

Esercitazione di Statistica
Indici di associazione
28/10/2015

La relazione tra caratteri

Indipendenza logica

Quando si suppone che tra due caratteri non ci sia alcuna relazione di causa-effetto.

Indipendenza statistica

- Studio scientifico che verifica l'esistenza o meno di regolarità nell'associazione tra due caratteri
- Indipendenza è l'assenza di un qualsiasi legame tra i due caratteri
- Nello studio dell'associazione è spesso utilizzata la tabella a doppia entrata

Quale indice utilizzare?

- Connessione o indipendenza assoluta se si tratta di mutabili (Chi-quadrato di Pearson, Indice di contingenza quadratica media, l'indice V di Cramer)
- Indipendenza in media per caratteri misti (Rapporto di correlazione)
- Correlazione o associazione se abbiamo due caratteri quantitativi (Covarianza, Correlazione di Pearson)

Esercizi

1. Si vuole controllare l'effetto di un farmaco sperimentale su due gruppi di pazienti malati. Calcolare l'indice di associazione tra farmaco e stato di salute dei pazienti.

	Presentano malattia	guariti
Agente A	52	34
Agente B	65	26

Soluzione

Si calcolano le frequenze teoriche:

$$n_{ij}^* = \frac{n_i \cdot n_j}{n} = \frac{\text{tot. riga} * \text{tot. colonna}}{\text{tot generale}}$$

	Presentano malattia	guariti	
Agente A	52	34	86
Agente B	65	26	91
TOT	117	60	177

	Presentano malattia	Guariti
Agente A	$(86 * 117) / 177 = 56.85$	$(86 * 60) / 177 = 29.15$

Agente B	$(91 \cdot 117) / 177 = 60.15$	$(91 \cdot 60) / 177 = 30.85$
-----------------	--------------------------------	-------------------------------

	Presentano malattia	Guariti
Agente A	56.85	29.15
Agente B	60.15	30.85

si calcola il chi-quadrato (mutabili):

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

$$\chi^2 = \frac{(52 - 56.85)^2}{56.85} + \frac{(34 - 29.15)^2}{29.15} + \frac{(65 - 60.15)^2}{60.15} + \frac{(26 - 30.85)^2}{30.85} = 2.40$$

- Poiché il $\chi^2 \neq 0$ possiamo affermare che c'è associazione tra i due caratteri, assume valori pari a zero se vi è indipendenza e cresce al crescere dell'associazione, tuttavia tale indice dipende dalla numerosità:
- Indice di contingenza quadratica

$$\varphi^2 = \frac{\chi^2}{n} = \frac{2.40}{177} = 0.01$$

- Quando il min tra numero di righe e colonne è pari a 2 (
- $0 \leq \varphi^2 \leq \min(h-1, k-1)$ dove h=numero di righe e k=numero di colonne) tale indice assume il suo valore massimo pari a 1. In questo caso l'associazione tra i due caratteri è debole.

- Una società di marketing ha monitorato la lunghezza dei post su FB e la loro capacità di generare like o commenti di 20 marchi. In particolare ha rilevato il numero di caratteri del post e il numero di like o commenti lasciati dagli utenti. Valutare il tipo di relazione tra i due caratteri.

X	Y
112	18
160	15
70	45
45	29
152	4
120	6
82	50
162	12
80	44
83	36
79	52
40	41
111	16

74	62
36	35
84	61
102	5
15	26
89	41
69	53

Soluzione

Per verificare la presenza di una relazione lineare tra i due caratteri possiamo calcolare la covarianza:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1765}{20} = 88.25$$

$$\bar{Y} = \frac{651}{20} = 32.55$$

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = -\frac{7997.75}{20} = -399.9$$

X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
112	18	23.75	-14.55	-345.563	564.0625	211.7025
160	15	71.75	-17.55	-1259.21	5148.063	308.0025
70	45	-18.25	12.45	-227.213	333.0625	155.0025
45	29	-43.25	-3.55	153.5375	1870.563	12.6025
152	4	63.75	-28.55	-1820.06	4064.063	815.1025
120	6	31.75	-26.55	-842.963	1008.063	704.9025
82	50	-6.25	17.45	-109.063	39.0625	304.5025
162	12	73.75	-20.55	-1515.56	5439.063	422.3025
80	44	-8.25	11.45	-94.4625	68.0625	131.1025
83	36	-5.25	3.45	-18.1125	27.5625	11.9025
79	52	-9.25	19.45	-179.913	85.5625	378.3025
40	41	-48.25	8.45	-407.713	2328.063	71.4025
111	16	22.75	-16.55	-376.513	517.5625	273.9025
74	62	-14.25	29.45	-419.663	203.0625	867.3025
36	35	-52.25	2.45	-128.013	2730.063	6.0025
84	61	-4.25	28.45	-120.913	18.0625	809.4025
102	5	13.75	-27.55	-378.813	189.0625	759.0025
15	26	-73.25	-6.55	479.7875	5365.563	42.9025
89	41	0.75	8.45	6.3375	0.5625	71.4025
69	53	-19.25	20.45	-393.663	370.5625	418.2025
1765	651			-7997.75	30369.75	6774.95

Quindi esiste una dipendenza lineare ed è negativa, una misura dell'intensità di tale dipendenza può essere misurata con il coefficiente di correlazione lineare

$$r = \frac{Codev(X,Y)}{\sqrt{Dev(X)Dev(Y)}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n y_i^2 - n\bar{y}^2)}}$$

Calcoliamo le rispettive deviazioni standard:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n x_i^2 - \bar{x}^2}{n}} = \sqrt{\frac{4328.332}{5}} = 29.42$$

$$\sigma_y = \sqrt{\frac{177.57}{5}} = 5.96$$

$$r = \frac{-399.9}{38.97 * 18.4} = -0.56$$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
-22.04	-4.438	97.81352	485.7616	19.69584
3.26	1.462	4.76612	10.6276	2.137444
-16.44	-3.868	63.58992	270.2736	14.96142
-20.74	-4.238	87.89612	430.1476	17.96064
55.96	11.082	620.1487	3131.522	122.8107
TOT		874.2144	4328.332	177.5661

3. Un campione di 5 lavoratori ha fornito i dati relativi al reddito e alle tasse pagate:

Reddito	Tasse
6.9	0.69
32.2	6.59
12.5	1.26
8.2	0.89
84.9	16.21

Calcolare l'associazione tra i due caratteri e la correlazione

Soluzione

Una misura del grado di associazione tra caratteri quantitativi è rappresentata dalla covarianza

$$cov(X,Y) = \frac{Codev(X,Y)}{n} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$$

Se la covarianza è positiva esiste concordanza, vale a dire che a valori crescenti di X corrispondono valori crescenti di Y, discordanza se a valori crescenti di X corrispondono valori decrescenti di Y oppure se gli scostamenti positivi e negativi di X e Y dalle rispettive medie si compensano e la covarianza è pari a zero non esiste associazione tra i caratteri.

Gli elementi necessari per il calcolo sono:

	X	Y	X^2	Y^2	XY
	6.9	0.69	47.61	0.4761	4.761
	32.2	6.59	1036.84	43.4281	212.198
	12.5	1.26	156.25	1.5876	15.75
	8.2	0.89	67.24	0.7921	7.298
	84.9	16.21	7208.01	262.7641	1376.229
TOT	144.7	25.64	8515.95	309.048	1616.236
Media	28.94	5.128			323.2472
Var	6370.486	236.4677			
Cov	174.8429				
r	0.997189				

$$r = \frac{174.84}{29.42 * 5.96} = 0.997$$

4. Nella tabella seguente sono riportati i voti finali in Statistica (x) ed Econometria (y) conseguiti da dieci studenti, si valuti la correlazione.

Studente	1	2	3	4	5	6	7	8	9	10
x	25	28	21	25	27	30	28	27	29	30
y	19	29	20	26	25	30	29	26	28	29

Soluzione

$$r = \frac{Codev(X,Y)}{\sqrt{Dev(X)Dev(Y)}} = \frac{8.1}{2.61 * 3.65} = 0.85$$

5. Un piccolo comune di 500 abitanti vuole valutare quanta parte della variabilità del reddito dei suoi cittadini è attribuibile alla soddisfazione di un dato servizio. Le due distribuzioni sono indipendenti?

X = soddisfazione	Classi di reddito dei cittadini				
	8- 15	15- 25	25- 35	35- 55	55- 80
SI	12	24	39	142	73
NO	55	20	24	49	62

X = soddisfazione	Classi di reddito dei cittadini					
	8- 15	15- 25	25- 35	35- 55	55- 80	TOT
SI	12	24	39	142	73	290

NO	55	20	24	49	62	210
TOT	67	44	63	191	135	500

- Un carattere quantitativo Y (reddito) è dipendente in media da un carattere X (soddisfazione) se le medie condizionate di Y rispetto ad X non sono tutte uguali tra loro e non sono uguali alla media non condizionata di Y.
- Si calcolano dunque le medie condizionate di Y rispetto ad X.
- Essendo Y un carattere raggruppato in classi, dovremo individuare il valore centrale della classe da utilizzare come valore rappresentativo nel calcolo delle medie condizionate

	11.5	20	30	45	67.5	TOT
SI	(12*11.5)=138	(24*20)=480	(39*30)=1170	(142*45)=6390	(73*67.5)=4927.5	13105.5
NO	(55*11.5)=632.5	(20*20)=400	(24*30)=720	(49*45)=2205	(62*67.5)=4185	8142.5
TOT	770.5	880	1890	8595	9112.5	21248

$$M(Y|X = x_i) = \frac{1}{n_i} \sum_{j=1}^c y_j n_{ij}$$

$$M(Y|X = x_1) = \frac{1}{n_1} \sum_{j=1}^c y_j n_{1j} = \frac{13105.5}{290} = 45.19$$

$$M(Y|X = x_2) = \frac{1}{n_2} \sum_{j=1}^c y_j n_{2j} = \frac{8142.5}{210} = 38.77$$

$$M(Y) = \frac{1}{n} \sum_{j=1}^c \bar{y}_j n_j = \frac{21248}{500} = 42.50$$

Dove la media di Y si può calcolare:

(67*11.5)=770.5	(44*20)=880	(63*30)=1890	(191*45)=8595	(135*67.5)=9112.5	21248
-----------------	-------------	--------------	---------------	-------------------	--------------

$$\eta^2 = \frac{Dev(Y)_{tra}}{Dev(Y)_{TOT}} = \frac{\sum_{i=1}^r (\bar{y}|x_i - \bar{y})^2 n_i}{\sum_{j=1}^c (y_j - \bar{y})^2 n_j}$$

	$M(Y X = x_i)$	$\bar{y} x_i - \bar{y}$	$(\bar{y} x_i - \bar{y})^2$	n_i	$(\bar{y} x_i - \bar{y})^2 n_i$
$M(Y X = x_1)$	45.19	(45.19-42.5)=2.69	7.2361	290	(290*7.23)=2098.47
$M(Y X = x_2)$	38.77	(38.77-42.5)=-3.73	13.91	210	(210*13.91)=2921.1

5019.57

y_j	$y_j - \bar{y}$	$(y_j - \bar{y})^2$	n_j	$(y_j - \bar{y})^2 n_j$
11.5	-31	961	67	64387
20	-22.5	506.25	44	22275
30	-12.5	156.25	63	9843.75
45	2.5	6.25	191	1193.75
67.5	25	625	135	84375

182074.5

$$\eta^2 = \frac{5019.57}{182074.5} = 0.028$$

6. Sono stati raccolti i dati relativi al numero di turisti di strutture ricettive di varia grandezza rispetto alle condizioni climatiche. Valutare con un opportuno indice se il meteo e la dimensione delle strutture sono indipendenti.

	Piccole	Medie	Grandi
Sereno	110	120	82
Nuvoloso	43	16	65
Pioggia	12	30	49

Frequenze assolute:

	Piccole	Medie	Grandi	TOT
Sereno	110	120	82	312
Nuvoloso	43	16	65	124
Pioggia	12	30	49	91
TOT	165	166	196	527

Frequenze teoriche:

	Piccole	Medie	Grandi	TOT
Sereno	97.68	98.28	116.04	312
Nuvoloso	38.82	39.06	46.12	124
Pioggia	28.49	28.66	33.84	91
TOT	165	166	196	527

$$\chi^2 = \frac{(110 - 97.68)^2}{97.68} + \frac{(120 - 98.28)^2}{98.28} + \frac{(82 - 116.04)^2}{116.04} + \frac{(43 - 38.82)^2}{38.82} + \frac{(16 - 39.06)^2}{39.06} + \frac{(65 - 46.12)^2}{46.12} + \frac{(12 - 28.49)^2}{28.49} + \frac{(30 - 28.66)^2}{28.66} + \frac{(49 - 33.84)^2}{33.84} = 85.97$$