Statistica Sociale

29/03/2022

Indipendenza statistica

Attraverso le tabelle di contingenza è possibile studiare l'eventuale dipendenza di una variabile dall'altra

È possibile verificare ciò attraverso la presenza o meno di indipendenza. Se due variabili categoriali non sono indipendenti, allora sono associate

Due variabili categoriali sono statisticamente indipendenti se nella popolazione le distribuzioni condizionate di una rispetto a ciascuna categoria dell'altra sono identiche

Due variabili categoriali sono statisticamente dipendenti se nella popolazione le distribuzioni condizionate di una rispetto a ciascuna categoria dell'altra non sono identiche

Molto utile, nella fase iniziale di una ricerca, è la trasformazione delle frequenze in percentuali, per agevolare la comprensione.

	Orientamento Politico			
Gruppo Etnico	Dem	Indip	Repub	Totale
Bianchi	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Neri	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Ispanici	110 (44%)	35 (14%)	105 (42%)	250 (100%)

Che tipo di distribuzioni sono?

Due variabili categoriali sono statisticamente indipendenti se nella popolazione le distribuzioni condizionate di una rispetto a ciascuna categoria dell'altra sono identiche

	Orientamento Politico				
Sesso	Dem	Indip	Repub	Totale	*n*
Femmine	38%	34%	28%	100%	1511
Maschi	31%	38%	32%	101%	1260

	Orientamento Politico			
Gruppo Etnico	Dem	Indip	Repub	Totale
Bianchi	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Neri	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Ispanici	110 (44%)	35 (14%)	105 (42%)	250 (100%)

L'indipendenza statistica è una proprietà simmetrica per le due variabili: se le distribuzioni condizionate per ogni riga sono identiche, sono identiche anche quelle per colonna

- Ad esempio per l'orientamento ed il gruppo etnico le distribuzioni condizionate del gruppo etnico rispetto alle modalità dell'orientamento politico sono:

- Dall'evidenza empirica, l'orientamento politico dipende dal sesso perché le distribuzioni condizionate percentuali sono diverse
- Nel campione osservato, l'orientamento politico non dipende dal gruppo etnico perché le distribuzioni condizionate percentuali sono uguali

Test chi-quadro di indipendenza

Il concetto di indipendenza statistica è riferito alla popolazione, in genere noi disponiamo di dati campionari. Le distribuzioni condizionate campionarie possono essere diverse pur essendo le variabili indipendenti a livello di popolazione

Per verificare statisticamente la reale esistenza di indipendenza tra due variabili categoriali (a livello di popolazione), possiamo applicare il test chi-quadro per l'indipendenza

Le ipotesi saranno:

- H_0: le variabili sono statisticamente indipendenti.
- H_a: le variabili sono statisticamente dipendenti.

Requisiti minimi per l'applicazione del test: campionamento casuale o esperimento randomizzato e campione sufficientemente grande.

Frequenze attese per l'indipendenza

Il test del chi-quadro si basa sul confronto tra frequenze osservate e frequenze attese

La frequenza attesa n'_{ij} è quella che potremmo osservare in presenza di indipendenza tra le due variabili, corrisponde cioè alla numerosità attesa in una cella se le due variabili sono indipendenti:

$$n'_{ij} = \frac{n_{i.} n_{.j}}{n} = \frac{totale\ di\ riga\ per\ la\ modalità\ i\ *\ totale\ di\ colonna\ per\ la\ modalità\ j}{numerosità\ campionaria\ totale}$$

Nel caso di indipendenza tra sesso e orientamento politico avremo:

$$n'_{11} = \frac{1511 * 959}{2771} = 522{,}9$$

$$n'_{12} = \frac{1511 * 991}{2771} = 540{,}4$$

$$n'_{13} = \frac{1511 * 821}{2771} = 447{,}7$$

$$n'_{21} = \frac{1260 * 959}{2771} = 436{,}1$$

$$n'_{22} = \frac{1260 * 991}{2771} = 450{,}6$$

$$n'_{23} = \frac{1260 * 821}{2771} = 373{,}3$$

	Orientamento Politico			
Sesso	Dem	Indip	Repub	Totale
Donne	573 (522.9)	516 (540.4)	422 (447.7)	1511
Uomini	386 (436.1)	475 (450.6)	399 (373.3)	1260
Totale	959	991	821	2771

Se c'è indipendenza, ci aspettiamo che le frequenze osservate siano «vicine» a quelle attese

Calcoliamo le differenze tra le frequenze osservate e quelle attese:

	Orientamento Politico			
Sesso	Dem	Indip	Repub	Totale
Donne	573 (522.9)	516 (540.4)	422 (447.7)	1511
Uomini	386 (436.1)	475 (450.6)	399 (373.3)	1260
Totale	959	991	821	2771

$$n_{11} - n'_{11} = 573 - 522{,}9 = 50{,}1$$
$$n_{12} - n'_{12} = 516 - 540{,}4 = -24{,}4$$
$$n_{13} - n'_{13} = 422 - 447{,}7 = -25{,}7$$

$$n_{21} - n'_{21} = 386 - 436{,}1 = -50{,}1$$
$$n_{22} - n'_{22} = 475 - 450{,}6 = 24{,}4$$
$$n_{23} - n'_{23} = 399 - 373{,}3 = 25{,}7$$

Riusciamo a valutare se le differenze sono grandi o piccole?

La statistica test del chi-quadro

Poichè in H_0 si è ipotizzata l'indipendenza tra le due variabili, il test statistico viene costruito con l'intenzione di evidenziare l'allontanamento da H_0

Si basa sulle differenze tra frequenze osservate e frequenze attese

La statistica test è la statistica chi-quadro χ^2 data da

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Sommiamo per ogni cella, il rapporto tra differenza al quadrato tra frequenza attesa e osservata e la frequenza attesa

- Se $\chi^2 = 0$ le due variabili sono indipendenti (applicando il test a dati campionari, può bastare che sia sufficientemente piccolo)
- Al crescere del valore di χ^2 aumenta l'evidenza contro H_0
- χ^2 non può essere negativo

Questo valore viene confrontato con i valori della distribuzione teorica sotto l'ipotesi di indipendenza

Si ottiene il p-value, un valore che misura quanto i dati osservati sono compatibili con l'ipotesi di indipendenza

Valori soglia per il p-value sono in genere: 0,1; 0,05;0,01

$$n_{11} - n'_{11} = 573 - 522{,}9 = 50{,}1$$
$$n_{12} - n'_{12} = 516 - 540{,}4 = -24{,}4$$
$$n_{13} - n'_{13} = 422 - 447{,}7 = -25{,}7$$
$$n_{21} - n'_{21} = 386 - 436{,}1 = -50{,}1$$
$$n_{22} - n'_{22} = 475 - 450{,}6 = 24{,}4$$
$$n_{23} - n'_{23} = 399 - 373{,}3 = 25{,}7$$

$$\chi^2 = \left(\frac{50{,}1^2}{522{,}9} + \frac{(-24{,}4)^2}{540{,}4} + \cdots + \frac{24{,}4^2}{450{,}6} + \frac{25{,}7^2}{373{,}3}\right) = 4{,}8 + \cdots + 1{,}8 = 16{,}2$$

Il p-value in questo caso è pari a 0,0003

Concludiamo che c'è una forte evidenza empirica contro l'ipotesi di indipendenza H_0, quindi le due variabili sesso e orientamento politico sembrano essere associate nella popolazione

Se le variabili fossero indipendenti, dovrebbe essere davvero inusuale per un campione casuale avere un valore della statistica χ^2 così elevato

Chi-quadro e associazione

Il test chi-quadro risponde alla domanda «C'è associazione?»

Esistono misure di associazione che sintetizzano la forza di dipendenza tra due variabili

Nel caso A della tabella vediamo un caso di indipendenza

Nel caso B vediamo un caso di dipendenza forte

Caso A		Opinione			Caso B	Opinione		
	Razze	Favorevole	Contrario	Totale		Favorevole	Contrario	Totale
	Bianchi	360	240	600		600	0	600
	Neri	240	160	400		0	400	400
	Totale	600	400	1000		600	400	1000

La statistica chi-quadro indica quanta evidenza c'è a favore della dipendenza, non ne misura la forza (valori più grandi si verificano quando la numerosità è grande)

	A			B			C		
	Sì	**No**	**Totale**	**Sì**	**No**	**Totale**	**Sì**	**No**	**Totale**
Bianchi	49	51	100	98	102	200	4900	5100	10 000
Neri	51	49	100	102	98	200	5100	4900	10 000
	100	100	200	200	200	400	10 000	10 000	20 000
	$\chi^2 = 0.08$ P-valore $= 0.78$			$\chi^2 = 0.16$ P-valore $= 0.69$			$\chi^2 = 8.0$ P-valore $= 0.005$		