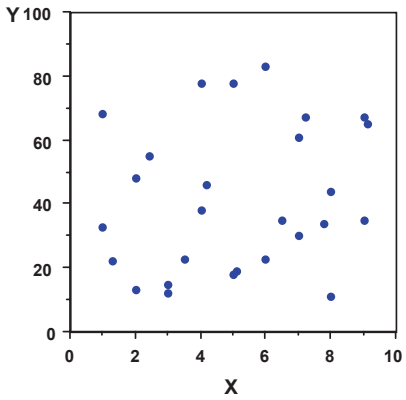
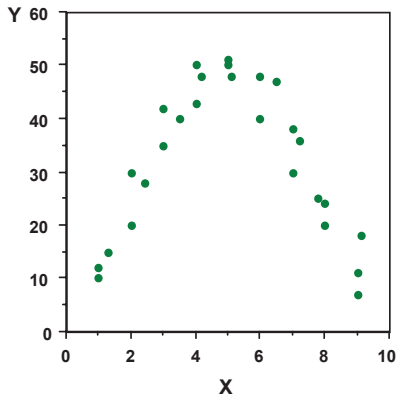
LA CORRELAZIONE LINEARE

C'è correlazione tra due variabili X e Y quando queste tendono a variare insieme, cioè quando all'aumentare dei valori di X, i valori di Y mostrano un andamento regolare (e viceversa). Se questo andamento assume una forma rettilinea, si parla di **correlazione lineare**, mentre se assume una forma curvilinea si parla di **correlazione non lineare**. C'è assenza di correlazione quando i valori di Y variano in modo casuale al variare dei valori di X (e viceversa). La **correlazione lineare** è **positiva** se all'aumentare dei valori di X, i valori di Y aumentano, mentre è **negativa** se all'aumentare dei valori di X i valori di Y diminuiscono. Esempi (**diagrammi di dispersione**):

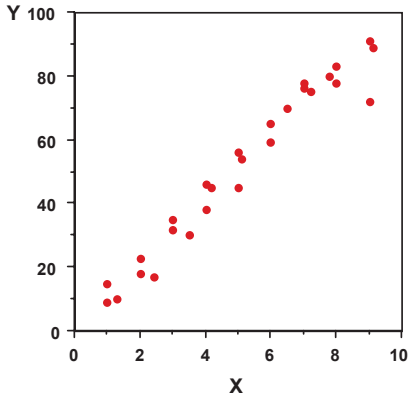
ASSENZA DI CORRELAZIONE



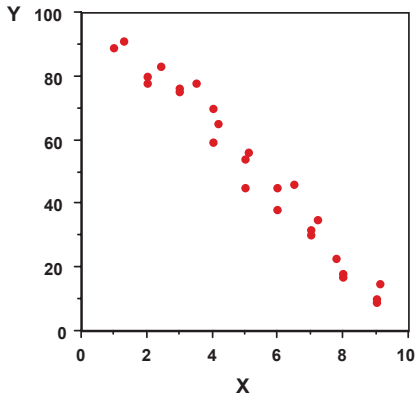
CORRELAZIONE NON LINEARE



CORRELAZIONE LINEARE POSITIVA



CORRELAZIONE LINEARE NEGATIVA



Sebbene esistano metodi statistici per valutare anche la significatività di correlazioni non lineari, quelli più utilizzati riguardano la correlazione lineare, per cui ci occuperemo solo di questo tipo di correlazione.

Coefficienti di correlazione lineare

Per esprimere la relazione lineare tra due variabili si usa un indice chiamato **coefficiente di correlazione**. Esistono diversi coefficienti di correlazione lineare, il loro uso dipende dal livello di misura delle variabili; essi variano nelle formule di calcolo, ma la loro interpretazione è esattamente la stessa.
Il coefficiente di correlazione può assumere **tutti i valori compresi tra -1 e +1** e fornisce due informazioni, una relativa alla **forza** o **intensità** della relazione tra X e Y, indicata dal **valore assoluto**, e l'altra relativa alla **direzione** della relazione, indicata dal **segno** positivo o negativo. Maggiore è il valore assoluto del coefficiente di correlazione (cioè più vicino ad 1), maggiore sarà la forza della relazione tra X e Y, minore è il suo valore assoluto (cioè più vicino a 0), più debole sarà la relazione tra X e Y. Così, per esempio, un coefficiente di correlazione di -0.88 indica una correlazione elevata negativa, mentre un coefficiente di 0.12 indica una bassa correlazione positiva.
Se il coefficiente di correlazione è **esattamente 1** (oppure **-1**), la correlazione tra X e Y è **perfetta**, cioè all'aumentare di una unità di X, Y aumenta (o diminuisce) **di una quantità costante**.
Se un coefficiente di correlazione lineare è molto vicino al valore 0, non si può affermare che non c'è correlazione tra X e Y, perchè le due variabili potrebbero essere legate da una relazione non lineare, che questi coefficienti non sono in grado di evidenziare. In caso di **correlazione 0** (o vicina a 0), è più corretto dire che **non c'è correlazione lineare**.
Quando si formula una ipotesi di correlazione tra due variabili, non si fa riferimento ad un rapporto di causa ed effetto ed infatti in genere non c'è una Variabile Indipendente ed una Variabile Dipendente. Si ipotizza cioè che le variazioni di Y possano essere spiegate, in una certa misura, dalle variazioni di X e, allo stesso modo, che le variazioni di X possano essere spiegate dalle variazioni di Y. Così se si dimostra l'esistenza di una correlazione tra, per esempio, velocità di lettura e correttezza di lettura, memoria a breve termine e comprensione di testi, ecc., ciò **non implica** che una delle due variabili sia causa dell'altra.

Il coefficiente r di Pearson

E' il coefficiente di correlazione più usato; si utilizza quando X e Y sono entrambe su **scala a intervalli** o **rapporti equivalenti**. Corrisponde alla media dei prodotti dei punti z delle variabili X e Y. Date cioè due variabili X e Y misurate sullo stesso gruppo di soggetti, volendo calcolare a mano il coefficiente **r** di Pearson bisogna trasformare in punti z tutti i valori X e Y, poi moltiplicare i due punti z per X e per Y di ogni soggetto, sommare tutti i prodotti ed infine dividere per il numero di soggetti. In formula:

$$r = \frac{\Sigma z_x z_y}{N}$$

Questa formula non è in genere utilizzata per il calcolo di **r** in quanto esistono formule semplificate, che però derivano tutte da questa. La formula originaria, comunque, serve a chiarire il significato di **r**, infatti poichè z rappresenta la distanza del punteggio rispetto alla media del gruppo, il prodotto tra z_x e z_y di un soggetto è una misura della concordanza tra X e Y per quel soggetto. Dividendo per N, poi, si ottiene una misura media della concordanza dei valori X e Y per tutto il campione. E' evidente che se tutti i soggetti hanno punti z simili (in valore assoluto) per X e per Y, il valore di **r** sarà più elevato, positivo se le coppie di punti z hanno segno uguale, negativo se hanno segno diverso. Se tutti i soggetti hanno esattamente lo stesso punto z per X e per Y, il coefficiente **r** sarà uguale a 1 (correlazione perfetta).

Esempio 1

SOGG.	X	Y	z_x	z_y	$z_x z_y$
1	2	21	-1,609	-1,487	2,392
2	8	78	0,866	0,749	0,649
3	6	54	0,041	-0,192	-0,008
4	3	33	-1,196	-1,016	1,215
5	5	48	-0,371	-0,428	0,159
6	4	32	-0,784	-1,055	0,827
7	9	96	1,279	1,456	1,861
8	6	64	0,041	0,200	0,008
9	7	75	0,454	0,632	0,287
10	9	88	1,279	1,142	1,460
				$\Sigma z_x z_y$ =	8,850

r = 8.850 /10 = **0.88** (correlazione positiva elevata)

Esempio 2

SOGG.	X	Y	z_x	z_y	$z_x z_y$
1	1	54	-1,429	0,282	-0,404
2	5	21	0,075	-0,882	-0,066
3	3	65	-0,677	0,671	-0,454
4	7	11	0,828	-1,235	-1,022
5	4	58	-0,301	0,424	-0,127
6	8	92	1,204	1,623	1,954
7	6	51	0,451	0,176	0,080
8	9	10	1,580	-1,271	-2,007
9	2	23	-1,053	-0,812	0,855
10	3	75	-0,677	1,023	-0,693
				$\Sigma z_x z_y$ =	-1,885

r = -1.885 /10 = **- 0.19** (correlazione negativa debole)

Il valore di **r** calcolato è un indicatore **descrittivo** della relazione tra X e Y nel campione. Perché sia possibile generalizzare alla popolazione il risultato ottenuto, si fa riferimento alla distribuzione campionaria di **r** tabulata nelle tavole (v. Ercolani & Areni, "Statistica per la ricerca in psicologia").

Per H_0: $\rho = 0$; per H_1: $\rho \neq 0$

Il simbolo ρ (che si legge "ro") indica il coefficiente di correlazione nella popolazione. Per H**0** nella popolazione vi è assenza di correlazione tra X e Y (quindi $\rho = 0$). Per H**1** nella popolazione vi è correlazione tra X e Y, quindi □ è significativamente diverso da 0 (**ipotesi bidirezionale**). Dal momento che **r** può assumere sia valori positivi che negativi, è possibile formulare anche ipotesi monodirezionali. Se per H**1** si prevede una correlazione positiva, allora

H_1: $\rho > 0$ (**ipotesi monodirezionale destra**); se si prevede invece una correlazione negativa, allora **H_1: $\rho < 0$** (**ipotesi mono-direzionale sinistra**).

Per individuare **r** critico sulla tavola, bisogna considerare **N - 2**= 8 gradi di libertà. Poniamo inoltre $\alpha = 0.05$; H**1** è bidirezionale. Poiché sulla tavola (v. Ercolani & Areni) i valori di **r** indicati sono riferiti ad un'ipotesi monodirezionale, dobbiamo dividere α a metà, quindi $\alpha/2 = 0.025$. In corrispondenza di $\alpha = 0.025$ e g.d.l. = 8 troviamo **r** critico = **0.632** (se si calcola **r** con un programma di statistica non è necessario consultare la tavola perché di solito la probabilità associata ad **r** viene fornita dal programma).

Nel caso del primo esempio **r** = 0.88 supera il valore critico (**r** = 0.632) quindi possiamo rifiutare l'ipotesi nulla ed affermare (con una probabilità di errore del 5%) che nella popolazione esiste una correlazione significativa tra X e Y.
Nel caso del secondo esempio **r** = -0.19 non supera il valore critico, quindi non possiamo rifiutare l'ipotesi nulla.
Se si eleva al quadrato il coefficiente di correlazione **r**, si ottiene il **coefficiente di determinazione** (r^2) che, moltiplicato per 100, indica la percentuale di variabilità in comune tra le due variabili o, in altre parole, la percentuale di variabilità di Y dovuta alla variazione di X, e viceversa.
Nel caso del primo esempio, **r** = 0.88; $r^2 = 0.88^2 = 0.77$. La percentuale di variabilità in comune tra X e Y è uguale al **77%**.
Nel caso del secondo esempio, **r** = -0.19; $r^2 = -0.19^2 = 0.036$. La percentuale di variabilità in comune tra X e Y è uguale al **3.6%**.
Nel secondo caso, in cui si otteneva un coefficiente di correlazione basso (e non significativo), la percentuale di variabilità condivisa da X e Y è irrilevante, le due variabili variano quindi in modo indipendente l'una dall'altra.

12.b) Il coefficiente r_S di Spearman

Si utilizza quando **almeno una** delle due variabili da correlare è **su scala ordinale**, e quindi non è possibile utilizzare l'indice **r** di Pearson. Si tratta di un coefficiente di correlazione tra **ranghi**, perchè consente di verificare l'esistenza di una relazione tra **graduatorie**.

Esempio

Si somministra ad 8 bambini un test di abilità verbale (punteggi in termini di risposte corrette) e poi si chiede all'insegnante della classe di formare una graduatoria degli stessi bambini in base alla capacità di espressione verbale dimostrata in classe durante l'anno scolastico. Si ottengono i seguenti risultati:

Bambini:	A	B	C	D	E	F	G	H
Punteggi al test (X):	8	10	7	12	6	11	3	5
Graduat. Y (ranghi):	1°	2°	3°	4°	5°	6°	7°	8°

Si vuole verificare se c'è concordanza tra la prestazione al test ed il giudizio dell'insegnante.

Una delle due variabili è su scala ordinale (la graduatoria dell'insegnante), mentre l'altra è su scala a rapporti (i punteggi al test). Per poter calcolare il coefficiente r_S di Spearman è necessario trasformare in graduatoria anche i punteggi ottenuti al test, assegnando ad ogni bambino un ordine di rango (attribuendo cioè il rango 1 al punteggio più elevato, 2 al successivo e così via). La formula per il calcolo del coefficiente di correlazione per ranghi è la seguente:

$$r_s = 1 - \frac{6 \Sigma d_i^2}{N \times (N^2 - 1)}$$

Dove d_i è la differenza tra i ranghi per il soggetto i-esimo. Quindi:

Bambini:	A	B	C	D	E	F	G	H
Punteggi al test (X):	8	10	7	12	6	11	3	5
Graduat. Y (ranghi):	1°	2°	3°	4°	5°	6°	7°	8°
Graduat. X (ranghi):	4°	3°	5°	1°	6°	2°	8°	7°
d_i:	-3	-1	-2	3	-1	4	-1	1
d_i^2:	9	1	4	9	1	16	1	1

$\Sigma d_i^2 = 9 + 1 + 4 + 9 + 1 + 16 + 1 + 1 = \mathbf{42}$

$r_S = 1 - (6 \times 42) / 8 (8^2 - 1) = 1 - 252/504 = 1 - 0.5 = \mathbf{0.5}$

N.B. Nell'assegnare i ranghi ai valori di X bisogna tenere presente che se ci sono punteggi uguali, va assegnato il rango medio, inoltre quando sono presenti molti ranghi uguali è preferibile usare il coefficiente "tau di Kendall" (v. Areni, Ercolani, Scalisi "Introduzione all'uso della statistica in psicologia", pag. 95-97).

Per verificare la significatività di r_S si consulta la relativa tavola, considerando N - 2 gradi di libertà. Nel nostro caso N - 2 = 6.
Per H_0: $\rho = 0$; per H_1: $\rho > 0$
Ponendo $\alpha = 0.05$ (monodirezionale destra), r_S critico = 0.829.
Poiché r_S calcolato è inferiore a r_S critico, non possiamo rifiutare l'ipotesi nulla e concludiamo che i risultati del test di abilità verbale e la valutazione dell'insegnante non sono significativamente correlati. C'è da osservare che anche in questo caso, come in esempi precedenti, si pone un problema di **scarsa potenza del test**, dovuto alla bassa numerosità del campione. Infatti con 8 soggetti per rifiutare l'ipotesi nulla è necessario un valore di r_S molto elevato (0.829). Già con 16 soggetti (14 g.d.l.), ad esempio, il valore critico al 5% è 0.456. Se avessimo ottenuto r_S = 0.5 con un campione di 16 soggetti invece che di 8, avremmo potuto rifiutare l'ipotesi nulla.

Il coefficiente di correlazione r_{pbis} (punto biseriale)

Si utilizza quando una delle due variabili è su **scala a intervalli** o **rapporti equivalenti** e l'altra è una **variabile dicotomica**, cioè una variabile categoriale con solo due modalità (vero-falso, giusto-sbagliato, maschio-femmina, destra-sinistra, favorevole-contrario ecc.). Questo coefficiente è molto usato per calcolare la correlazione tra le risposte date da un gruppo di soggetti ad una domanda di un test (con due alternative di risposta) ed il punteggio ottenuto dagli stessi soggetti a tutto il test.

Esempio

Un test è composto da 50 domande con due alternative di risposta, di cui una corretta. Si vuole verificare la correlazione tra la risposta fornita alla domanda n. 9 (variabile X) ed il punteggio complessivo al test (variabile Y). Il test viene somministrato a 30 soggetti, ottenendo i seguenti risultati:
20 soggetti (N_g) hanno dato la risposta giusta alla domanda n. 9 ed hanno ottenuto una media ($\overline{Y}_g$) di 42 al test;
10 soggetti (N_s) hanno dato la risposta sbagliata alla domanda n. 9 ed hanno ottenuto una media ($\overline{Y}_s$) di 35 al test;
lo scarto quadratico medio di tutto il campione al test (s_y) è 6.5.
La formula di calcolo è la seguente:

$$r_{pbis} = \frac{\overline{Y}_g - \overline{Y}_s}{s_y} \times \sqrt{\frac{N_g \times N_s}{N(N-1)}} = \frac{42 - 35}{6.5} \times \sqrt{\frac{20 \times 10}{30(29)}} =$$

= 1.08 x 0.48 = **0.52**

Per la verifica della significatività del coefficiente punto-biseriale si utilizza la distribuzione campionaria del coefficiente **r** di Pearson con N - 2 gradi di libertà.
Per H_0: $\rho = 0$; per H_1: $\rho \neq 0$ (ipotesi bidirezionale).
Con N - 2 = 28 g.d.l. e $\alpha = 0.05$ (per H**1** bidirezionale dobbiamo considerare $\alpha / 2 = 0.025$), **r** critico = 0.361. Rifiutiamo l'ipotesi nulla: la domanda n. 9 risulta significativamente correlata al punteggio complessivo del test.

N.B. Se le risposte (giusta - sbagliata) alla domanda del test vengono codificate con 0 e 1 (e non, per esempio, con 1 e 2) si può utilizzare direttamente la formula del coefficiente **r** di Pearson (per un esempio v. Ercolani & Areni, "Statistica per la ricerca in psicologia").

Il coefficiente di correlazione r_{phi}

Si utilizza per calcolare la correlazione tra **due variabili dicotomiche** (cioè entrambe del tipo vero-falso, giusto-sbagliato ecc.).

Esempio

Si vuole verificare se esiste una correlazione tra due esami universitari (X e Y) di materia affine. Vengono intervistati 12 studenti che hanno sostenuto entrambi gli esami e viene loro chiesto se sono stati promossi o bocciati. Attribuendo la codifica B alla risposta "bocciato", e P alla risposta "promosso", si ottengono i seguenti risultati:

Sog.:	1	2	3	4	5	6	7	8	9	10	11	12
X:	P	P	B	P	B	B	P	B	P	P	B	B
Y:	P	P	P	B	B	P	P	B	B	P	B	P

Si costruisce quindi una tabella a doppia entrata 2 x 2 in cui si conta il numero di studenti promossi all'esame X e bocciati a Y, bocciati a X e promossi a Y, promossi ad entrambi e bocciati ad entrambi (le lettere identificano le celle della tabella):

	ESAME Y				
ESAME X	**Promosso**		**Bocciato**		Totali
Promosso	a)	**4**	b)	**2**	p **6**
Bocciato	c)	**3**	d)	**3**	q **6**
Totali	p'	**7**	q'	**5**	N **12**

La formula di calcolo è la seguente:

$$r_{phi} = \frac{f_a \times f_d - f_b \times f_c}{\sqrt{p \times p' \times q \times q'}} = \frac{4 \times 3 - 2 \times 3}{\sqrt{6 \times 7 \times 6 \times 5}} = \frac{6}{35.5} = 0.17$$

Per verificare la significatività di r_{phi} bisogna calcolare il Chi^2 sulla stessa tabella dei dati, se il Chi^2 è significativo, anche r_{phi} è significativo. In questo caso, con un campione così piccolo ed un coefficiente di correlazione così basso, è inutile applicare il test del Chi^2 (ricordiamo che quando un test è poco potente, la conclusione di non significatività è di scarso valore).