

Distribuzioni doppie

- Quando vengono considerate congiuntamente due colonne di una matrice di dati si ha una **distribuzione doppia disaggregata** (o *unitaria*). Si tratta dell'elencazione delle modalità di due caratteri (X e Y) osservate per ogni unità statistica del collettivo considerato:

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Nome (i)	Sesso (x _i)	Regione (y _i)
Verdi M.	M	Marche
Bianchi C.	F	Calabria
Rossi V.	F	Umbria
Gialli L.	M	Piemonte
Grandi A.	F	Marche
Pini B.	F	Umbria

1

Esempi

- Si consideri la seguente *distribuzione doppia disaggregata* per due caratteri qualitativi (Sesso, Regione)

Nome (i)	Sesso (x _i)	Regione (y _i)
Verdi M.	M	Marche
Bianchi C.	F	Calabria
Rossi V.	F	Umbria
Gialli L.	M	Piemonte
Grandi A.	F	Marche
Pini B.	F	Umbria

- La corrispondente *distribuzione doppia di frequenza* è

		Regione			
Sesso		Calabria	Marche	Umbria	Piemonte
M		0	1	0	1
F		1	1	2	0
					3

- L'informazione contenuta in una distribuzione doppia disaggregata è solitamente *sintetizzata* tramite una **distribuzione doppia di frequenza** che viene rappresentata tramite una tabella a doppia entrata in cui per ogni coppia di modalità dei due caratteri

$(x_i, y_j), \quad i = 1, \dots, H, \quad j = 1, \dots, K$

viene indicata la corrispondente *frequenza congiunta* (n_{ij}).

X/Y	y ₁	y ₂	...	y _j	...	y _K
x ₁	n ₁₁	n ₁₂	...	n _{1j}	...	n _{1K}
x ₂	n ₂₁	n ₂₂	...	n _{2j}	...	n _{2K}
...	...	...	...	...	...	...
x _i	n _{i1}	n _{i2}	...	n _{ij}	...	n _{iK}
...	...	...	...	...	...	...
x _H	n _{H1}	n _{H2}	...	n _{Hj}	...	n _{HK}

Quando il carattere è quantitativo con molte modalità (tipicamente continuo) possono essere utilizzate le **classi** al posto delle modalità.

- Esempio di distribuzione in cui il carattere quantitativo è in classi

Reddito annuo (x 1.000€)			
	0-10	10-30	30-100
Titolo di studio			
Lic. Media	88	142	120
Diploma	9	39	38
Laurea	3	19	42

Distribuzioni marginali

- Sommando le frequenze congiunte *per colonna* si ottengono le **frequenze marginali** di X che corrispondono al numero di soggetti che presentano una certa modalità di questo carattere a prescindere dalla modalità di Y:

$$n_{i\circ} = \sum_{j=1}^K n_{ij}$$

- Analogamente, le **frequenze marginali** di Y si ottengono sommando le frequenze congiunte per riga:

$$n_{\circ j} = \sum_{i=1}^H n_{ij}$$

- La somma di tutte frequenze congiunte (o di tutte le frequenze marginali) corrisponde alla **numerosità del collettivo**

$$n = \sum_{j=1}^K \sum_{i=1}^H n_{ij} = \sum_{j=1}^K n_{\circ j} = \sum_{i=1}^H n_{i\circ}$$

X/Y	Y ₁	Y ₂	...	Y _j	...	Y _K	Totale
X ₁	n ₁₁	n ₁₂	...	n _{1j}	...	n _{1K}	n _{1°}
X ₂	n ₂₁	n ₂₂	...	n _{2j}	...	n _{2K}	n _{2°}
...	...	...	...	...	...	...	...
X _i	n _{i1}	n _{i2}	...	n _{ij}	...	n _{iK}	n _{i°}
...	...	...	...	...	...	...	...
X _H	n _{H1}	n _{H2}	...	n _{Hj}	...	n _{HK}	n _{H°}
Totale	n _{°1}	n _{°2}	...	n _{°j}	...	n _{°K}	n

Esempi

- Alla prima distribuzione doppia considerata nell'esempio precedente

	Regione				
Sesso	Calabria	Marche	Umbria	Piemonte	Totale
M	0	1	0	1	2
F	1	1	2	0	4
Totale	1	2	2	1	6

si hanno le seguenti distribuzioni marginali di X e Y

Sesso (x _i)	Frequenze (n _{i°})
M	2
F	4
Totale	6

Regione (y _j)	Frequenze (n _{°j})
Calabria	1
Marche	2
Umbria	2
Piemonte	1
Totale	6

- Associando a ogni modalità del carattere X la corrispondente frequenza marginale, si ottiene la **distribuzione marginale** di X. E' la stessa distribuzione che avremmo ottenuto osservando il carattere singolarmente.

Modalità (x _i)	Frequenze (n _{i°})
x ₁	n _{1°}
x ₂	n _{2°}
...	...
x _i	n _{i°}
...	...
x _H	n _{H°}
Totale	n

- In modo analogo si ottiene la **distribuzione marginale** di Y.
- Entrambe le distribuzioni possono essere *lette* direttamente dalla tabella a doppia entrata, quando sono presenti i totali (**margini**) di riga e di colonna.

- Per la seconda distribuzione doppia considerata nell'esempio precedente,

	Reddito annuo (x 1.000€)			
	0-10	10-30	30-100	Totale
Lic. Media	88	142	120	350
Diploma	9	39	38	86
Laurea	3	19	42	64
Totale	100	200	200	500

si hanno le seguenti distribuzioni marginali di X e Y

Titolo di studio (x_i)	Frequenze ($n_{i\cdot}$)	Reddito (y_j)	Frequenze ($n_{\cdot j}$)
Lic. media	350	0-10	100
Diploma	86	10-30	200
Laurea	64	30-100	200
Totale	500	Totale	500

Distr. condizionate relative e percentuali

- Per la distribuzione condizionata di Y dato $X = x_i$, le frequenze relative e percentuali possono essere calcolate come

$$f_{j|i} = \frac{n_{ij}}{n_{i\cdot}} \quad e \quad p_{j|i} = 100 \frac{n_{ij}}{n_{i\cdot}} = 100 f_{j|i}$$

- Associando alla distribuzione condizionata di Y dato $X = x_i$ le corrispondenti frequenze relative (o percentuali) si ottiene la distribuzione condizionata relativa (o percentuale) di Y dato $X = x_i$. Questa distribuzione permette di capire come X influenza Y.

Modalità (y_j)	Frequenze (n_{ij})	Freq. relative ($f_{j i}$)	Freq. percentuali ($p_{j i}$)
y_1	n_{1i}	$f_{1 i}$	$p_{1 i}$
y_2	n_{2i}	$f_{2 i}$	$p_{2 i}$
...	...	...	...
y_j	n_{ji}	f_{ji}	p_{ji}
...	...	...	...
y_K	n_{Ki}	f_{Ki}	p_{Ki}
Totale	$n_{i\cdot}$	1	100

Distribuzioni condizionate

- La **distribuzione condizionata di Y dato $X = x_i$** è la distribuzione di Y limitatamente ai soggetti che presentato la modalità x_i di X. Si ottiene associando a ogni modalità y_j di Y la frequenza congiunta di (x_i, y_j).

Modalità (y_j)	Frequenze (n_{ij})
y_1	n_{1i}
y_2	n_{2i}
...	...
y_j	n_{ji}
...	...
y_K	n_{Ki}
Totale	$n_{i\cdot}$

- Ogni riga della tabella a doppia entrata corrisponde a una distribuzione condizionata di Y per una certa modalità X.
- In modo analogo possono essere ottenute le **distribuzioni condizionate di X dato $Y = y_j$** . Ognuna di queste distribuzioni¹⁰ corrisponde a una diversa colonna della tabella a doppia entrata.

Esempio

- Distribuzioni condizionate del reddito al titolo di studio.

Reddito (y_j)	Frequenze (n_{ij})	Relative ($f_{j i}$)	Percentuali ($p_{j i}$)
0-10	88	0,2514	251,4
10-30	142	0,4057	405,7
30-100	120	0,3429	342,9
Totale	350	1	100

Licenza media

Reddito (y_j)	Frequenze (n_{ij})	Relative ($f_{j i}$)	Percentuali ($p_{j i}$)
0-10	9	0,1047	104,7
10-30	39	0,4535	453,5
30-100	38	0,4419	441,9
Totale	86	1	100

Diploma

Reddito (y_j)	Frequenze (n_{ij})	Relative ($f_{j i}$)	Percentuali ($p_{j i}$)
0-10	3	0,0469	46,9
10-30	19	0,2969	296,9
30-100	42	0,6562	656,2
Totale	64	1	100

Laurea

Analisi dell'associazione tra caratteri

- Lo scopo principale dell'analisi di una distribuzione doppia è usualmente quello di stabilire se tra i due caratteri considerati esiste una relazione e se, in particolare, uno dei due (tipicamente X) ha influenza sull'altro (Y). Esempi:
 - relazione tra la provincia di residenza e spesa per beni alimentari;
 - relazione tra voto di maturità e voto a un certo esame universitario;
 - relazione tra sesso e reddito;

- Se X non ha alcuna influenza su Y, allora si dice che Y è indipendente da X. In termini statistici questa situazione si ha quando le distribuzioni condizionate di Y sono equivalenti per ogni modalità di X, cioè hanno le stesse frequenze relative:

f_{j1} = f_{j2} = \dots = f_{jH} \quad j = 1, \dots, K

13

- Quando Y non è indipendente da X, Y dipende da X e quindi i due caratteri si dicono connessi. In pratica, ciò accade ogni volta che la tabella osservata non coincide con quella di indipendenza.
- In particolare, Y dipende perfettamente da X quando la modalità di X determina automaticamente la modalità di Y. Ciò accade quando K ≤ H e si ha una sola frequenza positiva in ogni riga della tabella a doppia entrata mentre le altre frequenze sono tutte nulle.

- Si può dimostrare che si ha indipendenza statistica se e solo se le frequenze congiunte osservate corrispondono alle frequenze teoriche sotto indipendenza

n'_{ij} = \frac{n_{i0} n_{0j}}{n}, \quad i = 1, \dots, H, \quad j = 1, \dots, K

- La tabella di indipendenza si ottiene sostituendo a ogni frequenza osservata (n_{ij}) la corrispondente frequenza di indipendenza (n'_{ij}).
- Sotto indipendenza si hanno le stesse distribuzioni marginali di quelle osservate e la stessa frequenza totale

\sum_{j=1}^K n'_{ij} = n_{i0}, \quad i = 1, \dots, H

\sum_{i=1}^H n'_{ij} = n_{0j}, \quad j = 1, \dots, K

\sum_{j=1}^K \sum_{i=1}^H n'_{ij} = n

14

Esempio

- Per la distribuzione doppia dei caratteri titolo di studio e reddito si ha la seguente distribuzione di indipendenza

	Reddito annuo (x 1.000€)			
Titolo di studio	0-10	10-30	30-100	Totale
Lic. Media	70	140	140	350
Diploma	17,2	34,4	34,4	86
Laurea	12,8	25,6	25,6	64
Totale	100	200	200	500

- Siccome la tabella osservata non coincide con quella di indipendenza, i due caratteri sono dipendenti e quindi si può ragionevolmente ritenere che il titolo di studio influenzi il reddito

- Se la distribuzione doppia fosse come la seguente, si avrebbe *perfetta dipendenza del reddito dal titolo di studio*

	Reddito annuo (x 1.000€)			
Titolo di studio	0-10	10-30	30-100	Totale
Lic. Media	100	0	0	100
Diploma	0	200	0	200
Laurea	0	0	200	200
Totale	100	200	200	500

- Dividendo le *contingenze assolute* per le corrispondenti frequenze sotto indipendenza si ottengono le **contingenze relative**
- Un indice sintetico di connessione è l'**indice chi-quadrato** che è una *somma ponderata* delle contingenze relative al quadrato
- Una *formula alternativa*, che non richiede il calcolo delle contingenze, per l'indice *chi-quadrato* è

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

Misura della connessione tra Y e X

- Il livello di connessione tra i due caratteri è tanto più elevato quanto più la tabella osservata si discosta da quella di indipendenza. Per misurare il livello di connessione si fa quindi uso delle **contingenze (assolute)**

$$c_{ij} = n_{ij} - n'_{ij} = n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}, \quad i = 1, \dots, H, \quad j = 1, \dots, K$$

- La **tabella delle contingenze** si ottiene indicando in una tabella a doppia entrata la contigenza (c_{ij}) corrispondente a ogni coppia di modalità (x_i, y_j). Un'importante proprietà di questa tabella è che la somma delle celle in ogni riga o colonna della tabella è nulla

$$\sum_{j=1}^K c_{ij} = 0, \quad i = 1, \dots, H \quad e \quad \sum_{i=1}^H c_{ij} = 0, \quad j = 1, \dots, K$$

- L'indice **chi-quadrato** assume valori tra 0 (nel caso di indipendenza) e $n \cdot \min[(H - 1), (K - 1)]$ nel caso di perfetta dipendenza di Y da X o di X da Y. Quindi è un indice di connessione bilaterale.
- Un indice di connessione alternativo a χ^2 è la media quadratica ponderata delle contingenze relative

$$\Phi^2 = \frac{1}{n} \sum_{i=1}^H \sum_{j=1}^K \left(\frac{c_{ij}}{n'_{ij}} \right)^2 n'_{ij} = \frac{1}{n} \chi^2$$

che assume valori tra 0 e $\min[(H - 1), (K - 1)]$

- E' quindi possibile costruire un indice *relativo* (che varia tra 0 e 1), chiamato **indice di connessione di Cramér**, come

$$C = \frac{\Phi}{\sqrt{\min[(H - 1), (K - 1)]}} = \sqrt{\frac{\chi^2}{n \cdot \min[(H - 1), (K - 1)]}}$$

Esempio

- Per la distribuzione doppia dei carattere *titolo di studio* e *reddito* si ha la seguente *tabella delle contingenze assolute*

Titolo di studio	Reddito annuo (x 1.000€)			Totale
	0-10	10-30	30-100	
Lic. Media	18	2	-20	0
Diploma	-8,2	4,6	3,6	0
Laurea	-9,8	-6,6	16,4	0
Totale	0	0	0	0

- Il numero di soggetti con *reddito elevato* e *licenza media* è minore di quello atteso sotto indipendenza ($c_{13} = -20$) mentre quello dei soggetti con *laurea* è superiore ($c_{33} = 16,4$).
- Gli indici di connessione sono pari a
$$\chi^2 = 32,13 \quad \Phi^2 = 0,0643$$
- Il massimo dell'indice χ^2 è $2 \times 500 = 1000$ mentre quello dell'indice Φ è 2 e quindi l'indice di Cramér è
$$C = \sqrt{32,13/1000} = 0,179$$