

Analisi congiunta di più fenomeni

Dati relativi al disastro del Titanic:

Classe	Sesso	Età	Morti	Sopravvissuti
1 ^a	Uomini	Bambini	0	5
		Adulti	118	57
	Donne	Bambini	0	1
		Adulti	4	140
2 ^a	Uomini	Bambini	0	11
		Adulti	154	14
	Donne	Bambini	0	13
		Adulti	13	80
3 ^a	Uomini	Bambini	35	13
		Adulti	387	75
	Donne	Bambini	17	14
		Adulti	89	76
Equipaggio	Uomini	Bambini	0	0
		Adulti	670	192
	Donne	Bambini	0	0
		Adulti	3	20
Totale			1490	711

Si tratta di una tabella a 4 vie. Sono stati rilevati 4 caratteri:

- la classe (1^a , 2^a , 3^a , equipaggio)
- il sesso (M, F)
- l'età (Bambini, Adulti)
- lo status (Morto, Sopravvissuto)

Problema:

C'è stata discriminazione per i passeggeri di terza classe?

Sono state salvate prima le donne e poi i bambini?

La connessione

Su 15 individui è stato effettuato un test per rilevare l'*la capacità di analisi* (X) e quella *a lavorare in gruppo* (Y)

modalità: sufficiente (S), buona (B) e ottima (O)

X	O	O	S	B	S	O	B	B	S	B	O	B	B	O	S
Y	O	B	B	B	S	S	O	O	B	B	O	S	B	S	B

Costruiamo la tabella delle frequenze

Dobbiamo contare quante volte si presenta ogni coppia di valori

	Y	S	B	O	
X					
S		1	3	0	4
B		1	3	2	6
O		2	1	2	5
		4	7	4	15

Si tratta di una tabella a **doppia entrata**, detta anche **tabella di contingenza**

Rileviamo su n individui i caratteri:

X con modalità x_1, \dots, x_h

Y con modalità y_1, \dots, y_k

In corrispondenza della riga i e della colonna j andiamo a inserire il numero di volte n_{ij} che si presenta la coppia (x_i, y_j)

Y	y_1	y_2	\dots	y_j	\dots	y_k	
X							
x_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2k}	$n_{2.}$
\vdots							\vdots
x_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{ik}	$n_{i.}$
\vdots							\vdots
x_h	n_{h1}	n_{h2}	\dots	n_{hj}	\dots	n_{hk}	$n_{h.}$
	$n_{.1}$	$n_{.2}$	\dots	$n_{.j}$	\dots	$n_{.k}$	n

n_{ij} è detta **frequenza congiunta**

Da questa tabella si ricavano altre distribuzioni utili a comprendere i fenomeni studiati e le loro relazioni.

Le frequenze marginali

Quando sommiamo per riga:

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ik} = \sum_{j=1}^k n_{ij}$$

è detta **frequenza marginale** di X

Quando sommiamo per colonna:

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{hj} = \sum_{i=1}^h n_{ij}$$

è detta **frequenza marginale** di Y

Abbiamo quindi due distribuzioni marginali:

La distribuzione di frequenza marginale di Y

y_1	y_2	\dots	y_k
$n_{.1}$	$n_{.2}$	\dots	$n_{.k}$

La distribuzione di frequenza marginale di X

x_1	x_2	\dots	x_k
$n_{1.}$	$n_{2.}$	\dots	$n_{k.}$

Per l'esempio sono rispettivamente

Y	S	B	O
	4	7	4

X	S	B	O
	4	6	5

Le distribuzioni condizionate assolute

Si ottengono dal *corpo* della tabella.

Se teniamo fisso un valore della variabile X otteniamo la **distribuzione condizionata di Y dato $X = x_i$** . Sono le righe della tabella.

$Y X = O$	S	B	O	
$n_{j i=3}$	2	1	2	5

Se teniamo fisso un valore della variabile Y otteniamo la **distribuzione condizionata di X dato $Y = y_i$** . Sono le colonne della tabella.

$X Y = S$	$n_{i j=1}$
S	1
B	1
O	2
	4

Ci si restringe a una sotto-popolazione. Negli esempi nel primo caso ai 5 individui che presentano attitudine all'analisi Ottimo, nel secondo ai 4 individui che hanno capacità a lavorare in gruppo Sufficiente

Distribuzioni relative congiunte e marginali

Si ottengono dividendo per n ogni frequenza congiunta assoluta. Otteniamo:

le frequenze congiunte relative

$$f_{ij} = n_{ij}/n$$

frequenze relative marginali di Y

$$f_{.j} = n_{.j}/n$$

frequenze relative marginali di X

$$f_{i.} = n_{i.}/n$$

Calcoliamo la tabella delle frequenze relative per l'esempio

	Y	S	B	O	
X					
S		$\frac{1}{15}$	$\frac{3}{15}$	0	$\frac{4}{15}$
B		$\frac{1}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{6}{15}$
O		$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{5}{15}$
		$\frac{4}{15}$	$\frac{7}{15}$	$\frac{4}{15}$	1

Distribuzioni relative condizionate

Sono utili per confrontare le distribuzioni

Le frequenze condizionate relative si ottengono dividendo ogni frequenza condizionata per il totale di riga o colonna.

Abbiamo le frequenze relative della variabile Y condizionata a X

Y	S	B	O	
$X = S$	$1/4$	$3/4$	0	1
$X = B$	$1/6$	$3/6$	$2/6$	1
$X = O$	$2/5$	$1/5$	$2/5$	1
	$4/15$	$7/15$	$4/15$	1

Nell'ultima riga abbiamo la distribuzione marginale relativa di Y

Le frequenze relative della X condizionata a Y

X	$Y = S$	$Y = B$	$Y = O$	
S	$1/4$	$3/7$	0	$4/15$
B	$1/4$	$3/7$	$2/4$	$6/15$
O	$2/4$	$1/7$	$2/4$	$5/15$
	1	1	1	1

Nell'ultima colonna abbiamo la distribuzione marginale relativa di X

Indipendenza

Se tutte le distribuzioni condizionate di X da Y sono uguali allora sono necessariamente uguali alla distribuzione marginale di X

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

Se tutte le distribuzioni condizionate di Y da X sono uguali allora sono necessariamente uguali alla distribuzione marginale di Y

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

Nei due casi deve essere

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

e nel qual caso si dice che i fenomeni X e Y sono **indipendenti**

In termini di frequenze relative possiamo riscrivere così

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

ovvero

$$f_{ij} = f_{i.} \cdot f_{.j}$$

Vale a dire

$$\text{Freq}(X = x_i \cap Y = y_j) = \text{Freq}(X = x_i) \cdot \text{Freq}(Y = y_j)$$

Le quantità

$$n_{ij}^* = \frac{n_{i.} \cdot n_{.j}}{n} \quad \text{per ogni } i = 1, \dots, h \quad j = 1, \dots, k$$

sono le **frequenze teoriche di indipendenza**

Per verificare se due fenomeni sono indipendenti basta calcolare le frequenze teoriche di indipendenza e confrontarle con le frequenze osservate.

Esempio Consideriamo $n_{11} = 1$ La frequenza teorica di indipendenza è

$$n_{11}^* = \frac{n_{1.} \cdot n_{.1}}{15} = \frac{4 \cdot 4}{15} = \frac{16}{15} = 0.94$$

Quindi i due fenomeni non sono indipendenti.

Attenzione: quando c'è una frequenza pari a zero e le marginali non sono nulle siamo sempre in presenza di caratteri non indipendenti.

Come misuriamo il grado di dipendenza per questo tipo di fenomeni?

Costruiamo un indice basato sulle differenze $n_{ij} - n_{ij}^*$. Più queste differenze sono piccole più i due fenomeni sono vicini all'indipendenza.

Queste differenze si chiamano *contingenze*

$$c_{ij} = n_{ij} - n_{ij}^*$$

Sulle contingenze si basa l'indice per misurare il grado di connessione tra due fenomeni χ^2 (Chi-quadrato)

$$\chi^2 = \sum_{i=1}^h \sum_{j=1}^k \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

Che diventa (per i conti)

$$\chi^2 = n \cdot \left(\sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}^2}{n_{i.} \cdot n_{.j}} - 1 \right)$$

Per effettuare confronti è bene avere un indice relativo (che vari tra 0 e 1)

$$\tilde{\chi}^2 = \frac{\sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}^2}{n_{i.} \cdot n_{.j}} - 1}{\min(h-1, k-1)}$$

L'indice relativo è ottenuto dividendo l'indice χ^2 per

$$\chi_{\max}^2 = \max \chi^2 = n \cdot \min(h-1, k-1)$$

che è il massimo valore che può assumere tale indice.

Esempio: Calcoliamo l'indice di connessione per i dati dell'esempio. Costruiamo la tabella con le frequenze teoriche di indipendenza

X	$Y = S$	$Y = B$	$Y = O$	
S	1.07	1.86	1.07	4
B	1.6	2.8	1.6	6
O	1.33	2.34	1.33	5
	4	7	4	15

Costruiamo la tabella con le contingenze

X	$Y = S$	$Y = B$	$Y = O$
S	1-0.94	3-1.87	0-0.94
B	1-1.6	3-2.8	2-1.6
O	2-1.33	1-2.33	2-1.33

Costruiamo la tabella con $\frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$

X	$Y = S$	$Y = B$	$Y = O$	
S	0.004	0.683	0.94	
B	0.225	0.014	0.100	
O	0.338	0.759	0.338	
				$\chi^2 = 3.400$

L'indice relativo lo otteniamo come

$$\tilde{\chi}^2 = 3.4 / (15 * 2) = 0.11$$

Esempio Verifichiamo se lo status (morto o vivo) è legato più all'età, al sesso o alla classe di imbarco

Età	Morti	Sopravvissuti	Totale
Bambini	52	57	109
Adulti	1438	654	2092
	1490	711	2201

$$\tilde{\chi}^2 = \frac{\frac{52^2}{109 \cdot 1490} + \frac{57^2}{109 \cdot 711} + \frac{1438^2}{2092 \cdot 1490} + \frac{654^2}{2092 \cdot 711} - 1}{1}$$

$$= 0.009$$

Sesso	Morti	Sopravvissuti	Totale
Uomini	1364	367	1731
Donne	126	344	470
	1490	711	2201

$$\tilde{\chi}^2 = \frac{\frac{1364^2}{1731 \cdot 1490} + \frac{367^2}{1731 \cdot 711} + \frac{126^2}{470 \cdot 1490} + \frac{344^2}{470 \cdot 711} - 1}{1}$$

$$= 0.208$$

Classe	Morti	Sopravvissuti	Totale
Prima	122	203	325
Seconda	167	118	285
Terza	528	178	706
Equipaggio	673	212	885
	1490	711	2201

$$\tilde{\chi}^2 = 0.086$$

Come si vede la variabile “Sopravvivenza” (o morte) sembra essere influenzata più dal sesso ($\tilde{\chi}^2 = 0.21$) che non dalla classe ($\tilde{\chi}^2 = 0.09$) o dall’età ($\tilde{\chi}^2 = 0.009$)

Massima connessione

Il concetto opposto all'indipendenza è quello di massima connessione. Si ha massima connessione quando la conoscenza dell'esito di una variabile determina completamente l'esito dell'altra.

La massima connessione può essere bilaterale (solo con tabelle quadrate!!!)

X	Y	y_1	y_2	y_3
x_1		×	0	0
x_2		0	0	×
x_3		0	×	0

Quando le tabelle sono rettangolari si ha massima connessione unilaterale

X	Y	y_1	y_2	y_3
x_1		×	0	0
x_2		0	×	×

(a)

X	Y	y_1	y_2
x_1		×	0
x_2		0	×
x_3		0	×

(b)

In (a) abbiamo massima dipendenza di X da Y in (b) viceversa

Esercizio. Molti anni fa venne condotto uno studio epidemiologico per studiare gli effetti positivi dell'uso di aspirina sulla prevenzione degli attacchi cardiaci. Da un insieme di 22071 medici volontari vennero formati due gruppi: il gruppo di trattamento e quello di controllo. Gli individui del gruppo di trattamento ricevevano una dose quotidiana di aspirina mentre quelli di controllo un farmaco *placebo*, cioè identico all'aspirina e non contenente alcun principio attivo. Lo studio venne condotto per un periodo di 5 anni osservando il numero di decessi per infarto. Si ottennero i seguenti risultati

Farmaco	Esito	Infartuati	Non Infartuati	Totali
Placebo		239	10795	11034
Aspirina		139	10898	11037
		378	21693	22071

Verificare se l'esito è indipendente dal trattamento e in caso negativo calcolare l'indice di connessione assoluto e relativo.