

STATISTICA DESCRITTIVA B

(prof. Fortunato Pesarin)

FACOLTA' DI ASTRONOMIA
Corso di Laurea in **Astronomia**

Correzione Prova scritta del 11 Luglio 2006

- 1) Per valutare l'area di attrazione della Facoltà di Scienze Statistiche su studenti di genere diverso, si è rilevata la residenza delle 180 matricole dell'A.A. 2003-04. I dati sono sintetizzati nella seguente tabella:

X Genere	Y Residenza		
	nella provincia di Padova	nel Veneto, fuori provincia di Padova	fuori del Veneto
maschio	30	48	22
femmina	38	25	17

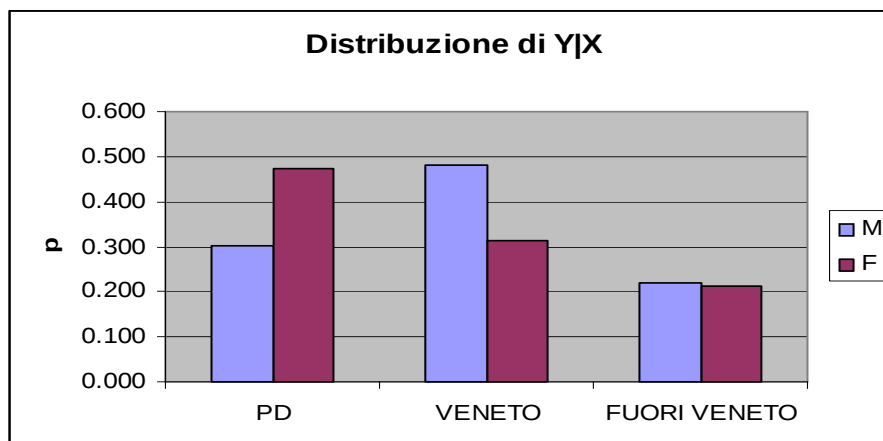
- 1.1) Qual è l'unità statistica, quale la natura delle variabili e su quale scala sono misurate.

L'unità statistica è la matricola iscritta alla facoltà di Scienze statistiche nell' A.A. 2003-2004. Le variabili rilevate sono entrambe qualitative su scala sconnessa.

- 1.2) Si ottengano le frequenze relative della variabile **Y**.

Y Residenza			
nella provincia di Padova	nel Veneto, fuori provincia di Padova	fuori del Veneto	Totale
0.378	0.406	0.217	1

- 1.3) Rappresentare graficamente la distribuzione percentuale della variabile **Y** condizionatamente alle modalità di **X**.



1.4) Si può affermare che la mutabilità della variabile Y sia la stessa tra i maschi e le femmine?

Indice di Gini: $G = 1 - \sum_{i=1}^3 p_i^2$: $G_M = 0.6312$ $G_F = 0.6316$

Indice di Gini standardizzato: $kG/(k-1)$ $G'_M = 0.9468$ $G'_F = 0.9473$

Entropia di Shannon $H = - \sum_{i=1}^3 p_i \log(p_i)$: $H_M = 0.4545$ $H_F = 0.4543$

Entropia di Shannon standardizzata: $H/\log(k)$ $H'_M = 0.9526$ $H'_F = 0.9523$

Entrambi gli indici confermano che la mutabilità di Y nei due sessi sia praticamente la stessa.

1.5) Utilizzando un indice opportuno, si può affermare che il genere e la residenza siano dipendenti?

Frequenze attese in ipotesi di indipendenza

37.78	40.56	21.67	100
30.22	32.44	17.33	80
68	73	39	180

Indice χ^2 : $6.689137 > (3-1) \times (2-1) = 2 \Rightarrow$ **esiste dipendenza tra sesso e residenza.**

Indice di Cramèr: $V = \sqrt{\frac{\chi^2}{n \cdot \min(R-1, C-1)}} = 0.1927$

- 2) Sugli studenti del corso di laurea in SGI si sono osservate tre variabili: X= età (in anni); Y= voto di Algebra (in trentesimi) e Z = voto di Statistica (in trentesimi). La matrice di correlazione tra le tre variabili è risultata uguale a:

	X	Y	Z
X	0	+0,48	+0,75
Y	+0,48	1	+0,81
Z	+0,75	+0,81	1

2.1) La tabella sopra contiene un dato impossibile, quale? Perché?

Il dato impossibile è nella prima cella della prima riga: $\rho(X, X) = 0$, perchè questo equivale ad affermare che $COV(X, X) = 0$, ma per ciascuna variabile non degenera si ha $COV(X, X) = V(X) > 0$ e $\rho(X, X) = 1$.

2.2) Quale dei seguenti due modelli:

I°) $Z = \alpha_1 + \beta_1 X + \varepsilon$ oppure II°) $Z = \alpha_2 + \beta_2 Y + \varepsilon$

conviene usare per prevedere Z e perché?

Conviene usare il secondo perchè Z risulta maggiormente correlata a Y di quanto lo sia X, perchè $\rho(Z, Y) = 0.81 > 0.75 = \rho(Z, X)$

2.3) Sono note le seguenti quantità: $M(X) = 27,3$; $M(Y) = 25,4$; $M(Z) = 21,6$; $M(X^2) = 751,59$; $M(Y^2) = 654,96$; $M(Z^2) = 471,26$, dove:

$$M(X) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{e} \quad M(X^2) = \frac{1}{n} \sum_{i=1}^n x_i^2.$$

Disegnare, su un sistema di assi cartesiani, la retta relativa al modello prescelto in (2.2).

	VAR	SQM	
X	6.3	2.50998	COV(X,Y) 3.771591
Y	9.8	3.130495	COV(X,Z) 4.08113
Z	4.7	2.167948	COV(Y,Z) 5.497269

Le covarianze si ottengono tenendo conto della definizione di correlazione:

$$\rho(X, Y) = \frac{\text{COV}(X, Y)}{\sqrt{V(X)V(Y)}} \Rightarrow \text{COV}(X, Y) = \rho(X, Y) \sqrt{V(X)V(Y)}$$

Le varianze si ottengono dalla relazione: $V(X) = M(X^2) - M(X)^2$.

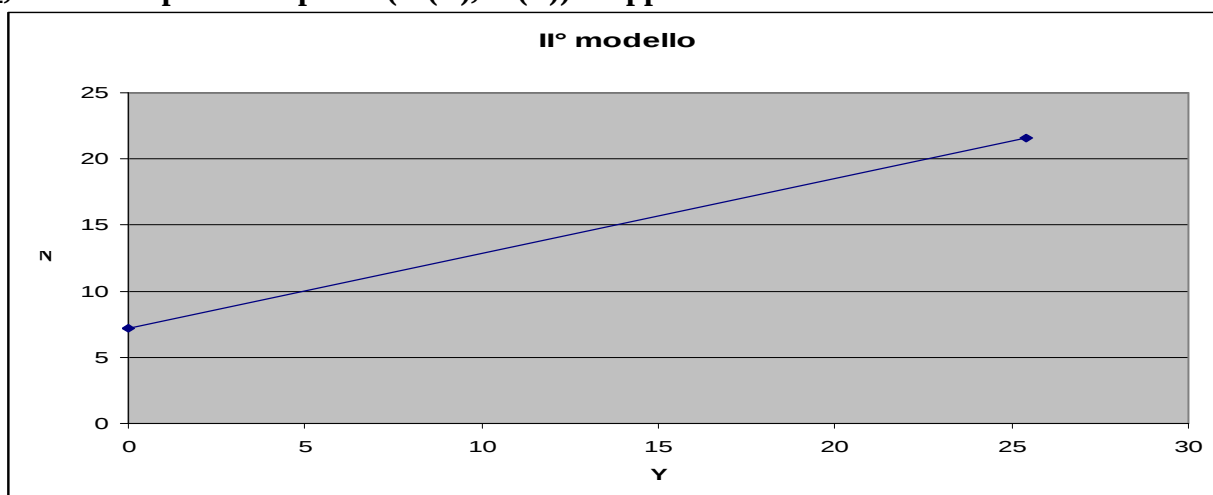
I° modello: $Z = \alpha_1 + \beta_1 X + \varepsilon$

$$\begin{aligned} a_1 &= M(Z) - b_1 M(X) &= 3.915 \\ b_1 &= \text{COV}(Z, X) / \text{VAR}(X) &= 0.648 \end{aligned}$$

II° modello: $Z = \alpha_2 + \beta_2 Y + \varepsilon$

$$\begin{aligned} a_2 &= M(Z) - b_2 M(Y) &= 7.352 \\ b_2 &= \text{COV}(Z, Y) / \text{VAR}(Y) &= 0.561 \end{aligned}$$

Esiste un'unica retta passante per due punti: il primo punto è rappresentato dall'intercetta α_2 , il secondo punto è il punto $(M(Y), M(Z))$. Rappresentazione II° modello:



2.4) Per quale punto passa sicuramente la retta di regressione?

Passa sicuramente per i punti $(0, a_2)$ e $(M(Y), M(Z))$. Il primo per definizione di intercetta di una retta, il secondo perchè la stima di Z in corrispondenza di $Y = M(Y)$ è:

$$Z' = a_2 + b_2 M(Y) = M(Z) - b_2 M(Y) + b_2 M(Y) = M(Z).$$

2.5) Si ottenga una misura della bontà del modello scelto.

$$\text{I° modello: } R^2 = \rho^2(Z, X) = 0.5625$$

$$\text{II° modello: } R^2 = \rho^2(Z, Y) = 0.6561$$

2.6) Si ottenga il coefficiente di correlazione parziale tra voto in statistica descrittiva e voto in algebra lineare al netto dell'età.

La correlazione parziale tra Z e Y al netto di X è la correlazione tra Z' e Y', dove Z' e Y' sono i residui delle regressioni di Z su X e Y su X. In questo modo si evidenzia la correlazione dei dati “depurata” dall'effetto di X. Il calcolo dei residui è il seguente:

$$y_i' = y_i - \left(\bar{y} - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x} + \frac{\sigma_{xy}}{\sigma_x^2} x_i \right) = (y_i - \bar{y}) - \frac{\sigma_{xy}}{\sigma_x^2} (x_i - \bar{x})$$

Analogamente:

$$z_i' = (z_i - \bar{z}) - \frac{\sigma_{xz}}{\sigma_x^2} (x_i - \bar{x})$$

segue che:

$$\begin{aligned} V(Y') &= \sigma_{Y'}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{\sigma_{xy}^2}{\sigma_x^4} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - 2 \frac{\sigma_{xy}}{\sigma_x^2} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \sigma_Y^2 + \frac{\sigma_{xy}^2}{\sigma_x^4} \sigma_x^2 - 2 \frac{\sigma_{xy}}{\sigma_x^2} \sigma_{xy} = \sigma_Y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_Y^2 \left(1 - \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_Y^2} \right) = \sigma_Y^2 (1 - \rho_{xy}^2) \end{aligned}$$

Analogamente:

$$\sigma_{Z'}^2 = \sigma_Z^2 (1 - \rho_{xz}^2)$$

$$\sigma_{Y'Z'} = \sigma_{YZ} - \sigma_{XY} \sigma_{XZ} / \sigma_X^2$$

Infine:

$$\rho_{YZ|X} = \frac{\rho_{YZ} - \rho_{XY} \rho_{XZ}}{\sqrt{(1 - \rho_{XY}^2)(1 - \rho_{XZ}^2)}} = 0.79275$$