

Correlazione e regressione per problemi

di Luciano Corso

Presidente della sezione di Verona della Mathesis

Direttore della Rivista MatematicaMente

Email: lcorso@iol.it

CASTELLAMMARE DI STABIA 20180717

Due variabili statistiche X e Y , su base sperimentale, hanno presentato la seguente distribuzione delle frequenze (va inteso che la coppia $(x=5, y=4)$ è stata osservata 3 volte):

		x					
		1	2	3	5	6	8
y	1	2	1	0	0	0	0
	3	0	1	4	1	0	0
	4	0	0	1	3	1	0
	6	0	0	0	0	0	2

Quesiti

- 1) Si chiede se i due caratteri sono statisticamente indipendenti.
- 2) Si verifichi l'ipotesi di indipendenza statistica tra i due caratteri al livello di significatività del $\alpha = 5\%$.
- 3) Si calcolino medie aritmetiche e varianze totali e si valuti quale delle due variabili ha una maggiore variabilità.
- 4) Si determini la media aritmetica di x condizionata da $y = 4$ e la media aritmetica di y condizionata da $x = 2$.
- 5) Si calcolino la covarianza e il coefficiente di correlazione lineare.
- 6) Si determini la retta (interpolante) di regressione.
- 7) Si valuti con un opportuno indice la bontà dell'accostamento fatto tra fenomeno osservato e modello teorico.
- 8) Si tracci il grafico del fenomeno presentato in tabella e del modello teorico interpolato.

Obiettivo didattico

- L'obiettivo della verifica è di valutare se uno studente possiede i concetti di indipendenza statistica di due variabili, di verifica delle ipotesi (inferenza statistica), di correlazione lineare tra due variabili, di dipendenza lineare e di regressione, di bontà di un accostamento tra dati sperimentali e modelli teorici, di medie totali e condizionate e di misure comparabili della variabilità.

Verificare l'indipendenza statistica tra X e Y

$$\text{Prob}(x_{.j} \cap y_{i.}) = \text{Prob}(x_{.j}) \cdot \text{Prob}(y_{i.}), \quad \forall i, j.$$

$$\frac{\hat{n}_{ij}}{n} = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}, \quad \forall i, j$$

$$\hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}, \quad \forall i, j$$

Calcolo delle frequenze assolute congiunte in ipotesi di indipendenza statistica

		x						fr marg. Y
		1	2	3	5	6	8	
y	1	2 (6/16)	1 (6/16)	0 (15/16)	0 (12/16)	0 (3/16)	0 (6/16)	3
	3	0 (12/16)	1 (12/16)	4 (30/16)	1 (24/16)	0 (6/16)	0 (12/16)	6
	4	0 (10/16)	0 (10/16)	1 (25/16)	3 (20/16)	1 (5/16)	0 (10/16)	5
	6	0 (4/16)	0 (4/16)	0 (10/16)	0 (8/16)	0 (2/16)	2 (4/16)	2
Freq. Assol. Marginali x		2	2	5	4	1	2	16

Test delle ipotesi

H_0 = “Dichiaro che è vero lo stato 0” $\alpha = Prob(H_1 / \Omega_0)$; $\beta = Prob(H_0 / \Omega_1)$

H_1 = “Dichiaro che è vero lo stato 1”

Ω_0 = “È vero lo stato 0”

Ω_1 = “È vero lo stato 1”

		STATI DI NATURA	
		Ω_0	Ω_1
IPOTESI	H_0	$H_0 \mid \Omega_0$	$H_0 \mid \Omega_1$
	H_1	$H_1 \mid \Omega_0$	$H_1 \mid \Omega_1$

Si fissano le ipotesi:

$$\begin{cases} H_0: n_{i,j} = \hat{n}_{i,j} & \forall i, j \\ H_1: n_{i,j} \neq \hat{n}_{i,j} , & \exists i, j \end{cases}$$

$$\alpha = 0,05$$

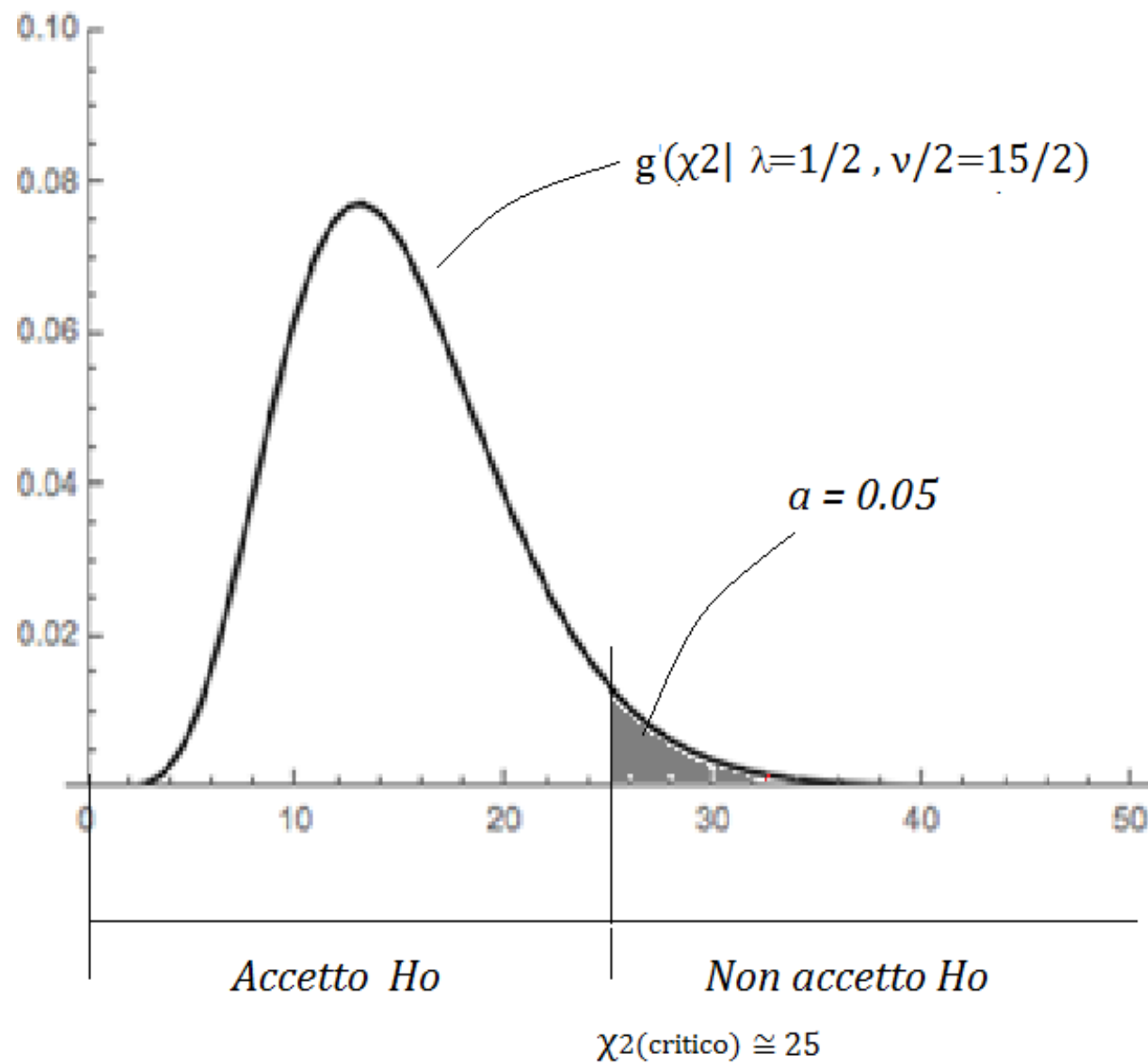
2) Per verificare se questa dipendenza è o no casuale si applica il test delle ipotesi sulla statistica

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c [(n_{i,j} - \hat{n}_{i,j})^2 / \hat{n}_{i,j}]$$

dove $n_{i,j}$ e $\hat{n}_{i,j}$ sono rispettivamente le frequenze assolute osservate e teoriche in ipotesi di indipendenza; dal calcolo ($r = 4, c = 6$), dove r è il numero di righe e c il numero delle colonne della tabella, risulta che $\chi^2 \cong 34.9067$.

Si dimostra che χ^2 ha una distribuzione di probabilità del tipo Gamma: $G(\chi^2 | \lambda = 1/2, v/2)$. Presentiamo la densità di probabilità e il grafico della distribuzione della varia-bile aleatoria χ^2

$$\bullet g\left(\chi^2 \middle| \lambda = \frac{1}{2}, \frac{v}{2}\right) = \begin{cases} \frac{2^{-\frac{v}{2}} \cdot e^{-\lambda \cdot \chi^2} \cdot (\chi^2)^{\left(\frac{v}{2}-1\right)}}{\Gamma\left(\frac{v}{2}\right)}, & \{\chi^2 | \chi^2 \in \mathbb{R}^+\} \\ 0 & \text{altrove} \end{cases} \quad (3)$$



Equazione integrale: $\chi^2(\text{critico}) \cong 25$.
Si respinge H_0 .

$$\int_0^{\chi^2_{critico}} \frac{2^{-\frac{\nu}{2}} \cdot e^{-\lambda \cdot \chi^2} \cdot \chi^{2\left(\frac{\nu}{2}-1\right)}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot d\chi^2 = 0.95.$$

3) Calcoliamo medie, varianze e coefficienti di variazione

$$M(x) = 63/16 ; \quad V(x) = M(x - \bar{x})^2 = 1135/256 ;$$

$$v(x) = \frac{\sqrt{V(x)}}{\bar{x}} \cong 0.5348;$$

$$M(y) = 53/16 ; \quad V(y) = M(y - \bar{y})^2 = 535/256 ;$$

$$v(y) = \frac{\sqrt{V(y)}}{\bar{y}} \cong 0.4364.$$

4) Medie condizionate:

$$M(x \mid y = 4) = 24/5, \quad M(y \mid x = 2) = 2.$$

5) Calcolo della covarianza

$$C(x, y) = M[(x - \bar{x}) \cdot (y - \bar{y})]$$

$$C(x, y) = M(x \cdot y) - M(x) \cdot M(y) = 709/256$$

Coefficiente di correlazione lineare di Bravais-Pearson

$$r(x, y) = \frac{C(x, y)}{\sqrt{V(x)} \cdot \sqrt{V(y)}} \cong 0.91.$$

$$-1 \leq r \leq +1$$

Dimostrazione di: $-1 \leq r \leq +1$

$$(x - \bar{x})t = (y - \bar{y}) \quad ==> \quad (y - \bar{y}) - (x - \bar{x})t = 0$$

$$M[(y - \bar{y}) - (x - \bar{x})t]^2 \geq 0$$

$$M(x - \bar{x})^2 t^2 - 2M[(x - \bar{x})(y - \bar{y})]t + M(y - \bar{y})^2 \geq 0$$

$$\sigma_x^2 \cdot t^2 - 2 \cdot \sigma_{xy} \cdot t + \sigma_y^2 \geq 0$$

$$\sigma_{xy}^2 - \sigma_x^2 \cdot \sigma_y^2 \leq 0 \quad ==> \quad -\sigma_x \sigma_y \leq \sigma_{xy} \leq +\sigma_x \sigma_y$$

$$-1 \leq \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \leq +1$$

Regola dei minimi quadrati

Modello: $\hat{y} = a + b \cdot x$

$$S(a, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

$$\begin{cases} \frac{\partial S}{\partial a} = 2 \sum_{i=1}^n (y_i - a - bx_i) \cdot (-1) \\ \frac{\partial S}{\partial b} = 2 \sum_{i=1}^n (y_i - a - bx_i) \cdot (-x_i) \end{cases}$$

Metodo dei Minimi Quadrati ponderati

$$S(a, b) = \sum_{j=1}^c \sum_{i=1}^r (y_{ij} - \hat{y}_{ij})^2 \cdot n_{ij}$$

$$S(a, b) = \sum_{j=1}^c \sum_{i=1}^r (y_{ij} - a - b \cdot x_{ij})^2 \cdot n_{ij}$$

$$\begin{cases} \frac{\partial S}{\partial a} = 2 \sum_{j=1}^c \sum_{i=1}^r (y_{ij} - a - b \cdot x_{ij}) \cdot (-1) \cdot n_{ij} \\ \frac{\partial S}{\partial b} = 2 \sum_{j=1}^c \sum_{i=1}^r (y_{ij} - a - b \cdot x_{ij}) \cdot (-x_{ij}) \cdot n_{ij} \end{cases}$$

$$\begin{cases} \sum_{j=1}^r \sum_{i=1}^c n_{ij} + \sum_{j=1}^r \sum_{i=1}^c x_{ij} \cdot n_{ij} = \sum_{j=1}^r \sum_{i=1}^c y_{ij} \cdot n_{ij} \\ \sum_{j=1}^r \sum_{i=1}^c x_{ij} \cdot n_{ij} + \sum_{j=1}^r \sum_{i=1}^c x_{ij}^2 \cdot n_{ij} = \sum_{j=1}^r \sum_{i=1}^c x_{ij} \cdot y_{ij} \cdot n_{ij} \end{cases}$$

$$\begin{cases} n \cdot a + \left(\sum_{i=1}^n x_i \right) \cdot b = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i \right) \cdot a + \left(\sum_{i=1}^n x_i^2 \right) \cdot b = \left(\sum_{i=1}^n x_i y_i \right) \end{cases}$$

$$\begin{cases} 16 \cdot a + 63 \cdot b = 53 \\ 63 \cdot a + 319 \cdot b = 253 \end{cases} \quad \begin{cases} a = \frac{968}{1135} \\ b = \frac{709}{1135} \end{cases} ; \quad \hat{y} = \frac{968}{1135} + \frac{709}{1135} \cdot x$$

Un modo statistico per arrivare allo stesso risultato:

$$a = M(y) - b \cdot M(x) \quad ==> \quad b = \frac{C(x,y)}{V(x)}$$

$$b = \frac{709/256}{1135/256} = \frac{709}{1135}$$

$$a = \frac{53}{16} - \frac{709}{1135} \cdot \frac{63}{16} = \frac{968}{1135}$$

$$\text{Dimostrazione: } \begin{cases} y = a + b x \\ \bar{y} = a + b \bar{x} \end{cases} \quad y - \bar{y} = b(x - \bar{x}) \rightarrow$$

$$(x - \bar{x})(y - \bar{y}) = b(x - \bar{x})^2 \quad \rightarrow \quad M(x - \bar{x})(y - \bar{y}) = b M(x - \bar{x})^2$$

Grafico

