STATISTICA DESCRITTIVA

La correlazione

Due caratteri quantitativi insieme: diagramma di dispersione, grafico dei quadranti, covarianza e coefficiente di correlazione lineare, con esercizi svolti

Grafico dei quadranti | Covarianza | Coefficiente r | r = 0 e indipendenza | Correlation is not causation

Fabrizio Antolini · Università di Teramo

PRIMA DI TUTTO · I DATI

Due caratteri quantitativi, sulle stesse unità

Studente	X · ore di studio	Y · voto
A	2	19
B	4	21
C	5	25
D	6	22
…	…	…

Entrambi quantitativi

La correlazione usa medie, scarti e prodotti di scarti: servono valori numerici, con distanze che abbiano un significato.

Rilevati sulle stesse unità

Ogni unità fornisce una coppia (xi, yi): i dati si leggono per righe, non come due colonne separate.

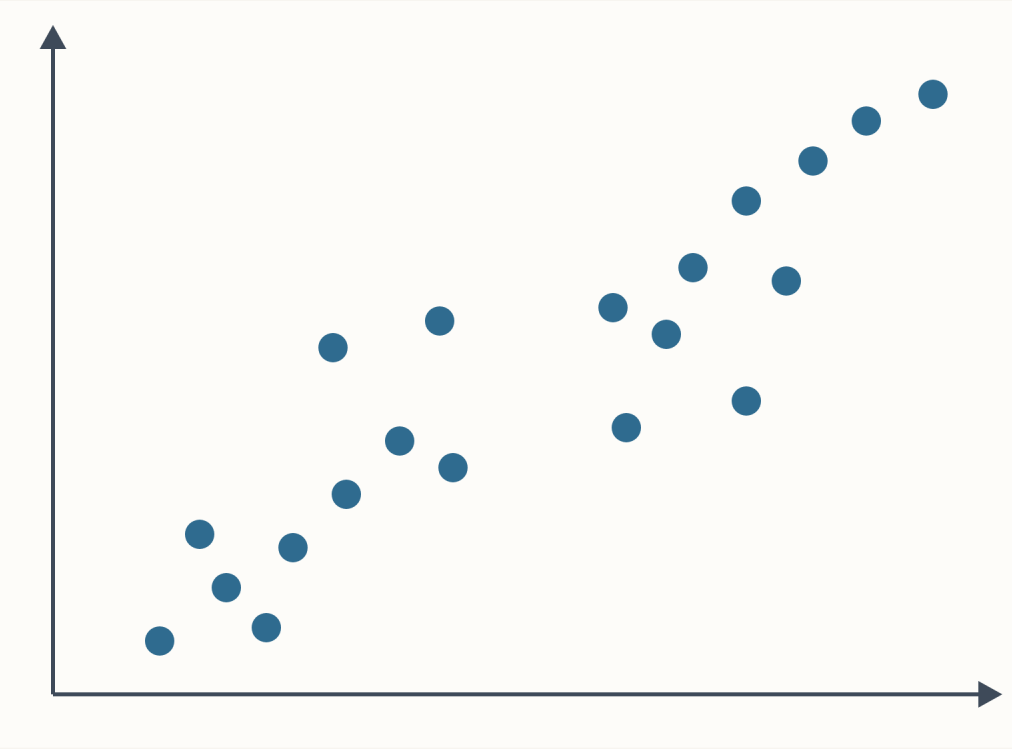
Due caratteri quantitativi
correlazione lineare (Bravais-Pearson)

Due caratteri ordinali
cograduazione (Spearman)

Caratteri qualitativi
connessione (indice χ^2)

CORRELAZIONE · IL GRAFICO

Il diagramma di dispersione



Ogni punto è un’unità: ascissa xi, ordinata yi. Si guarda **prima** di calcolare qualsiasi indice.

Direzione

Crescente se la nuvola sale da sinistra a destra, decrescente se scende.

Forma

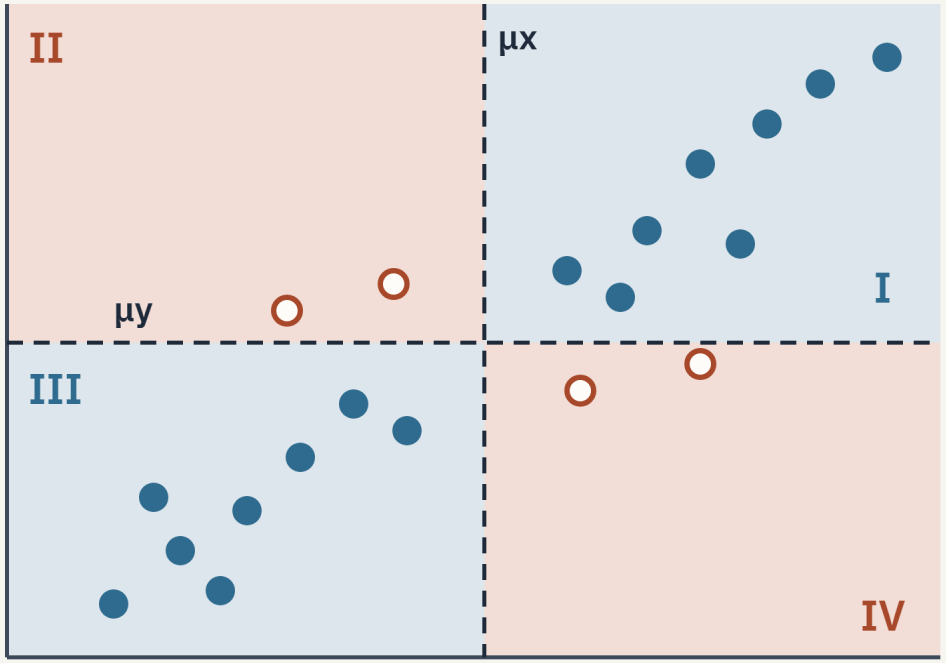
Lineare se i punti si dispongono attorno a una retta; altrimenti curva, o nessuna forma.

Intensità

Quanto i punti stanno vicini a quella retta.

CORRELAZIONE · IL GRAFICO DEI QUADRANTI

Le medie dividono la nuvola in quattro



Quadrante	xi − μx	yi − μy	Prodotto
I	+	+	+
II	−	+	−
III	−	−	+
IV	+	−	−

Punto pieno: prodotto degli scarti positivo. Punto vuoto: prodotto negativo.

Prevalgono I e III → somma dei prodotti positiva, legame crescente. Prevalgono II e IV → somma negativa, legame decrescente.

COVARIANZA · TEORIA

La media dei prodotti degli scarti

Definizione

$$\sigma_{xy} = \Sigma\,(x_i - \mu_x)(y_i - \mu_y) / N$$

Formula di calcolo

$$\sigma_{xy} = \Sigma\, x_i\, y_i / N - \mu_x\, \mu_y$$

$\sigma_{xy} > 0$	$\sigma_{xy} = 0$	$\sigma_{xy} < 0$
prevalgono I e III: legame crescente	i prodotti si compensano	prevalgono II e IV: legame decrescente

Il limite: dipende dalle unità di misura

Misurando lo studio in minuti invece che in ore la covarianza si moltiplica per 60. Il segno è affidabile, il valore non dice se il legame è forte o debole.

COEFFICIENTE DI CORRELAZIONE · TEORIA

La covarianza diventa un numero puro

Coefficiente di correlazione lineare di Bravais-Pearson

$$r = \sigma_{xy} / (\sigma_x \cdot \sigma_y)$$

la covarianza divisa per il prodotto dei due scarti quadratici medi

Con i punti standardizzati

$$r = \Sigma\, z_{xi} \cdot z_{yi} / N$$

con $z_{xi} = (x_i - \mu_x) / \sigma_x$ e $z_{yi} = (y_i - \mu_y) / \sigma_y$: r è la covarianza dei dati standardizzati.

Proprietà

- Non dipende dalle unità di misura
- Ha lo stesso segno della covarianza
- È simmetrico: r(X, Y) = r(Y, X)

COEFFICIENTE DI CORRELAZIONE · TEORIA

Perché r varia tra −1 e +1

1 · I punti z hanno varianza 1

$$\Sigma\, zx^2 / N = \Sigma\, zy^2 / N = 1$$

Standardizzare porta media a 0 e varianza a 1.

2 · Il tetto: r ≤ +1

$$\Sigma\,(zx - zy)^2 / N = 1 - 2r + 1 = 2\,(1 - r) \geq 0$$

Una somma di quadrati non è mai negativa.

3 · Il pavimento: r ≥ −1

$$\Sigma\,(zx + zy)^2 / N = 1 + 2r + 1 = 2\,(1 + r) \geq 0$$

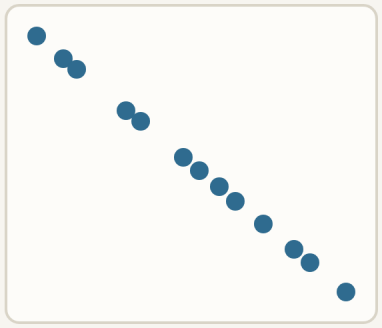
Lo stesso ragionamento con il segno più.

r = +1 solo se zx = zy per ogni unità: tutti i punti su una retta crescente. **r = −1** solo se zx = −zy: tutti su una retta decrescente.

Più i punti si allontanano da una retta, più r si avvicina a 0: |r| misura quanto il legame è vicino a una relazione lineare esatta.

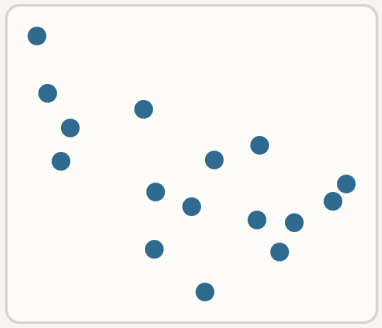
COEFFICIENTE DI CORRELAZIONE · INTERPRETAZIONE

Il segno dà la direzione, |r| l'intensità



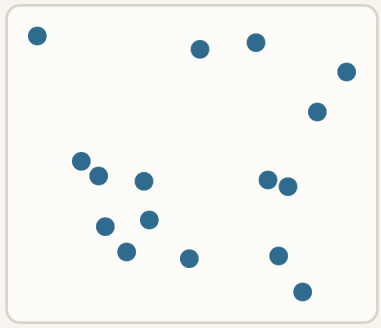
r = −1

lineare perfetto, decrescente



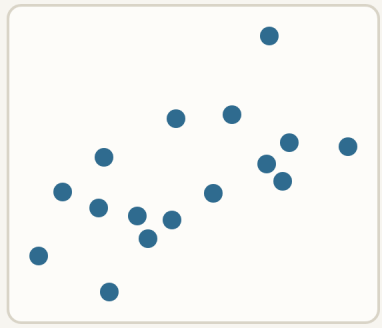
r ≈ −0,6

decrescente, moderato



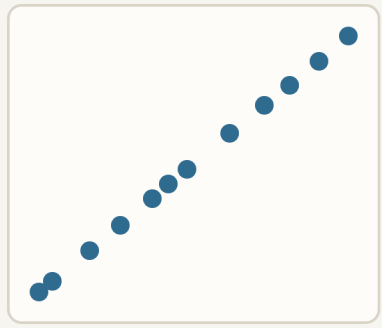
r ≈ 0

nessun legame lineare



r ≈ +0,6

crescente, moderato



r = +1

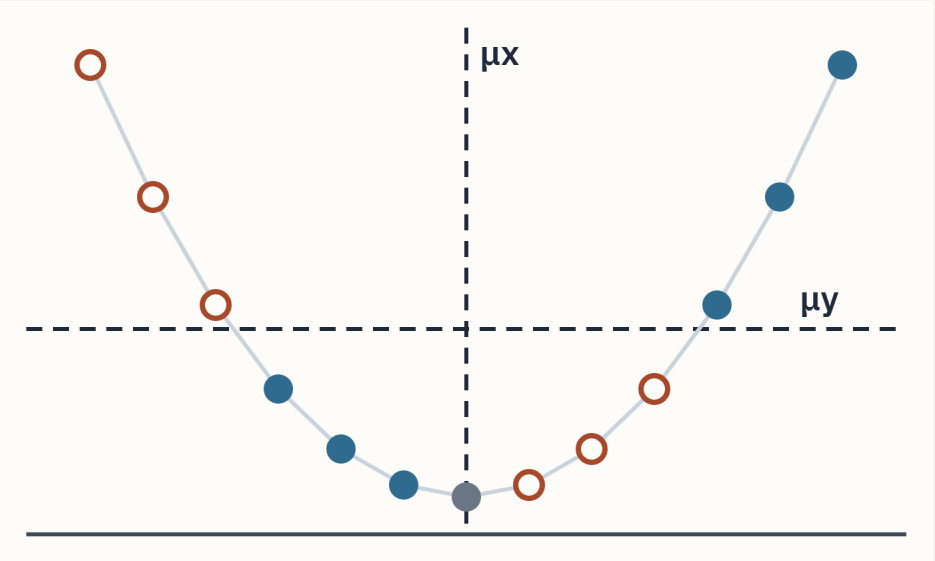
lineare perfetto, crescente

|r| < 0,3 · legame lineare debole	**0,3 ≤ |r| < 0,7** · moderato	**|r| ≥ 0,7** · forte

Soglie indicative: cambiano tra manuali e discipline, e vanno sempre lette insieme al grafico.

IL CASO R = 0

Indifferenti non vuol dire indipendenti



Y è determinata esattamente da X ($y = x^2$), eppure r = 0: i prodotti dei punti pieni e di quelli vuoti si annullano.

Indipendenza ⇒ r = 0

Se X e Y sono indipendenti la covarianza è nulla, quindi anche r.

r = 0 non implica indipendenza

Con r = 0 le variabili sono **indifferenti** (incorrelate): manca un legame lineare, ma può esistere un legame di altra forma.

r = 0 significa assenza di legame lineare, non assenza di legame.

ATTENZIONE

Correlation is not causation

La correlazione misura un’associazione tra due caratteri. Non dice se uno causa l’altro, né in quale direzione.

Una terza variabile

Gelati venduti e annegamenti salgono insieme: li spinge entrambi il caldo estivo.

Causalità inversa

Più agenti dove ci sono più reati: sono i reati ad attirare gli agenti, non il contrario.

Coincidenza

Due serie che crescono nel tempo risultano correlate anche senza alcun legame.

Per parlare di causa servono un disegno di ricerca o un esperimento, non un indice.

CORRELAZIONE · CAUTELE

Prima di fidarsi di r

Misura solo legami lineari

Un legame curvo forte può dare r vicino a 0, come nella parabola della slide precedente.

È sensibile ai valori anomali

Un solo punto lontano dalla nuvola può creare una correlazione che non c'è, o cancellarne una che c'è.

Attenzione ai sottogruppi

Unire gruppi diversi può far nascere una correlazione assente nei singoli gruppi, o invertirne il segno.

Non è una pendenza

y = x e y = 10x danno entrambe r = 1: quanto cresce Y lo dirà la regressione, non r.

Regola pratica: prima il diagramma di dispersione, poi il coefficiente.

SECONDA PARTE

Esercizi svolti

Due distribuzioni semplici. Nella prima r è vicino a +1; nella seconda r = 0, anche se le variabili sono tutt’altro che indipendenti.

1 · Ore di studio e voto: calcolo con gli scarti

1 · Grafico dei quadranti

1 · Formula di calcolo

2 · Temperatura e consumo di energia

ESERCIZIO SVOLTO 1 · DATI SEMPLICI · CALCOLO CON GLI SCARTI

Ore di studio settimanali e voto di 8 studenti

X = ore di studio, Y = voto (trentesimi) · dx = xi − μx · dy = yi − μy

xi	yi	dx	dy	dx · dy	dx²	dy²
2	19	−5	−5	+25	25	25
4	21	−3	−3	+9	9	9
5	25	−2	+1	−2	4	1
6	22	−1	−2	+2	1	4
8	23	+1	−1	−1	1	1
9	26	+2	+2	+4	4	4
10	28	+3	+4	+12	9	16
12	28	+5	+4	+20	25	16
56	192	0	0	69	78	76

1 · Medie

μx = 56 / 8 = 7 · μy = 192 / 8 = 24

2 · Covarianza

σxy = 69 / 8 = 8,625

3 · Scarti quadratici medi

σx = √(78 / 8) = √9,75 = 3,1225
σy = √(76 / 8) = √9,5 = 3,0822

4 · Coefficiente di correlazione

r = 8,625 / (3,1225 · 3,0822) = 8,625 / 9,6244 = **0,90**