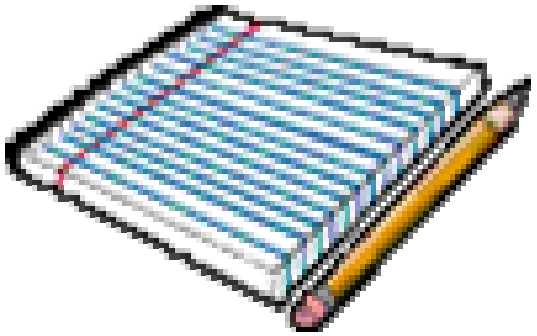
Lezione 8

Relazioni Statistiche



Struttura Generale

Distribuzioni doppie

c modalità del carattere Y

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}	$n_{2\circ}$
⋮	....	....	....	....	....	....	⋮
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{ic}	$n_{i\circ}$
⋮	....	....	....	....	....	....	⋮
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

Numero totale di unità statistiche

r modalità del carattere X

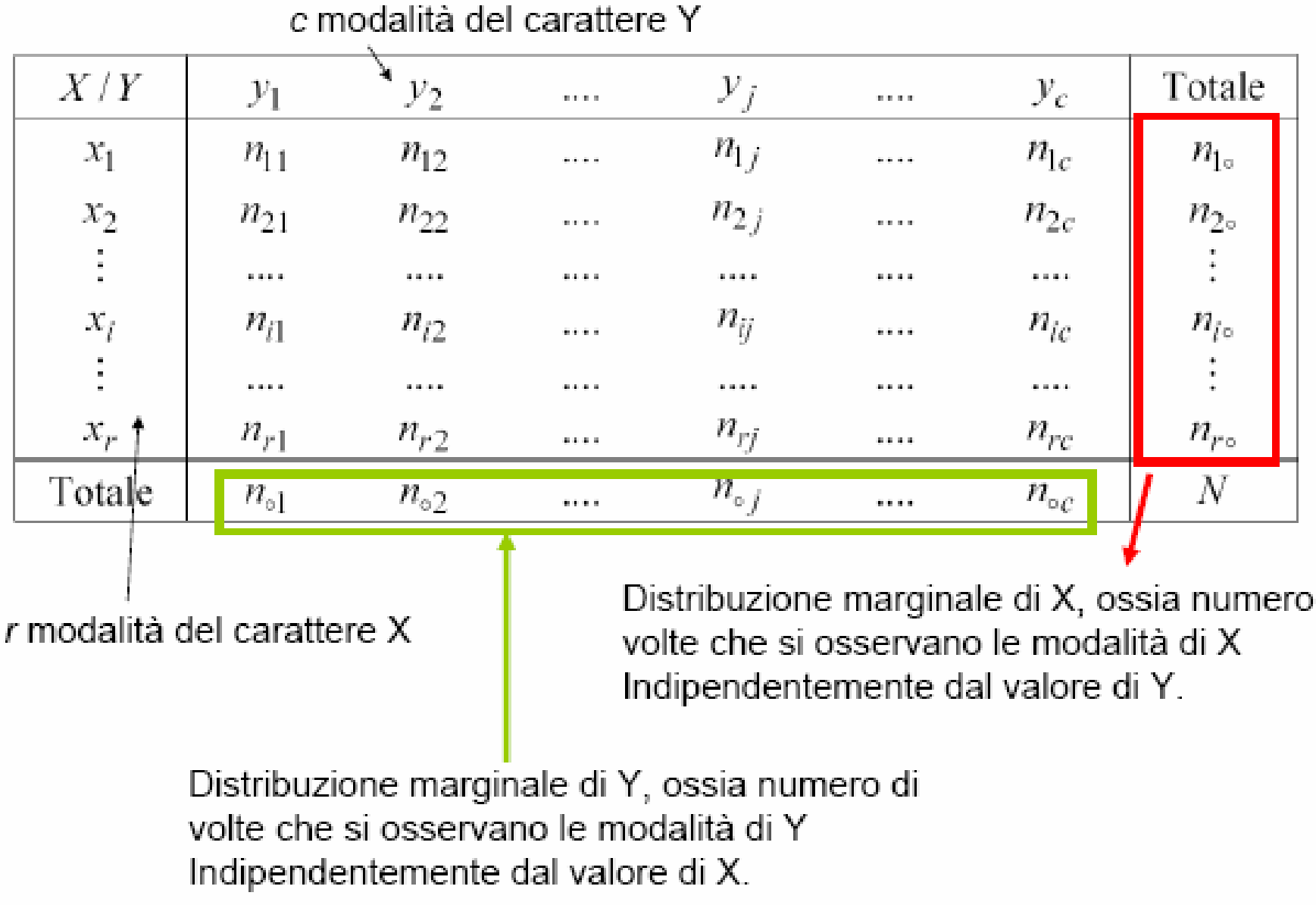
Frequenza congiunta (assoluta) con cui si osserva la modalità i-ima di X congiuntamente alla modalità j-ima di Y, ossia la coppia (*xi*,*yj*) è osservata *nij* volte!!

Tabelle a doppia entrata

Dopo il disastro, una commissione d'inchiesta del *British Board of Trade* ha compilato una lista di tutti i 1316 passeggeri del Titanic con alcune informazioni aggiuntive riguardanti: l'esito (salvato, non salvato), la classe (I, II, III) in cui viaggiavano, il sesso, l'eta,. . . .

Passeggero	Classe	Esito
nome 1	II	salvato
nome 2	III	non salvato
nome 3	I	non salvato
⋮	⋮	⋮
nome 1316	III	salvato

Distribuzioni marginali



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}	$n_{2\circ}$
⋮	....	....	....	....	....	....	⋮
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{ic}	$n_{i\circ}$
⋮	....	....	....	....	....	....	⋮
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

- Quindi avremo:

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

Frequenze marginali di X

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

Frequenze marginali di Y

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

La numerosità del colletivo è data dalla somma di tutte le frequenze congiunte, o di tutte le frequenze marginali.

Indici di Connessione

Un operatore statistico bivariato è una procedura di calcolo che considera due variabili e sintetizza l'informazione sulla loro distribuzione congiunta in uno scalare.

Gli operatori di connessione producono uno scalare sempre positivo; assumono valore zero in assenza di connessione e maggiore di zero in presenza di connessione tra le due variabili.

La natura della relazione deve essere ipotizzata dal ricercatore

Quello che dobbiamo notare è che la natura della relazione causale tra due o più variabili discende da un ragionamento che si pone a livello teorico e non a livello empirico: un operatore statistico di per sé non informa sulla natura logica di una relazione causale tra le variabili empiriche.

L’indice Chi-quadrato

Il più importante operatore di connessione che rappresenta l’intensità della relazione tra due variabili categoriali è chiamato “chi quadrato” (χ^2).

Per costruire un operatore di connessione tra le due variabili prendiamo come modello di riferimento *l’assenza di relazione* e calcoliamo quanto le frequenze osservate si discostano dalle frequenze teoriche calcolate sulla base dell’ipotesi di completa indipendenza.

L’indice Chi-quadrato

Più le frequenze empiriche si allontanano dalle frequenze teoriche più è elevato il grado di connessione tra le variabili.
L’operatore chi quadrato si basa proprio sulla differenza tra frequenze empiriche e teoriche.
Più le precisamente, l’operatore chi quadrato si calcola come:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

Frequenze Teoriche

Le frequenze teoriche $(\hat{n}_{ij})$ sono riportate tra parentesi e sono calcolate come:

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

Campo di variazione

- L'operatore chi quadrato assume come valore minimo zero e come valore massimo il minore dei seguenti due valori: $N\,(I - 1)$ e $N\,(J - 1)$.

- Il valore massimo dipende dunque dall'ampiezza del collettivo e dal numero di righe e colonne della tabella.

Misure basate sul chi quadrato

Una misura di connessione basata sul chi quadrato e indipendente dal numero dei casi è stata proposta da Pearson e si calcola come:

$$\Phi^2 = \frac{X^2}{N}$$

Questa misura di connessione viene chiamata phi quadrato, assume coma valore minimo 0. Il suo valore massimo è funzione del numero di modalità delle variabili:

- $\min (I - 1; J - 1)$

Misure basate sul chi quadrato

Un'altra misura di connessione è data dal *T* di Tschuprov:

$$T = \frac{\Phi^2}{\sqrt{(J-1)(I-1)}}$$

Il *T* di Tschuprov assume il valore di 1 (con tabelle quadrate) nel caso di dipendenza reciproca perfetta.

Misure basate sul chi quadrato

Un ultimo operatore basato sulla relativizzazione del chi quadrato è il *V* di Cramer:

$$V = \frac{\chi^2}{N \times \min\left[(J-1);(I-1)\right]}$$

Assume sempre valori compresi tra 0 e 1.

Dipendenza in media

Medie parziali

X/Y	10	20	30	40	
A	7	11	6	3	27
B	8	11	15	18	52
	15	22	21	21	79

E' possibile determinare la media totale e le medie parziali

$$\bar{y}_i = \sum \frac{y_j n_{ij}}{n_{i\bullet}}$$

$$\bar{y} = \frac{\sum_j y_j n_{\bullet j}}{n}$$

Medie parziali

X/Y	10	20	30	40	
A	7	11	6	3	27
B	8	11	15	18	52
	15	22	21	21	79

E' possibile determinare la media totale e le medie parziali

$$\overline{y}_A = \frac{(10\times 7)+(20\times 11)+(30\times 6)+(40\times 3)}{27} = 21.85$$

$$\overline{y}_B = \frac{(10\times 8)+(20\times 11)+(30\times 15)+(40\times 18)}{52} = 28.26$$

$$\overline{y} = \frac{(10\times 15)+(20\times 22)+(30\times 21)+(40\times 21)}{79} = 26.07$$

Medie parziali

Si può notare che…

$$\bar{y} = \frac{(\bar{y}_1 \times n_{1\bullet}) + (\bar{y}_2 \times n_{2\bullet})}{n} =$$

$$= \frac{(21.85 \times 27) + (28.26 \times 52)}{79} = 26.07$$

Esempio

X/Y	1	2	3	
M	8	3	3	14
F	5	3	0	8
	13	6	3	22

Y=numero universitari in famiglia
X= genere

Calcoliamo la media totale e le medie parziali

$$\bar{y} = \frac{13 + 12 + 9}{22} = 1.54$$

$$\bar{y}_M = \frac{8 + 6 + 9}{14} = 1.65 \qquad \bar{y}_F = \frac{5 + 6 + 0}{8} = 1.375$$

Devianza e sua scomposizione

$$Dev(Y) = \sum_i \sum_j (y_j - \bar{y})^2 n_{ij} =$$

$$= \sum_i \sum_j (y_j - \bar{y}_i + \bar{y}_i - \bar{y})^2 n_{ij} =$$

$$= \sum_i \sum_j (y_j - \bar{y}_i)^2 n_{ij} + \sum_i \sum_j (\bar{y}_i - \bar{y})^2 n_{ij} +$$

$$+ 2 \sum_i \sum_j (y_j - \bar{y}_i)(y_j - \bar{y}_i) n_{ij} =$$

$$0$$

Devianza e sua scomposizione

$$= \sum_i \left[\sum_j (y_j - \bar{y}_i)^2 n_{ij} \right] +$$

$$+ \sum_i (y_i - \bar{y})^2 \sum_j n_{ij} =$$

$$= \sum_i \left[Dev(Y \mid X = x_i) \right] + \sum_i (\bar{y}_i - \bar{y})^2 n_{i\bullet} \Rightarrow$$

$$\boxed{Dev(Y) = Dev(W) + Dev(B)}$$

Rapporto di correlazione

$$\eta^2_{y/x} = \frac{Dev(B)}{Dev(Y)} = \frac{\sum_i (\bar{y}_i - \bar{y}) n_{i\bullet}}{\sum_i (y_j - \bar{y}) n_{\bullet j}}$$

$$\eta^2_{x/y} = \frac{Dev(B)}{Dev(X)} = \frac{\sum_i (\bar{x}_j - \bar{x}) n_{\bullet j}}{\sum_i (x_j - \bar{x}) n_{i\bullet}}$$

$\eta^2_{y/x} \neq \eta^2_{x/y}$

In generale

Il rapporto di correlazione è un indice NON SIMMETRICO