Statistica: principi e metodi

Capitolo 11

Analisi delle distribuzioni doppie: correlazione

Correlazione

Data una **distribuzione doppia in forma disaggregata**, si dice che tra le due variabili X e Y

- vi è **correlazione positiva** o concordanza quando esse tendono a crescere (decrescere) insieme
- vi è **correlazione negativa** o discordanza quando al crescere di una variabile l'altra tende a decrescere.

Definizione del coefficiente di correlazione lineare di Bravais

In una **distribuzione doppia disaggregata**, il coefficiente di correlazione lineare di Bravais è

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}.\frac{y_i - \mu_y}{\sigma_y}\right).$$

Scarti standardizzati della X e della Y

$$z_{x_i} = \frac{x_i - \mu_X}{\sigma_X}; \quad z_{y_i} = \frac{y_i - \mu_y}{\sigma_y}$$

Si può ottenere anche come

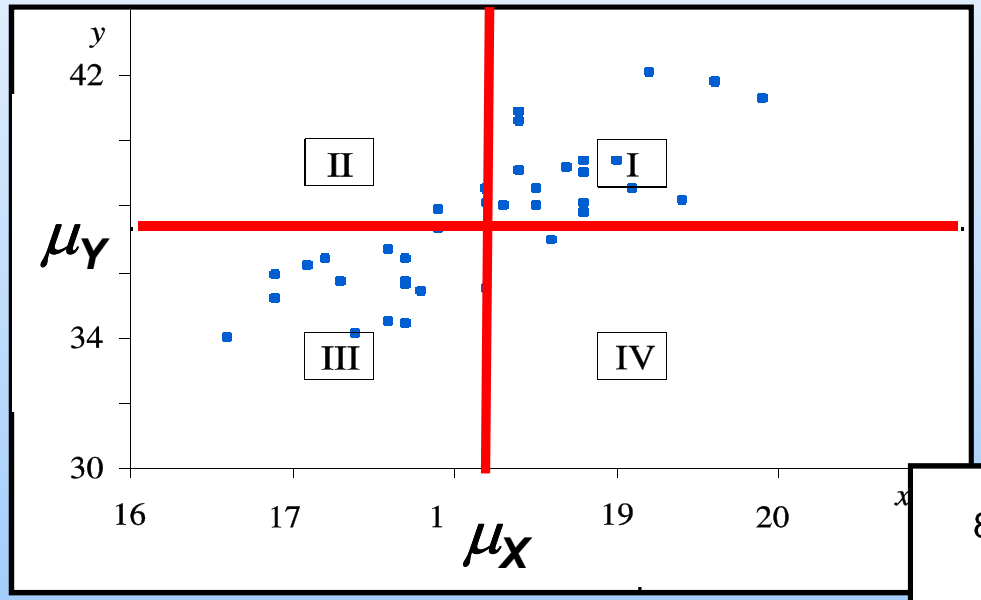
$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}} = \frac{C_{xy}}{\sqrt{D_x D_y}}$$

r coincide con la **radice quadrata dell'indice di determinazione** $r = \pm\sqrt{r^2}$

dove il segno è determinato dalla codevianza $C_{xy} = \sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)$

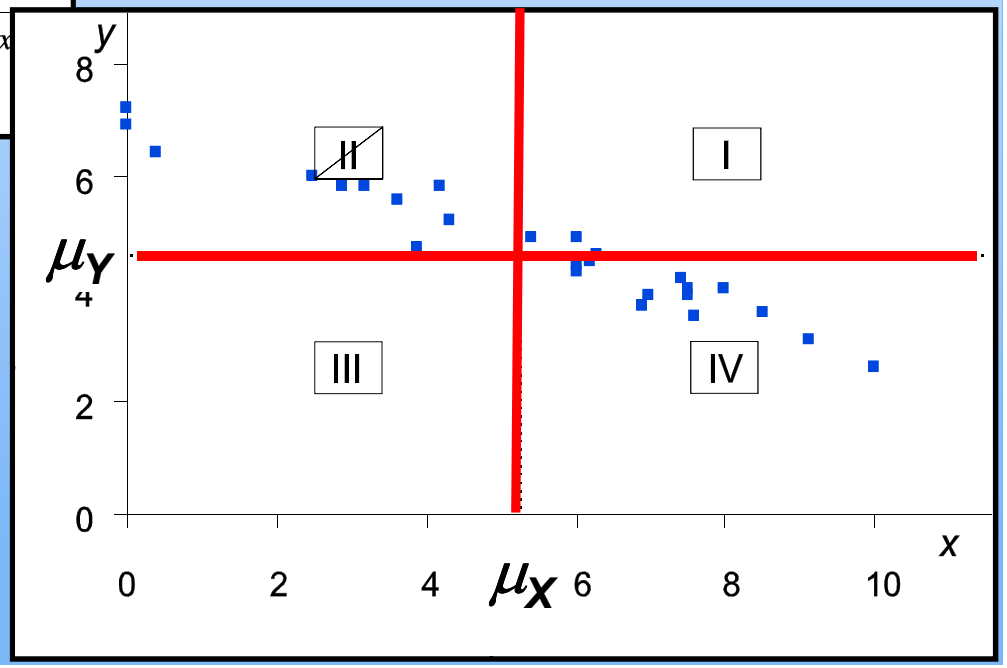
Interpretazione geometrica della formula

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right).$$



Correlazione positiva: i punti osservati sono collocati in prevalenza nel I e nel III quadrante dei nuovi assi cartesiani aventi origine nel punto (μ_X, μ_Y).

Correlazione negativa: i punti osservati sono collocati in prevalenza nel secondo e nel quarto quadrante dei nuovi assi cartesiani aventi origine nel punto (μ_X, μ_Y).



Interpretazione geometrica della formula

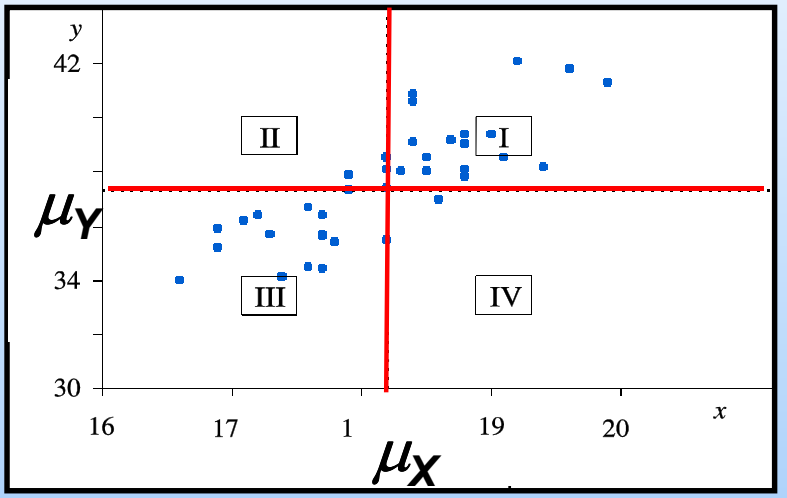
$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right).$$

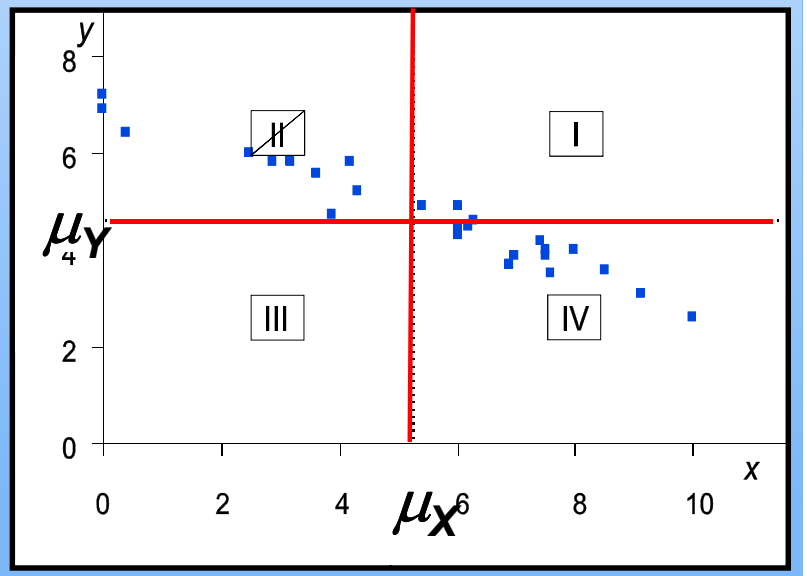
Ne segue che i prodotti

sono in prevalenza positivi nel primo caso e prevalentemente negativi nel secondo. Cosicché la quantità

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right),$$

media di tali prodotti, è positiva nel primo caso e negativa nel secondo.





Proprietà del coefficiente di correlazione lineare di Bravais

- Varia nell'intervallo [-1, 1].
  - È pari a **1 quando tutti i punti osservati si trovano su una retta con coefficiente angolare positivo.**
  - È uguale a **-1 quando tutti punti osservati si trovano su una retta con coefficiente angolare negativo.**

Il coefficiente di correlazione è la radice quadrata dell'indice di determinazione r^2 che assume valore minimo 0 (retta di regressione parallela all'asse delle ascisse) e valore massimo 1 (tutti i punti giacciono sulla retta).

Nel caso in cui tutti i punti giacciono su una retta inclinata positivamente, la codevianza sarà positiva e r=1; se tutti i punti giacciono su una retta inclinata negativamente, la codevianza sarà negativa e r=-1

Proprietà del coefficiente di correlazione lineare di Bravais

- È **positivo** quando la retta di regressione di *Y* su *X* ha coefficiente angolare positivo ($b_1>0$); è **negativo** nel caso opposto ($b_1<0$).
  - Il segno del coefficiente di regressione e del coefficiente di correlazione è determinato dal numeratore che rappresenta in entrambi i casi la codevianza

$$b_1 = \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_y)}{\sum_{i=1}^{N}(x_i-\mu_X)^2} = \frac{C_{xy}}{D_x}$$

$$r = \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_X)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = \frac{C_{xy}}{\sqrt{D_x D_y}}$$

Proprietà del coefficiente di correlazione lineare di Bravais

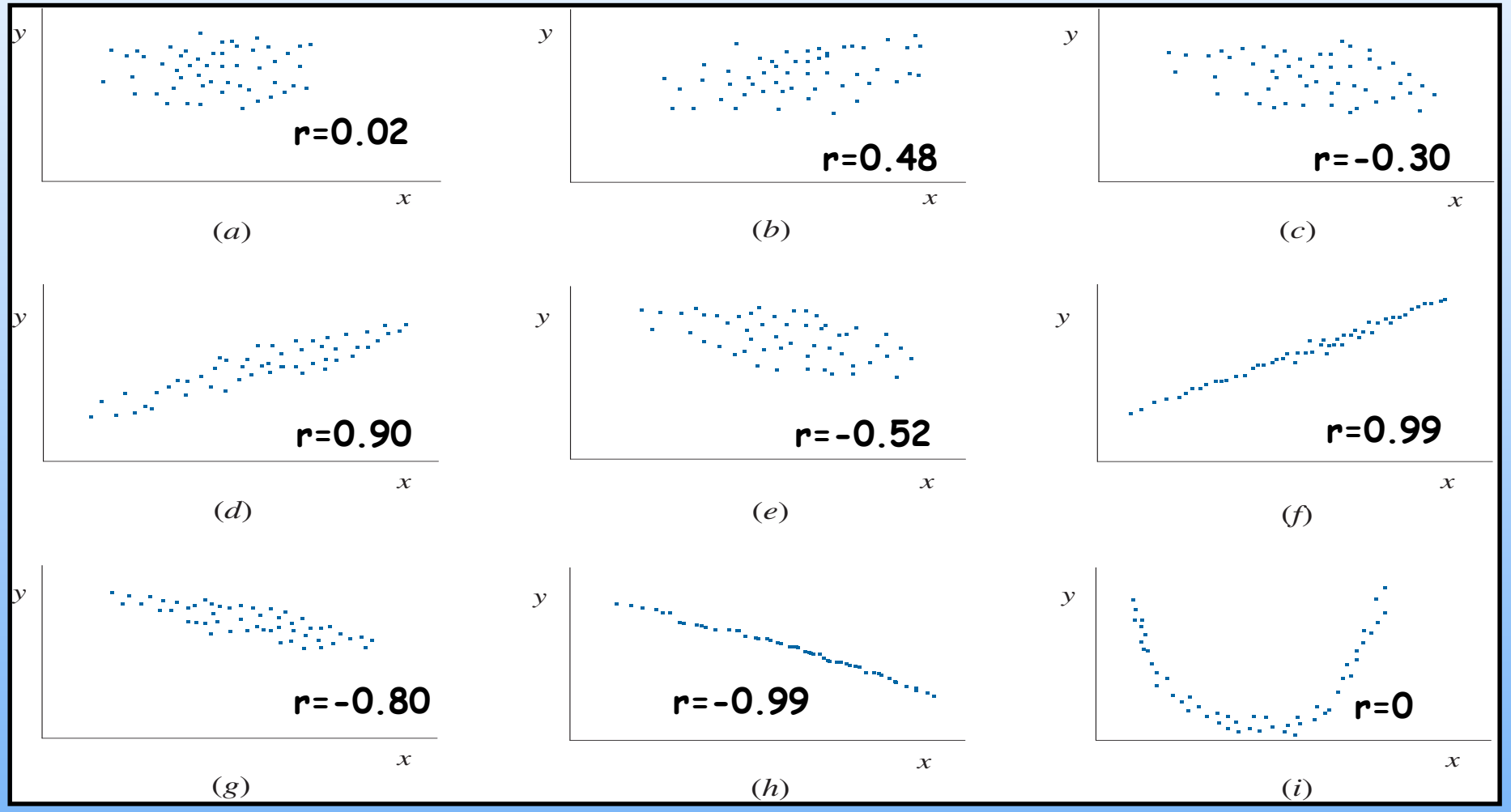
- **Non cambia** se le modalità della singola variabile vengono moltiplicate per una costante o aumentate (diminuite) di una costante positiva.
  - Se ai termini della distribuzione disaggregata $x_1, x_2, \ldots, x_N$, aggiungiamo una quantità costante a, $v_i=x_i+a$, si ha

$$r_{V,y} = \frac{\sum_{i=1}^{N}(v_i-\mu_V)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(v_i-\mu_V)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = \frac{\sum_{i=1}^{N}(x_i+a-\mu_X-a)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i+a-\mu_X-a)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = r_{X,y}$$

  - Se moltiplichiamo i termine della distribuzione disaggregata $x_1, x_2, \ldots, x_N$, per una quantità costante positiva non nulla b, $v_i=bx_i$, si ha

$$r_{V,y} = \frac{\sum_{i=1}^{N}(v_i-\mu_V)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(v_i-\mu_V)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = \frac{\sum_{i=1}^{N}(bx_i-b\mu_X)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(bx_i-b\mu_X)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = \frac{b\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_y)}{b\sqrt{\sum_{i=1}^{N}(x_i-\mu_X)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} = r_{X,y}$$

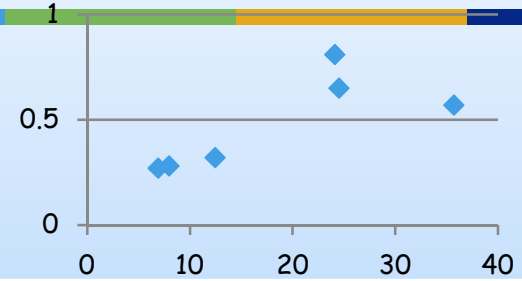
Diagrammi di dispersione e coefficienti di correlazione



Coefficiente di correlazione lineare di Bravais: calcolo

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

Tabella di calcolo dell'indice *r* per i dati che seguono (durezza e spessore di sei stoffe ritardanti di fiamma)



- indice di correlazione

Durezza (mg-cm) x_i	Spessore (mm) y_i	$x_i - \mu_X$	$y_i - \mu_y$	$(x_i - \mu_X)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_X) \cdot (y_i - \mu_y)$
7.98	0.28	-10.64	-0.20	113.17	0.04	2.16
24.52	0.65	5.90	0.17	34.83	0.03	0.98
12.47	0.32	-6.15	-0.16	37.80	0.03	1.00
6.92	0.27	-11.70	-0.21	136.85	0.05	2.50
24.11	0.81	5.49	0.33	30.16	0.11	1.79
35.71	0.57	17.09	0.09	292.13	0.01	1.48
			Totale	**644.94**	**0.26**	**9.92**

Le medie sono:

$$\mu_X = 18.62; \mu_Y = 0.48$$

$$r = \frac{9.92}{\sqrt{644.94 \cdot 0.26}} = 0.77$$

Il caso delle distribuzioni doppie di frequenze

Per quanto abbiamo visto riguardo all'indice di determinazione nel caso delle distribuzioni di frequenze nel capitolo 10, abbiamo

$$r = \frac{\sum_{i=1}^{s}\sum_{j=1}^{t}(x_i - \mu_X)(y_j - \mu_Y)n_{ij}}{\sqrt{\sum_{i=1}^{s}(x_i - \mu_X)^2 n_{i0}\sum_{j=1}^{t}(y_j - \mu_Y)^2 n_{0j}}}.$$

Quando uno o entrambi i caratteri sono divisi in intervalli, l'indice r si calcola prendendo i valori centrali di classe.

Il caso delle distribuzioni doppie di frequenze: calcolo di *r*

Distribuzione doppia di frequenze di un campione di coniugi classificati secondo l'età:

I numeri in rosso sono i valori centrali

Età marito	Età della moglie				Totale
	18-30 (24)	31-40 (35.5)	41-50 (45.5)	51-65 (58)	
20-30 (25.0)	14	0	0	0	14
31-40 (35.5)	5	23	0	0	28
41-50 (45.5)	0	5	17	1	23
51-65 (58.0)	0	0	9	26	35
Totale	19	28	26	27	100

μ_X	44.21
μ_Y	41.99
D_X	13984.55
D_Y	14569.49
C_{XY}	13153.46
r	0.92

- L'indice di correlazione è $r = \dfrac{13153.46}{\sqrt{13984.55 \cdot 14569.49}} = 0.92$