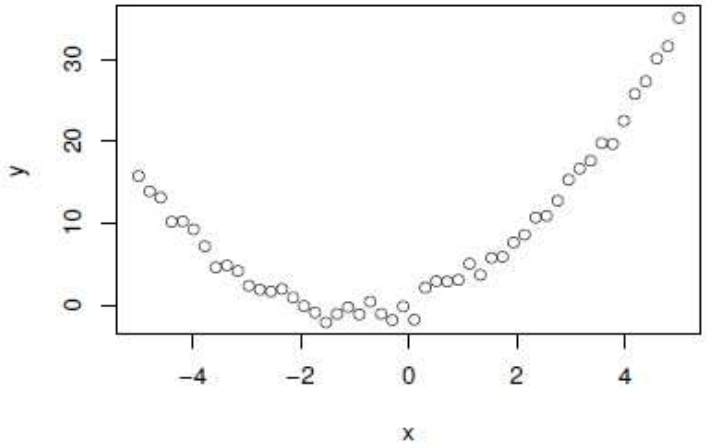
0 indica la mancanza di qualunque relazione fra le due variabile.

Guida all'interpretazione di r

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = +1$: correlazione positiva perfetta (tutti i punti su una retta: concordi)
- $r_{xy} = -1$: correlazione negativa perfetta (tutti i punti su una retta: discordi)
- $r_{xy} > 0$: correlazione positiva
- $r_{xy} < 0$: correlazione negativa
- $r_{xy} \cong 0$: assenza di relazione lineare

Quando tra X e Y non vi è una relazione lineare o essa è estremamente debole, il valore dell'indice r_{xy} è zero o circa zero, e le variabili sono dette incorrelate.

ATTENZIONE: Il coefficiente di correlazione misura una associazione lineare. Il valore $r_{xy} = 0$ non indica tuttavia un'assenza di relazione tra le due variabili. Può esserci una relazione curvilinea.



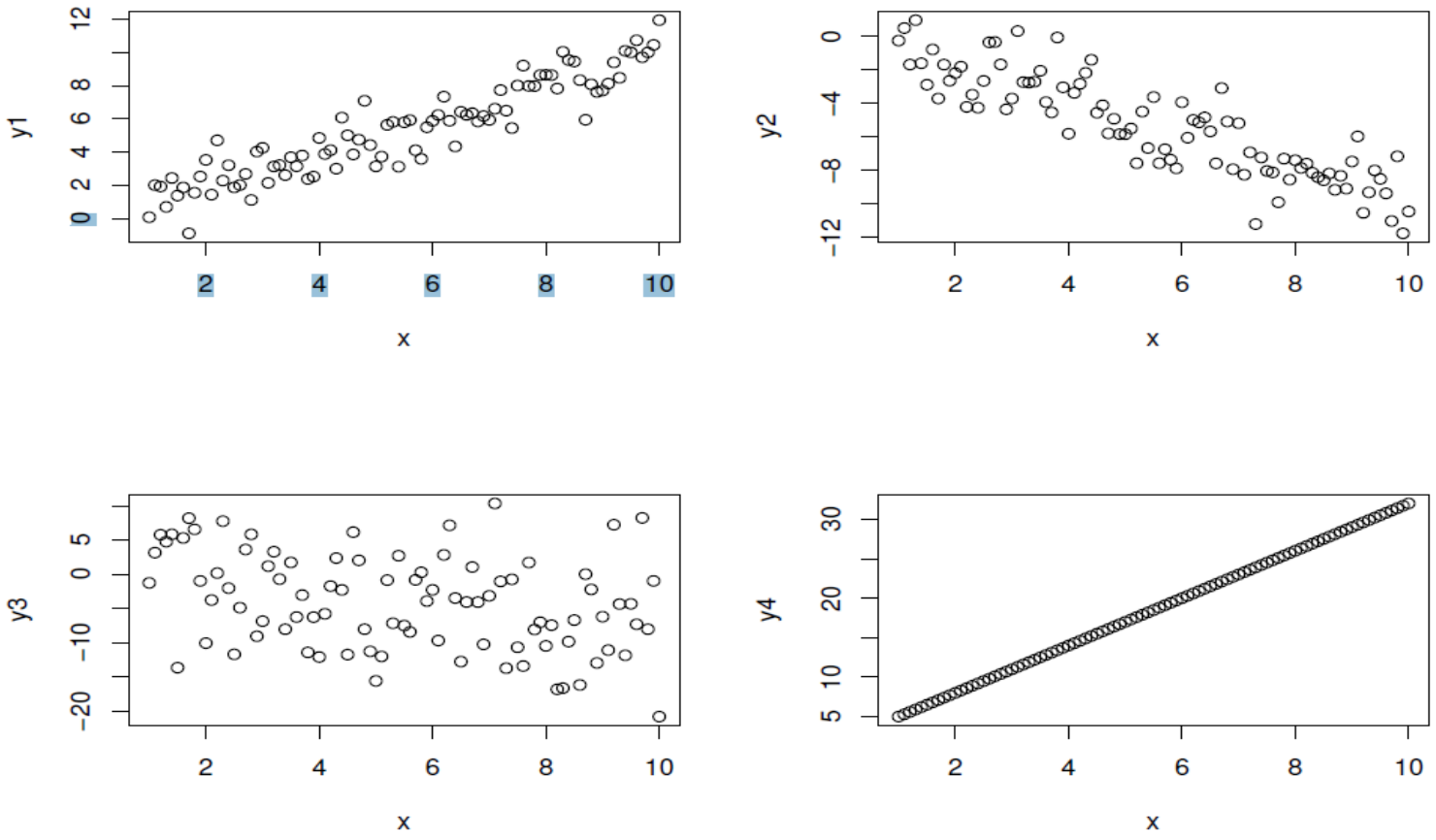
Guida all'interpretazione di r

L'interpretazione si applica al valore della correlazione indipendentemente dal segno.

Valore di r	Correlazione	Relazione
0.00-0.20	Piccola	Molto poco intensa, quasi inesistente
0.20-0.40	Bassa	Piccola, appena appena apprezzabile
0.40-0.60	Regolare	Considerevole
0.60-0.80	Alta	Intensa
0.80-1.00	Molto alta	Molto intensa

Il segno indica solo la relazione proporzionale (positiva) o inversamente proporzionale (negativa).

Quale correlazione……



Coefficiente di correlazione lineare di Bravais-Pearson: Esempio

La Tabella mostra i punteggi ottenuti da un gruppo di 10 studenti agli esami di un college (X) e ad un test di comprensione verbale (Y).

Studente	Esami di ammissione *X*	Test di comprensione verbale *Y*
A	52	49
B	28	34
C	70	45
D	51	49
E	49	40
F	65	50
G	49	37
H	49	49
I	63	52
J	32	32

Coefficiente di correlazione lineare di Bravais-Pearson: calcolo

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

- indice di correlazione

Esame di ammissione X	Test di comprensione verbale Y	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - y_x)^2$	$(x_i - \mu_x) * (y_i - \mu_Y)$
52	49	1.2	5.3	1.44	28.09	6.36
28	34	-22.8	-9.7	519.84	94.09	221.16
70	45	19.2	1.3	368.64	1.69	24.96
51	49	0.2	5.3	0.04	28.09	1.06
49	40	-1.8	-3.7	3.24	13.69	6.66
65	50	14.2	6.3	201.64	39.69	89.46
49	37	-1.8	-6.7	3.24	44.89	12.06
49	49	-1.8	5.3	3.24	28.09	-9.54
63	52	12.2	8.3	148.84	68.89	101.26
32	32	-18.8	-11.7	353.44	136.89	219.96
				1603.6	**484.1**	**673.4**

Le medie sono:

$\mu_X = 50.8$ $\mu_Y = 43.7$

$$r = \frac{673.4}{\sqrt{1603.6 * 484.1}} = 0.76$$

Il caso delle distribuzioni doppie di frequenze

Nel **caso delle distribuzioni di frequenze** abbiamo

$$r = \frac{\sum_{i=1}^{s}\sum_{j=1}^{t}(x_i - \mu_X)(y_j - \mu_Y)n_{ij}}{\sqrt{\sum_{i=1}^{s}(x_i - \mu_X)^2 n_{i0}\sum_{j=1}^{t}(y_j - \mu_Y)^2 n_{0j}}}.$$

N.B:

Quando uno o entrambi i caratteri sono divisi in intervalli, l'indice *r* si calcola prendendo i valori centrali di classe.

Il caso delle distribuzioni doppie di frequenze: calcolo di r

Distribuzione doppia di frequenze di un campione di coniugi classificati secondo l'età:

I numeri in rosso sono i valori centrali

Età marito	Età della moglie				Totale
	18-30 24	31-40 35.5	41-50 45.5	51-65 58	
20-30 25.0	14	0	0	0	14
31-40 35.5	5	23	0	0	28
41-50 45.5	0	5	17	1	23
51-65 58.0	0	0	9	26	35
Totale	19	28	26	27	100

μ_X	44.21
μ_Y	41.99
D_X	13984.55
D_Y	14569.49
C_{XY}	13153.46
r	0.92

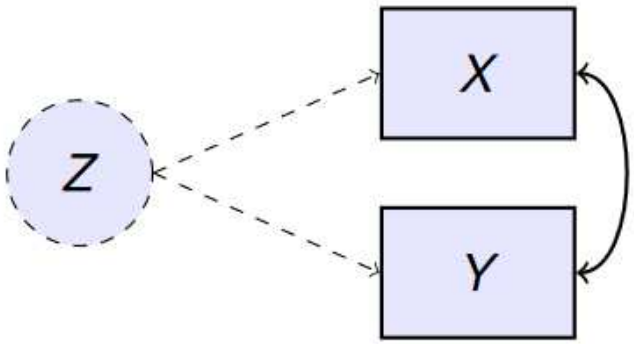
- L'indice di correlazione è $r = \frac{13153.46}{\sqrt{13984.55 \cdot 14569.49}} = 0.92$

Legame tra le variabili

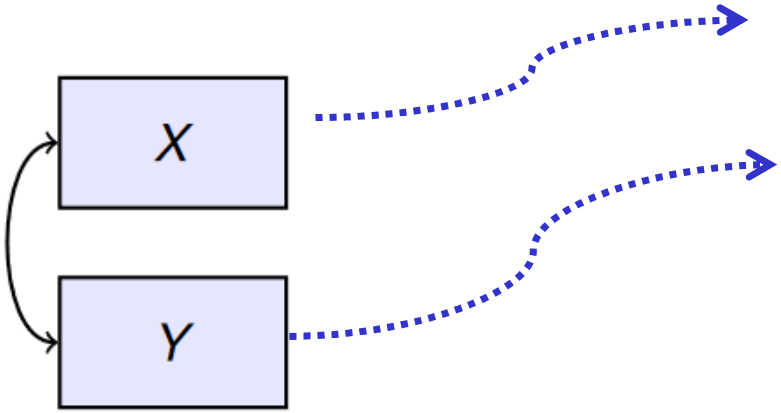
- È importante ricordare che se esiste una correlazione fra due variabili (che calcoliamo con r), questo indice non ci dà nessuna informazione sui legami di causa-effetto.

- Le due variabili "si muovono assieme". STOP!

- È possibile che esista una terza variabile che ha influenza su entrambe e che la correlazione che abbiamo calcolato sia dovuta a questa influenza.

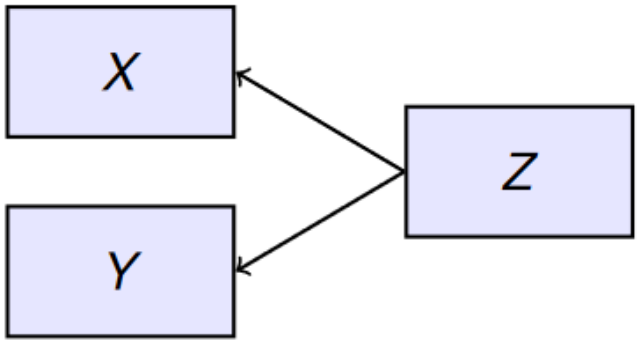


Correlazioni spurie



- X è il numero di vigili del fuoco mandato a spegnere un incendio
- Y è l'entità del danno prodotto dall'incendio

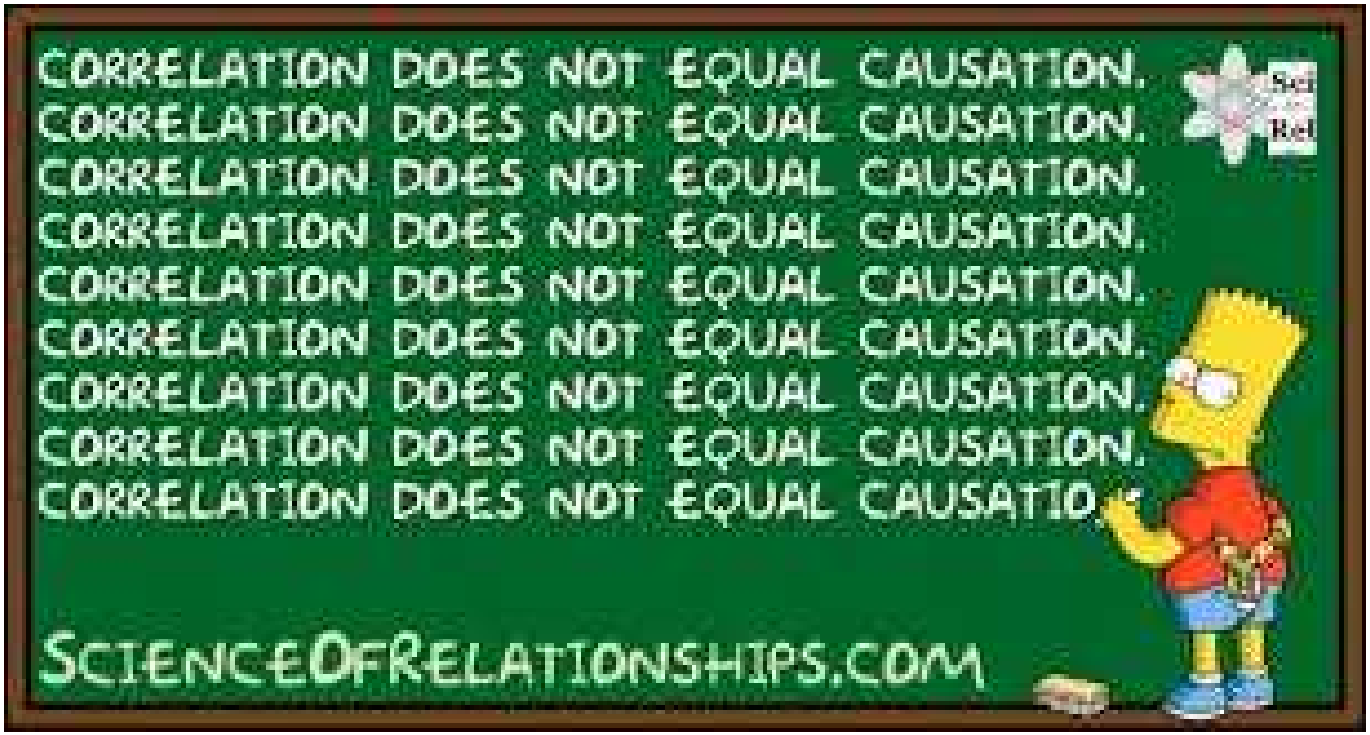
La loro correlazione vuol dire che più vigili del fuoco producono più danni?



Nel momento in cui si identifica una variabile antecedente ad entrambe, la correlazione spuria acquista senso.

Z è l'ampiezza dell'incendio

Più ampio l'incendio, più vigili del fuoco vengono inviati a spegnerlo più ampio l'incendio, più danni prodotti


CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATIO
SCIENCEOFRELATIONSHIPS.COM