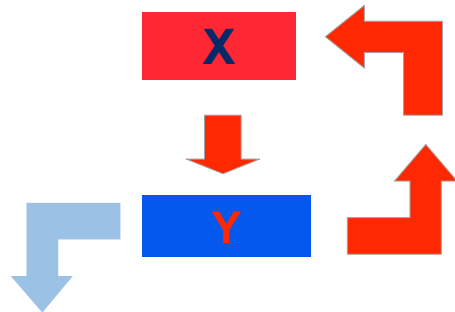


Studio delle relazioni statistiche (bivariate)



Problema_1: è possibile sapere che succede alla “Y” se varia la “X” (in modo spontaneo o indotto)?

Problema_2: si ritiene ci sia un legame tra la Y e la X. E' possibile dimostrare il contrario?

Dalla matrice dei dati alla tabella doppia

Operaio	Importo	Livello	Operaio	Importo	Livello	Operaio	Importo	Livello
1	133754	A	41	139537	B	81	156488	A
2	177321	D	42	196198	C	82	191405	A
3	198093	B	43	183375	B	83	117894	F
4	198951	F	44	148518	F	84	161926	A
5	128050	A	45	126191	B	85	102978	B
6	107152	B	46	148488	C	86	171470	A
7	168502	B	47	129230	B	87	131906	A
8	185872	C	48	193780	F	88	179658	C
9	174107	A	49	141154	B	89	146534	A
10	127670	F	50	100256	B	90	137011	B
11	171307	B	51	140573	A	91	112452	D
12	135016	A	52	191271	A	92	117509	A
13	116721	B	53	194093	B	93	185801	C
14	138590	E	54	109994	B	94	172984	A
15	122672	C	55	177444	A	95	103235	B
16	191676	D	56	100239	F	96	195622	B
17	174958	B	57	176015	B	97	127726	D
18	187423	D	58	170692	C	98	121094	A
19	111110	C	59	167677	E	99	193272	B
20	136503	E	60	199348	E	100	148265	B
21	120768	C	61	123781	B			
22	191648	D	62	179708	D			
23	101570	D	63	139825	A			
24	145044	A	64	148948	C			
25	102990	F	65	146901	D			
26	187028	E	66	136471	D			
27	124437	D	67	104697	A			
28	122079	C	68	152657	E			
29	163468	E	69	170503	B			
30	140935	A	70	135280	D			
31	146843	A	71	107743	B			
32	172497	C	72	171517	D			
33	122209	D	73	193946	C			
34	135783	D	74	170884	A			
35	150789	C	75	181407	B			
36	121587	A	76	124571	E			
37	133415	D	77	139906	A			
38	194731	F	78	142344	A			
39	176619	B	79	190776	A			
40	104960	A	80	141811	B			

Su n=100 operai è stato rilevato l'importo dello straordinario settimanale e la classe stipendiale.

In questa forma i dati non sono leggibili;

Organizziamo gli importi in classi:

Excel: Tabella pivot

Count of Operaio	Livello						
Imp.MGL	A	B	C	D	E	F	Grand Total
<120	3	7	1	2	0	3	16
120-140	8	5	3	7	3	1	27
140-160	7	3	3	1	1	1	16
>160	9	12	7	6	4	3	41
Grand Total	27	27	14	16	8	8	100

La tabella rivela che il 41% si colloca nella 4ª classe; che il 12% si trova nella combinazione (4,B) e che il livello “A” fa più straordinari (27%) rispetto a tutti gli altri.

Dipendenza statistica

Riguarda l'analisi delle relazioni tra due VARIABILI

Se le variabili sono entrambi quantitative lo studio dà origine all'analisi della **CORRELAZIONE**.

Se, sono anche incluse variabili qualitative o trattate come tali, occorrono metodi che prescindano da considerazioni metriche e applicheremo l'analisi della **CONNESSIONE** o **ASSOCIAZIONE**

Si parlerà di dipendenza statistica, diversa dalla dipendenza matematica, facendo riferimento a modifiche più o meno rilevanti della distribuzione di una variabile se l'altra subisce delle variazioni.

Trattazione generale

Partiamo dalla variabile doppia: $(X_i, Y_i); i = 1, 2, \dots, n$

Supponiamo che siano state organizzate in una tabella con “r” modalità distinte per la variabile sulle righe (X) e “c” modalità per la variabile sulle colonne (Y)

Dove:

$$\begin{array}{c|cccc|c}
 & Y_1 & Y_2 & \dots & Y_c & \\
 \hline
 X_1 & n_{11} & n_{12} & & n_{1c} & n_{1.} \\
 X_2 & n_{21} & n_{22} & & n_{2c} & n_{2.} \\
 & & & & & \\
 X_r & n_{r1} & n_{r2} & & n_{rc} & n_{r.} \\
 \hline
 & n_{.1} & n_{.2} & \dots & n_{.c} & n
 \end{array}$$

$$n_{i.} = \sum_{j=1}^c n_{ij} = n_{i1} + n_{i2} + \dots + n_{ic} = \text{totale di riga}$$

$$n_{.j} = \sum_{i=1}^r n_{ij} = n_{j1} + n_{j2} + \dots + n_{rj} = \text{totale di colonna}$$

$$n = \sum_{i=1}^r n_{i.} = \sum_{j=1}^c n_{.j} = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$$

il punto indica l'indice rispetto a cui si è sommato

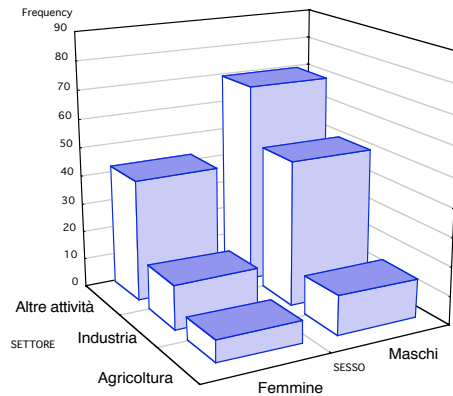
Esempio

Occupati per settori di attività economica (media annua). Dati in migliaia

Settori	Sesso		Totale
	Maschi	Femmine	
Agricoltura	1.485	812	2.297
Industria	5.270	1.626	6.896
Terziario	7.232	4.318	11.550
Totale	13.987	6.756	20.743

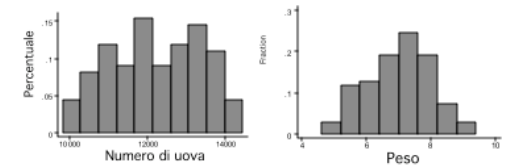
$r=3$; $c=2$; $n=20'743$

La diversa struttura delle due componenti è evidente dal grafico

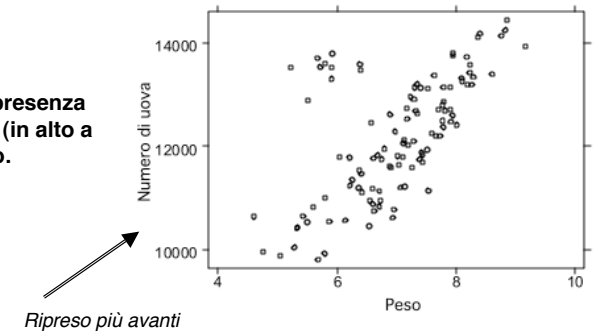


Effetti della multidimensionalità

La presentazione congiunta delle due variabili rivela aspetti che rimangono oscurati nella rappresentazione separata dei due aspetti.



Lo scatterplot indica la presenza di un gruppo di soggetti (in alto a sinistra) diversi dal resto.



Distribuzione congiunta di due variabili

Anche nella tabella doppia possiamo usare le frequenze relative:

	Y_1	Y_2	...	Y_c		$0 \leq f_{ij} \leq 1$
X_1	f_{11}	f_{12}		f_{1c}	$f_{1.}$	$f_{i.} = \sum_{j=1}^c f_{ij}$
X_2	f_{21}	f_{22}		f_{2c}	$f_{2.}$	$f_{.j} = \sum_{i=1}^r f_{ij}$
$\vdots$						
X_r	f_{r1}	f_{r2}		f_{rc}	$f_{r.}$	$\sum_{i=1}^r \sum_{j=1}^c f_{ij} = 1$
	$f_{.1}$	$f_{.2}$	...	$f_{.c}$	1	

Le f_{ij} sono le frequenze relative congiunte;

Le " $f_{i.}$ " e le " $f_{.j}$ " sono le frequenze relative marginali.

L'insieme delle coppie (X_i, Y_j) e delle rispettive frequenze relative f_{ij} costituisce la distribuzione congiunta delle variabili X ed Y .

Essa associa ad ogni combinazione di modalità (X_i, Y_j) un numero in $(0,1)$ e la cui somma è pari ad uno

Distribuzioni marginali

A partire dalla distribuzione congiunta si definiscono le distribuzioni per ciascuna delle variabili a prescindere dall'altra

$$f(X = x_i) = \sum_{j=1}^c f(X = x_i, Y = y_j) = \sum_{j=1}^c f_{ij} = f_{i.}; \quad i = 1, 2, \dots, r$$

$$f(Y = y_j) = \sum_{i=1}^r f(X = x_i, Y = y_j) = \sum_{i=1}^r f_{ij} = f_{.j}; \quad j = 1, 2, \dots, c$$

Per ottenere la distribuzione marginale si somma rispetto alla variabile che NON interessa

Distribuzioni condizionate

Per studiare il comportamento della "Y" rispetto alla "X" dividiamo la distribuzione congiunta in tante sottodistribuzioni

La distribuzione della Y CONDIZIONATA (PARZIALE) dal fatto che "X" è ad un dato livello è

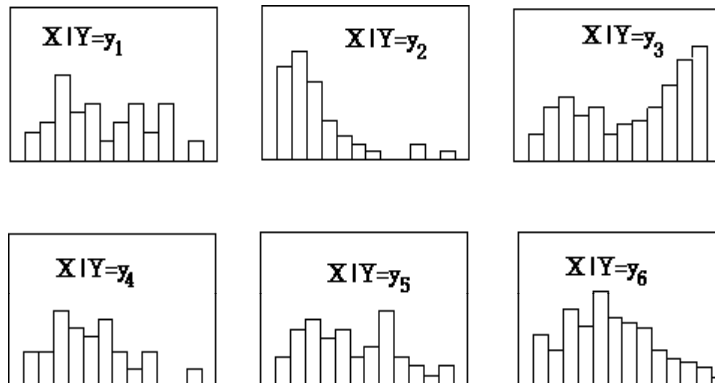
$$f(Y = y_j | X = x_i) = \frac{f(X = x_i, Y = y_j)}{f(X = x_i)}; \quad j = 1, 2, \dots, c$$

cioè un riscalamento pro-quota delle righe per assicurare la somma unitaria

Analogamente, la distribuzione della X dato che Y è ad un livello prefissato è:

$$f(X = x_i | Y = y_j) = \frac{f(X = x_i, Y = y_j)}{f(Y = y_j)}; \quad i = 1, 2, \dots, r$$

Multiplot



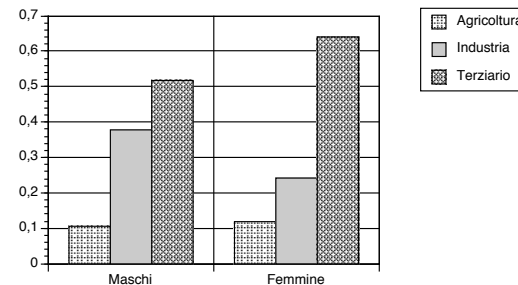
Per ogni modalità della Y è rappresentato il corrispondente l'istogramma della X CONDIZIONATO ai vari valori della Y

Ovviamente il ruolo delle variabili può essere scambiato.

Esempio

Distribuzione congiunta

Settori	Sesso		Totale
	Maschi	Femmine	
Agricoltura	7,16%	3,91%	11,07%
Industria	25,41%	7,84%	33,24%
Terziario	34,86%	20,82%	55,68%
Totale	67,43%	32,57%	100,00%



Distribuzione condizionale
Donne

Settori	Femmine
Agricoltura	12,02%
Industria	24,07%
Terziario	63,91%
Totale	100,00%

Distribuzione condizionale
maschi

Settori	Maschi
Agricoltura	10,62%
Industria	37,68%
Terziario	51,71%
Totale	100,00%

Indipendenza di variabili

Perché abbia senso lo studio CONGIUNTO esso deve essere più informativo dello studio SEPARATO delle due componenti

Se la "X" assume valori in relazione ad eventi indipendenti da quelli che generano i valori della "Y" non esiste alcun legame statistico interessante

ESEMPIO

Lancio di due dadi di diverso colore

X: punteggio del dado rosso;

Y: punteggio del dado blu;



Sapere che lanciando i due dadi, X= 4 e, contemporaneamente, Y= 3 è come sapere che X=4 (ignorando "Y") e che Y=3 (ignorando "X")

Indipendenza in distribuzione

Se la condizionata di YIX non cambia al variare di X allora Y è **INDIPENDENTE IN DISTRIBUZIONE** da X.

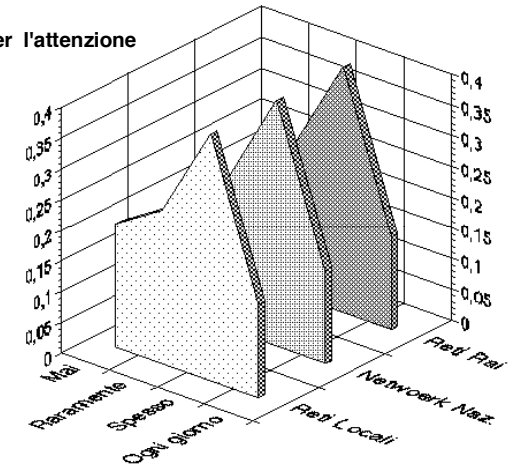
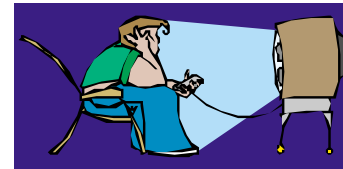
$$f(X = x_i | Y = y_j) = f(X_i); \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c$$

L'indipendenza è una relazione simmetrica: Se X è indipendente da Y anche Y è indipendente da X

$$f(Y = y_j | X = x_i) = f(Y_j) \quad j = 1, 2, \dots, s; \quad i = 1, 2, \dots, r$$

Rappresentazione grafica

Campione di famiglie classificato per l'attenzione ai programmi televisivi



Indipendenza significa che si guardano con la stessa frequenza tutti i network ovvero la frequenza con cui si guarda la TV prescinde dal network

Esempio

Reddito familiare e rendimento scolastico

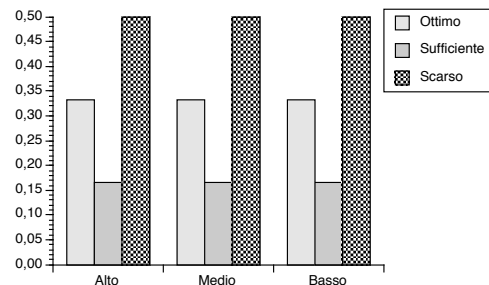
Rendimento	Alto	Medio	Basso	Totale
Ottimo	16	32	40	88
Sufficiente	8	16	20	44
Scarso	24	48	60	132
Totale	48	96	120	264

	Reddito familiare			
Rendimento	Alto	Medio	Basso	Totale
Ottimo	0.3333	0.3333	0.3333	0.3333
Sufficiente	0.1667	0.1667	0.1667	0.1667
Scarso	0.5000	0.5000	0.5000	0.5000
Totale	1.0000	1.0000	1.0000	1.0000

Le frequenze assolute sono diverse, ma quelle relative coincidono per ogni distribuzione condizionata del rendimento.

Verifica:

$$40 = \frac{88 * 120}{264}; \quad 16 = \frac{44 * 96}{264}$$



Conseguenza della definizione

Se fra le due variabili c'è indipendenza, la frequenza congiunta è pari al prodotto delle frequenze marginali diviso per il totale frequenze:

$$f(X = x_i | Y = y_j) = f(X_i) \Rightarrow \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} \Rightarrow n_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

$$= \frac{\left(\frac{n_{i.}}{n}\right) \left(\frac{n_{.j}}{n}\right)}{1} = f_{i.} * f_{.j}$$

Questa relazione costituisce una definizione alternativa della relazione di indipendenza

Solo in caso di **indipendenza statistica** la frequenza congiunta è ricavabile dalla conoscenza delle frequenze marginali (è pari al loro prodotto)

Esempio

Verificare se fra Y ed X c'è indipendenza

	Y1	Y2	Y3	
X1	3	9	18	30
X2	2	6	12	20
X3	5	15	30	50
	10	30	60	100

$$n_{11} = \frac{n_{1.} * n_{.1}}{n} = \frac{30 * 10}{100} = 3$$

$$n_{21} = \frac{n_{2.} * n_{.1}}{n} = \frac{20 * 10}{100} = 2$$

$$n_{31} = \frac{n_{3.} * n_{.1}}{n} = \frac{50 * 10}{100} = 5$$

$$n_{13} = \frac{n_{1.} * n_{.3}}{n} = \frac{30 * 60}{100} = 18$$

$$n_{33} = \frac{n_{3.} * n_{.3}}{n} = \frac{50 * 60}{100} = 30$$

$$n_{12} = \frac{n_{1.} * n_{.2}}{n} = \frac{30 * 30}{100} = 9$$

$$n_{22} = \frac{n_{2.} * n_{.2}}{n} = \frac{20 * 30}{100} = 6$$

$$n_{32} = \frac{n_{3.} * n_{.2}}{n} = \frac{50 * 30}{100} = 15$$

$$n_{23} = \frac{n_{2.} * n_{.3}}{n} = \frac{20 * 60}{100} = 12$$

Le frequenze riportate sono identiche a quelle ottenibili in caso di indipendenza

Osservazioni

La condizione di indipendenza è molto stringente: è sufficiente che si verifichi discrasia in una sola cella (ad esempio uno zero) perché ci sia dipendenza.

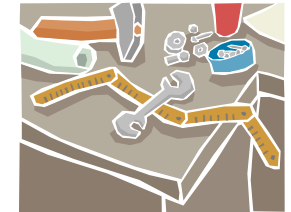
Infatti, è difficile trovare casi in cui si sia perfetta indipendenza, anche per variabili molto remote e logicamente non collegate

Ne consegue che nel valutare il grado di dipendenza dovremo guardarci dai disturbi dovuti a

Errori di misurazione

Fluttuazioni campionarie

e che non dipende da un nesso di causalità.



Le contingenze

La misura del grado di dipendenza si basa sullo scarto tra frequenza osservata in una cella e la frequenza teorica che si osserverebbe se fra le variabili ci fosse perfetta indipendenza

$$c_{ij} = n_{ij} - n'_{ij} \quad \text{dove } n'_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

che è detta CONTINGENZA (assoluta o relativa secondo le frequenze utilizzate)

In caso di indipendenza le contingenze sono tutte nulle per cui se si ha

$c_{ij} > 0$ nella cella "i,j" si riscontra un addensamento di frequenze rispetto alla situazione di indipendenza dei due fenomeni.

$c_{ij} < 0$ nella cella "i,j" si riscontra una rarefazione di frequenze rispetto alla situazione di indipendenza dei due fenomeni.

Esempio

Pazienti classificati per durata dello stato febbrile e per il tipo di trattamento subito

Frequenze osservate

	Durata della febbre (in gg)				
Trattam.	1-4	5-6	7-8	9-12	
A	45	27	20	12	104
B	25	10	9	10	54
C	56	47	30	18	151
	126	84	59	40	309

concordano i totali di colonna

Frequenze teoriche

concordano i totali di riga

	Durata della febbre (in gg)				
Trattam.	1-4	5-6	7-8	9-12	
A	42.41	28.27	19.86	13.46	104.00
B	22.02	14.68	10.31	6.99	54.00
C	61.57	41.05	28.83	19.55	151.00
	126.00	84.00	59.00	40.00	309.00

Esempio (continua)

Tabella delle contingenze

Trattam.	Durata della febbre (ingg)				
	1-4	5-6	7-8	9-12	
A	2.59	-1.27	0.14	-1.46	0.00
B	2.98	-4.68	-1.31	3.01	0.00
C	-5.57	5.95	1.17	-1.55	0.00
	0.00	0.00	0.00	0.00	0.00

La somma per colonne delle contingenze è sempre nulla

La somma per righe delle contingenze è sempre nulla

Massimo scostamento negativo

Massimo scostamento positivo

Proprietà della tabella di contingenza

PROPRIETA': La somma delle contingenze di riga o di colonna è pari a zero.

Dimostrazione per le contingenze di riga

$$\sum_{i=1}^r c_{ij} = \sum_{i=1}^r \left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right) = \sum_{i=1}^r n_{ij} - \sum_{i=1}^r \frac{n_{i.} n_{.j}}{n} = n_{.j} - \frac{n_{.j}}{n} \sum_{i=1}^r n_{i.} = n_{.j} - n_{.j} = 0$$

Dimostrazione per le contingenze di colonna

$$\sum_{i=1}^r c_{ij} = \sum_{j=1}^c \left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right) = \sum_{j=1}^c n_{ij} - \sum_{j=1}^c \frac{n_{i.} n_{.j}}{n} = n_{i.} - \frac{n_{i.}}{n} \sum_{j=1}^c n_{.j} = n_{i.} - n_{i.} = 0$$

Connessione tra variabili

In questo ambito gli eventuali legami di dipendenza si riflettono esclusivamente nella classificazione delle unità.

Se si scambiano tra di loro le righe o le colonne, l'associazione non cambia

Una variabile è connessa ad un'altra se, al modificarsi delle sue modalità, cambia la proporzione con cui si verificano le modalità di quella condizionata.

La difformità della o delle parziali rispetto alla marginale può verificarsi per una sola modalità o per tutte; può inoltre essere di poco conto oppure di grande entità.

Esiste una gradualità della connessione che procede da un minimo (la condizione di indipendenza) ad un massimo.

Esempio

Un'impresa commercializza 4 bibite tipo cola in diverse aree geografiche

Area	Prodotto				Totale
	Moka-Cola	Neocafé	Arabeira	Decaf	
Sud	72	8	12	23	115
Nord	7	10	14	19	50
Centro	26	10	16	33	85
Totale	105	28	42	75	250

L'ufficio marketing si domanda se c'è un legame tra il tipo consumato e l'area di residenza del consumatore.

La tabella classifica le unità di assaggio - simultaneamente- per regione e per prodotto preferito.

La risposta deve essere data usando in modo efficace le informazioni così raccolte



Connessione massima

Tra Y ed X esiste la massima connessione se nota una qualsiasi modalità di X è univocamente determinata la modalità di Y ad essa corrispondente

Se la tabella è rettangolare non è possibile la reciprocità della dipendenza massima

$$r < s$$

	y_1	y_2	y_3	y_4	
x_1	7	0	3	14	24
x_2	0	4	0	0	4
	7	4	3	14	28

Se si fissa la Y, diciamo al livello y_2 , la modalità di X è necessariamente x_2 . Ogni volta che si sceglie Y risulta subito scelta anche X. Il contrario non è vero.

$$r > s$$

	y_1	y_2	y_3	
x_1	6	0	0	6
x_2	0	0	9	9
x_3	0	4	0	4
x_4	2	0	0	2
	8	4	9	21

Analogia situazione, ma con ruoli invertiti. Una volta scelta X risulta automaticamente scelta anche Y, ma non viceversa.

Misure sintetiche della Connessione

La contingenza è un indicatore, in valore ed in segno, dello scostamento tra le frequenze osservate e quelle attese nel caso di indipendenza delle due variabili.

Possiamo considerare i rapporti di contingenza

$$\delta_{ij} = \frac{(f_{ij} - f'_{ij})}{f'_{ij}} = \frac{\frac{n_{ij}}{n} - \frac{n_{i.}n_{.j}}{n^2}}{\frac{n_{i.}n_{.j}}{n^2}} = \frac{n_{ij} - \frac{n_{i.}n_{.j}}{n}}{\frac{n_{i.}n_{.j}}{n}}$$

che misurano lo scarto percentuale delle frequenze (assolute o relative) osservate rispetto alle teoriche

Per misurare il grado di connessione useremo medie aritmetiche dei rapporti di contingenza.

Statistica del Mortara

E' una media ponderata dei rapporti di contingenza presi in valore assoluto

$$M = \sum_{i=1}^r \sum_{j=1}^s \left| \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \right| f_{i.}f_{.j} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s |C_{ij}|$$

e corrisponde alla media aritmetica semplice delle contingenze in valore assoluto

L'indice del Mortara è normalizzato: $0 \leq M \leq 2$.

$$M = \sum_{i=1}^r \sum_{j=1}^s \left| \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \right| f_{i.}f_{.j} = \sum_{i=1}^r \sum_{j=1}^s |f_{ij} - f_{i.}f_{.j}| \leq \sum_{i=1}^r \sum_{j=1}^s |f_{ij} + f_{i.}f_{.j}| \leq \sum_{i=1}^r \sum_{j=1}^s f_{ij} + \sum_{i=1}^r \sum_{j=1}^s f_{i.}f_{.j} = 2$$

Assume valore nullo se e solo se le contingenze sono tutte nulle ovvero se c'è perfetta indipendenza.

Assume il valore massimo in caso di perfetta dipendenza (che di solito è <2)

Esempio di calcolo di M

	Y1	Y2	Y3	Y4	
X1	8	2	10	10	30
X2	5	4	6	5	20
	13	6	16	15	50

Frequenze attese in caso di indipendenza

	Y1	Y2	Y3	Y4	
X1	2.6	3.6	9.6	9.0	30
X2	5.2	2.4	6.4	6.0	20
	13.0	6.0	16.0	15.0	50

Valore assoluto delle contingenze

	Y1	Y2	Y3	Y4	
X1	5.4	1.6	0.4	1.0	8.4
X2	0.2	1.6	0.4	1.0	3.2
	5.6	3.2	0.8	2.0	11.6

$$M = \frac{11.6}{50} = 0.232$$

Statistica χ^2 (chi quadrato)

Questo indice si basa sulla media ponderata dei rapporti di contingenza al quadrato

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \left(\frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}} \right)^2 f_{i.}f_{.j} = \left[\sum_{i=1}^r \sum_{j=1}^s \left(\frac{f_{ij}^2}{f_{i.}f_{.j}} \right) \right] - 1 = n \left[\sum_{i=1}^r \sum_{j=1}^s \left(\frac{n_{ij}^2}{n_{i.}n_{.j}} \right) \right] - 1$$

Il chi-quadrato è nullo se e solo se c'è perfetta indipendenza tra le due variabili.

Aumenta se aumenta la differenza tra frequenze teoriche ed osservate.

L'indice, per come è definito, può un valore valori arbitrariamente grandi.

Valori estremi del χ^2

Se le variabili fossero indipendenti allora $f_{ij} = (f_{i.})(f_{.j})$ e quindi

$$\begin{aligned} \chi^2 &= \left[\sum_{i=1}^r \sum_{j=1}^s \left(\frac{f_{ij}^2}{f_{i.}f_{.j}} \right) \right] - 1 = \left[\sum_{i=1}^r \sum_{j=1}^s \left(\frac{(f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \right) \right] - 1 = \left[\sum_{i=1}^r \sum_{j=1}^s f_{i.}f_{.j} \right] - 1 \\ &= \left(\sum_{i=1}^r f_{i.} \right) \left(\sum_{j=1}^s f_{.j} \right) - 1 = (1)(1) - 1 = 0 \end{aligned}$$

In caso di perfetta dipendenza sarebbero nulle tutte le celle fuori diagonale.

$$\chi^2 = n \begin{cases} \sum_{i=1}^r \left(\frac{n_{ii}^2}{n_{i.}n_{.i}} \right) + \sum_{j=r+1}^s \left(\frac{n_{(j-r)j}^2}{n_{(j-r).}n_{.j}} \right) - 1 & \text{se } r \leq s \\ \sum_{j=1}^s \left(\frac{n_{jj}^2}{n_{.j}n_{j.}} \right) + \sum_{i=s+1}^r \left(\frac{n_{i(i-s)}^2}{n_{i.}n_{(i-s).}} \right) - 1 & \text{se } r \geq s \end{cases}$$

Il massimo cambia da tabella a tabella.

Esempio

Produzione di palloni di cuoio. Per il controllo della qualità i prodotti sono classificati rispetto a: X=pressione interna e Y=superficie esterna.



Y/X	Rifutati	Imperfetti	Standard	
Rifutati	12	23	89	124
Imperfetti	8	12	62	82
Standard	21	30	119	170
	41	65	270	376

Y/X	R.	I.	S.	
R.	$\frac{14 = 124 \cdot 41}{376}$	21	89	124
I.	9	14	59	82
S.	18	30	122	170
	41	65	270	376

$$\chi_c^2 = \frac{(12-14)^2}{14} + \frac{(23-21)^2}{21} + \dots + \frac{(119-122)^2}{122} = 1.599$$

Il valore dell'indice sembra basso, ma è abbastanza basso?

Esempio

Un'indagine ha classificato i rivenditori di hardware di una regione secondo il tipo di società ed il tipo di collocazione



Negozio	Tipologia società			Totale
	Persone	Cooperativa	Impresa	
Autonomo	34	16	4	54
Supermercato	4	2	3	9
Misto proprio	17	21	32	70
Misto altri	13	5	6	24
Totale	68	44	45	157

$$M = \frac{55.8471}{157} = 0.3557, \quad \chi^2 = 1.1771$$

Dovremo ricorrere all'inferenza statistica per stabilire se ci troviamo di fronte ad una associazione significativa

Esercizio (Excel)

Indagine sulla mobilità di voto. Uso dello strumento PivotTable

Soggetto	Ha votato	Voterà			
Adua	Centro	Destra	Iris	Centro	Centro
Aida	Sinistra	Sinistra	Irma	Destra	Destra
Alda	Destra	Destra	Jula	Sinistra	Centro
Alea	Centro	Centro	Kara	Sinistra	Destra
Alfa	Destra	Centro	Lara	Destra	Sinistra
Anna	Sinistra	Sinistra	Leda	Centro	Centro
Asia	Centro	Destra	Lena	Sinistra	Sinistra
Atte	Sinistra	Centro	Lisa	Centro	Centro
Beba	Sinistra	Sinistra	Lory	Sinistra	Centro
Bice	Centro	Destra	Mara	Centro	Destra
Cira	Centro	Sinistra	Mena	Centro	Sinistra
Cleo	Destra	Destra	Mina	Sinistra	Sinistra
Cora	Sinistra	Destra	Mira	Sinistra	Sinistra
Demi	Centro	Destra	Olga	Centro	Destra
Dina	Centro	Centro	Pina	Centro	Centro
Dora	Destra	Destra	Rina	Destra	Centro
Edda	Centro	Destra	Rita	Destra	Destra
Elsa	Destra	Sinistra	Rosa	Sinistra	Sinistra
Emma	Sinistra	Sinistra	Sara	Destra	Destra
Enza	Centro	Destra	Teti	Centro	Destra
Etta	Centro	Centro	Tina	Sinistra	Sinistra
Fede	Destra	Destra	Vega	Sinistra	Sinistra
Gina	Sinistra	Centro	Vera	Centro	Destra
Gisa	Centro	Destra	Zita	Destra	Destra
Ines	Destra	Destra	Zora	Centro	Centro

Count	Voterà			
Ha votato	Centro	Destra	Sinistra	Totale
Centro	8	11	2	21
Destra	2	9	2	13
Sinistra	4	2	10	16
Totale	14	22	14	50

5.88	9.24	5.88	21	=I3*\$F\$6/\$I\$6
3.64	5.72	3.64	13	
4.48	7.04	4.48	16	
14	22	14	50	

0.764	0.335	2.560	3.660	=(F3-F8)^2/F8
0.739	1.881	0.739	3.359	
0.051	3.608	6.801	10.461	
C hi-quadrato		17.480		
Gdl		4	=(3-1)(3-1)	
p-Value		0.0016	=Distrib.Chi(I16;I17)	

Chiarite più avanti

Rapporto di verosimiglianza

Sono coinvolti i logaritmi naturali delle frequenze osservate e teoriche

$$G^2 = 2 \sum_{i=1}^r \sum_{j=1}^s n_{ij} \log \left(\frac{n_{ij}}{n_{ij}^e} \right)$$

Il G-quadro è nullo se e solo se c'è perfetta indipendenza tra le due variabili.

Aumenta se aumenta la differenza tra frequenze teoriche ed osservate.

L'indice, per come è definito, può un valore valori arbitrariamente grandi.

N.B. $x \log(x)$ tende a zero se x tende a zero

Esempio

	Otite			
	Peggior	Uguale	Meglio	
Peggior	13	5	6	24
Brochite	1	19	4	24
Meglio	4	0	8	12
	18	24	18	60

Rilevazione dell'effetto di un antibiotico su pazienti affetti sia da bronchite che da otite dell'orecchio medio



	Otite			
	Peggior	Uguale	Meglio	
Peggior	7,2	9,6	7,2	24
Brochite	7,2	9,6	7,2	24
Meglio	3,6	4,8	3,6	12
	18	24	18	60

	Otite			
	Peggior	Uguale	Meglio	
Peggior	7,6813	-3,2616	-1,0939	3,3257
Brochite	-1,9741	12,9708	-2,3511	8,6456
Meglio	0,4214	0,0000	6,3881	6,8095
				37,5617

Il valore dell'indice sembra elevato, ma lo è abbastanza per concludere sul doppio spettro del farmaco?

Esercizio

Una ricerca sulla disponibilità ad andare in vacanza da sole per un campione di donne ha prodotto i seguenti risultati

	Certo che no	Forse no	Non sa	Forse si	Certo che si	Totale
Laureata	52	79	124	342	226	823
Semilaureata	62	153	136	417	262	1030
Diplomata	53	213	184	629	375	1454
Scuola sup.	54	231	221	571	244	1321
Lic.Media	43	175	319	439	190	1166
	264	851	984	2398	1297	5794

Calcolare il Mortara, il χ^2 ed il rapporto di verosimiglianza

Analisi della media

Questo tipo di studio si attiva se una delle variabili è metrica ed un'altra è qualitativa oppure quantitativa, ma con modalità non metriche



Costo di un appartamento e zona di residenza



Quantità di principio attivo e stadio della malattia

Si parlerà di dipendenza o indipendenza in media facendo riferimento a modifiche più o meno rilevanti della media di una variabile se l'altra subisce delle variazioni (indotte o spontanee)

Valore atteso della marginale

Le distribuzioni marginali sono delle vere e proprie distribuzioni univariate.

In particolare, ci interessa il valore atteso (o media aritmetica) della variabile metrica. Supponiamo sia la "Y"

$$E(Y) = \sum_{j=1}^c Y_j f_{.j} = \mu_y$$

Esempio

Y/X	A	B	C	
1	4/20	2/20	4/20	10/20
3	4/20	1/20	5/20	10/20
	8/20	3/20	9/20	1

$$\mu_y = 1\left(\frac{10}{20}\right) + 3\left(\frac{10}{20}\right) = \frac{40}{20} = 2$$

La scala della "X" è tale da non consentire il calcolo logico della media aritmetica

Valore atteso delle condizionate

Anche le distribuzioni condizionate sono delle distribuzioni univariate.

Per calcolare il valore atteso della variabile metrica, fissata la modalità della variabile di controllo, abbiamo

$$E(Y|X = x_i) = \sum_{j=1}^c Y_j \frac{f_{ij}}{f_{i.}} = \mu_y(x_i)$$

Esempio

Y/X	A	B	C	
1	4/20	2/20	4/20	10/20
3	4/20	1/20	5/20	10/20
	8/20	3/20	9/20	1

C'è una media di Y per ogni fissata X

$$\mu_y(A) = 1\left(\frac{4}{8}\right) + 3\left(\frac{4}{8}\right) = \frac{20}{8} = 2.5$$

$$\mu_y(B) = 1\left(\frac{2}{3}\right) + 3\left(\frac{1}{3}\right) = \frac{5}{3} = 0.667$$

$$\mu_y(C) = 1\left(\frac{4}{9}\right) + 3\left(\frac{5}{9}\right) = \frac{24}{9} = 2.667$$

Relazione tra i valori attesi

Al variare della variabile condizionante la condizionata assume un certo valore atteso. Quindi

$$E(Y|X) = \text{funzione}(x)$$

Il valor atteso della Y è quindi una funzione delle modalità della X.

Qual è la media della Y se vogliamo prescindere dai valori della X?

$$\begin{aligned} \sum_{i=1}^r E(Y|X = x_i) * f_{i.} &= \sum_{i=1}^r \left(\sum_{j=1}^c Y_j \frac{f_{ij}}{f_{i.}} \right) * f_{i.} = \sum_{i=1}^r \left(\sum_{j=1}^c Y_j f_{ij} \right) \\ &= \sum_{j=1}^c Y_j \left(\sum_{i=1}^r f_{ij} \right) = \sum_{j=1}^c Y_j f_{.j} = E(Y) \end{aligned}$$

La marginale della Y coincide con la media ponderata delle medie parziali della stessa Y.

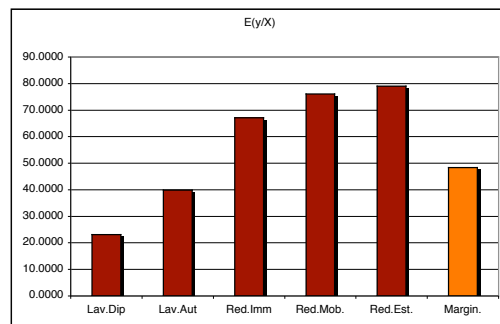
Esempio

Campione di contribuenti classificato per livello di reddito e tipologia di reddito.

Calcolo di medie condizionate e media marginale

	Yi						
	<18	18-24	24-32	32-48	49-81	>81	
Ci	12	20	29	50	70	100	Totale
Lav.Dip	140	120	20	18	12	10	320
Lav.Aut	90	75	60	55	50	43	373
Red.Imm	5	12	19	26	35	58	155
Red.Mob.	2	3	11	27	56	78	177
Red.Est.	0	1	6	16	34	54	111
	249	231	145	192	257	343	1136

Categ.	E(y/X)	fx	E(y/X)*fx
Lav.Dip	23.1250	320	7400
Lav.Aut	39.8660	373	14870
Red.Imm	67.1032	155	10401
Red.Mob.	76.1186	177	13473
Red.Est.	79.0450	111	8774
Margin.		1136	48.3433



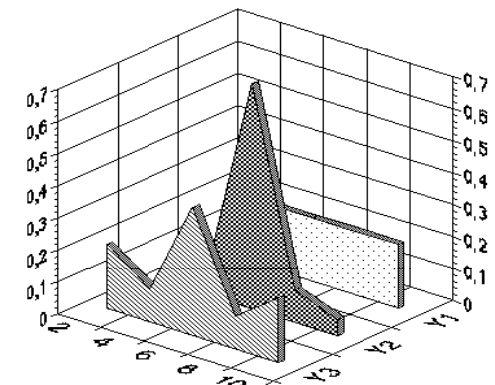
Indipendenza in media

Una variabile è indipendente in media da un'altra se le sue medie condizionali sono tutte uguali alla media marginale.

In questo caso non ci interessa se al variare di un carattere si modifichi o resti costante l'intera distribuzione. La nostra attenzione è limitata alla media.

Tre diverse distribuzioni parziali che hanno la stessa media

	Y1	Y2	Y3
2	0.20	0.00	0.20
4	0.20	0.15	0.10
6	0.20	0.70	0.40
8	0.20	0.15	0.10
10	0.20	0.00	0.20
	1.00	1.00	1.00



l'indipendenza in media non implica l'indipendenza distributiva

Considerazioni aggiuntive

- 1) L'indipendenza in media non necessariamente è simmetrica, cioè se la "Y" è indipendente in media dalla "X" nulla si può affermare sulla dipendenza in media della X rispetto alla Y
- 2) L'indipendenza assoluta implica l'indipendenza in media ovvero se fra la "Y" e la "X" si riscontra indipendenza assoluta allora ci sarà anche indipendenza in media.
- 3) L'indipendenza in media non può implicare l'indipendenza assoluta dato che lo stesso media può essere associata a distribuzioni molto diverse per altri aspetti

Sintesi delle medie condizionali

Le medie condizionali " μ_{y/x_i} " e le frequenze marginali $f_{.i}$ formano una distribuzione di frequenza:

Modalità	Frequenza
$\mu_{y(x_1)}$	$f_{1.}$
$\mu_{y(x_2)}$	$f_{2.}$
...	...
$\mu_{y(x_i)}$	$f_{i.}$
...	...
$\mu_{y(x_r)}$	$f_{r.}$
	1

Per la quale possiamo calcolare gli usuali indicatori di sintesi: media e varianza in particolare.

Ad esempio la media di questa distribuzione è data dalla media della marginale della Y.

Varianza delle medie condizionali

$$\text{Var}[E(y|X)] = \sum_{i=1}^r [\mu_y(x_i) - \mu_y]^2 f_i.$$

Esprime il valore medio dello scarto al quadrato tra le medie condizionali e quella marginale.

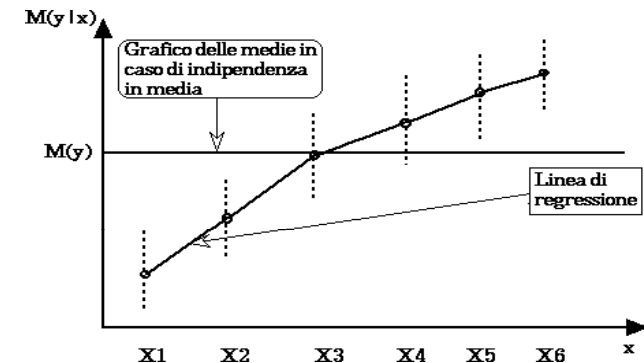
Misura la distanza tra le medie condizionali osservate e la distribuzione ed il valore (costante) che esse avrebbero in caso di indipendenza in media

La varianza delle medie condizionali è nulla se fra i caratteri c'è indipendenza assoluta. Infatti si ha

$$\mu_y(x_i) - \mu_y = 0 \text{ per } i = 1, 2, \dots, r$$

Rappresentazione grafica

Questo tipo di grafico può subito suggerire l'esistenza o meno della dipendenza in media tra le due variabili



Ancora sul valore atteso condizionale

il valore atteso della distribuzione condizionale è in genere funzione della v.c. che condiziona.

Esempio: $P(X_1, X_2) = \frac{X_1 + 2X_2}{18}$; per $X_1, X_2 = 1, 2$ con $P(X_1) = \frac{2X_1 + 6}{18}$ per $X_1 = 1, 2$;

Ne consegue che

$$P(X_2|X_1) = \frac{P(X_1, X_2)}{P(X_1)} = \frac{\frac{X_1 + 2X_2}{18}}{\frac{2X_1 + 6}{18}} = \frac{X_1 + 2X_2}{2X_1 + 6}$$

Vediamo il valore atteso delle distribuzioni condizionali

$$E(X_2|X_1) = \sum_{X_2=1}^2 X_2 \cdot \frac{X_1 + 2X_2}{2X_1 + 6} = \frac{X_1 + 1}{2X_1 + 6} + \frac{2X_1 + 4}{2X_1 + 6} = \frac{3X_1 + 5}{2X_1 + 6}$$

Che, come si vede, è funzione del valore di X_1 : cambiando quest'ultimo si altera la distribuzione condizionale e perciò dovrebbe cambiarne il valore atteso.

Se questo non succede c'è **INDIPENDENTI IN MEDIA**.

Misura della dipendenza in media

La misura più ovvia è la varianza delle medie parziali intorno alla media marginale.

Si annulla solo nel caso di indipendenza in media ed aumenta all'aumentare del grado di dipendenza

E' massima se fissata una qualunque della condizionante si può risalire con certezza alla media della condizionata.

Questo succede solo quando per ogni riga o colonna della tabella doppia entrata c'è una sola cella diversa da zero.

	y_1	y_2	y_3	
x_1	6	0	0	6
x_2	0	0	9	9
x_3	0	4	0	4
x_4	2	0	0	2
	8	4	9	21

In tali casi la media parziale di $Y|x$ coincide con la modalità di Y corrispondente ad x .

La varianza delle medie condizionali coincide con la varianza marginale della Y

Il rapporto di correlazione di K. Pearson

$$\eta_{y/x} = \sqrt{\frac{\sum_{i=1}^r [\mu_y(x_i) - \mu_y]^2 f_i}{\sum_{i=1}^r [y_i - \mu_y]^2 f_i}}$$

Il deponente segnala che l'indice è costruito per la Y dato che è la X a condizionare

L'indice eta mette a confronto la variabilità tra le medie condizionali con la variabilità marginale del carattere condizionato.

Poiché eta è costruito come rapporto di una quantità positiva al suo massimo avrà valori compresi nell'intervallo [0, 1]

L'indice è invariante rispetto a trasformazioni lineari della variabile condizionata

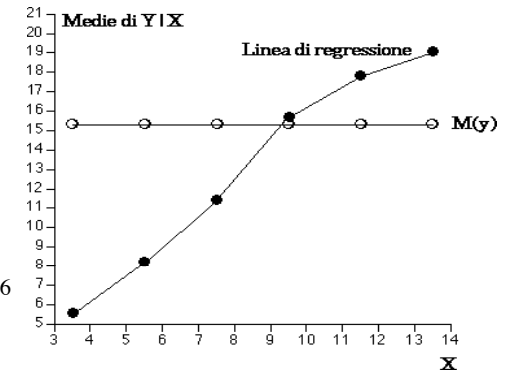
Esempio

Scolarità e reddito in un campione di soggetti

Scolarità	Reddito	
x	M(y x)	f(y x)
3.5	5.5000	0.0165
5.5	8.1667	0.0744
7.5	11.3947	0.1570
9.5	15.6622	0.3058
11.5	17.7609	0.3802
13.5	19.0000	0.0661
m.margin.	15.2851	1.0000

$$\text{Var}[E(y|x)] = 11.0143, \quad \text{Var}(y) = 23.2596$$

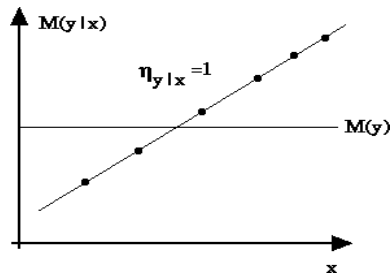
$$\eta_{y/x} = \sqrt{\frac{11.0143}{23.2596}} = 0.6881$$



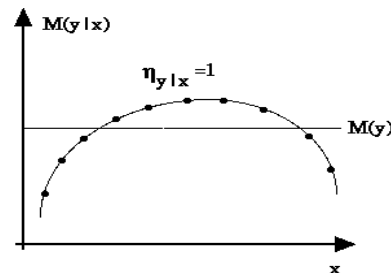
Esiste una dipendenza in media di tipo diretto: all'aumentare della X aumenta, in media, anche la y.

Senza ulteriori sviluppi inferenziali non possiamo stabilire fino a che punto ciò che si è riscontrato nel campione sia vero per l'intera popolazione

Casi particolari



All'aumentare della x il livello medio della y cresce in modo lineare esatto



All'aumentare della x il livello medio della y ha un andamento parabolico: cresce fino ad un certo livello per poi diminuire

Esercizio

Percentuale di incremento degli incentivi per un campione di lavoratori a "progetto" classificati in base al livello di specializzazione

	5	10	15	20	
L ₁	16	10	8	6	40
L ₂	10	13	17	23	63
L ₃	20	12	11	7	50
L ₄	22	9	5	2	38
L ₅	18	14	11	9	52
L ₆	14	12	6	3	35
	100	70	58	52	280



Calcolare il rapporto di correlazione.

Due variabili quantitative

Entrambe le variabili rilevate sulle unità sono misurate con scala metrica

Sebbene sia possibile effettuare lo studio della connessione o quello della dipendenza in media questo è sconsigliato.

Nel primo caso si perdono tutte le informazioni relative alle modalità delle variabili.

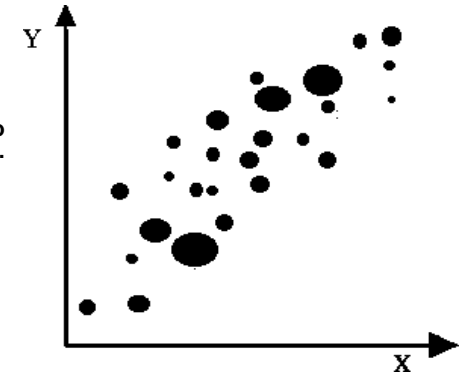
Nel secondo si trascura gran parte delle informazioni contenute nelle modalità della variabile condizionante.

Questo spreco è inopportuno, a meno che le misurazioni sulla condizionante o su entrambe le variabili non siano tanto contaminate da errori da costringere ad ignorare la loro scala.

Diagramma di dispersione (Scatterplot)

Su due assi coordinati ed in scala opportuna si riportano i valori delle due variabili ed ogni combinazione (X,Y) è rappresentata da un punto.

Per ogni combinazione (X,Y) si visualizza la frequenza relativa ad essa assegnata dalla distribuzione congiunta con cerchi di raggio ad essa proporzionali



Questo è il grafico più noto ed è di realizzazione e lettura molto semplice evidenziando la tendenza ad abbinarsi delle due variabili.

Una lettura attenta permette anche di stabilire, con buona approssimazione, il tipo di legame tra la Y e la X.

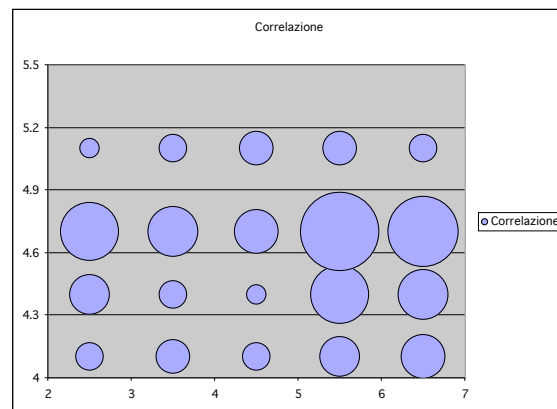
Tabella a doppia entrata

	4.1	4.4	4.7	5.1
2.5	2	4	8	1
3.5	3	2	6	2
4.5	2	1	5	3
5.5	4	8	15	3
6.5	5	6	12	2

Sviluppo in coppie di valori

X	Y	frequenze
2.5	4.1	2
2.5	4.4	4
2.5	4.7	8
2.5	5.1	1
3.5	4.1	3
3.5	4.4	2
3.5	4.7	6
3.5	5.1	2
4.5	4.1	2
4.5	4.4	1
4.5	4.7	5
4.5	5.1	3
5.5	4.1	4
5.5	4.4	8
5.5	4.7	15
5.5	5.1	3
6.5	4.1	5
6.5	4.4	6
6.5	4.7	12
6.5	5.1	2

Esempio- Excel

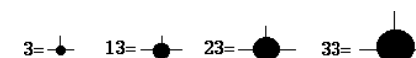


Esercizio (scatterplot)

Ampiezza famiglia	Numero di percettori di reddito					
	1	2	3	4	5	
1	5	-	-	-	-	5
2	12	3	-	-	-	15
3	15	9	2	-	-	26
4	18	12	7	3	-	40
5	20	19	14	5	4	62
6	13	28	20	12	3	76
7	6	24	27	18	8	83
8	-	16	32	21	12	81
	89	111	102	59	27	388

Costruite lo scatterplot

NB: per la rappresentazione grafica adoperate una combinazione del grafico a girasole (per le unità) e dei cerchi di raggio proporzionale (per le decine)



Valore atteso delle marginali

In questo caso possiamo considerare il valore atteso di entrambe le variabili

$$E(X) = \sum_{i=1}^r X_i f_{i.} = \mu_x ; \quad E(Y) = \sum_{j=1}^c Y_j f_{.j} = \mu_y$$

Esempio

Y/X	2	4	6	
1	$\frac{4}{20}$	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{10}{20}$
3	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{5}{20}$	$\frac{10}{20}$
	$\frac{8}{20}$	$\frac{3}{20}$	$\frac{9}{20}$	1

$$\mu_x = 2\left(\frac{8}{20}\right) + 4\left(\frac{3}{20}\right) + 6\left(\frac{9}{20}\right) = \frac{82}{20} = 4.1$$

$$\mu_y = 1\left(\frac{10}{20}\right) + 3\left(\frac{10}{20}\right) = \frac{40}{20} = 2$$

Valori attesi nelle distribuzioni doppie

Nel caso di variabili quantitative metriche siamo interessati anche al ...



Media o valore atteso della somma

$$E(X + Y) = \sum_{i=1}^r \sum_{j=1}^c (X_i + Y_j) f_{ij}$$



Media o valore atteso del prodotto

$$E(XY) = \sum_{i=1}^r \sum_{j=1}^c (X_i Y_j) f_{ij}$$

Media o valore atteso della somma

$$\begin{aligned} E(X + Y) &= \sum_{i=1}^r \sum_{j=1}^c X_i f_{ij} + \sum_{j=1}^c \sum_{i=1}^r Y_j f_{ij} = \sum_{i=1}^r X_i \sum_{j=1}^c f_{ij} + \sum_{j=1}^c Y_j \sum_{i=1}^r f_{ij} \\ &= \sum_{i=1}^r X_i f_{i.} + \sum_{j=1}^c Y_j f_{.j} = \mu_x + \mu_y \end{aligned}$$

Esempio

Y/X	2	4	6	
1	$\frac{4}{20}$	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{10}{20}$
3	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{5}{20}$	$\frac{10}{20}$
	$\frac{8}{20}$	$\frac{3}{20}$	$\frac{9}{20}$	1

$$\begin{aligned} E(X + Y) &= (2+1)\left(\frac{4}{20}\right) + (4+1)\left(\frac{2}{20}\right) + (6+1)\left(\frac{4}{20}\right) + \\ &\quad (2+3)\left(\frac{4}{20}\right) + (4+3)\left(\frac{1}{20}\right) + (6+3)\left(\frac{5}{20}\right) \\ &= \frac{12 + 10 + 28 + 20 + 7 + 45}{20} = \frac{122}{20} = 6.1 \end{aligned}$$

$$\mu_x + \mu_y = 4.1 + 2 = 6.1$$

Media o valore atteso del prodotto

$$E(XY) = \sum_{i=1}^r \sum_{j=1}^c X_i Y_j f_{ij}$$

Esempio

Y/X	2	4	6	
1	$\frac{4}{20}$	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{10}{20}$
3	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{5}{20}$	$\frac{10}{20}$
	$\frac{8}{20}$	$\frac{3}{20}$	$\frac{9}{20}$	1

$$\begin{aligned} E(XY) &= (2)\left(\frac{4}{20}\right) + (4)\left(\frac{2}{20}\right) + (6)\left(\frac{4}{20}\right) + \\ &\quad (6)\left(\frac{4}{20}\right) + (12)\left(\frac{1}{20}\right) + (18)\left(\frac{5}{20}\right) \\ &= \frac{8 + 8 + 24 + 24 + 12 + 90}{20} = 8.3 \end{aligned}$$

E(XY) in caso di indipendenza

$$E(XY) = \sum_{i=1}^r \sum_{j=1}^c X_i Y_j f_{i,j} = \sum_{i=1}^r X_i f_{i.} \sum_{j=1}^c Y_j f_{.j} = \mu_x \mu_y$$

In questo caso, la media dei prodotti è pari al prodotto delle medie.

Esempio

Y/X	2	4	6	
1	3/20	3/20	6/20	12/20
3	2/20	2/20	4/20	8/20
	5/20	5/20	10/20	1

$$E(XY) = (2)\left(\frac{3}{20}\right) + (4)\left(\frac{3}{20}\right) + (6)\left(\frac{6}{20}\right) + (6)\left(\frac{2}{20}\right) + (12)\left(\frac{2}{20}\right) + (18)\left(\frac{4}{20}\right) = \frac{6 + 12 + 36 + 12 + 24 + 72}{20} = 8.1$$

$$\mu_x = (2)\left(\frac{5}{20}\right) + (4)\left(\frac{5}{20}\right) + (6)\left(\frac{10}{20}\right) = 4.5$$

$$\mu_y = (1)\left(\frac{12}{20}\right) + (3)\left(\frac{8}{20}\right) = 1.8; \quad \mu_x \mu_y = 4.5 * 1.8 = 8.1$$

Esempio

Se non sono dipendenti si ha $E(X*Y) \neq E(X)*E(Y)$

		Variabile Y			
		2	4	6	
Var.	1	3/20	2/20	4/20	9/20
X	3	4/20	1/20	6/20	11/20
		7/20	3/20	10/20	1

$$[1*2]*\left(\frac{3}{20}\right) + [1*4]*\left(\frac{2}{20}\right) + [1*6]*\left(\frac{4}{20}\right) + [3*2]*\left(\frac{4}{20}\right) + [3*4]*\left(\frac{1}{20}\right) + [3*6]*\left(\frac{6}{20}\right) = \frac{182}{20}$$

$$E(y*X)=9.1$$

$$E(y) = 2*\left(\frac{7}{20}\right) + 4*\left(\frac{3}{20}\right) + 6*\left(\frac{10}{20}\right) = \frac{86}{20}$$

$$E(y)*E(x) = \frac{86}{20} * \frac{42}{20} = 9.03$$

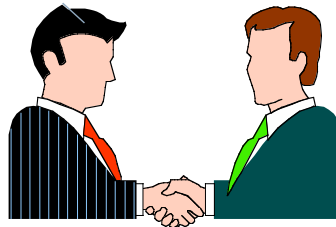
$$E(x) = 1*\left(\frac{9}{20}\right) + 3*\left(\frac{11}{20}\right) = \frac{42}{20}$$

Molto vicina, ma comunque diversa

La concordanza

Un aspetto essenziale della dipendenza tra due variabili su scala almeno intervallare è la concordanza, cioè la ricerca della direzione della dipendenza tra Y ed X.

Ci si chiede se valori inferiori (superiori) alla media si accompagnano con valori inferiori (superiori) alla media nell'altra



Per ognuna delle combinazioni di possibili valori si può averne una indicazione dagli SCARTI MISTI:

$$S_{ij} = (X_i - \mu_x)(Y_j - \mu_y)$$

Significato della concordanza

Il segno degli scarti è utile per sapere se, per la combinazione dei valori "X_i" e "Y_j" l'andamento delle due variabili è concorde oppure discorde:

 CONCORDANZA

$$S_{ij} > 0 \Rightarrow (X_i > \mu_x) \text{ e } (Y_j > \mu_y) \quad \text{oppure} \quad (X_i < \mu_x) \text{ e } (Y_j < \mu_y)$$

 DISCORDANZA

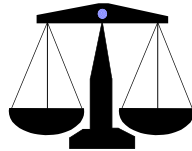
$$S_{ij} > 0 \Rightarrow (X_i > \mu_x) \text{ e } (Y_j < \mu_y) \quad \text{oppure} \quad (X_i < \mu_x) \text{ e } (Y_j > \mu_y)$$

E' difficile cogliere il senso della concordanza analizzando uno per uno TUTTI gli scarti misti.

La covarianza

La sintesi più semplice di tutti gli scarti misti è il loro valore atteso che costituisce la covarianza tra Y ed X

$$Cov(X, Y) = \sum_{j=1}^c \sum_{i=1}^r (X_i - \mu_x)(Y_j - \mu_y) f_{ij}$$



Se $Cov(Y, X) > 0$; Predominano gli scarti di segno concorde. Ci si aspetta X e Y tendano a cambiare nella stessa direzione



Se $Cov(Y, X) < 0$; Predominano gli scarti di segno discorde. Ci si aspetta X e Y tendano a cambiare in direzioni opposte



Se $Cov(Y, X) = 0$; le forze di discordanza e di concordanza sono bilanciate e le due variabili si dicono INCORRELATE

Esempio di calcolo della covarianza

Y/X	2	4	6	
1	$\frac{1}{9}$	$\frac{3}{9}$	$\frac{1}{9}$	$\frac{5}{9}$
3	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{4}{9}$
	$\frac{3}{9}$	$\frac{4}{9}$	$\frac{2}{9}$	1

Medie marginali

$$\mu_x = 2 * \frac{3}{9} + 4 * \frac{4}{9} + 6 * \frac{2}{9} = \frac{34}{9} = 3.78$$

$$\mu_y = 1 * \frac{5}{9} + 3 * \frac{4}{9} = \frac{17}{9} = 1.89$$

$(y - \mu_y) / (y - \mu_x)$	-1.78	0.22	2.22	
-0.89	$\frac{1}{9}$	$\frac{3}{9}$	$\frac{1}{9}$	
1.11	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	

$$Cov(x, y) = -1.78 * (-0.89) * \frac{1}{9} + 0.22 * (-0.89) * \frac{3}{9} + 2.22 * (-0.89) * \frac{1}{9} +$$

$$-1.78 * 1.11 * \frac{2}{9} + 0.22 * 1.11 * \frac{1}{9} + 2.22 * 1.11 * \frac{1}{9}$$

$$Cov(x, y) = -0.2469$$

Dominano gli scarti discordi

Calcolo con il foglio elettronico

	A	B	C	D	E	F	G
1	Guidatrice	Percorso	Velocità media				
2	Anna	87.05	53.77		94.6790		
3	Gina	76.82	9.11		=COVARIANZA(B2:B21;C2:C21)		
4	Lena	87.56	64.02				
5	Lara	98.04	96.39				
6	Bice	86.28	50.05				
7	Zora	82.97	38.70				
8	Olga	85.38	53.60				
9	Rosa	88.01	71.17				
10	Pina	81.80	39.04				
11	Dina	86.74	51.71				
12	Dora	85.19	56.92				
13	Fede	84.28	53.07				
14	Beba	97.44	97.43				
15	Tina	85.96	55.53				
16	Sara	83.97	48.46				
17	Elsa	89.04	63.51				
18	Enza	94.77	81.92				
19	Edda	92.06	78.58				
20	Mara	88.06	64.08				
21	Emma	85.54	53.99				

Dominano gli scarti concordi

Esempio di calcoli per dati in classi

	1 - 3	3 - 5	5 - 7	
0 - 2	0	1	4	5
2 - 4	2	1	2	5
4 - 6	4	3	0	7
6 - 10	5	3	0	8
	11	8	6	25

	2	4	6
1	0.00	-0.62	-14.98
3	-1.41	0.18	2.11
5	-15.62	2.93	0.00
8	-43.52	6.53	0.00
	-58.54	13.01	-6.86
E(X)	2.56	Cov(X, Y)=	-2.10
E(Y)=	3.60		

In questi casi si utilizzano i valori centrali delle classi, ma con risultati più approssimati

Formula semplificata per la covarianza

Usando le proprietà delle sommatorie si ottiene

$$\begin{aligned} \text{Cov}(X, Y) &= \sum_{j=1}^c \sum_{i=1}^r (X_i - \mu_x)(Y_j - \mu_y) f_{ij} = \sum_{j=1}^c \sum_{i=1}^r (X_i - \mu_x) Y_j f_{ij} - \sum_{j=1}^c \sum_{i=1}^r (X_i - \mu_x) \mu_y f_{ij} \\ &= \sum_{j=1}^c Y_j \left[\sum_{i=1}^r (X_i - \mu_x) f_{ij} \right] - \mu_y \sum_{j=1}^c \left[\sum_{i=1}^r (X_i - \mu_x) f_{ij} \right] \\ &= \sum_{j=1}^c \sum_{i=1}^r X_i Y_j f_{ij} - \mu_x \sum_{j=1}^c Y_j f_{.j} = \sum_{j=1}^c \sum_{i=1}^r X_i Y_j f_{ij} - \mu_x \mu_y = E(XY) - \mu_x \mu_y \end{aligned}$$

che semplifica il calcolo e soprattutto l'interpretazione della covarianza

Se c'è indipendenza la covarianza è zero dato che in questo caso $E(XY) = \mu_x \mu_y$

Esempio

Supponiamo che due variabili abbiano frequenze congiunte date da

$$f(X_1, X_2) = \frac{X_1 + 2X_2}{18}; \quad \text{per } X_1, X_2 = 1, 2$$

Con distribuzioni marginali

$$f(X_1) = \frac{2X_1 + 6}{18} \quad \text{per } X_1 = 1, 2; \quad f(X_2) = \frac{3 + 4X_2}{18} \quad \text{per } X_2 = 1, 2;$$

che hanno medie: $E(X_1) = \frac{14}{9}; \quad E(X_2) = \frac{29}{18}$

La covarianza è:
$$\text{Cov}(X_1, X_2) = \left(\sum_{x_1=1}^2 \sum_{x_2=1}^2 X_1 X_2 \frac{X_1 + 2X_2}{18} \right) - \frac{14}{9} \cdot \frac{29}{18} = \left(1 \cdot 1 \cdot \frac{3}{18} + 2 \cdot 1 \cdot \frac{4}{18} + 1 \cdot 2 \cdot \frac{5}{18} + 2 \cdot 2 \cdot \frac{6}{18} \right) - \frac{406}{162}$$

In medie le due variabili sono discordi
$$= \frac{45}{18} - \frac{406}{162} = -\frac{1}{162}$$

Esempio/2

Supponiamo che le frequenze congiunte siano date da

$$f(X, Y) = \begin{cases} 1/4 & \text{per } (x, y) \in \{(-4, 1); (4, -1); (2, 2); (-2, -2)\} \\ 0 & \text{altrove} \end{cases}$$

Dato che si possono presentare solo le quattro coppie cui è associata una frequenza positiva la X e la Y sono dipendenti in senso funzionale

Noto che X=4 solo Y=-1 è possibile (frequenza positiva)

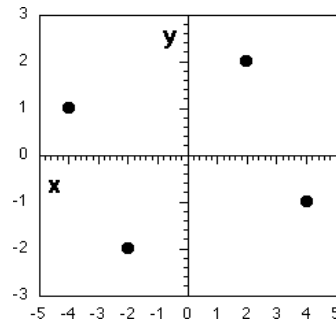
D'altra parte si ha

$$E(X) = -4 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} - 2 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} = 0;$$

$$E(Y) = 1 \cdot \frac{1}{4} - 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} - 2 \cdot \frac{1}{4} = 0; \quad E(XY) = E(X) \cdot E(Y) = 0$$

$$E(XY) = -4 \cdot \frac{1}{4} - 4 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} = 0;$$

Quindi, le due variabili pur essendo dipendenti (in senso funzionale) risultano incorrelate



Esercizi



A partire dalla seguente tabella

Y/X	-1	0	-1	
-2	1/8	1/8	0	2/8
0	1/8	0	2/8	3/8
-2	1/8	1/8	1/8	3/8
	3/8	2/8	3/8	1

Calcolare

1. $E(X)$, $E(Y)$;
2. $\text{Var}(X)$, $\text{Var}(Y)$;
3. $E(X \cdot Y)$
4. $E(X + Y)$
5. $\text{Cov}(X, Y)$



A partire dalla seguente distribuzione doppia

$$P(X_1, X_2) = \frac{X_1 + X_2}{32}; \quad X_1 = 1, 2; \quad X_2 = 1, 2, 3, 4.$$

Ripetere gli stessi calcoli dell'esercizio precedente

Covarianza e trasformazioni lineari

La covarianza risente delle trasformazioni moltiplicative, ma non di quelle additive

$$W_i = a + bX_i; \quad Z_j = c + dY_j$$

$$\begin{aligned} \text{Cov}(X, Y) &= \sum_{j=1}^c \sum_{i=1}^r W_i Z_j f_{ij} - \mu_w \mu_z = \sum_{j=1}^c \sum_{i=1}^r (a + bX_i)(c + dY_j) f_{ij} - [a + b\mu_x][c + d\mu_y] \\ &= \sum_{j=1}^c \sum_{i=1}^r [ac + bcX_i + adY_j + bdX_i Y_j] f_{ij} - [ac + ad\mu_y + bc\mu_x + bd\mu_x \mu_y] \\ &= ac \sum_{j=1}^c \sum_{i=1}^r [1] f_{ij} + bc \sum_{j=1}^c \sum_{i=1}^r [X_i] f_{ij} + ad \sum_{j=1}^c \sum_{i=1}^r [Y_j] f_{ij} + bd \sum_{j=1}^c \sum_{i=1}^r [X_i Y_j] f_{ij} \\ &\quad - ac - ad\mu_y - bc\mu_x - bd\mu_x \mu_y \\ &= ac + bc\mu_x + ad\mu_y + bd \sum_{j=1}^c \sum_{i=1}^r [X_i Y_j] f_{ij} - ac - ad\mu_y - bc\mu_x - bd\mu_x \mu_y \\ &= bd \sum_{j=1}^c \sum_{i=1}^r [X_i Y_j] f_{ij} - bd\mu_x \mu_y = bd \text{Cov}(X, Y) \end{aligned}$$

i parametri additivi "a" e "c" sono scomparsi, quelli moltiplicativi compaiono come fattore

Disuguaglianza Cauchy-Schwartz

Consideriamo una relazione che lega linearmente gli scarti medi di Y agli scarti medi di X

$$\begin{aligned} \sum_{j=1}^c \sum_{i=1}^r [(Y_j - \mu_y) - b(X_i - \mu_x)]^2 f_{ij} &= \sum_{j=1}^c \sum_{i=1}^r [(Y_j - \mu_y)^2 + b^2(X_i - \mu_x)^2 - 2b(Y_j - \mu_y)(X_i - \mu_x)] f_{ij} \geq 0 \\ &= \text{Var}(Y) + b^2 \text{Var}(X) - 2\text{Cov}(X, Y) \geq 0 \end{aligned}$$

Perché tale disequazione di 2° grado in "b" sia sempre soddisfatta, il discriminante deve essere negativo e cioè:

$$[2\text{Cov}(X, Y)]^2 - 4\text{Var}(Y)\text{Var}(X) \geq 0 \Rightarrow [\text{Cov}(X, Y)]^2 \leq \text{Var}(Y)\text{Var}(X)$$

La covarianza, al quadrato, è inferiore o uguale al prodotto delle varianze delle distribuzioni marginali

Limiti della Covarianza

La covarianza ha tutti i difetti delle misure di variabilità assoluta (dipendenza dall'unità di misura, mancanza di limiti predefiniti, etc.)

E' però legata alla variabilità dei due caratteri nel senso che non può superare, in valore assoluto, il prodotto degli SQM di Y e X.

Per ottenere un indice normalizzato e standardizzato covarianza è calcolata sulle variabili standardizzate. L'indice risultante si chiama coefficiente di correlazione

$$\text{Cov}(X^*, Y^*) = \sum_{j=1}^c \sum_{i=1}^r \left(\frac{X_i - \mu_x}{\sigma_x} \right) \left(\frac{Y_j - \mu_y}{\sigma_y} \right) f_{ij}$$

Coefficiente di correlazione



E' compreso tra -1 e +1 perché espresso come rapporto di una quantità (la covarianza) al suo massimo (in valore assoluto)

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) * \text{Var}(Y)}}$$



E' standardizzato. Se una o entrambe le variabili subiscono una trasformazione lineare il coefficiente rimane lo stesso:

$$r(a+bX, c+dY) = r(X, Y)$$



E' simmetrico rispetto alle due variabili: $r(Y, X) = r(X, Y)$



E' uguale a zero se c'è indipendenza tra le due variabili (il numeratore in questo caso è infatti zero)

Coefficiente di correlazione/2

Assume i valori estremi solo in caso di relazione lineare esatta

$$\begin{aligned} \text{Cov}(X, a + bX) &= \sum_{j=1}^c \sum_{i=1}^r [X_i(a + bX_j) - \mu_x(a + b\mu_x)] f_{ij} = \sum_{j=1}^c \sum_{i=1}^r aX_j f_{ij} + b \sum_{j=1}^c \sum_{i=1}^r X_i X_j f_{ij} - a\mu_x - b\mu_x^2 \\ &= \sum_{j=1}^c \sum_{i=1}^r aX_j f_{ij} + b \sum_{j=1}^c \sum_{i=1}^r X_i X_j f_{ij} - a\mu_x - b\mu_x^2 = a\mu_x - a\mu_x + b[\text{Var}(x)] \\ &= a \sum_{j=1}^c \mu_x f_{.j} + b \sum_{j=1}^c X_j^2 f_{.j} - a\mu_x - b\mu_x^2 = b\text{Var}(X) \end{aligned}$$

$Y_{ij}=0 \text{ se } i \neq j$

Ne consegue che

$$r(X, a + bX) = \frac{b\text{Var}(x)}{\sqrt{\text{Var}(x)\text{Var}(a + bX)}} = \frac{b\text{Var}(x)}{\sqrt{\text{Var}(x)b^2\text{Var}(x)}} = \frac{b}{|b|} = \begin{cases} -1 & \text{se } b < 0 \\ +1 & \text{se } b > 0 \end{cases}$$

il coefficiente di correlazione misura, quindi, l'intensità del legame lineare che sussiste tra le due variabili.

Esempio

Consideriamo la distribuzione congiunta: $f(X_1, X_2) = \frac{X_1 + 2X_2}{18}$; per $X_1, X_2 = 1, 2$

Con $E(X_1) = \frac{14}{9}$; $E(X_2) = \frac{29}{18}$; $\sigma^2(X_1) = \frac{20}{81}$; $\sigma^2(X_2) = \frac{77}{324}$; $\text{Cov}(X_1, X_2) = -\frac{1}{162}$;

il coefficiente di correlazione è $r(X_1, X_2) = \frac{-\frac{1}{162}}{\sqrt{\frac{20}{81} + \frac{77}{324}}} = \frac{1}{\sqrt{1540}} = -0.025$

Che un qualche legame di dipendenza ci fosse era chiaro dal fatto che

$$f(X_1, X_2) \neq f(X_1) \cdot f(X_2)$$

il valore di $r(X_1, X_2)$ conferma che i valori delle variabili sono discordi e il legame lineare è molto tenue

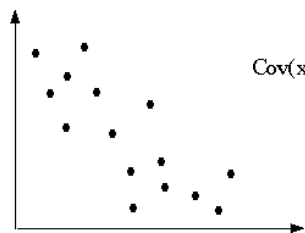
Semplificazioni per coppie di valori

Non sempre è opportuno e conveniente organizzare la variabile doppia in una tabella soprattutto se le coppie hanno la stessa probabilità.

Quando per le due v.c. siano osservabili "n" coppie di valori, ciascuna con frequenza pari a (1/n)

$$X_i, Y_i, \quad i = 1, 2, \dots, n \quad \text{con} \quad p_{ij} = \frac{1}{n}$$

In questo caso le formule si semplificano ed in particolare sparisce una sommatoria

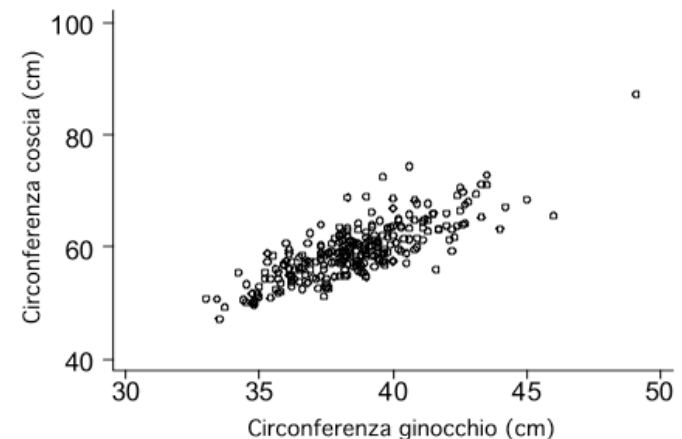


$$\text{Cov}(x, y) = \frac{\sum_{i=1}^n x_i y_i}{n} - \mu_x \mu_y$$

$$\begin{aligned} \mu_x &= \frac{\sum_{i=1}^n x_i}{n}; \quad \mu_y = \frac{\sum_{i=1}^n y_i}{n} \\ \sigma_x^2 &= \frac{\sum_{i=1}^n x_i^2}{n} - \mu_x^2; \quad \sigma_y^2 = \frac{\sum_{i=1}^n y_i^2}{n} - \mu_y^2; \end{aligned}$$

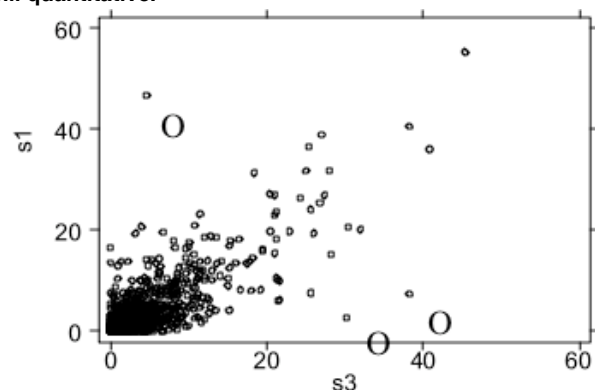
Scatterplot (valori singoli)

Su due assi coordinati ed in scala opportuna si riportano i valori delle due variabili ed ogni combinazione (X,Y) è rappresentata da un punto.



Scatterplot/2

Lo scatterplot offre una comoda rappresentazione delle possibili relazioni tra due variabili quantitative.

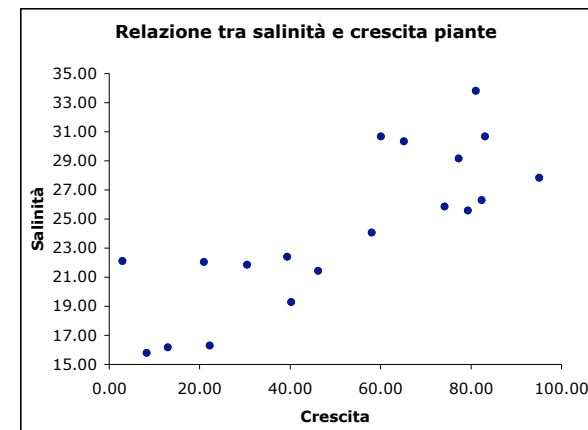


Il grafico evidenzia il gradiente dei dati, l'intensità del legame nonché i possibili valori anomali (outliers) cioè osservazioni lontane, a prima vista, dal centro della relazione.

Esempio

C'è una relazione tra il tasso di crescita delle mangrovie e la salinità del suolo?

Prelievi	Salinity	Crescita
1	2.90	22.12
2	40.25	19.29
3	60.05	30.69
4	8.24	15.80
5	58.05	24.08
6	95.07	27.85
7	79.31	25.58
8	8.35	14.59
9	12.93	16.17
10	22.21	16.31
11	77.23	29.17
12	74.11	25.87
13	20.91	22.05
14	83.08	30.68
15	81.02	33.82
16	82.31	26.30
17	46.19	21.45
18	65.12	30.34
19	30.46	21.86
20	39.31	22.42



Appare evidente una relazione diretta

Calcolo di $r(x,y)$ per coppie di valori

il calcolo è molto semplice purché opportunamente organizzato.

X	Y	X ²	Y ²	XY
0	0	0	0	0
1	0	1	0	0
1	0	1	0	0
1	1	1	1	1
1	1	1	1	1
2	1	4	1	2
Σ=	6	8	3	4

$$\mu_x = \frac{6}{6} = 1; \quad \mu_y = \frac{3}{6} = \frac{1}{2};$$

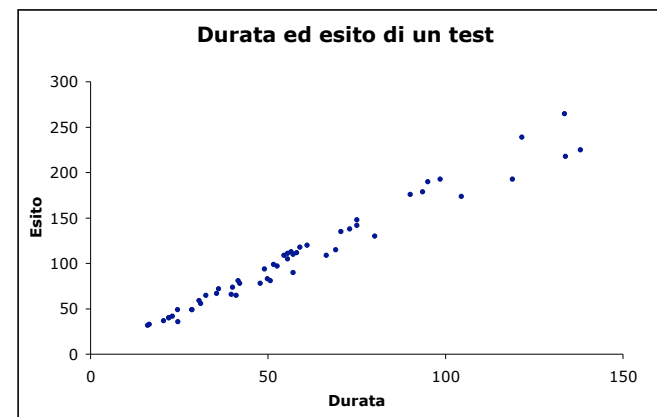
$$\sigma^2(X) = \frac{8}{6} - 1^2 = \frac{1}{3}; \quad \sigma^2(Y) = \frac{3}{6} - \left(\frac{1}{2}\right)^2 = \frac{1}{4};$$

$$\text{Cov}(X,Y) = \frac{4}{6} - 1 \cdot \frac{1}{2} = \frac{1}{6}; \quad r(X,Y) = 0.5773$$

Le due variabili presentano una correlazione positiva tendendo a presentare insieme i valori più grandi

Esercizio - Excel

Esito test	Durata
134	265 53 97
122	239 49 94
138	225 57 90
134	218 50 83
119	193 42 81
99	193 51 81
95	190 42 78
94	179 48 78
90	176 40 74
104	174 36 72
75	148 36 67
75	142 40 66
73	138 41 65
71	135 33 65
80	130 31 59
61	120 31 56
59	118 25 49
69	115 29 49
57	113 29 49
58	112 23 42
56	111 22 40
57	110 22 40
66	109 21 37
55	109 25 36
56	105 17 33
52	99 16 32



=CORREL(B2:B53;C2:C53)

0.95

Significato di $r(x,y)$

Dalle proprietà di " $r(x,y)$ " si deduce il suo significato:

Quanto più i suoi valori si avvicinano, in modulo, ad uno tanto più i valori delle variabili risultano collegabili con una retta.

D'altra parte, quanto più " r " è vicino a " ± 1 " tanto più la conoscenza di una delle variabili permette, attraverso la relazione lineare, di conoscere l'altra.

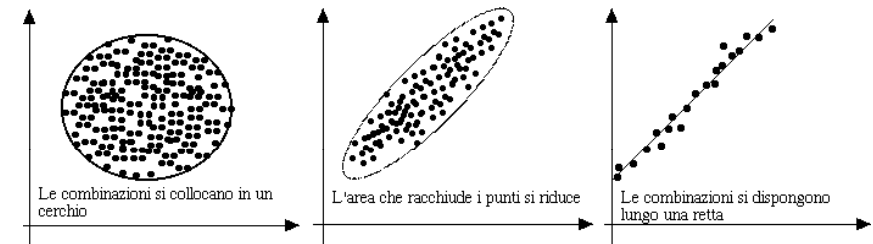
In questo senso " r " è una misura del grado di concordanza tra i valori della variabile doppia (X,Y)

- INTENSITA' DEL LEGAME LINEARE
- PREVEDIBILITA' DI UNA VARIABILE CONOSCENDO L'ALTRA
- GRADO DI CONCORDANZA

Scatterplot e correlazione

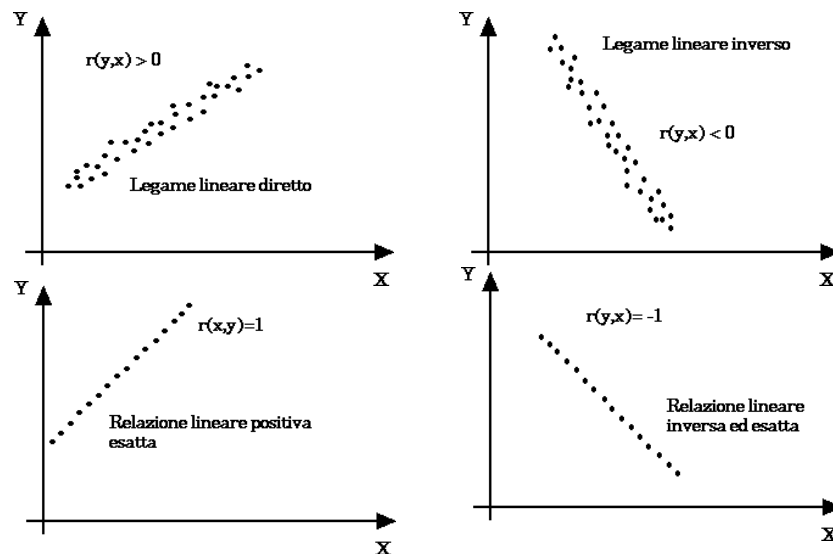
Lo scatterplot fornisce una idea immediata della intensità del legame che vige tra le due variabili

Si realizza riportando -in scala opportuna- le combinazioni osservate dei valori

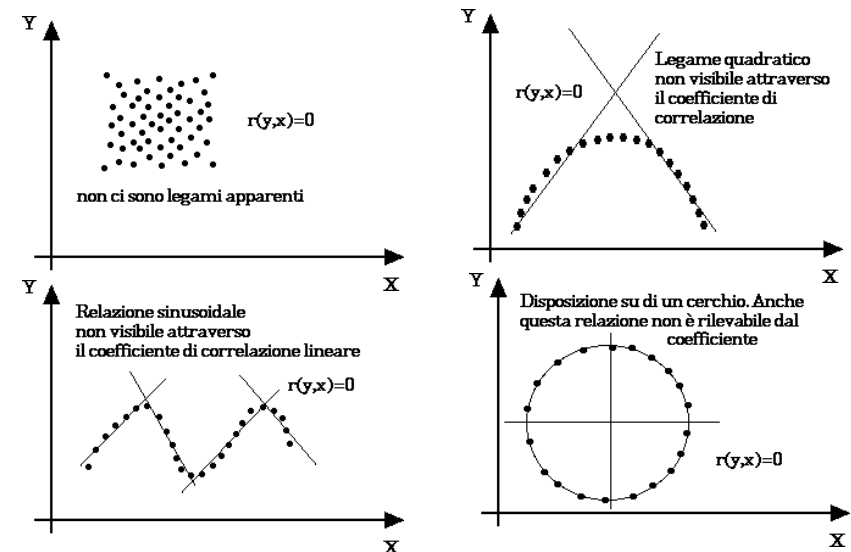


La relazione tra due variabili tende a divenire più stretta ma mano che la nube di punti passa dalla forma circolare, alla ellisse ed alla retta

Scatterplot e correlazione/2



Assenza di legami lineari



Correlazione e causa-effetto

L'esistenza di correlazione, per quanto intensa, non implica una relazione di causa ed effetto.

LEGAME PLAUSIBILE

Il tasso di criminalità è fortemente legato al tasso di disoccupazione.

LEGAME SPURIO

Nei bambini, la misura delle scarpe è molto correlata con la capacità di lettura.

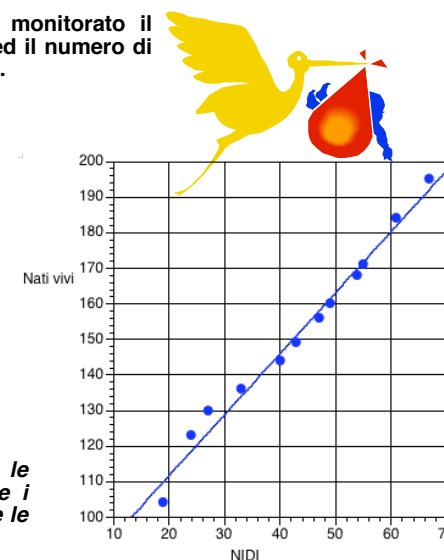
La correlazione indica solo che l'andamento di una variabile tende a disporsi secondo una retta se rappresentato insieme all'altra. I "perchè?" di questa tendenza vanno cercati al di fuori della statistica.

Il coefficiente di correlazione misura solo la co-variazione tra valori standardizzati

Esempio

In una zona del Nord Europa è stato monitorato il numero di nidi costruiti dalle cicogne ed il numero di nati vivi nel loro periodo di permanenza.

Anno	Nidi di cicogne	Nati vivi
1972	19	104
1973	24	123
1974	27	130
1975	33	136
1976	40	144
1977	43	149
1978	47	156
1979	49	160
1980	54	168
1981	55	171
1982	61	184
1983	67	195

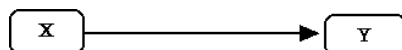


Dal punto di vista della correlazione le ipotesi che siano le cicogne a portare i bambini o che siano i bambini a portare le cicogne sono equivalenti.

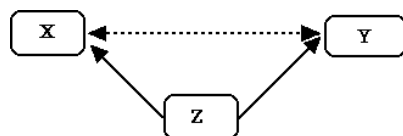
Correlazione spuria

Spesso, il valore di $r(y,x)$ altro non è che l'apparenza di un legame la cui sostanza è invece dovuta a fenomeni esterni.

La situazione di causalità tra X e Y:



Non è distinguibile dal legame spurio che fra di esse si pone a causa della comune dipendenza da una terza variabile Z



L'apprendimento di nuove parole non rende i piedi più grandi ovvero avere piedi più grandi non aiuta a conoscere nuove parole.
C'è un terzo fattore nascosto dietro la correlazione: l'età

Questo si verifica spesso a causa dell'esistenza di fenomeni tendenziali di lungo periodo che incidono allo stesso modo su variabili diverse

Esempio (vintage)

Prezzo del grano e tasso di matrimonialità.

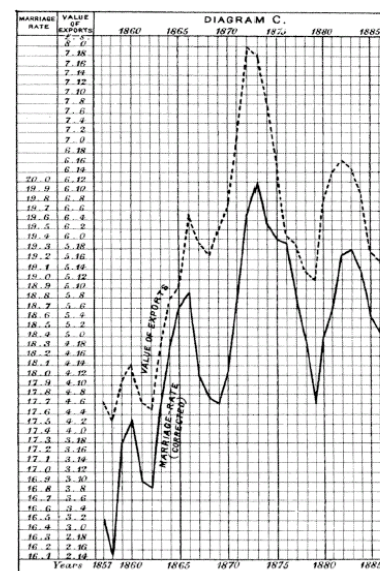
E' evidente un andamento sincrono che induce una correlazione molto elevata.

La spiegazione è semplice: un comune fattore esterno

Ci si sposa quando le condizioni economiche sono brillanti: c'è un aumento degli scambi e quindi delle importazioni

Se le importazioni aumentano tendono ad aumentare i noli ed i dazi.

Questo si riflette sui prezzi del grano che tendono aumentare (Ogle, 1890).



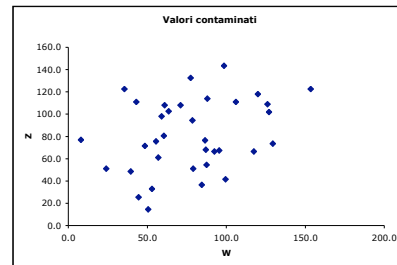
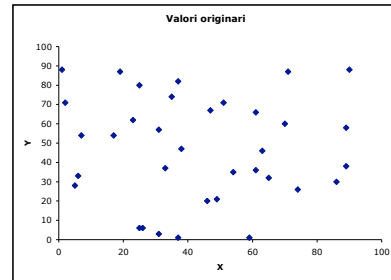
Esercizio

X	Y	W	Z	A
54	35	87.0	68.0	22
63	46	126.0	109.0	42
71	87	78.5	94.5	5
35	74	59.0	98.0	16
1	88	35.5	122.5	23
33	37	57.0	61.0	16
61	66	106.0	111.0	30
49	21	79.0	51.0	20
86	30	129.5	73.5	29
25	6	44.5	25.5	13
65	32	87.5	54.5	15
70	60	86.5	76.5	11
61	36	127.0	102.0	44
38	47	39.5	48.5	1
5	28	48.5	71.5	29
47	67	60.5	80.5	9
19	87	43.0	111.0	16
74	26	84.5	36.5	7
26	6	53.0	33.0	18
89	38	117.5	66.5	19
25	80	77.5	132.5	35
2	71	8.0	77.0	4
7	54	61.0	108.0	36
31	3	95.5	67.5	43
17	54	71.0	108.0	36
31	57	88.0	114.0	38
37	82	98.5	143.5	41
6	33	24.0	51.0	12
59	1	99.5	41.5	27
89	58	153.5	122.5	43
90	88	120.0	118.0	20
46	20	92.5	66.5	31
51	71	55.5	75.5	3
37	1	50.5	14.5	9
23	62	63.5	102.5	27

Corr(X,Y) -0.037
Corr(W,Z) 0.309

$$W = X + 1.2A$$

$$Z = Y + 1.2A$$



Dipendenza dei ranghi

Riguarda le variabili riportate in scala quantitativa ordinale.

> Perché non esiste una vera misura, ma solo un punteggio o valutazione

> Perché le misurazioni su sono imprecise o vizzate da errore

> Perché sono presenti dei valori remoti

Le modalità sono poste in corrispondenza con dei numeri naturali (ranghi)



Per ogni unità si osserva una coppia di modalità che si trasforma poi in una coppia di ranghi

$$(X_i, Y_i) \longleftrightarrow (r_i, s_i)$$

Esempio

Un gruppo di clienti di una banca classificato per reddito e per importo del prestito. Convertiamo i valori osservati in ranghi.

Cliente	Reddito (x)	Rango (x)	Prestito (y)	Rango (y)
A	25 600	6	8 600	5
B	17 800	9	8 800	4
C	167 200	1	500	8
D	44 200	3	6 600	6
E	36 400	4	10 500	3
F	27 400	5	74 400	1
G	83 600	2	0	9
H	18 600	8	6 300	7
I	24 500	7	12 100	2

E' evidente la perdita di informazione. Lo scarto tra i ranghi in X per i clienti H ed I è $9-7=2$ e sarebbe questo per qualunque coppia di valori compresi tra 18'600 e 24'500.

In breve, conoscere i ranghi poco ci dice sui valori originari

rho di Spearman

La misura forse più popolare della dipendenza tra i ranghi è la seguente

$$r_s = \frac{\sum_{i=1}^n \left(r_i - \frac{n+1}{2} \right) \left(s_i - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^n \left(r_i - \frac{n+1}{2} \right)^2 \sum_{i=1}^n \left(s_i - \frac{n+1}{2} \right)^2}} \quad \text{detto rho di Spearman}$$

Caso delle n coppie di valori senza posizioni di parità.

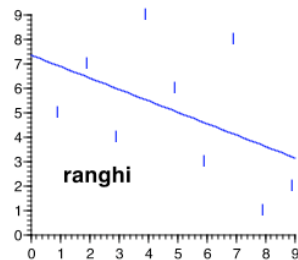
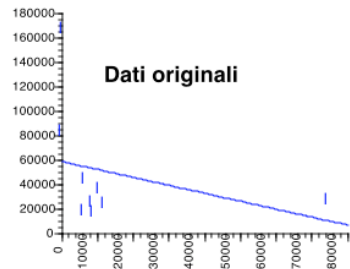
La definizione di r_s è la stessa del coefficiente di correlazione. Comunque il particolare tipo di dati coinvolti consente delle semplificazioni

$$r_s = 1 - \frac{6 \sum_{i=1}^n (r_i - s_i)^2}{n(n^2 - 1)}$$

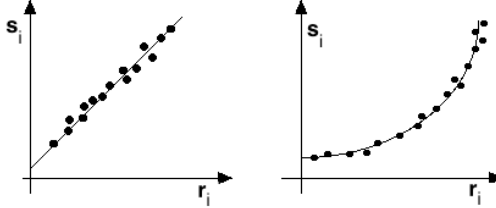
Esempio

Cliente	Reddito	r_i	Prestito	s_i	$(r_i - s_i)^2$
A	25600	6	8600	5	1
B	17800	9	8800	4	25
C	167200	1	500	8	49
D	44200	3	6600	6	9
E	36400	4	10500	3	1
F	27400	5	74400	1	16
G	83600	2	0	9	49
H	18600	8	6300	7	1
I	24500	7	12100	2	25
					176

$$r_s = 1 - \frac{6 \cdot 176}{9 \cdot (81 - 1)} = -0.4667$$



Considerazione sul rho di Spearman

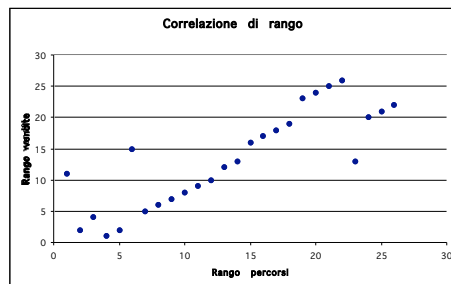
- Per costruzione l'indice rho varia tra -1 ed 1
 - Misura la dipendenza monotonica tra le due variabili. Assume il valore massimo (minimo) quando gli ordinamenti sono perfettamente concordi (discordi)
- 

$r_s = \pm 1$ se tra la Y e X esiste una relazione monotona, crescente o decrescente, senza vincolo di linearità
- Rho mette tutte le osservazioni sullo stesso piano (considera solo l'ordine) e perciò annulla l'effetto dei valori remoti
 - L'indice rimane invariato se una o entrambe le variabili subiscono una trasformazione monotona, ad esempio $Y' = \text{Log}(Y)$ e/o $X' = \text{exp}(X)$

Esempio

Unità	X Percorsi	Y Vendite	Rank(X)	Rank(Y)
A	121.5	373	21	25
B	151.5	314	25	21
C	146.2	301	24	20
D	106.7	263	16	17
E	98.9	204	11	9
F	95.1	176	9	7
G	90.1	138	4	1
H	115.5	329	19	23
I	71.7	225	1	11
J	111.7	300	18	19
K	93.6	164	7	5
L	109.6	284	17	18
M	105.3	252	15	16
N	125.0	400	22	26
O	91.7	239	6	15
P	88.7	161	3	4
Q	101.9	226	13	12
R	162.3	322	26	22
S	96.4	185	10	8
T	90.7	143	5	2
U	100.0	212	12	10
V	102.6	232	14	13
X	94.5	171	8	6
Y	88.6	143	2	2
W	119.4	358	20	24
Z	142.9	232	23	13

Venditori porta-a-porta per vendite e km percorsi



La correlazione è elevata sebbene si notino diversi disturbi

rho= 0.850084703
gdl= 25
tc= 7.907679188
p-Value 2.90161E-08

Esercizio

Calcolare il rho per le graduatorie dei fattori essenziali per la localizzazione di un nuovo impianto nel Mezzogiorno. Disegnare lo scatterplot

Fattori	Imprese Meridionali	Imprese non Meridionali
Costo del lavoro	1	1
Incentivi	2	2
Trasporti	7	5
Preferenze Personali	13	13
Servizi alle imprese	11	6
Finanziamenti	4	12
Materie prime	10	9
Sicurezza	8	3
Tassazione	6	7
Mercato	5	4
Commesse pubbliche	3	10
Collegamenti intern.	12	11
Qualità della vita	9	8

Rilevazione diretta dei ranghi

Un certo insieme di n oggetti o situazioni sono ordinate secondo il grado con cui presentano una certa caratteristica X .

Supponiamo ...

Che la caratteristica sia un *mix* di immaterialità graduabile, ma non misurabile.

Che le valutazioni siano espresse con i voti $\{1, 2, \dots, n\}$ così ottenendo la permutazione $\{s_1, s_2, \dots, s_n\}$

Ripetiamo la rilevazione per una Y rilevata allo stesso modo e che produce la permutazione: $\{r_1, r_2, \dots, r_n\}$



*Condizione di ansia e stress
Prima e dopo una separazione*

Il rho di Spearman cerca di quantificare l'intensità del legame tra i due insiemi di giudizi

Esempio: giudizi degli esperti



Ad un esperto è stato chiesto di pronunciarsi sulla posizione che le 20 squadre di un campionato di calcio occuperanno alla fine: $\{s_1, s_2, \dots, s_{20}\}$.

Alla fine della stagione i giudizi sono comparati con le posizioni reali: $\{r_1, r_2, \dots, r_{20}\}$.

Per semplificare il calcolo possiamo disporre le due serie di posizioni secondo l'ordine crescente della prima

Squadra	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	Totale
Prima	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	210
Dopo	9	2	4	7	5	1	3	8	6	11	13	10	14	18	15	12	16	20	17	19	210

$\rho=0.87$ (p -value 0.000001). L'esperto ha dato un buon giudizio sebbene sembri più in grado di indovinare le squadre che avranno una cattiva stagione rispetto a quelle che l'avranno buona

Esercizio

Ad un campione di consumatori è stato chiesto di giudicare la qualità di un servizio con un voto da 0 a 12.

E' anche stato chiesto di valutare con un voto da 0 a 12 la reputazione dell'azienda che forniva il servizio

Azienda	Rating servizio	Reputazione azienda
Alfa01	8	9
Alfa02	9	12
Alfa03	2	0
Alfa04	5	10
Alfa05	6	6
Alfa06	4	11
Alfa07	7	4
Alfa08	10	3
Alfa09	3	5
Alfa10	1	2
Alfa11	0	1
Alfa12	12	7
Alfa13	11	8

Vi sembra che ci sia un legame tra le due valutazioni?

Rho-Spearman= 0.4890
Tc= 1.8593
p-value 0.0899

tau di Kendall

E' una misura alternativa di dipendenza lineare tra ranghi

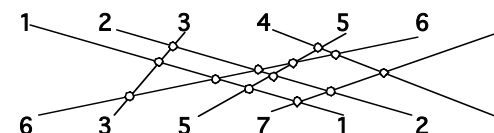
$$\begin{matrix} r_1 & r_2 & \dots & r_i & \dots & r_n \\ s_1 & s_2 & \dots & s_i & \dots & s_n \end{matrix}$$

$$\tau = 1 - \frac{4C}{n(n-1)} \quad \text{con} \quad -1 \leq \tau \leq 1$$

"C" è il numero minimo di scambi necessari per trasformare una graduatoria nell'altra. Gli estremi sono interpretabili come nel rho di Spearman

ESEMPIO

Calcolo con il metodo di Holmes (1920)

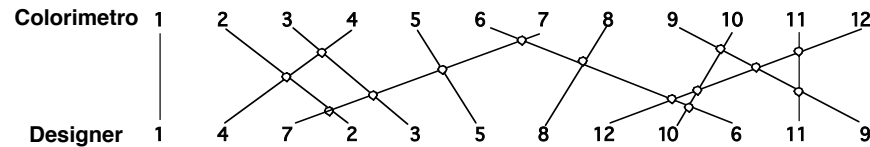


Le linee che congiungono i ranghi nelle due graduatorie si incrociano C volte

$$\tau = 1 - \frac{4(13)}{7(6)} = 1 - 1.2381 = -0.2381$$

Esempio

N=12 dischi hanno nuance del blu disposte secondo un colorimetro ed una candidata designer è chiamata a ricostruire la graduatoria



$$\tau = 1 - \frac{4(14)}{12(11)} = 1 - 0.4242 = 0.5758$$

Secondo il colorimetro c'è correlazione positiva ed abbastanza grande, ma che sia significativa dovrà essere stabilito con l'inferenza

Presenza di valori uguali

Se ci sono degli ex-aequo sorge il problema di assegnare il rango ai valori uguali

In genere si assegna a ciascun elemento di un gruppo di parità, la media dei ranghi che sarebbero loro spettati se fossero stati distinti.

Esempio

intensità	101	108	108	117	117	117	141	142	154	154	154	154	154	173
ranghi teorici	1	2	3	4	5	6	7	8	9	10	11	12	13	14
medie		(2+3)/2=2.5		(4+5+6)/3=5					(9+10+11+12+13)/5=11					
ranghi reali	1	2.5	2.5	5	5	5	7	8	11	11	11	11	11	14

Ai fini del calcolo nulla è cambiato se non la perdita delle semplificazioni possibili solo in caso di assenza di parità

Formula di rho in caso di parità

$$\rho_S = \frac{(n^3 - 3) - 6 \sum_{i=1}^n d_i^2 - \frac{1}{2} \left\{ \sum_{j=1}^{n_x} \left[(t_j^x)^3 - (t_j^x) \right] + \sum_{j=1}^{n_y} \left[(t_j^y)^3 - (t_j^y) \right] \right\}}{\sqrt{\left[(n^3 - 3) - \sum_{j=1}^{n_x} \left[(t_j^x)^3 - (t_j^x) \right] \right] \left[(n^3 - 3) - \sum_{j=1}^{n_y} \left[(t_j^y)^3 - (t_j^y) \right] \right]}}$$

dove

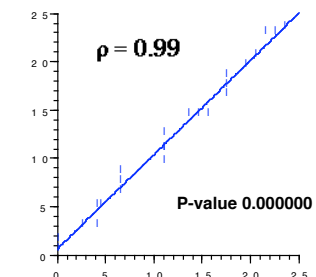
- n_x = numero di gruppi di X con parità
- t_j^x = numero di valori uguali per la j-esima parità in X
- n_y = numero di gruppi di Y con parità
- t_j^y = numero di valori uguali per la j-esima parità in Y

Esempio

E	E-or	rE	B	B-or	rB	(rE) ²	(rB) ²	rE*rB
10.3	9.2	0.5	-1.3	-1.8	1	0.25	1	0.5
9.9	9.2	0.5	-1.1	-1.6	2	0.25	4	1
10.3	9.3	3.0	1.3	-1.5	3.5	9	12.25	10.5
9.7	9.4	4.5	0.5	-1.5	3.5	20.25	12.25	15.75
9.6	9.4	4.5	1.8	-1.3	5.5	20.25	30.25	24.75
9.5	9.5	5.0	-0.4	-1.3	5.5	25	30.25	27.5
10.8	9.6	7.0	-1.6	-1.1	7	49	49	49
9.7	9.6	7.0	-1.8	-0.8	8	49	64	56
9.4	9.6	7.0	1.0	-0.7	9	49	81	63
10.1	9.7	11.5	-1.5	-0.5	10	132.25	100	115
9.4	9.7	11.5	-0.8	-0.4	11.5	132.25	132.25	132.25
10.4	9.7	11.5	-1.3	-0.4	11.5	132.25	132.25	132.25
9.6	9.7	11.5	-0.7	-0.2	13	132.25	169	149.5
9.7	9.8	14.0	-0.4	0.5	15	196	225	210
10.6	9.9	15.0	0.5	0.5	15	225	225	225
9.3	10.1	16.0	1.3	0.5	15	256	225	240
10.5	10.3	18.0	1.1	0.7	17	324	289	306
10.7	10.3	18.0	0.9	0.8	18	324	324	324
9.2	10.3	18.0	0.5	0.9	19	324	361	342
9.7	10.4	20.0	0.8	1.0	20	400	400	400
9.2	10.5	21.0	-1.5	1.1	21	441	441	441
9.6	10.6	22.0	-0.2	1.3	23.5	484	552.25	517
9.8	10.7	23.0	0.7	1.3	23.5	529	552.25	540.5
10.3	10.8	24.0	-0.5	1.8	24	576	576	576
						4830	4988	4898.5

Accertamento di una relazione d'ordine tra il tasso di interesse effettivo "E" dei BOT trimestrali e l'indice di borsa "B"

$$r_s = \frac{\frac{4898.5}{24} - \frac{(25)^2}{4}}{\sqrt{\left[\frac{4830}{24} - \frac{(25)^2}{4} \right] \left[\frac{4988}{24} - \frac{(25)^2}{4} \right]}}$$



Esempio

Distanza da un punto inquinante e concentrazione dell'agente nell'aria

Distanza (X)	Concen. (Y)	ranghi(X)	ranghi (Y)	d(x,y)
0	510	1	12	121
50	380	2	9	49
300	450	3.5	10	42.25
300	480	3.5	11	56.25
800	300	5	7.5	6.25
900	300	6	7.5	2.25
1000	170	7	6	1
1500	94	9	3.5	30.25
1500	94	9	3.5	30.25
1500	108	9	5	16
2000	45	11	1	100
5000	89	12	2	100
				554.5

$$\rho_S = \frac{1725 - 3327 - 0.5[(8-2) + (27-3)] + [(8-2) + (8-2)]}{\sqrt{[1725-30][1725-12]}} = -0.95$$

Esercizio

Voti in due discipline per un campione di studenti.

C'è un legame tra i due voti?



rho-Spearman 0.7836
Tc 7.5669
p-value 0.0000

Matricola	Disciplina A	Disciplina B
50825	18	18
64506	18	18
64289	18	18
31136	18	18
81016	20	19
91817	20	19
42720	20	19
92614	21	20
33491	21	20
31947	21	21
56554	21	21
83355	22	21
95516	22	21
44659	22	22
93637	22	22
70350	22	22
53806	23	24
44509	23	24
92149	23	24
86848	23	24
35750	24	24
95748	24	25
76681	25	25
70776	25	25
43071	26	26
42950	26	26
45653	26	26
56123	28	27
53240	28	27
91805	28	27
69069	28	27
77209	29	27
84099	29	27
55360	30	29
48820	30	29
76747	30	30
92951	30	30
66366	30	30

Formula di tau in caso di parità

$$\tau_b = \frac{S}{\sqrt{\left[\binom{n}{2} - \sum_{j=1}^{n_x} \binom{t_j^x}{2}\right] \left[\binom{n}{2} - \sum_{j=1}^{n_y} \binom{t_j^y}{2}\right]}}$$

dove $\begin{cases} n_x = \text{numero di gruppi di } X \text{ con parità} \\ t_j^x = \text{numero di valori uguali per la } j\text{-esima parità in } X \\ n_y = \text{numero di gruppi di } Y \text{ con parità} \\ t_j^y = \text{numero di valori uguali per la } j\text{-esima parità in } Y \\ s = \text{numero minimo di interscambi che trasforma } X \text{ in } Y \end{cases}$

$$S = \sum_{i=1}^{n-1} \sum_{j=1}^i \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) \quad \text{dove} \quad \text{sgn}(x) = \begin{cases} 1 & \text{se } x > 0 \\ 0 & \text{se } x = 0 \\ -1 & \text{se } x < 0 \end{cases}$$

Calcolo del tau-b di Kendall

Una delle graduatorie è disposta in ordine ascendente (con eventuali parità). L'altra segue per abbinamento.

Per ogni rango del secondo si contano quanti, tra quelli alla sua destra, ne sono superiori

Il totale di questi conteggi darà il valore di "S" nel numeratore del tau.

ESEMPIO	A	B	C	D	E	F	G	H	I	J
	1	2	4.5	4.5	4.5	4.5	8	8	8	10
	1	2.5	2.5	4.5	4.5	6.5	6.5	8	9.5	9.5

$$\tau_k = \frac{33}{\sqrt{[45 - (6+3)][45 - (1+1+1+1)]}} = \frac{33}{\sqrt{41*36}} = 0.859$$

Esercizio

Qualità e costo dei servizi di un resort secondo il giudizio concorde della coppia più facoltosa.

Riesprimete i giudizi in ranghi

b) Calcolate il tau di Kendall



Qualità'	Costo
3	7
3	7
3	6
1	4
3	5
4	3
2	3
2	3
6	7
6	1
6	8
7	9
9	7
8	6
5	2

Tabelle doppie ed ordinamenti

Quando la rilevazione si graduatorie si ripete per un numero elevato di i dati sono raccolti in una tabella a doppia entrata con modalità ordinate

ESEMPIO:
capacità visiva

Occhio destro	Occhio sinistro				
	1° grado	2° grado	3° grado	Inferiore	
1° grado	821	112	85	35	1053
2° grado	116	494	145	27	782
3° grado	72	151	583	87	893
Inferiore	43	34	106	331	514
	1052	791	919	480	3242

Ai fini del numeratore S del tau-b avremo contributi positivi da celle che stanno sotto e a destra di quella considerata

Inoltre, avremo contributi negativi da celle che stanno sotto e a sinistra di quella considerata

$$112(145 + 27 + 583 + 87 + 106 + 331 - 116 - 72 - 43) = 112 * 1048 = 117376$$

Esempio-Continua

Occhio destro	Occhio sinistro					
	1° grado	2° grado	3° grado	Inferiore		
1° grado	821	112	85	35	1053	553878
2° grado	116	494	145	27	782	305371
3° grado	72	151	583	87	893	398278
Inferiore	43	34	106	331	514	131841
	1052	791	919	480	3242	5253661
	552826	312445	421821	114960	5253661	3864293
						3851609
	1607518	117376	-39525	-61040		
	149872	490048	17110	-26703		0.64288695
	33912	59494	148082	-15921		
					2480223	

Per il denominatore i contributi verranno da

$$\frac{(3242)3241}{2} - \frac{(1053)1052}{2} - \frac{(782)781}{2} - \frac{(893)892}{2} - \frac{(514)513}{2} = 3864293$$

$$\frac{(3242)3241}{2} - \frac{(1052)1051}{2} - \frac{(791)790}{2} - \frac{(919)918}{2} - \frac{(480)479}{2} = 3864293$$

$$\tau_b = \frac{2480223}{\sqrt{(3864293)3851609}} = 0.643$$

Tau-c

Il tau-b ha il difetto di non raggiungere il valore massimo se la tabella è rettangolare.

In questi casi Kendall propone di usare

$$\tau_c = \frac{2S}{n^2 \frac{(m-1)}{m}} \quad \text{dove} \quad m = \text{Min}(r, s)$$

ESEMPIO: Competenza e stipendio

Competenza	Classe stipendiale					
	1°	2°	3°	4°		
1	99	84	44	40	267	35511
2	47	20	10	26	103	5253
3	59	60	55	9	183	16653
	205	164	109	75	553	152628
	20910	13366	5886	2775	152628	95211
						109691
	17820	-504	-6644	-10040	Tau-b	0.0092
	5828	100	-1100	-4524	Tau-c	0.0122
					936	
	r=	2	m=	2		
	s=	4				

Entrambi gli indici riscontrano assenza di legame.

Non vi preoccupate.
E' un esempio ipotetico

Goodman-Kruskal

E' un indice in grado di raggiungere il valore massimo anche per tabelle rettangolari

$$\gamma = \frac{N_c - N_d}{N_c + N_d} \text{ dove } \begin{cases} N_c & \text{contributi positivi} \\ N_d & \text{contributi negativi} \end{cases}$$

Per costruzione

$$-1 \leq \gamma \leq 1$$

Il numeratore è lo stesso del Tau-b.

Conviene comunque separare il calcolo delle coppie discordi e di quelle concordi

Occhio sinistro					
Occhio destro	1° grado	2° grado	3° grado	Inferiore	
1° grado	821	112	85	35	1'053
2° grado	116	494	145	27	782
3° grado	72	151	583	87	893
Inferiore	43	34	106	331	514
	1'052	791	919	480	3'242

1'607'518	143'248	37'825		
149'872	546'858	60'610	Gamma	0.7757
33'912	65'987	192'973		
			2'838'803	
	-25'872	-77'350	-61'040	
	-56'810	-43'500	-26'703	
	-6'493	-44'891	-15'921	
			-358'580	
			2'480'223	

Goodman-Kruskal/2

Competenza	Classe stipendiale				
	1°	2°	3°	4°	
1	99	84	44	40	267
2	47	20	10	26	103
3	59	60	55	9	183
	205	164	109	75	553
	17820	8400	1540	Gamma=	0.0136
	5828	1280	90		
		-8904	-8184	-10040	34958
		-1180	-1190	-4524	
					-34022

L'indice di Goodman-Kruskal esprime la riduzione dell'errore che si commette nel prevedere come una coppia di unità si ordinerà rispetto ad una variabile allorché si apprende come sia ordinata rispetto all'altra.

Essendo prossimo allo zero ciò implica che conoscendo la classe stipendiale non si può essere conclusivi rispetto alla competenza.