

Corso di Statistica Facoltà di Economia

a.a. 2010-2011

francesco mola

Lezione n° 9

Contenuti

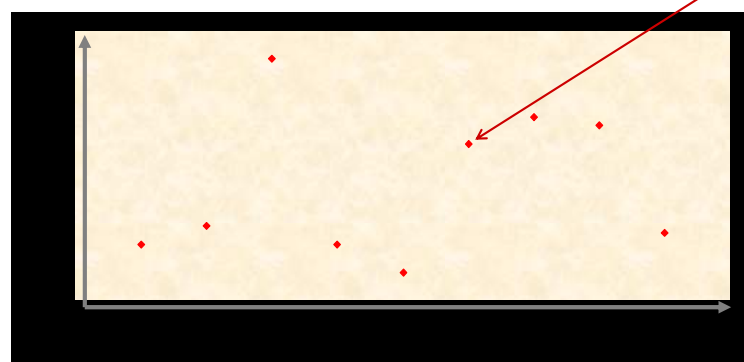
- Correlazione lineare
 - Scatterplot e analisi grafica
 - L'uso delle variabili standardizzate
 - La covarianza
 - Il coefficiente di correlazione lineare di Bravais-Pearson
- Formule alternative
 - Per dati grezzi
 - Per tabelle a doppia entrata con variabili numeriche (tabelle di correlazione)

Lez9_a.a. 2010-2011

statistica-francesco mola

2

In caso di variabili quantitative possiamo rappresentare graficamente la variabile doppia con
uno SCATTER PLOT (diagramma a dispersione, nube di punti, grafico XY)



Lez9_a.a. 2010-2011

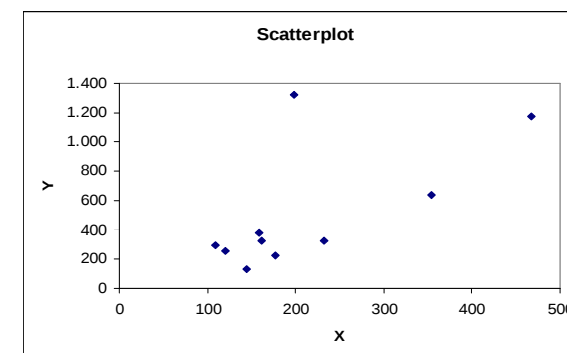
statistica-francesco mola

3

Scatterplot o diagramma a dispersione

- Analisi grafica della relazione tra due variabili numeriche

osserv.	(X)	(Y)
1	121	260
2	109	292
3	233	323
4	199	1.320
5	354	640
6	145	135
7	467	1.176
8	177	225
9	161	326
10	158	378



Lez9_a.a. 2010-2011

statistica-francesco mola

4

L'uso di variabili standardizzate

- Siano X e Y due variabili numeriche
(non necessariamente con la stessa scala di misura)
- La standardizzazione fa sì che abbiamo

$$Z(X) = \frac{X - \mu_X}{\sigma_X} \quad Z(Y) = \frac{Y - \mu_Y}{\sigma_Y}$$

Variabile standardizzata

$$X = x_1, x_2, x_3, \dots, x_n \Rightarrow Z = Z(X)$$

$$Z(X) = \frac{X - \mu}{\sigma}$$

X	Z = Z(X)
x_1	$\frac{x_1 - \mu}{\sigma} = z_1$
x_2	$\frac{x_2 - \mu}{\sigma} = z_2$
x_3	$\frac{x_3 - \mu}{\sigma} = z_3$

Proprietà



$$\mu = 0$$

$$\mu = \frac{1}{n} \sum z_i = \frac{1}{n} \left(\frac{x_i - \mu}{\sigma} \right) = \frac{1}{n} \frac{1}{\sigma} \sum (x_i - \mu) = \frac{0}{n\sigma} = 0$$



$$\sigma^2 = 1$$

$$\sigma^2 = \frac{1}{n} \sum (z_i - \mu_z)^2 = \frac{1}{n} \sum z_i^2 = \frac{1}{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 = \frac{1}{n} \frac{1}{\sigma^2} \sum (x_i - \mu)^2 = \frac{1}{\sigma^2} \sigma^2 = 1$$

Come si standardizzano i dati

(X)	(Y)
121	260
109	292
233	323
199	1.320
354	640
145	135
467	1.176
177	225
161	326
158	378

$$Z(X) = \frac{X - \mu_X}{\sigma_X}$$

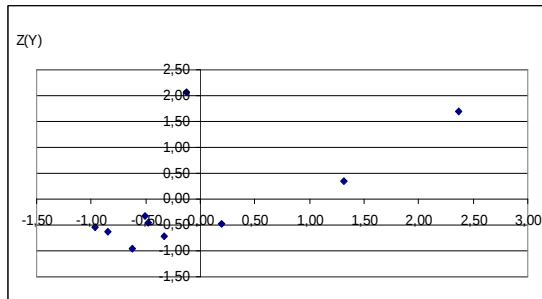
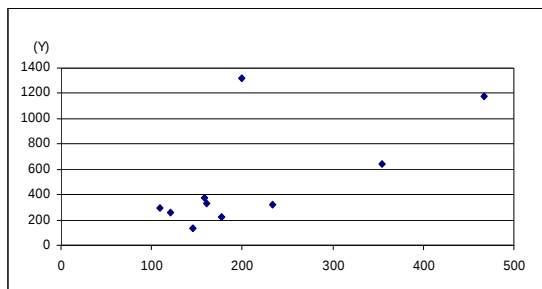
$$Z(Y) = \frac{Y - \mu_Y}{\sigma_Y}$$

Media X =	212,4
Media Y =	507,5
Varianza X =	11592
Varianza Y =	153462
Sqm X =	107,67
Sqm Y =	391,74

Z(X)	Z(Y)
-0,85	-0,68
-0,96	-0,55
0,19	-0,47
-0,12	2,07
1,32	0,34
-0,63	-0,95
2,36	1,71
-0,33	-0,72
-0,48	-0,46
-0,51	-0,33

$$\frac{225 - 507,5}{391,74}$$

$$\frac{121 - 212,4}{107,67}$$



Se rappresentiamo
graficamente sia le
distribuzioni originali
X e **Y**

che quelle
standardizzate
Z(X) e **Z(Y)**

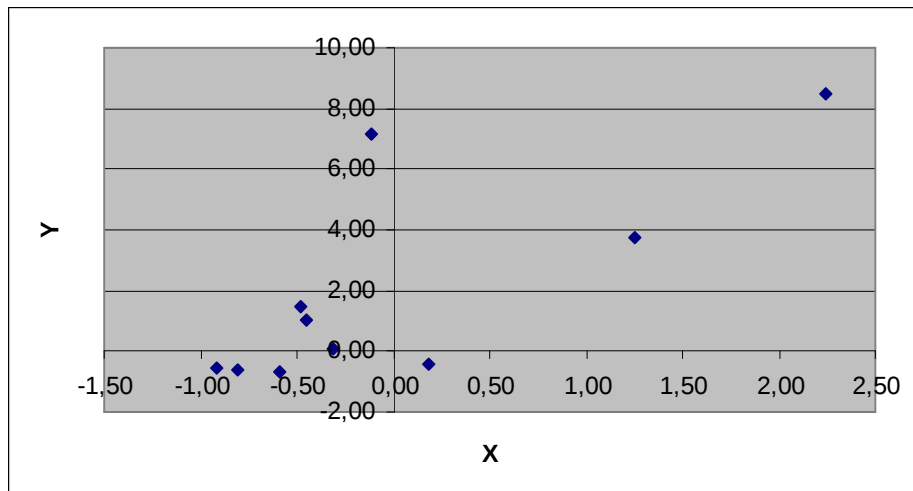
notiamo che la
relazione tra i dati
non muta ma cambia
Il **centro** di riferimento
(le medie di **Z(X)**
e **Z(Y)** sono pari a 0)

I dati

utile per i
calcoli
successivi

(X)	(Y)	Z_X	Z_Y	y^2
121	260	-0,85	-0,63	67.600
109	292	-0,96	-0,55	85.264
233	323	0,19	-0,47	104.329
199	1.320	-0,12	2,07	1.742.400
354	640	1,32	0,34	409.600
145	135	-0,63	-0,95	18.225
467	1.176	2,36	1,71	1.382.976
177	225	-0,33	-0,72	50.625
161	326	-0,48	-0,46	106.276
158	378	-0,51	-0,33	142.884

Analisi grafica



Il coefficiente di correlazione lineare

- Si definisce come

$$\rho_{XY} = \frac{1}{n} \sum_{i=1}^n z_{x_i} z_{y_i} = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \mu_X}{\sigma_X} \frac{y_i - \mu_Y}{\sigma_Y} =$$

$$= \frac{\frac{1}{n} \sum (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

covarianza

- Si dimostra che $-1 \leq \rho_{XY} \leq 1$

Il coefficiente di correlazione come espressione della devianza

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \longrightarrow \quad \rho_{XY} = \frac{Cod(X, Y)}{\sqrt{Dev(X)Dev(Y)}}$$

$$Cod(X, Y) = n\sigma_{XY}$$

$$Dev(X) = n\sigma_X^2$$

$$Dev(Y) = n\sigma_Y^2$$

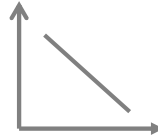
Formulazioni alternative

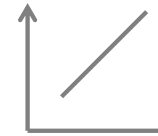
$$\begin{aligned} \rho_{XY} &= \frac{\sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_Y)^2}} = \\ &= \frac{\sum_{i=1}^n x_i y_i - n\mu_X \mu_Y}{\sqrt{\sum_{i=1}^n x_i^2 - n\mu_X^2} \sqrt{\sum_{i=1}^n y_i^2 - n\mu_Y^2}} = \frac{Cod(X, Y)}{\sqrt{Dev(X)Dev(Y)}} \end{aligned}$$

Proprietà di ρ

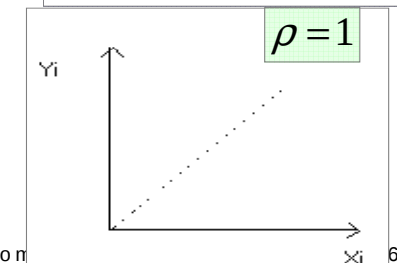
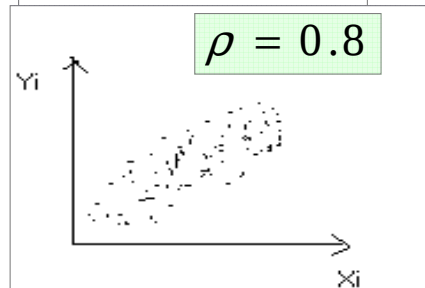
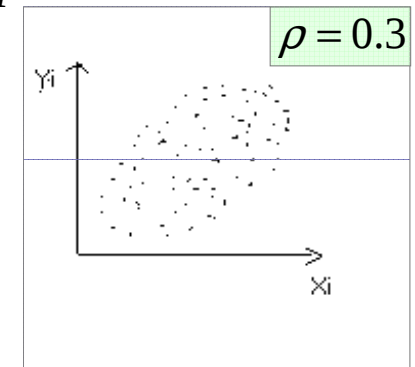
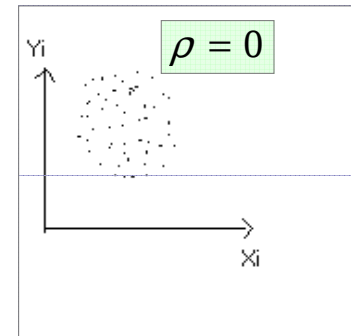
Il segno di ρ_{XY} dipende da σ_{XY}

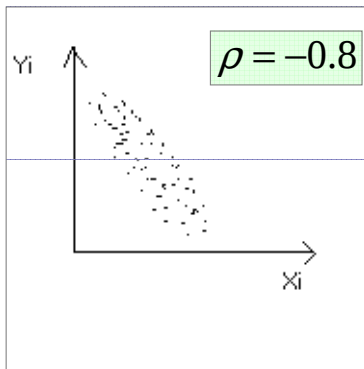
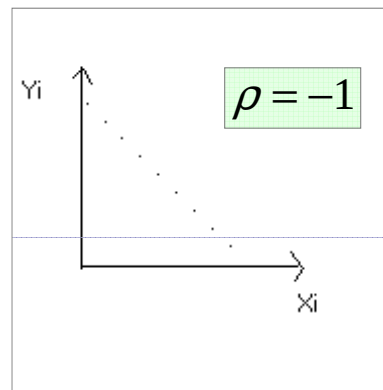
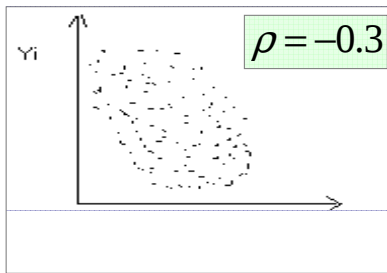
Se X e Y sono indipendenti allora: $\rho_{XY} = 0$

$$\rho_{XY} = -1 \Rightarrow Y = \alpha - \beta X$$


$$\rho_{XY} = 1 \Rightarrow Y = \alpha + \beta X$$


$$-1 \leq \rho_{XY} \leq 1$$





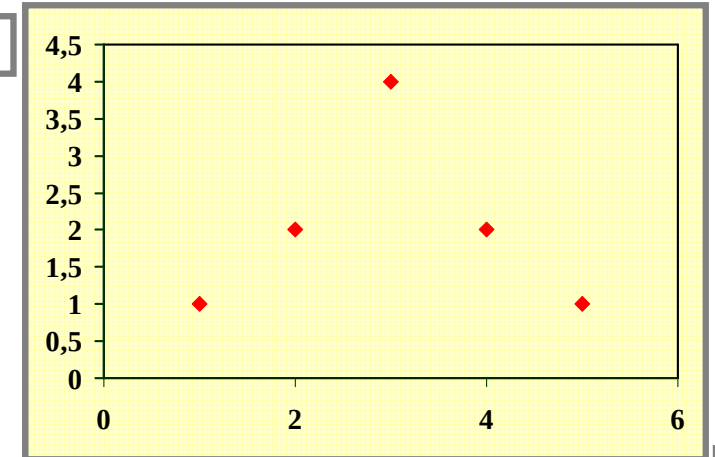
statistica-francesco mola

17

Esempio

X	Y
1	1
2	2
3	4
4	2
5	1

Yi



Xi

$$\mu_X = 3$$

$$\mu_Y = 2$$

Esiste una relazione funzionale

Lez9_a.a. 2010-2011

statistica-francesco mola

18

I calcoli

X	Y	$X - \mu_X$	$(X - \mu_X)^2$	$Y - \mu_Y$	$(Y - \mu_Y)^2$	$(X - \mu_X)(Y - \mu_Y)$
1	1	-2	4	-1	1	2
2	2	-1	1	0	0	0
3	4	0	0	2	4	0
4	2	1	1	0	0	0
5	1	2	4	-1	1	-2
			10		6	0

DEV(X)

statistica-francesco mola

DEV(Y) COD(X,Y)

19

Lez9_a.a. 2010-2011

Esempio cont.

I calcoli fanno sì che $\rho_{XY} = \frac{0}{\sqrt{10 \cdot 6}} = 0$

Le variabili sono **incorrelate** ma **non sono indipendenti**.

E' evidente dall'analisi grafica che esiste una **relazione funzionale** tra le variabili ma questa relazione **non è di tipo lineare**!



Il coefficiente di correlazione misura la **relazione lineare** tra variabili e non riesce ad intercettare altre relazioni!

Diciamo quindi che:

$$X, Y = ind \not\Rightarrow \rho_{XY} = 0$$

Lez9_a.a. 2010-2011

statistica-francesco mola

20

Per tabelle a doppia entrata

$$\rho_{XY} = \frac{\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h (x_i - \mu_X)(y_j - \mu_Y) n_{ij}}{\sqrt{\frac{1}{n} \sum_i (x_i - \mu_X)^2 n_{i\cdot}} \sqrt{\frac{1}{n} \sum_j (y_j - \mu_Y)^2 n_{\cdot j}}} =$$

$$= \frac{\frac{1}{n} \sum_i \sum_j x_i y_j n_{ij} - \mu_X \mu_Y}{\sqrt{\frac{1}{n} \sum_i x_i^2 n_{i\cdot} - \mu_X^2} \sqrt{\frac{1}{n} \sum_j y_j^2 n_{\cdot j} - \mu_Y^2}}$$

Esempio su dati grezzi

Calcolare la **covarianza** e la **correlazione** tra i caratteri "X" e "Y" per le 10 osservazioni.

osserv.	(X)	(Y)
1	121	260
2	109	292
3	233	323
4	199	1.320
5	354	640
6	145	135
7	467	1.176
8	177	225
9	161	326
10	158	378

$$\sigma_{xy} = \frac{1}{n} \sum x_i y_i - \mu_x \mu_y$$

$$\rho = \frac{\frac{1}{n} \sum x_i y_i - \mu_x \mu_y}{\sqrt{\frac{1}{n} \sum x_i^2 - \mu_x^2} \sqrt{\frac{1}{n} \sum y_i^2 - \mu_y^2}}$$

Esempio su dati grezzi

osserv.	(X)	(Y)	xy	x ²	y ²
1	121	260	31.460	14.641	67.600
2	109	292	31.828	11.881	85.264
3	233	323	75.259	54.289	104.329
4	199	1.320	262.680	39.601	1.742.400
5	354	640	226.560	125.316	409.600
6	145	135	19.575	21.025	18.225
7	467	1.176	549.192	218.089	1.382.976
8	177	225	39.825	31.329	50.625
9	161	326	52.486	25.921	106.276
10	158	378	59.724	24.964	142.884
Totale	2.124	5.075	1.348.589	567.056	4.110.179

Calcolo della covarianza

$$\mu_x = \frac{1}{n} \sum x_i = \frac{1}{10} 2.124 = 212,4$$

$$\mu_y = \frac{1}{n} \sum y_i = \frac{1}{10} 5.075 = 507,5$$

$$\mu_{xy} = \frac{1}{n} \sum x_i y_i = \frac{1}{10} 1.348.589 = 134.858,9$$

$$\sigma_{xy} = \mu_{xy} - \mu_x \mu_y = 134.858,9 - (212,4 \cdot 507,5) = 27.065,9$$

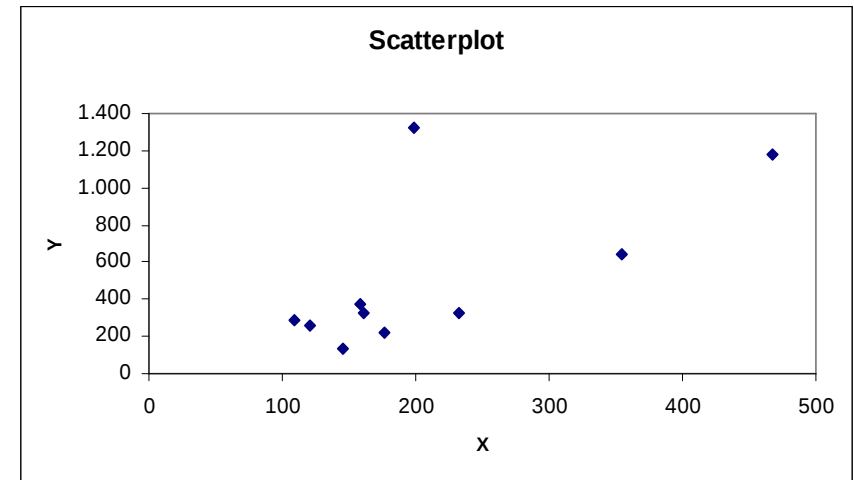
Calcolo della correlazione

$$\mu_{x^2} = \frac{1}{n} \sum x_i^2 = \frac{1}{10} 567.056 = 56.705,6$$

$$\mu_{y^2} = \frac{1}{n} \sum y_i^2 = \frac{1}{10} 4.110.179 = 411.017,9$$

$$\rho = \frac{\frac{1}{n} \sum x_i y_i - \mu_x \mu_y}{\sqrt{\left(\frac{1}{n} \sum x_i^2 - \mu_x^2\right) \left(\frac{1}{n} \sum y_i^2 - \mu_y^2\right)}} = \frac{27.065,9}{\sqrt{(56.705,6 - 212,4^2)(411.017,9 - 507,5^2)}} = 0,642$$

Scatterplot



Correlazione lineare in tabelle a doppia entrata

Data la seguente distribuzione doppia:

(X)	(Y)					Totale
	≤ 200	200- 300	300- 550	550- 950	>950	
≤ 200	5	9	7	2	2	25
200 - 300	1	1	1	2	0	5
300 - 400	1	2	1	3	0	7
400 - 500	0	0	0	2	3	5
>500	0	1	0	2	5	8
Totale	7	13	9	11	10	50

Calcolare la covarianza e la correlazione.

Correlazione lineare in tabelle a doppia entrata

$$\rho = \frac{\frac{1}{n} \sum_i \sum_j x_i y_j n_{ij} - \mu_x \mu_y}{\sqrt{\left(\frac{1}{n} \sum_i x_i^2 n_{i\cdot} - \mu_x^2\right) \left(\frac{1}{n} \sum_j y_j^2 n_{\cdot j} - \mu_y^2\right)}}$$

Per il calcolo della covarianza si costruisce la tabella degli elementi $x_i y_j n_{ij}$

	100	250	425	750	3.608
100	50.000	225.000	297.500	150.000	721.600
250	25.000	62.500	106.250	375.000	0
350	35.000	175.000	148.750	787.500	0
450	0	0	0	675.000	4.870.800
1.256	0	314.000	0	1.884.000	22.658.240

La somma degli elementi all'interno di tale tabella è pari a 33.561.140

Correlazione lineare in tabelle a doppia entrata

Per il calcolo delle medie e delle varianze è utile considerare la seguente tabella:

X	n	Y	n	$x_i n_{i\bullet}$	$y_j n_{\bullet j}$	x_i^2	$x_i^2 n_{i\bullet}$	y_j^2	$y_j^2 n_{\bullet j}$
100	25	100	7	2.500	700	10.000	250.000	10.000	70.000
250	5	250	13	1.250	3.250	62.500	312.500	62.500	812.500
350	7	425	9	2.450	3.825	122.500	857.500	180.625	1.625.625
450	5	750	11	2.250	8.250	202.500	1.012.500	562.500	6.187.500
1.256	8	3.608	10	10.048	36.080	1.577.536	12.620.288	13.017.664	130.176.640
				18.498	52.105		15.052.788		138.872.265

Calcolo della covarianza

$$\mu_X = \frac{1}{n} \sum_i x_i n_{i\bullet} = \frac{1}{50} 18.498 = 369,96$$

$$\mu_Y = \frac{1}{n} \sum_j y_j n_{\bullet j} = \frac{1}{50} 52.105 = 1.042,1$$

$$\sigma_{XY} = \frac{1}{50} 33.561.140 - 369,96 \times 1.042,1 =$$

$$= 671.223 - 385.536 = 285.687$$

Calcolo della correlazione

$$\sigma_X^2 = \frac{1}{n} \sum_i x_i^2 n_{i\bullet} - \mu_X^2 = \frac{1}{50} 15.052.788 - 369,96^2$$

$$\sigma_Y^2 = \frac{1}{n} \sum_j y_j^2 n_{\bullet j} - \mu_Y^2 = \frac{1}{50} 138.872.265 - 1.042,1^2$$

$$\rho_{XY} = \frac{\frac{1}{n} \sum_i \sum_j x_i y_j n_{ij} - \mu_X \mu_Y}{\sqrt{\frac{1}{n} \sum_i x_i^2 n_{i\bullet} - \mu_X^2} \sqrt{\frac{1}{n} \sum_j y_j^2 n_{\bullet j} - \mu_Y^2}} =$$

$$= \frac{285.687}{\sqrt{(301.056 - 369,96^2)} \sqrt{(2.777.445 - 1.042,1^2)}} = 0,542$$