Analisi di relazioni tra variabili

- Correlazione: analizza se esiste una relazione tra due variabili (come e quanto due variabili variano insieme)
- Regressione: analizza la forma della relazione tra variabili

MISURE DI ASSOCIAZIONE TRA 2 VARIABILI QUANTITATIVE

- Covarianza
- Coefficiente di correlazione

Esempio: Consideriamo i dati relativi alla pressione sistolica arteriosa e al peso di 59 soggetti:

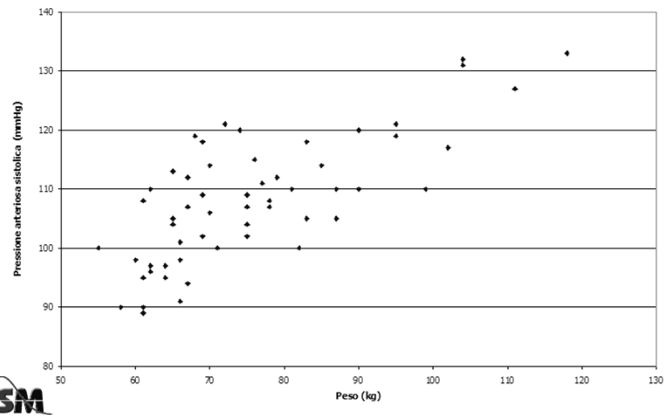
Peso	Pressione arteriosa sistolica
55	100
58	90
60	98
61	95
61	108
61	89
61	90
62	97
62	96
62	110
64	95
64	97
65	105
65	104
65	113
66	98
66	101
66	91
67	107
67	112
67	94
68	119
69	102
69	102
69	118
69	109
70	114
70	106
71	100

Peso	Pressione Arteriosa sistolica
72	121
74	120
75	104
75	102
75	109
75	107
76	115
77	111
78	108
78	107
79	112
81	110
82	100
83	105
83	118
85	114
87	105
87	110
90	110
90	120
95	121
95	119
99	110
102	117
104	131
104	132
111	127
118	133

Quale relazione tra i dati?

DIAGRAMMA DI DISPERSIONE

- Riportiamo su un diagramma cartesiano in ascissa (X) i valori del peso e in ordinata (Y) i valori della pressione arteriosa



Commenti

Il grafico precedente ci mostra che:

- Peso e pressione sistolica arteriosa sono "positivamente associate": i soggetti che hanno peso più elevato, hanno anche valori della pressione arteriosa maggiori

•La relazione tra le due variabili, ad una prima osservazione, sembra essere lineare

→ Quanto sono "associate"?
Qual è la forza della relazione?
Che tipo di relazione tra le variabili?

SESM

Covarianza

$$Cov(X,Y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

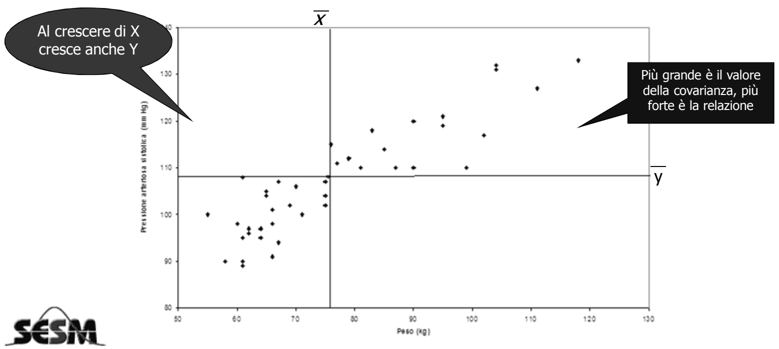
Dove (x_i, y_i) sono i dati disponibili per due variabili numeriche
$\bar{x}, \bar{y}$ indicano le due medie aritmetiche

SESM

Covarianza positiva

- Considera i valori: $(x_i - \bar{x})(y_i - \bar{y})$

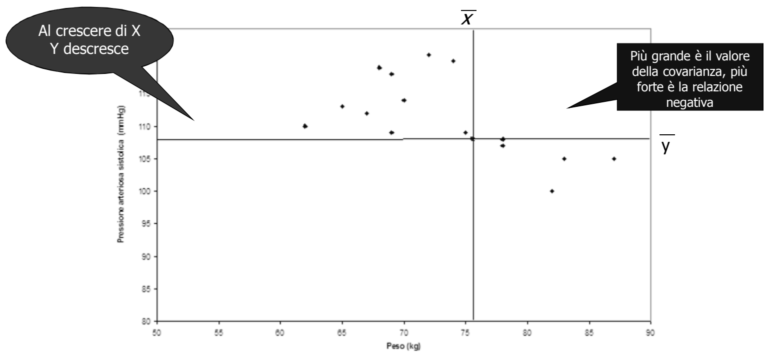
$(x_i - \bar{x})>0$ e $(y_i - \bar{y})>0$
$(x_i - \bar{x})<0$ e $(y_i - \bar{y})<0$ → Il prodotto è positivo → La covarianza è positiva



SESM

Covarianza negativa

$(x_i - \bar{x})>0$ e $(y_i - \bar{y})<0$
$(x_i - \bar{x})<0$ e $(y_i - \bar{y})>0$ → Il prodotto è negativo → La covarianza è negativa



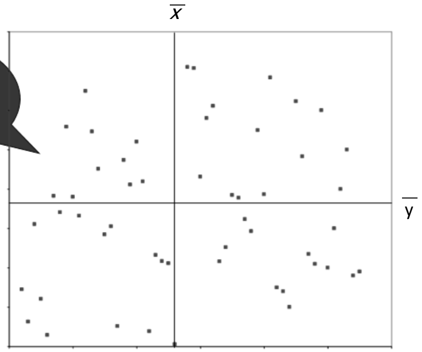
SESM

Covarianza nulla

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

La somma dei prodotti è nulla → La covarianza è nulla **Nessuna relazione**

Se X cresce, Y può crescere o decrescere



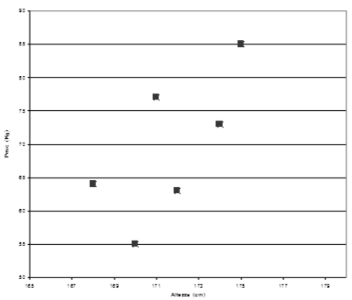
I valori non si concentrano in alcun quadrante in particolare

SESM

Esercizio: rappresentare graficamente i dati riportati in tabella e ricavare il valore della covarianza. Commentare i risultati ottenuti.

	statura (cm) ((X)	*peso (Kg) (Y)*	*(x-171.7)*	*(y-69.5)*	*(x-171.7)(y-69.5)*	*xy*
	172	63	0.3	-6.5	-1.95	10836
	174	73	2.3	3.5	8.05	12702
	171	77	-0.7	7.5	-5.25	13167
	175	85	3.3	15.5	51.15	14875
	168	64	-3.7	-5.5	20.35	10752
	170	55	-1.7	-14.5	24.65	9350
Totale:	1030	417			97	71682
Media:	171.7	69.5				

SESM



Cov(X,Y)=19,4

- La covarianza ha segno positivo: peso e altezza sono due variabili positivamente associate: a valori maggiori dell'altezza, corrispondono valori maggiori del peso.

NB: Se l'altezza fosse stata espressa in m, anziché in cm la covarianza sarebbe risultata pari a **0,194**

La covarianza è dipende dalla <u>scala di misura</u> in cui sono espresse le variabili

SESM

Coefficiente di correlazione di Bravais-Pearson

r(XY) non dipende dalla scala di misura delle variabili

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} = covarianza tra *X* e *Y*

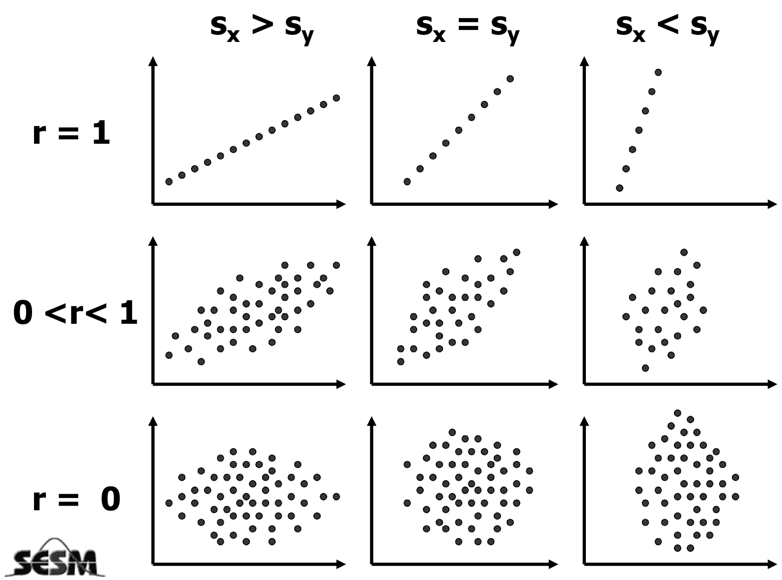
s_x = deviazione standard di *X*

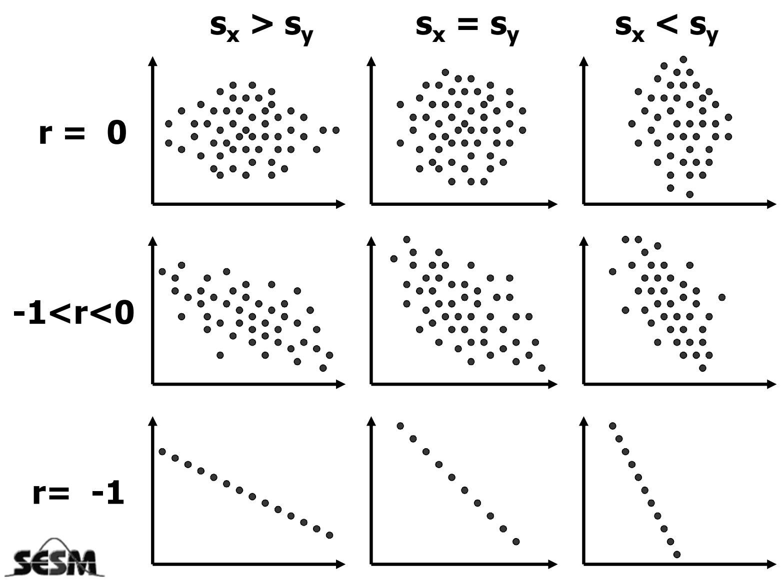
s_y = deviazione standard di *Y*

Massima correlazione lineare negativa ← $-1 \leq r_{xy} \leq 1$ → Massima correlazione lineare positiva

$r_{xy} = 0$ → Nessuna correlazione

SESM





Esempio: Utilizzando i dati dell'esempio relativo a peso e pressione arteriosa, ricaviamo il valore della covarianza e del coefficiente di correlazione:

Sapendo che: ∑x=4310, n=57
∑y=6158, s(x)=14.3
∑xy=471976 s(y)=10.6

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) = \frac{1}{57-1}(471976 - (1/57)*6158*4310) = 113.3$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{113.3}{14.3*10.6} = 0.75$$

- Come si poteva osservare graficamente, peso e pressione sono positivamente associate.
- r è molto elevato: tra le due variabili c'è relazione lineare

Coefficiente di correlazione di Bravais-Pearson

La correlazione è significativa?
- Il valore di **r è stato calcolato da un** campione e non dalla popolazione (**ρ)**
- Ipotesi nulla: ρ = 0 (ρ è il coefficiente di correlazione della popolazione, **r del** campione).
- Il valore calcolato indica una correlazione significativa?

Coefficiente di correlazione di Bravais-Pearson

OK: la correlazione è significativa ma….

- Le 2 variabili sono distribuite normalmente?
- La relazione tra le 2 variabili è lineare?
- Ricordarsi che anche se c’e’ correlazione non vuol dire che c’e’ nesso di causa-effetto …