# LEZIONI DI STATISTICA MEDICA

***Prof. Roberto de Marco***

***Lezione n.6***

- ***Distribuzione bivariata***

***- Misure di associazione tra due variabili quantitative***

*Sezione di Epidemiologia & Statistica Medica*
*Università degli Studi di Verona*

# DISTRIBUZIONE BIVARIATA (CROSS-TABULATION)

**Permette la rappresentazione congiunta della distribuzione di frequenza di due variabili qualitative**

**Permette di capire la relazione tra le due variabili**

Esempio: *distribuzione dell'abitudine al fumo e della broncopneumopatia cronico-ostruttiva (GOLD-BPCO: 0+) in adulti italiana di età 20-44 anni (indagine ISAYA).*

| Fumo | $n_i$ | $p_i$ |
|---|---|---|
| non fumatore | 9667 | 51.9% |
| ex fumatore | 2743 | 14.7% |
| fumatore | 6228 | 33.4% |
| Totale | 18638 | 100.0% |

| BPCO | $n_i$ | $p_i$ (%) |
|---|---|---|
| assente | 16622 | 89.2% |
| presente | 2016 | 10.8% |
| Totale | 18638 | 100.0% |

# DISTRIBUZIONE CONGIUNTA ASSOLUTA

distribuzione congiunta del fumo e della BPCO ($n_{ij}$)

| | BPCO | | |
|---|---|---|---|
| **FUMO** | **assente** | **presente** | **TOTALE** |
| **non fumatore** | **9042** | **625** | **9667** |
| **ex fumatore** | **2472** | **271** | **2743** |
| **fumatore** | **5108** | **1120** | **6228** |
| **TOTALE** | **16622** | **2016** | **18638** |

distribuzione marginale del fumo ($n_i$)

dimensione campionaria (n)

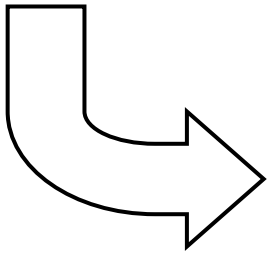
distribuzione marginale della BPCO ($n_j$)

# DISTRIBUZIONE CONGIUNTA RELATIVA (%)

| FUMO | BPCO assente | BPCO presente | TOTALE |
|---|---|---|---|
| non fumatore | 9042 | 625 | 9667 |
| ex fumatore | 2472 | 271 | 2743 |
| fumatore | 5108 | 1120 | 6228 |
| TOTALE | 16622 | 2016 | 18638 |

**non fumatori con BPCO ($n_{12}$)**

dimensione campionaria (n)



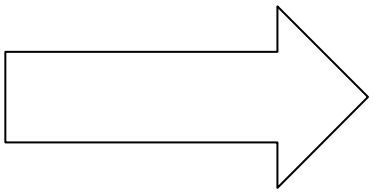
$(n_{ij} / n) * 100$

(625 / 18638) * 100

| FUMO | BPCO assente | BPCO presente | TOTALE |
|---|---|---|---|
| non fumatore | 48.5% | 3.4% | 51.9% |
| ex fumatore | 13.3% | 1.4% | 14.7% |
| fumatore | 27.4% | 6.0% | 33.4% |
| TOTALE | 89.2% | 10.8% | 100.0% |

SESM

# DISTRIBUZIONI CONDIZIONALI



- **Rappresentano la distribuzione di una variabile all'interno delle modalità dell'altra variabile**

**N.B.** **Se le distribuzioni condizionali sono differenti, si può supporre che esista una <u>relazione</u> tra le due variabili**

# **DISTRIBUZIONI CONDIZIONALI AI MARGINALI DI RIGA:** DISTRIBUZIONE DELLA BPCO PER LIVELLO DI FUMO

| | BPCO | | |
|---|---|---|---|
| **FUMO** | **assente** | **presente** | **TOTALE** |
| non fumatore | 9042 | 625 | 9667 |
| ex fumatore | 2472 | 271 | 2743 |
| fumatore | 5108 | 1120 | 6228 |
| TOTALE | 16622 | 2016 | 18638 |

**marginali di riga ($n_i$)**

(625 / 9667) * 100

**$(n_{ij} / n_i) * 100$**

| | BPCO | | |
|---|---|---|---|
| **FUMO** | **assente** | **presente** | **TOTALE** |
| non fumatore | 93.5% | 6.5% | 100.0% |
| ex fumatore | 90.1% | 9.9% | 100.0% |
| fumatore | 82.0% | 18.0% | 100.0% |
| TOTALE | 89.2% | 10.8% | 100.0% |

SESM

# DISTRIBUZIONI CONDIZIONALI AI MARGINALI DI COLONNA: DISTRIBUZIONE DEL FUMO PER LIVELLO DELLA BPCO

| | BPCO | | |
|---|---|---|---|
| **FUMO** | **assente** | **presente** | **TOTALE** |
| non fumatore | 9042 | 625 | 9667 |
| ex fumatore | 2472 | 271 | 2743 |
| fumatore | 5108 | 1120 | 6228 |
| TOTALE | 16622 | 2016 | 18638 |

**marginali di colonna ($n_j$)**

**$(n_{ij} / n_j) * 100$**

(625 / 2016) * 100

| | BPCO | | |
|---|---|---|---|
| **FUMO** | **assente** | **presente** | **TOTALE** |
| non fumatore | 54.4% | 31.0% | 51.9% |
| ex fumatore | 14.9% | 13.4% | 14.7% |
| fumatore | 30.7% | 55.6% | 33.4% |
| TOTALE | 100.0% | 100.0% | 100.0% |

SESM

# ESERCIZIO

**In un'indagine, è stato chiesto ad un gruppo di 101 consumatori e ad un gruppo di 124 dentisti se erano favorevoli alla pubblicità fatta dai dentisti per attrarre nuovi pazienti.**

**Si sono ottenuti i seguenti risultati:**

| | GIUDIZIO | | | | | |
|---|---|---|---|---|---|---|
| **CATEGORIA** | **molto favorevole** | **abbastanza favorevole** | **indifferente** | **abbastanza sfavorevole** | **molto sfavorevole** | **TOTALE** |
| **consumatore** | **34** | **49** | **9** | **4** | **5** | **101** |
| **dentista** | **9** | **18** | **23** | **28** | **46** | **124** |
| **TOTALE** | **43** | **67** | **32** | **32** | **51** | **225** |

1. C'è differenza tra il giudizio espresso dai consumatori e dai dentisti? C'è relazione tra la categoria e il giudizio?
2. Cercate di interpretare il risultato

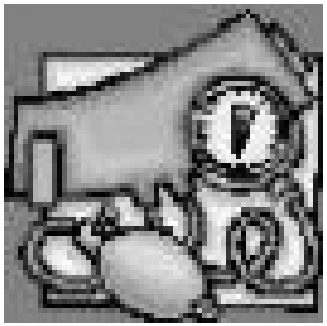
# MISURE DI ASSOCIAZIONE TRA 2 VARIABILI QUANTITATIVE

- 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 |

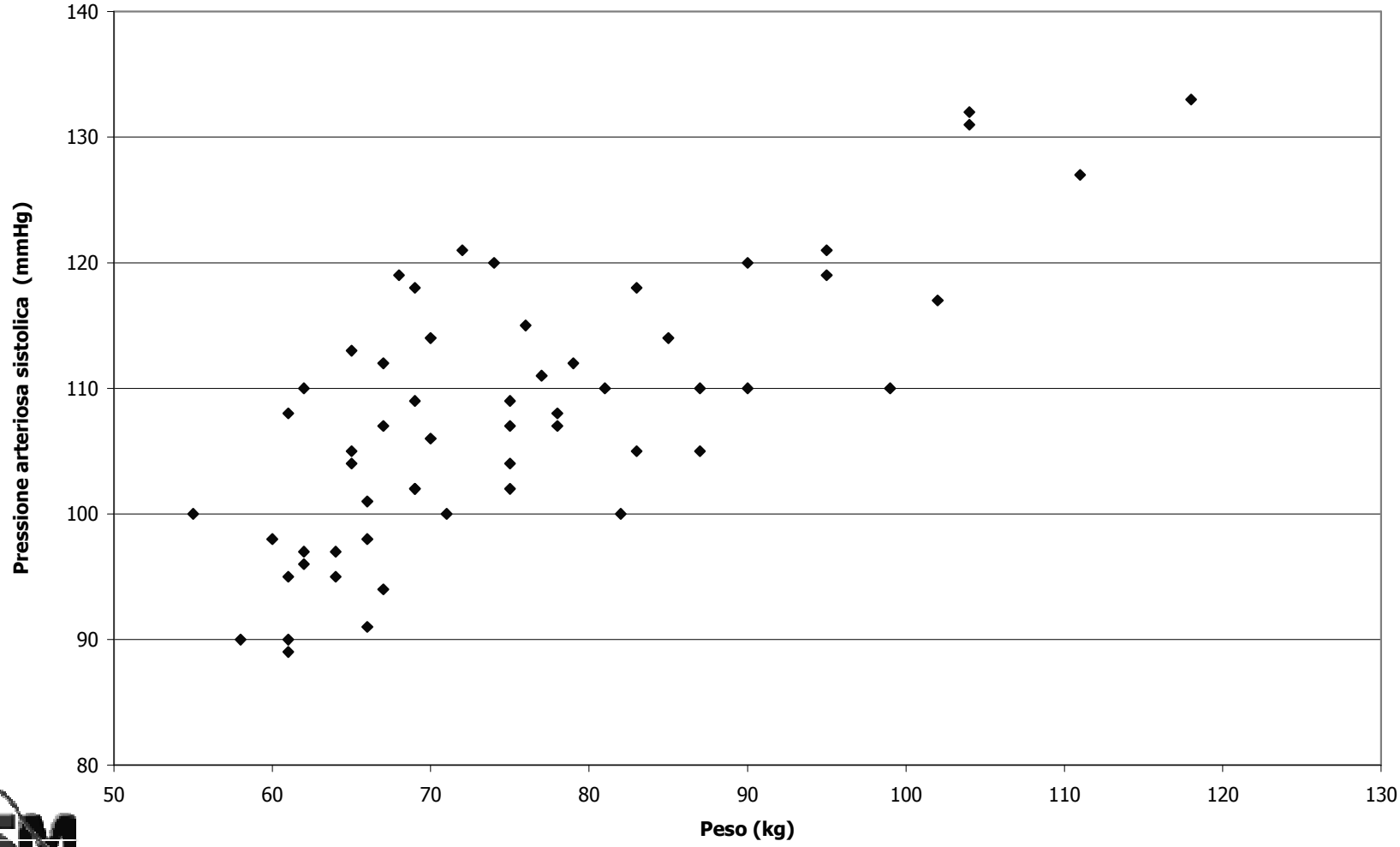


Quale relazione tra i dati?

SESM

# 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

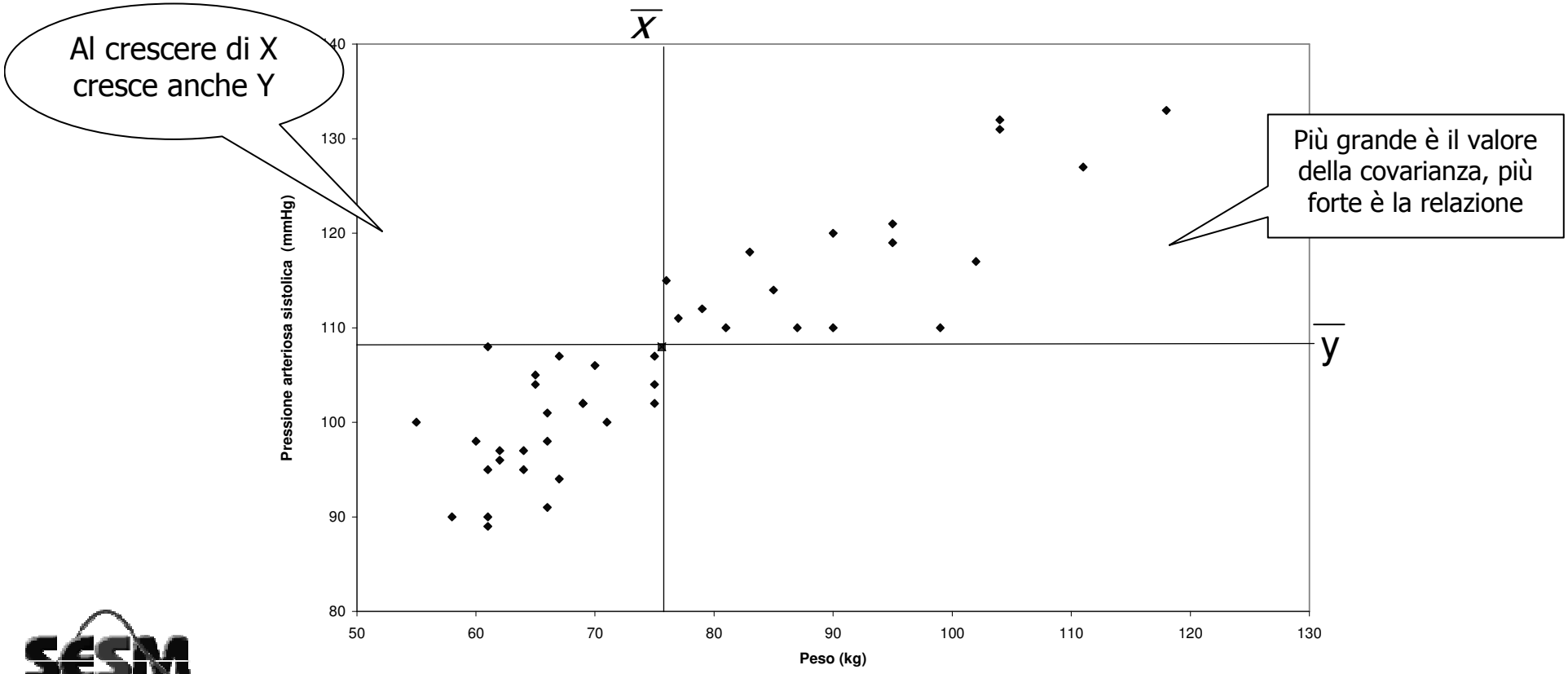
# Covarianza positiva

- 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

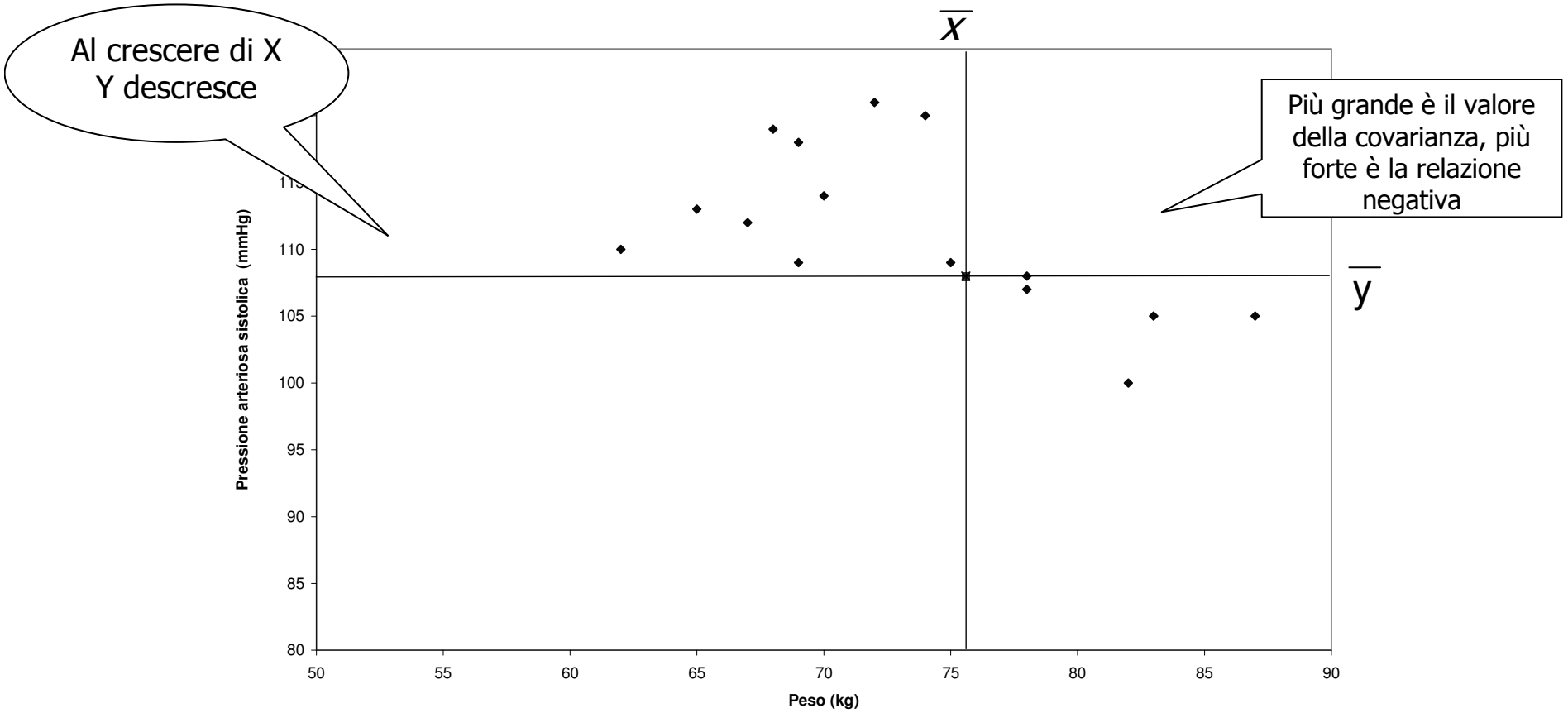


# Covarianza negativa

$(x_i-\bar{x})>0$ 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

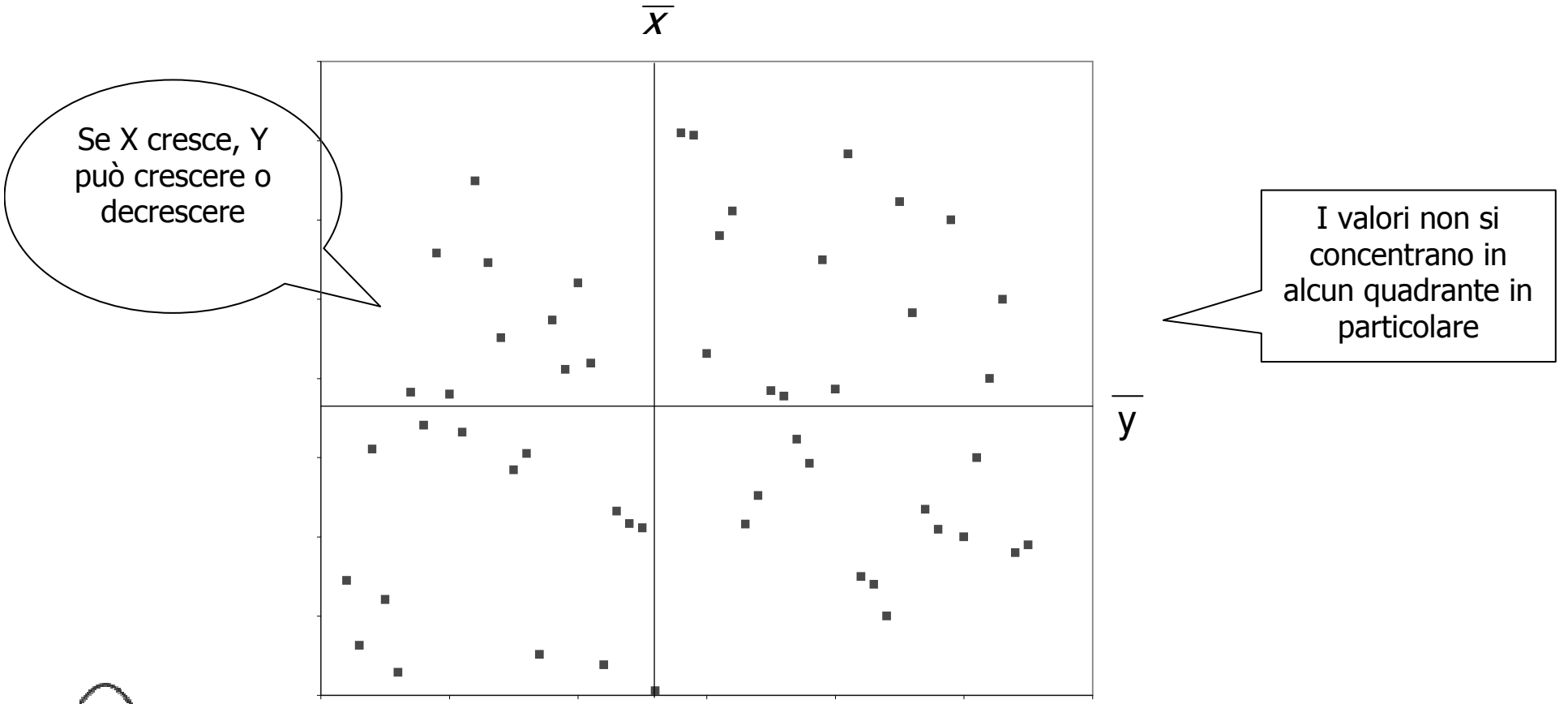


# Covarianza nulla

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

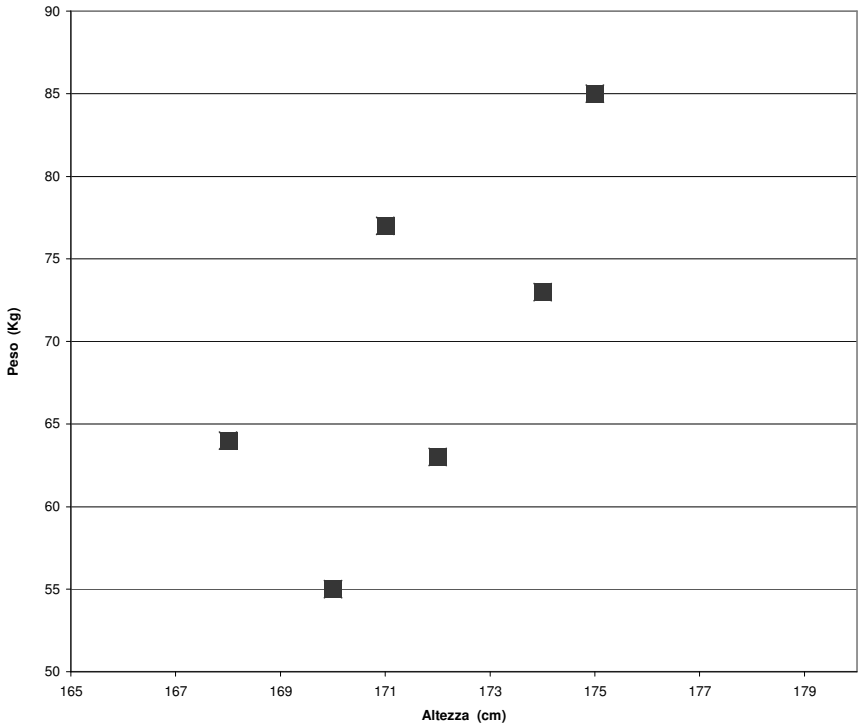
La somma dei prodotti è nulla

La covarianza è nulla
**<u>Nessuna relazione</u>**



**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 | | | | |




**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 scala di misura in cui sono espresse le variabili

# 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*

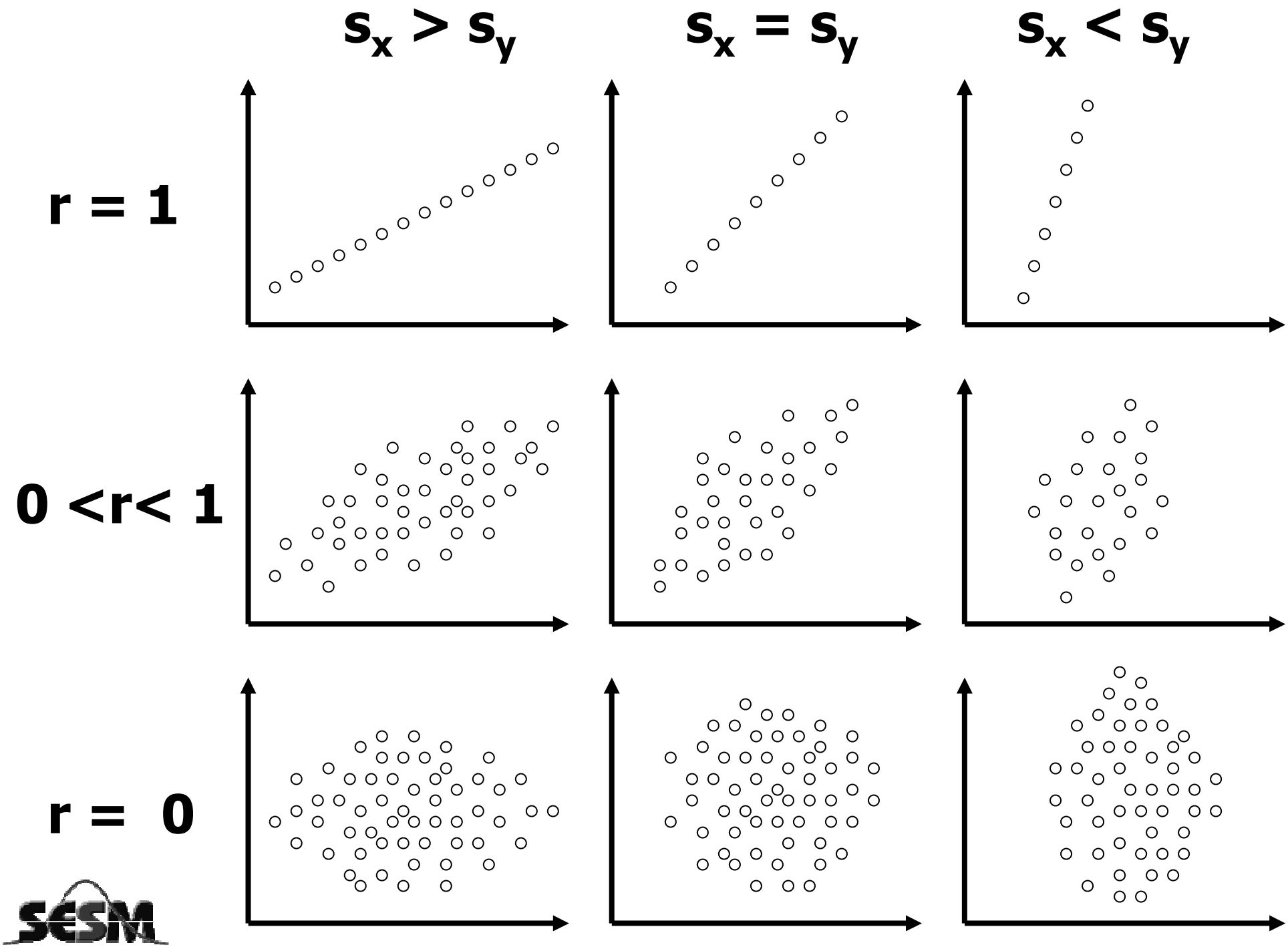
Massima correlazione lineare negativa

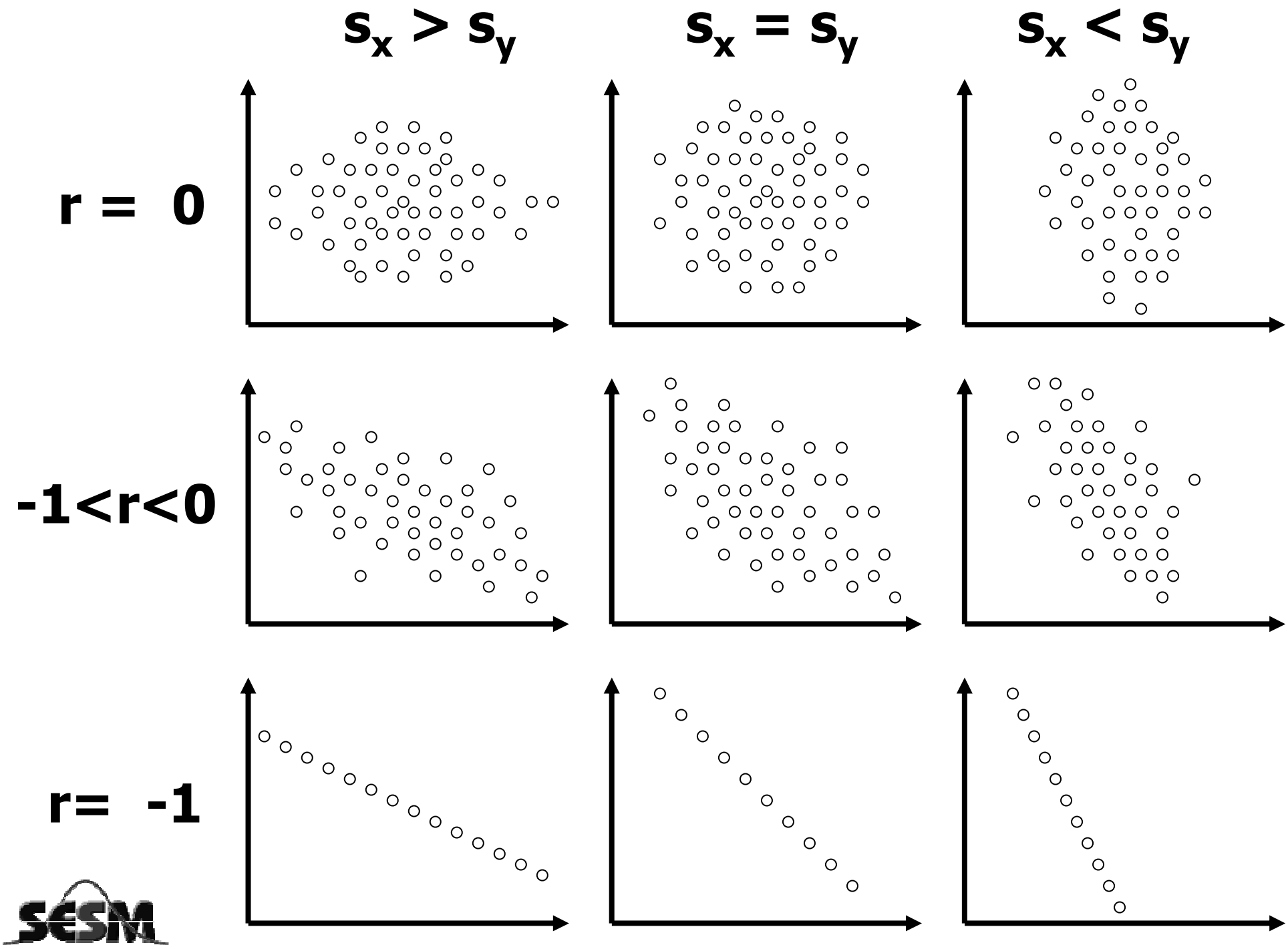
$-1 \leq r_{xy} \leq 1$

Massima correlazione lineare positiva

$r_{xy} = 0$

Nessuna correlazione


$s_x > s_y$
$s_x = s_y$
$s_x < s_y$
r = 1
0 <r< 1
r = 0
SESM


$s_x > s_y$
$s_x = s_y$
$s_x < s_y$
r = 0
-1<r<0
r= -1
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

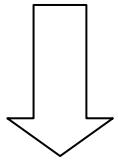
Esercizio: dai dati dell'esempio relativo al peso e all'altezza, calcolare il coefficiente di correlazione di Bravais Pearson e commentare il risultato.

| | ***statura (cm) ( X)*** | ***peso (Kg) (Y)*** |
|---|---|---|
| | 172 | 63 |
| | 174 | 73 |
| | 171 | 77 |
| | 175 | 85 |
| | 168 | 64 |
| | 170 | 55 |
| **tot** | 1030 | 417 |
| media | 171.7 | 69.5 |

s(X,Y)=16.2

s(X)= 2,58

s(Y)= 10,9



# CONCLUSIONI

- La statistica descrittiva è un metodo che ci permette di riassumere **l'informazione** contenuta in un insieme di dati campionari in poche **misure sintetiche**

- Tali misure prendono il nome di **STATISTICHE**. Le più note ed utilizzate sono: $\overline{x}$, $s^2$, $s$, $P$,$r$,$s_{xy}$

- L'importanza delle statistiche campionarie risiede nel fatto che esse ci permettono di **stimare caratteristiche ignote** della popolazione da cui provengono i campioni in studio

# CONCLUSIONI

- Tali caratteristiche prendono il nome di **PARAMETRI** della popolazione e, quando possibile, sono calcolati in modo del tutto analogo.

I parametri vengono indicati con lettere greche:

| | | |
|---|---|---|
| $\mu$ = media | *stimata da* | x |
| $\sigma^2$ = varianza | *stimata da* | $s^2$ |
| $\sigma$ = deviazione standard | *stimata da* | s |
| $\pi$ = probabilità | *stimata da* | P |

SESM