Università di Cassino

Esercitazioni di Statistica 1 del 26 Febbraio 2010

Dott. Mirko Bevilacqua

ESERCIZIO n° 1

Considerando le classi di altezza 160 |–| 164; 164 –| 170; 170 –| 178; 178 –| 186 per un collettivo di 20 persone, si può affermare che l'ALTEZZA dipende in media dal Paese di provenienza?

Altezza (Y) / Nazione (X)	160 |-| 164	164 -| 170	170 -| 178	178 -| 186	**Totale**
CINA	5	1	1	0	**7**
ITALIA	0	1	1	3	**5**
FRANCIA	0	0	1	4	**5**
SVEZIA	0	0	1	2	**3**
Totale	**5**	**2**	**4**	**9**	**20**

In questo caso lo studio riguarda l'associazione tra un carattere quantitativo (Y) e un carattere qualitativo (X). L'analisi della dipendenza può essere condotta confrontando le distribuzioni del carattere Y in corrispondenza delle diverse modalità del carattere X.
Y (carattere quantitativo) è indipendente in media da X (carattere qualitativo) se tutte le medie condizionate della variabile Y sono fra loro uguali e uguali quindi anche alla media marginale:

$\mu_{Y/X=xi}=\mu_Y$ per ogni i = 1,2,…,k

Considerando la scomposizione della varianza si può introdurre il seguente indice relativo di dipendenza in media:

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

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

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

1. Media generale di Y:

$$\mu_Y = \frac{\sum_{j=1}^{4} y_j n_{.j}}{N} = \frac{(162 \cdot 5) + (167 \cdot 2) + (174 \cdot 4) + (182 \cdot 9)}{20} = 173,9$$

2. Medie di Y condizionate alle modalità di X

$$\mu_{CINA} = \frac{\sum_{j=1}^{4} y_{CINA} n_{1j}}{n_{1.}} = \frac{(162 \cdot 5) + (167 \cdot 1) + (174 \cdot 1) + (182 \cdot 0)}{7} = 164$$

$$\mu_{ITA} = \frac{\sum_{j=1}^{4} y_{ITA} n_{2j}}{n_{2.}} = \frac{(162 \cdot 0) + (167 \cdot 1) + (174 \cdot 1) + (182 \cdot 3)}{5} = 177,4$$

$$\mu_{FRA} = \frac{\sum_{j=1}^{4} y_{FRA} n_{3j}}{n_{3.}} = \frac{(162 \cdot 0) + (167 \cdot 0) + (174 \cdot 1) + (182 \cdot 4)}{5} = 180,4$$

$$\mu_{SVE} = \frac{\sum_{j=1}^{4} y_{SVE} n_{4j}}{n_{4.}} = \frac{(162 \cdot 0) + (167 \cdot 0) + (174 \cdot 1) + (182 \cdot 2)}{3} = 179,3$$

3. Calcolo del numeratore dell'indice:

$$\sum_{i=1}^{r} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 \cdot n_{i.} =$$
$$(164 - 173,9)^2 \cdot 7 + (177,4 - 173,9)^2 \cdot 5 +$$
$$+ (180,4 - 173,9)^2 \cdot 5 + (179,3 - 173,9)^2 \cdot 3 = 989,02$$

4. Calcolo del denominatore dell'indice:

$$\sum_{j=1}^{c} \left(y_j - \mu_Y\right)^2 \cdot n_{.j} =$$
$$(162 - 173,9)^2 \cdot 5 + (167 - 173,9)^2 \cdot 2 +$$
$$+ (174 - 173,9)^2 \cdot 4 + (182 - 173,9)^2 \cdot 9 = 1393,8$$

5. Calcolo dell'indice:

$$\eta_{Y|X} = \frac{\sigma^2_{ext_Y}}{\sigma^2_Y} = \frac{\sum_{i=1}^{r} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 \cdot n_{i.}}{\sum_{j=1}^{c} \left(y_j - \mu_Y\right)^2 \cdot n_{.j}} = \frac{989,02}{1393,8} = 0,71$$

Il valore dell'indice indica un grado di dipendenza in media abbastanza elevato

ESERCIZIO n°2:

Da un collettivo di 20 studenti della facoltà di economia sono stati rilevati i voti ottenuti nella prova scritta e orale dell'esame di statistica Si calcoli la concordanza o la discordanza tra i due caratteri mediante il calcolo della covarianza. Le votazioni sono riportate nella seguente tabella.

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Prova scritta	29	29	27	24	24	23	23	23	21	21	19	19	18	18	18	18	18	18	18	17
Prova orale	22	28	30	25	27	23	22	25	24	24	25	24	25	23	27	26	22	25	23	18

La covarianza tra due caratteri quantitativi è definita come la media dei prodotti degli scostamenti delle variabili X e Y dalle rispettive medie:

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

Quest'indice misura la concordanza o la discordanza tra due caratteri quantitativi. È definita sull'insieme dei numeri reali ed è positiva se al numeratore prevalgono i prodotti degli scostamenti concordi (tutti e due positivi o tutti e due negativi), mentre è negativa se prevalgono i prodotti degli scostamenti discordi. Il numeratore della covarianza è detto **codevianza**.
Se due caratteri sono statisticamente indipendenti la loro covarianza è zero. Tuttavia, se la covarianza è nulla, non è detto che i due caratteri siano indipendenti. Infatti, la covarianza si annulla se i prodotti degli scostamenti dalla media si compensano tra loro, ma ciò può avvenire anche se tra i due caratteri sussiste una relazione di dipendenza non di tipo lineare.

1. **Media voto scritto**

$$\mu_x = \frac{29 + 29 + 27 + \ldots\ldots\ldots\ldots\ldots + 18 + 18 + 17}{20} = 21,25$$

2. **Media voto orale**

$$\mu_y = \frac{22 + 28 + 30 + \ldots\ldots\ldots\ldots\ldots + 25 + 23 + 18}{20} = 24,4$$

X	*Y*	*x*Y*	$x-\mu_x$	$y-\mu_y$	$(x-\mu_x)(y-\mu_y)$
29	**22**	638	7,75	-2,4	-18,6
29	**28**	812	7,75	3,6	27,9
27	**30**	810	5,75	5,6	32,2
24	**25**	600	2,75	0,6	1,65
24	**27**	648	2,75	2,6	7,15
23	**23**	529	1,75	-1,4	-2,45
23	**22**	506	1,75	-2,4	-4,2
23	**25**	575	1,75	0,6	1,05
21	**24**	504	-0,25	-0,4	0,1
21	**24**	504	-0,25	-0,4	0,1
19	**25**	475	-2,25	0,6	-1,35
19	**24**	456	-2,25	-0,4	0,9
18	**25**	450	-3,25	0,6	-1,95
18	**23**	414	-3,25	-1,4	4,55
18	**27**	486	-3,25	2,6	-8,45
18	**26**	468	-3,25	1,6	-5,2
18	**22**	396	-3,25	-2,4	7,8
18	**25**	450	-3,25	0,6	-1,95
18	**23**	414	-3,25	-1,4	4,55
17	**18**	306	-4,25	-6,4	27,2
		$\sum_{i=1}^{n} x_i y_i = 10441$			$\sum_{i=1}^{n}(y_i-\mu_y)\cdot(x_i-\mu_x) = 71$

$$cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu_y\right)\cdot\left(x_i - \mu_x\right) = \frac{71}{20} = 3,55$$

<u>Formula alternativa per la covarianza:</u>

$$cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu_y\right)\cdot\left(x_i - \mu_x\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_x \cdot \mu_y$$

Media dei prodotti:

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{10441}{20} = 522,05$$

Calcolo della covarianza con la formula alternativa:

$$cov\,(x,y) = 522,05 - (21,25 \cdot 24,4) = 3,55$$

ESERCIZIO n°3:

Le quantità di precipitazioni e le temperature medie registrate in 10 stazioni meteorologiche sono state le seguenti:

stazione meteorologica	1	2	3	4	5	6	7	8	9	10
Precipitazioni	29	35	87	32	112	14	26	120	190	85
Temperatura	18	16	14	19	11	20	17	12	9	13

Determinare con il metodo dei minimi quadratica la retta di regressione relativa alla quantità di precipitazione (Y) in funzione della temperatura media (X).
Commentare il valore del coefficiente di regressione ottenuto.

Fin'ora abbiamo illustrato alcuni indici in grado di misurare la relazione statistica esistente tra due caratteri. Quando si utilizzano due o più caratteri quantitativi si può cercare di individuare una funzione che descriva in modo dettagliato la relazione che emerge tra i dati. Se una delle variabili è considerata dipendente dall'altra si utilizzerà un modello di regressione. Tale modello può avere diversi scopi: descrittivo, interpretativo e previsivo. La sua importanza all'interno della teoria statistica deriva dalla sua semplicità unità a una formalizzazione rigorosa che consentono di ricercare, a partire da dati osservati e da assunzioni in larga misura verificabili, una relazione statistica tra la variabile dipendente e le altre variabili indipendenti (chiamate anche variabili esplicative).
Una relazione statistica tra una variabile indipendente X e una variabile dipende Y può essere descritta dall'equazione:

$$Y = \alpha + \beta X + \varepsilon$$

in cui $\alpha+\beta X$ definisce il contributo della variabile esplicativa al valore della variabile dipendente Y mentre ε rappresenta il contributo di tutti gli altri fattori in grado di influenzare la risposta (variabile dipendente Y) .
Il modello di regressione lineare si dice semplice quando si considera una sola variabile esplicativa (o indipendente).
Per ciascuna osservazione si ha

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

dove α e β corrispondono all'intercetta e al coefficiente angolare di una retta sul piano e sono chiamati coefficienti di regressione.
Occorre a questo punto definire un metodo di stima dei coefficienti di regressione. In altri termini, occorre individuare una retta che per ogni x_i restituisca un valore di Y_i che sia più vicino possibile ai valori osservati y_i.
Il metodo dei minimi quadrati consiste nel ricercare le stime di α e β mediante a e b che rendono minima la somma dei quadrati dei residui e_i (differenza tra il valore osservato y_i e il valore fornito dalla retta di regressione $\hat{y}$).
Le stime dei MINIMI QUADRATI dei coefficienti di regressione sono date da:

$$b = \frac{\sum_{i=1}^{n} (x_i - \mu_x) \cdot (y_i - \mu_y)}{\sum_{i=1}^{n} (x_i - \mu_x)^2} = \frac{cod(x,y)}{dev(x)} = \frac{cov(x,y)}{var(x)}$$

$$a = \mu_y - b \cdot \mu_x$$

$$\hat{y}_i = a + b \cdot x_i$$

Y	X	x*y	$(x-\mu_x)$	$(x-\mu_x)^2$
29	18	522	3,1	9,61
35	16	560	1,1	1,21
87	14	1218	-0,9	0,81
32	19	608	4,1	16,81
112	11	1232	-3,9	15,21
14	20	280	5,1	26,01
26	17	442	2,1	4,41
120	12	1440	-2,9	8,41
190	9	1710	-5,9	34,81
85	13	1105	-1,9	3,61
				Tot 120,9

Calcolo della media di x:

$$\mu_x = \frac{18 + 16 + \ldots\ldots\ldots + 9 + 13}{10} = 14,9$$

Calcolo della media di y:

$$\mu_y = \frac{29 + 35 + \ldots\ldots + 190 + 85}{10} = 73$$

Calcolo della covarianza tra x e y:

$$cov(x,y) = 911,7 - (14,9 \cdot 73) = -176$$

Calcolo della varianza di x:

$$var(x) = \frac{120,9}{10} = 12,09$$

Stima della retta di regressione: calcolo di a e b.

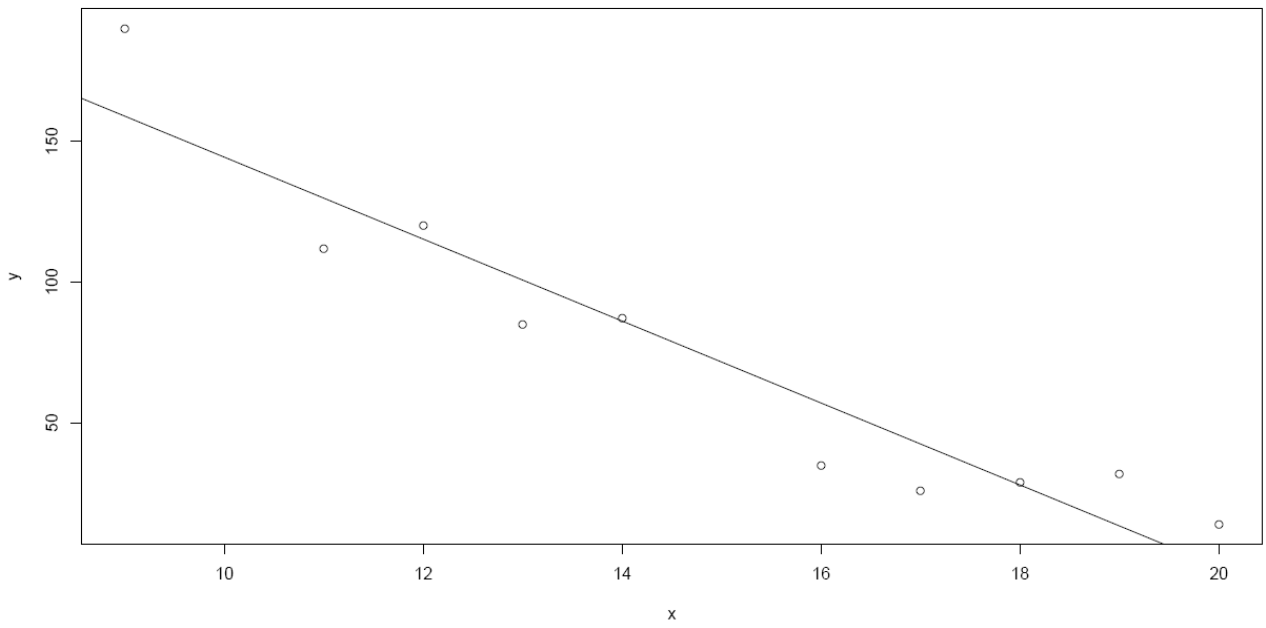
$$b = \frac{cov(x,y)}{var(x)} = -\frac{176}{12,09} = -14,56$$

$$a = \mu_y - b \cdot \mu_x = 73 - (14,56 \cdot 14,9) = 289,91$$

$$\hat{y}_i = a + b \cdot x_i$$

$$\hat{y}_i = 289,91 - 14,56 x_i$$

All'aumentare di una unità della temperatura media (variabile X) corrisponde una riduzione della quantità di precipitazione (Y) pari a 14,56.


y
150
100
50
10
12
14
16
18
20
x