Esercitazione 9

ESERCIZIO 1

La seguente tabella riporta la distribuzione di 75 irlandesi, divisa in classi rispetto al numero di pinte consumate in una sera al pub.

$x_l - \|x_u$	n_i
$0-2$	50
$3-5$	12
$6-7$	7
$8-10$	6
Totale	75

Si calcoli il valore del rapporto di concentrazione di Gini. Si calcoli inoltre l'indice di concentrazione di Zenga.

SOLUZIONE

i	$x_l - \|x_u$	n_i	C_i	x_i^C	$n_i \times x_i^C$	$(2 \times C_i - n - n_i)$	$n_i \times x_i^C \times (2 \times C_i - n - n_i)$
1	$0-2$	50	50	1	50	-25	-1250
2	$3-5$	12	62	4	48	37	1776
3	$6-7$	7	69	6,5	45,5	56	2548
4	$8-10$	6	75	9	54	69	3726
	Totale	75					

Il rapporto di concentrazione di Gini può essere calcolato come

$$R = \frac{Area\,concentrazione}{Area\,max\,concentrazione} = \frac{\Delta}{2 \times M_1}$$

Per i dati in analisi:

$$S = 2 \times \sum_{i=1}^{k} n_i \times x_i^C \times (2 \times C_i - n - n_i) = 2 \times 6800 = 13600$$

$$\Delta = \frac{S}{n\,(n-1)} = \frac{12794}{75 \times 74} = 2,450$$

La media è pari a

$$M_1 = \frac{1}{n} \sum_{i=1}^{k} n_i \times x_i^C = \frac{1}{75} \times 191 = 2,633$$

quindi il rapporto di concentrazione di Gini è uguale a

$$R = \frac{2,305}{2 \times 2,547} = 0,465$$

i	$x_l - \|x_u$	n_i	C_i	x_i^C	$n_i \times x_i^C$	q_i'	$M_i^-(x)$	$T - q_i'$	$M_i^+(x)$	$I_i(x)$
1	$0-2$	50	50	1	50	50	1	147,5	5,900	0,831
2	$3-5$	12	62	4	48	98	1,581	99,5	7,654	0,793
3	$6-7$	7	69	6,5	45,5	143,5	2,080	54	9	0,769
4	$8-10$	6	75	9	54	197,5	2,633	0	9	0,707
	Totale	75								

Calcoliamo le pinte consumate cumulate:

$$q'_i = \sum_{j=1}^{i} n_j \times x_j^C$$

Poniamo inoltre

$$T = \sum_{j=1}^{k} n_j \times x_j^C$$

pari alle pinte totali. Le medie inferiori possono essere calcolate come

$$M_i^- (x) = \frac{1}{C_i} \sum_{j=1}^{i} n_j \times x_j^C = \frac{q'_i}{C_i}$$

mentre le medie superiori

$$M_i^+ (x) = \frac{1}{n - C_i} \sum_{j=i+1}^{k} n_j \times x_j^C = \frac{T - q'_i}{n - C_i}$$

E per convenzione $M_i^+ (x) = M_{n-1}^+ (x)$
Le ineguaglianze puntuali sono calcolate partendo dalle precedenti quantità

$$I_i (x) = \frac{M_i^+ (x) - M_i^- (x)}{M_i^+ (x)} = 1 - \frac{M_i^- (x)}{M_i^+ (x)}$$

e rappresenta la variazione relativa della media del gruppo inferiore rispetto al gruppo superiore.
L'indice di ineguaglianza di Zenga:

$$I (x) = \frac{1}{n} \sum_{i=1}^{k} n_i \times I_i (x) = \frac{1}{75} \sum_{i=1}^{4} n_i \times I_i (x) = \frac{1}{75} \times (60,674) = 0,809$$

ESERCIZIO 2

La seguente serie storica riporta il fatturato di un'azienda in 6 anni differenti (valori in migliaia di euro), il valore per l'anno 2003 è ignoto.

Anno	*Fatturato*
2001	205
2002	206
2003	non disponibile
2004	197
2005	200
2006	199
2007	194

1. si determinino i parametri della retta a minimi quadrati quadrati che interpola il fatturato al variare degli anni trascorsi dal 2001
2. si valuti la bontà di adattamento della retta trovata al punto 1
3. si fornisca il valore del fatturato dell'anno 2003 secondo la retta trovata al punto 1 e si commenti.

SOLUZIONE

i	x_i	y_i	$x_i - \bar{x}$	$y_i \times (x_i - \bar{x})$	x_i^2	y_i^2
1	2001	205	-3,167	-649,167	4004001	42025
2	2002	206	-2,167	-446,333	4008004	42436
3	2004	197	-0,167	-32,833	4016016	38809
4	2005	200	0,833	166,667	4020025	40000
5	2006	199	1,833	364,833	4024036	39601
6	2007	194	2,833	549,667	4028049	37636

La retta hai minimi quadrati è la retta della forma

$$y = \alpha_0 + \alpha_1 \times x$$

dove α_0 e α_1 minimizzano

$$\sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 \times x_i)$$

Per calcolare i parametri dellla retta ai minimi quadrati abbiamo bisogno della covarianza (o codevianza) tra x e y e della varianza (o devianza) di x. La covarianza può essere calcolata come

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \times y_i$$

per i dati in analisi otteniamo

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{6} (x_i - \bar{x}) \times y_i = -7,861$$

Mentre

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = M_2^X - \left[M_1^X\right]^2$$

dove

$$M_1^X = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{6} \times 12025 = 2004,167$$

$$M_2^X = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{6} \times 24100131 = 4016688,500$$

e quinti

$$Var(X) = M_2^X - \left[M_1^X\right]^2 = 4016688,500 - 2004,167^2 = 4,472$$

Possiamo ottenere quindi la stima ai minimi quadrati per α_0 e α_1 come

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)} = \frac{-7,861}{4,472} = -1,758$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x} = 200,167 + 1,758 \times 2004,167 = 3723,018$$

e quindi il modello stimato è

$$y = 3723,018 - 1,758 \times x$$

Per valutare la bontà del modello possiamo utilizzare l'indice di determinazione

$$I_d^2 = \frac{Dev\,Spiegata}{Dev\,Totale} = 1 - \frac{Dev\,Residua}{Dev\,totale} = r^2 = \left[\frac{cov(X,Y)}{\sigma_X \sigma_Y}\right]^2$$

Calcoliamo σ_X e σ_Y

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{4,472} = 2,115$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{M_2^Y - \left[M_1^Y\right]^2} = \sqrt{40084,5 - 200,167^2} = \sqrt{17,806} = 4,220$$

Otteniamo quindi che l'indice di determinazione è pari a

$$I_d^2 = r^2 = \left[\frac{-7,861}{2,115 \times 4,220}\right]^2 = 0,776$$

Per calcolare il valore previsto dalla retta stimata, per l'anno 2003, dato il modello

$$y = 3723,018 - 1,758 \times x$$

sostituiamo a x il valore 2003 e calcoliamo il valore corrispondente per y

$$\hat{y}_{2003} = 3723,018 - 1,758 \times 2003 = 201,774$$

ESERCIZIO 3

I 200 clienti che hanno frequentato un supermercato in una certa fascia oraria di una data giornata sono stati riclassificati secondo il tempo X (in minuti) trascorso nel supermercato e la spesa Y (in euro):

Y\X	$0 \vert - 10$	$10 \vert - 30$	$30 \vert - 60$	Totale
$0 \vert - 10$	15	9	0	24
$10 \vert - 20$	12	20	41	73
$20 \vert - 50$	35	0	33	68
$50 \vert - 200$	0	7	28	35
Totale	62	36	102	200

1. Si determinino le distribuzioni parziali di frequenze relative di Y date le modalità di X.
2. Si determinino i parametri della retta di regressione che spiega Y in funzione di X e si commenti il valore del coefficiente angolare della retta trovata.
3. Si calcoli il coefficiente di correlazione.
4. Si valuti il grado di correlazione lineare tra X ed Y e si commenti il valore ottenuto.
5. Si stabilisca, motivando la risposta, se il coefficiente di correlazione precedentemente calcolato varierebbe qualora il tempo X venisse calcolato in ore e la spesa Y venisse espressa in centinaia di euro.

SOLUZIONE

Consideriamo per comodità la tabella a doppia entrata con valori centrali

Y\X	5	20	45	Totale
5	15	9	0	24
15	12	20	41	73
35	35	0	33	68
125	0	7	28	35
Totale	62	36	102	200

Le distribuzioni parziali di frequenze relative di Y date le modalità di X sono ottenute considerando ogni colonna come distribuzione, e calcolando le frequenze relative (quindi dividendo per il totale di colonna)

Y\X	5	20	45
5	0,242	0,250	0,000
15	0,194	0,556	0,402
35	0,565	0,000	0,324
125	0,000	0,194	0,275
Totale	1,000	1,000	1,000

Per il calcolo dei parametri della retta di regressione, abbiamo che, rispetto a X

x_i	n_i	x_i^2	$x_i \times n_i$	$x_i^2 \times n_i$
5	62	25	310	1550
20	36	400	720	14400
45	102	2025	4590	206550

quindi

$$M_1^X = \frac{1}{n}\sum_{i=1}^{k} n_i \times x_i = 28,1$$

$$M_1^X = \frac{1}{n}\sum_{i=1}^{k} n_i \times x_i^2 = 1112,5$$

$$\sigma_X^2 = M_2^X - \left[M_1^X\right]^2 = 322,89$$

$$\sigma_X = \sqrt{\sigma_X^2} = 17,969$$

rispetto a Y

y_i	n_i	y_i^2	$y_i \times n_i$	$y_i^2 \times n_i$
5	24	25	120	600
15	73	225	1095	16425
35	68	1225	2380	83300
125	35	15625	4375	546875

quindi

$$M_1^Y = \frac{1}{n}\sum_{i=1}^{k} n_i \times x_i = 39,85$$

$$M_1^Y = \frac{1}{n}\sum_{i=1}^{k} n_i \times x_i^2 = 3236$$

$$\sigma_Y^2 = M_2^Y - \left[M_1^Y\right]^2 = 1647,978$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = 40,595$$

Per il calcolo della covarianza abbiamo che

$x_i \times y_j \times n_{ij}$	5	20	45
5	375	900	0
15	900	6000	27675
35	6125	0	51975
125	0	17500	157500

La covarianza può essere calcolata come

$$M_1^{XY} = \frac{1}{n}\sum_{i=1}^{k_X}\sum_{i=1}^{k_Y} x_i \times y_j \times n_{ij} = \frac{1}{200}268950 = 1344,75$$

$$cov(X,Y) = M_1^{XY} - M_1^X \times M_1^Y = 1344,75 - (28,1 \times 39,85) = 224,965$$

La retta può essere quindi stimata come

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)} = \frac{224,965}{322,89} = 0,697$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x} = 39,85 - 0,697 \times 28,1 = 20,272$$

Il coefficiente di correlazione di Pearson è pari a

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{224,965}{17,969 \times 40,595} = 0,308$$

Il grado di correlazione lineare tra X ed Y è ottenuto come

$$I_d^2 = r^2 = \left[\frac{cov(X,Y)}{\sigma_X \sigma_Y}\right]^2 = 0,308^2 = 0,095$$

Il coefficiente di correlazione non varierebbe qualora il tempo X venisse calcolato in ore e la spesa Y venisse espressa in centinaia di euro, infatti è invariante per trasformazioni di scala delle variabili e le trasformazioni non cambiano il segno delle variabili X e Y.