

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:

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

- 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

$$(x_i, y_j), \quad i = 1 \dots, r, \quad j = 1 \dots, c$$

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

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

X / Y	y_1	y_2	y_j	y_c
x_1	n_{11}	n_{12}	n_{1j}	n_{1c}
x_2	n_{21}	n_{22}	n_{2j}	n_{2c}
\vdots
x_i	n_{i1}	n_{i2}	n_{ij}	n_{ic}
\vdots
x_r	n_{r1}	n_{r2}	n_{rj}	n_{rc}

Esempio

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

Nome	Sesso	Regione
M. Rossi	M	Marche
A. Bianchi	F	Calabria
A. Franchi	F	Umbria
G. Gini	M	Piemonte
A. Grandi	F	Marche
P. Lini	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

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

	Reddito annuo (x 1.000€)		
Titolo di studio	0–10	10–30	30–100
Lic. media	88	143	120
Diploma	9	38	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 :

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

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

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

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

$$N = \sum_{i=1}^r \sum_{j=1}^c n_{ij} = \sum_{i=1}^r n_{i\circ} = \sum_{j=1}^c n_{\circ j}$$

X / Y	y_1	y_2	y_j	y_c	Totale
x_1	n_{11}	n_{12}	n_{1j}	n_{1c}	$n_{1\circ}$
x_2	n_{21}	n_{22}	n_{2j}	n_{2c}	$n_{2\circ}$
\vdots	\vdots
x_i	n_{i1}	n_{i2}	n_{ij}	n_{ic}	$n_{i\circ}$
\vdots	\vdots
x_r	n_{r1}	n_{r2}	n_{rj}	n_{rc}	$n_{r\circ}$
Totale	$n_{\circ 1}$	$n_{\circ 2}$	$n_{\circ j}$	$n_{\circ c}$	N

- 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 ($n_{i\circ}$)
x_1	$n_{1\circ}$
x_2	$n_{2\circ}$
\vdots	\vdots
x_i	$n_{i\circ}$
\vdots	\vdots
x_r	$n_{r\circ}$
Totale	N

- In modo analogo si ottiene la ***distribuzione marginale di Y***

Modalità (y_j)	Frequenze ($n_{\circ j}$)
y_1	$n_{\circ 1}$
y_2	$n_{\circ 2}$
\vdots	\vdots
y_j	$n_{\circ j}$
\vdots	\vdots
y_c	$n_{\circ c}$
Totale	N

- Entrambe le distribuzioni possono essere *lette* direttamente dalla tabella a doppia entrata, quando sono presenti i totali (***margini***) di riga e di colonna.

Esempio

- Alla prima distribuzione doppia considerata nell’esempio precedente

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

corrispondono le seguenti distribuzioni marginali di X e Y

Sesso (x_i)	Frequenze ($n_{i\cdot}$)
M	2
F	4
Totale	6

Regione (y_j)	Frequenze ($n_{\cdot j}$)
Calabria	1
Marche	2
Umbria	2
Piemonte	1
Totale	6

- Per la seconda distribuzione doppia considerata nell’esempio precedente, si hanno le seguenti distribuzioni marginali

Titolo di studio (x_i)	Frequenze ($n_{i\cdot}$)
Lic. media	351
Diploma	85
Laurea	64
Totale	500

Reddito (y_j)	Frequenze ($n_{\cdot j}$)
0-10	100
10-30	200
30-100	200
Totale	500

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_{i1}
y_2	n_{i2}
\vdots	\vdots
y_j	n_{ij}
\vdots	\vdots
y_c	n_{ic}
Totale	$n_{i\circ}$

- 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 distribuzioni corrisponde a una diversa colonna della tabella a doppia entrata.

Distribuzioni condizionate relative e percentuali

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

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

- 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_{i1}	$f_{1 i}$	$p_{1 i}$
y_2	n_{i2}	$f_{2 i}$	$p_{2 i}$
⋮	⋮	⋮	⋮
y_j	n_{ij}	$f_{j i}$	$p_{j i}$
⋮	⋮	⋮	⋮
y_c	n_{ic}	$f_{c i}$	$p_{c i}$
Totale	$n_{i\circ}$	1	100

- Analogamente si può ottenere la ***distribuzione condizionata relativa*** (e ***percentuale***) di X dato $Y = y_j$.

Esempio

- Per la seconda distribuzione doppia considerata nell’esempio precedente si hanno le seguenti distribuzioni condizionate del *reddito* dato il *titolo di studio* dalle quali si può dedurre che il secondo carattere è influenzato dal primo.

Titolo di studio = licenza media

Reddito (y_j)	Frequenze ($n_{\circ j}$)	Relative ($f_{j i}$)	Percentuali ($p_{j i}$)
0-10	88	0,2507	25,07
10-30	143	0,4074	40,74
30-100	120	0,3419	34,19
Totale	351	1,000	100,00

Titolo di studio = diploma

Reddito (y_j)	Frequenze ($n_{\circ j}$)	Relative ($f_{j i}$)	percentuali ($p_{j i}$)
0-10	9	0,1058	10,58
10-30	38	0,4471	44,71
30-100	38	0,4471	44,71
Totale	85	1,000	100,00

Titolo di studio = laurea

Reddito (y_j)	Frequenze ($n_{\circ j}$)	Relative ($f_{j i}$)	Percentuali ($p_{j i}$)
0-10	3	0,0469	4,69
10-30	19	0,2969	29,69
30-100	42	0,6562	65,62
Totale	64	1,000	100,00

Analisi dell’associazione tra due 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* (o *percentuali*):

$$f_{j|1} = f_{j|2} = \dots = f_{j|r}, \quad j = 1, \dots, c.$$

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

$$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

- La ***tabella di indipendenza*** si ottiene sostituendo a ogni frequenza osservata (n_{ij}) la corrispondente frequenza di indipendenza (\hat{n}_{ij}).

- Sotto *indipendenza* si hanno le stesse distribuzioni marginali di quelle osservate e la stessa frequenza totale

$$\sum_{j=1}^c \hat{n}_{ij} = n_{i\circ}, \quad i = 1, \dots, r$$

$$\sum_{i=1}^r \hat{n}_{ij} = n_{\circ j}, \quad j = 1, \dots, c.$$

$$\sum_{i=1}^r \sum_{j=1}^c \hat{n}_{ij} = N$$

- 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 $c \leq r$ e si ha una sola frequenza positiva in ogni riga della tabella a doppia entrata mentre le altre frequenze sono tutte nulle.

Esempio

- Per la distribuzione doppia del carattere *titolo di studio* e *reddito* si ha la seguente *distribuzione di indipendenza*

Titolo di studio	Reddito annuo (x 1.000€)			Totale
	0-10	10-30	30-100	
Lic. media	70,2	140,4	140,4	351
Diploma	17,0	34,0	34,0	85
Laurea	12,8	25,6	25,6	64
Totale	100,0	200,0	200,0	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*

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

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)*

$$c_{ij} = n_{ij} - \hat{n}_{ij} = n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

- La *tabella delle contingenze* si ottiene indicando in una tabella a doppia entrata la contingenza (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

$$\sum_{j=1}^c c_{ij} = 0, \quad i = 1, \dots, r \quad \text{e} \quad \sum_{i=1}^r c_{ij} = 0, \quad j = 1, \dots, c.$$

- Dividendo le *contingenze assolute* per le corrispondenti frequenze sotto indipendenza si ottengono le *contingenze (relative)*

$$\frac{c_{ij}}{\hat{n}_{ij}} = \frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

- Un indice sintetico di connessione è l’*indice chi-quadro* che è una *somma ponderata* delle contingenze relative al quadrato

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \left(\frac{c_{ij}}{\hat{n}_{ij}} \right)^2 \hat{n}_{ij} = \sum_{i=1}^r \sum_{j=1}^c \frac{c_{ij}^2}{\hat{n}_{ij}}$$

- Una *formula alternativa*, che non richiede il calcolo delle contingenze, per l’indice *chi-quadro* è

$$\chi^2 = N \left(\sum_{i=1}^r \sum_{j=1}^c \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right)$$

- L’indice *chi-quadro* assume valori tra 0 (nel caso di indipendenza) e

$$N \cdot \min[(r-1), (c-1)]$$

nel caso di perfetta dipendenza di Y da X o di X da Y . Quindi è un indice di connessione *bilaterale*.

- Per rendere l’indice chi-quadro di Person relativo (con valori compresi tra zero ed uno) lo si divide per il suo massimo teorico, ottenendo:

$$\tilde{\chi}^2 = \frac{\sum_{i=1}^r \sum_{j=1}^c \frac{c_{ij}^2}{\hat{n}_{ij}}}{N \cdot \min[(r-1), (c-1)]} = \frac{\left(\sum_{i=1}^r \sum_{j=1}^c \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right)}{\cdot \min[(r-1), (c-1)]}$$

Esempio

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

Titolo di studio	Reddito annuo (x 1.000€)			Totale
	0-10	10-30	30-100	
Lic. media	17,8	2,6	-20,4	0
Diploma	-8,0	4,0	4,0	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 ($c_{13} = -20,4$) mentre quello dei soggetti con *laurea* è superiore ($n_{33} = 16,4$).
- L’ indice di connessione chi-quadro è pari a

$$\chi^2 = 31,9$$

Il massimo dell’indice χ^2 è $2 \cdot 500 = 1000$, il che indica una moderato livello di connessione.

L’indice chi-quadro di Pearson relativo assume un valore prossimo allo zero, confermando quanto detto sopra:

$$\tilde{\chi}^2 = \frac{31,9}{500 * 2} = 0,0319$$