Esercitazione 12

ESERCIZIO 1

La seguente tabella riporta la distribuzione di 100 persone, suddivisi per classi d'età (Y) e a cui è stato chiesto quanti cellulari posseggono

X\Y	Giovane	Adulto	Anziano	TOTALE
0	0	0	20	20
1	35	0	0	35
2	0	45	0	45
TOTALE	35	45	20	100

1. Si stabilisca, giustificando la risposta, se fra i due caratteri considerati esiste indipendenza distributiva.
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

Tra i due caratteri considerati non esiste indipendenza distributiva. Siamo in un caso di dipendenza perfetta. Per misurare la connessione calcoliamo le frequenze teoriche

X\Y	Giovane	Adulto	Anziano	TOTALE
0	7	9	4	20
1	12,25	15,75	7	35
2	15,75	20,25	9	45
TOTALE	35	45	20	100

e le contingenze

X\Y	Giovane	Adulto	Anziano	TOTALE
0	-7	-9	16	0
1	22,75	-15,75	-7	0
2	-15,75	24,75	-9	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

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}| = \frac{127}{100} = 1,27$$

Si poteva decidere di misurare il grado di dipendenza applicando l'indice di connessione di Pearson

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

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

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}} = \sqrt{\frac{1}{100}\times 200} = \sqrt{2} = 1,414$$

Non possiamo calcolare le medie parziali della variabile Y, in quanto qualitativa. Possiamo studiare se X è indipendente in media da Y. Per calcolare l'intensità della dipendenza in media calcoliamo media e varianza della distribuzione marginale di X

x_i	$n_{i\cdot}$	$x_i \times n_{i\cdot}$	$x_i^2 \times n_{i\cdot}$
0	20	0	0
1	35	35	35
2	45	90	180

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{3} x_i \times n_{i\cdot} = \frac{125}{100} = 1,25$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{3} x_i^2 \times n_{i\cdot} = \frac{215}{100} = 2,15$$

$$\sigma^2 = \overline{x^2} - \overline{x}_1^2 = 3,05 - 1,45^2 = 0,5875$$

la devianza totale è pari a $DT = n \times \sigma^2 = 58,75$. La devianza fra i gruppi è

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

essendo le distribuzioni parziali concentrate tutte in una modalità, abbiamo che la devianza fra i gruppi coincide con la devianza totale, e quindi

$$\eta^2_{(X|Y)} = \frac{DF}{DT} = \frac{58,75}{58,75} = 1$$

ESERCIZIO 2

La seguente tabella riporta la distribuzione di 100 lavoratori, sui quali viene rilevato il reddito medio mensile X, espresso in migliaia di euro, ed il numero di week end dedicati a viaggiare mediamente in un mese Y.

X\Y	0	1	2	3	4	TOTALE
1	42	12	0	0	0	54
2	4	12	3	0	0	19
3	1	6	5	0	0	12
4	0	7	8	0	0	15
TOTALE	47	37	16	0	0	100

1. Si fornisca un indice che misuri il grado di connessione tra X e Y, commentando il risultato.
2. Misurare la dipendenza in media di X da Y mediante un indice opportuno.
3. Si determinino i parametri della retta di regressione che spiega Y in funzione di X e si commenti il valore del coefficiente angolare della retta trovata.
4. Si calcoli il coefficiente di correlazione. Si calcoli inoltre l'indice di determinazione lineare.

SOLUZIONE

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	0	1	2	TOTALE
1	16,62	-7,98	-8,64	54
2	-4,93	4,97	-0,04	19
3	-4,64	1,56	3,08	12
4	-7,05	1,45	5,6	15
TOTALE	47	37	16	100

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

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}| = \frac{66,56}{100} = 0,666$$

in media aritmetica le frequenze osservate differiscono (in valore assoluto) del 66% 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	0	1	2
1	10,884	3,187	8,640
2	2,722	3,514	0,001
3	3,817	0,548	4,941
4	7,050	0,379	13,067

usando la seconda formula otteniamo

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}} = \sqrt{\frac{1}{100} \times 58,748} = \sqrt{0,587} = 0,766$$

in media quadratica le frequenze osservate differisconodel 76% dalle frequenze teoriche.
Per calcolare la dipendenza in media di X da Y, abbiamo bisogno delle medie parziali di X, dato un valore di Y.

$$\overline{x}_1 = M_1^{X|Y=0} = \frac{1}{n_{\cdot 1}}\sum_{i=1}^{k_X} n_{i1}x_i = \frac{1}{47}(53) = 1,128$$

$$\overline{x}_2 = M_1^{X|Y=1} = \frac{1}{n_{\cdot 2}}\sum_{i=1}^{k_X} n_{i2}x_i = \frac{1}{37}(82) = 2,216$$

$$\overline{x}_3 = M_1^{X|Y=2} = \frac{1}{n_{\cdot 3}} \sum_{i=1}^{k_X} n_{i3} x_i = \frac{1}{16} (53) = 3,313$$

La media marginale è

$$\overline{x} = M_1 = \frac{1}{n} \sum_{i=1}^{k_X} n_{i\cdot} x_i = \frac{1}{100} (188) = 1,88$$

La devianza fra i gruppi è pari a

$$DF = \sum_{j=1}^{k_Y} (\bar{x}_j - \bar{x})^2 \times n_{\cdot j} = 63,618$$

La devianza totale corrisponde alla devianza della variabile X, ovvero

$$DT = \sum_{i=1}^{k_X} (x_i - \bar{x})^2 \times n_{i\cdot} = 124,56$$

quindi possiamo calcolare il rapporto di correlazione di Pearson come

$$\eta^2_{(X|Y)} = \frac{DF}{DT} = \frac{63,618}{124,56} = 0,511$$

Dobbiamo stimare una retta di regressione

$$y = \alpha_0 + \alpha_1 x$$

La cui soluzione è data dalle equazioni di stima

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)}$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x}$$

Il valore $\overline{x}$ è stato calcolato precedentemente, la varianza di X possiamo ottenerla partendo dalla devianza totale calcolata precedentemente:

$$\sigma^2_X = \frac{1}{100} DT = 1,246$$

Per Y abbiamo

$$M_1^Y = \frac{1}{100} \sum_{j=1}^{k_Y} n_{\cdot j} \times y_j = 0,69$$

$$M_2^Y = \frac{1}{100} \sum_{j=1}^{k_Y} n_{\cdot j} \times y_j^2 = 1,01$$

$$\sigma^2_Y = M_2^Y - \left[M_1^Y\right]^2 = 1,01 - 0,69^2 = 0,534$$

Per il calcolo della covarianza abbiamo che

$x_i \times y_j \times n_{ij}$	0	1	2
1	0	12000	0
2	0	24000	12000
3	0	18000	30000
4	0	28000	64000

La covarianza può essere calcolata come

$$M_1^{XY} = \frac{1}{n} \sum_{i=1}^{k_X} \sum_{j=1}^{k_Y} x_i \times y_j \times n_{ij} = \frac{1}{100} 188 = 1,88$$

$$cov(X,Y) = M_1^{XY} - M_1^X \times M_1^Y = 1,88 - (1,88 \times 0,69) = 0,523$$

La retta può essere quindi stimata come

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)} = \frac{0,523}{1,246} = 0,468$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x} = 0,69 - 0,468 \times 1,88 = -4,402$$

Il coefficiente di correlazione di Pearson è pari a

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{0,523}{1,246 \times 0,534} = 0,801$$

C'è alta correlazione tra le due variabili ($-1 \leq r \leq 1$). Possiamo calcolare l'indice di determinazione lineare partendo da r

$$I_d^2 = r^2 = \left[\frac{cov(X,Y)}{\sigma_X \sigma_Y} \right]^2 = 0,801^2 = 0,642$$

Con il modello spieghiamo il 64,2% della variabilità del fenomeno

ESERCIZIO 3

La seguente tabella riporta la distribuzione di 100 persone, sui quali viene rilevato l'essere fumatore (3 modalità: non fumatore - ex fumatore - fumatore) ed il sesso.

X\Y	Uomo	Donna	TOTALE
Non fumatore	20	25	45
Ex fumatore	25	5	30
Fumatore	10	15	25
TOTALE	55	45	100

1. Calcolare le distribuzioni di frequenze relative parziali di X e di Y
2. Si calcolino le contingenze assolute e si commentino quelle della prima colonna.
3. Si misuri la connessione tra X e Y mediante un indice basato su un'opportuna media quadratica delle contingenze relative e mediante un indice basato sulle medie aritmetiche delle contingenze.

SOLUZIONE

Le distribuzioni parziali di frequenze relative del carattere X si ottengono calcolando le distribuzioni di frequenze relative, fissata una modalità del carattere Y, quindi ogni valore è ottenuto come

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

ed otteniamo

X\Y	Uomo	Donna	TOTALE
Non fumatore	0,364	0,556	0,450
Ex fumatore	0,455	0,111	0,300
Fumatore	0,182	0,333	0,250
TOTALE	1,000	1,000	1,000

Le distribuzioni parziali di frequenze relative del carattere Y si ottengono calcolando le distribuzioni di frequenze relative, fissata una modalità del carattere X, quindi ogni valore è ottenuto come

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

ed otteniamo

X\Y	Uomo	Donna	TOTALE
Non fumatore	0,444	0,556	1,000
Ex fumatore	0,833	0,167	1,000
Fumatore	0,400	0,600	1,000
TOTALE	0,550	0,450	1,000

Le frequenze teoriche sono

X\Y	Uomo	Donna	TOTALE
Non fumatore	24,75	20,25	45
Ex fumatore	16,5	13,5	30
Fumatore	13,75	11,25	25
TOTALE	55	45	100

Le contingenze sono definite come la differenza dalle

X\Y	Uomo	Donna	TOTALE
Non fumatore	-4,75	4,75	0
Ex fumatore	8,5	-8,5	0
Fumatore	-3,75	3,75	0
TOTALE	0	0	0

La prima colonna ci dice che abbiamo osservato 4,75 non fumatori uomini in meno rispetto alla situazione di indipendenza, 8,5 ex fumatori uomini in più rispetto alla situazione di indipendenza e 3,75 fumatori uomini in meno rispetto alla situazione di indipendenza.
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

$$M_1(|\rho|) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}|c_{ij}| = \frac{34}{100} = 0,34$$

in media aritmetica le frequenze osservate differiscono (in valore assoluto) del 34% 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	Uomo	Donna
Non fumatore	0,912	1,114
Ex fumatore	4,379	5,352
Fumatore	1,023	1,250

usando la seconda formula otteniamo

$$M_2(|\rho|) = \sqrt{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{c_{ij}^2}{\hat{n}_{ij}}} = \sqrt{\frac{1}{100} \times 14,029} = \sqrt{0,140} = 0,375$$

in media quadratica le frequenze osservate differisconodel 37% dalle frequenze teoriche.