Esercitazione 7 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

27 Novembre 2008

Stipendio percepito	Età	N. di anni di servizio	Qualifica funzionale	Regime di impiego	Genere	Stato Civile	Abitazione di Proprietà
1750	34	12	Impiegato	Tempo pieno	F	Coniugato	Sì
1950	26	6	Impiegato	Tempo pieno	M	Non coniugato	No
3400	34	8	Operaio	Collaboratori esterni	F	Vedovo	Sì
2500	41	10	Operaio	Tempo pieno	F	Non coniugato	Sì
1150	38	9	Impiegato	Collaboratori esterni	M	Coniugato	No
2400	29	11	Operaio	Tempo pieno	F	Vedovo	Sì
2900	36	15	Impiegato	Tempo pieno	M	Non coniugato	Sì
2000	32	10	Impiegato	Tempo pieno	F	Vedovo	Sì
2150	36	7	Impiegato	Part time	M	Non coniugato	Sì
3900	48	8	Impiegato	Tempo pieno	F	Vedovo	Sì
1550	29	13	Dirigente	Collaboratori esterni	M	Coniugato	No
2000	31	7	Operaio	Collaboratori esterni	F	Coniugato	Sì
1800	33	8	Operaio	Tempo pieno	M	Coniugato	No
1850	42	9	Impiegato	Part time	M	Vedovo	Sì
1350	26	12	Impiegato	Collaboratori esterni	F	Coniugato	No
2450	30	15	Operaio	Tempo pieno	M	Non coniugato	Sì
2550	41	13	Impiegato	Collaboratori esterni	F	Non coniugato	Sì
2000	28	7	Impiegato	Collaboratori esterni	F	Vedovo	No
2400	33	8	Operaio	Tempo pieno	M	Non coniugato	Sì
1500	37	12	Impiegato	Collaboratori esterni	F	Coniugato	No

Esercizio 1

Vogliamo scoprire se esiste una qualche relazione (e se esiste di che natura è) tra il *numero anni di servizio* e l'*età*, per i primi 8 dipendenti. In questo caso il *numero anni di servizio* diventa la variabile dipendente e l'*età* la variabile esplicativa.

Età (x)	Anni di servizio (y)	x_i^2	y_i^2	$x \cdot y$
34	12	1156	144	408
26	6	676	36	156
34	8	1156	64	272
41	10	1681	100	410
38	9	1444	81	342
29	11	841	121	319
36	15	1296	225	540
32	10	1024	100	320
270	**81**	**9274**	**871**	**2767**

$\mu_x = \frac{270}{8} = 33{,}75 \qquad \mu_y = \frac{81}{8} = 10{,}125 \qquad \sum_{1=1}^{n} x_i^2 = 9274 \quad \sum_{1=1}^{n} y_i^2 = 871$

$\sigma^2_x = VAR(X) = E[X^2] - \mu^2 = \frac{9274}{8} - 33{,}75^2 = 1159{,}25 - 1139{,}06) = 20{,}19 \rightarrow \sigma = \sqrt{20{,}19} = 4{,}49$

$\sigma^2_y = VAR(Y) = E[Y^2] - \mu^2 = \frac{871}{8} - 10{,}125^2 = 108{,}875 - 102{,}5 = 6{,}375 \rightarrow \sigma = \sqrt{6{,}375} = 2{,}52$

$\mu(x \cdot y) = 2767/8 = 345{,}875$

$Cov_{x,y} = \mu(x \cdot y) - (\mu_x \cdot \mu_y) = 345{,}875 - (33{,}75 * 10{,}125) = 4{,}165$

$Corr_{x,y} = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{4{,}165}{4{,}49 \cdot 2{,}52} = \frac{4{,}165}{11{,}31} = 0{,}368$ Correlazione positiva

$\rho^2_{x,y} = R^2 = 0{,}368^2 = 0{,}1356$

Utilizziamo l'R^2 per capire in che misura la variazione della variabile *Numero anni di servizio* viene spiegata utilizzando la variabile *Età*. In questo caso l'R^2 vale 0,1356 e indica che approssimativamente il 13% della variazione degli anni di servizio può essere spiegato da un modello lineare utilizzando l'età.

STIMA DELLA RETTA DI REGRESSIONE

$\hat{y} = \hat{\alpha} + \hat{\beta} x_i$

$\hat{\beta} = \frac{Cov_{x,y}}{\sigma^2_x} = \frac{4{,}165}{20{,}19} = 0{,}20$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 10{,}125 - (0{,}20 \cdot 33{,}75) = 3{,}16$

$\hat{y} = \hat{\alpha} + \hat{\beta} x_i = 3{,}16 + 0{,}20 x_i$

Esercizio 2

Considerando il DATASET DIPENDENTI, si calcoli la correlazione tra i caratteri STIPENDIO PERCEPITO ED ETA'. A tale riguardo si suddividano entrambi i caratteri in tre classi equi-ampie.

Soluzione

L'ampiezza costante delle tre classi per il carattere "stipendio percepito" (espresso in migliaia di euro) si ottiene come:

range (STIPENDIO PERCEPITO) = 3900 – 1150 = 2750

L'ampiezza delle classi della prima distribuzione di frequenza è pari a:

Ampiezza **stipendio percepito** = range/3 = 2750/3 = 923,67 (per cui avremo due classi di ampiezza 925 e una di ampiezza 900).
L'ampiezza costante delle tre classi per il carattere "ETA'" si ottiene come:

range (ETA') = 48-26 = 22

L'ampiezza delle classi della prima distribuzione di frequenza è pari a:

Ampiezza **stipendio percepito** = range/3 = 22/3 = 7,3 (per cui avremo due classi di ampiezza 7 e una di ampiezza 8)

Stipendio percepito / ETA'	[1150, 2075]	]2075, 3000]	]3000, 3900]	**Totale**
[26,33]	7	3	0	**10**
]33,40]	3	2	1	**6**
]40,48]	1	2	1	**4**
Totale	**11**	**7**	**2**	**20**

Valori Centrali di y = 1612,5; 2537,5; 3450
Valori Centrali di x = 29,5 36,5; 44

$\hat{y}_i$	n_j	$\hat{x}_i$	n_i	$\hat{y}_i n_j$	$\hat{x}_i n_i$	$\hat{y}_j^2$	$\hat{y}_j^2 n_i$	$\hat{x}_i^2$	$\hat{x}_i^2 n_i$
1612,5	11	29,5	10	17737,5	295	2600156,3	28601718,75	870,25	8702,5
2537,5	7	36,5	6	17762,5	219	6438906,3	45072343,75	1332,25	7993,5
3450	2	44	4	6900	176	11902500	23805000	1936	7744
				42400	**690**		**97479062,5**		**24440**

Media delle y = μ_y = 42400/20 = **2120**
Media delle x = μ_x = 690/20 = **34,5**

Per calcolare il termine $\sum_i \sum_j \hat{x}_i \hat{y}_j n_{ij}$ è necessario costruire la tabella delle $\hat{x}_i \hat{y}_j n_{ij}$ come segue:

$\hat{x}_i$ \ $\hat{y}_i$	1612,5	2537,5	3450
29,5	332981,25	224568,75	0
36,5	176568,75	185237,5	125925
44	70950	223300	151800

La somma degli elementi all'interno della tabella è pari a 1491331, da cui:

$$\frac{\sum_i \sum_j \hat{x}_i \hat{y}_j n_{ij}}{n} = \frac{1491331}{20} = 74566,55$$

Cov = $\sigma_{x,y} = \mu_{x,y} - (\mu_x * \mu_y) = 74566,55 - (2120*34,5) = 74566,55 - 73140 = 1426,55$

$$\sigma^2{}_x = VAR(X) = [\hat{X}^2 \cdot n_i] - \mu^2 = \frac{24440}{20} - 34,5^2 = 31,75 \rightarrow \sigma = \sqrt{31,75} = 5,63$$

$$\sigma^2{}_y = VAR(Y) = [\hat{Y}^2 \cdot n_j] - \mu^2 = \frac{97479062,5}{20} - 2120^2 = 379553,125 \rightarrow \sigma = \sqrt{379553,125} = 616$$

Corr = $\rho_{x,y} = \dfrac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \dfrac{1426,55}{5,63 \cdot 616} = \dfrac{1426,55}{3468,08} = 0,41$ <u>Correlazione positiva</u>