

Corso di Statistica Facoltà di Economia

a.a. 2010-2011

francesco mola

Lezione n° 8

Sommario

- Variabili doppie
- L'indipendenza
 - In distribuzione
 - In media
- La connessione
 - Unilaterale
 - Bilaterale

lez8_a.a. 2010-2011

statistica-francesco mola

2

Variabili doppie (generalità)

$Y \backslash X$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_h	
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$	n

lez8_a.a. 2010-2011

statistica-francesco mola

3

Variabili doppie (generalità)

$Y \backslash X$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_h	
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}	$n_{2\bullet}$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$	n

variabile di colonna

lez8_a.a. 2010-2011

statistica-francesco mola

4

Variabili doppie (generalità)

$Y \backslash X$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_h	
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$	n

variabile di riga

Variabili doppie (generalità)

$Y \backslash X$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_h	
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$	n

j - esima modalità di Y

Variabili doppie (generalità)

$Y \backslash X$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_h	
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$	n

i - esima modalità di X

Variabili doppie (generalità)

frequenza del contemporaneo verificarsi di x_i e y_j

x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$	n

Variabili doppie (generalità)

$Y \backslash X$	frequenza marginale di riga					
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$
						n

Variabili doppie (generalità)

$Y \backslash X$	frequenza marginale di colonna					
x_1	n_{11}	n_{12}	$\vdots$	n_{1j}	$\vdots$	n_{1h}
x_2	n_{21}	n_{22}	$\vdots$	n_{2j}	$\vdots$	n_{2h}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{ih}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\vdots$	n_{kj}	$\vdots$	n_{kh}
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet h}$
						n

...per frequenza relative

- Se dividiamo ogni cella per n otteniamo la tabella doppia per frequenze relative
- Valgono per estensione tutte le proprietà viste per le distribuzioni semplici

Caratteristiche principali

n_{ij} = frequenza congiunta

$$n_{i\bullet} = \text{totali di riga} = \sum_{j=1}^h n_{ij}$$

$$n_{\bullet j} = \text{totali di colonna} = \sum_{i=1}^k n_{ij}$$

$$n = \sum_{i=1}^k \sum_{j=1}^h n_{ij} = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$$

...per frequenze relative

f_{ij} = frequenza relativa congiunta

$f_{i\bullet}$ = totali di riga = $\sum_{j=1}^h f_{ij}$

$f_{\bullet j}$ = totali di colonna = $\sum_{i=1}^k f_{ij}$

$$1 = \sum_{i=1}^k \sum_{j=1}^h f_{ij} = \sum_i f_{i\bullet} = \sum_j f_{\bullet j}$$

Indici di connessione

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Formulazione alternativa

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{ij}} + \sum_i \sum_j \frac{\hat{n}_{ij}^2}{\hat{n}_{ij}} - 2 \sum_i \sum_j \frac{n_{ij} \hat{n}_{ij}}{\hat{n}_{ij}} =$$

...formulazione alternativa (cont.)

$$= \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{i\bullet} \cdot \hat{n}_{\bullet j}} + \sum_i \sum_j \hat{n}_{ij} - 2 \sum_i \sum_j n_{ij} =$$

$$= n \sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}} + n - 2n =$$

$$= n \left[\frac{\sum_i \sum_j n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}} - 1 \right]$$

Esempio 1

X/Y	y1	y2	
x1	40	10	50
x2	25	18	43
x3	12	9	21
	77	37	114

Costruiamo la tabella delle frequenze teoriche

X/Y	y1	y2	
x1	33,8	16,2	50
x2	29,0	14,0	43
x3	14,2	6,8	21
	77	37	114

$$\frac{50 \times 77}{114}$$

..esempio (cont.)

$$\chi^2 = \left[\frac{(40 - 33.8)^2}{33.8} \right] + \left[\frac{(10 - 16.2)^2}{16.2} \right] + \dots = 6.26$$

oppure con la formulazione alternativa

$$\chi^2 = \frac{40^2}{50 \times 77} + \frac{10^2}{50 \times 37} + \dots = 6.26$$

...il fondamento teorico

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

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

Quindi le frequenze teoriche sono:

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

$$c_{ij} = n_{ij} - \hat{n}_{ij}$$

...il fondamento teorico (cont...)

$$\sum_i \sum_j c_{ij} = \sum_i \sum_j (n_{ij} - \hat{n}_{ij}) = \sum_i \sum_j n_{ij} - \sum_i \sum_j \hat{n}_{ij} =$$

$$= n - \frac{1}{n} \sum_i \sum_j n_{i\cdot} n_{\cdot j} = n - \frac{1}{n} \sum_i n_{i\cdot} \sum_j n_{\cdot j} =$$

$$= n - \frac{1}{n} n \times n$$

$$= n - n = 0$$

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

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

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

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

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

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

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

Medie parziali

Si può notare che...

$$\begin{aligned} \bar{y} &= \frac{(\bar{y}_1 \times n_{1\cdot}) + (\bar{y}_2 \times n_{2\cdot})}{n} \\ &= \frac{(21.85 \times 27) + (28.26 \times 52)}{79} = 26.07 \end{aligned}$$

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

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

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

Devianza e sua scomposizione

$$\begin{aligned} Dev(Y) &= \sum_i \sum_j (y_j - \bar{y})^2 n_{ij} = \\ &= \sum_i \sum_j (y_j - \bar{y}_i + \bar{y}_i - \bar{y})^2 n_{ij} = \\ &= \sum_i \sum_j (y_j - \bar{y}_i)^2 n_{ij} + \sum_i \sum_j (\bar{y}_i - \bar{y})^2 n_{ij} + \\ &\quad + 2 \sum_i \sum_j (y_j - \bar{y}_i)(\bar{y}_i - \bar{y}) n_{ij} = \end{aligned}$$

0

Devianza e sua scomposizione

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

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

$$Dev(Y) = Dev(W) + Dev(B)$$

Interpretazione

- La devianza “Between” descrive la variabilità “tra” i gruppi, ossia la variabilità delle medie parziali di Y rispetto alla media generale.
- La devianza “Within” descrive la variabilità “interna” ai gruppi, ossia la somma delle variabilità della Y in ciascun gruppo.
- Quanto più i gruppi sono ben discriminati tanto maggiore è la componente di variabilità esterna rispetto a quella interna. Ciò implica che la variabile X “spiega” il comportamento della Y.
- La variabile X è detta di stratificazione in quanto dalle sue modalità si determinano gli strati o gruppi parziali del collettivo.

Rapporto di correlazione

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

$$\eta_{x/y}^2 = \frac{Dev(B)}{Dev(X)} = \frac{\sum_i (\bar{x}_j - \bar{x})^2 n_{\bullet j}}{\sum_i \sum_j (x_i - \bar{x})^2 n_{ij}}$$

$$\eta_{y/x}^2 \neq \eta_{x/y}^2$$

In generale

Il rapporto di correlazione è un indice NON SIMMETRICO

Rapporto di correlazione (cont.)

$$\eta_{y/x} = \sqrt{\eta_{y/x}^2}$$

Riconsideriamo l'esempio

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

$$\bar{y}_A = 21.85 \quad \bar{y}_B = 28.26 \quad \bar{y} = 26.07$$

Calcoliamo il rapporto di correlazione

$$\text{Dev}(B) = (21.85 - 26.07)^2 \times 27 + (28.26 - 26.07)^2 \times 52 = 480.83 + 249.4 = 730.22$$

$$\text{Dev}(Y) = (10 - 26.07)^2 \times 15 + (20 - 26.07)^2 \times 22 + \dots = 9083.54$$

Riconsideriamo l'esempio

$$\eta^2_{y/x} = \frac{730.22}{9083.54} = 0.08039$$

$$\eta_{y/x} = \sqrt{\frac{730.22}{9083.54}} = 0.28353$$

Quando non c'è indipendenza, si parla di connessione

Tipi di connessione

X/Y	y1	y2	y3	
x1	0	5	0	5
x2	2	0	0	2
x3	0	0	4	4
	2	5	4	11

Connessione Perfetta Bilaterale (solo per tabelle quadrate!)

Ad ogni modalità di X corrisponde una ed una sola modalità di Y e viceversa

Tipi di connessione (cont.)

X/Y	y1	y2	y3	
x1	0	5	2	7
x2	2	0	1	3
x3	0	2	4	6
	2	7	7	16

In questo esempio, NON ESISTE connessione perfetta!

Tipi di connessione (cont.)

X/Y	y1	y2	
x1	5	0	5
x2	0	1	1
x3	2	0	2
	7	1	8

Connessione Unilaterale di Y da X

Per ogni modalità della X esiste una sola modalità della Y con frequenze diverse da 0

Tipi di connessione (cont.)

X/Y	y1	y2	y3	
x1	2	0	5	7
x2	0	4	0	4
	2	4	5	11

Connessione Unilaterale di X da Y

Per ogni modalità della Y esiste una sola modalità della X con frequenze diverse da 0

La connessione Unilaterale è solo per tabelle rettangolari!

Tipi di connessione (cont.)

X/Y	1	2	4	
M	4	1	2	7
F	2	4	1	7
	6	5	3	14

Esiste Indipendenza in distribuzione?

Verifichiamo:

Tipi di connessione (cont.)

X/Y	1	2	4	
M	4/7	1/7	2/7	
F	2/7	4/7	1/7	
	6/14	5/14	3/14	

X/Y	1	2	4	
M	0,57	0,14	0,29	
F	0,29	0,57	0,14	
	0,43	0,36	0,21	

Le Distribuzioni Parziali sono diverse dalla distribuzione Marginale quindi non esiste indipendenza in distribuzione

NON ESISTE indipendenza in distribuzione, verifichiamo se esiste l'indipendenza in media

$$\bar{y}_M = \frac{(1 \times 4) + (2 \times 1) + (4 \times 2)}{7} = 2$$

$$\bar{y}_F = \frac{(1 \times 2) + (2 \times 4) + (4 \times 1)}{7} = 2$$

$$\bar{y} = \frac{(1 \times 6) + (2 \times 5) + (4 \times 3)}{14} = 2$$

Le MEDIE PARZIALI sono uguali alla media generale, per cui esiste indipendenza in media