Università di Cassino

Esercitazione di Statistica 1 del 20 novembre 2006

Dott.ssa Simona Balzano

Considerando il DATASET STUDENTI risolvere i seguenti quesiti:

1) Esiste connessione tra i caratteri CORSO LAUREA e ATTIVITÀ SPORTIVA?
2) Considerando le classi di altezza 160 |–| 170; 170 –| 180; 180 –| 190, si può affermare che l'ALTEZZA dipende in media dal SESSO?
3) In quale misura i caratteri PESO e ALTEZZA sono tra loro correlati?
4) In quale misura gli stessi caratteri sono correlati se espressi attraverso la seguente distribuzione doppia?

Peso \ Altezza	160 |-| 170	170 -| 180	180 -| 190	**totale**
46 |-| 56	6	1		**7**
56 -| 66	2	1	2	**5**
66 -| 76	2	1	2	**5**
76 -| 86		1	2	**3**
Totale	**10**	**4**	**6**	**20**

Soluzione

1)

La distribuzione doppia dei caratteri CORSO LAUREA e ATTIVITÀ SPORTIVA è la seguente:

Corso laurea \ Attività sportiva	Nulla	Media	Alta	**Totale**
Biologia	0	1	0	**1**
Informatica	4	7	1	**12**
Matematica	2	5	0	**7**
Totale	**6**	**13**	**1**	**20**

Trattandosi di due caratteri qualitativi, il loro grado di connessione si misura attraverso l'indice χ^2:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

Le frequenze teoriche $\hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{N}$ sono raccolte nella seguente tabella:

Attività sportiva / Corso laurea	Nulla	Media	Alta	**Totale**
Biologia	0,3	0,65	0,05	**1**
Informatica	3,6	7,8	0,6	**12**
Matematica	2,1	4,55	0,35	**7**
Totale	**6**	**13**	**1**	**20**

Sostituendo nella formula si ha:

$$\chi^2 = \frac{(0-0,3)^2}{0,3} + \frac{(1-0,65)^2}{0,65} + \frac{(0-0,05)^2}{0,05} + \cdots + \frac{(5-4,55)^2}{4,55} + \frac{(0-0,35)^2}{0,35} = \mathbf{1,331}$$

Quindi si ha:

$$\phi^2 = \frac{\chi^2}{N} = \frac{1,331}{20} = \mathbf{0,066}$$

Tale valore va confrontato con l'intervallo [0, 2], in quanto
$0 \le \varphi^2 \le \min(r-1;\ c-1)$

Formula alternativa

$$\chi^2 = n\left[\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

n_{ij}^2	Nulla	Media	Alta
Biologia	0	1	0
Informatica	16	49	1
Matematica	4	25	0

$n_{i.} \times n_{.j}$	Nulla	Media	Alta
Biologia	6	13	1
Informatica	72	156	12
Matematica	42	91	7

$$\chi^2 = 20 \times \left[\left(\frac{0}{6} + \frac{1}{13} + \frac{0}{1} + \cdots + \frac{4}{42} + \frac{25}{91} + \frac{0}{7} \right) - 1 \right] = \mathbf{1,331}$$

2)

L'indice per misurare il grado di indipendenza in media di un carattere quantitativo y da uno qualitativo (o quantitativo) x è il rapporto di correlazione o $\eta^2_{y|x}$

Sesso \ Altezza	160 |-| 170	170 -| 180	180 -| 190	**totale**
Femmine	7	1	0	**8**
Maschi	3	8	1	**12**
Totale	**10**	**9**	**1**	**20**

$$\eta^2_{Y|X} = \sqrt{\frac{\sigma^2_{est}}{\sigma^2}} = \sqrt{\frac{dev_{est}}{dev(Y)}} = \sqrt{\frac{\sum_{i=1}^{r}(\mu_i - \mu)^2 n_{i.}}{\sum_{j=1}^{c}(y_j - \mu)^2 n_{.j}}}$$

Considerando che:

valori centrali: $y_1 = 165$; $y_2 = 175$; $y_3 = 185$

$$\mu = \frac{\sum_{j=1}^{3} y_j n_{.j}}{N} = \frac{(165\times 10)+(175\times 9)+(185\times 1)}{20} = 170,5\,;$$ altezza media generale

$$\mu_1 = \mu_{femine} = \frac{\sum_{j=1}^{3} y_j n_{1j}}{n_{1.}} = \frac{(165\times 7)+(175\times 1)}{8} = 166,25\,;$$ altezza media femmine

$$\mu_2 = \mu_{maschi} = \frac{\sum_{j=1}^{3} y_j n_{2j}}{n_{2.}} = \frac{(165\times 3)+(175\times 8)+(185\times 1)}{12} = 173,33\,;$$ altezza media maschi

$$\eta^2_{altezza\ |\ sesso} = \sqrt{\frac{\left[(166,25-170,5)^2\times 8\right]+\left[(173,33-170,5)^2\times 12\right]}{\left[(165-170,5)^2\times 10\right]+\left[(175-170,5)^2\times 9\right]+\left[(185-170,5)\times 1\right]}} =$$

$$= \sqrt{\frac{144,5+96,3}{302,5+182,25+210,25}} = \sqrt{\frac{240,83}{695}} = \sqrt{0,436} = \mathbf{0,59}$$

Il valore va confrontato con l'intervallo [0, 1] in cui è compreso $\eta^2_{y|x}$, quindi indica un discreto grado di dipendenza in media.

2)
La presenza di dipendenza lineare si verifica attraverso il coefficiente di correlazione lineare di Bravais-Pearson ρ:

$$\rho_{XY} = \frac{\mathrm{cod}(x,y)}{\sqrt{\mathrm{dev}(x)\cdot \mathrm{dev}(y)}} = \frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

Le quantità necessarie al calcolo dell'indice sono racchiuse nella tabella seguente:

Altezza X	Peso Y	Xi2	Yi2	xy
179	65	32.041	4.225	11.635
180	62	32.400	3.844	11.160
165	50	27.225	2.500	8.250
160	49	25.600	2.401	7.840
160	47	25.600	2.209	7.520
160	48	25.600	2.304	7.680
164	56	26.896	3.136	9.184
170	59	28.900	3.481	10.030
180	73	32.400	5.329	13.140
186	86	34.596	7.396	15.996
170	66	28.900	4.356	11.220
180	68	32.400	4.624	12.240
180	85	32.400	7.225	15.300
176	56	30.976	3.136	9.856
170	72	28.900	5.184	12.240
180	65	32.400	4.225	11.700
170	75	28.900	5.625	12.750
172	70	29.584	4.900	12.040
178	80	31.684	6.400	14.240
162	49	26.244	2.401	7.938
3.442	**1.281**	**593.646**	**84.901**	**221.959**

Da cui si ricava:

$$\mu_x = \frac{3342}{20} = 172,1$$

$$\mu_y = \frac{1281}{20} = 64,05$$

$$\mathrm{cod}(x,y) = \sum_{i=1}^{N} x_i y_i - n\mu_x\mu_y =$$

$$= 221.959 - 20 \times 172,1 \times 64,05 = 1498,9$$

$$\mathrm{dev}(x) = \sum_{i=1}^{20} x_i^2 - n\mu_x^2 = 593.646 - 20 \times 172,1^2 = 1.277,8$$

$$dev(y) = \sum_{i=1}^{20} y_i^2 - n\mu_y^2 = 84.901 - 20 \times 64,05^2 = 2.852,95$$

$$\rho_{XY} = \frac{cod(x,y)}{\sqrt{dev(x) \cdot dev(y)}} = \frac{1.498,9}{\sqrt{1.277,8 \times 2.852,95}} = \mathbf{0,785}$$

Tale valore va confrontato con l'intervallo [-1, 1], quindi indica una correlazione lineare positiva molto forte.

Calcolo di ρ per i primi 5 individui (svolto in aula):

Altezza x_i	**Peso** y_i	x_i^2	y_i^2	$x_i y_i$
179	65	32.041	4.225	11.635
180	62	32.400	3.844	11.160
165	50	27.225	2.500	8.250
160	49	25.600	2.401	7.840
160	47	25.600	2.209	7.520
844	**273**	**142.866**	**15.179**	**46.405**

$$\mu_x = \frac{844}{5} = 168,8$$

$$\mu_y = \frac{273}{5} = 54,6$$

$$cod(x,y) = \sum_{i=1}^{n} x_i y_i - n\mu_x\mu_y =$$

$$= 46.405 - 5 \times 168,8 \times 54,6 = 322,6$$

$$dev(x) = \sum_{i=1}^{5} x_i^2 - n\mu_x^2 = 142.866 - 5 \times 168,8^2 = 398,8$$

$$dev(y) = \sum_{i=1}^{5} y_i^2 - n\mu_y^2 = 15.179 - 5 \times 54,6^2 = 273,2$$

$$\rho_{XY} = \frac{cod(x,y)}{\sqrt{dev(x) \cdot dev(y)}} = \frac{322,6}{\sqrt{398,8 \times 273,2}} = \mathbf{0,977}$$

Tale valore va confrontato con l'intervallo [-1, 1], quindi indica l'esistenza di una relazione lineare positiva praticamente perfetta.

3)
Quandoi caratteri sono espressi come distribuzione doppia di frequenza il coefficiente di correlazione si calcola come:

$$\rho = \frac{\mu_{XY} - \mu_X\mu_Y}{\sqrt{\left(\mu_{2X} - \mu_X^2\right)\left(\mu_{2Y} - \mu_Y^2\right)}} = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}\hat{x}_i\hat{y}_j n_{ij} - \mu_X\mu_Y}{\sqrt{\frac{1}{n}\sum_{i=1}^{k}x_i^2 n_{i.} - \mu_X^2}\sqrt{\frac{1}{n}\sum_{j=1}^{h}y_j^2 n_{.j} - \mu_Y^2}}$$

A partire dalla tabella delle frequenze:

Peso (y) \ Altezza (x)	160 |-| 164	164 -| 170	170 -| 178	178 -| 186	**Totale**
46 |-| 56	5	1	1	0	**7**
56 -| 66	0	2	0	3	**5**
66 -| 76	0	2	1	2	**5**
76 -| 86	0	0	1	2	**3**
Totale	**5**	**5**	**3**	**7**	**20**

Valori centrali per l'altezza: $x_1 = 162$; $x_2 = 167$; $x_3 = 174$; $x_4 = 182$
Valori centrali per il peso: $y_1 = 51$; $y_2 = 61$; $y_3 = 71$; $y_4 = 81$

Le quantità $\hat{x}_i\hat{y}_j n_{ij}$ sono raccolte nella tabella che segue:

$\hat{x}_i\hat{y}_j n_{ij}$	**162**	**167**	**174**	**182**
51	41.310	8.517	8.874	0
61	0	20.374	0	33.306
71	0	23.714	12.354	25.844
81	0	0	14.094	29.484

Totale generale: 217.871

Da cui deriva:

$$\mu_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}x_i y_j n_{ij}}{n} = \frac{217.871}{20} = 10.894$$

Le altre quantità necessarie sono contenute nella seconda tabella:

x_i	$n_{i.}$	y_i	$n_{.j}$	$x_i n_{i.}$	$y_j n_{.j}$	x_i^2	$x_i^2 n_{i.}$	y_j^2	$y_j^2 n_{.j}$
162	5	51	7	810	357	2601	131220	26244	18207
167	5	61	5	835	305	3721	139445	27889	18605
174	3	71	5	522	355	5041	90828	30276	25205
182	7	81	3	1274	243	6561	231868	33124	19683
Totali	**20**		**20**	**3.441**	**1.260**		**593.361**		**81.700**

$$\mu_X = \frac{1}{n}\sum_{i=1}^{k} x_i n_{i.} = \frac{3.441}{20} = 172,05$$ altezza media

$$\mu_Y = \frac{1}{n}\sum_{j=1}^{h} y_j n_{.j} = \frac{1260}{20} = 63$$ peso medio

$$\mu_{2X} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_{i.} = \frac{593.361}{20} = 29.668,05$$

$$\mu_{2Y} = \frac{1}{n}\sum_{j=1}^{h} y_j^2 n_{.j} = \frac{81.700}{20} = 4.085$$

Sostituendo i valori ottenuti nella formula:

$$\operatorname{cov}(x,y) = \mu_{XY} - \mu_x\mu_y = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h} \hat{x}_i \hat{y}_j n_{ij} - \mu_X\mu_Y = 10.894 - 172,05 \times 63 = 54,4$$

$$\rho = \frac{\mu_{XY} - \mu_X\mu_Y}{\sqrt{\left(\mu_{2X} - \mu_X^2\right)\left(\mu_{2Y} - \mu_Y^2\right)}} = \frac{54,4}{\sqrt{\left(29.668,05 - 172,05^2\right) \times \left(4.085 - 63^2\right)}} = \frac{54,4}{88,06} = \mathbf{0,62}$$

Tale valore va confrontato con l'intervallo [-1, 1], quindi indica una correlazione lineare positiva abbastanza forte.