Esercitazione 6

A. Iodice

Esercitazione 6

Statistica

Alfonso Iodice D'Enza

iodicede@unicas.it

Università degli studi di Cassino

Outline

Misura del legame

Data una variabile doppia (X, Y), la misura del legame che caratterizza le componenti X ed Y si definisce

- connessione se X e Y sono mutabili
- correlazione se X e Y sono variabili

Interdipendenza e dipendenza

Se le componenti di una variabile doppia (X, Y) oggetto di studio rivestono lo stesso ruolo ai fini dell'analisi si studia l'interdipendenza tra X e Y. Se si vuole studiare, invece, l'andamento della variabile Y rispetto ad X, si farà riferimento alla dipendenza di Y da X.

- Y si definisce variabile dipendente
- X si definisce variabile indipendente

Frequenze condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Distribuzione condizionata del carattere A rispetto alla j-sima modalità del carattere B

Frequenze condizionate

A \ B	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Distribuzione condizionata del carattere B rispetto alla i-sima modalità del carattere A

Frequenze relative condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

9

La distribuzione delle frequenze relative condizionate della variabile A (k modalità) rispetto alla $j-$sima modalità della variabile B (h modalità) si ottiene dividendo ciascun elemento dell'$j-$ma colonna (frequenza assoluta) per il rispettivo totale di di colonna $n_{ij}/n_{.j}$ per $i = 1, \ldots, k$.

Frequenze relative condizionate

B / A	b_1	b_2	...	b_j	...	b_q	totale
a1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
a2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
a_K	n_{K1}	n_{K2}	...	n_{kj}	...	n_{Kq}	$n_{K.}$
totale	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

1

La distribuzione delle frequenze relative condizionate della variabile B (h modalità) rispetto alla $i-$sima modalità della variabile A (k modalità) si ottiene dividendo ciascun elemento dell'$i-$ma riga (frequenza assoluta) per il rispettivo totale di riga $n_{ij}/n_{i.}$ per $j = 1, \ldots, h$.

Esempio di tabella a doppia entrata

Si consideri di aver registrato la meta del viaggio e il mezzo di trasporto di un collettivo di 592 persone. I risultati sono raccolti nella seguente tabella

occhi/capelli	*Italia*	*Spagna*	*Portogallo*	*Francia*	*Tot*
macchina	68	119	26	7	220
aereo	20	84	17	94	215
treno	15	54	14	10	93
nave	5	29	14	16	64
Tot	108	286	71	127	592

Distribuzioni relative condizionate

Esercitazione 6

A. Iodice

Relazioni tra variabili

Indipendenza

Indici di connessione

Dipendenza in variabili miste

Frequenze condizionate della variabile *destinazione* rispetto alle modalità della variabile *mezzo*

mezzo/destinazione	*Italia*	*Spagna*	*Portogallo*	*Francia*	*Tot*
macchina	0.309	0.541	0.118	0.032	1
aereo	0.093	0.391	0.079	0.437	1
treno	0.161	0.581	0.151	0.108	1
nave	0.078	0.453	0.219	0.250	1

Frequenze condizionate della variabile *mezzo* rispetto alle modalità della variabile *destinazione*

mezzo/destinazione	*Italia*	*Spagna*	*Portogallo*	*Francia*
macchina	0.630	0.416	0.366	0.055
aereo	0.185	0.294	0.239	0.740
treno	0.139	0.189	0.197	0.079
nave	0.046	0.101	0.197	0.126
Tot	1	1	1	1

Indipendenza e distribuzioni condizionate

Le componenti di una variabile doppia (X, Y) sono indipendenti se le distribuzioni di frequenze relative condizionate $Y|X$ e $X|Y$ sono costanti.
Formalmente dovrà risultare per $Y|X$

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \frac{n_{i3}}{n_{.3}} = \ldots = \frac{n_{ih}}{n_{.h}}$$

e per $X|Y$

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \frac{n_{3j}}{n_{3.}} = \ldots = \frac{n_{kj}}{n_{k.}}$$

Indipendenza

Si supponga che nel precedente esempio sia stata osservata la seguente distribuzione doppia.

mezzo/destinazione	*Italia*	*Spagna*	*Portogallo*	*Francia*	*Tot*
macchina	40	106	26	47	220
aereo	39	104	26	46	215
treno	17	45	11	20	93
nave	12	31	8	14	64
Tot	108	286	71	127	592

Indipendenza

In questo caso le frequenze condizionate della variabile *destinazione* rispetto alle modalità della variabile *mezzo*

mezzo/destinazione	*Italia*	*Spagna*	*Portogallo*	*Francia*	*Tot*
macchina	0.182	0.483	0.120	0.215	1
aereo	0.182	0.483	0.120	0.215	1
treno	0.182	0.483	0.120	0.215	1
nave	0.182	0.483	0.120	0.215	1
Tot	0.182	0.483	0.120	0.215	1

Mentre le frequenze condizionate della variabile *mezzo* rispetto alle modalità della variabile *destinazione*

mezzo/destinazione	*Italia*	*Spagna*	*Portogallo*	*Francia*	*Tot.*
macchina	0.372	0.372	0.372	0.372	0.372
aereo	0.363	0.363	0.363	0.363	0.363
treno	0.157	0.157	0.157	0.157	0.157
nave	0.108	0.108	0.108	0.108	0.108
Tot	1	1	1	1	1

Indipendenza

Se le componenti di una variabile doppia (X, Y) sono indipendenti (le distribuzioni di frequenze relative condizionate $Y|X$ e $X|Y$ sono costanti), allora vale la seguente relazione

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

con $i = 1, \ldots, k; j = 1, \ldots, h$

Pertanto, data una distribuzione doppia di frequenze, il legame tra le due componenti (mutabile) varierà tra una situazione di indipendenza (assenza di legame) e un qualche grado di connessione

Indice quadratico di connessione (X^2)

Gli indici per la misura della connessioni sono basati sulle differenze tra le frequenze osservate sul collettivo n_{ij} e le frequenze teoriche $\hat{n}_{ij}$, che si osserverebbero sul collettivo se le mutabili considerate fossero indipendenti.
Indice quadratico di connessione (X^2) è dato dalla seguente relazione

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$$

- in caso di indipendenza, essendo $n_{ij} = \hat{n}_{ij}$, risulta $X^2 = 0$
- il massimo valore dell'indice è dato dalla seguente espressione: $n \times min(k-1, h-1)$

Indice quadratico di connessione (X^2)

Per calcolare l'indice quadratico di connessione che caratterizza le variabili $mezzo$ e $destinazione$, con distribuzione congiunta di frequenze

n_{ij} :

mezzo/destinazione	$Italia$	$Spagna$	$Portogallo$	$Francia$	$Tot.$
$macchina$	68	119	26	7	220
$aereo$	20	84	17	94	215
$treno$	15	54	14	10	93
$nave$	5	29	14	16	64
Tot	108	286	71	127	592

si deve calcolare la distribuzione di frequenze che si osserverebbero in caso di indipendenza

$\hat{n}_{ij}$:

mezzo/destinazione	$Italia$	$Spagna$	$Portogallo$	$Francia$	$Tot.$
$macchina$	40.135	106.284	26.385	47.196	220
$aereo$	39.223	103.868	25.785	46.123	215
$treno$	16.966	44.929	11.154	19.951	93
$nave$	11.676	30.919	7.676	13.730	64
Tot	108	286	71	127	592

Indice quadratico di connessione (X^2)

Esercitazione 6

A. Iodice

Relazioni tra variabili

Indipendenza

Indici di connessione

Dipendenza in variabili miste

$\frac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$:

mezzo/destinazione	*Italia*	*Spagna*	*Portogallo*	*Francia*
macchina	19.346	1.521	0.006	34.234
aereo	9.421	3.800	2.993	49.697
treno	0.228	1.831	0.726	4.963
nave	3.817	0.119	5.211	0.375

L'indice X^2 è dato dunque dalla somma degli elementi in tabella

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = 19.346 + 1.521 + 0.006 + 34.234 + 9.421 + 3.800 + 2.993 +$$

$$+ 49.697 + 0.228 + 1.831 + 0.726 + 4.963 + 3.817 + 0.119 + 5.211 + 0.375 = 138.29$$

Indice ν di Cramer

Esercitazione 6

A. Iodice

Relazioni tra variabili

Indipendenza

Indici di connessione

Dipendenza in variabili miste

avendo definito $n \times min(k-1, h-1)$ come valore massimo che X^2 può assumere, è possibile ottenere una versione normalizzata dell'indice di connessione. Viene definito indice ν di Cramer.

$$\nu = \sqrt{\frac{X^2}{n \times min(k-1, h-1)}}$$

con k e h numero di modalità delle componenti della mutabile doppia.

L'indice è normalizzato, quindi $0 \leq \nu \leq 1$.

Indice ν di Cramer

Con riferimento ai dati dell'esercizio, si ha che $X^2 = 138.29$, $n = 592$, $h = 4$ e $k = 4$

$$\nu = \sqrt{\frac{X^2}{n \times min(k-1, h-1)}} = \sqrt{\frac{138.29}{592 \times min(3,3)}} = 0.28$$

Connessione in media

Esercitazione 6

A. Iodice

Relazioni tra variabili

Indipendenza

Indici di connessione

Dipendenza in variabili miste

Data una distibuzione doppia di un carattere misto (X, Y), si dir che la componente Y indipendente in media da X se al variare delle modalità di X le medie condizionate di X rimangono costanti (vale il viceversa).
Il fatto che Y sia indipendente in media da X non implica che sia vero il contrario (come invece accade per l'indipendenza in distribuzione).

Connessione in media

Data una distibuzione doppia di un carattere misto (X, Y), si dir che la componente Y indipendente in media da X se al variare delle modalità di X le medie condizionate di X rimangono costanti (vale il viceversa).
Il fatto che Y sia indipendente in media da X non implica che sia vero il contrario (come invece accade per l'indipendenza in distribuzione).

$$\mu_y = \overline{y} = \frac{1}{n}\sum_{j=1}^{h} y_j n_{.j}$$

Rappresenta la media di Y e si ottiene considerando la distribuzione marginale di Y.

$$\overline{y}_i = \overline{y}|x_i = \frac{1}{n_{i.}}\sum_{j=1}^{h} y_j n_{ij}$$

Rappresenta la media di Y condizionata alla $i-ma$ modalità della variabile X.

Decomposizione della devianza

Ricordando che la devianza il numeratore della varianza...

$$
\begin{aligned}
Dev_y = & \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y})^2 n_{ij} = \\
& = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y}_i+\overline{y}_i-\overline{y})^2 n_{ij} = \\
& = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y}_i)^2 n_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{h}(\overline{y}_i-\overline{y})^2 n_{ij} + \\
& + 2\sum_{i=1}^{k}\sum_{j=1}^{h}(y_j-\overline{y}_i)(\overline{y}_i-\overline{y}) n_{ij}
\end{aligned}
$$

Decomposizione della devianza

$$= \sum_{i=1}^{k}\left[\sum_{j=1}^{h}(y_j-\overline{y}_i)^2 n_{ij}\right]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.}+$$

$$+2\sum_{i=1}^{k}(y_j-\overline{y}_i)\sum_{j=1}^{h}(\overline{y}_i-\overline{y})n_{ij}=$$

$$=\sum_{i=1}^{k}[Dev(Y\mid X=x_i)]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.}=$$

$$=Dev(W)+Dev(B)$$

Decomposizione della devianza

Esercitazione 6

A. Iodice

Relazioni tra variabili

Indipendenza

Indici di connessione

Dipendenza in variabili miste

$$= \sum_{i=1}^{k} \left[\sum_{j=1}^{h} (y_j - \overline{y}_i)^2 n_{ij} \right] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} +$$

$$+ 2 \sum_{i=1}^{k} (y_j - \overline{y}_i) \sum_{j=1}^{h} (\overline{y}_i - \overline{y}) n_{ij} =$$

$$= \sum_{i=1}^{k} [Dev(Y \mid X = x_i)] + \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2 n_{i.} =$$

$$= Dev(W) + Dev(B)$$

Decomposizione della devianza

$$= \sum_{i=1}^{k}\left[\sum_{j=1}^{h}(y_j-\overline{y}_i)^2 n_{ij}\right]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.}+$$

$$+2\sum_{i=1}^{k}(y_j-\overline{y}_i)\sum_{j=1}^{h}(\overline{y}_i-\overline{y})n_{ij}=$$

$$=\sum_{i=1}^{k}[Dev(Y\mid X=x_i)]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.}=$$

$$=Dev(W)+Dev(B)$$

$$= \sum_{i=1}^{k}\left[\sum_{j=1}^{h}(y_j-\overline{y}_i)^2 n_{ij}\right]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.}+$$

$$+2\sum_{i=1}^{k}(y_j-\overline{y}_i)\sum_{j=1}^{h}(\overline{y}_i-\overline{y})n_{ij} =$$

$$= \sum_{i=1}^{k}[Dev(Y \mid X = x_i)]+\sum_{i=1}^{k}(\overline{y}_i-\overline{y})^2 n_{i.} =$$

$$= Dev(W) + Dev(B)$$

Rapporto di correlazione di Pearson (η^2)

$Dev(W)$ rappresenta la varianza all'interno dei gruppi definiti dalle modalità di X. $Dev(B)$ rappresenta invece la variabilità tra i gruppi: ovvero la variabilità delle medie condizionate rispetto alla media generale.

Rapporto di correlazione di Pearson (η^2)

$Dev(W)$ rappresenta la varianza all'interno dei gruppi definiti dalle modalità di X. $Dev(B)$ rappresenta invece la variabilità tra i gruppi: ovvero la variabilità delle medie condizionate rispetto alla media generale.
Se Y indipendente in media da X, allora le medie condizionate $\overline{y}_i$ saranno tutte costanti, la variabilità ad esse associate sar uguale a zero. In particolare risulter $Dev(B) = 0$ quindi

$$Dev(Y) = Dev(W) + 0$$

Rapporto di correlazione di Pearson (η^2)

$Dev(W)$ rappresenta la varianza all'interno dei gruppi definiti dalle modalità di X. $Dev(B)$ rappresenta invece la variabilità tra i gruppi: ovvero la variabilità delle medie condizionate rispetto alla media generale.
Se Y indipendente in media da X, allora le medie condizionate $\overline{y}_i$ saranno tutte costanti, la variabilità ad esse associate sar uguale a zero. In particolare risulter $Dev(B) = 0$ quindi

$$Dev(Y) = Dev(W) + 0$$

Quindi, per quantificare la dipendenza in media di Y da X occorre un indice basato su $Dev(B)$.

$$\eta^2 = \frac{Dev(B)}{Dev(Y)}$$

Calcolo del rapporto di correlazione: valori in classi

Si consideri l'esempio della variabile doppia reddito/grado di anzianità

Zona\reddito(in migliaia)	[10,15]	[15, 20[	[20,25[	[25,30[	Totale
Nord	0	7	34	5	46
Centro	1	18	5	1	25
Sud	31	1	0	0	32
Totale	32	26	39	6	103

Calcolo del rapporto di correlazione

Esercitazione 6

A. Iodice

Ai fini del calcolo del rapporto di correlazione necessario calcolare la devianza totale della variabile $Dev(Y)$ e la devianza tra le classi $Dev(B)$ (ovvero la devianza tra le medie condizionate $Y \mid X = x_i, i = 1, 2, \ldots, k$ e la media globale). Dunque

$$\mu(Y) = \frac{1}{103}(12.5 \times 32) + (17.5 \times 26) + \\ + (22.5 \times 39) + (27.5 \times 6) = 18.42$$

Calcolo del rapporto di correlazione

$$\mu(Y \mid x_i = Nord) = \frac{1}{46}(12.5 \times 0) + (17.5 \times 7) + \\ + (22.5 \times 34) + (27.5 \times 5) = 22.28$$

$$\mu(Y \mid x_i = Centro) = \frac{1}{25}(12.5 \times 1) + (17.5 \times 18) + \\ + (22.5 \times 5) + (27.5 \times 1) = 18.7$$

$$\mu(Y \mid x_i = Sud) = \frac{1}{32}(12.5 \times 31) + (17.5 \times 1) + \\ + (22.5 \times 0) + (27.5 \times 0) = 12.66$$

Calcolo del rapporto di correlazione

$$dev(Y) = (12.5 - 14.9)^2 \times 32 + (17.5 - 14.9)^2 \times 26 + \\ + (22.5 - 14.9)^2 \times 39 + (27.5 - 14.9)^2 \times 6 = 2287.39$$

$$dev(B) = (22.28 - 14.9)^2 \times 46 + (18.7 - 14.9)^2 \times 25 + \\ + (12.66 - 14.9)^2 \times 32 = 1749.02$$

$$\eta^2 = \frac{dev(B)}{dev(Y)} = \frac{1749.02}{2287.39} = 0.76$$