

L'analisi bivariata

L'analisi bivariata permette di analizzare e modellizzare le relazioni tra coppie di variabili

Lo strumento fondamentale dell'analisi bivariata è la matrice di varianza-covarianza e/o correlazione

CORRELAZIONE FRA VARIABILI

ANALISI DI TIPO COMPARATIVO per stabilire se esiste una connessione tra due caratteri rilevati.

Lo studio di correlazione si effettua utilizzando un **diagramma di dispersione** (scatter plot)

Dall'analisi si può concludere che:

- **Esiste una correlazione diretta**
- **Esiste una correlazione inversa**
- **Non esiste alcuna correlazione**

La rilevazione contemporanea di due variabili X e Y ha essenzialmente lo scopo di evidenziare le eventuali relazioni esistenti fra loro, ossia di verificare se esiste una certa **dipendenza** di una variabile dall'altra. In una condizione di dipendenza di Y da X , al variare delle determinazioni assunte da X si modifica una qualche caratteristica della distribuzione di Y .

Se **una delle due variabili è di tipo qualitativo**, si ha una situazione di dipendenza, che in questo caso è detta anche **connessione**, quando a una particolare determinazione di una di esse si accompagna, in genere, una specifica determinazione dell'altra.

Se invece **entrambe le variabili sono quantitative**, si è in presenza di un certo grado di dipendenza, che in questo caso è detta **correlazione**, quando al crescere dei valori assunti dalla X anche la Y tende a crescere, oppure quando al crescere dei valori assunti dalla X la Y tende a diminuire.

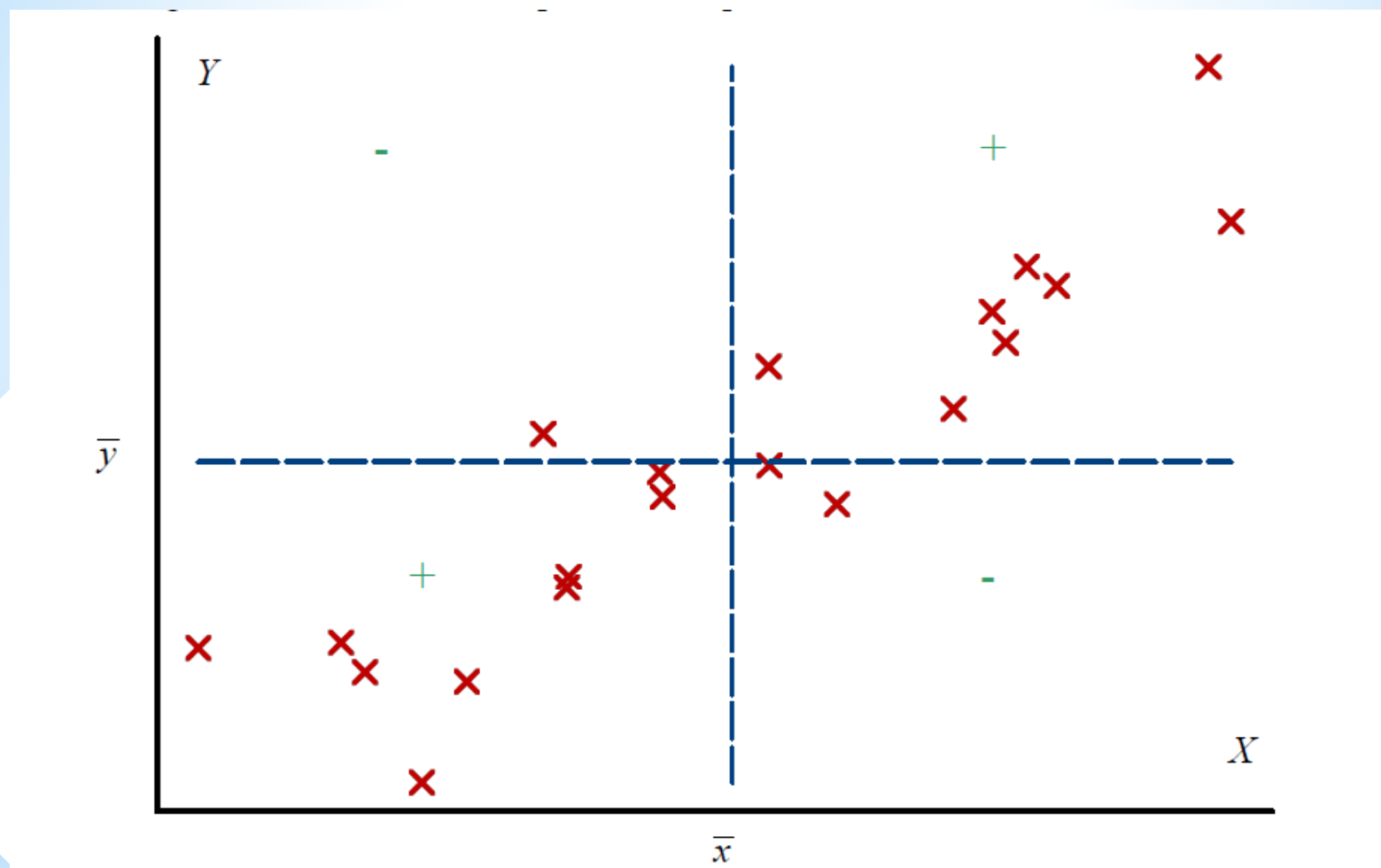
	Attività febbraio 2015		Attività maggio 2015		Attività dicembre 2015	
	CSE020215	CSE020215B	CSE140515lap	CSE150515	CSE071215B	CSE081215A
	lava	lava	lapilli	lava	lava	lava
SiO ₂	46.95	46.99	47.19	47.15	47.20	47.12
TiO ₂	1.90	1.91	1.92	1.90	1.88	1.91
Al ₂ O ₃	17.00	16.97	17.07	17.06	17.01	16.83
Fe ₂ O ₃ _{tot}	11.43	11.48	11.48	11.48	11.77	11.79
MnO	0.18	0.18	0.18	0.18	0.19	0.19
MgO	5.49	5.37	5.42	5.42	5.44	5.49
CaO	10.83	10.83	10.71	10.81	10.51	10.83
Na ₂ O	3.36	3.41	3.17	3.15	3.12	2.98
K ₂ O	1.89	1.92	1.92	1.91	1.94	1.92
P ₂ O ₅	0.58	0.55	0.55	0.54	0.54	0.53
L.O.I.	0.80**	0.86**	0.86*	0.86*	0.80**	0.80**

Analisi degli elementi maggiori

In numerose situazioni reali, quando entrambe le variabili X e Y sono di tipo quantitativo, si vuole valutare se al crescere dei valori assunti da una variabile anche i valori dell'altra tendono a crescere oppure se tendono a diminuire.

Il diagramma di dispersione mette in evidenza il campo di variazione delle due variabili, i punti intorno ai quali sono concentrati i loro valori, il tipo di legame che esiste tra le variabili

Un indice in grado di valutare se le due variabili in esame sono linearmente legate fra loro in modo diretto o inverso è la covarianza. Se esiste una relazione di dipendenza lineare diretta fra X e Y , al crescere dei valori assunti da una variabile anche l'altra tende a crescere, per cui gli scarti delle due variabili dalla propria media avranno tendenzialmente segno concorde: a scarti negativi/positivi della X tenderanno a essere associati scarti negativi/positivi della Y .



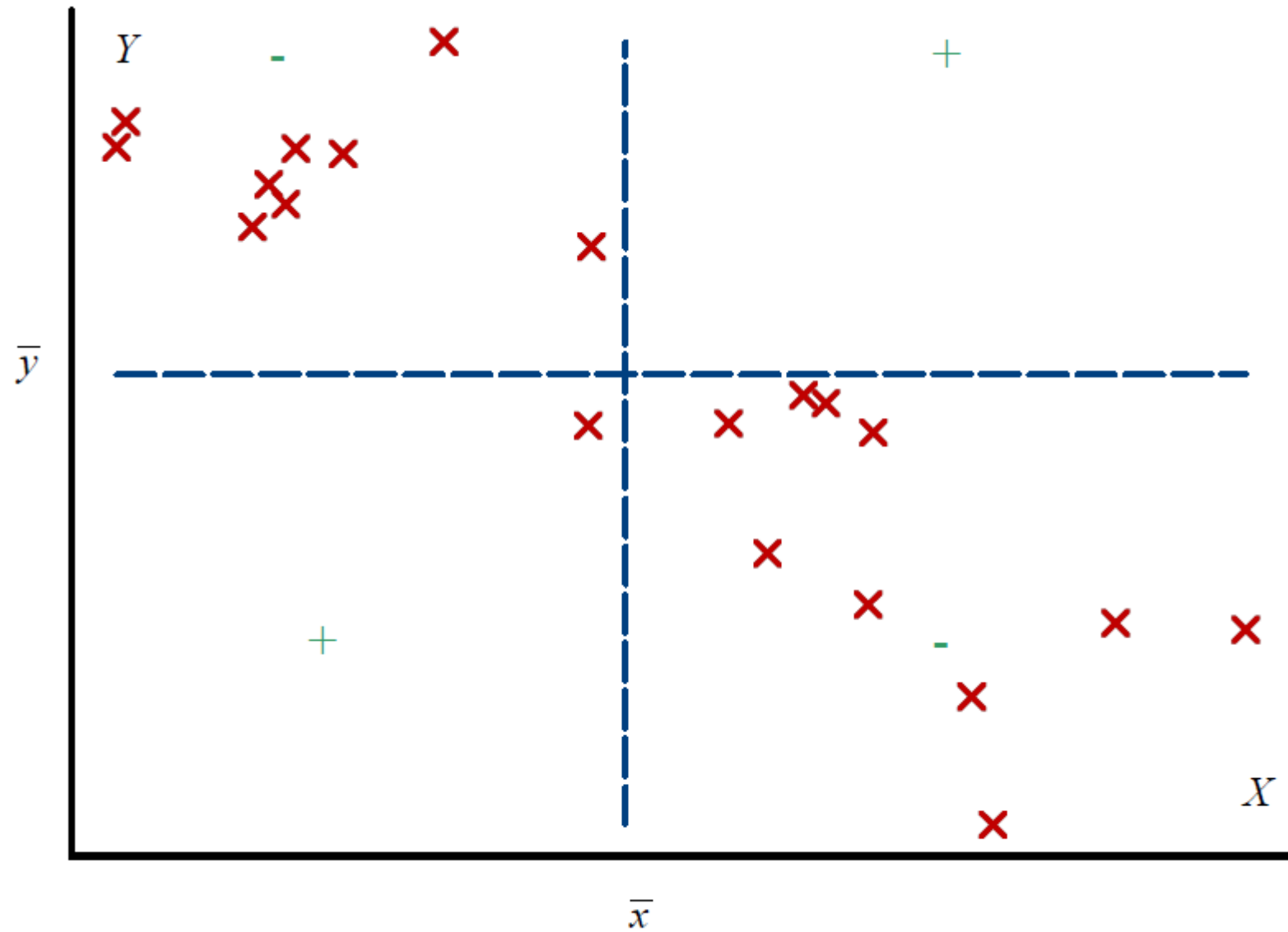
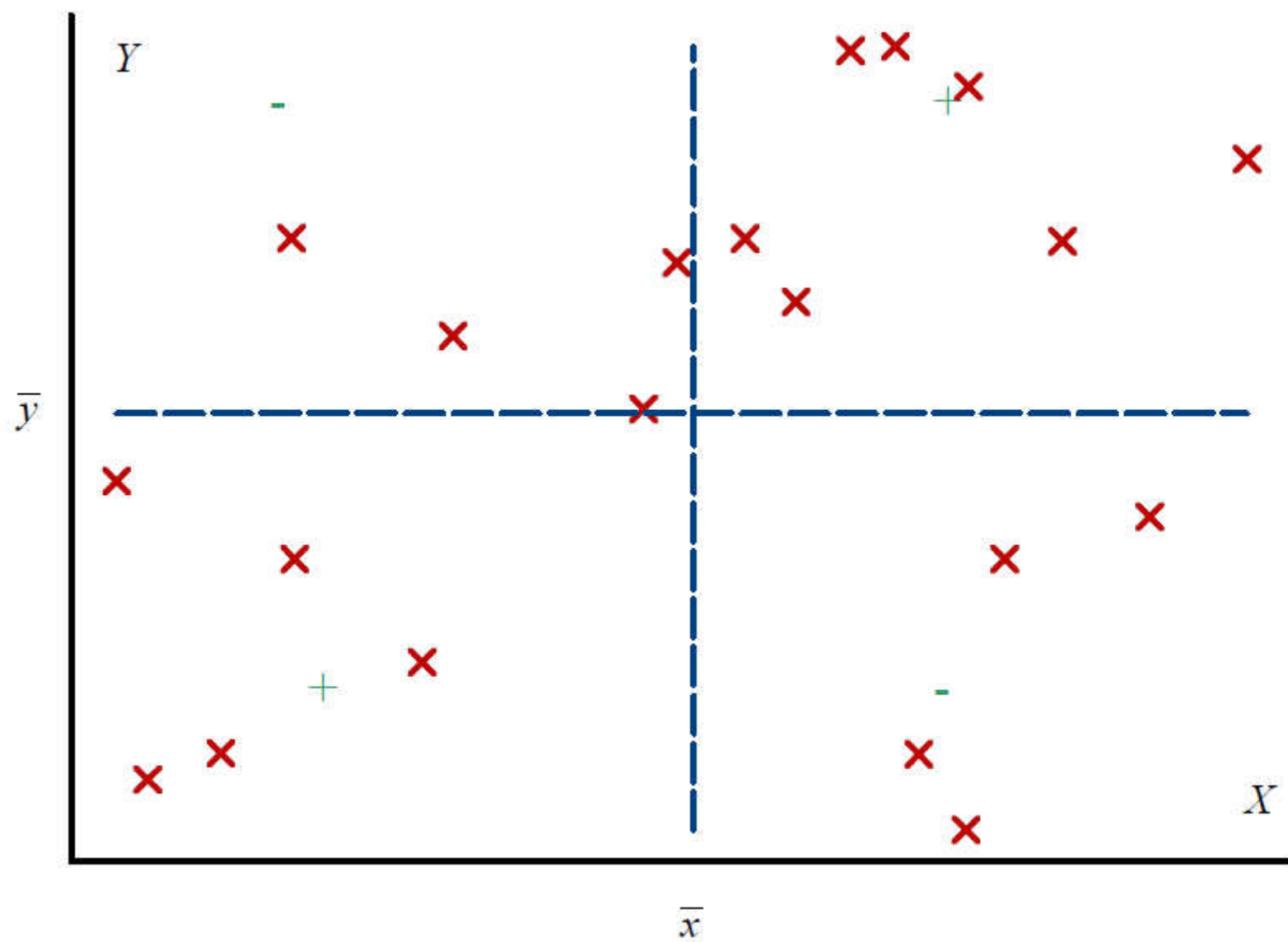


Figura 5.4 – Grafico a dispersione in caso di indipendenza



La **COVARIANZA** è la media dei prodotti degli scostamenti dalla media

5.2 Covarianza

La covarianza consente di verificare se fra due variabili statistiche X e Y esiste un legame lineare e, in caso affermativo, se vi è concordanza oppure discordanza. Siano date le osservazioni $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ rilevate su n unità statistiche e si indichino con $\bar{x}$ e $\bar{y}$ le medie di X e Y rispettivamente. La covarianza è data da

$$(5.2.1) \quad \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = s_{xy}.$$

Essa è la media dei prodotti degli scarti delle osservazioni delle variabili X e Y dalle loro medie.

Se vi è concordanza vi è una maggioranza di coppie di osservazioni nelle quali gli scarti delle variabili X e Y dalla loro media presentano lo stesso segno. Quindi i prodotti $(x_i - \bar{x})(y_i - \bar{y})$ saranno in prevalenza positivi e la covarianza risulterà positiva.

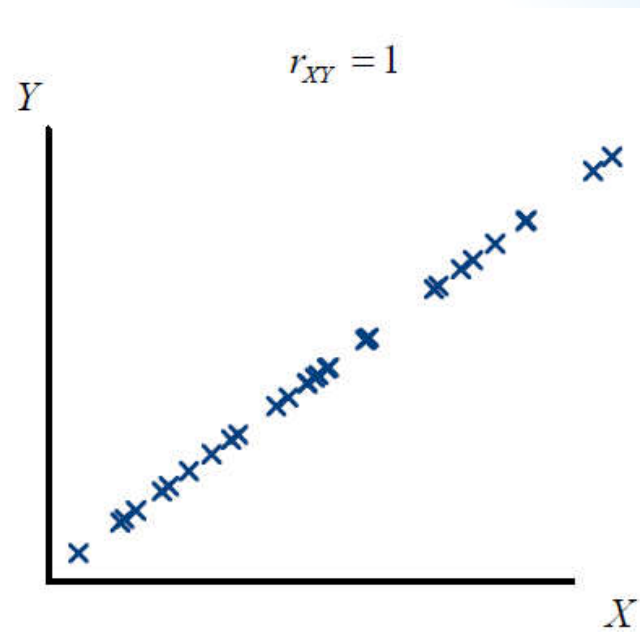
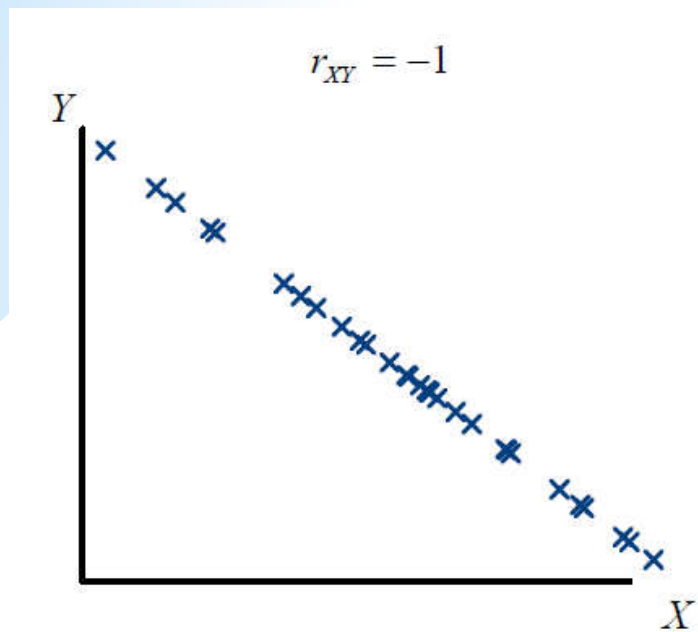
Se invece vi è discordanza vi è una maggioranza di coppie di osservazioni nelle quali gli scarti delle variabili X e Y dalla loro media presentano segno opposto. Quindi i prodotti $(x_i - \bar{x})(y_i - \bar{y})$ saranno in prevalenza negativi e la covarianza risulterà negativa.

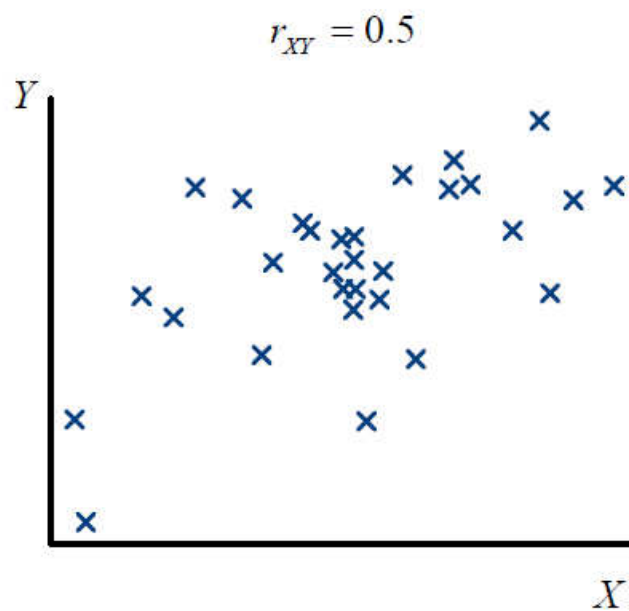
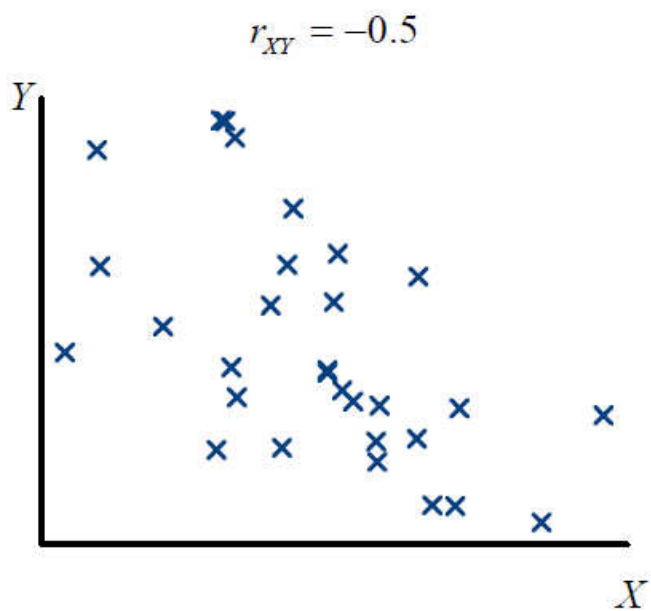
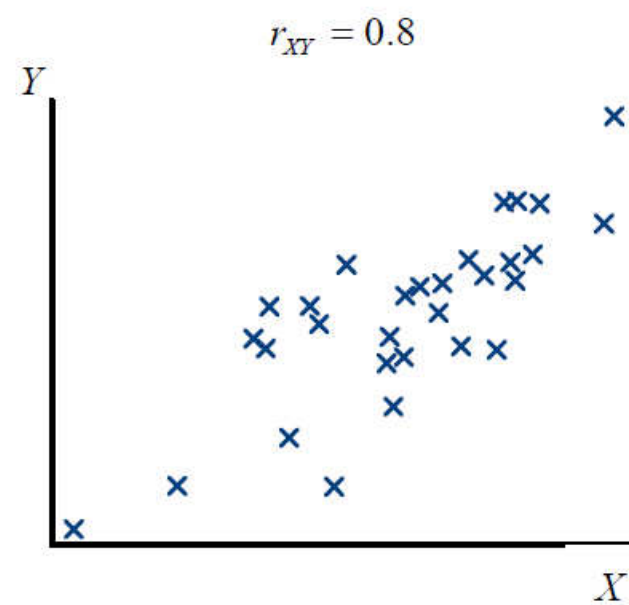
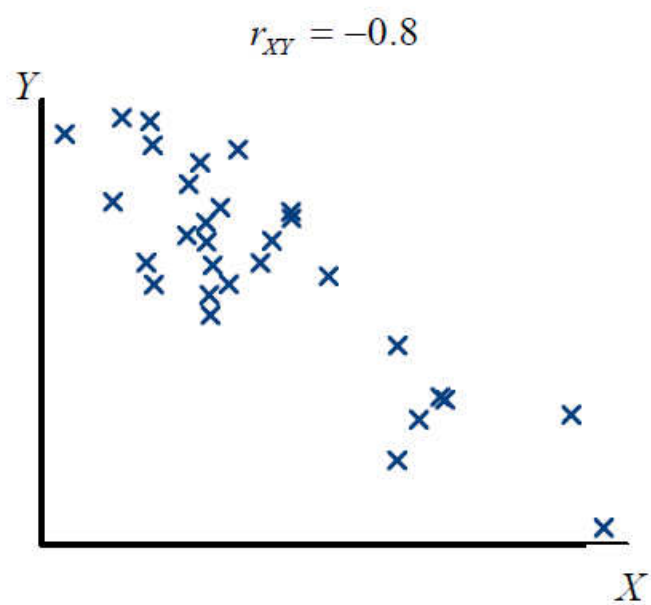
5.6 Calcolo del coefficiente di correlazione

Sostituendo l'espressione della covarianza e degli scarti quadratici medi nella formula del coefficiente di correlazione si ottiene la formula seguente

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Il **COEFFICIENTE DI CORRELAZIONE** è il rapporto tra l'indice di **COVARIANZA** e il prodotto delle **DEVIAZIONI STANDARD** delle variabili X ed Y



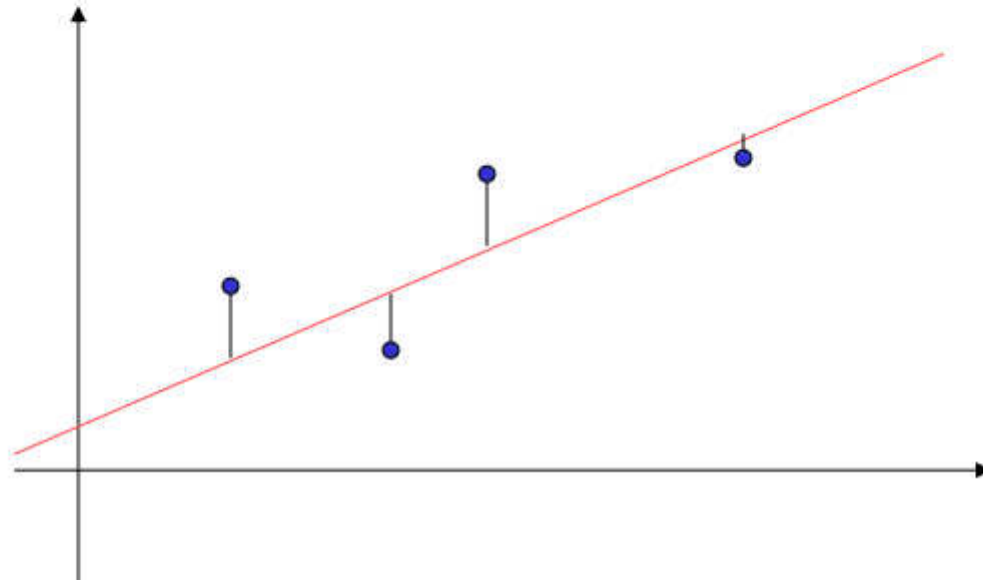


- Una covarianza positiva indica che esiste una relazione di dipendenza lineare diretta
- Una covarianza negativa indica che la relazione lineare è inversa
- Una covarianza pari a zero indica che fra le due variabili non esiste alcuna relazione di dipendenza di tipo lineare.

Regressione lineare: retta di regressione

Si vuole cercare la relazione lineare tra due variabili x e y .

Date n osservazioni congiunte di 2 variabili
cerco due coefficienti a e b tali che $y=ax+b$ passi il più
possibile vicino a questi punti.



Minimi quadrati – regressione lineare

Metodo dei minimi quadrati: consente di determinare un'unica retta di regressione per ogni insieme di dati.

Si richiede che sia minima la somma dei quadrati delle lunghezze dei segmenti che costituiscono le distanze verticali dei punti dalla retta.

