Correlazione

I dati di un'indagine per riflettere

Cominciamo con i dati di un'indagine svolta in una quinta classe di scuola superiore.
Dopo l'Esame di Stato 12 studenti si sono iscritti a corsi di laurea umanistici e 12 a corsi di laurea scientifici.
Gli studenti sono stati seguiti nel loro percorso universitario e si è rilevata la media degli esami sostenuti durante i primi due anni.

Le tabelle seguenti mostrano i dati raccolti.

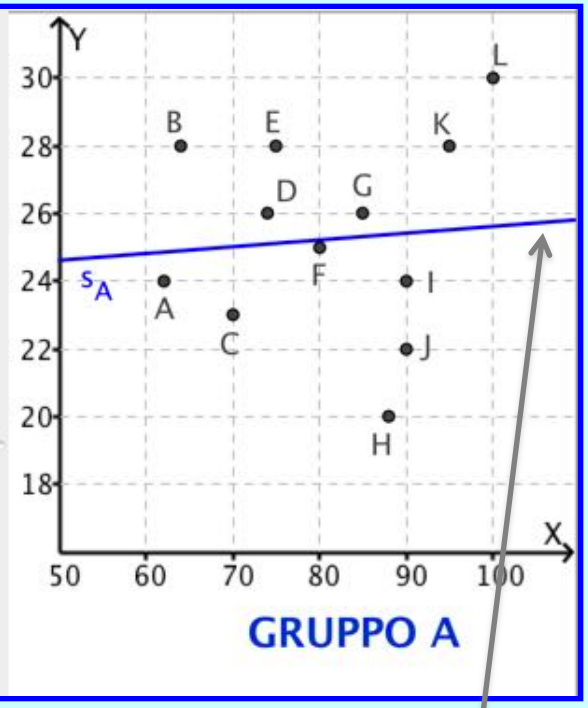
I dati di un'indagine per riflettere

Gruppo A	
Voto di maturità X_A	Media esami universitari Y_A
62	24
64	28
70	23
74	26
75	28
80	25
85	26
88	20
90	24
90	22
95	28
100	30

Gruppo B	
Voto di maturità X_B	Media esami universitari Y_B
70	22
75	23
78	24
80	25
83	24
85	26
88	25
90	27
95	28
98	29
99	30
100	30

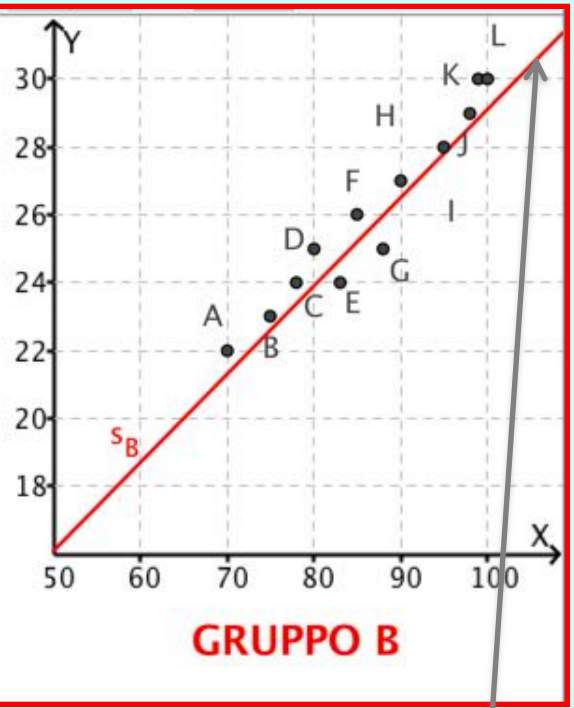
I dati trattati con un foglio di calcolo

A	B
X_A	Y_A
62	24
64	28
70	23
74	26
75	28
80	25
85	26
88	20
90	24
90	22
95	28
100	30



Retta di regressione

A	B
X_B	Y_B
70	22
75	23
78	24
80	25
83	24
85	26
88	25
90	27
95	28
98	29
99	30
100	30

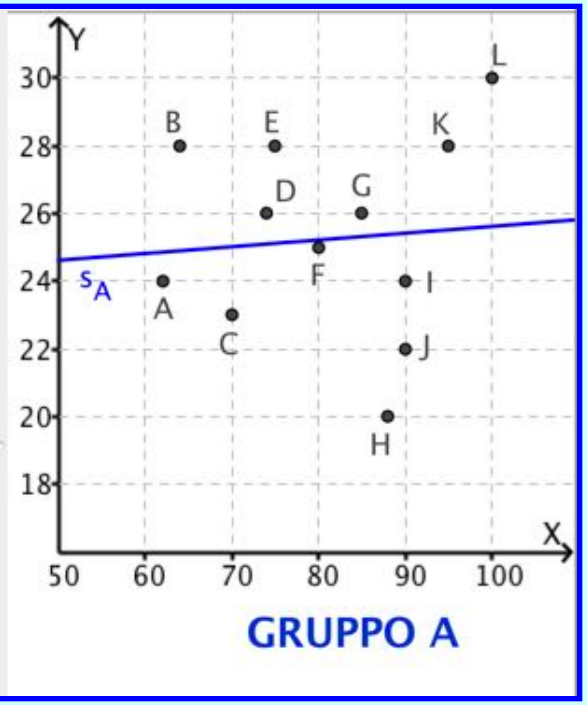


Retta di regressione

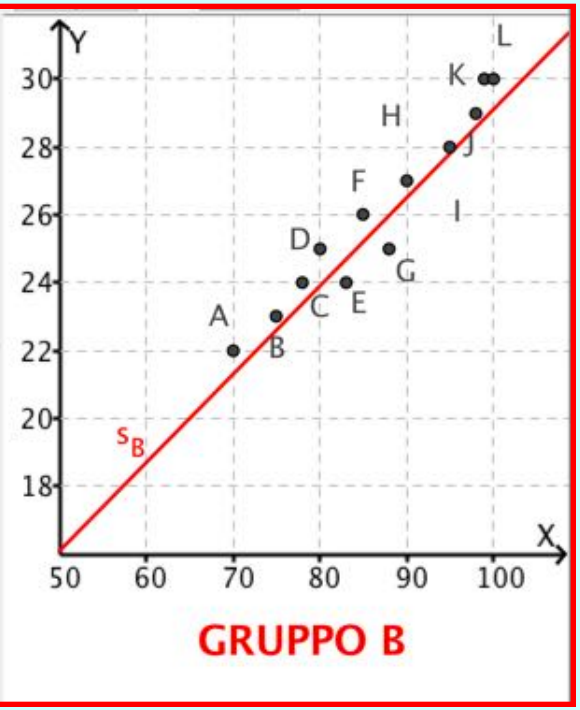
I due gruppi di dati sono molto diversi

I dati trattati con un foglio di calcolo

A	B
X_A	Y_A
62	24
64	28
70	23
74	26
75	28
80	25
85	26
88	20
90	24
90	22
95	28
100	30



A	B
X_B	Y_B
70	22
75	23
78	24
80	25
83	24
85	26
88	25
90	27
95	28
98	29
99	30
100	30

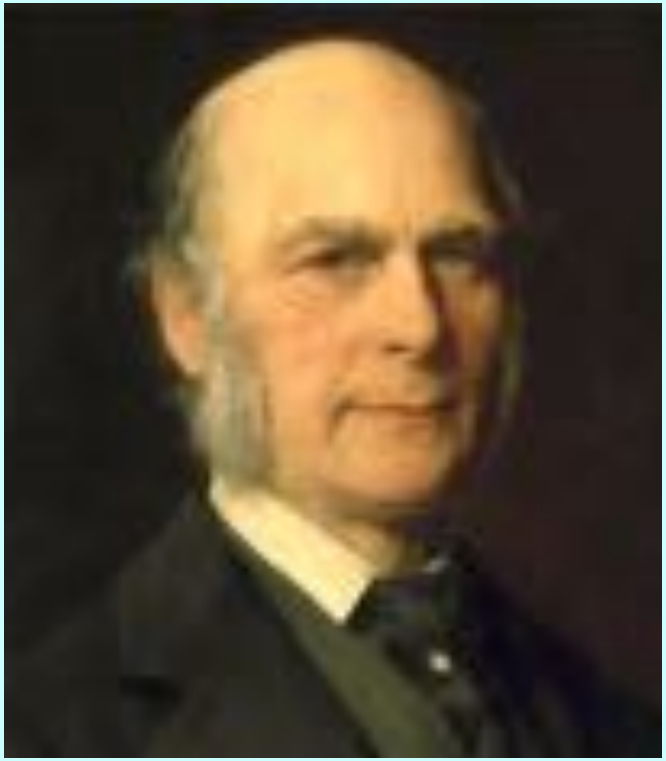


Punti 'sparsi' e lontani dalla retta di regressione

Punti tutti molto vicini alla retta di regressione

C'è un indicatore statistico per stabilire quanto i punti sono vicini alla retta di regressione?

Studi di Galton e Pearson



F. Galton, Regno Unito
1822 - 1911



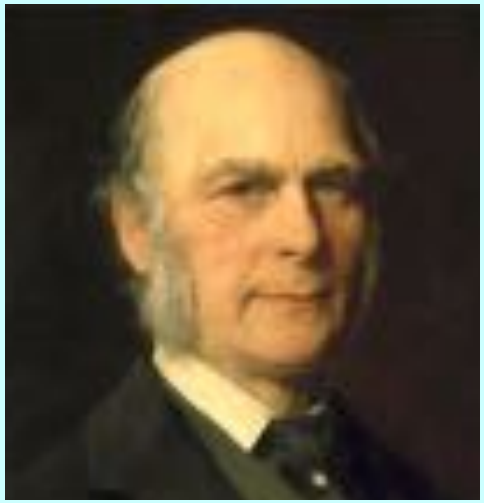
K. Pearson, Regno Unito
1857 - 1936

Il termine '*correlazione*': origini

I. "Co-relations and their Measurement, chiefly from Anthropometric Data." By FRANCIS GALTON, F.R.S. Received December 5, 1888.

"Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but 1 am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.

Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. Thus the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long log, and conversely. If the correlation be



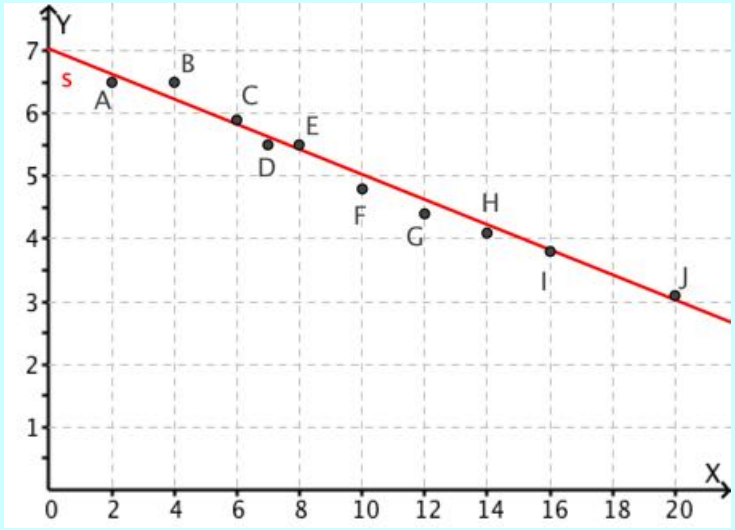
F. Galton

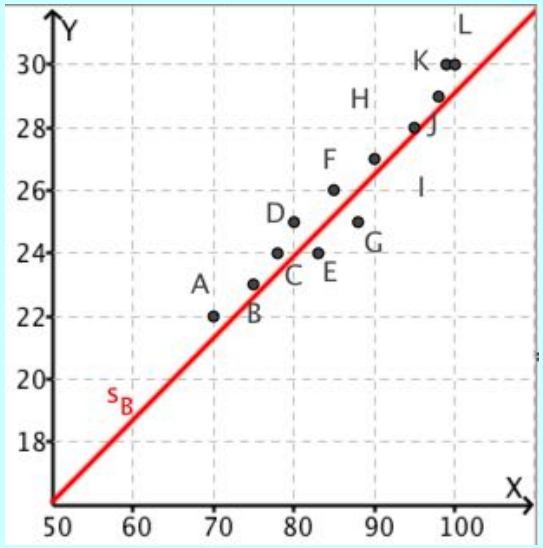
Galton riprende un termine già usato nelle scienze con un largo significato di *'relazione fra due variabili, in modo che a variazioni di una si accompagnino variazioni dell'altra'*.

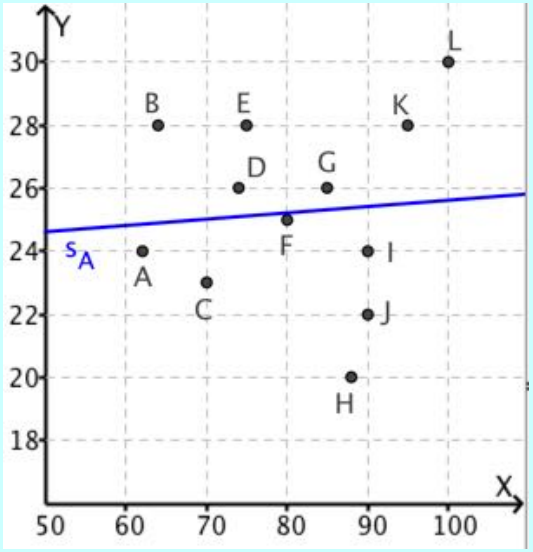
I lavori di Galton, Pearson e dei successivi ricercatori portano a precisare il significato statistico del termine.

Correlazione: il cammino delle idee

Retta di regressione e correlazione lineare







Forte correlazione lineare fra le due variabili

Variabili non correlate linearmente

Lunghe ricerche per trovare *un indice di correlazione lineare*, cioè un numero adatto a misurare l'intensità della correlazione lineare.

L'indice di correlazione lineare *r*

Galton

Il numero *r* è legato alla pendenza m_s della retta di regressione, ma anche alla deviazione standard σ_X dei dati X e σ_Y dei dati Y.

$$r = m_s \cdot \frac{\sigma_X}{\sigma_Y}$$

La formula di Galton sembra snella, ma ogni simbolo sintetizza calcoli che diventano molto lunghi se è grande il numero N di dati:

$$m_s = \frac{\sum_{k=1}^{k=N}(x_k - M_x)(y_k - M_y)}{\sqrt{\sum_{k=1}^{k=N}(x_k - M_x)^2}} \quad , \sigma_x = \sqrt{\frac{\sum_{k=1}^{k=N}(x_k - M_x)^2}{N}} \quad , \sigma_y = \sqrt{\frac{\sum_{k=1}^{k=N}(y_k - M_y)^2}{N}}$$

L'indice di correlazione lineare *r*

Se 'espando' ogni simbolo, ottengo una formula che mostra esplicitamente i calcoli da eseguire a partire dalle N coppie (X, Y) di dati.

Indice di correlazione lineare *r*

Indice di correlazione di Pearson

Il numero *r* è espresso tramite i dati e le medie M_X ed M_Y

$$r = \frac{\sum_{k=1}^{k=N} (x_k - M_x) \cdot (y_k - M_y)}{\sqrt{\sum_{k=1}^{k=N} (x_k - M_x)^2 \cdot \sum_{k=1}^{k=N} (y_k - M_y)^2}}$$

M_X è la media degli N dati X

M_Y è la media degli N dati Y

Indice di correlazione lineare e covarianza

Indice di correlazione di Pearson

Il numero r è espresso tramite i dati e le medie M_X ed M_Y

$$r = \frac{\sum_{k=1}^{k=N} (x_k - M_x) \cdot (y_k - M_y)}{\sqrt{\sum_{k=1}^{k=N} (x_k - M_x)^2} \cdot \sqrt{\sum_{k=1}^{k=N} (y_k - M_y)^2}}$$

La covarianza σ_{xy} è un indice di variabilità congiunta che porta a scrivere anche la formula di Pearson in una forma più sintetica.

$$\sigma_{xy} = \frac{\sum_{k=1}^{k=N} (x_k - M_x) \cdot (y_k - M_y)}{N}$$

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

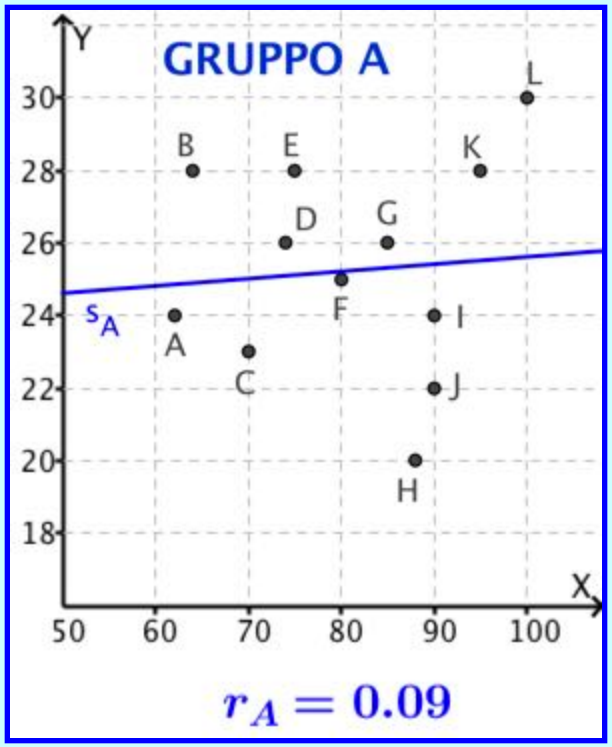
Attività
Indice di correlazione lineare

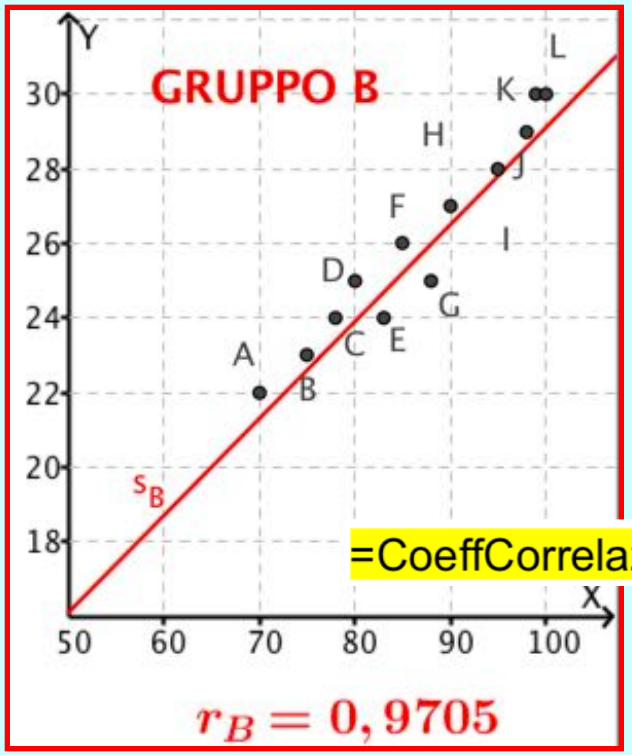
Completa la scheda di lavoro per studiare e applicare l'indice di correlazione lineare.

Che cosa hai ottenuto

Valori dell'indice *r*

Quesiti 1- 2 a, b





=CoeffCorrelazione(Lista2)

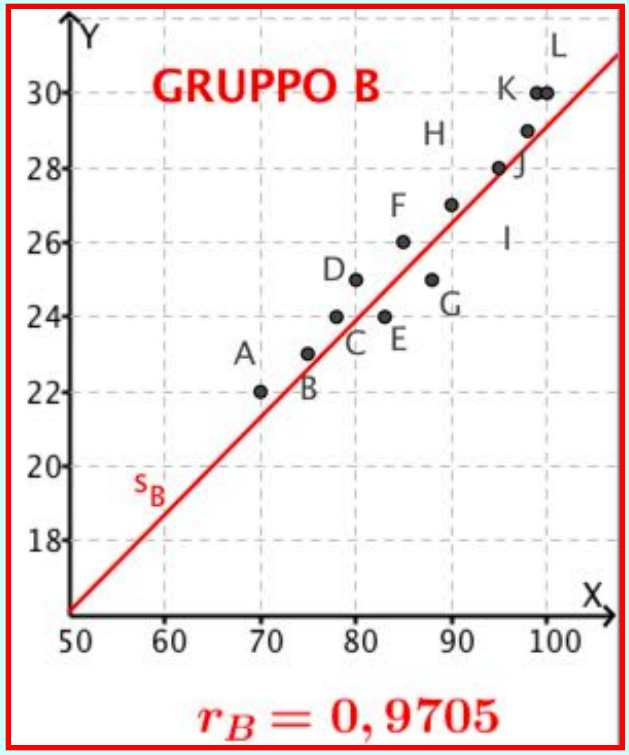
2c. Confronta r_B con r_A e scrivi qui sotto le tue osservazioni.

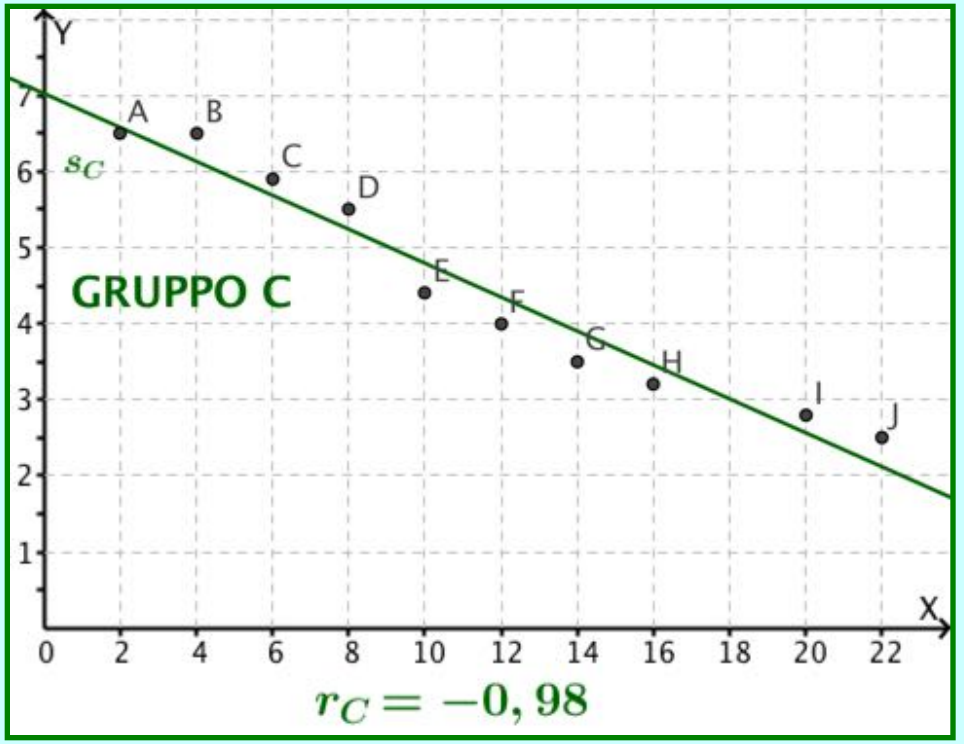
Varie risposte possibili, ad esempio:

r_A positivo e vicino a zero, r_B maggiore di r_A e vicino a uno.

Valori dell'indice *r*

Quesito 3a



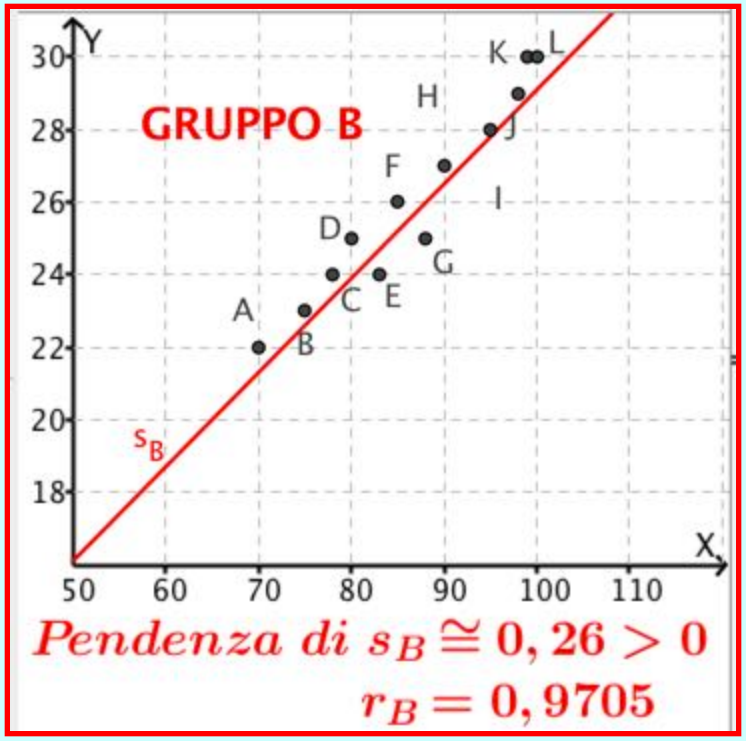


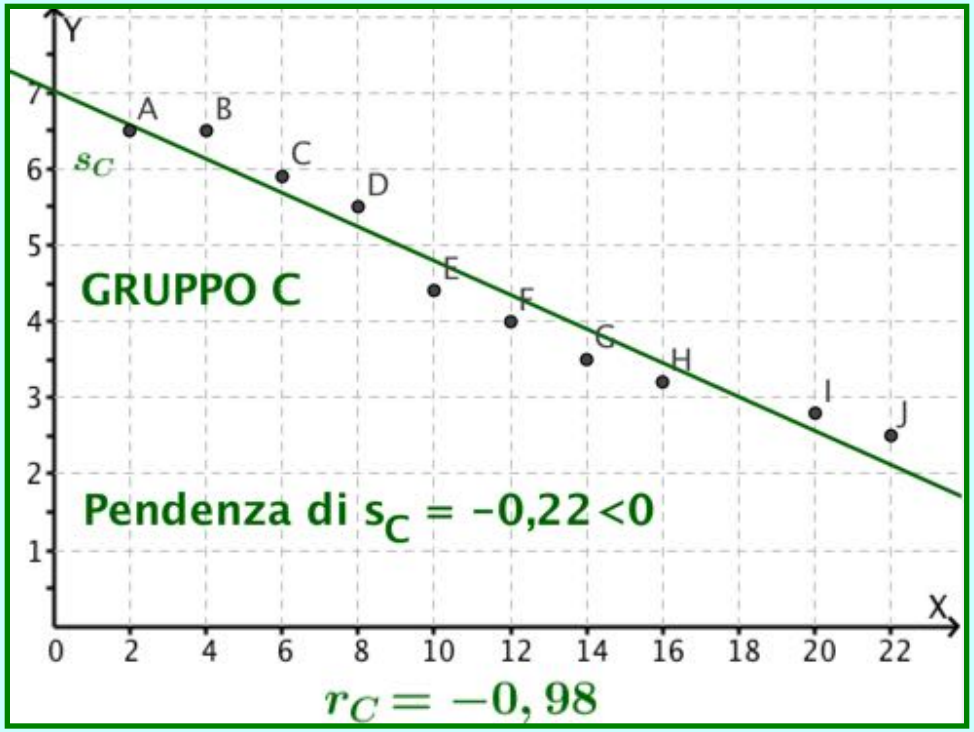
3b. Confronta r_C con r_B e scrivi qui sotto le tue osservazioni.

Varie risposte possibili, ad esempio:

r_B positivo e vicino a 1; r_C negativo e vicino a -1.

Il linguaggio della statistica

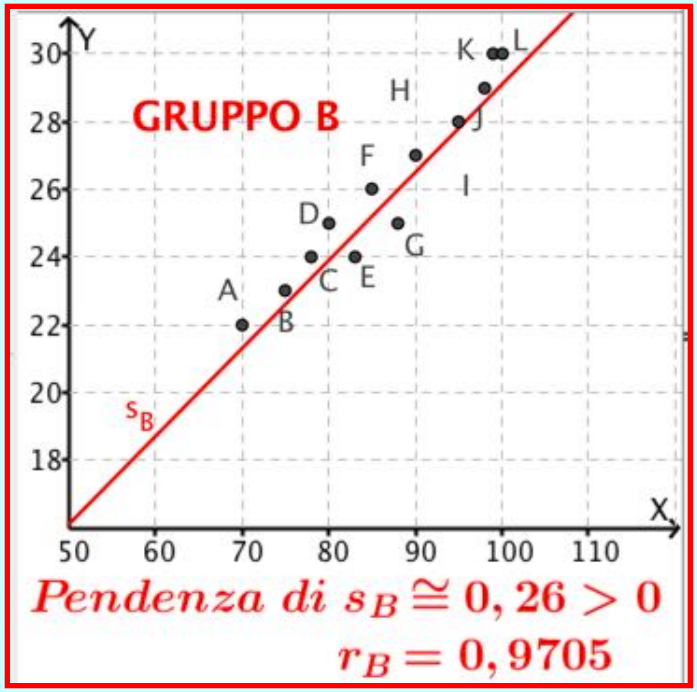


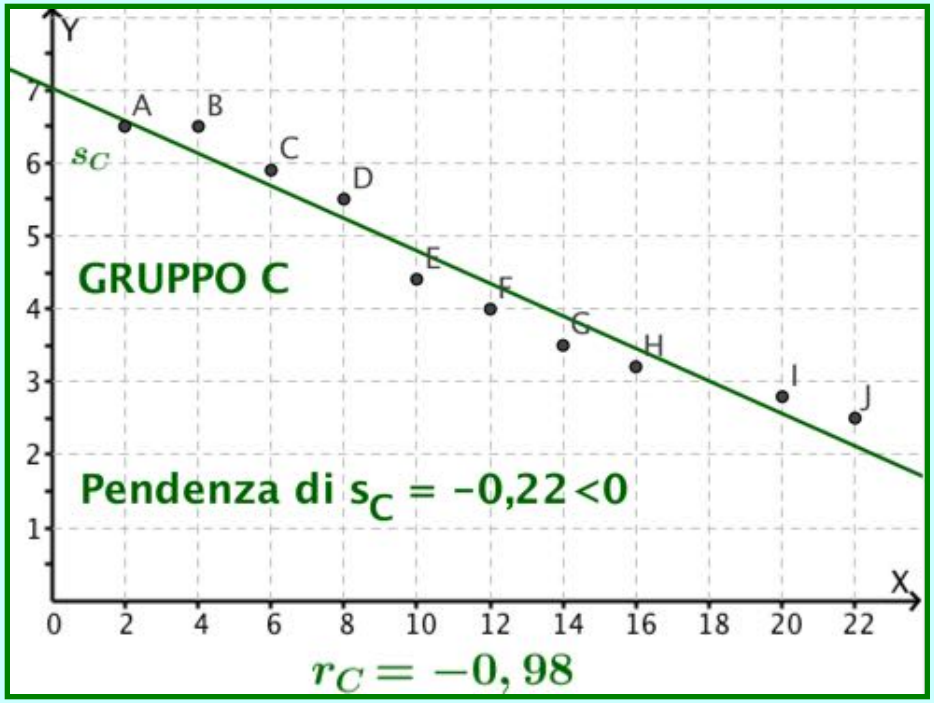


$r > 0$ correlazione positiva
Se cresce X, cresce anche Y.

$r < 0$ correlazione negativa
Se cresce X, decresce Y.

Il segno dell'indice *r*





$$r = m_s \cdot \frac{\sigma_X}{\sigma_Y}$$

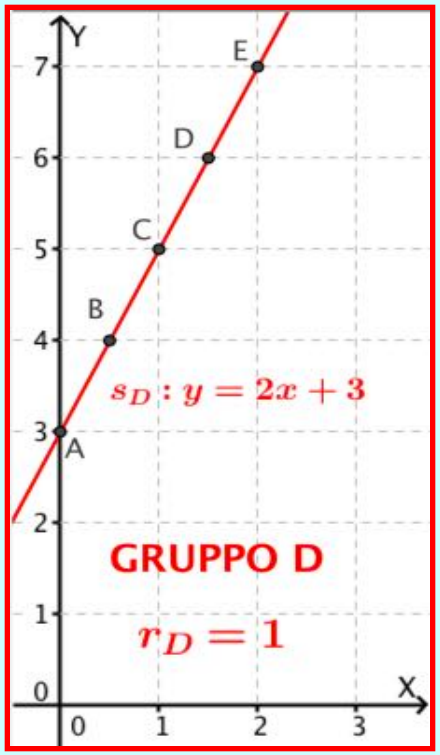
positivo

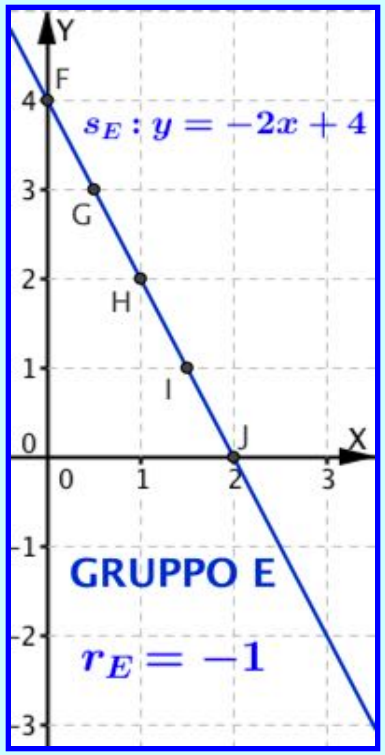
r ha lo stesso segno di m_s

m_s = pendenza della retta di regressione *s*

Valori dell'indice *r*

Quesito 4a





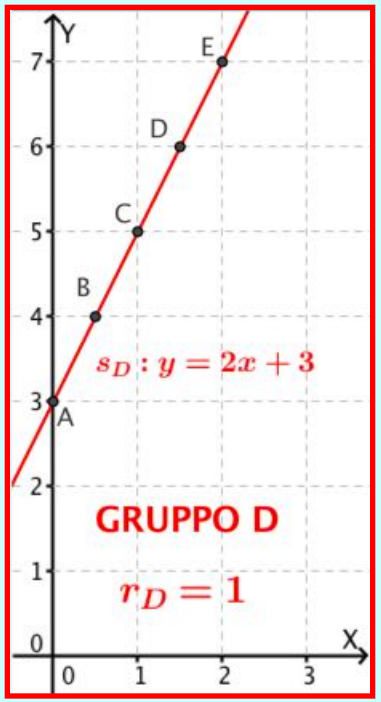
4b. Scrivi qui sotto le tue osservazioni sui risultati ottenuti.

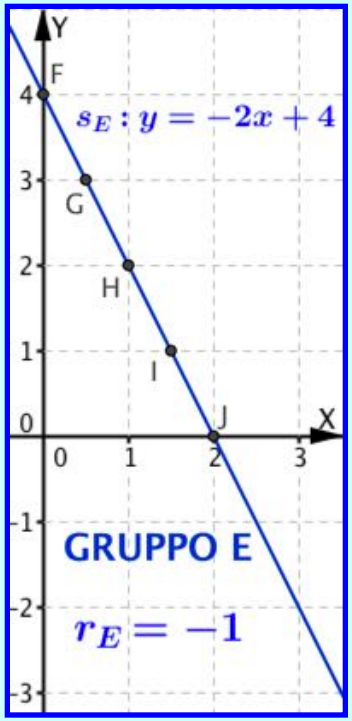
Varie risposte possibili, ad esempio:

$r_D = 1$ e tutti i punti sono sulla retta di regressione con pendenza positiva;

$r_E = -1$ e tutti i punti sono sulla retta di regressione con pendenza negativa.

Il linguaggio della statistica





$r = 1$

Correlazione lineare perfettamente positiva

I punti appartengono tutti alla retta di regressione con pendenza $m_s > 0$.

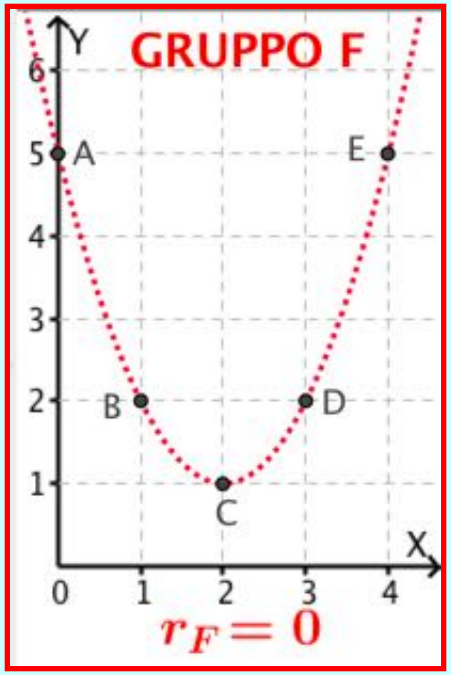
$r = -1$

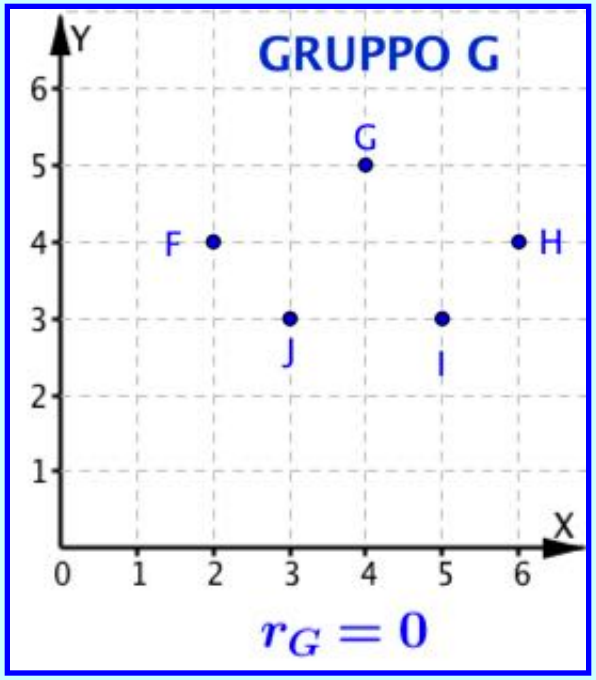
Correlazione lineare perfettamente negativa

I punti appartengono tutti alla retta di regressione con pendenza $m_s < 0$.

Valori dell'indice *r*

Quesito 5a





5b. Scrivi qui sotto le tue osservazioni sui risultati ottenuti

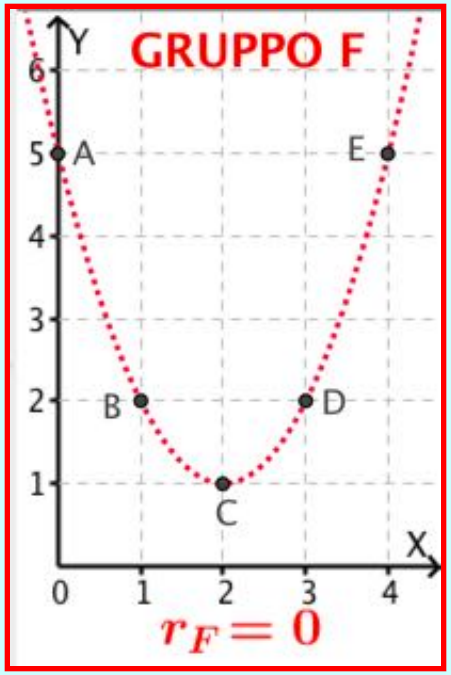
Varie risposte possibili, ad esempio:

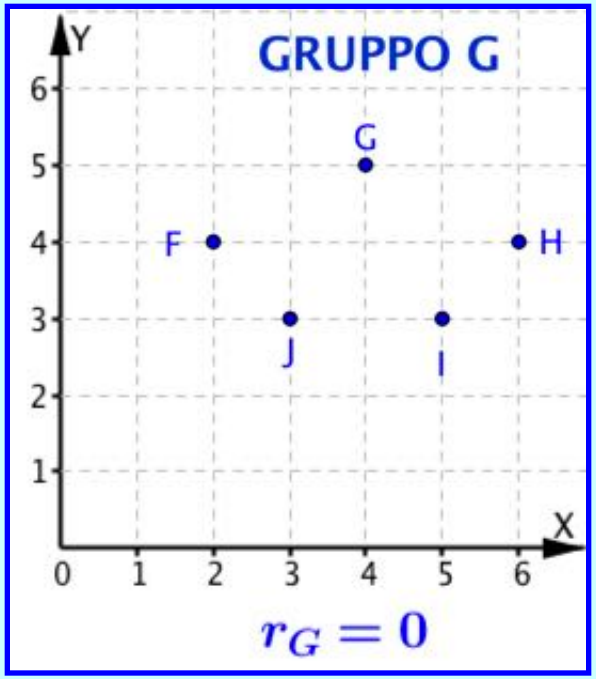
$r_F = 0$ e i punti si trovano su una parabola;

$r_G = 0$ e i punti sono vertici di un poligono con un asse di simmetria parallelo all'asse Y.

In tutti e due i casi i punti non sono 'vicini ad una retta'.

Il linguaggio della statistica





$r = 0$

Assenza di correlazione lineare

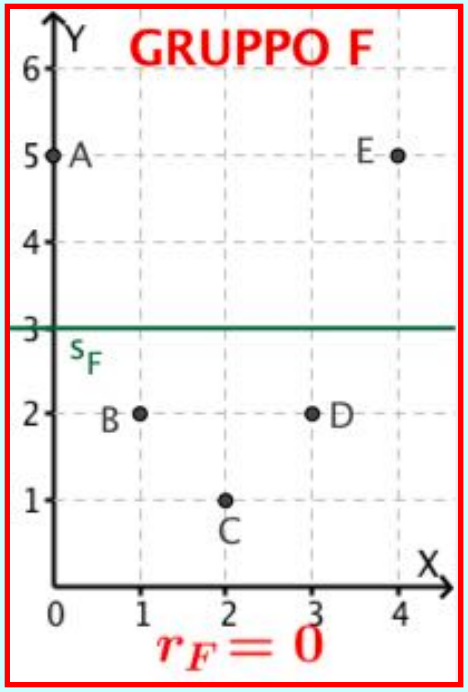
Dati non correlati linearmente

Correlazione lineare e valori di *r*



r = 0 e retta di regressione

X_F	Y_F
0	5
1	2
2	1
3	2
4	5



Quesito 5c

5c. Descrivi un procedimento per ottenere in questo caso l'equazione della retta di regressione s_F con carta e penna.

Retta di regressione s con equazione $y = m_s x + q_s$

Da $r = m_s \cdot \frac{\sigma_x}{\sigma_y}$ e ricavo $m_s = r \cdot \frac{\sigma_y}{\sigma_x}$

Perciò se $r = 0$, risulta $m_s = 0$

Da $q_s = M_y - m_s M_x$ ricavo $q_s = M_y$

La media dei dati Y è $M_y = \frac{5+2+1+2+5}{5} = \frac{15}{5} = 3$

Concludo che l'equazione della retta s_F è

Y= 3

Ma ora la retta di regressione s_F non è utile a fini statistici, perché r = 0 indica che i punti sono sparsi lontano da s_F.

Regressione e correlazione lineari.

Le formule studiate e l'uso del computer permettono di esaminare dati anche senza rappresentarli come punti del piano OXY.

I calcoli sono svolti rapidamente, ma quando i dati sono molto numerosi, come spesso avviene nella applicazioni, come conviene programmare il computer?

Regressione e correlazione lineari: come conviene a esaminare dati?

Conviene programmare il computer in modo da:

1. **Calcolare l'indice di correlazione *r*.**
2. **Stabilire se il valore di *r* indica una correlazione significativa per l'indagine.**
3. **Solo in caso affermativo determinare la retta di regressione.**