Università di Cassino

Esercitazioni di Statistica 1 del 4 Marzo 2011

Dott. Mirko Bevilacqua

ESERCIZIO n° 1

Su un collettivo di individui sono stati rilevati i caratteri **X** Peso(in kg) e **Y** Altezza (in cm) otteniamo la seguente distribuzione di frequenza congiunta:

Y / X	165	170	175	Tot
60	2	0	0	**2**
70	0	1	0	**1**
80	1	0	1	**2**
Tot	**3**	**1**	**1**	**5**

a) Ricostruire la successione dell'altezza
b) Calcolare la media e la mediana dell'altezza
c) Calcolare il peso medio per gli individui che hanno un'altezza di 165 cm
d) Calcolare il coefficiente di correlazione lineare tra peso e altezza

SOLUZIONI

a) 165; 165; 165; 170; 175

b) Essendo N=5 la mediana è la modalità che occupa il terzo posto nella successione ordinata: Me=165

MEDIA DI Y (ALTEZZA):

$$\mu_y = \frac{\sum_{i=1}^{5} y_i \cdot n_i}{n} = \frac{(165 \cdot 3) + 170 + 175}{5} = 168$$

c) Il peso medio degli individui con altezza pari a 165 cm è

$$\mu_{x/y=165} = \frac{\sum_{j=1}^{3} x_i n_{.i}}{n_{.1}} = \frac{(60 \cdot 2) + 80}{3} = 66,6$$

d)

Il coefficiente di correlazione lineare, a differenza della covarianza, non dipende dall'unità di misura delle osservazioni e permette di confrontare diverse distribuzioni doppie.

$$\rho_{xy} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$$

- $-1 \leq \rho_{XY} \leq 1$
  - $\rho_{XY}=1$ Tra la Y e la X sussiste un perfetto legame lineare e i due caratteri sono concordi
  - $\rho_{XY}=-1$ Tra la Y e la X sussiste un perfetto legame lineare e i due caratteri sono discordi
  - $\rho_{XY}=0$ I due caratteri sono indipendenti, oppure se la loro relazione non è lineare

MEDIA DI X:

$$\mu_x = \frac{\sum_{i=1}^{5} x_i \cdot n_i}{n} = \frac{(60 \cdot 2) + 70 \cdot 1 + 80 \cdot 2}{5} = 70$$

VARIANZA DI X

$$\sigma_x^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2 n_i}{n} = \frac{(60-70)^2 \cdot 2 + (70-70)^2 \cdot 1 + (80-70)^2 \cdot 2}{5} = \frac{400}{5} = 80$$

SCARTO QUADRATICO MEDIO DI X

$$\sigma_x = \sqrt{80} = 8,94$$

MEDIA DI Y

$$\mu_y = 168$$

VARIANZA DI Y

$$\sigma_Y^2 = \frac{\sum_{J=1}^{k} (Y_J - \mu_Y)^2 n_J}{n} = \frac{(165-168)^2 \cdot 3 + (170-168)^2 \cdot 1 + (175-168)^2 \cdot 1}{5} = \frac{80}{5} = 16$$

SCARTO QUADRATICO MEDIO DI X

$$\sigma_Y = \sqrt{16} = 4$$

COVARIANZA (X;Y):

$$cov(x,y) = \frac{1}{5} \cdot \left[\begin{array}{l} (60-70) \cdot (165-168) \cdot 2 + (80-70) \cdot (165-168) \cdot 1 + \\ + (70-70) \cdot (170-168) \cdot 1 + (80-70) \cdot (175-168) \cdot 1 \end{array} \right] =$$

$$= \frac{100}{5} = 20$$

In alternativa la covarianza tra x e y si può calcolare come segue:

X \ Y	165	170	175
60	19.800	0	0
70	0	11.900	0
80	13.200	0	14.000

$$\mu_{xy} = \frac{1}{n} \sum_i \sum_j \hat{x}_i \hat{y}_j n_{ij} = \frac{19800 + 11900 + 13200 + 14000}{5} = 11780$$

$$cov(x,y) = \mu_{xy} - \mu_x \cdot \mu_y = 11780 - 168 \cdot 70 = 20$$

COEFFICIENTE DI CORRELAZIONE

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{20}{8,94 \cdot 4} = 0,56$$

ESERCIZIO n° 2

La seguente tabella riporta i dati del prezzo (migliaia di euro) e potenza (kw) di 10 auto monovolume a benzina.

X: potenza	*108*	*55*	*55*	*80*	*103*	*67*	*76*	*76*	*76*	*56*
Y: prezzo	*32,6*	*14,2*	*17,2*	*18,0*	*25,9*	*13,9*	*17*	*15,8*	*17,3*	*15,2*

a) Stimare la retta di regressione che pone la il prezzo (Y) in funzione della potenza (X).
b) Valutare la bontà di adattamento del modello ai dati osservati.
c) Disegnare il grafico di dispersione e tracciare la retta stimata.
d) A quale delle 10 automobili corrisponde il residuo più elevato (in valore assoluto).
e) Se il prezzo delle automobili fosse espresso in euro e non in migliaia di euro cosa accadrebbe alle stime dei parametri della retta di regressione.
f) L'indice di regressione per questa retta spiega più del 90 % della variabilità totale?

SOLUZIONI

Quando si analizzano due (o più) caratteri quantitativi si può cercare di individuare una funzione che descriva in modo dettagliato la relazione che emerge tra i dati. Si ha così un modello di regressione, ossia una relazione statistica tra la variabile dipendente e le altre variabili indipendenti (chiamate anche variabili esplicative). Tale modello può avere diversi scopi: descrittivo, interpretativo e previsivo.
Una relazione statistica tra una variabile indipendente X e una variabile dipende Y può essere descritta dall'equazione:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

in cui $\alpha+\beta X$ definisce il contributo della variabile esplicativa X al valore della variabile dipendente Y mentre ε rappresenta il contributo di tutti gli altri fattori in grado di influenzare la risposta (variabile dipendente Y) .
Il modello di regressione lineare si dice semplice quando si considera una sola variabile esplicativa (o indipendente). α e β corrispondono all'intercetta e al coefficiente angolare di una retta sul piano e sono chiamati coefficienti di regressione.
Occorre a questo punto definire un metodo di stima dei coefficienti di regressione α e β. In altri termini, occorre individuare una retta che per ogni x_i restituisca un valore di Y_i che sia più vicino possibile ai valori osservati y_i. Il metodo dei minimi quadrati consiste nel ricercare le stime di α e β mediante a e b che rendono minima la somma dei quadrati dei residui e_i (differenza tra il valore osservato y_i e il valore fornito dalla retta di regressione $\hat{y}$).

Le stime dei MINIMI QUADRATI dei coefficienti di regressione sono date da:

$$b = \frac{\sum_{i=1}^{n} (x_i - \mu_x) \cdot (y_i - \mu_y)}{\sum_{i=1}^{n} (x_i - \mu_x)^2} = \frac{cod(x,y)}{dev(x)} = \frac{cov(x,y)}{var(x)}$$

$$a = \mu_y - b \cdot \mu_x$$

$$\hat{y}_i = a + b \cdot x_i$$

X	Y	X^2	Y^2	X·Y
108	32,60	11.664	1062,76	3520,80
55	14,20	3.025	201,64	781,00
55	17,20	3.025	295,84	946,00
80	18,00	6.400	324,00	1440,00
103	25,90	10.609	670,81	2667,70
67	13,90	4.489	193,21	931,30
76	17,00	5.776	289,00	1292,00
76	15,80	5.776	249,64	1200,80
76	17,30	5.776	299,29	1314,80
56	15,20	3.136	231,04	851,20
752	**187**	**59.676**	**3817,23**	**14945,60**

$$\mu_x = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{752}{10} = 75,2 \qquad \mu_{x^2} = \frac{\sum_{i=1}^{10} x_i^2}{n} = \frac{59676}{10} = 5967,6$$

$$\mu_y = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{187,1}{10} = 18,71 \qquad \mu_{y^2} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{3817}{10} = 381,7$$

$$\mu_{xy} = \frac{\sum_{i=1}^{10} x_i \cdot y_i}{n} = \frac{14945,6}{10} = 1494,56$$

$$cov(x,y) = \mu_{xy} - \mu_x \cdot \mu_y = 1494,56 - 18,71 \cdot 75,2 = 87,6$$

$$\sigma_x^2 = 5967,6 - (75,2)^2 = 312,56 \qquad \sigma_x = 17,68$$

a) ***Stima della retta di regressione che pone la Y in funzione di X:***

$$b = \frac{cov(x,y)}{var(x)} = -\frac{87,6}{312,56} = 0,28$$

$$a = \mu_y - b \cdot \mu_x = 18,71 - (0,28 \cdot 75,2) = -2,36$$

$$\hat{y}_i = -2,36 + 0,28 x_i$$

b) **Bontà di adattamento del modello ai dati osservati.**

Il coefficiente di determinazione R^2 indica la porzione di variabilità di Y spiegata dalla variabile esplicativa X attraverso il modello di regressione:

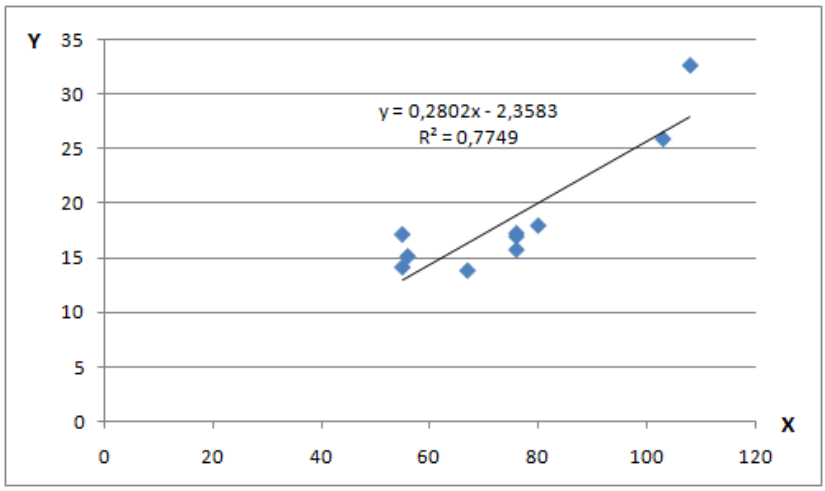
$$R^2 = \left(\frac{cov(x;y)}{\sigma_x \sigma_y} \right)^2$$

Il coefficiente di determinazione varia tra 0 e 1: vale 0 in assenza di relazione statistica di tipo lineare tra le osservazioni; vale 1 in presenza di perfetta dipendenza lineare (al crescere di R^2 diminuiscono le distanze dei punti osservati y_i dalla retta di regressione).

$$\sigma_y^2 = 381,7 - (18,71)^2 = 31,66 \qquad \sigma_y = 17,68$$

$$R^2 = \left(\frac{cov\,(x;y)}{\sigma_x \sigma_y} \right)^2 = \left(\frac{87,6}{5,63 \cdot 17,68} \right)^2 = 0,775$$

c) Grafico di dispersione e retta di regressione



d) Determinazione del residuo più elevato

Data la retta di regressione

$\hat{y}_i = a + b \cdot x_i$ si definisce residuo (*e*) la differenza tra il valore osservato y_i e il valore fornito dalla retta di regressione $\hat{y}_i$: $\hat{e}_i = y_i - \hat{y}_i$.

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
108	32,60	27,90	4,70
55	14,20	13,05	1,15
55	17,20	13,05	4,15
80	18,00	20,05	-2,05
103	25,90	26,50	-0,60
67	13,90	16,41	-2,51
76	17,00	18,93	-1,93
76	15,80	18,93	-3,13
76	17,30	18,93	-1,63
56	15,20	13,33	1,87

Guardando alla tabella dei residui riportata di seguito, si osserva che alla prima auto corrisponde il maggior residuo. Si noti che la somma dei residui vale zero (a meno di una piccola differenza dovuta alle approssimazioni.

e) Retta di regressione con prezzi espressi in euro.

Entrambi i coefficienti di regressione vengono moltiplicati per 1000

$$\hat{y}_i = -2360 + 280 x_i$$

f) Proporzione di variabilità di Y spiegata dalla variabili esplicativa x

No, la porzione di variabilità di Y spiegata dalla variabile esplicativa X è pari al 77,5% (R^2).