Analisi delle distribuzioni doppie: dipendenza

Cicchitelli Cap. 9

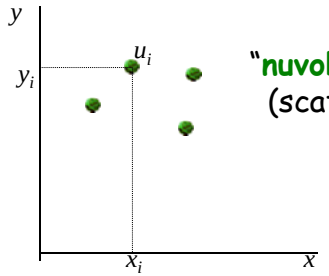
Variabili statistiche bivariate

Unità statistiche	X	Y
u_1	x_1	y_1
u_2	x_2	y_2
:	:	:
u_i	x_i	y_i
:	:	:
u_n	x_n	y_n

v.s. bivariata quantitativa

$$(X,Y): U \to \mathcal{X} \times \mathcal{Y} \subseteq \Re^2$$

X,Y sono le COMPONENTI di (X,Y)



"nuvola" di punti (scatter plot)

Distribuzioni bivariate

Supporto (v.s. quantitativa)

$$S_{x,y} = \{(X,Y) \in \mathcal{X} \times \mathcal{Y} \subseteq \Re^2 : x = X(u), y = Y(u), u \in U\}$$

Distribuzione di frequenza bivariata

$$(X,Y): \{(x_i, y_j, f_{ij}), i = 1,2,\ldots I,\ j = 1,2,\ldots J\}$$

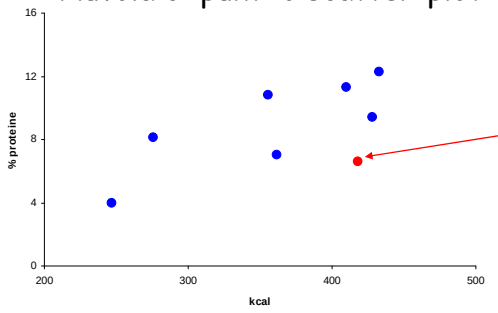
Frequenze congiunte

$$f_{ij} = \frac{|\{u \in U : X(u) = x_i, Y(u) = y_j\}|}{N} = pr(X = x_i, Y = y_j)$$

$n_{ij}=Nf_{ij}$ Frequenze assolute congiunte

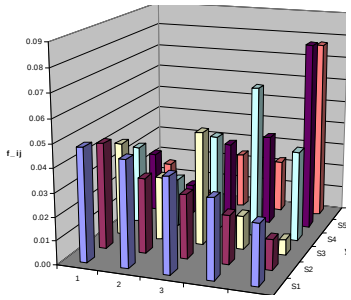
Rappresentazione grafica di una successione bivariata

Nuvola di punti o scatter-plot



alimento	*energia kcal*	*%proteine*
pizza	247	4
pane	276	8.1
pasta	356	10.8
riso	362	7
fette	410	11.3
biscotti	418	6.6
crackers	428	9.4
grissini	433	12.3
Totale	2930	69.5

Rappresentazione grafica di una serie bivariata



Altezza proporzionale alle frequenze congiunte

x	y						
	1	*2*	*3*	*4*	*5*	*6*	*TOT*
1	4.82	4.49	4.01	3.37	2.57	1.61	20.87
2	4.49	3.21	2.73	2.09	1.28	0.32	14.13
3	4.01	2.73	4.82	4.17	3.37	2.41	21.51
4	3.37	2.09	1.44	6.42	3.85	2.25	19.42
5	2.57	1.28	0.64	3.85	8.03	7.70	24.08
TOTALE	19.26	13.80	13.64	19.90	19.10	14.29	100.00

Per (X,Y) quantitative: stereogramma con volume proporzionale alle frequenze congiunte

Funzione di ripartizione congiunta

$$F_{xy}(x,y) = pr(X \le x, Y \le y)$$

Distribuzione marginale

$$S_x = \{X \in \Re : (x,y) \in S_{xy}, \text{ per qualche } y \in \Re\}$$

$$f_x(x) = \begin{cases} \int_{-\infty}^{+\infty} f_{xy}(x,y)dy & \text{v.s. continua} \\ \sum_{y \in S_y} f_{xy}(x,y) = \sum_{j=1}^{J} f_{ij} & \text{v.s. discreta} \end{cases}$$

Distribuzioni condizionate

$$Y|X=x_i : \{(y_j, f_{j|i}), j=1,2,\ldots,J\}$$

$n_{j|i} = n_{ij}$ *freq. assolute condizionate*

$f_{j|i} = n_{ij}/n_i$ *freq. relative condizionate*

$\Sigma_j f_{j|i} = 1$

Analogamente possono essere costruite le *J* distribuzioni condizionate di *X*.

Indipendenza in distribuzione

$$\forall x_i \in S_x : Y \mid X = x_i \sim Y \mid X = x_{i'}$$

TEOREMA

Condizione necessaria e sufficiente perché Y sia indipendente da X

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

Oppure: $n_{ij} = (n_{i.} \times n_{.j})/n$

Dipendenza funzionale univoca

Y dipende funzionalmente da *X* quando tutte le distribuzioni condizionate di *Y* sono degeneri, con $J \leq I$

J n. modalità di *Y*
I n. modalità di *X*

		y	
X	*y1*	*y2*	TOT
x1	4	0	4
x2	0	2	2
x3	3	0	3
TOT	7	2	9

Viceversa, *X* dipende funzionalmente da *Y* quando tutte le distribuzioni condizionate di *X* sono degeneri, con $I \leq J$

Dipendenza funzionale biunivoca

I caratteri *X* e Y dipendono funzionalmente uno dall'altro, con *I*=J

		y		
X	*y1*	*y2*	*y3*	TOT
x1	0	5	0	5
x2	3	0	0	3
x3	0	0	1	1
TOT	3	5	1	9

Indice di contingenza quadratica media (Pearson)

$$\psi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i.} \times n_{.j}} - 1$$

- In tabelle 2*2

  $-1 \leq \psi \leq +1$ segno di $(n_{11}n_{22}\text{-}n_{12}n_{21})$

- In tabelle I*J

  $0 \leq \psi^2 \leq \min[(I\text{-}1),(J\text{-}1)]$

Dipendenza funzionale esatta

(1) $X|Y=y_j \sim D(x_0)$, j=1,2,…,J

X	y			
	y_1	y_2	y_3	Tot
x_1	n_{11}	0	n_{13}	$n_{1.}$
x_2	0	n_{22}	0	$n_{2.}$
Tot	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

⇒ ψ^2=I-1

(2) $Y|X=x_i \sim D(y_0)$, i=1,2,…,I

⇓

ψ^2=J-1

X	y		
	y_1	y_2	Tot
x_1	n_{11}	0	$n_{1.}$
x_2	0	n_{22}	$n_{2.}$
X_3	n_{31}		$n_{3.}$
Tot	$n_{.1}$	$n_{.2}$	N

Indice C di Cramèr

$$C = \frac{\psi}{\sqrt{\min\left[(I-1),(J-1)\right]}}$$

$0 \le C \le 1$

C =0 indipendenza in distribuzione

C =1 dipendenza funzionale esatta

Indice χ^2 di Pearson

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$$

$$n_{ij}^* = \left(n_{i.} \times n_{.j}\right)/n \qquad \chi^2 = N\psi^2$$

χ^2 =0 se X e Y sono indipendenti

$\chi^2 \Rightarrow \infty$ per N$\rightarrow\infty$

Formula per il calcolo

$$\chi^2 = N\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i.} \times n_{.j}} - 1\right]$$

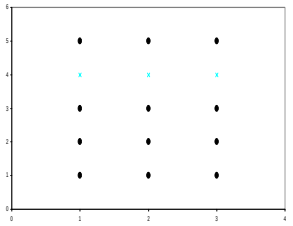
$$N = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$$

χ^2 =0 se X e Y sono indipendenti

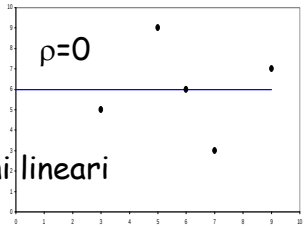
$\chi^2 \Rightarrow \infty$ per N$\rightarrow\infty$

Relazione tra caratteri

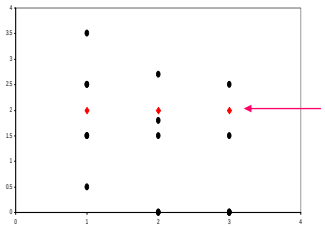
Al variare di X in S_x, si osserva una variazione sistematica della distribuzione di Y?



$f_{Y|X=x}$ costante



No variazioni lineari



M(Y|X=x) costante

Studio della relazione tra caratteri

Studio della dipendenza di Y da X:
come varia Y|X=x al variare di x in S_x?

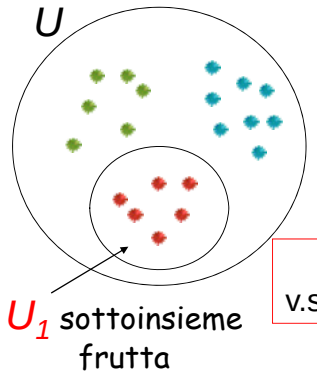
- Indici di posizione
- Variabilità
- Quantili

Piccolo D. (1999) §7.4 (pagg. 189-194)

Dipendenza in media

X gruppo
Y proteine

medie per gruppo	
frutta	0,67
verdura	1,78
cereali	8,69
TOTALE	4,21



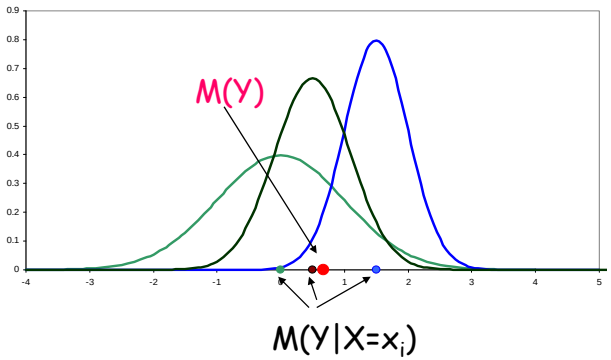
Y|X=1
v.s. condizionata

alimento	gruppo	%proteine
mele	1	0.2
uva	1	0.5
limoni	1	0.6
arance	1	0.7
pesche	1	0.8
banane	1	1.2
pomodori	2	1
carote	2	1.1
zucchine	2	1.3
lattuga	2	1.8
patate	2	2.1
spinaci	2	3.4
pizza	3	4
biscotti	3	6.6
riso	3	7
pane	3	8.1
crackers	3	9.4
pasta	3	10.8
fette bisc	3	11.3
grissini	3	12.3

Media marginale e medie condizionate

$$M(Y)=\frac{1}{N}\sum_{i=1}^{I}M(Y|X=x_i)n_{i.}=\sum_{i=1}^{I}M(Y|X=x_i)\,f_{i.}$$

$$n_{i.}=\sum_{j=1}^{J}n_{ij},\ f_{i.}=n_{i.}/N$$



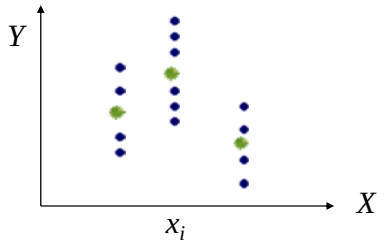
Calcolo della media marginale attraverso le medie condizionate

Frequenze relative				
n. figli	NORD	CENTRO	SUD	TOTALE
0	0.30	0.20	0.10	0.21
1	0.30	0.40	0.20	0.32
2	0.20	0.20	0.25	0.21
3	0.10	0.10	0.30	0.14
4	0.05	0.10	0.10	0.08
5	0.05	0.00	0.05	0.03
TOTALE	1.00	1.00	1.00	1.00
$n_{i.}$	150	200	100	450
$f_{i.}$	0.33	0.44	0.22	1.00

fj*yj				
n. figli	NORD	CENTRO	SUD	TOTALE
0	0.00	0.00	0.00	0.00
1	0.30	0.40	0.20	0.32
2	0.40	0.40	0.50	0.42
3	0.30	0.30	0.90	0.43
4	0.20	0.40	0.40	0.33
5	0.25	0.00	0.25	0.14
TOTALE	1.45	1.50	2.25	1.65
$n_{i.}$	150	200	100	450
$f_{i.}$	0.33	0.44	0.22	1.00

M(Y)=1.45×0.33+ 1.50×0.44+ 2.25×0.22=1.65

Indipendenza in media



Y indipendente in media da X
se $M(Y|X=x)=c$, $\forall\ x \in S_x$

Misura dell'indipendenza in media

- Misura della diversità tra le medie condizionate
  - dato $\bar{y} = \frac{1}{N}\sum_{i=1}^{I} \bar{y}_i n_{i.}$ $\quad \bar{y} = M(Y),\ \bar{y}_i = M(Y|X = x_i) n_{i.}$
  - Consideriamo la devianza $D_I = \sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2 n_{i.}$

D_I=0 *sse* tutte le medie condizionate sono uguali tra loro

  - Conviene *normalizzare* la devianza per ottenere un indice che varia tra 0 e 1

Scomposizione della devianza

- Dev tot=media dev condizionate+dev medie

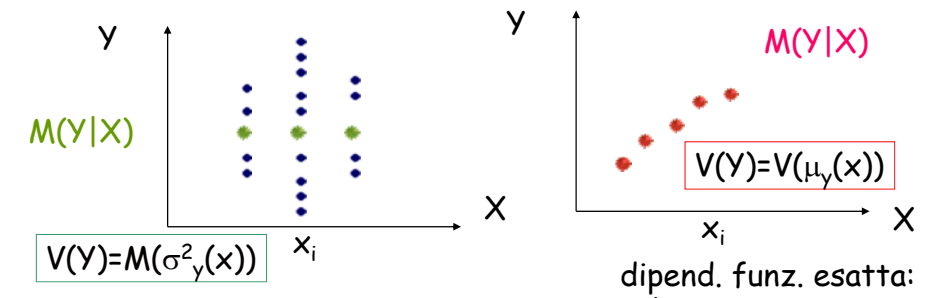
$$\sum_{i=1}^{I}\sum_{j=1}^{J} (y_j - \bar{y})^2 n_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} (y_j - \bar{y}_i)^2 n_{ij} + \sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2 n_{i.}$$

$$D_Y = D_R + D_S$$

Scomposizione della varianza

$$V(Y) = M[\sigma_y^2(x)] + V[\mu_y(x)]$$

Varianza residua — Varianza spiegata



Indipendenza in media: $\mu_y(x) \sim D(\mu_y)$

dipend. funz. esatta: $Y|X \sim D(\mu_y(x))$

Rapporto di correlazione

$$\eta^2_{y|x} = \frac{V(M(Y \mid x))}{V(Y)}, 0 \le \eta^2_{y|x} \le 1$$

Indipendenza in media

se M(Y|X=x)=M(Y) $\forall\ x \in S_x \Rightarrow \eta^2_{y|x}=0$

Dipendenza funzionale esatta

se V(Y|X=x)=0 $\forall\ x \in S_x \Rightarrow \eta^2_{y|x}=1$

Indagini ISTAT nel 1982 secondo l'area di studio e il numero di questionari per indagine

Area di studio	<=1	1-|5	5-|10	10-|50	>50	TOT	M(Y|X=x)	V(Y|X=x)
Demografica	0	1	1	0	6	8	76.31	1684.56
Sociale	12	12	10	11	14	59	31.37	1568.34
Economica	39	24	8	14	12	97	18.38	1036.18
Altre	5	0	0	1	0	6	5.67	118.42
Totale	56	37	19	26	32	170	25.17	1395.11

$$\eta^2_{y|x} = \frac{V(M(Y \mid X))}{V(Y)} = \frac{176.1}{1395.11} = 0.126$$

$$\eta = \sqrt{\eta^2} = 0.355$$

Rapporto di correlazione