

CORRELAZIONE e CONNESSIONE

Correlazione (r)

Il coefficiente di correlazione è dato da : $r = \sqrt{b \cdot b'}$ con $-1 \leq r \leq +1$.

Se $r=1$ c'è correlazione e le due rette sono coincidenti.

Se $r=0$ non c'è correlazione.

In particolare se si è vicini ad 1 si parla di correlazione **diretta**, se si è vicini a -1 si parla di correlazione **inversa** .

Esercizio 1.

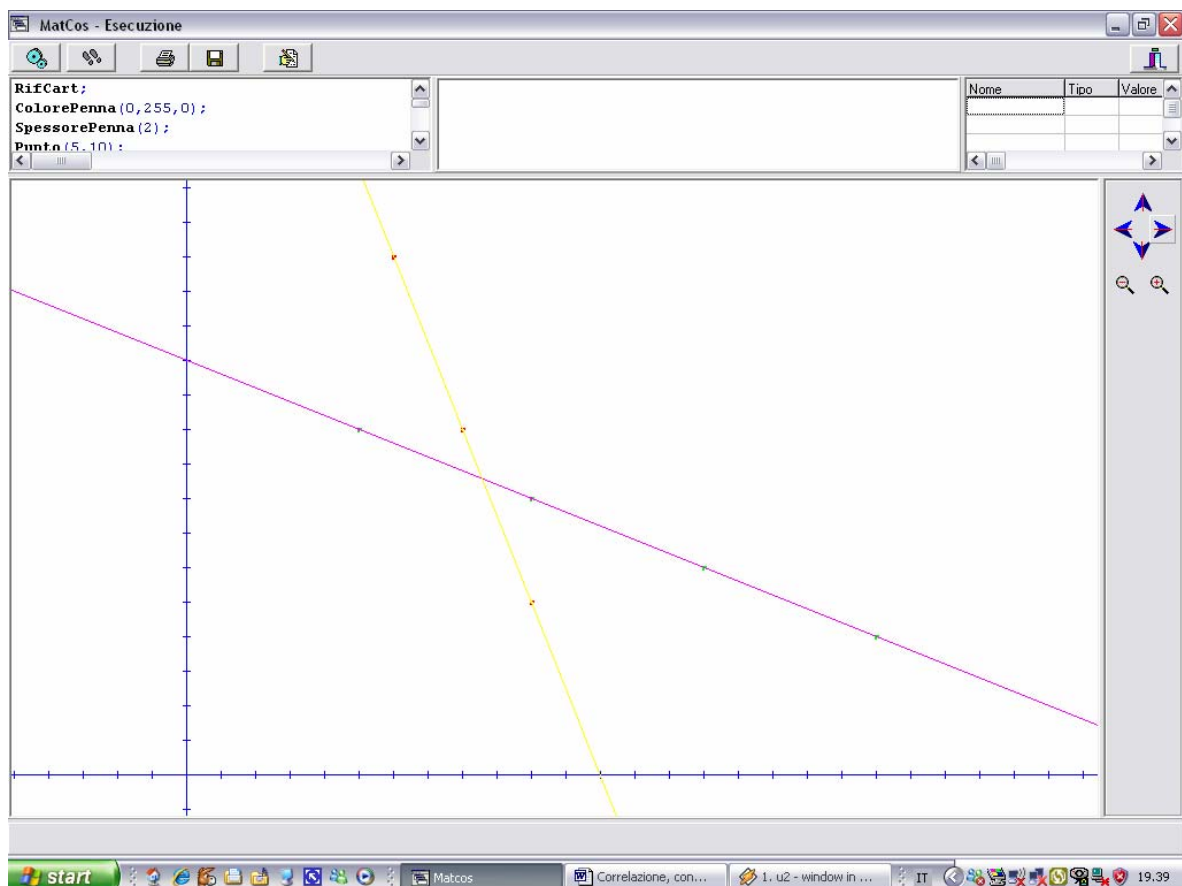
Dato un bene sono messi in relazione il prezzo e la quantità.

PREZZI	QUANTITA'
5	10
10	8
15	6
20	4

Determinare r cioè verificare se c'è correlazione tra il fattore “prezzo” e il fattore “quantità”.

Soluzione:

Facendo variare i prezzi e le quantità, cioè le x_i e le y_i , si ottengono due rette di regressione i cui coefficienti angolari sono b e b' .



Per determinare le rette si devono calcolare a e b , facendo variare le x_i e le y_i , dal seguente sistema di equazioni normali:

$$\begin{cases} na + b \sum_i X_i = \sum_i Y_i \\ a \sum_i X_i + b \sum_i X_i^2 = \sum_i X_i Y_i \end{cases}$$

Il coefficiente b si può calcolare anche nel modo seguente:

$$b = \frac{n \sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{n \sum_i X_i^2 - \left(\sum_i X_i \right)^2}$$

Ma esiste un metodo per semplificare tutto calcolando delle medie e cioè r si ricava direttamente con la seguente formula:

$$r = \sqrt{b \cdot b'} \quad \text{con } b = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \quad \text{e } b' = \frac{\sum_i y_i x_i}{\sum_i y_i^2}$$

$$\text{Allora } r \text{ diventa: } r = \sqrt{b \cdot b'} = \sqrt{\frac{\sum_i x_i y_i}{\sum_i x_i^2} \cdot \frac{\sum_i y_i x_i}{\sum_i y_i^2}} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}}$$

Ma cosa sono x_i e y_i ?

$$x_i = X_i - \text{media}(X_i) = X_i - m_x \quad \text{e } y_i = Y_i - \text{media}(Y_i) = Y_i - m_y$$

Poichè $m_x = \frac{50}{4} = 12.5$ e $m_y = \frac{28}{4} = 7$ si avrà che:

X_i	Y_i	$x_i = X_i - m_x$	$y_i = Y_i - m_y$
5	10	5-12.5=-7.5	10-7=+3
10	8	10-12.5=-2.5	8-7=+1
15	6	15-12.5=+2.5	6-7=-1
20	4	20-12.5=+7.5	4-7=-3
50	28		

$$r = \frac{(-7.5)(3) + (-2.5)(1) + (2.5)(-1) + (7.5)(-3)}{\sqrt{\left[(-7.5)^2 + (-2.5)^2 + (2.5)^2 + (7.5)^2 \right] \cdot \left[(3)^2 + (1)^2 + (-1)^2 + (-3)^2 \right]}} = \dots$$

$$\dots \frac{-22.50 - 2.50 - 2.50 - 22.50}{\sqrt{125 \cdot 20}} = \frac{-50}{\sqrt{2500}} = \frac{-50}{50} = -1$$

Abbiamo ottenuto un tipo di correlazione inversa poichè $r = -1$.

Connessione (η di Pearson)

Esercizio 1.

Verificare se esiste connessione statistica tra il fattore “titolo di studio” e il fattore “affluenza al cinema/mese”.

Titolo di studio	1 volta/mese	2 volte/mese	3 volte/mese
Nessuno	2	3	1
Licenza elementare	1	3	1
Licenza media inf	1	1	4
Licenza media sup	1	3	2
Laurea	1	1	5
	6	11	13

Il coefficiente di connessione è dato da $0 \leq \eta = \sqrt{\frac{\sum_i (\bar{Y}_i - \bar{Y})^2 \cdot N_i}{\sum_j (\bar{Y}_j - \bar{Y})^2 \cdot N_j}} \leq 1$ con \bar{Y}_i media di riga e \bar{Y}

media ponderata.

Se $\eta = 0$ non esiste connessione statistica.

Se $\eta = 1$ si ha connessione statistica massima.

Soluzione:

Calcoliamo le medie di riga \bar{Y}_i .

Titolo di studio	1 volta/m	2 volte/m	3 volte/m	Tot.	\bar{Y}_i
Nessuno	2	3	1	6	$\bar{Y}_1 = (2*1)+(3*2)+(1*3)/6=11/6=1.8$
Lic. Elem.	1	3	1	5	$\bar{Y}_2 = (1*1)+(3*2)+(1*3)/6=10/5=2$
Lic. M.I.	1	1	4	6	$\bar{Y}_3 = (1*1)+(1*2)+(4*3)/6=15/6=2.5$
Lic. M.S.	1	3	2	6	$\bar{Y}_4 = (1*1)+(3*2)+(2*3)/6=13/6=2.2$
Laurea	1	1	5	7	$\bar{Y}_5 = (1*1)+(1*2)+(5*3)/7=18/7=2.6$
	6	11	13	30	

Calcoliamo la media ponderata \bar{Y} .

$$\bar{Y} = \text{media_ponderata} = \frac{(6*1)+(11*2)+(13*3)}{30} = \frac{6+22+39}{30} = \frac{67}{30} = 2.2.$$

Il coefficiente η sarà allora uguale a :

$$\eta = \sqrt{\frac{(\bar{Y}_1 - \bar{Y})^2 \cdot N_1 + (\bar{Y}_2 - \bar{Y})^2 \cdot N_2 + (\bar{Y}_3 - \bar{Y})^2 \cdot N_3 + (\bar{Y}_4 - \bar{Y})^2 \cdot N_4 + (\bar{Y}_5 - \bar{Y})^2 \cdot N_5}{(Y_1 - \bar{Y})^2 \cdot N_1 + (Y_2 - \bar{Y})^2 \cdot N_2 + (Y_3 - \bar{Y})^2 \cdot N_3}} \quad \text{cioè}$$

$$\eta = \sqrt{\frac{(1.8 - 2.2)^2 \cdot 6 + (2 - 2.2)^2 \cdot 5 + (2.5 - 2.2)^2 \cdot 6 + (2.2 - 2.2)^2 \cdot 6 + (2.6 - 2.2)^2 \cdot 7}{(1 - 2.2)^2 \cdot 6 + (2 - 2.2)^2 \cdot 11 + (3 - 2.2)^2 \cdot 13}} \dots\dots$$

$$\dots\dots \eta = \sqrt{\frac{0.96 + 0.2 + 0.54 + 0 + 1.12}{8.64 + 0.44 + 8.32}} = \sqrt{\frac{2.82}{17.4}} = \sqrt{0.16} = 0.4.$$

Essendo $\eta = 0.4$ possiamo affermare che esiste connessione quasi media tra il fattore “titolo di studio” e il fattore “affluenza al cinema/mese”.

Esercizio 2.

Una indagine statistica relativa ai costi medi mensili delle saune di una beauty-farm ha prodotto i risultati inseriti nella seguente tabella di frequenza:

	Costo medio 50 €	Costo medio 100 €	Costo medio 150 €
Città di Milano	35	60	55
Città di Venezia	71	56	33
Città di Catania	50	32	24

Calcolare l’Eta di Pearson e commentare il risultato.

Soluzione:

Calcoliamo le medie di riga \bar{Y}_i .

	50 €	100 €	150 €	Totali	\bar{Y}_i
Milano	35	60	55	150	$\bar{Y}_1 = \frac{(50 * 35) + (100 * 60) + (150 * 55)}{150} = 106.67$
Firenze	71	56	33	160	$\bar{Y}_2 = \frac{(71 * 50) + (56 * 100) + (33 * 150)}{160} = 88.12$
Catania	50	32	24	106	$\bar{Y}_3 = \frac{(50 * 50) + (32 * 100) + (24 * 150)}{106} = 87.73$
	156	148	112	416	

Calcoliamo la media ponderata \bar{Y} .

$$\bar{Y} = \text{media_ponderata} = \frac{(156 * 50) + (148 * 100) + (112 * 150)}{416} = \frac{39400}{416} = 94.71.$$

Il coefficiente η sarà allora uguale a :

$$\eta = \sqrt{\frac{(\bar{Y}_1 - \bar{Y})^2 \cdot N_1 + (\bar{Y}_2 - \bar{Y})^2 \cdot N_2 + (\bar{Y}_3 - \bar{Y})^2 \cdot N_3}{(Y_1 - \bar{Y})^2 \cdot N_1 + (Y_2 - \bar{Y})^2 \cdot N_2 + (Y_3 - \bar{Y})^2 \cdot N_3}} \quad \text{cioè}$$

$$\eta = \sqrt{\frac{(106.67 - 94.71)^2 \cdot 150 + (88.12 - 94.71)^2 \cdot 160 + (87.73 - 94.71)^2 \cdot 106}{(50 - 94.71)^2 \cdot 156 + (100 - 94.71)^2 \cdot 148 + (150 - 94.71)^2 \cdot 112}} \dots\dots$$

$$\dots\dots \eta = \sqrt{\frac{21456.24 + 6948.496 + 5164.3624}{311841.5196 + 4141.6468 + 342382.2192}} = \sqrt{\frac{33569.0984}{658365.3856}} = \sqrt{0.050988553} = 0.2258.$$

Essendo $\eta = 0.2258$ possiamo affermare che esiste connessione minima tra i due fattori esaminati.