ISTITUZIONI DI STATISTICA – A. A. 2007/2008
Marco Minozzo e Annamaria Guolo
Laurea in Economia del Commercio Internazionale
Laurea in Economia e Amministrazione delle Imprese
Università degli Studi di Verona (sede di Vicenza)

Statistica Descrittiva
Soluzioni 8. Dipendenza

Introduzione

Si consideri l'informazione sui due caratteri X e Y, di modalità rispettivamente $x_1, \ldots, x_r$ e $y_1, \ldots, y_c$, racchiusa nella seguente tabella di frequenza a doppia entrata

	y_1	$\ldots$	y_j	$\cdots$	y_c	Totale
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1c}	$f_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ic}	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rc}	$f_{r.}$
Totale	$f_{.1}$	$\cdots$	$f_{.j}$	$\cdots$	$f_{.c}$	N

Nella tabella, f_{ij} rappresenta la frequenza assoluta congiunta rilevata per la modalità x_i di X e y_j di Y, $f_{i.}$ rappresenta la frequenza assoluta marginale della modalità x_i di X, $f_{.j}$ rappresenta la frequenza assoluta marginale della modalità y_j di Y e N rappresenta il totale delle osservazioni rilevate. Si ha

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij} = \sum_{i=1}^{r} f_{i.} = \sum_{j=1}^{c} f_{.j}.$$

A partire dalla tabella a doppia entrata, si ricavano la distribuzione relativa marginale di X

X	$p_{x.}$
x_1	$p_{1.} = f_{1.}/N$
$\vdots$	$\vdots$
x_i	$p_{i.} = f_{i.}/N$
$\vdots$	$\vdots$
x_r	$p_{r.} = f_{r.}/N$

e la distribuzione relativa marginale di Y

Y	$p_{y.}$
y_1	$p_{.1} = f_{.1}/N$
$\vdots$	$\vdots$
y_j	$p_{.j} = f_{.j}/N$
$\vdots$	$\vdots$
y_c	$p_{.c} = f_{.c}/N$

Fissata una qualunque modalità di Y, y_j, la distribuzione relativa di X condizionata a y_j è data da

X	$p_{x\|Y=y_j}$
x_1	$p_{1\|j} = f_{1j}/f_{.j}$
$\vdots$	$\vdots$
x_i	$p_{i\|j} = f_{ij}/f_{.j}$
$\vdots$	$\vdots$
x_r	$p_{r\|j} = f_{rj}/f_{.j}$

In modo analogo si ricava la distribuzione relativa di Y condizionata alla modalità generica x_i

Y	$p_{y\|X=x_i}$
y_1	$p_{1\|i} = f_{i1}/f_{i.}$
$\vdots$	$\vdots$
y_j	$p_{j\|i} = f_{ij}/f_{i.}$
$\vdots$	$\vdots$
y_c	$p_{c\|i} = f_{ic}/f_{i.}$

Solitamente è di interesse valutare se esiste una indipendenza tra i due caratteri X e Y. In caso di indipendenza, le distribuzione relative di un carattere condizionate a ciascuna delle modalità assunte dall'altro carattere sono identiche. Ciò si traduce nella possibilità di scrivere le frequenze assolute congiunte come segue

$$f^*_{ij} = \frac{f_{i.} f_{.j}}{N},$$

o, equivalentemente, in termini di frequenze relative congiunte

$$p^*_{ij} = p_{i.} p_{.j}.$$

A partire dal confronto tra le frequenze congiunte osservate f_{ij} e quelle teoriche f^*_{ij}, si può valutare se esista indipendenza tra X e Y. A tal fine, si costruisce l'indice di dipendenza χ^2

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(f_{ij} - f^*_{ij}\right)^2}{f^*_{ij}}.$$

Il valore nullo dell'indice è compatibile con una situazione di indipendenza tra X e Y. Solitamente, a partire da χ^2 si costruisce un indice normalizzato, il coefficiente di contingenza C

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N} \cdot \frac{k}{k-1}},$$

dove k è il valore minimo tra il numero delle righe r ed il numero delle colonne c. Il coefficiente di contingenza C assume valore 0 (valore minimo) in caso di indipendenza tra X e Y e valore 1 (valore massimo) in caso di massima dipendenza (massima connessione).
Nelle situazioni in cui la tabella di contingenza non è quadrata, vale a dire se il numero di modalità di X e Y è diverso, allora non può esistere una dipendenza massima di X da Y e di Y da X. Vale solo la massima dipendenza di X da Y oppure di Y da X.

Esercizio A.
a) La distribuzione marginale del carattere *Tipo di occupazione* è:

Tipo di occupazione	p
Agricoltura	$445/8406 = 0,0529$
Industria	$2859/8406 = 0,3401$
Altri servizi	$5102/8406 = 0,6069$

b) Le distribuzioni condizionate relative richieste sono le seguenti:

	Agricoltura	Industria	Altre attività	Totale
Piemonte	0,0514 =87/1693	0,3993	0,5493	1
Veneto	0,0541=100/1849	0,4143	0,5316	1
Emilia Romagna	0,0714=121/1694	0,3495	0,5791	1
Toscana	0,0421=57/1353	0,3422	0,6157	1
Lazio	0,0440=80/1817	0,1992	0,7568	1

Essendo queste distribuzioni non tutte uguali tra loro, possiamo affermare che esiste dipendenza tra i due caratteri.
c) Per il calcolo degli indici richiesti, si considerino le seguenti frequenze teoriche di indipendenza ottenute come $f^*_{ij} = f_{i\cdot} f_{\cdot j}/N$:

	Agricoltura	Industria	Altre attività	Totale
Piemonte	89,625	575,813	1027,562	1693
Veneto	97,883	628,871	1122,246	1849
Emilia Romagna	89,678	576,153	1028,169	1694
Toscana	71,625	460,175	821,200	1353
Lazio	96,189	617,988	1102,823	1817
Totale	445	2859	5102	8406

dove, ad esempio, $89,625 = 1693 \cdot 445/8406$. Da qui si ricava l'indice χ^2 che risulta pari a 266,645. Il coefficiente di contingenza in questo caso risulta pari a

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N} \cdot \frac{k}{k-1}} = \sqrt{\frac{266,645}{266,645 + 8406} \cdot \frac{3}{3-1}} = 0,215,$$

dove $k = \min\{\text{numero di righe; numero di colonne}\}$. L'indice C indica una dipendenza bassa del tipo di occupazione dalla regione.

Esercizio B.
a) La distribuzione di frequenza assoluta marginale della reazione al test è pari a: 25, 36, 31, 20, dove i valori sono ottenuti sommando le frequenze di ciascun gruppo, ad esempio: $25 = 14+8+3$. Quindi, la distribuzione marginale relativa e le distribuzioni condizionate relative della reazione al test sono date da

	elevata	discreta	moderata	negativa	Totale
Gruppo 1	0,326=14/43	0,279	0,302	0,093	1
Gruppo 2	0,167=8/48	0,458	0,250	0,125	1
Gruppo 3	0,143=3/21	0,095	0,286	0,476	1
Marginale Y	0,223=25/112	0,321	0,277	0,179	1

b) Le distribuzioni condizionate relative del tipo di patologia data la risposta al test sono le seguenti

	elevata	discreta	moderata	negativa	Marginale X
Gruppo 1	0,56=14/25	0,33=12/36	0,42=13/31	0,20=4/20	0,38=43/112
Gruppo 2	0,32	0,61	0,39	0,30	0,43
Gruppo 3	0,12	0,06	0,19	0,50	0,19
Totale	1	1	1	1	1

Ovviamente, essendo le distribuzioni condizionate relative non tutte uguali tra loro, c'è dipendenza tra i caratteri X ed Y.
c) Per una tabella di dimensione 3 righe (X) per 4 colonne (Y) ci può essere perfetta dipendenza di X da Y, ma non di Y da X.
d) Per il calcolo dell'indice di dipendenza χ^2 consideriamo la seguente tabella

$\frac{(f_{ij}-f^*_{ij})^2}{f^*_{ij}}$					
	2,0187	0,2400	0,1013	1,7623	
	0,6876	2,7989	0,1244	0,7714	
	0,6075	3,3426	0,0060	10,4167	
					22,8776

dove $f^*_{ij} = f_{i\cdot}f_{\cdot j}/N$. Quindi, l'indice di dipendenza χ^2 è pari a 22,8776. Il coefficiente di contingenza C risulta pari a 0,5044. Secondo questo indice, nella tabella c'è un grado di dipendenza intermedio.

Esercizio C.
a) Le distribuzioni relative marginali dei due caratteri sono

Tipo di produzione	p
Cereali	$376/856 = 0,439$
Ortaggi	$0,366$
Latte	$0,195$

e

Nazione	p
Grecia	$89/856 = 0,104$
Italia	$0,521$
Spagna	$0,375$

b) La distribuzione relativa condizionata è

	Cereali	Ortaggi	Latte	
Grecia	$42/89 = 0,472$	0,449	0,079	1
Italia	$201/446 = 0,451$	0,323	0,226	1
Spagna	$133/321 = 0,414$	0,402	0,184	1

c) Considerando che le frequenze teoriche f^*_{ij} in caso di indipendenza sono

	Cereali	Ortaggi	Latte
Grecia	$39,093$	32,543	17,363
Italia	$195,906$	163,082	87,012
Spagna	141	117,375	62,625

si ha che $\chi^2 = 14,5390$. Di conseguenza

$$C = \sqrt{\frac{14,5390}{14,5390 + 856}\frac{3}{2}} = 0,158,$$

il che significa che la dipendenza tra i due caratteri è bassa.

Esercizio D.
a) Nel caso in cui si abbia la massima dipendenza tra i due caratteri, la distribuzione delle frequenze nelle celle della tabella deve essere tale per cui la realizzazione di una modalità di un carattere implichi l'esatta conoscenza della osservazione del secondo carattere. Ciò significa che per ogni riga e per ogni colonna, vi sarà una sola celle con valori diversi da 0 e le rimanenti avranno frequenze osservate pari a 0. Considerando $N = 20$, una possibile tabella compatibile con la richiesta di massima dipendenza tra i caratteri è

	y_1	y_2	y_3	Totale
x_1	0	9	0	9
x_2	4	0	0	4
x_3	0	0	7	7
Totale	4	9	7	20

b) Per soddisfare la richiesta di indipendenza tra i caratteri, è necessario costruire la tabella di modo che le frequenze assolute congiunte siano $f^*_{ij} = f_{i.}f_{.j}/N$. Una possibile soluzione è

	y_1	y_2	y_3	Totale
x_1	4	1	7	12
x_2	4	1	7	12
x_3	8	2	14	24
Totale	16	4	28	48

Si noti, infatti, che in tal caso le distribuzioni relative di X condizionate alle modalità di Y sono uguali tra loro, così come lo sono le distribuzioni relative di Y condizionate alle modalità di X.