Esercitazione 6 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

25 Febbraio 2011

N	SESSO	ETA'	PESO	ALTEZZA	DIPLOMAI	COMPONENTI	OCCHIALI	FUMO	REDDITO
1	0	20,6	65	180	Ist.Tecnico	6	0	1	0,7
2	0	20,2	75	180	Liceo	4	0	0	0,2
3	0	20,3	60	173	Ist.Tecnico	4	1	0	1,6
4	0	23,9	93	187	Liceo	8	0	1	2,5
5	0	21,4	66	164	Ist.Tecnico	5	0	0	3,2
6	0	25	84	186	Ist.Tecnico	4	0	0	0,1
7	0	20,8	67	175	Altro dipl.	4	0	1	3,8
8	0	20,6	89	170	Liceo	3	1	0	1,3
9	0	27,1	71	180	Liceo	1	0	1	1,2
10	0	23,3	63	170	Liceo	4	0	0	1,7
11	1	20,5	51	161	Ist.Tecnico	4	0	1	1,9
12	1	19,1	58	167	Ist.Tecnico	5	1	1	0,8
13	1	22,1	67	165	Altro dipl.	5	1	1	0,4
14	1	21,8	51	156	Ist.Tecnico	4	0	0	1,8
15	1	19,2	60	170	Ist.Tecnico	5	1	1	1,9
16	1	20,8	55	165	Liceo	4	1	1	3,2
17	1	21	55	158	Liceo	5	1	0	2,1
18	1	20,9	58	170	Liceo	5	1	1	0,1
19	1	22,7	76	170	Liceo	6	1	0	1,6
20	1	21	55	165	Liceo	7	0	0	2,2

Esercizio 1

Vogliamo scoprire se esiste una qualche relazione (e se esiste di che natura è) tra il peso e l'altezza dei primi 5 uomini e delle prime 5 donne.
Poichè sospettiamo che il peso possa dipendere dall'altezza, il peso diventa la variabile dipendente e l'altezza la variabile esplicativa.

Altezza(x)	Peso(y)	x_i^2	y_i^2	$x \cdot y$
180	65	32400	4225	11700
180	75	32400	5625	13500
173	60	29929	3600	10380
187	93	34969	8649	17391
164	66	26896	4356	10824
161	51	25921	2601	8211
167	58	27889	3364	9686
165	67	27225	4489	11055
156	51	24336	2601	7956
170	60	28900	3600	10200
1703	**646**	**290865**	**43110**	**110903**

$$\mu_x = \frac{1703}{10} = 170{,}3 \qquad \mu_y = \frac{646}{10} = 64{,}6 \qquad \sum_{1=1}^{n} x_i^{\,2} = 290865 \qquad \sum_{1=1}^{n} y_i^{\,2} = 43110$$

$$\sigma^2{}_x = VAR(X) = E[X^2] - \mu^2 = \frac{290865}{10} - 170{,}3^2 = 29086{,}5 - 29002{,}9 = 83{,}6 \rightarrow \sigma = \sqrt{83{,}6} = 9{,}14$$

$$\sigma^2{}_y = VAR(YX) = E[Y^2] - \mu^2 = \frac{43110}{10} - 64{,}6^2 = 4311 - 4173{,}16 = 137{,}84 \rightarrow \sigma = \sqrt{137{,}84} = 11{,}7$$

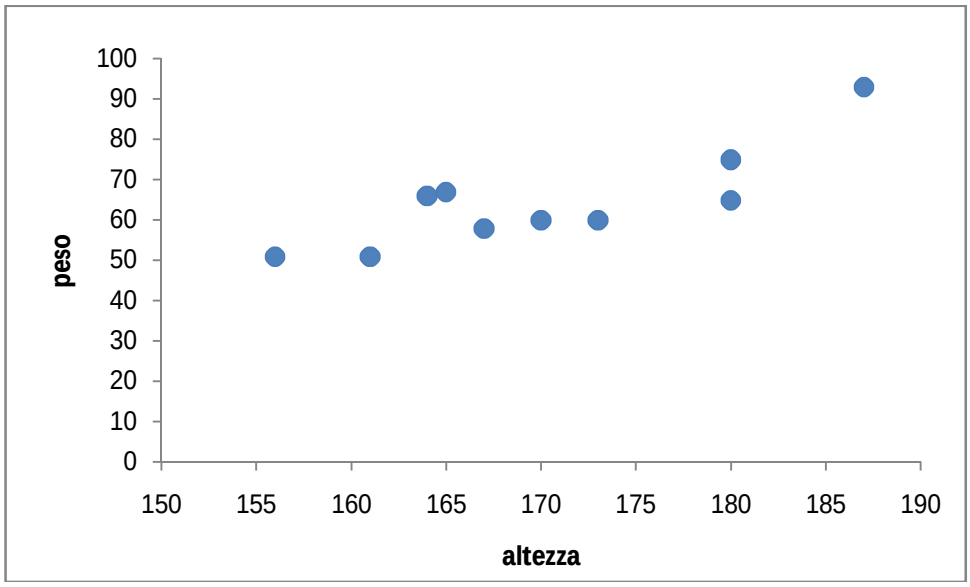
$$\mu(x \cdot y) = 110903/10 = 11090{,}3$$

$$Cov_{x,y} = \mu(x \cdot y) - (\mu_x \cdot \mu_y) = 11090{,}3 - (170{,}3 \cdot 64{,}6) = 88{,}92$$

$$Corr_{x,y} = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{88{,}92}{9{,}14 \times 11{,}7} = 0{,}83$$ **<u>Correlazione positiva</u>**

$$\rho^2_{x,y} = R^2 = 0{,}83^2 = 0{,}69$$

Diagramma di dispersione



Esercizio n 2

Calcolare la correlazione tra le variabili Peso ed Altezza ripartite in 3 classi equiampie

Altezza= 187-156=31/3=10,33 Essendo 20 le modalità per cui avremo 2 classi da 10 e 1 classe da 11 modalità

[156,166] [166,176] [176,187]

Peso=93-51=42/3=14 [51, 65] [65, 79] [79, 93]

Peso / Altezza	[51, 65]	[65, 79]	[79, 93]	**tot**
[156,166]	5	2	0	**7**
[166,176]	5	2	1	**8**
[176,187]	1	2	2	**5**
tot	**11**	**6**	**3**	**20**

Valori centrali delle x = 161; 171; 181,5

Valori centrali delle y = 58; 72; 86

Media delle x = (161*7+171*8+181,5*5)/20 = 170,125
Media delle y = (58*11+72*6+86*3)/20 = 66,4

$$(x \cdot y) = (58 \times 161 \times 5) + (58 \times 171 \times 5) + (58 \times 181{,}5 \times 1) + (72 \times 161 \times 2) + (72 \times 171 \times 2) + (72 \times 181{,}5 \times 2) + (86 \times 161 \times 0) + (86 \times 171 \times 1) + (86 \times 181{,}5 \times 2) =$$

46690	23184	0	
49590	24624	14706	somma
10527	26136	31218	**226675**

$$Cov = \sigma_{x,y} = \bar{x}_{x,y} - (\bar{x} \cdot \bar{y}) = \frac{226675}{20} - (170{,}125 \cdot 66{,}4) = 11333{,}75 - 11296{,}3 = 37{,}45$$

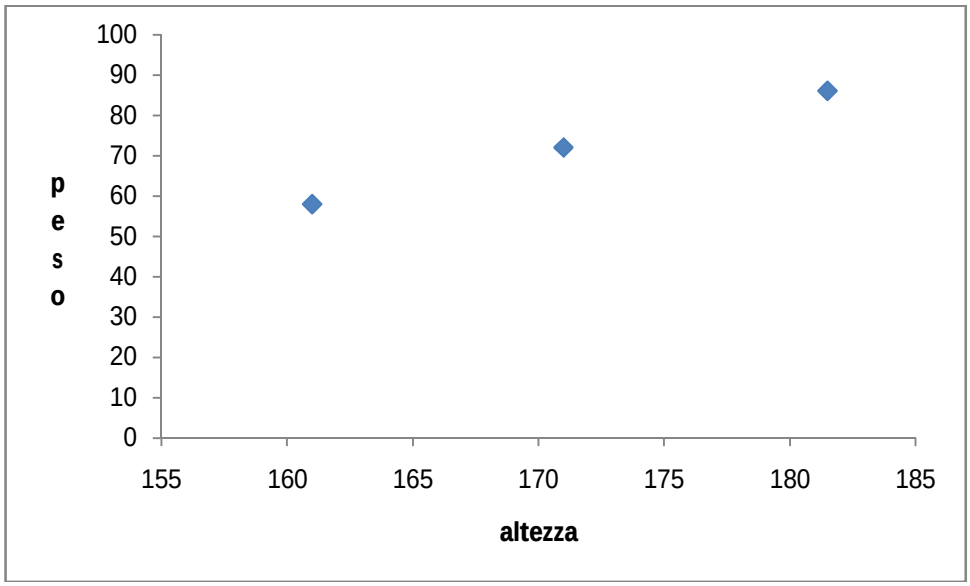
$$Var(x) = \sigma_x^2 = \frac{1}{N} \sum (\hat{c}_i - \bar{x})^2 \cdot n_i = (161 - 170{,}125)^2 \cdot 7 + (171 - 170{,}125)^2 \cdot 8 + (181{,}5 - 170{,}125)^2 \cdot 5 \Big/ 20$$

$$Var(x) = \sigma_x^2 = \frac{582{,}86 + 6{,}125 + 646{,}95}{20} = 61{,}8 \rightarrow \sqrt{\sigma^2} = \sqrt{61{,}8} = 7{,}86$$

$$Var(y) = \sigma_y^2 = \frac{1}{N} \sum (\hat{c}_i - \bar{x}_y)^2 \cdot n_i = (58 - 66{,}4)^2 \cdot 11 + (72 - 66{,}4)^2 \cdot 6 + (86 - 66{,}4)^2 \cdot 3 \Big/ 20$$

$$Var(y) = \sigma_y^2 = \frac{776{,}16 + 188{,}16 + 1152{,}48}{20} = 105{,}84 \rightarrow \sqrt{\sigma^2} = \sqrt{105{,}84} = 10{,}28$$

$$Corr = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{37{,}45}{7{,}86 \cdot 10{,}28} = \frac{37{,}45}{80{,}80} = 0{,}46$$ **Correlazione positiva media**



Esercizio 3

Considerando i caratteri “tipo di diploma” e ”Reddito”, si può affermare che il Reddito dipende in media dal “tipo di diploma”?

Utilizziamo l’indice $\eta_{y|x}$ per misurare il grado di indipendenza in media del carattere qualitativo *y* dal carattere quantitativo *x*.

Ordiniamo i dati del carattere Reddito e suddividiamolo in 2 classi equiampie

REDDITO	*DIPLOMA*
0,1	Ist.Tecnico
0,1	Liceo
0,2	Liceo
0,4	Altro dipl.
0,7	Ist.Tecnico
0,8	Ist.Tecnico
1,2	Liceo
1,3	Liceo
1,6	Ist.Tecnico
1,6	Liceo
1,7	Liceo
1,8	Ist.Tecnico
1,9	Ist.Tecnico
1,9	Ist.Tecnico
2,1	Liceo
2,2	Liceo
2,5	Liceo
3,2	Ist.Tecnico
3,2	Liceo
3,8	Altro dipl.

Range = 3,8-0,1 = 3,7
2 classi equiampie = 3,7/2= 1,85

REDDITO / TIPO DI DIPLOMA	[0,1,1,95]	]1,95,3,7]	**totale**
Liceo	6	4	**10**
Ist. tecnici	7	1	**8**
Altro diploma	1	1	**2**
totale	**14**	**6**	**20**

Quando y è quantitativo

$$\eta_{Y|X} = \frac{\sigma^2_{EXT_Y}}{\sigma^2_Y} = \frac{\sum_{i=1}^{k}\left(\mu_{Y|X=x_i} - \mu_Y\right)^2 n_{i\bullet}}{\sum_{j=1}^{h}\left(\hat{y}_j - \mu_Y\right)^2 n_{\bullet j}}$$

Considerando che:

valori centrali: $y_1 = 1{,}025$; $y_2 = 2{,}825$

$$\bar{x} = \frac{\sum_{j=1}^{3} y_j n_{\cdot j}}{N} \quad \frac{1{,}025*14+2{,}825*6}{20} \quad \textbf{= 1,565 reddito medio}$$

$$\bar{x}_1 = Liceo = \frac{\sum_{j=1}^{2} y_j n_{1j}}{n_{1\cdot}} = \frac{1{,}025 * 6 + 2{,}825 * 4}{10} = 1{,}745 \text{ reddito Liceo}$$

$$\bar{x}_2 = Ist.tecnici = \frac{\sum_{j=1}^{2} y_j n_{1j}}{n_2} = \frac{1{,}025*7 + 2{,}825*1}{8} = 1{,}25 \text{ n. reddito Ist. tecnici}$$

$$\bar{x}_3 = Altrodiploma = \frac{\sum_{j=1}^{2} y_j n_{1j}}{n_3} = \frac{1{,}025*1 + 2{,}825*1}{2} = 1{,}925 \text{ n. anni di servizio Coll. Est.}$$

Commento: si può vedere che le medie delle distribuzioni condizionate differiscono dalla media generale di Y, anche se non di molto, quindi i due caratteri **non sono indipendenti in media**.

Ma quanto è forte il legame di dipendenza in media?

Calcolo del numeratore dell'indice

$\sum (\mu_{y|x} - \mu_y)^2 * n_{i.} = (1{,}565 - 1{,}745)^2 * 10 + (1{,}565 - 1{,}25)^2 * 8 + (1{,}565 - 1{,}925)^2 * 2 =$

$= 0{,}324 + 0{,}7938 + 0{,}2592 = 1{,}377$

Calcolo del denominatore dell'indice

$\sum (\acute{y}_j - \mu_y)^2 * n_{.j} = (1{,}025 - 1{,}565)^2 * 14 + (2{,}825 - 1{,}565)^2 * 6 =$

$= 4{,}0824 + 9{,}5256 = 13{,}608$

Calcolo dell'indice

$$\eta_{y|x} = \frac{\sum_{i=1}^{k} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 n_{i\bullet}}{\sum_{j=1}^{h} \left(\hat{y}_j - \mu_Y\right)^2 n_{\bullet j}} = \frac{1{,}377}{13{,}608} = 0{,}10$$

Considerando che **$0 \leq \eta_{y/x} \leq 1$** abbiamo un bassissimo grado di dipendenza in media, piuttosto siamo vicini ad un'ipotesi di *indipendenza in media*.