

Studio dell'associazione tra due variabili numeriche (correlazione)

	Task Description	Due Date	Status
☐	Studio del legame tra due variabili: differenza tra approccio simmetrico ed approccio asimmetrico		
☐	Studio dell'interdipendenza tra due variabili quantitative: concordanza-discordanza, uso degli scarti come segnali di presenza di concordanza-discordanza, la covarianza come media del prodotto degli scarti.		
☐	Contributo di un punto alla covarianza: interpretazione geometrica.		
☐	Casi di covarianza nulla: casi di assenza totale di relazione e casi di assenza di relazione lineare.		
☐	Casi di relazione lineare perfetta		
☐	Normalizzazione della covarianza in $[-1, +1]$: il coefficiente di correlazione lineare		
☐			
☐			
☐			

Studio
univariato

sintesi distribuzione

posizione

variabilità

forma

Studio
bivariato

studiare relazione (legame, associazione,
connessione) tra due variabili X e Y

approccio

tipologia di
variabili

simmetrico

asimmetrico

$X \longleftrightarrow Y$

$X \rightarrow Y$

(interdipendenze)

(dipendenze)

STUDIO DEL LEGAME TRA DUE VARIAB. NUMERICHE

↓
simmetrico

$X \longleftrightarrow Y$ giocano lo stesso ruolo (interdipendenza)

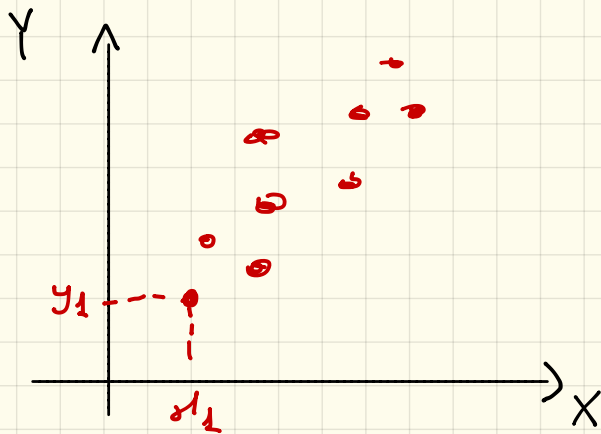
① comportamento delle variabili rispetto alle medie

↓

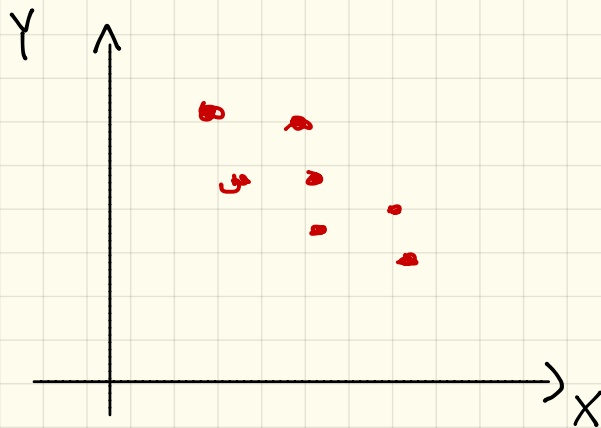
concorrenza: le due variabili X e Y tendono a muoversi

insieme $\left(\begin{array}{l} \text{al crescere di } X, \text{ cresce anche } Y \\ \text{al decrescere di } X, \text{ decresce anche } Y \end{array} \right)$

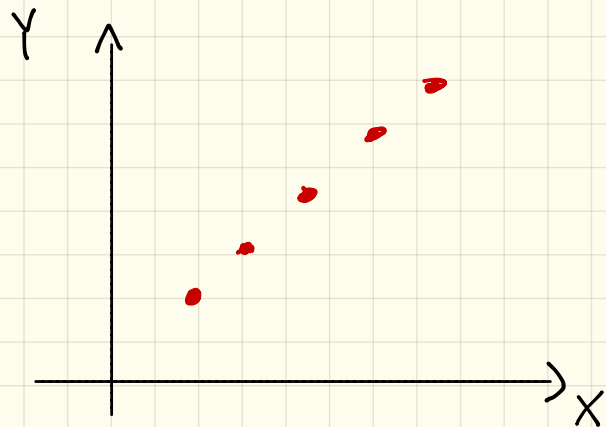
discorrenza: le due variabili X e Y si muovono in direzione opposta: al crescere di X , Y decresce
al decrescere di X , Y cresce



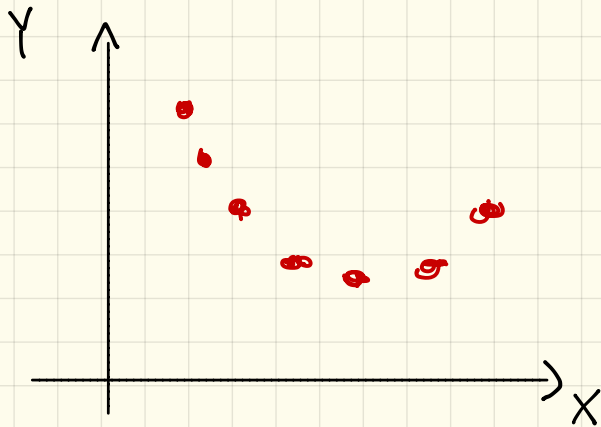
concorrenza
tra X e Y



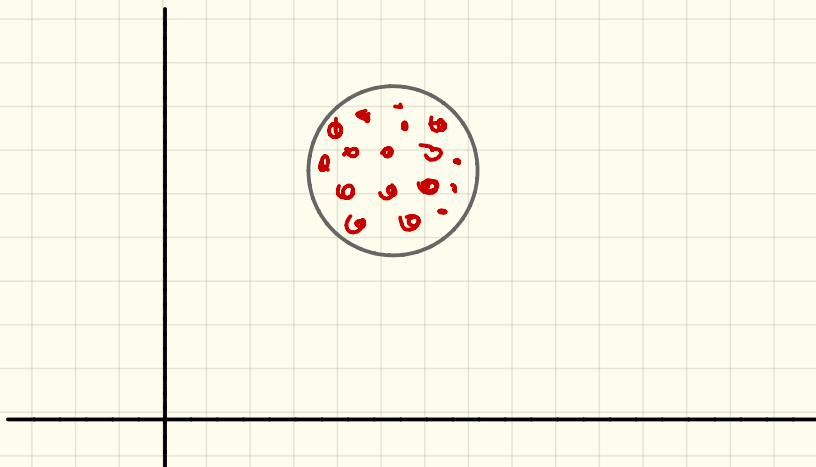
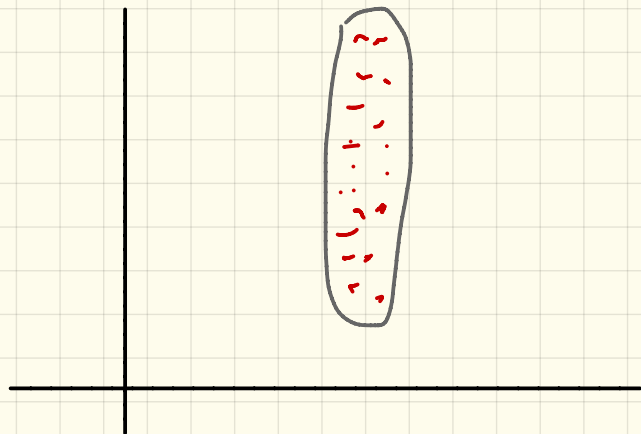
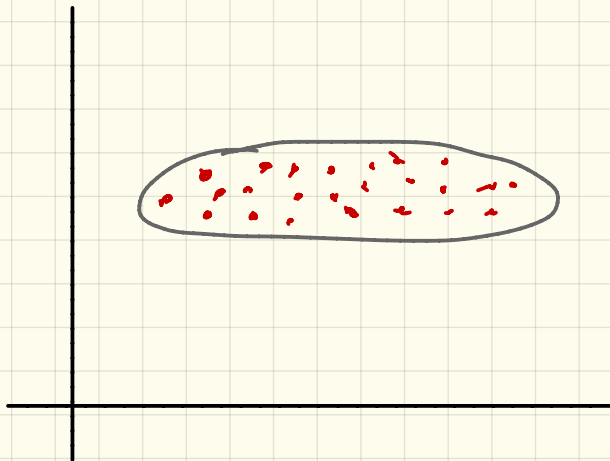
discorrenza
tra X e Y



legame "perfetto"
 } tra X e Y
 √
 lineare



legame "perfetto"
 ↓
 non lineare



Per studiare la presenza di concordanza / discordanza
si va a studiare come X e Y si comportano contemporaneamente
rispetto a μ_X e μ_Y

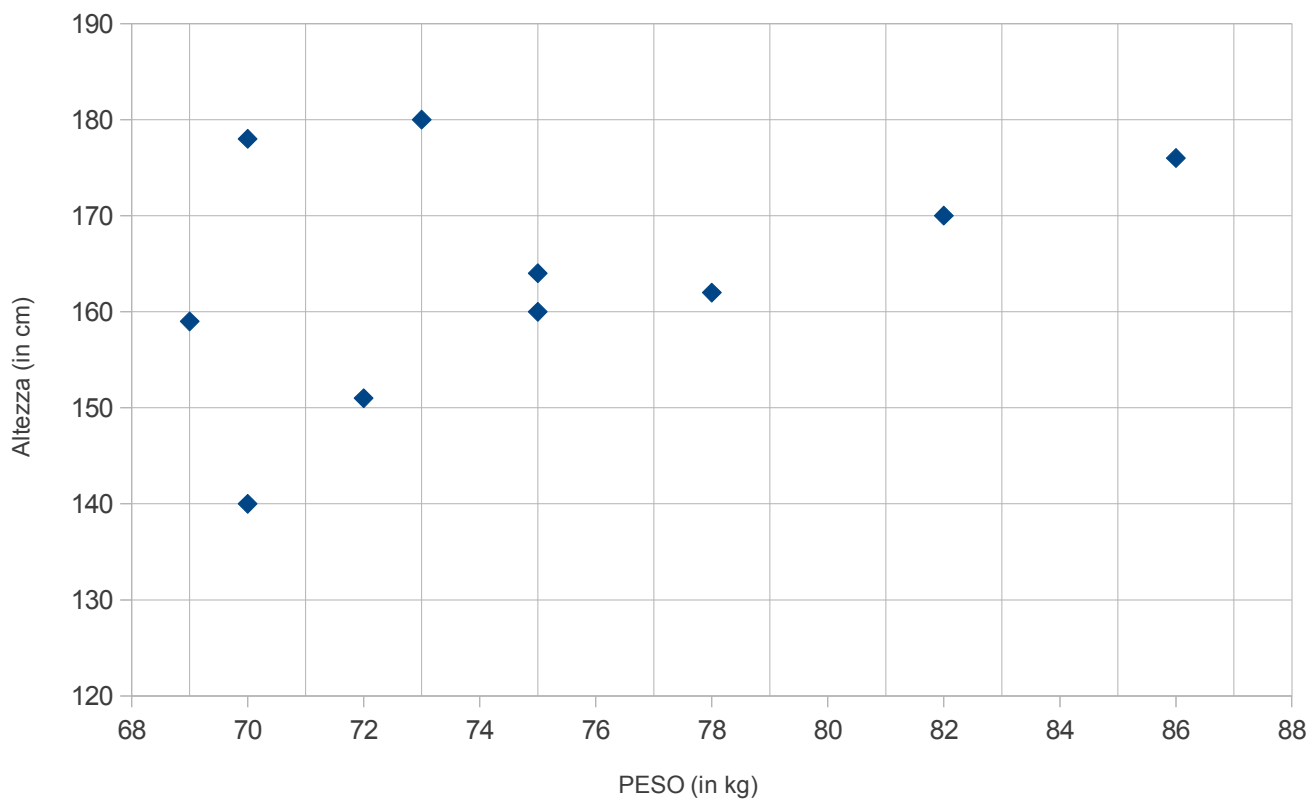


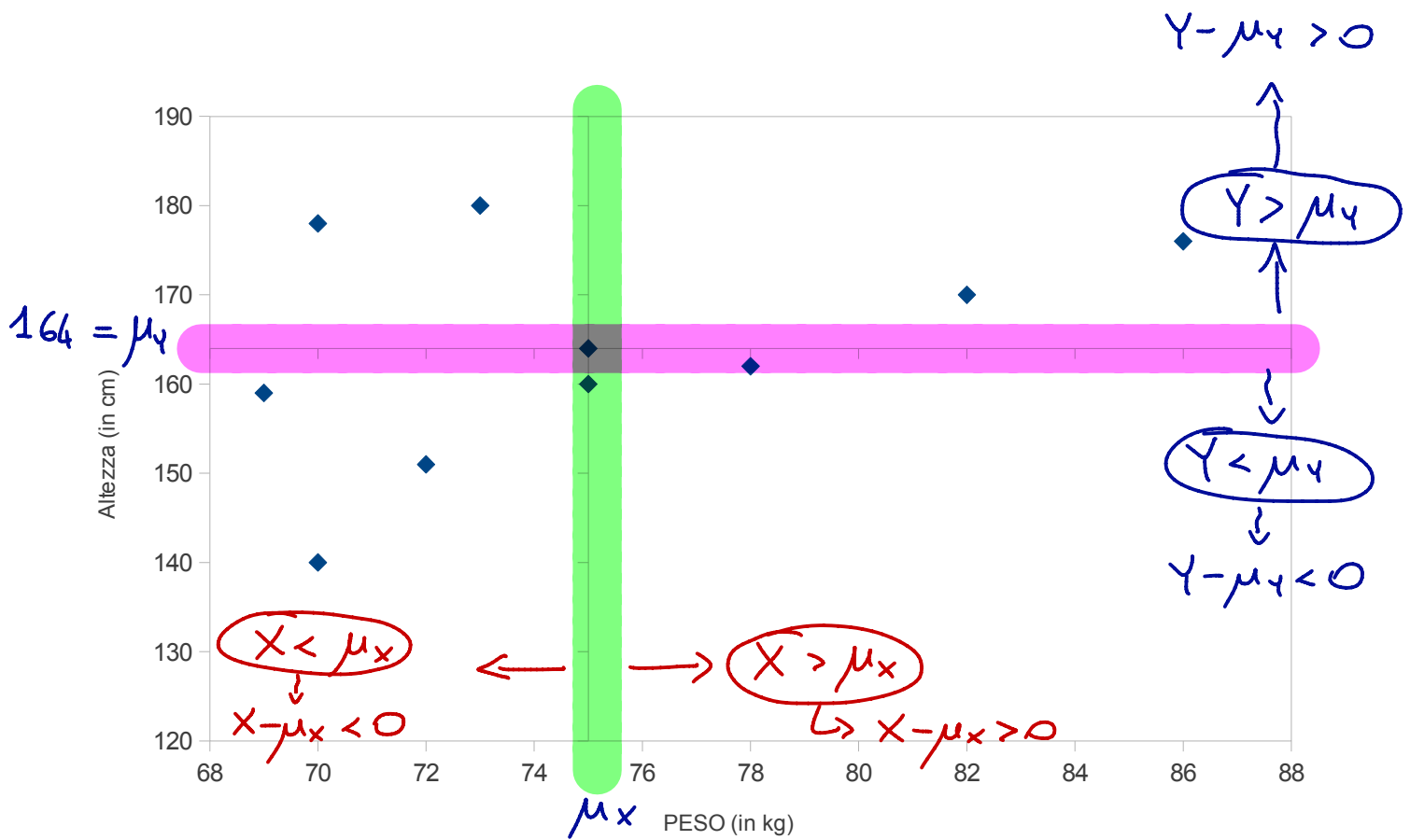
quando X assume valori maggiori di μ_X
cosa succede a Y ?

Un esempio numerico...

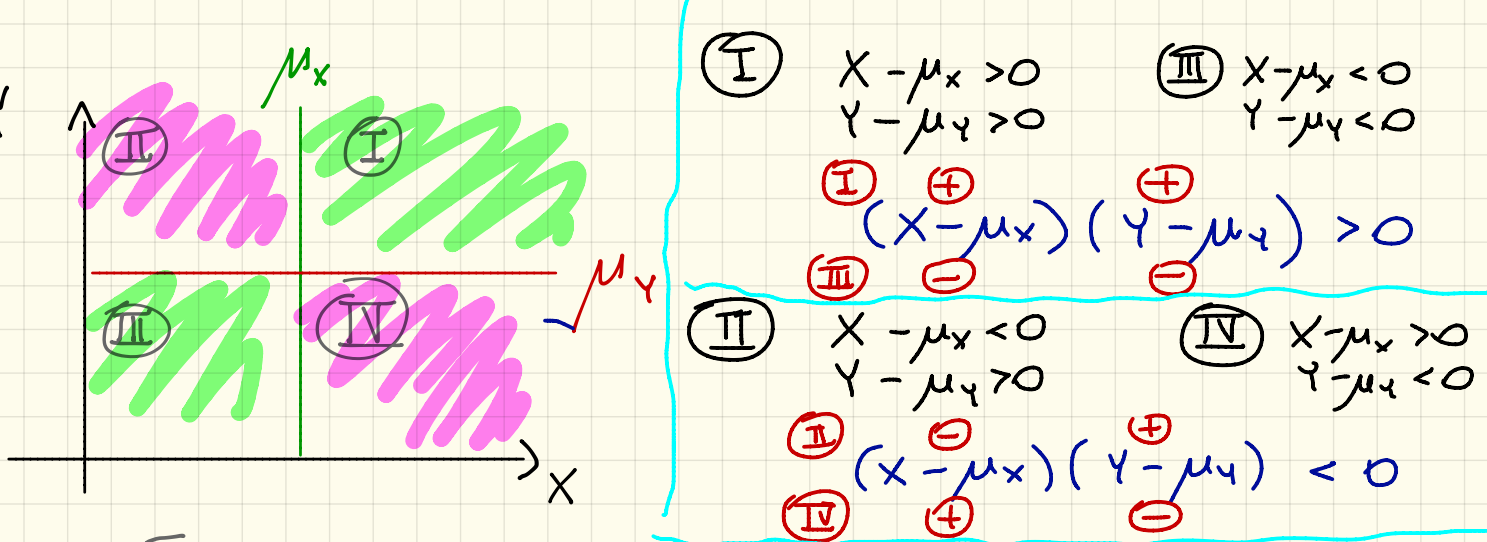
Peso (kg)	Altezza (cm)
70	140
72	151
69	159
75	160
78	162
75	164
82	170
86	176
70	178
73	180

... e grafico: il diagramma di dispersione





	Peso (kg)	Altezza (cm)	scarti X	scarti Y
	70	140	-5	-24
	72	151	-3	-13
	69	159	-6	-5
	75	160	0	-4
	78	162	3	-2
	75	164	0	0
	82	170	7	6
	86	176	11	12
	70	178	-5	14
	73	180	-2	16
TOTALE	750	1640		
MEDIA	75	164		



quadranti con segnali di concordanza tra X e Y

quadranti con segnali di discordanza tra X e Y

→ la covarianza cerca di sintetizzare quali segnali sono più "forti"

X	Y	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
x_1	y_1	$x_1 - \mu_x$	$y_1 - \mu_y$	$(x_1 - \mu_x)(y_1 - \mu_y)$
x_2	y_2	$x_2 - \mu_x$	$y_2 - \mu_y$	$(x_2 - \mu_x)(y_2 - \mu_y)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	y_m	$x_m - \mu_x$	$y_m - \mu_y$	$(x_m - \mu_x)(y_m - \mu_y)$

la somma di questa colonna
mi fornisce informazioni sulla concordanza / discordanza

→ media del prodotto
degli scarti

Covarianza

$$\text{cov}(X, Y) = \sigma_{xy} = \sum_{i=1}^n \frac{(x_i - \mu_x)(y_i - \mu_y)}{n} \begin{matrix} \leq \\ \geq \end{matrix} 0$$

$\sigma_{xy} < 0$ quando i segnali negativi (II e IV quadrante) prevalgono su quelli positivi

$\sigma_{xy} > 0$ quando i segnali positivi (I e III quadrante) prevalgono su quelli negativi

$\sigma_{xy} = 0$ quando i segnali di segno opposto si compensano

covarianza tra X e Y ← [media del prodotto degli scarti]

	Peso (kg)	Altezza (cm)	scarti X	scarti Y	Prodotto scarti
	70	140	-5	-24	120
	72	151	-3	-13	39
	69	159	-6	-5	30
	75	160	0	-4	-0
	78	162	3	-2	-6
	75	164	0	0	0
	82	170	7	6	42
	86	176	11	12	132
	70	178	-5	14	-70
	73	180	-2	16	-32
TOTALE	750	1640			255
MEDIA	75	164			25.5

$Cov(X, Y)$ oppure σ_{XY}

278	Dev(X)
1422	Dev(Y)
0.41	Correlazione

covarianza tra X e Y

Se prevalgono i segnali del I e III quadranti (concordi) avrà un valore positivo;
se invece sono più forti quelli discordi (II e IV) avrà un valore negativo

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

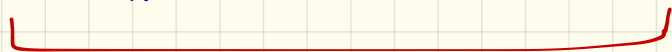
Caso di serie
bivariate
prezze

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^c (x_i - \mu_x)(y_j - \mu_y) n_{ij}$$

CASO TABELLA
A DOPPIA
ENTRATA

nel calcolare σ_{xy}
devo pesare per la frequenza
della coppia (x_i, y_j)

$$\sigma_{xy} = \frac{1}{n} \sum_i (x_i - \mu_x)(y_i - \mu_y)$$



sviluppando i prodotti si ottiene
una formula alternativa più agevole
per i calcoli

$$\sigma_{xy} = \frac{\sum_i x_i y_i}{n_i} - \mu_x \mu_y$$

media dei
prodotti

prodotto delle
due medie

$$\sigma_{xy} = \frac{\sum_i \sum_j x_i y_j n_{ij}}{n} - \mu_x \mu_y$$

Dimostrazione

$$\tilde{\sigma}_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n} =$$

$$= \frac{\sum (x_i y_i - x_i \mu_y - \mu_x y_i + \mu_x \mu_y)}{n} =$$

$$= \sum \frac{x_i y_i}{n} - \mu_y \overset{(\mu_x)}{\sum \frac{x_i}{n}} - \mu_x \overset{(\mu_y)}{\sum \frac{y_i}{n}} + \frac{n \mu_x \mu_y}{n} =$$

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

c.v.d.

media dei
prodotti

prodotto delle
medie

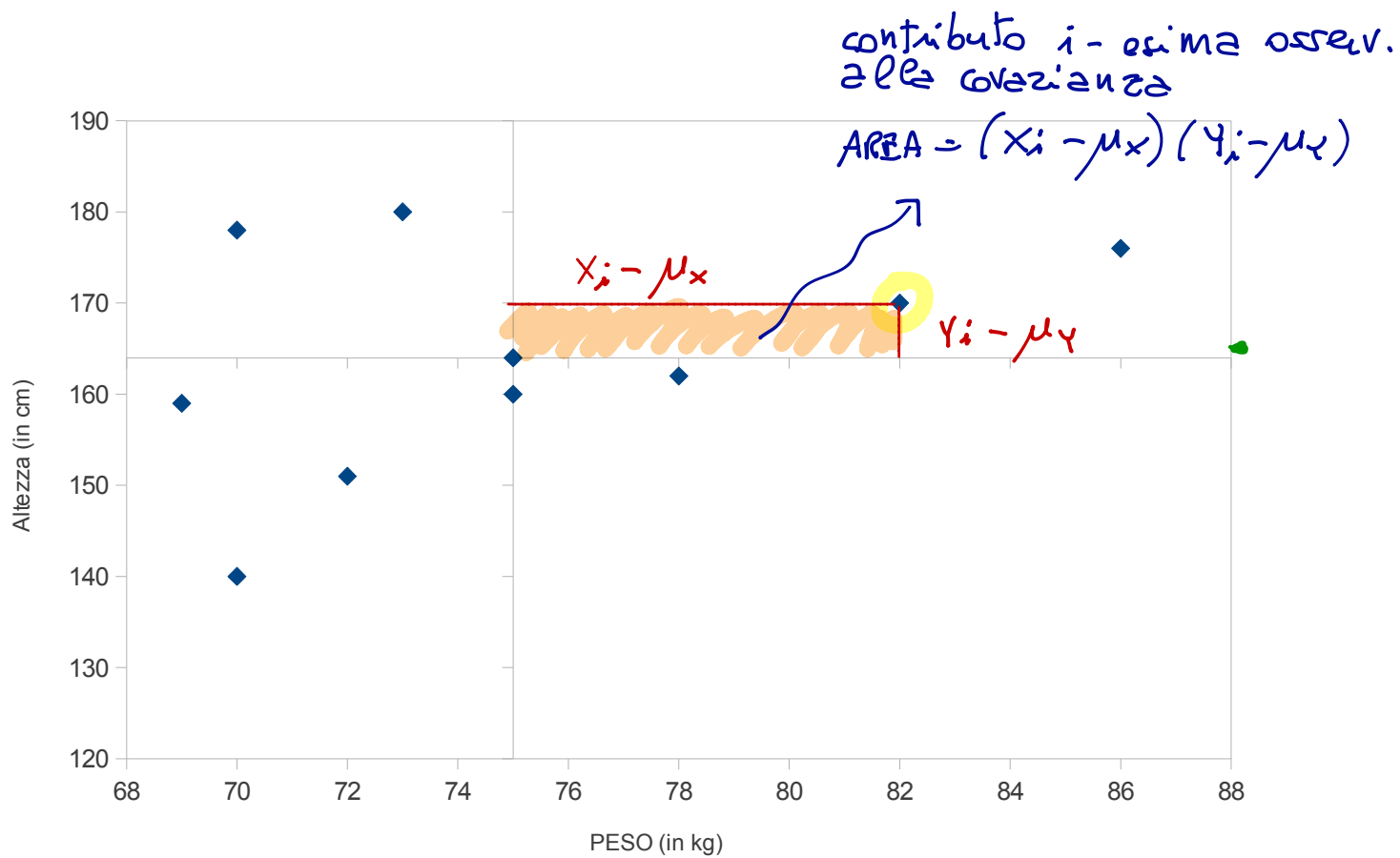
		Y	
		y _j	
X	x _i	m _{ij}	

$$\sigma_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^c (x_i - \mu_x)(y_j - \mu_y) m_{ij}}{n}$$

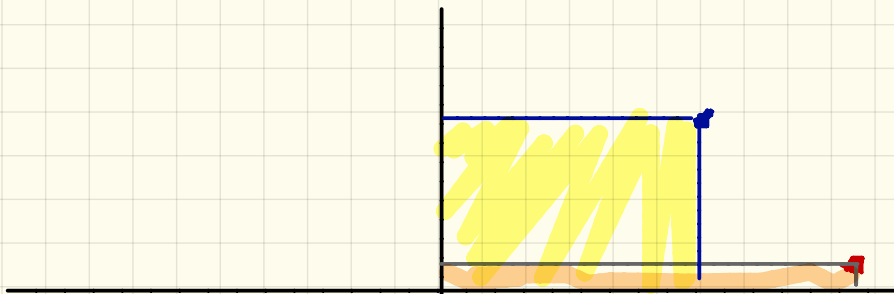
		Y	
		y _j	
X	x _i	x _i y _j m _{ij}	

$$= \sum_{i=1}^n \sum_{j=1}^c \underbrace{x_i y_j m_{ij}}_{\uparrow} - \mu_x \mu_y$$

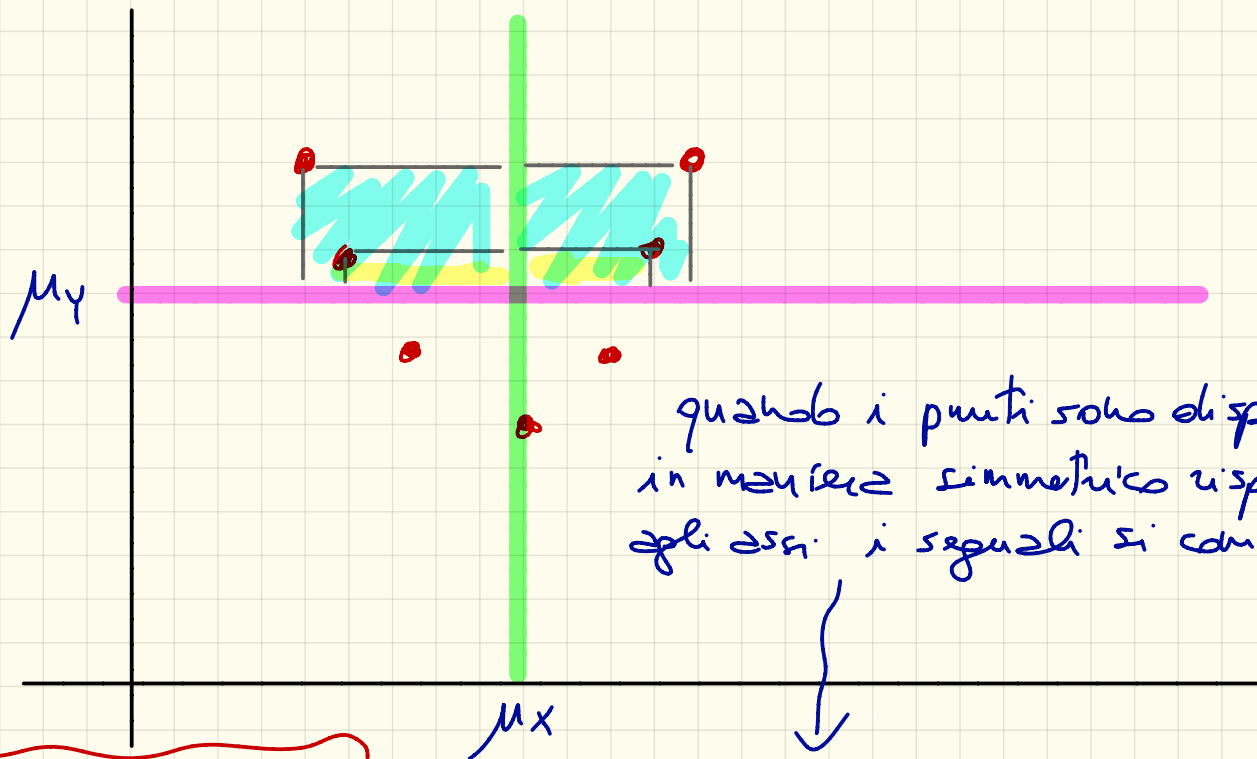
→ la media di questa tabella è il primo termine



	Peso (kg)	Altezza (cm)	scarti X	scarti Y
	70	140	-5	-24
	72	151	-3	-13
	69	159	-6	-5
	75	160	0	-4
	78	162	3	-2
	75	164	0	0
	82	170	7	6
	86	176	11	12
	70	178	-5	14
	73	180	-2	16
TOTALE	750	1640		
MEDIA	75	164		



un punto contribuisce in misura
maggiore a σ_{xy} quanto più è
distante contemporaneamente da
entrambi gli assi



quando i punti sono disposti
in maniera simmetrica rispetto
agli assi i segnali si compensano



← la covarianza è in grado di
cogliere solo legami di tipo
lineare

$$\sigma_{xy} = 0$$



Assenza legame
lineare

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

$\overbrace{\quad\quad\quad}^{kg \times cm}$
 $\overbrace{\quad\quad}^{kg} \quad \overbrace{\quad\quad}^{cm}$

l'indice dipende dalle due unità di misura in gioco

↓
disuguaglianza di Cauchy-Schwarz

$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq +\sigma_x \sigma_y$$

dividendo per $\sigma_x \sigma_y$
disuguaglianza per $\sigma_x \sigma_y$

$$-1 \leq \frac{\sigma_{xy}}{\sigma_x \sigma_y} \leq +1$$

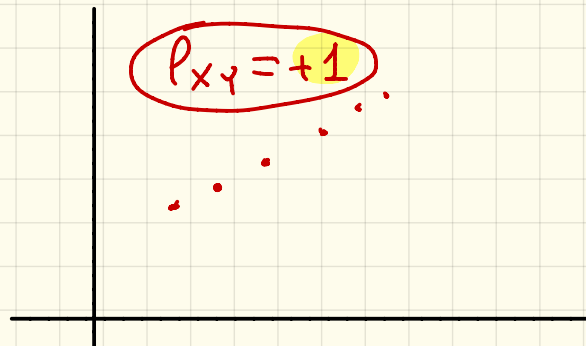
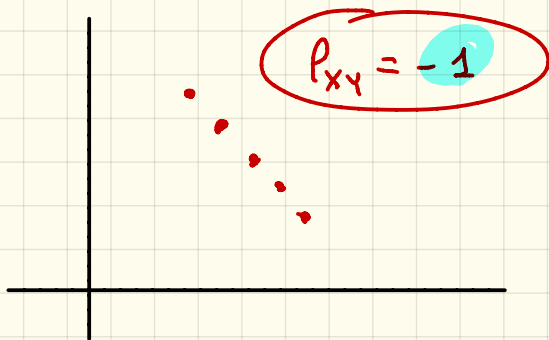
$\boxed{r_{xy}}$
 σ_{xy}

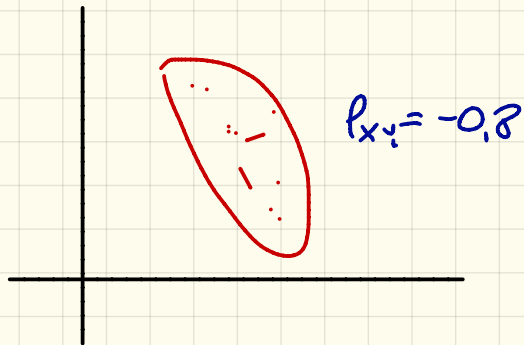
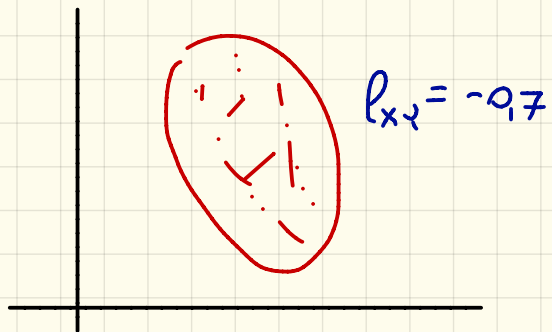
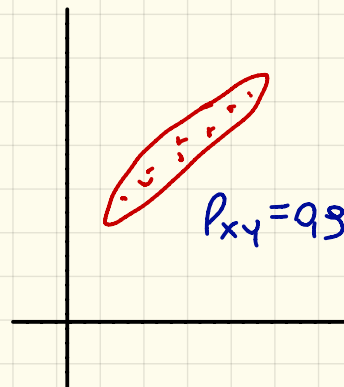
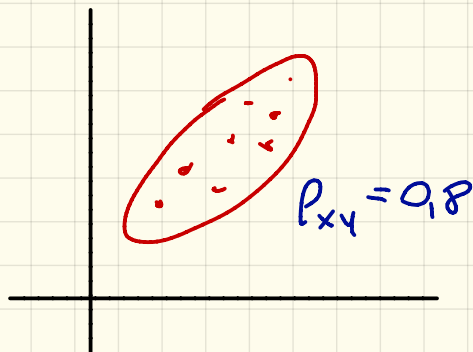
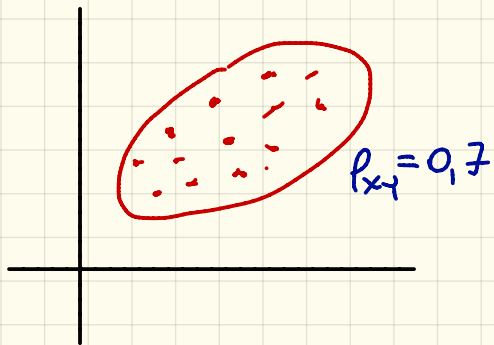
indice di correlazione
lineare

Pearson
[Bravais-Pearson]

$$\text{Corr}(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

normalizzazione di σ_{XY}
in $[-1, +1]$





$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq +\sigma_x \sigma_y$$

$$-1 \leq \rho_{xy} \leq +1$$



$$\frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

σ_x

σ_y

media dei
prodotti delle
variabili
standardizzate

standardizzazione
di X

standardizzazione
di Y

ρ_{xy} va interpretato in segno e grandezza

tipo di legame

discordanza

< 0

concordanza

> 0

grandezza

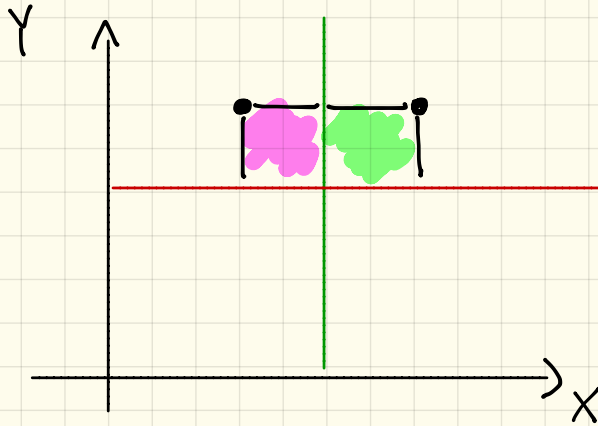
forza del
legame

$$\sum (x_i - \mu_x)(y_i - \mu_y)$$

$\text{CODEV}(X, Y)$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{\text{CODEV}(X, Y)}{n}}{\sqrt{\frac{\text{Dev}(X)}{n}} \sqrt{\frac{\text{Dev}(Y)}{n}}} = \frac{\text{CODEV}(X, Y)}{\sqrt{\text{Dev}(X) \text{Dev}(Y)}}$$

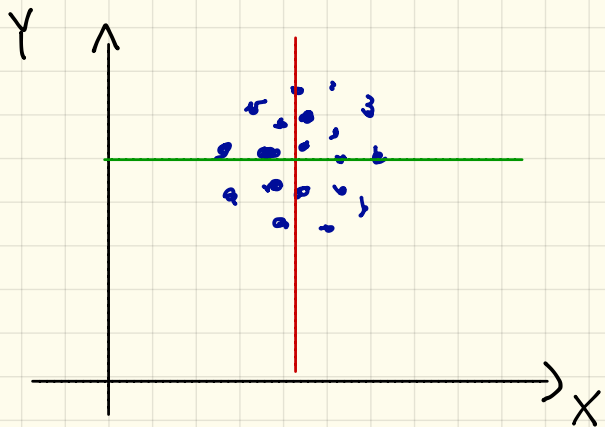
$$\frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2}}$$



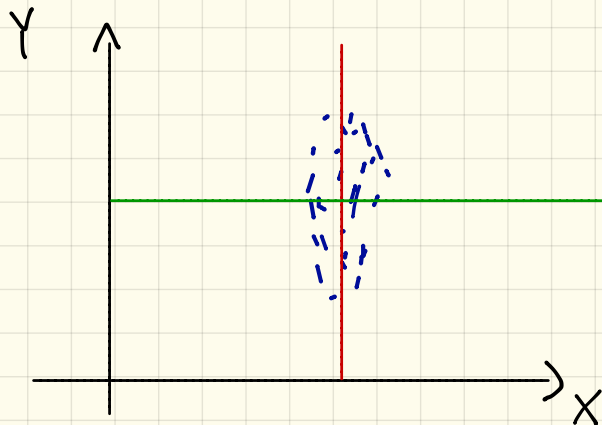
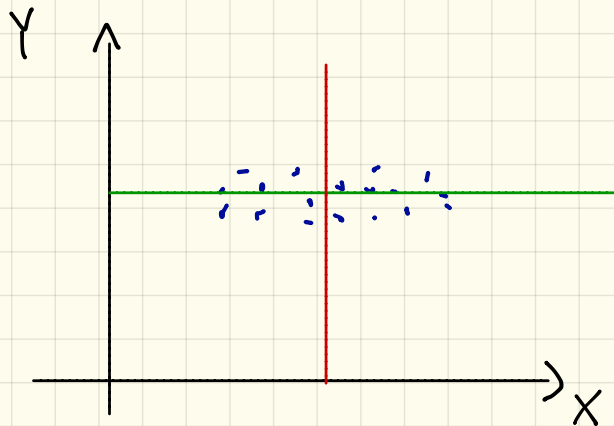
i due segnali si compensano
perfettamente

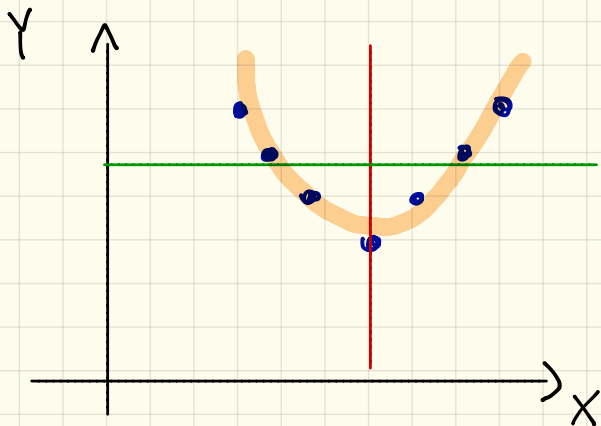
↓
stessa grandezza ma
segno opposto

$\sigma_{XY} = 0 \leadsto$ i segnali concordi (I e III) sono
perfettamente bilanciati da quelli
disconcordi (II e IV)



$\sigma_{xy} = 0$
 tra X e Y non sembra
 esserci un legame che vale
 la pena "sfruttare"



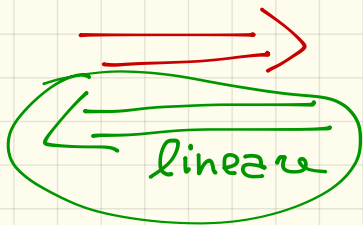


$$\sigma_{xy} = 0$$

↓
~~assenza relazione~~

↓
 la covarianza è in grado di cogliere solo la presenza di forme lineari

ASSENZA
 RELAZIONE
 TRA X e Y



$$\sigma_{xy} = 0$$



NOTA: il numeratore della covarianza

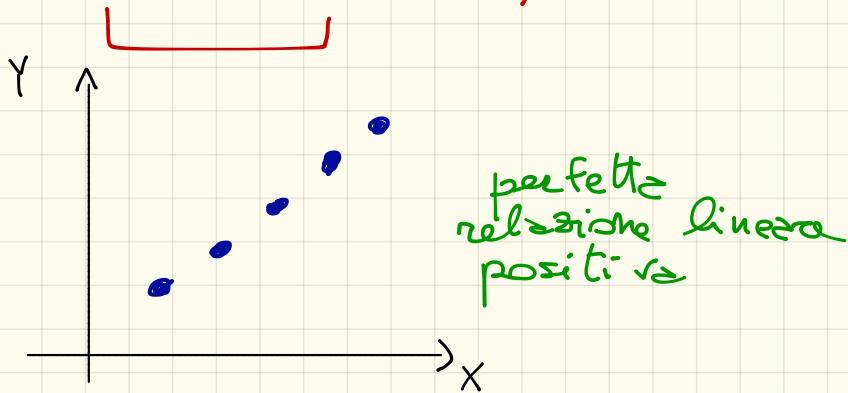
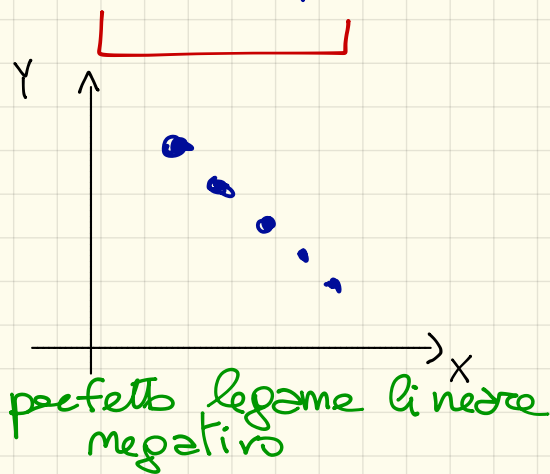
$$\sum_i (x_i - \mu_x)(y_i - \mu_y)$$

si chiama covarianza $[cov(x, y)]$

La covarianza assume valori tra:

$$- \sigma_x \sigma_y \leq \sigma_{xy} \leq + \sigma_x \sigma_y$$

disuguaglianze di
Cauchy-Schwarz



$$\frac{-\sigma_X \sigma_Y}{\sigma_X \sigma_Y} \leq \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \leq \frac{+\sigma_X \sigma_Y}{\sigma_X \sigma_Y}$$

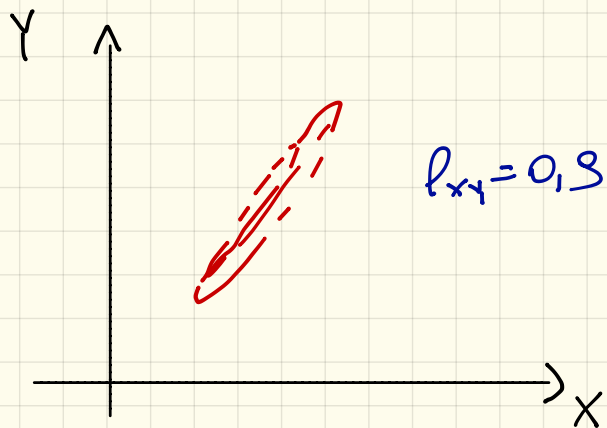
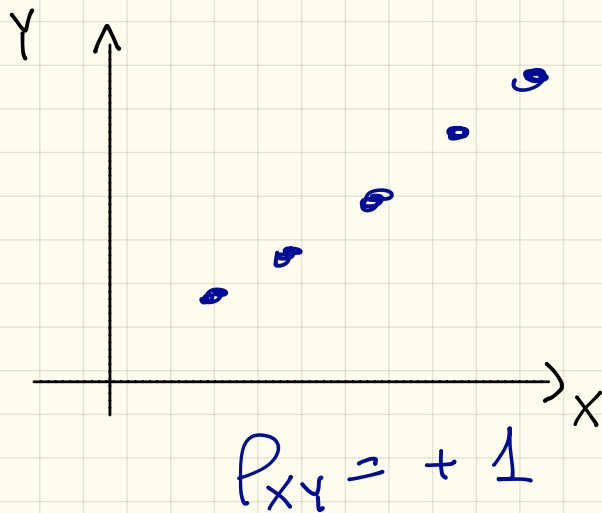
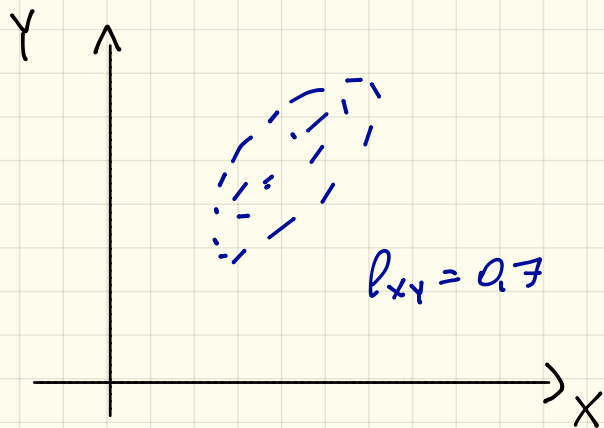
$$-1 \leq \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \leq +1$$

normalizzazione della
covarianza in $[-1, +1]$

ρ_{XY}
 $\text{cor}(X, Y)$

coefficiente di
correlazione di Pearson
(o di Bravais-Pearson)

non dipende
dall'unità di
misura



$$\rho_{xy} = \frac{\frac{1}{n} \sum_i (x_i - \mu_x) (y_i - \mu_y)}{\sigma_x \sigma_y} =$$

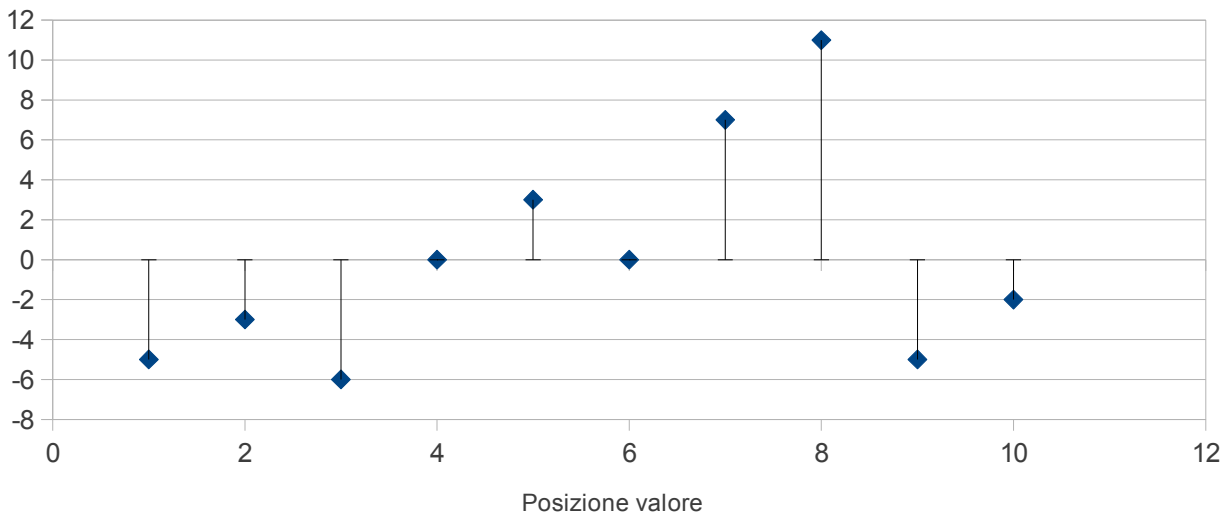
$$= \frac{1}{n} \sum_i \left(\frac{x_i - \mu_x}{\sigma_x} \right) \left(\frac{y_i - \mu_y}{\sigma_y} \right)$$

media dei prodotti delle due
variabili standardizzate

Differente interpretazione grafica della covarianza usando i diagrammi ad aste di:

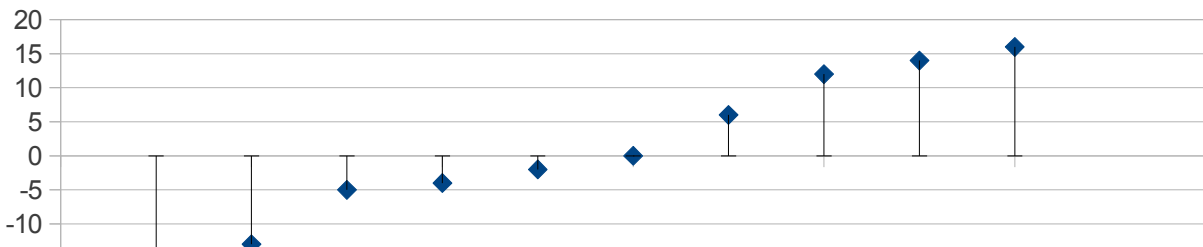
Scarti X

$$\leadsto X - \mu_x$$



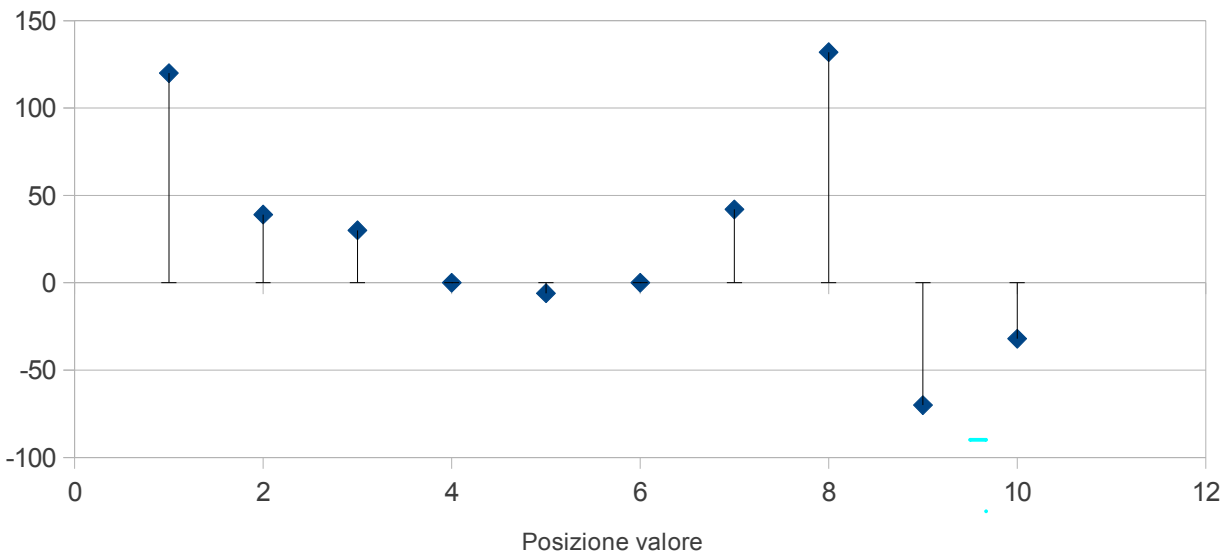
Scarti Y

$$\leadsto Y - \mu_y$$



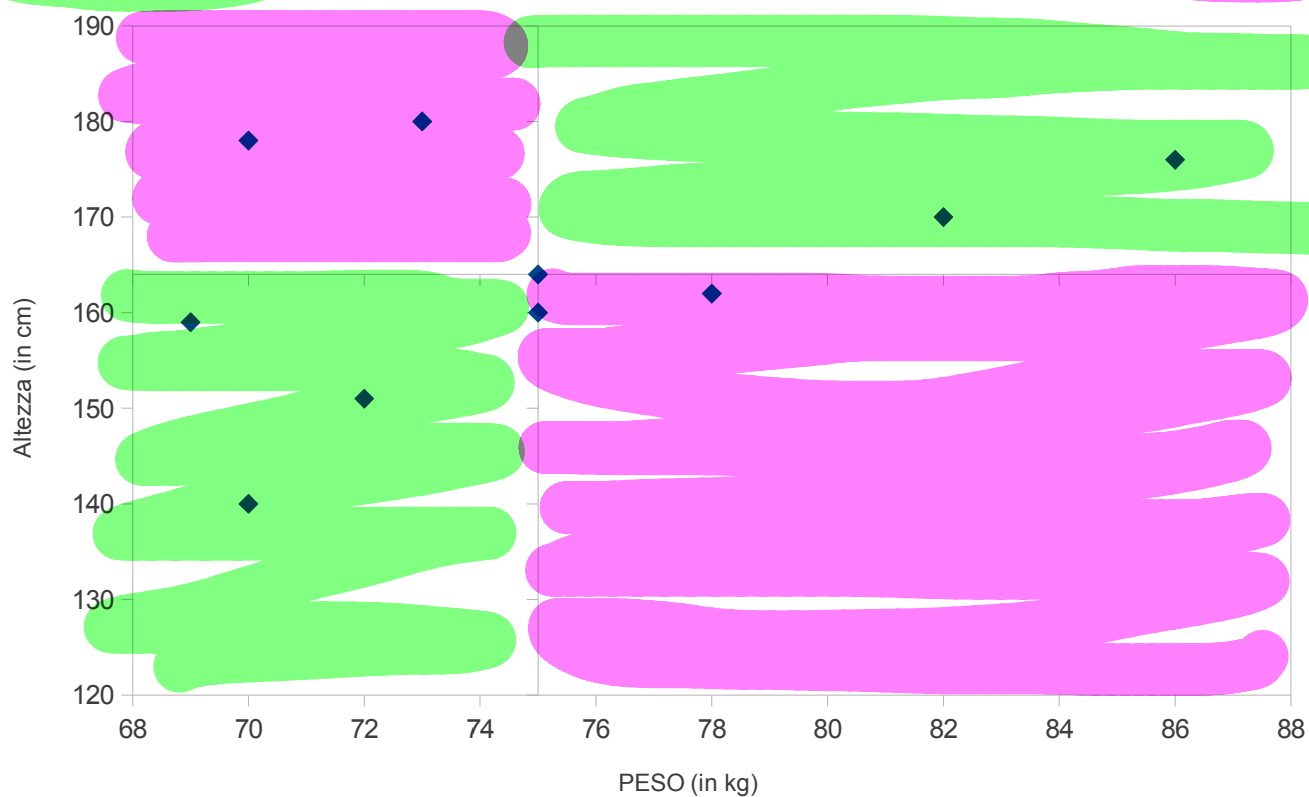
Prodotto degli scarti

$$\leadsto (X - \mu_x)(Y - \mu_y)$$



Segnali
concorrenza

Segnali
discordanza



Prodotto degli scarti

