CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 5

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Misura dell'associazione tra due caratteri

Uno store manager è interessato a studiare la relazione tra il numero di volte che viene trasmesso un annuncio pubblicitario durante il week end (X) e l'ammontare ricavato (in dollari) a fronte delle vendite realizzate nella settimana successiva (Y).A tale proposito, decide di effettuare un'indagine statistica e raccoglie le seguenti informazioni:

Settimana	X	Y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46
Totale	30	510

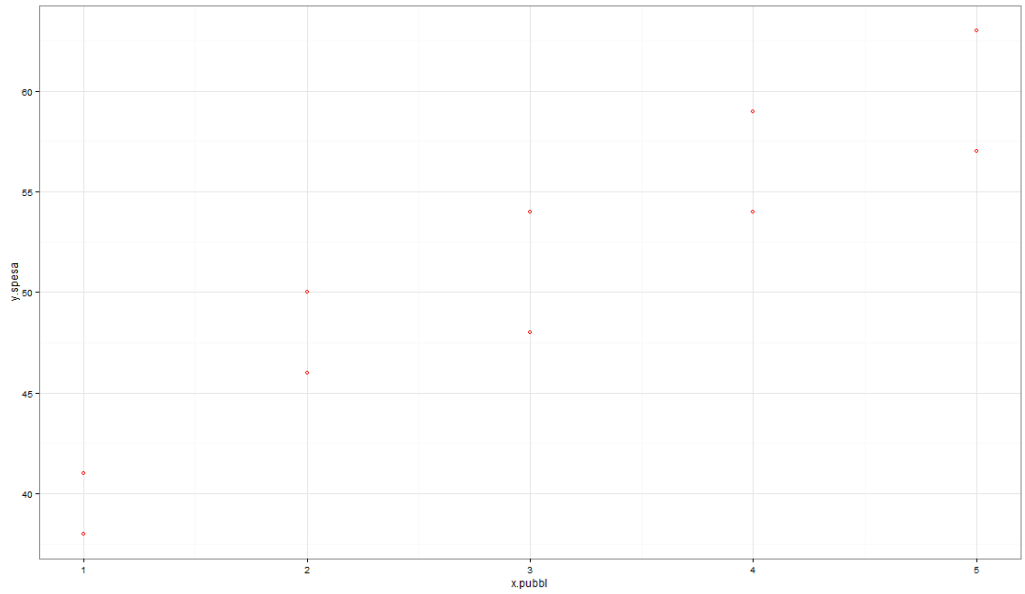
Valutare il legame di associazione e l'eventuale relazione lineare tra i due caratteri.

Sol.

Nello studio della relazione tra due variabili quantitative, risulta di grande utilità la rappresentazione dei valori di X e di Y in una diagramma a dispersione (o scatterplot). Questo grafico mette in evidenza, con una buona approssimazione, il tipo di legame tra Y e X.

Dal grafico risulta che esiste una associazione (interdipendenza) lineare tra X e Y.

Scatterplot di Y e X



.
Una misura del grado di associazione tra le variabili in esame è rappresentata dalla covarianza:

$$cov(X,Y) = \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Il numeratore della covarianza è detto codevianza. Dal segno del numeratore è possibile individuare se tra i caratteri X e Y esiste *concordanza* ossia se a valori crescenti di X corrispondono valori crescenti di Y ($\sigma_{xy} > 0$), *discordanza* ossia se a valori crescenti di X corrispondono valori decrescenti di Y ($\sigma_{xy} < 0$) oppure se gli scostamenti positivi e negativi di X e Y dalle rispettive medie si compensano ($\sigma_{xy} = 0$) per cui non esiste associazione tra i caratteri.

Formula alternativa della covarianza:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} xy - \mu_x\mu_y = \mu_{xy} - \mu_x\mu_y$$

Settimana	X	Y	XY	X^2	Y^2
1	2	50	100	4	2500
2	5	57	285	25	3249
3	1	41	41	1	1681
4	3	54	162	9	2916
5	4	54	216	16	2916
6	1	38	38	1	1444
7	5	63	315	25	3969
8	3	48	144	9	2304
9	4	59	236	16	3481
10	2	46	92	4	2116
Totale	30	510	1629	110	26576

$\mu_x = 3$

$\mu_y = 51$

$\sigma_{xy} = \frac{1}{10}(1629) - 3(51) = 9.9$

La covarianza è un indice assoluto di correlazione la cui unità di misura è pari al prodotto delle unità di misura in cui sono espressi i caratteri osservati. Pertanto, essa non può essere utilizzata per confrontare l'interdipendenza tra coppie di caratteri espressi in unità di misura differenti.

Il coefficiente di correlazione di Bravais-Pearson misura l'intensità del legame lineare che sussiste tra due variabili X e Y:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

ρ_{xy} è un indice che varia tra –1 e 1 e non dipende dall'unità di misura dei due caratteri.

• Se tra due caratteri non vi è correlazione lineare, si ha che $\rho_{xy} = 0$

• Se tra i due caratteri sussiste un legame lineare perfetto (Y= a$\pm$ bX) allora $\rho_{xy} = \pm 1$

Sapendo che:

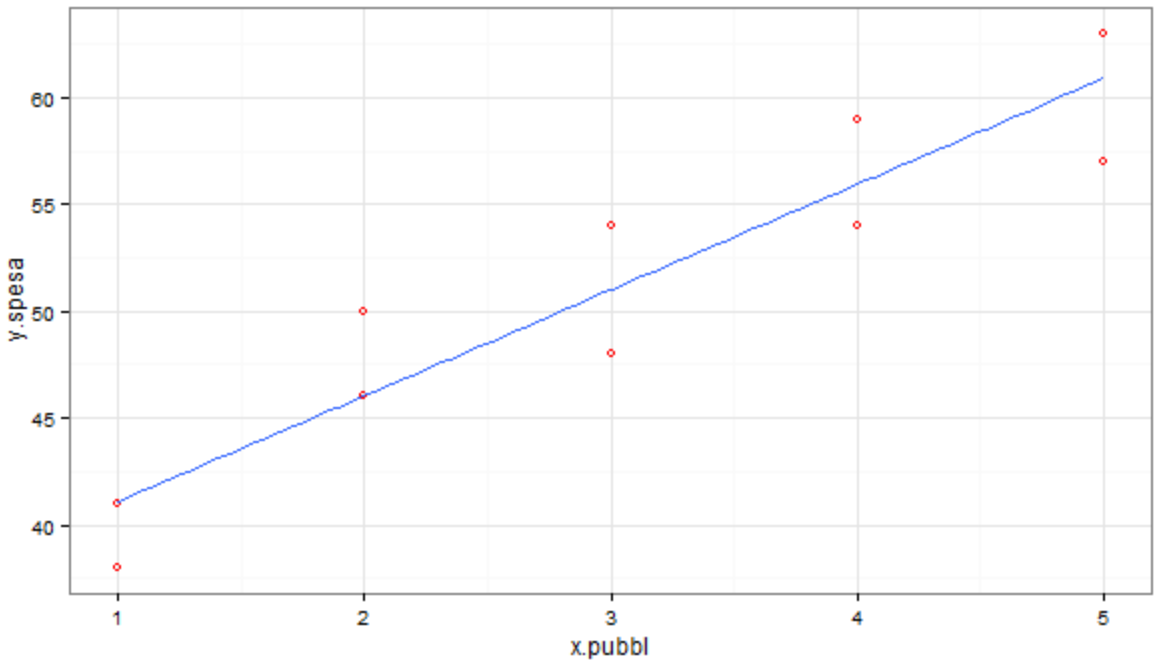
$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu_x^2 = \frac{1}{10}(110) - 9 = 2$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \mu_y^2 = \frac{1}{10}(26576) - 2601 = 56.6$$

allora $\sigma_x = \sqrt{2} = 1.41$ e $\sigma_y = \sqrt{56.6} = 7.52$

$$\rho_{xy} = \frac{9.9}{1.41 \times 7.52} = \frac{9.9}{10.6032} = 0.93$$

Il valore del coefficiente di correlazione è prossimo a 1, quindi indica un legame lineare molto forte tra i due caratteri. In particolare:



Esercizio 2. Covarianza e correlazione lineare per dati in tabella

Consideriamo il seguenti dati relativi alla distribuzione delle famiglie italiane del 2008 per numero di componenti e spesa mensile (in migliaia di euro):

X=spesa mensile|Y=nr.componenti	1	2	3	4	Totale (ni.)
(0,2]	1201	670	118	45	2034
(2,4]	773	1582	1018	550	3923
(4,6]	253	928	1160	791	3132
(6,8]	84	446	834	651	2015
Totale (n.j)	2311	3626	3130	2037	11104

Notazione:

n_{ij} frequenza congiunta i-esima riga (i=1,…,h) j-esima colonna (j=1,…,k).

$n_{i.} = \sum_{i=1}^{h} n_{ij}$ frequenza marginale di riga. Esprime il numero di soggetti che possiedono la modalità x_i indipendentemente da quello che avviene per il carattere Y

$n_{.j} = \sum_{j=1}^{k} n_{ij}$ frequenza marginale di colonna. Esprime il numero di soggetti che possiedono la modalità y_i a prescindere da quello che avviene per il carattere X.

$\mathbf{n_{..}} = N = 11104$

Vale inoltre la seguente relazione: $n_{..} = \sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{h} n_{.j} = \sum_{j=1}^{k} n_{i.}$

a) Costruire la distribuzione congiunta in frequenze relative dei caratteri X e Y
b) Costruire la distribuzione marginale di Y e X
c) Valutare, se esiste, una correlazione lineare tra i caratteri X e Y. Commentare i risultati ottenuti.

Sol.

a)Distribuzione congiunta di X e Y in frequenze relative

X=spesa mensile|Y=nr.componenti	1	2	3	4	Totale (ni.)
(0,2]	0.108	0.060	0.011	0.004	0.183
(2,4]	0.070	0.142	0.092	0.049	0.353
(4,6]	0.023	0.083	0.104	0.071	0.282
(6,8]	0.008	0.040	0.075	0.058	0.181
Totale (n.j)	0.208	0.326	0.282	0.183	1

b)

Distribuzione marginale di Y

Y=nr.componenti	n.j
1	2311
2	3626
3	3130
4	2037
	11104

$$\mu_y = \frac{1}{n}\sum_{j=1}^{k} y_j n_{.j} = \frac{27101}{11104} = 2.44$$

Distribuzione marginale di X

X=spesa mensile (valori centrali ci)	ni.
1	2034
3	3923
5	3132
7	2015
	11104

$$\mu_x = \frac{1}{n}\sum_{i=1}^{h} c_i n_{i.} = \frac{43568}{11104} = 3.92$$

d)Per calcolare il coefficiente di correlazione lineare dobbiamo calcolare la covarianza e le varianze di X e Y rispettivamente. Per comodità calcoliamo:

Matrice dei valori $x_i y_j$

X|Y	1	2	3	4
1	1	2	3	4
3	3	6	9	12
5	5	10	15	20
7	7	14	21	28

Matrice dei valori $x_i y_j n_{ij}$

X|Y	1	2	3	4
1	1201	1340	354	180
3	2319	9492	9162	6600
5	1265	9280	17400	15820
7	588	6244	17514	18228

$$\frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij} = \frac{1}{11104}(116987) = 10.54$$

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij} - \mu_x\mu_y = \mu_{xy} - \mu_x\mu_y = 10.54 - 3.92 \times 2.44 = 0.96$$

Calcoliamo quindi le varianze di X e Y:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{h} c_i^2 n_{i.} - \mu_x^2 = 3.911$$

$$\sigma_x = 1.978$$

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{k} y_j n_{.j} - \mu_y^2 = 1.030$$

$$\sigma_x = 1.015$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{0.96}{1.015 \times 1.978} = 0.48$$

Esercizio 3. La covarianza in caso di trasformazioni lineari

La tabella seguente contiene le distribuzioni dei caratteri X=spesa mensile per il tempo libero (in euro) sostenuta dalle famiglie e Y=peso (in kg) dei figli. L'indagine ha coinvolto un collettivo di 5 famiglie residenti a Milano con un solo figlio a carico in età scolare.

X=spesa	Y=Peso	XY
90	55	4950
110	72	7920
130	65	8450
170	62	10540
200	58	11600
Totale		43460

Un economista italiano è interessato a valutare, se esiste, un legame/associazione tra la spesa per il tempo libero sostenuta dalle famiglie e il peso dei figli in età scolare. A tale proposito calcola la covarianza come misura dell'interdipendenza tra i caratteri:

$\mu_x = 140$
$\mu_y = 62.4$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=}^{n} xy - \mu_x\mu_y = \frac{1}{5}(43460) - 140(62.4) = -44$$

Un economista americano è interessato ad effettuare lo stesso studio sullo stesso set di dati. Tuttavia, ritiene opportuno esprimere i valori della spesa in dollari anziché in euro. Il tasso di cambio di 1 euro è di 1.34 dollari. Inoltre, effettua una trasformazione di scala sulla distribuzione del peso registrato in chilogrammi nel corrispondente valore in libbre (unità di misura utilizzata negli USA, per cui 1kg=2.2 lbs).
Quale sarà il valore della covarianza tra i due caratteri calcolato dall'economista americano?

Sol.

Le variabili X e Y vengono opportunamente trasformate:

Z=1.34 X	W=2.2Y	ZW
120.6	121	14592.6
147.4	158.4	23348.16
174.2	143	24910.6
227.8	136.4	31071.92
268	127.6	34196.8
Totale		128120.08

$\mu_z = 1.34\,\mu_x = 187.6$

$\mu_w = 2.2\,\mu_y = 137.28$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=}^{n} xy - \mu_x\mu_y = \frac{1}{5}(128120.08) - 187.6(137.28) = -129.71$$

Nota: per la proprietà della covarianza in caso di trasformazioni lineari potevamo ottenere esattamente lo stesso valore sfruttando la seguente:

$$Cov(aX, bY) = \left[\frac{1}{n}\sum_{i=1}^{n} ax_i, cy_i\right] - a\mu_x c\mu_y =$$

$$= ac\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \mu_x\mu_y\right] = acCov(X,Y)$$

Quindi:

$$Cov(Z,W) = 1.34 \times 2.2 Cov(X,Y) = -129.71$$