Esercitazione 11

ESERCIZIO 1

La seguente tabella riporta la distribuzione di 300 studenti, laureati in una sessione di una scuola di Economia e Statistica, classificati secondo la materia argomento della tesi magistrale (M) e il genere (G).

M\G	Maschio	Femmina	TOTALE
Contabilità	37	67	104
Ec. Aziendale	42	48	90
Ec. Politica	3	8	11
Finanza	56	30	86
Met. Quantitativi	2	7	9
TOTALE	140	160	300

1. Si determini la moda del carattere M e se ne discuta la rappresentatività.
2. Si calcolino le distribuzioni parziali di frequenze relative del carattere M.
3. Si calcolino le contingenze assolute e si commentino quelle della prima colonna.
4. Si misuri la connessione tra M e G mediante un indice basato su un'opportuna media quadratica delle contingenze relative e mediante un indice basato sulle medie aritmetiche delle contingenze.

SOLUZIONE

La moda del carattere M (materia argomento della tesi magistrale) è la modalità la cui frequenza è massima. Consideriamo dunque la distribuzione marginale del carattere M

M	$n_{i\cdot}$
Contabilità	104
Ec. Aziendale	90
Ec. Politica	11
Finanza	86
Met. Quantitativi	9

La moda è la modalità "Contabilità". Essendo che la frequenza marginale relativa, per la modalità "Contabilità" è pari a

$$f_{1\cdot} = \frac{n_{1\cdot}}{n} = \frac{104}{300} = 0,347$$

e che $0,347 < 0,5$ allora la moda non è rappresentativa.
Le distribuzioni parziali di frequenze relative del carattere M si ottengono calcolando le distribuzioni di frequenze relative, fissata una modalità del carattere G, quindi ogni valore è ottenuto come

$$f_{i|j} = \frac{n_{ij}}{n_{\cdot j}}$$

M\G	Maschio	Femmina	TOTALE
Contabilità	0,264	0,419	0,347
Ec. Aziendale	0,300	0,300	0,300
Ec. Politica	0,021	0,050	0,037
Finanza	0,400	0,188	0,287
Met. Quantitativi	0,014	0,044	0,030
TOTALE	1,000	1,000	1,000

Per calcolare le contingenze dobbiamo prima calcolare le frequenze teoriche. Se due caratteri sono indipendenti, le frequenze relative congiunte sono uguali al prodotto delle frequenze marginali. Le frequenze teoriche assolute si ottengono come

$$\hat{n}_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

quindi per esempio:

$$\hat{n}_{11} = \frac{n_{1\cdot} \times n_{\cdot 1}}{n} = \frac{104 \times 140}{300} = 48,533$$

$$\hat{n}_{12} = \frac{n_{1\cdot} \times n_{\cdot 2}}{n} = \frac{104 \times 160}{300} = 55,467$$

$$\hat{n}_{21} = \frac{n_{2\cdot} \times n_{\cdot 1}}{n} = \frac{90 \times 140}{300} = 42$$

M\G	Maschio	Femmina	TOTALE
Contabilità	48,533	55,467	104
Ec. Aziendale	42,000	48,000	90
Ec. Politica	5,133	5,867	11
Finanza	40,133	45,867	86
Met. Quantitativi	4,200	4,800	9
TOTALE	140	160	300

Le contingenze sono la differenza tra la frequenza osservata e la frequenza teorica (con segno).

M\G	Maschio	Femmina	TOTALE
Contabilità	-11,533	11,533	0
Ec. Aziendale	0,000	0,000	0
Ec. Politica	-2,133	2,133	0
Finanza	15,867	-15,867	0
Met. Quantitativi	-2,200	2,200	0
TOTALE	0	0	0

Relativamente alla prima colonna: la frequenza osservata dei maschi che hanno fatto una tesi nel tema della contabilità è minore di 11,533 osservazioni rispetto al valore atteso in situazione di indipendenza. Il numero di maschi che hanno deciso di scrivere una tesi nell'ambito dell'economia aziendale è pari al numero atteso in situazione di indipendenza....
Un indice opportuno per il misurare la connessione tra M e G, basato su una media quadratica, è l'indice di connessione di Pearson. Siano $\rho_{ij} = \frac{c_{ij}}{\hat{n}_{ij}} = \frac{(n_{ij} - \hat{n}_{ij})}{\hat{n}_{ij}}$ le contingenze relative, l'indice di connessione di Pearson è la media quadratica del valore assoluto di tali quantità.

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|\rho_{ij}|^2 \times \hat{n}_{ij}} = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}}$$

avendo a disposizione le contingenze assolute, usiamo la seconda formula, ed otteniamo

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}} = \sqrt{\frac{1}{300} \times 20,723} = \sqrt{0,069} = 0,263$$

Le frequenze osservate differiscono da quelle teoriche in media (quadratica) del 26,3%.

Un indice opportuno per il misurare la connessione tra M e G, basato su una media aritmetica, è l'indice di connessione di Mortara, ovvero la media aritmetica del valore assoluto delle contingenze relative

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|\rho_{ij}| \times \hat{n}_{ij} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}|$$

avendo a disposizione le contingenze assolute, usiamo la seconda formula, ed otteniamo

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}| = \frac{1}{300} \times (63,467) = 0,212$$

Le frequenze osservate differiscono da quelle teoriche in media (aritmetica) del 21,2%.

ESERCIZIO 2

La seguente tabella riporta le distribuzione di frequenze di 100 famiglie classificate secondo il titolo di godimento dell'abitazione (Y) e il numero di figli (X):

X\Y	Proprietà	Affitto	Altro	TOTALE
0	5	15	0	20
1	10	5	15	30
2	5	15	15	35
3	5	10	0	15
TOTALE	25	45	30	100

1. Si stabilisca, giustificando la risposta, se fra i due caratteri considerati esiste indipendenza distributiva. In caso di risposta negativa, si costruisca la tabella di frequenze congiunte teoriche in modo tale che i due caratteri risultino indipendenti.
2. Si fornisca un indice che misuri il grado di connessione tra X e Y, commentando il risultato.
3. In relazione alla natura di X e Y, si analizzi, giustificando la scelta, la dipendenza in media che si ritiene più idonea e se ne misuri l'intensità con un opportuno indice commentando il risultato ottenuto.

SOLUZIONE

Per vedere se tra i caratteri esiste indipendenza distributiva senza calcoli, controlliamo le distribuzioni parziali. Non sono proporzionali tra loro, quindi non c'è indipendenza distributiva.
Calcoliamo le frequenze relative teoriche, come il prodotto delle frequenze marginali (corrette per la numerosità):

$$\hat{n}_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

otteniamo

X\Y	Proprietà	Affitto	Altro	TOTALE
0	5	9	6	20
1	7,5	13,5	9	30
2	8,75	15,75	10,5	35
3	3,75	6,75	4,5	15
TOTALE	25	45	30	100

Per calcolare la connessione tra X ed Y dobbiamo costruire la tabella di contingenze, dove ogni valore corrisponde a $c_{ij} = (n_{ij} - \hat{n}_{ij})$ e rappresenta quanto si discostano le frequenze osservate dalla situazione di indipendenza.

X\Y	Proprietà	Affitto	Altro	TOTALE
0	0	6	-6	0
1	2,5	-8,5	6	0
2	-3,75	-0,75	4,5	0
3	1,25	3,25	-4,5	0
TOTALE	0	0	0	0

Un indice opportuno per il misurare la connessione tra X e Y, basato su una media aritmetica, è l'indice di connessione di Mortara, ovvero la media aritmetica del valore assoluto delle contingenze relative

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|\rho_{ij}| \times \hat{n}_{ij} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}|$$

avendo a disposizione le contingenze assolute, usiamo la seconda formula, ed otteniamo

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}| = \frac{47}{100} = 0,47$$

in media aritmetica le frequenze osservate differiscono (in valore assoluto) del 47% dalle frequenze teoriche. Si poteva decidere di misurare il grado di dipendenza applicando l'indice di connessione di Pearson

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|\rho_{ij}|^2 \times \hat{n}_{ij}} = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}}$$

Le contingenze al quadrato, diviso le frequenze teoriche, sono

X\Y	Proprietà	Affitto	Altro
0	0,000	4,000	6,000
1	0,833	5,352	4,000
2	1,607	0,036	1,929
3	0,417	1,565	4,500

avendo a disposizione le contingenze assolute, usiamo la seconda formula, ed otteniamo

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}} = \sqrt{\frac{1}{100} \times 30,238} = \sqrt{0,302} = 0,550$$

Non possiamo calcolare le medie parziali della variabile Y, in quanto qualitativa. Possiamo studiare se X è indipendente in media da Y. Calcoliamo le medie e le varianze delle distribuzioni parziali di X. Per Y uguale a "Proprietà" abbiamo

x_i	n_{i1}	$x_i \times n_{i1}$	$x_i^2 \times n_{i1}$
0	5	0	0
1	10	10	10
2	5	10	20
3	5	15	45

Otteniamo quindi che la media e la varianza della distribuzione parziale, per Y uguale a "Proprietà", sono

$$\bar{x}_1 = \frac{1}{n_{\cdot 1}}\sum_{i=1}^{4} x_i \times n_{i1} = \frac{35}{25} = 1,4$$

$$\overline{x_1^2} = \frac{1}{n_{\cdot 1}} \sum_{i=1}^{4} x_i^2 \times n_{i1} = \frac{75}{25} = 3$$

$$\sigma_1^2 = \overline{x_1^2} - \overline{x}_1^2 = 1,04$$

Analogamente otteniamo per Y uguale a "Affitto"

$$\bar{x}_2 = \frac{1}{n_{\cdot 2}} \sum_{i=1}^{4} x_i \times n_{i2} = \frac{65}{45} = 1,444$$

$$\overline{x_2^2} = \frac{1}{n_{\cdot 2}} \sum_{i=1}^{4} x_i^2 \times n_{i2} = \frac{155}{45} = 3,444$$

$$\sigma_2^2 = \overline{x_2^2} - \overline{x}_2^2 = 1,358$$

e per Y uguale a "Altro"

$$\bar{x}_3 = \frac{1}{n_{\cdot 3}} \sum_{i=1}^{4} x_i \times n_{i3} = \frac{45}{30} = 1,5$$

$$\overline{x_3^2} = \frac{1}{n_{\cdot 3}} \sum_{i=1}^{4} x_i^2 \times n_{i3} = \frac{75}{30} = 2,5$$

$$\sigma_3^2 = \overline{x_3^2} - \overline{x}_3^2 = 0,25$$

Per la distribuzione marginale abbiamo

x_i	$n_{i\cdot}$	$x_i \times n_{i\cdot}$	$x_i^2 \times n_{i\cdot}$
0	20	0	0
1	30	30	30
2	35	70	140
3	15	45	135

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{4} x_i \times n_{i\cdot} = \frac{145}{100} = 1,45$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{4} x_i^2 \times n_{i\cdot} = \frac{305}{100} = 3,05$$

$$\sigma^2 = \overline{x^2} - \overline{x}_1^2 = 3,05 - 1,45^2 = 0,9475$$

La devianza totale è pari a $DT = n \times \sigma^2 = 94,75$. La devianza fra i gruppi è

$$DF = \sum_{j=1}^{3} (\bar{x}_j - \bar{x})^2 \times n_{\cdot j} = 0,139$$

quindi possiamo calcolare il rapporto di correlazione di Pearson come

$$\eta^2_{(X|Y)} = \frac{DF}{DT} = \frac{0,139}{94,75} = 0,00146$$

Il rapporto è molto vicino al valore 0, vi è indipendenza in media.