

4 La correlazione

[0.4]

4.1 Cos'è la correlazione

La correlazione è un indice che misura l'associazione fra due variabili, più in particolare, misura il grado in cui due variabili si “muovono assieme”. Esistono diversi indici di correlazione, applicabili a tipi diversi di variabili e a diversi livelli di misura. Prenderemo in considerazione la correlazione lineare prodotto-momento di Pearson, per capire il concetto di correlazione e vedremo quindi altri indici di correlazione.

Il concetto di correlazione è relativamente semplice, ma, da un punto di vista formale (ovvero matematico) ha molte relazioni con altre tecniche (come ad esempio la regressione lineare, i punti standard...). Il percorso che seguirò per spiegare questa tecnica statistica, è solo uno dei possibili, spero, il più semplice.

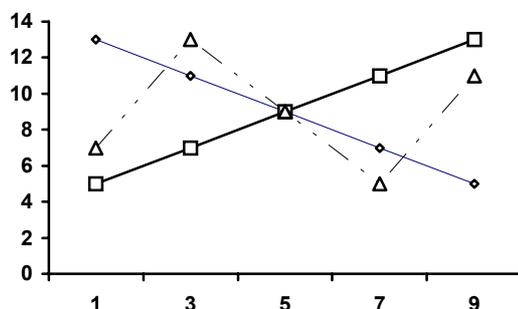
4.2 Correlazione lineare di Pearson

Immaginiamo di avere una serie di quattro variabili (del tutto fittizie, con valori scelti appositamente per evidenziare determinate relazioni), che chiameremo con le lettere finali dell'alfabetico, e i dati di alcuni soggetti (che chiameremo con le prime cinque lettere dell'alfabeto).

Tabella 4.1

	X	Y	Z	W
A	1	5	13	7
B	3	7	11	13
C	5	9	9	9
D	7	11	7	5
E	9	13	5	11
\bar{X}	5	9	9	9
s	2.828	2.828	2.828	2.828

Queste variabili sono state scelte in modo da avere uguale deviazione standard e media (per lo meno le variabili Y, Z e W). Come possiamo vedere, a piccoli valori di X, corrispondono piccoli valori di Y e grandi valori di Z, a valori grandi di X, corrispondono grandi valori di Y e piccoli valori di Z; non sembra esistere una vera relazione fra X e W. Possiamo rappresentare graficamente queste relazioni, in un grafico.



Il grafico evidenzia bene come:

- la relazione fra X e Y è una relazione lineare crescente;
- la relazione fra X e Z è lineare decrescente;
- la relazione fra X e W non è riconducibile ad una regola.

Se riscriviamo la Tabella 4.1 in modo da usare gli scarti dalla media (ovvero $x - \bar{X}$), possiamo notare qualcosa di ancora più significativo:

Tabella 4.2

	X	Y	Z	W
A	-4	-4	4	-2
B	-2	-2	2	4
C	0	0	0	0
D	2	2	-2	-4
E	4	4	-4	2

Quello che possiamo notare è che la relazione lineare crescente fra X e Y è caratterizzata dal fatto che tutti gli scarti dalla media hanno lo stesso segno, la relazione inversa fra X e Z corrisponde a scarti che hanno segno opposto, mentre la relazione non definita fra X e W ha scarti i cui segni si associano “casualmente”.

Con questi dati possiamo tentare di costruire una statistica, che chiameremo indice di correlazione lineare. In teoria, questo indice, dovrebbe avere un valore positivo per indicare relazioni lineari positive (come quella fra X e Y), un valore negativo per relazioni lineari negative o inverse (X e Z) e un valore nullo per relazioni inesistenti o nulle (X e W). Inoltre dovremmo cercare di standardizzare l'indice affinché oscilli sempre fra valori predefiniti, qualunque siano i numeri che costituiscono le variabili. Una possibilità è quella che oscilli fra -1 e $+1$.

Un primo passo potrebbe essere quello di moltiplicare i valori delle variabili che vogliamo mettere in relazione e poi di sommare questi valori:

Tabella 4.3

XY	XZ	XW
16	-16	8
4	-4	-8
0	0	0
4	-4	-8
16	-16	8
40	-40	0

Se, a questo punto, dividiamo i totali per la numerosità, otteniamo qualcosa che assomiglia alla formula della varianza e che chiameremo covarianza:

$$\text{cov} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N}$$

E se dividiamo la covarianza per il prodotto delle deviazioni standard, otteniamo un valore standardizzato, che oscilla fra -1 e $+1$. Questa è una delle formule che esprime la *correlazione di Pearson*.

$$r = \frac{\text{cov}}{s_x s_y}$$

Tabella 4.4

	XY	XZ	XW
Cov	40	-40	0
Cov / n	8	8	0
$s_x s_y$	8	8	8

r	1	-1	0
-----	-----	------	-----

4.3 Formule alternative

Una formula alternativa per la correlazione di Pearson è facilmente derivabile dalla precedente, se consideriamo che nella formula della covarianza abbiamo le somme degli scarti dalla media e che queste vengono poi divise per le deviazioni standard. La formula (la più facile da ricordare) esprime la correlazione come media dei prodotti dei punti z (cfr. la dimostrazione 6.1.1 in Appendice):

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

Una seconda formula alternativa, è:

$$r = \frac{\frac{\sum xy}{N} - \bar{X}\bar{Y}}{s_x s_y}$$

Una terza formula alternativa (generalmente usata per i calcoli, anche se è più complessa da ricordare), utilizza solo i dati grezzi (cfr. la dimostrazione 6.1.2 in Appendice) e può esprimersi in due modi leggermente diversi:

$$r = \frac{\frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}}{\sqrt{[N\sum XY - \sum X \sum Y] \sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}}$$

E il calcolo finale della correlazione fra x e y , secondo le due formule, risulta:

Tabella 4.5

X	Y	X^2	Y^2	XY
1	5	1	25	5
3	7	9	49	21
5	9	25	81	45
7	11	49	121	77
9	13	81	169	117
25	45	165	445	265

$$r = \frac{265 - \frac{25 \cdot 45}{5}}{\sqrt{(165 - \frac{(25)^2}{5})(445 - \frac{(45)^2}{5})}} = \frac{265 - \frac{1125}{5}}{\sqrt{(165 - \frac{625}{5})(445 - \frac{2025}{5})}}$$

$$= \frac{265 - 225}{\sqrt{(165 - 125)(445 - 405)}} = \frac{40}{\sqrt{40 \cdot 40}} = \frac{40}{40} = 1$$

$$r = \frac{5 \cdot 265 - 25 \cdot 45}{\sqrt{[5 \cdot 165 - (25)^2][5 \cdot 445 - (45)^2]}} = \frac{1235 - 1125}{\sqrt{(825 - 625)(2225 - 2025)}}$$

$$= \frac{200}{\sqrt{200 \cdot 200}} = \frac{200}{200} = 1$$

4.4 Interpretazione

Non vi è un criterio matematico o statistico per interpretare la forza della relazione fra le due variabili. La prassi ha stabilito una serie di convenzioni:

Tabella 4.6

<i>Valore di r</i>	<i>Correlazione</i>	<i>Relazione</i>
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

Una particolare attenzione va posta nell'interpretare il significato stesso di correlazione.

Innanzitutto è necessario ricordare che la formula, generalmente utilizzata (quella di Pearson), è relativa ad una relazione lineare e che quindi tutte le forme diverse di relazione, possono produrre risultati anomali. Consideriamo i due grafici della figura seguente. Entrambi rappresentano dati che si distribuiscono in modo curvilineo; tuttavia, nel primo esempio, vi sono dati sufficienti per portare ad una correlazione lineare di .54 (una correlazione considerata considerevole).

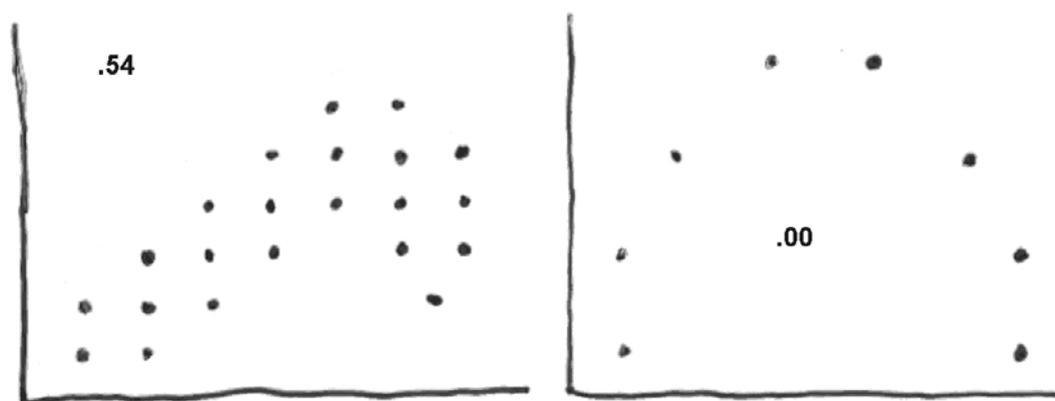


Figura 4.1 – Esempi di relazioni curvilinee e corrispondenti valori di r

Un secondo problema riguarda il rapporto di causalità dell'indice di correlazione, ci dice solo e soltanto che le due variabili hanno un andamento comune. Non ci dice mai che una variabile è la causa e che l'altra è l'effetto.

Un terzo problema è legato al fatto che, qualsiasi siano le variabili utilizzate, sempre si ottiene un qualunque valore di r . Tuttavia, non sempre questa correlazione avrà un significato logico. Ad esempio, io posso correlare il mio consumo giornaliero di acqua con il numero di barattoli di yogurt venduti giornalmente da un certo negozio; troverò certamente un valore di r che potrebbe anche essere diverso da zero. Sarà comunque una correlazione priva di significato logico. Questo tipo di correlazione è chiamata "casuale" o "spuria".

4.5 Mi posso fidare?

Un quarto problema dipende appunto dal fatto che ottengo sempre e comunque un valore di correlazione. Consideriamo che, quando calcoliamo la correlazione fra due va-

riabili, stiamo lavorando su un campione che è stato estratto casualmente da una popolazione. Potrei imbartermi in due situazioni opposte (illustrate in Figura 4.2):

1. il campione su cui calcolo la correlazione presenta, casualmente, una relazione lineare (i puntini cerchiati), mentre la popolazione da cui l'ho estratto non lo è;
2. il campione su cui calcolo la correlazione non ha una struttura lineare, mentre la popolazione sì.

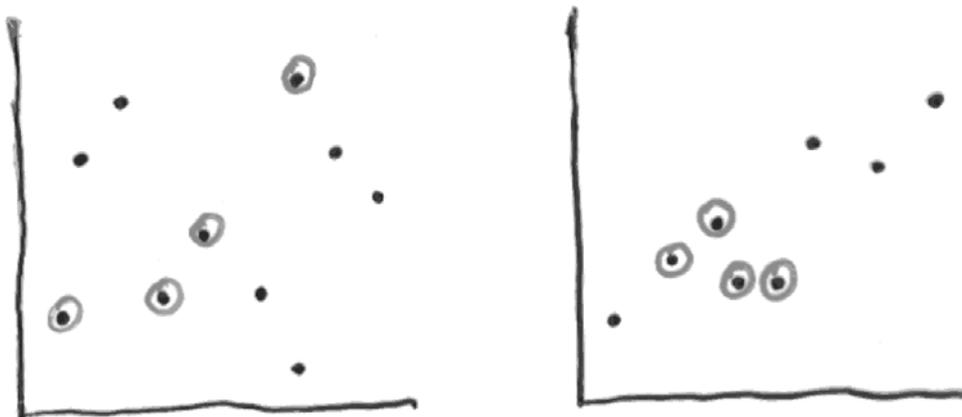


Figura 4.2

Si tratta a questo punto di attivare un processo di inferenza per sapere quanto possiamo fidarci della correlazione calcolata. Il ragionamento che sottostà all'inferenza è il seguente ed è comune a quello utilizzato con altre statistiche.

Ipotizziamo che la correlazione sia tratta da una popolazione che abbia correlazione nulla:

$$H_0: \rho = 0$$

Ovviamente formuliamo anche l'ipotesi alternativa:

$$H_1: \rho \neq 0$$

Come il solito, ragioniamo sulla base dell'ipotesi nulla.

Ipotizziamo una popolazione in cui la correlazione fra X e Y sia conosciuta e sia pari a 0 (o meglio $\rho = 0$). Estraggo un campione di ampiezza n e calcolo la correlazione fra le due variabili x e y . Estraggo un altro campione e, di nuovo, calcolo la correlazione. Ripeto il procedimento per un numero infinito di volte. Tutti i possibili valori di r calcolati su tutti i possibili campioni estratti da questa popolazione vanno a formare una distribuzione campionaria delle correlazioni, che avrà una sua propria media e una sua propria deviazione standard che tenderà ad approssimarsi alla distribuzione normale all'aumentare del valore di unità statistiche che uso per il calcolo (ovvero n):

Figura 4.3

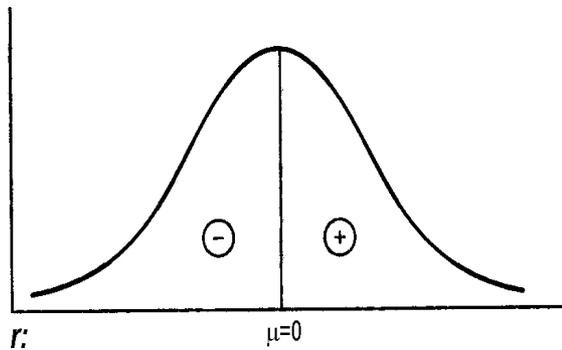


Tabella 4.7

$$\mu_r = 0$$

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

E' ovvio aspettarsi che, sebbene il valore zero (corrispondente alla media) sia il più frequente, sia possibile ottenere anche correlazioni diverse da zero, sia negative sia positive. Tanto più le correlazioni saranno lontane da zero (cioè grandi, in valore assoluto) tanto meno saranno frequenti e probabili (sempre nell'ipotesi che siano calcolate su campioni provenienti da una popolazione con correlazione nulla).

Per sapere se la probabilità della nostra correlazione r , calcolata su un singolo campione, è sufficiente calcolare il punto z corrispondente a r :

$$r_t = \frac{r - \rho}{s_r}$$

r_t si distribuisce secondo la *distribuzione t di Student* con $n-2$ gradi di libertà.

Senza bisogno di applicare le formule, si possono consultare delle apposite tavole, che forniscono la probabilità associata ad un certo valore di r , per un dato grado di libertà.

Tabella 4.8- Valori critici di correlazione di Pearson (estratto)

	.05	.025	.010	.005
<i>Mono-dir.</i>				
<i>Bi-dir.</i>	.10	.05	.025	.01
$d=1$.988	.997	.9995	.999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874

Usando le tavole, dobbiamo seguire il seguente procedimento:

- 1- fissiamo un livello α (di solito $\alpha = .05$)
- 2- stabiliamo l'ipotesi alternativa come monodirezionale o bidirezionale (quest'ultima è la scelta più comunemente usata)
- 3- calcoliamo i gradi di libertà ($d=n-2$)
- 4- troviamo sulle tavole la riga corrispondente ai gradi di libertà e la scorriamo fino alla colonna corrispondente al livello α

- 5- all'incrocio fra riga e colonna, troviamo il valore critico di r (r_c)
- 6- se la nostra correlazione è inferiore al valore critico, accettiamo l'ipotesi H_0
- 7- se è superiore, accettiamo l'ipotesi alternativa

Tabella 4.9

$r < r_c$	accetto H_0
$r \geq r_c$	accetto H_1

Accettare l'ipotesi nulla significa che, qualunque sia il valore di r che abbiamo trovato nel campione, esso è comunque pari a 0, poiché viene casualmente da una popolazione che ha correlazione zero. Non dobbiamo cercare di interpretare questa correlazione, perché è un errore di estrazione casuale del campione.

Accettare l'ipotesi alternativa, significa che il valore calcolato è effettivo poiché viene da un campione casualmente estratto da una popolazione che ha correlazione diversa da zero. A questo punto possiamo cercare di interpretare la correlazione trovata.

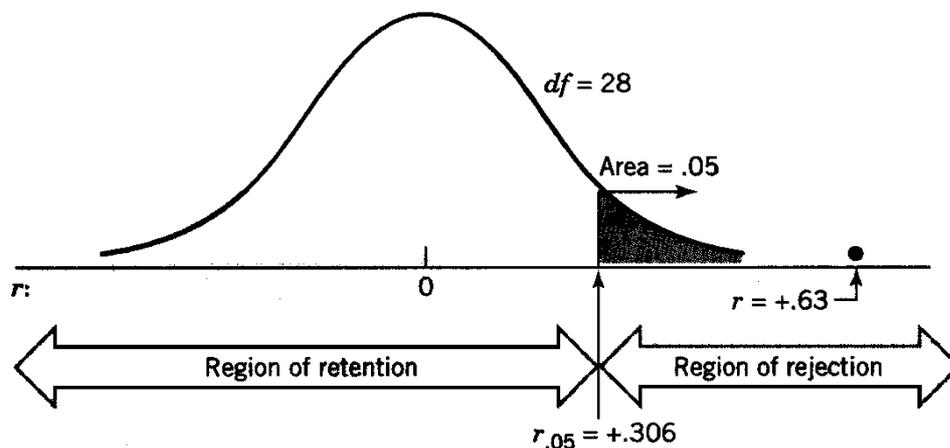
Esempi:

$r = .63$; $N = 30$; $gl = 28$; $\alpha = .05$; ipotesi monodirezionale

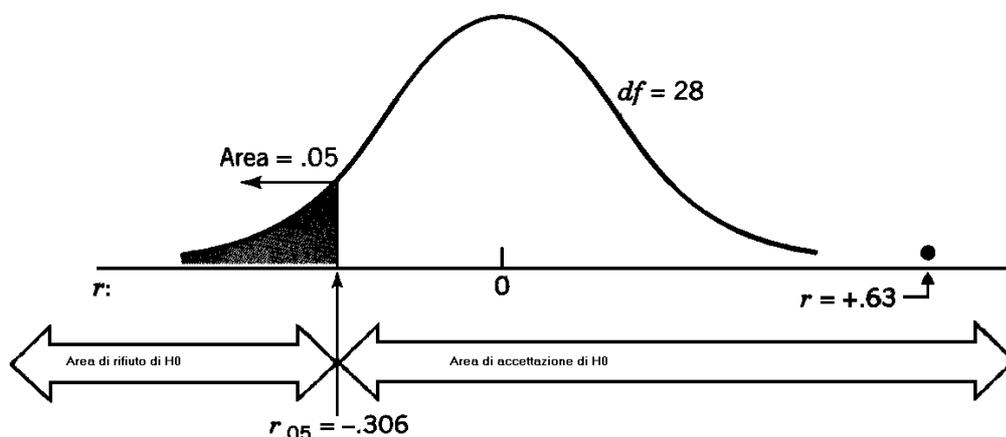
Consultando le tavole, trovo un r critico di .306

Poiché .63 è maggiore del valore critico, rifiuto H_0 e accetto H_1

Fare un'ipotesi monodirezionale equivale a dire che noi ci aspettiamo che la correlazione sia non solo diversa da zero, ma anche positiva o negativa. In tal caso, l'area di probabilità corrispondente ad alfa, che ci indicherà l'area di accettazione o di rifiuto dell'ipotesi nulla, sarà posta su un solo lato della curva (cfr.).



Nell'ipotesi bidirezionale, ipotizziamo soltanto che la correlazione sia diversa da zero, senza fare alcuna ipotesi sul suo andamento. In tal caso l'area di rifiuto dell'ipotesi nulla dovrà essere suddivisa fra i valori positivi e quelli negativi (cfr.).



Con i dati della Tabella 4.10, calcoliamo la correlazione lineare di Pearson e la sua significatività; per semplicità di calcolo, usiamo le formule alternative.

Le variabili x e y rappresentano rispettivamente l'“Abilità di comprensione di un testo scritto” e il “Quoziente di intelligenza”.

Tabella 4.10 – Esempio di correlazione lineare

x	y	x^2	y^2	xy
46	126	2116	15876	5796
49	110	2401	12100	5390
48	103	2304	10609	4944
42	128	1764	16384	5376
46	111	2116	12321	5106
49	128	2401	16384	6272
43	104	1849	10816	4472
45	101	2025	10201	4545
49	111	2401	12321	5439
42	125	1764	15625	5250
40	113	1600	12769	4520
45	115	2025	13225	5175
48	100	2304	10000	4800
41	124	1681	15376	5084
43	101	1849	10201	4343
40	102	1600	10404	4080
47	129	2209	16641	6063
48	112	2304	12544	5376
48	128	2304	16384	6144
46	123	2116	15129	5658
905	2294	41133	265310	103833

Applicando la seconda formula alternativa, risulterà:

$$\frac{20 \cdot 103833 - 905 \cdot 2294}{\sqrt{(20 \cdot 41133 - 905^2)(20 \cdot 265310 - 2294^2)}} =$$

$$\frac{2076660 - 2076070}{\sqrt{(822660 - 819025)(5306200 - 5262436)}} =$$

$$\frac{590}{\sqrt{3635 \cdot 43764}} = \frac{590}{\sqrt{159082140}} = \frac{590}{12612.77685524} = 0.0467$$

che è un valore troppo basso per essere significativo. Quindi possiamo tranquillamente considerarlo come proveniente da una popolazione che ha correlazione nulla.

Se però volessimo fare una verifica, dovremmo cercare sulle tavole il valore critico di r per $20-2=18$ gradi di libertà. Il valore dipenderà anche dal livello α che decidiamo di adottare e dal tipo di ipotesi alternativa: bidirezionale o mono-direzionale. Non avendo pre-conoscenze che ci portino ad esprimere un'ipotesi mono-direzionale, utilizziamo una un'ipotesi alternativa del tipo:

$$H_1: \rho \neq 0$$

Per 18 gradi di libertà, i valori critici di r sono:

α	.05	.01
r_c	.444	.561

Tutti i valori critici sono decisamente superiori al valore di r da noi calcolato che quindi non è significativo.

Attenzione: la correlazione si calcola sempre sui dati grezzi e mai usando la distribuzione di frequenza.

In effetti, se usassimo la tabella delle frequenze, potremmo imbarcarci in alcuni problemi:

+ se le due variabili non hanno lo stesso numero di categorie, come si può procedere al calcolo? Ovviamente è impossibile!

+ dal momento che le categorie in una tabella delle frequenze, sono state poste in ordine (crescente o decrescente), la correlazione sarebbe necessariamente sempre positiva.

4.6 La correlazione di Spearman

Vi sarete accorti che, negli esempi, abbiamo usato variabili che sono misurate a livello di scale intervallo o a rapporto. In effetti, la correlazione lineare di Pearson si applica su variabili quantitative. La stessa considerazione poteva scaturire considerando che, in una delle formule, abbiamo usato i punti z e la deviazione standard, statistiche che hanno significato se calcolate a livello di scale intervallo o a rapporto.

Esistono molte altre formule che stimano l'associazione di due variabili su scale diverse da quelle quantitative. Ad es. la correlazione a ranghi di Spearman (chiamata anche ρ_s) utilizza dati misurati a livello di scala ordinale.

Poiché la scala ordinale non permette calcoli, dobbiamo prima di tutto trasformare la variabile in qualcosa di metrico e di lineare. La trasformazione implicata dalla ρ_s di Spearman è l'ordinamento a ranghi (in inglese, *ranking*). I valori della variabile vengono dapprima ordinati fra di loro in modo crescente, quindi si procede ad associare a ciascun valore il rango che gli compete. Al valore più basso il rango 1, a quello immediatamente successivo il rango 2 e così via. Per mantenere la stessa struttura, a valori uguali dovremo associare ranghi uguali e quindi utilizzeremo la media dei ranghi. Facciamo un esempio: consideriamo le seguenti due variabili (x e y), di tipo ordinale.

X	Y
A	3
B	3
A	1

D	2
C	3
B	2

Per ciascuna, dobbiamo, per prima cosa, riordinare i valori:

A	A	B	B	C	D
1	2	2	3	3	3

e quindi assegnare i ranghi:

	<i>A</i>	<i>A</i>	<i>B</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>Pos.</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
	$\frac{1+2}{2}$	$\frac{1+2}{2}$	$\frac{3+4}{2}$	$\frac{3+4}{2}$	<i>5</i>	<i>6</i>
<i>rango</i>	<i>1,5</i>	<i>1,5</i>	<i>3,5</i>	<i>3,5</i>	<i>5</i>	<i>6</i>
	<i>1</i>	<i>2</i>	<i>2</i>	<i>3</i>	<i>3</i>	<i>3</i>
<i>Pos.</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
		$\frac{2+3}{2}$	$\frac{2+3}{2}$	$\frac{4+5+6}{3}$	$\frac{4+5+6}{3}$	$\frac{4+5+6}{3}$
		<i>2</i>	<i>2</i>	<i>3</i>	<i>3</i>	<i>3</i>
<i>rango</i>	<i>1</i>	<i>2,5</i>	<i>2,5</i>	<i>5</i>	<i>5</i>	<i>5</i>

Infine usare i singoli ranghi al posto dei valori.

La trasformazione in ranghi ha prodotto una nuova variabile che “quantifica” la posizione dei singoli valori. A questo punto, usando i ranghi possiamo applicare la formula di Spearman per il calcolo dell’associazione, che fa uso delle differenze fra i ranghi:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

dove d è la differenza fra i ranghi e n è la numerosità.

X	<i>X rango</i>	Y	<i>Y rango</i>	<i>d</i>	<i>d</i> ²
A	<i>1.5</i>	<i>3</i>	<i>5</i>	<i>-3,50</i>	<i>12,25</i>
B	<i>3.5</i>	<i>3</i>	<i>5</i>	<i>-1,50</i>	<i>2,25</i>
A	<i>1.5</i>	<i>1</i>	<i>1</i>	<i>0,50</i>	<i>0,25</i>
D	<i>5</i>	<i>2</i>	<i>2.5</i>	<i>2,50</i>	<i>6,25</i>
C	<i>6</i>	<i>3</i>	<i>5</i>	<i>1,00</i>	<i>1,00</i>
B	<i>3.5</i>	<i>2</i>	<i>2.5</i>	<i>1,00</i>	<i>1,00</i>

$$r_s = 1 - \frac{6 \cdot 23}{6(6^2 - 1)} = 1 - \frac{138}{6 \cdot 35} = 1 - \frac{138}{210} = 1 - 0,657143 = 0,342857$$

Anche la correlazione di Spearman produce valori che oscillano fra -1 e $+1$ e anche per questa correlazione esistono dei test statistici per verificare se la correlazione calcolata è stata casualmente estratta da una popolazione con correlazione nulla. Anche le ipotesi relative all’inferenza sono analoghe:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

I test statistici differiscono secondo la numerosità. Per $n < 30$, è possibile calcolare direttamente la probabilità associati ai valori r_s , per valori di $n > 30$, la distribuzione di r_s viene trasformata tramite la formula

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

che si approssima alla distribuzione t di Student con $n-2$ gradi di libertà.

La correlazione di Spearman, può essere usata con variabili a intervallo (previa trasformazione in ranghi) quando la numerosità del campione è inferiore a 30.

4.6.1 Tavola dei valori critici di rho

rho= coefficiente di correlazione a ranghi di Spearman

g.l = $n-2$

	Mono-	.05	.025	.01	.005
Bi-	.10	.05	.02	.01	
4	1.000				
5	.900	1.000	1.000		
6	.829	.886	.943	1.000	
7	.714	.786	.893	.929	
8	.643	.738	.833	.881	
9	.600	.683	.783	.833	
10	.564	.648	.746	.794	
12	.506	.591	.712	.777	
14	.456	.544	.645	.715	
16	.425	.506	.601	.665	
18	.399	.475	.564	.625	
20	.377	.450	.534	.591	
22	.359	.428	.508	.562	
24	.343	.409	.485	.537	
26	.329	.392	.465	.515	
28	.317	.377	.448	.496	
30	.306	.364	.432	.478	

4.7 Altri tipi di correlazione

Solo a livello informativo, diremo che esistono svariate altre formule di calcolo per stimare il grado di associazione fra due variabili. Alcune sono di tipo lineare, mentre altre presuppongono relazioni di tipo non lineare.

la maggior parte di queste formule sono state costruite in modo da oscillare fra -1 e $+1$, per poter rendere comprensibili e confrontabili i risultati ottenuti.