

Statistica - Esercizi di bivariata

22 novembre 2023

Associazione tra 2 variabili

Esercizio 1 In uno studio del 2002 è stato effettuato uno studio sulla relazione tra attitudine sul pensionamento anticipato e la soddisfazione sull'ambiente di lavoro (su scala da 0 a 10, dove 0=per nulla soddisfatto; 10=pienamente soddisfatto). la tabella doppia entrata mostra le frequenze assolute congiunte.

Attitudine pens. anticipato	Soddisfazione ambiente lavoro		
	0-3	3-7	7-10
A favore	267	152	52
Neutro	68	30	11
A sfavore	26	24	10
Total	361	206	73

- verificare se esiste una relazione (associazione) tra attitudine sul pensionamento e soddisfazione sull'ambiente di lavoro?

Soluzione (provate a risolvere prima da soli!) :

Nella tabella manca la distribuzione di frequenza marginale della variabile “attitudine sul pensionamento anticipato” (e anche il totale generale), che sarebbe:

Attitudine pens. anticipato	Frequenza
A favore	471
Neutro	109
A sfavore	60
Totale	640

la tabella di contingenza completa sarà:

Attitudine pens. anticipato	Soddisfazione ambiente lavoro			
	0-2	3-6	7-10	
A favore	267	152	52	471
Neutro	68	30	11	109
A sfavore	26	24	10	60
Totale	361	206	73	640

L'associazione tra due variabili poste in una tabella di contingenza (nel nostro caso, una qualitativa sconnessa: l'attitudine verso il pensionamento anticipato; l'altra quantitativa discreta in classi: attitudine sul pensionamento anticipato) si calcola in generale con il χ^2 (chi-quadro o x-quadro).

Il primo passaggio è calcolare le frequenze teoriche (T) dove la generica è: $T_{ij} = \frac{n_{i0}n_{0j}}{N}$

Per esempio la frequenza teorica corrispondente alla prima frequenza congiunta (incrocio tra prima riga e prima colonna, ossia posizione n_{11}) :

$$T_{11} = \frac{n_{10}n_{01}}{N} = \frac{361 \times 471}{640} = 265.67$$

La corrispondente tabella teorica (**sotto l'ipotesi di indipendenza**) sarà:

Attitudine pens. anticipato	Soddisfazione ambiente lavoro		
	0-2	3-6	7-10
A favore	265.67	151.60	53.72
Neutro	61.48	35.08	12.43
A sfavore	33.84	19.31	6.84

Per effettuare agevolmente i calcoli per il χ^2 si può usare il seguente schema:

freq. osservate O	freq. teoriche (T)	$O-T$	$(O-T)^2$	$\frac{(O-T)^2}{T}$
267	265.67	1.33	1.769	0.006
152	151.6	0.4	0.16	0.001
52	53.72	-1.72	2.95	0.05
68	61.48	6.52	42.51	0.69
30	35.08	-5.08	25.81	0.73
11	12.43	-1.43	2.04	0.18
26	33.84	-7.84	61.47	1.82
24	19.31	4.69	22	1.14
10	6.84	3.16	9.99	1.46
Total				$\chi^2 = 6.07$

Il corrispondente chi-quadro normalizzato è:

$$\bar{\chi}^2 = \sqrt{\frac{\chi^2}{N \times \min(\text{righe} - 1; \text{colonne} - 1)}} = \sqrt{\frac{6.07}{640 \times \min(3 - 1; 3 - 1)}} = \sqrt{\frac{6.07}{640 \times 2}} = 0.07$$

che è un valore molto vicino a 0 (cosa che si vede anche dall'entità degli scarti $\frac{(O-T)^2}{T}$). La risposta finale è: non possiamo concludere che vi sia associazione tra attitudine verso il pensionamento anticipato e soddisfazione sull'ambiente di lavoro.

Una regola per l'interpretazione del chi-quadro normalizzato è che già per un valore pari a 0.4, si può parlare di associazione tra le variabili poste in una tabella a doppia entrata.

Esercizio 2 Dato un diagramma di dispersione tra due variabili X e Y individuare i valori per ciascuna osservazione e riempire i valori mancanti. Attenzione: non si tratta di frequenze ma coppie di valori delle due variabili per ciascuna osservazione (in riga).

X	Y
1	4
-	3
3	-
-	1
-	-

Soluzione (provate a risolvere prima da soli!):

X	Y
1	4
2	3
3	1
4	1
4	2

Esercizio 3 Nella seguente matrice dati sono riportati i voti per $N = 8$ studenti agli esami di matematica (variabile X) e statistica (variabile Y).

Student	Matematica (X)	Statistica (Y)
1	24	23
2	27	28
3	30	30
4	26	27
5	29	30
6	18	20
7	21	20
8	22	25

Si vuole misurare la relazione (lineare) tra le due variabili X e Y.

Soluzione (provate a risolvere prima da soli!): Per valutare la relazione lineare tra due variabili quantitative occorre calcolare la **covarianza** (σ_{XY}) e il suo corrispondente normalizzato (tra -1 e +1), ossia il **coefficiente di correlazione** (r). Per fare ciò ci servono anzitutto le medie delle due variabili: voti in matematica (variabile X) e voti in statistica (variabile Y).

La media dei voti in matematica è:

$$\bar{X} = \frac{\sum_{i=1}^8 x_i}{n} = \frac{197}{8} = 24.625$$

La media dei voti in statistica è:

$$\bar{Y} = \frac{\sum_{i=1}^8 y_i}{n} = \frac{203}{8} = 25.375$$

Per il calcolo della covarianza si usi il seguente schema:

	x_i	y_i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$x_i - \bar{X} \times y_i - \bar{Y}$
1	24	23	-0.62	-2.38	1.48
2	27	28	2.38	2.62	6.23
3	30	30	5.38	4.62	24.86
4	26	27	1.38	1.62	2.23
5	29	30	4.38	4.62	20.23
6	18	20	-6.62	-5.38	35.61
7	21	20	-3.62	-5.38	19.48
8	22	25	-2.62	-0.38	0.98
<i>TOT</i>					111.125

la covarianza pertanto sarà:

$$\sigma_{XY} = \frac{\sum_{i=1}^8 (x_i - \bar{X})(y_i - \bar{Y})}{8} = \frac{111.125}{8} = 13.89$$

(quindi basta dividere la somma dell'ultima colonna dello schema per N).

La covarianza è positiva e per adesso possiamo concludere che vi è associazione positiva tra le due variabili.

In altre parole chi prende alti voti a matematica ottiene alti voti anche a statistica (e viceversa).

Per il calcolo della covarianza si consideri anche la seguente formula alternativa:

$$\sigma_{XY} = \frac{\sum_{i=1}^8 x_i y_i}{8} - \bar{X} \bar{Y}$$

	x_i	y_i	$x_i y_i$
1	24	23	552
2	27	28	756
3	30	30	900
4	26	27	702
5	29	30	870
6	18	20	360
7	21	20	420
8	22	25	550
<i>TOT</i>			5110

$$\sigma_{XY} = \frac{5110}{8} - (24.625 \cdot 25.375) = 638.75 - 624.86 = 13.89$$

Sembra conveniente utilizzare questa seconda formula.

Ma quanto forte è la relazione tra le due variabili?

Per rispondere a tale domanda ci serve il **coefficiente di correlazione**, per il cui calcolo abbiamo bisogno delle **deviazioni standard** delle due variabili (le radici quadrate della varianza!). Usando la formula computazionale della varianza (ci serve la media quadratica!) otteniamo facendo i calcoli:

$$\sigma_X = 3.87$$

$$\sigma_Y = 3.805$$

Il coefficiente di correlazione risulta:

$$r = \frac{\sigma_{XY}}{\sigma_x \sigma_y} = \frac{13.89}{3.87 \times 3.805} = 0.943$$

Dato che r varia tra -1 (perfetta relazione negativa) e $+1$ (perfetta relazione positiva), questo suggerisce che non solo la relazione tra X e Y è positiva ma è anche quasi perfetta (quindi molto forte). Verificate questa affermazione disegnando un diagramma di dispersione tra X e Y (cosa dovreste aspettarvi?)