Università di Cassino

Esercitazioni di Statistica 1 - *11 Marzo 2011*

Dott. Mirko Bevilacqua

ESERCIZIO n° 1

Si consideri la seguente tabella che riporta il numero di occupati (x1.000) secondo il tipo di occupazione (Agricoltura, Industria e Altre attività) per alcune regioni italiane.

Tipo occupazione / Regione	Agricoltura	Industria	Altre attività	Totale
Piemonte	87	676	930	1693
Emilia Romagna	121	592	981	1694
Lazio	80	362	1375	1817

1. Si ricavi la distribuzione marginale del carattere tipo di occupazione, riportando anche le frequenze cumulate assolute e relative;
2. Si ricavi la tabella di indipendenza.
3. Si calcolino gli indici di connessione chi quadrato di Pearson e Fisher.

Soluzioni:

1.1

Tipo occupazione	$n_{.j}$	$f_{.j}$	$N_{.j}$	$F_{.j}$
Agricoltura	288	0,06	288	0,06
Industria	1630	0,31	1918	0,37
Altre attività	3286	0,63	5204	1,00
Totale	5204			

1.2

Tabella di indipendenza ($\hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{N}$)

Tipo occupazione / Regione	Agricoltura	Industria	Altre attività
Piemonte	93,7	530,3	1069,0
Emilia Romagna	93,7	530,6	1069,7
Lazio	100,6	569,1	1147,3

1.3

Indice chi quadrato di Pearson

$$\chi^2 = \frac{(87-93,7)^2}{93,7} + \frac{(676-530,3)^2}{530,3} + \frac{(930-1069)^2}{1069} + \frac{(121-93,7)^2}{93,7} + \frac{(592-530,6)^2}{530,6} +$$
$$+ \frac{(981-1069,7)^2}{1069,7} + \frac{(80-100,6)^2}{100,6} + \frac{(362-569,1)^2}{569,1} + \frac{(1375-1147,3)^2}{1147,3} = 205,74$$

Indice di Fisher

$$\phi^2 = \frac{\chi^2}{n} = \frac{205,74}{5204} = 0,04$$

Il valore dell'indice di Fisher va confrontato con l'intervallo [0, 2], in quanto

$$0 \le \varphi^2 \le \min(r-1;\ c-1) \qquad 0 \le \varphi^2 \le 2$$

Risultato: basso grado di connessione tra le variabili; il valore dell'indice è prossimo a zero.

ESERCIZIO n°2

I giovani addetti all'agricoltura in due diverse regioni sono stati classificati per età; la distribuzione di frequenza congiunta è data dalla tabella seguente

Età / Regione	[14-16]	[17-18]	[19-24]	[25-34]
A	15	20	120	150
B	24	30	72	90

1. Determinare la classe modale dell'età per ciascuna regione.
2. Calcolare l'età media per ciascuna delle due regioni.
3. Stabilire se l'età è indipendente in media dalla regione.

Soluzioni:

2.1

Le classi contengono un diverso numero di modalità, quindi per determinare la classe modale dell'età è necessario considerare le densità di frequenza condizionate (ultima colonna nella tabella che segue):

Età *(in anni)*	n_i	a_i	n_i / a_i
Regione A			
[14-16]	15	3	5
[17-18]	20	2	10
[19-24]	120	6	20
[25-34]	150	10	15
Regione B			
[14-16]	24	3	8
[17-18]	30	2	15
[19-24]	72	6	12
[25-34]	90	10	9

Nota: a_i = numero di modalità comprese nella classe iesima.

Nella regione A la classe dell'età con densità di frequenza maggiore è [19; 24], a cui corrisponde una densità di frequenza pari a 20. Nella regione B la classe dell'età con densità di frequenza maggiore è [17;18], a cui corrisponde una densità di frequenza pari a 15. Quindi,

Classe modale dell'età nella regione A = [19; 24]
Classe modale dell'età nella regione B = [17; 18]

2.2

Età *(in anni)*	C_i	n_i
Regione A		
[14-16]	15,0	15
[17-18]	17,5	20
[19-24]	21,5	120
[25-35]	29,5	150
Totale frequenze:		*305*
Regione B		
[14-16]	15,0	24
[17-18]	17,5	30
[19-24]	21,5	72
[25-34]	29,5	90
Totale frequenze:		*216*

Nota: C_i = valore centrale della iesima classe di età.

$$\mu_{Età/Reg=A} = \frac{(15 \cdot 15) + (17,5 \cdot 20) + (21,5 \cdot 120) + (29,5 \cdot 150)}{305} = 24,85$$

$$\mu_{Età/Reg=B} = \frac{(15 \cdot 24) + (17,5 \cdot 30) + (21,5 \cdot 72) + (29,5 \cdot 90)}{216} = 23,56$$

2.3

$$\mu_{Età} = \frac{(15 \cdot 39) + (17,5 \cdot 50) + (21,5 \cdot 192) + (29,5 \cdot 240)}{521} = 24,32$$

L' indice relativo di dipendenza in media è

$$\eta_{Y|X} = \frac{\sigma^2_{ext_Y}}{\sigma^2_Y} = \frac{\sum_{i=1}^{r} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 \cdot n_{i.}}{\sum_{j=1}^{c} \left(y_j - \mu_Y\right)^2 \cdot n_{.j}}$$

$\eta_{Y|X} = 0$ perfetta indipendenza in media di Y da X

$\eta_{Y|X} = 1$ perfetta dipendenza in media di Y da X

Il numero dell'indice di dipendenza in media è

$$\sum_{i=1}^{2} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 \cdot n_{i.} =$$

$$(24,85 - 24,32)^2 \cdot 305 + (23,56 - 24,32)^2 \cdot 216 = 212,7$$

e il denominatore

$$\sum_{j=1}^{4} \left(y_j - \mu_Y\right)^2 \cdot n_{.j} =$$

$$(15 - 24,32)^2 \cdot 39 + (17,5 - 24,32)^2 \cdot 50 +$$

$$+ (21,5 - 24,32)^2 \cdot 192 + (29,5 - 24,32)^2 \cdot 240 = 13679$$

Indice relativo di dipendenza in media

$$\eta_{Y|X} = \frac{\sigma^2_{ext_Y}}{\sigma^2_Y} = \frac{\sum_{i=1}^{2} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 \cdot n_{i.}}{\sum_{j=1}^{4} \left(y_j - \mu_Y\right)^2 \cdot n_{.j}} = \frac{212,7}{13679} = 0,016$$

Risultato: bassa dipendenza in media; il valore dell'indice è prossimo a zero.

ESERCIZIO n°3

Lo stipendio medio annuo (Y), in migliaia di euro, dei dirigenti e il numero di dipendenti (X) di 5 aziende sono riportati nella tabella che segue:

Azienda	**1**	**2**	**3**	**4**	**5**
Stipendio (Y)	45	30	84	63	62
Numero dipendenti (X)	14	16	46	32	22

1. Determinare la mediana del numero di dipendenti;
2. Calcolare il coefficiente di correlazione lineare fra X e Y;
3. Calcolare i parametri della retta di regressione di Y su X;
4. Disegnare il grafico di dispersione tracciare la retta di regressione;
5. Valutare la bontà di adattamento della retta ai dati.

Soluzioni:

3.1

La mediana è l'intensità del carattere ordinabile posseduta dall'unità statistica che, nella successione ordinata delle modalità, è preceduta e seguita dallo stesso numero di unità statistiche del collettivo; per individuare quindi la mediana sarà necessario ordinare le unità statistiche in ordine crescente (decrescente) secondo il numero di dipendenti, la successione ordinata delle osservazioni date è la seguente:

Azienda	1	2	5	4	3
Numero dipendenti	14	16	22	32	46

l'unità statistica preceduta e seguita dallo stesso numero di unità statistiche è l' Azienda 5 che possiede un numero di dipendenti pari a 22, quindi la mediana è proprio 22.

3.2

Y	**X**	**Y^2**	**x^2**	**XY**
45	14	2025	196	630
30	16	900	256	480
84	46	7056	2116	3864
63	32	3969	1024	2016
62	22	3844	484	1364
SOMMA				
284	130	17794	4076	8354

$$\mu_y = \frac{\sum_{i=1}^{5} y_i}{n} = \frac{284}{5} = 56,8 \qquad \mu_{y^2} = \frac{\sum_{i=1}^{5} y_i}{n} = \frac{17794}{5} = 3558,8$$

$$\mu_x = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{130}{5} = 26 \qquad \mu_{x^2} = \frac{\sum_{i=1}^{5} x_i^2}{n} = \frac{4076}{5} = 815,2$$

$$\mu_{xy} = \frac{\sum_{i=1}^{5} x_i \cdot y_i}{n} = \frac{8354}{5} = 1670,8$$

$$cov(x,y) = \mu_{xy} - \mu_x \cdot \mu_y = 1670,8 - (56,8 \cdot 26) = 194$$

$$\sigma_x^2 = 815,2 - (26)^2 = 139,20 \qquad \sigma_x = 11,80$$

$$\sigma_y^2 = 3558,8 - (56,8)^2 = 332,56 \qquad \sigma_y = 18,24$$

Coefficiente di correlazione lineare fra X e Y

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{194}{11,8 \cdot 18,24} = 0,901$$

Risultato: X e Y sono correlati positivamente; il valore dell'indice è prossimo a 1.

3.3

Parametri della retta di regressione di Y su X

- b *(coefficiente angolare della retta)*

$$b = \frac{cov(x,y)}{var(x)} = -\frac{194}{139,2} = 1,394$$

- a *(intercetta)*

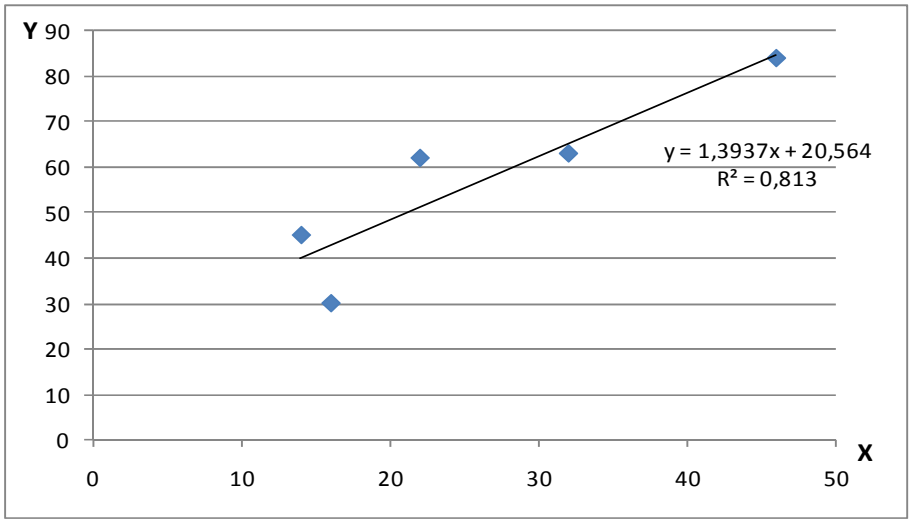
$$a = \mu_y - b \cdot \mu_x = 59,8 - (1,394 \cdot 26) = 20,56$$

Retta di regressione

$$\hat{y}_i = 20,56 + 1,394 \cdot x_i$$

3.4

Grafico di dispersione e retta di regressione



3.5

Bontà di adattamento del modello ai dati osservati

$$R^2 = \left(\frac{cov(x;y)}{\sigma_x \sigma_y} \right)^2 = \left(\frac{194}{11,8 \cdot 18,24} \right)^2 = (0,901)^2 = 0,813$$

Risultato: il valore del coefficiente di determinazione R^2 (moto vicino a 1) indica un elevato grado di adattamento della retta stimata ai dati osservati (c'è una forte relazione statistica di tipo lineare tra i valori delle variabili X e Y).