Settimana 2

G. M. Marchetti

2018

Mediana

- Un indice di posizione alternativo alla media e molto usato è la *mediana*
- È basato sull'**ordinamento** dei dati
- La mediana è il valore Me tale che la metà dei dati è minore di Me.

Statistiche ordinate

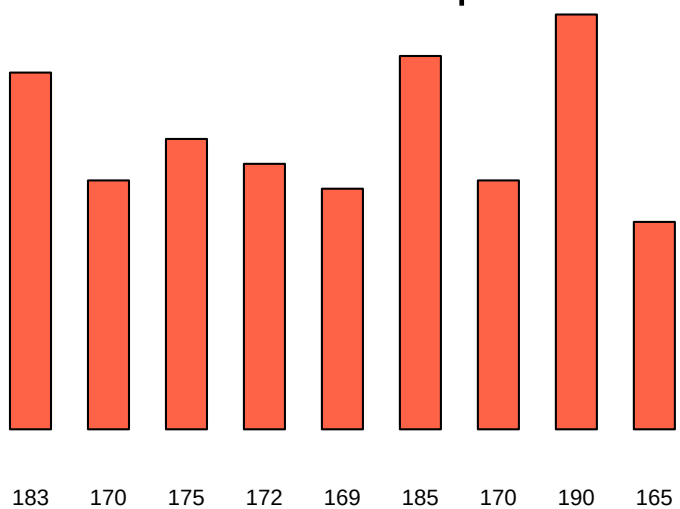
- Supponiamo di avere 9 individui adulti di altezze diverse

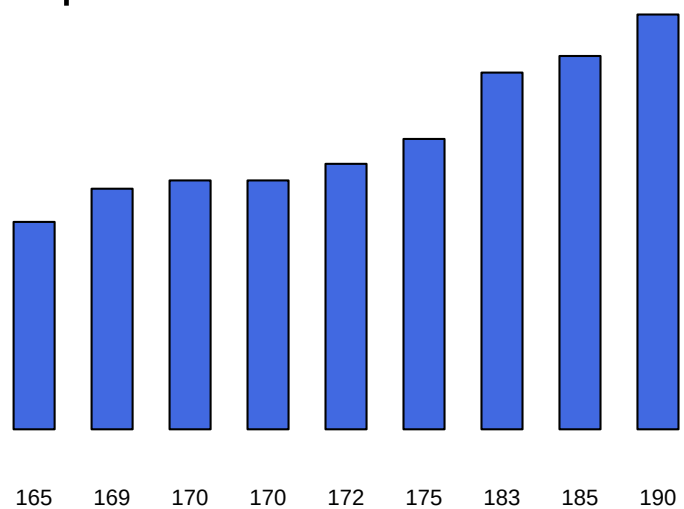
  183 170 175 172 169 185 170 190 165

- I dati ordinati

  165 169 170 170 172 175 183 185 190

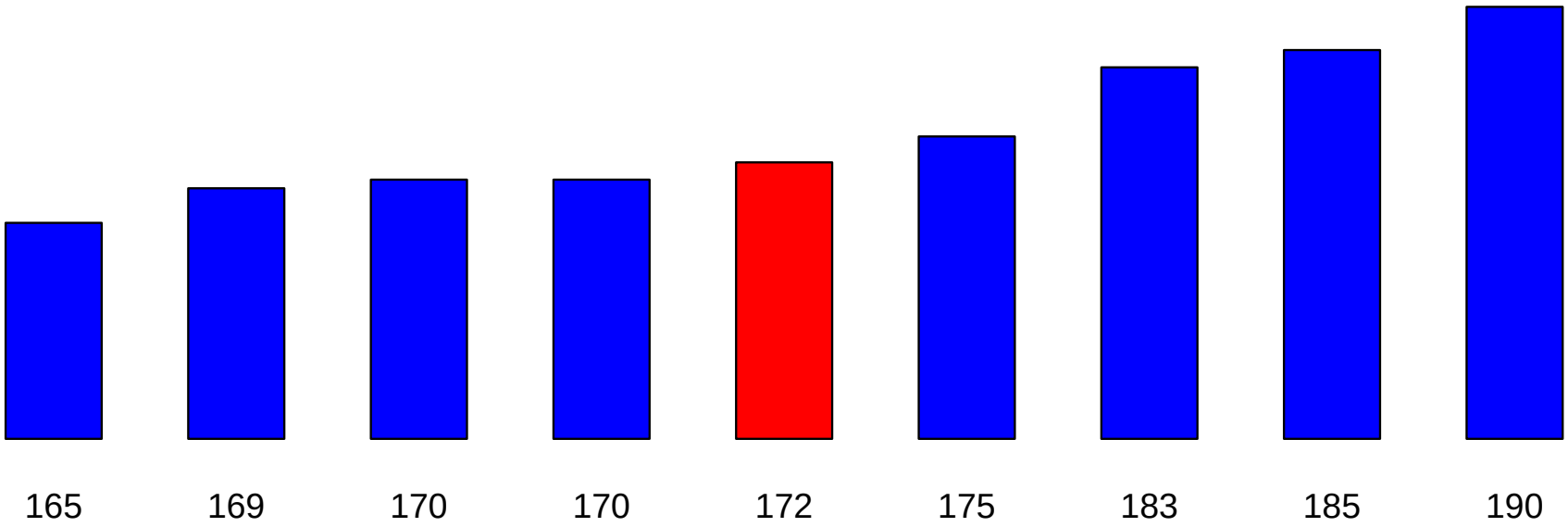
- Possiamo vederle prima disordinate e poi ordinate





Posto centrale

- Nella serie ordinata il valore del **posto centrale** ha un ruolo importante e si chiama **mediana**



- La metà degli individui sono più bassi di 172 e l'altra metà sono più alti
- La mediana è $Me = 172$.

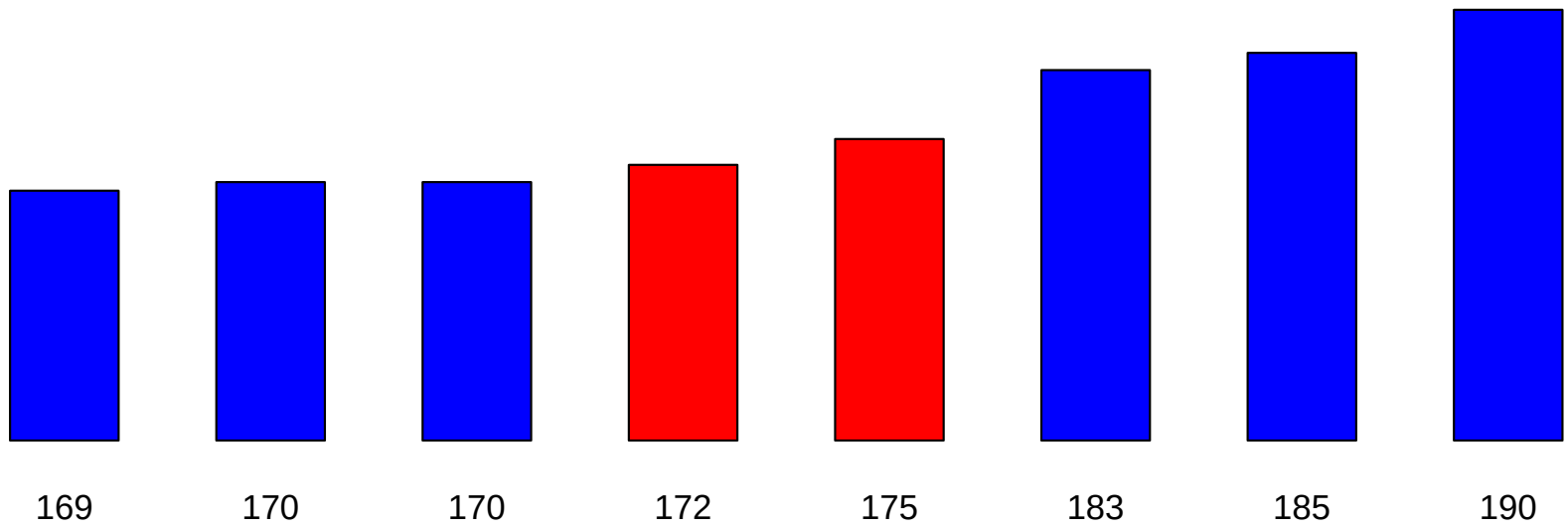
Pari e dispari

- Se n, il numero di unità, è **pari** ci sono **due posti centrali**.
- Altezze (n = 8):

  183 170 175 172 169 185 170 190

- I dati ordinati

  169 170 170 172 175 183 185 190



La mediana è

$$Me = (172 + 175)/2 = 173.5$$

Mediana

- È un indice di posizione basato sulla *successione ordinata* dei dati

$$x_{min} = x_{(1)}, x_{(2)}, \ldots, x_{(n)} = x_{max}.$$

- La mediana è definita come
  - il valore dell'unità centrale se il numero di osservazioni è dispari
  - la semisomma dei valori delle due unità centrali se n è pari.

$$Me = \begin{cases} x_{(n+1)/2} & \text{se } n \text{ è dispari} \\ (x_{(n/2)} + x_{(n/2)+1})/2 & \text{se } n \text{ è pari} \end{cases}$$

- Interpretazione: la metà dei dati hanno valori inferiori alla mediana.

Ve lo ripeto

- **La mediana va calcolata dopo aver ordinato!!**

Media o la mediana?

- Se i dati hanno una distribuzione simmetrica, la media e la mediana sono uguali
- Esempio banale

Altezza	Frequenza	Frequenza cumulata
160	10	10
170	20	30
180	10	40
Totale	40	

- Media e mediana $= 170$
- Media $= \frac{160 \cdot 10 + 170 \cdot 20 + 180 \cdot 10}{40} = 6800/40 = 170$
- Due posti centrali 20 e 21: entrambi nella seconda classe.

È meglio la media o la mediana?

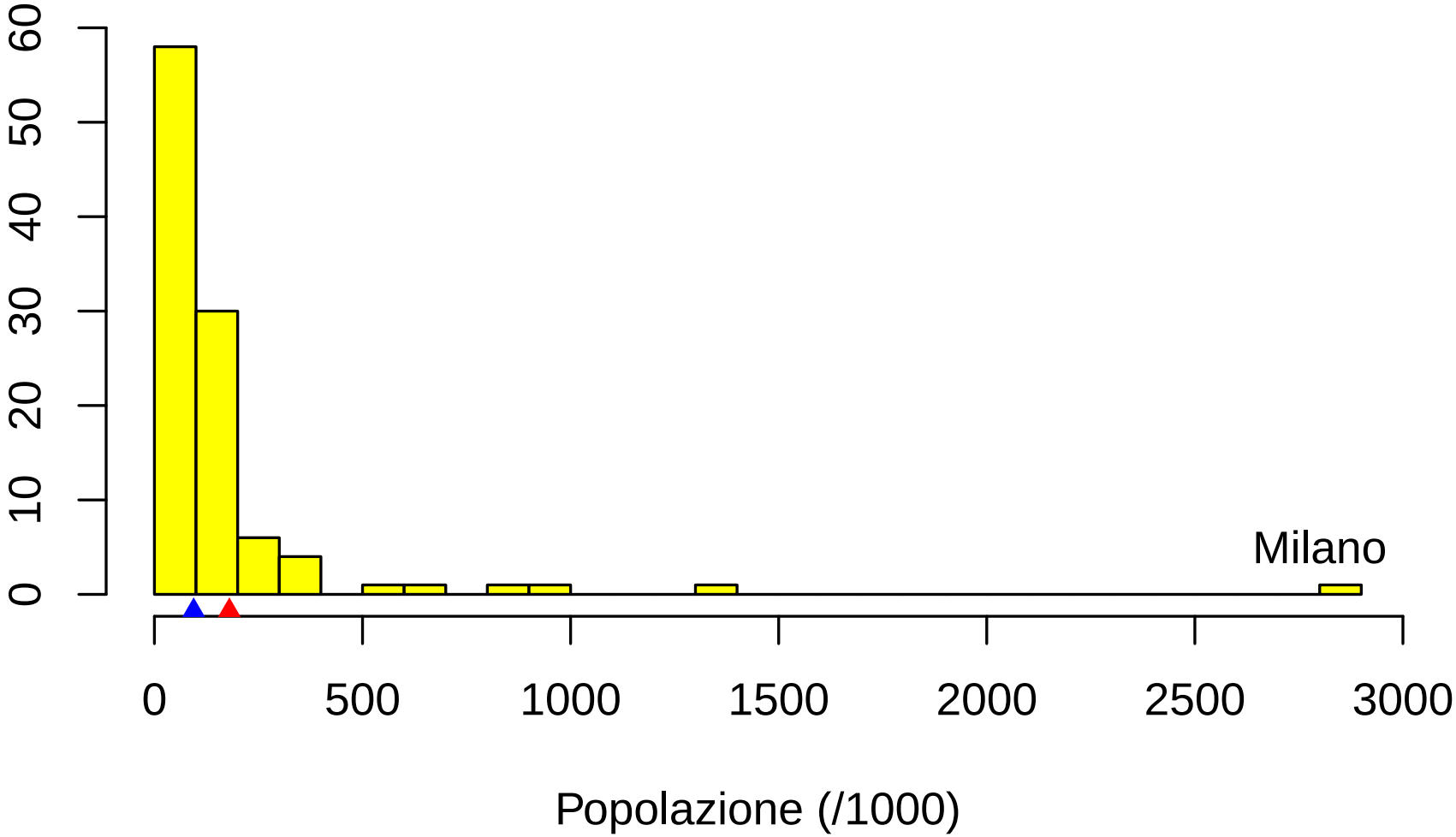
- Se ci sono valori **atipici** (outlier) è meglio la mediana.
- Esempio (in un asilo)

Età	Frequenza
3	9
50	1
Tot	10

- Età media $= (27 + 50)/10 = 7.7$ anni
- Età mediana $= 3$.
- La media è troppo sensibile ai valori anomali
- La mediana è più resistente e quindi fornisce un valore tipico migliore

Outlier in realtà

Un outlier è un dato che sta molto fuori rispetto al grosso dei dati



- Media = 180.000 abitanti. Mediana = 94.270 abitanti

Media complessiva e medie parziali

- In una classe ci sono 20 maschi e 10 femmine
- Il voto medio a statistica è 21 per i maschi e 28 per le femmine

Quant'è il voto medio complessivo?

Soluzione

Sesso	Voto medio	Frequenza	Totale voti
M	21	20	420
F	28	10	280
Tot		30	700

- Voto medio $= 700/30 = 23.3$
- Formula

$$\bar{x} = 21\left(\frac{20}{30}\right) + 28\left(\frac{10}{30}\right) = \bar{x}_M \cdot 0.67 + \bar{x}_F \cdot 0.33$$

Variabilità

- Senza variabilità non c'è statistica
- Misuriamo l'eterogeneità dei dati quantitativi
- Tre gruppi di persone di età diverse

Dati	Media	Variabilità
21, 21, 21, 21, 21, 21, 21	21	zero
14, 17, 20, 22, 23, 25, 26	21	presente
8, 10, 10, 20, 25, 32, 42	21	maggiore

Misure di variabilità

- **Varianza**: basata sugli scarti dalla media (maggiori gli scarti, maggiore la variabilità)
- **Scarto interquartile** basato sulla lunghezza di un intervallo che contiene il 50% dei dati

Varianza

- La varianza σ^2 è basata sulla somma degli scarti al quadrato dalla media

Dati	Scarti dalla media	Scarti^2
14	-7	49
17	-4	16
20	-1	1
22	1	1
23	2	4
25	4	16
26	5	25
Tot	0	112

$$\textit{varianza} = \sigma^2 = \frac{1}{N}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \frac{112}{7} = 16.$$

Varianza e deviazione standard

- La varianza ha il difetto di essere espressa in una **unità di misura al quadrato**
- età media = 21 anni,
- varianza dell'età = 16 anni al quadrato
- Perciò si rimedia facendone la radice quadrata chiamata **deviazione standard**

$$\textit{deviazione standard} = \sigma = \sqrt{\frac{1}{N}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}$$

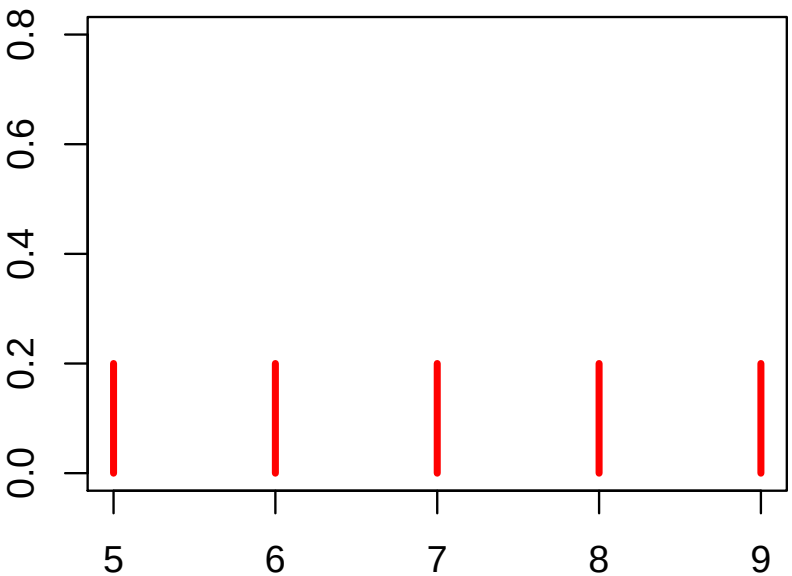
- Deviazione standard dell'età = $\sqrt{16} = 4$ anni.

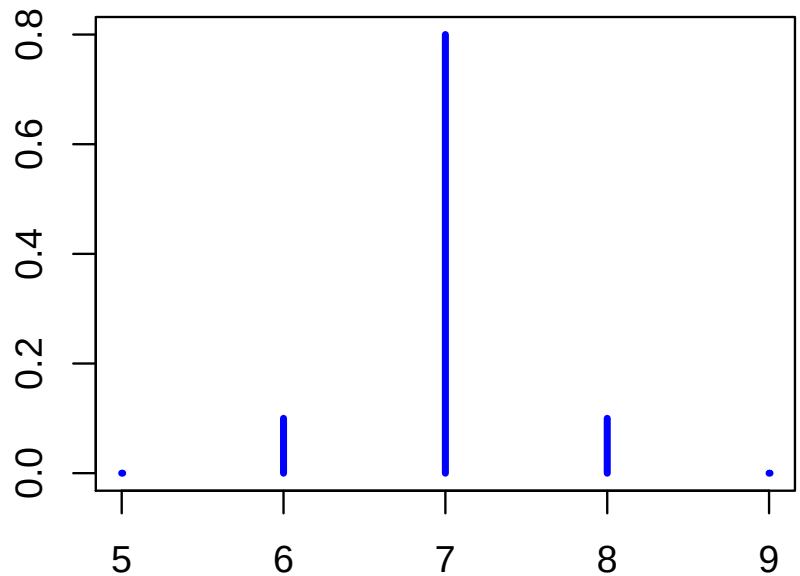
Valutare la variabilità

- È la cosa più importante in pratica
- L'inverso della variabilità è la **precisione**
- Significato
  - Finanza: rischio, volatilità
  - Ingegneria: precisione delle misure
  - Economia: accuratezza delle previsioni
  - Politica: incertezza dei sondaggi

Confronti di variabilità

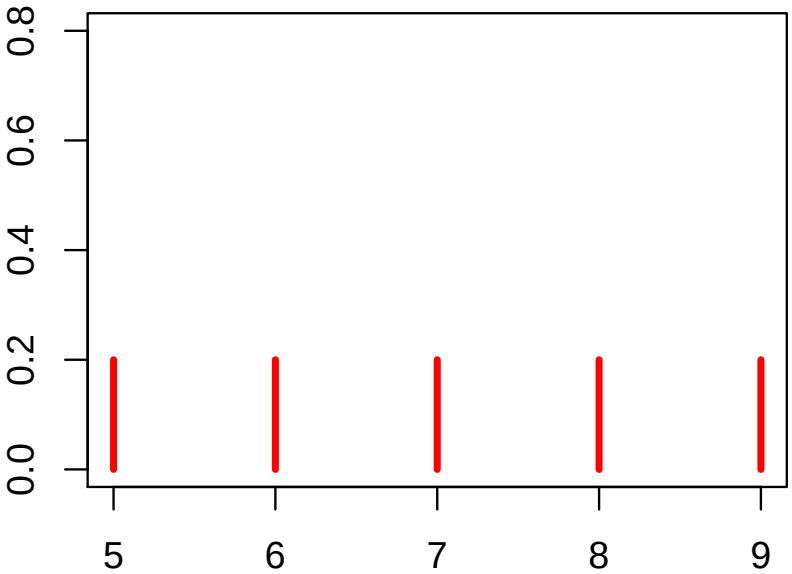
Fatturato	Gennaio freq. rel.	Giugno freq. rel.
5	0.2	0.0
6	0.2	0.1
7	0.2	0.8
8	0.2	0.1
9	0.2	0.0
Totale	1.0	1.0



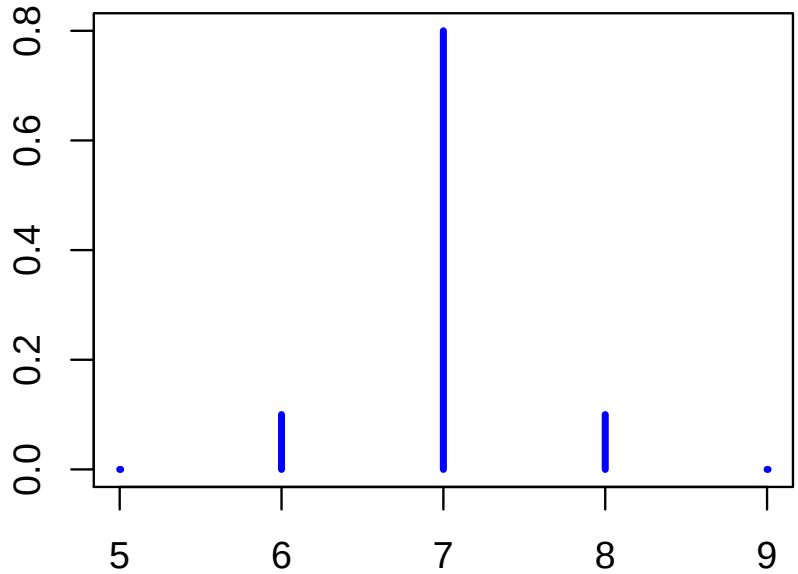


Confronti di variabilità

- Attenzione: la variabilità si valuta **in orizzontale**



Gennaio: Più variabile



Giugno: Meno variabile

Verificate: $\sigma^2_{gennaio} = 2$,$\sigma^2_{giugno} = 0.2$.

Come si calcola la varianza su una distribuzione di frequenza?

Fatturato	Frequenza
6	10
7	80
8	10
Totale	100

- Si calcola la media che è $\bar{x} = 7$ (come mai?)
- Si calcolano gli scarti al quadrato

Fatturato	Frequenza	Scarti	Scarti^2
6	10	-1	1
7	80	0	0
8	10	1	1
Totale	100		

(Continua)

- Si calcola il prodotto degli scarti al quadrato per la frequenza

Fatturato	Frequenza	Scarti	Scarti^2	Scarti^2 x freq.
6	10	-1	1	10
7	80	0	0	0
8	10	1	1	10
Totale	100			20

- La varianza è (k è il numero di classi)

$$\sigma^2 = \frac{1}{N}[(x_1 - \bar{x})^2 n_1 + \cdot + (x_k - \bar{x})^2 n_k] = 20/100 = 0.2$$

Varianza nella popolazione e nel campione

- Spessissimo i dati sono un **campione** estratto da una **popolazione**
- A seconda di come si considerano i dati la varianza viene indicata e calcolata diversamente
- Se i dati sono relativi a una popolazione di N unità

$$varianza = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Se i dati sono relativi a un campione di dimensione $n < N$ allora

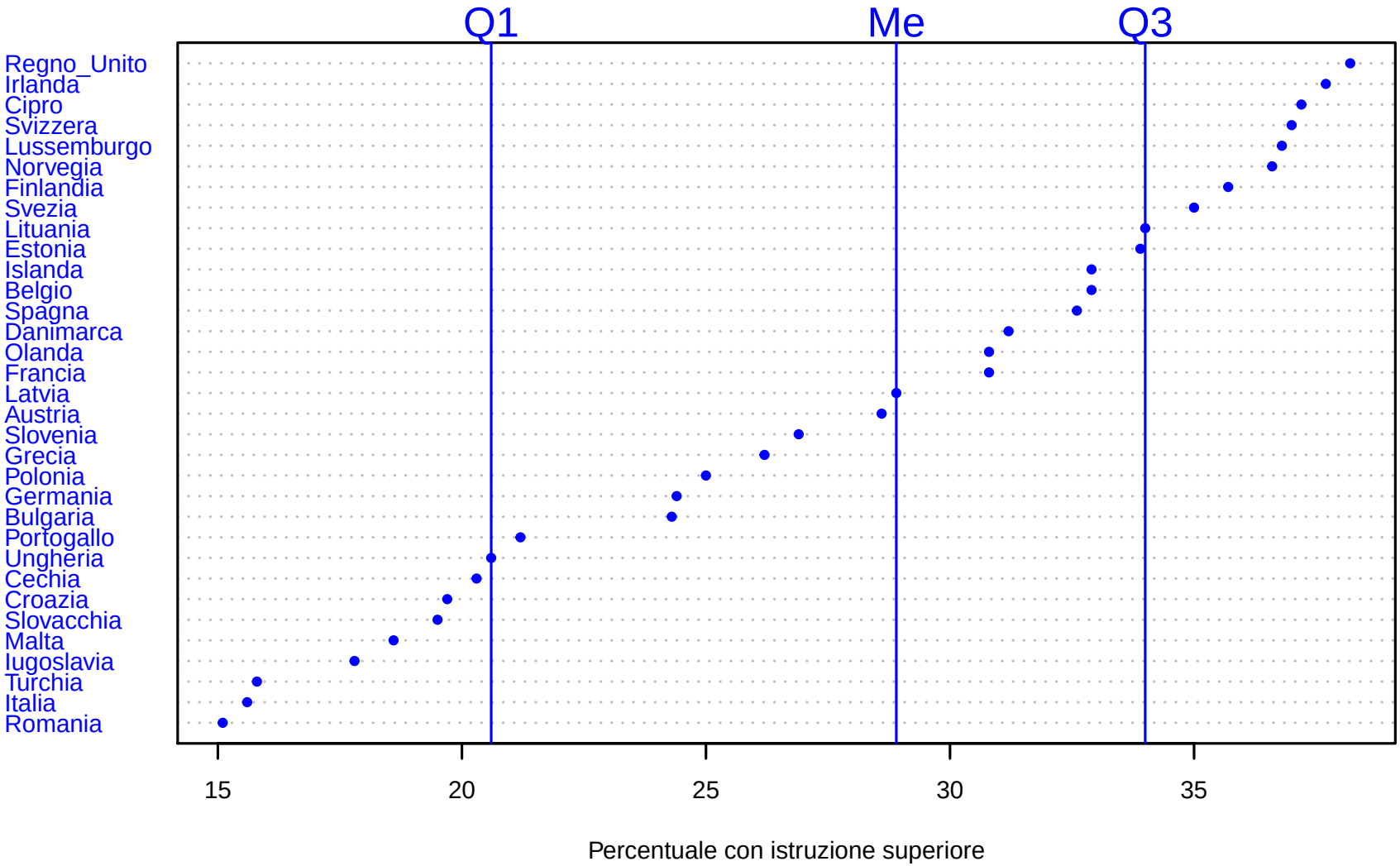
$$varianza = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Un indice di variabilità basato sull'ordinamento

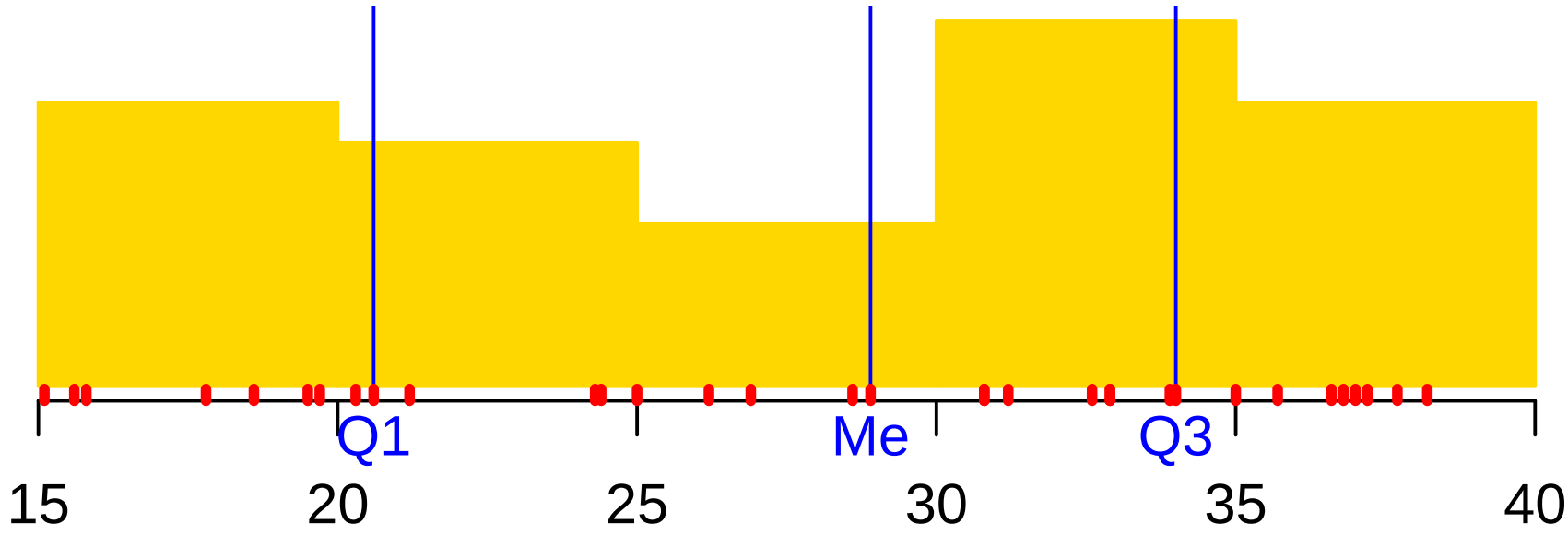
- **Quartili**
  - Il primo quartile Q_1 è il valore che ha prima di sé il 25% dei dati
  - Il secondo quartile Q_2 è il valore che ha prima di sé il 50% dei dati cioè è la mediana
  - Il terzo di quartile Q_3 è il valore che ha prima di sé il 75% dei dati
- **Differenza interquartile**
  - È la differenza $Q_3 - Q_1$.

Esempio

- percentuale di popolazione con educazione terziaria nei paesi europei



Esempio (segue)



$$IQR = 13.4$$

Come si calcolano i quartili

- **Ordinare i dati!**
- Q_1 è il dato della osservazione numero $(n+1)(1/4)$
- Q_3 è il dato della osservazione numero $(n+1)(3/4)$

Problema

- Come nel caso della mediana se n non è divisibile per 4 $(n+1)\frac{1}{4}$ o $(n+1)\frac{1}{4}$ sono numeri con la virgola.
- In quel caso si fa la semisomma dei dati delle osservazioni corrispondenti all'intero inferiore e superiore.

Calcolare i posti dei quartili

- $n = 23$: posto del primo quartile $= 24/4 = 6$.
- $n = 29$: posto del primo quartile $= 30/4 = 7.5$. Prendi la 7ma e l'8va
- $n = 20$: posto del primo quartile $= 21/4 = 5.25$. Prendi la 5a e la 6a

Esempio delle percentuali di istruzione superiore

```
15.1 15.6 15.8 17.8 18.6 19.5 19.7 20.3 20.6 21.2 24.3
24.4 25.0 26.2 26.9 28.6 28.9 30.8 30.8 31.2 32.6 32.9
32.9 33.9 34.0 35.0 35.7 36.6 36.8 37.0 37.2 37.7 38.2
```

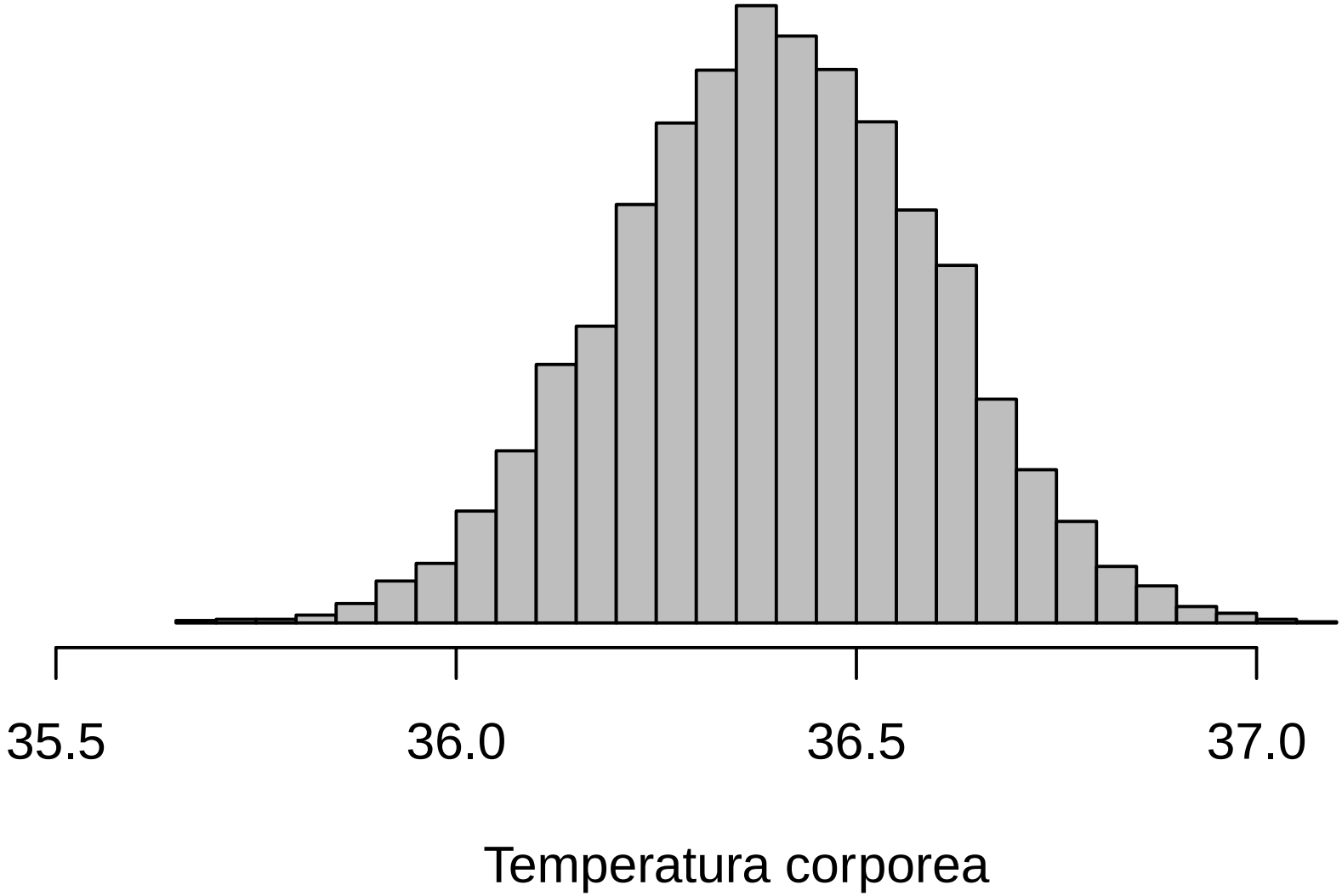
- Già ordinati, $n = 33$.
- Posto di Q_1: $(1 + 33)/4 = 8.5 \longrightarrow$ 8 e 9
- Posto di Q_2: $(1 + 33)/2 = 17$
- Posto di Q_3: $(1 + 33) \cdot (3/4) = 25.5 \longrightarrow$ 24 e 25

Quindi

- $Q_1 = (20.3 + 20.6)/2 = 20.45$
- $Q_3 = (33.9 + 34.0)/2 = 33.95$

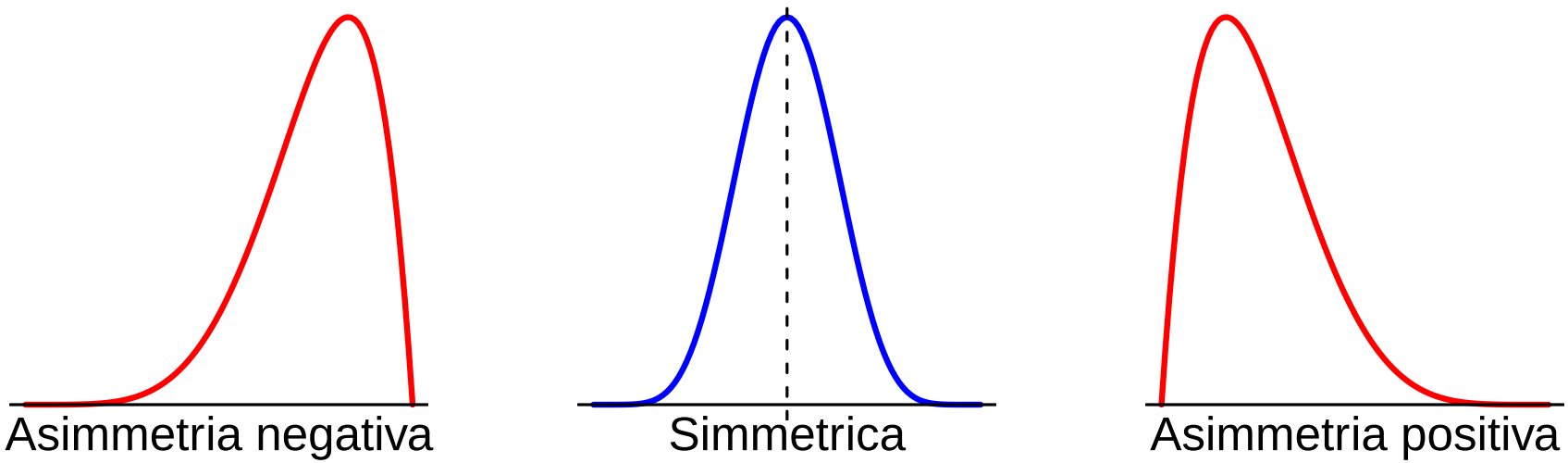
Dati con una distribuzione di frequenza “normale”

- Esempio: X = temperatura corporea di 10000 adulti sani



Forma della distribuzione

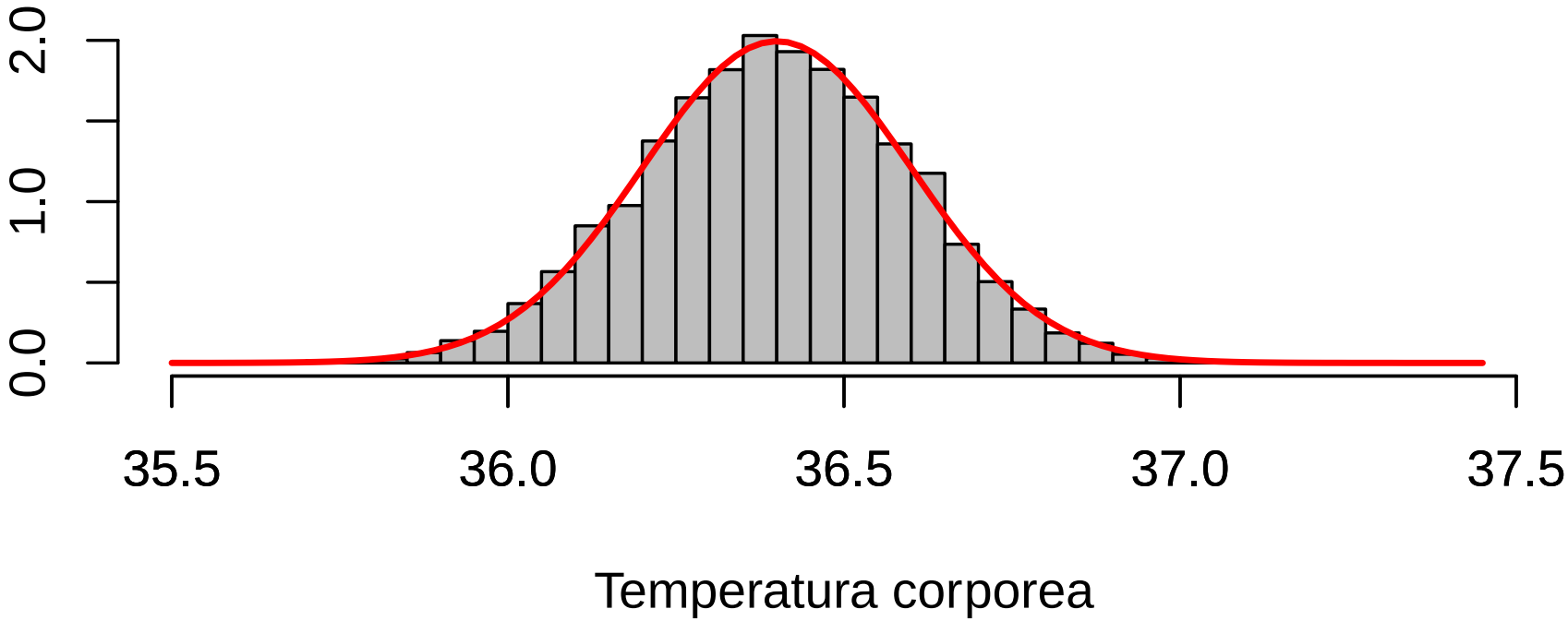
- È simmetrica e unimodale (con un unico massimo)



- Di forma “campanulare”

Curva normale o “Gaussiana”

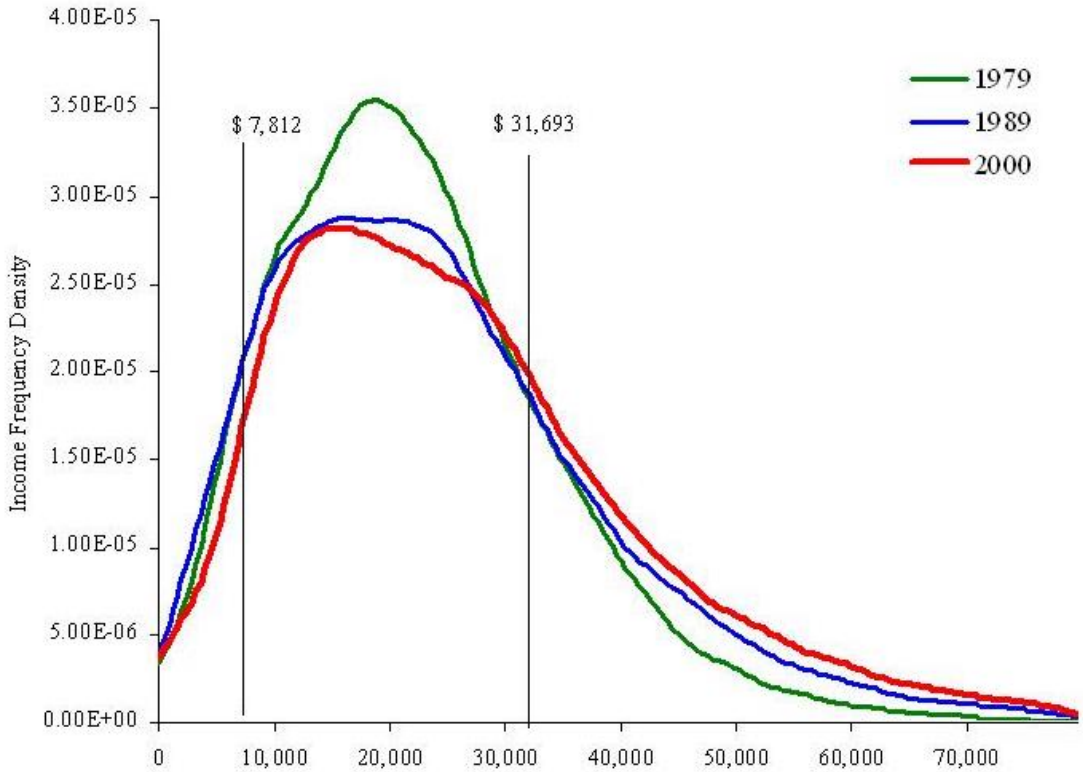
- La curva normale è il grafico di una funzione matematica



- La curva è una sorta di “istogramma ideale”
- L’area totale **è uguale a 1**.

Dati con asimmetria positiva

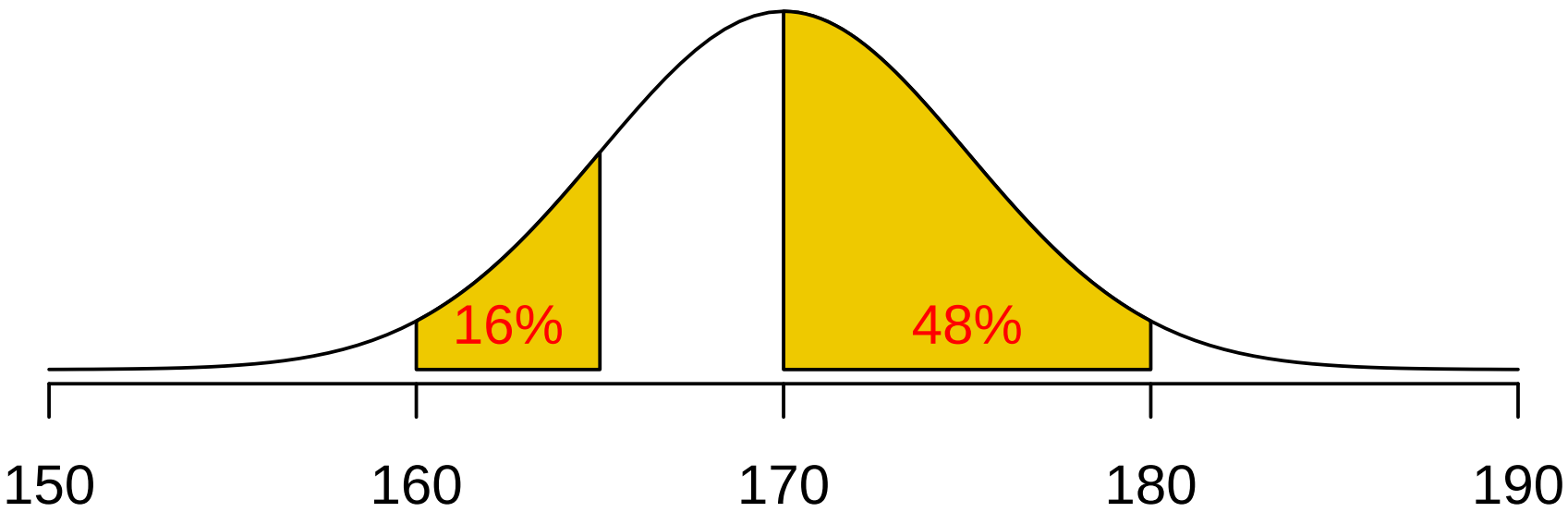
- Distribuzione dei redditi (USA)



- Si riconosce perché la **media è maggiore della mediana**

Proporzione di individui che appartengono a un intervallo

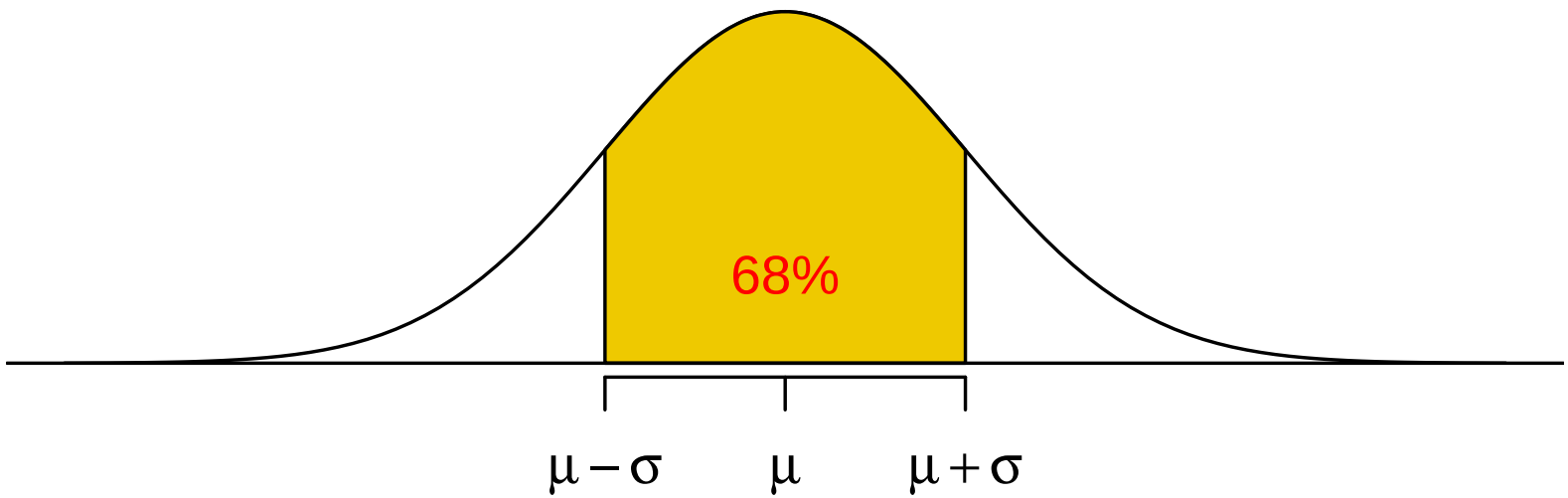
- Gli istogrammi hanno un area totale uguale a 1
- L'area sotto la curva compresa in un intervallo (a, b) è la frequenza relativa (la proporzione) di unità che hanno $a \leq X \leq b$



Regola empirica = Regola per la curva normale

- Se l'istogramma ha una forma "normale" (cioè Gaussiana) l'intervallo

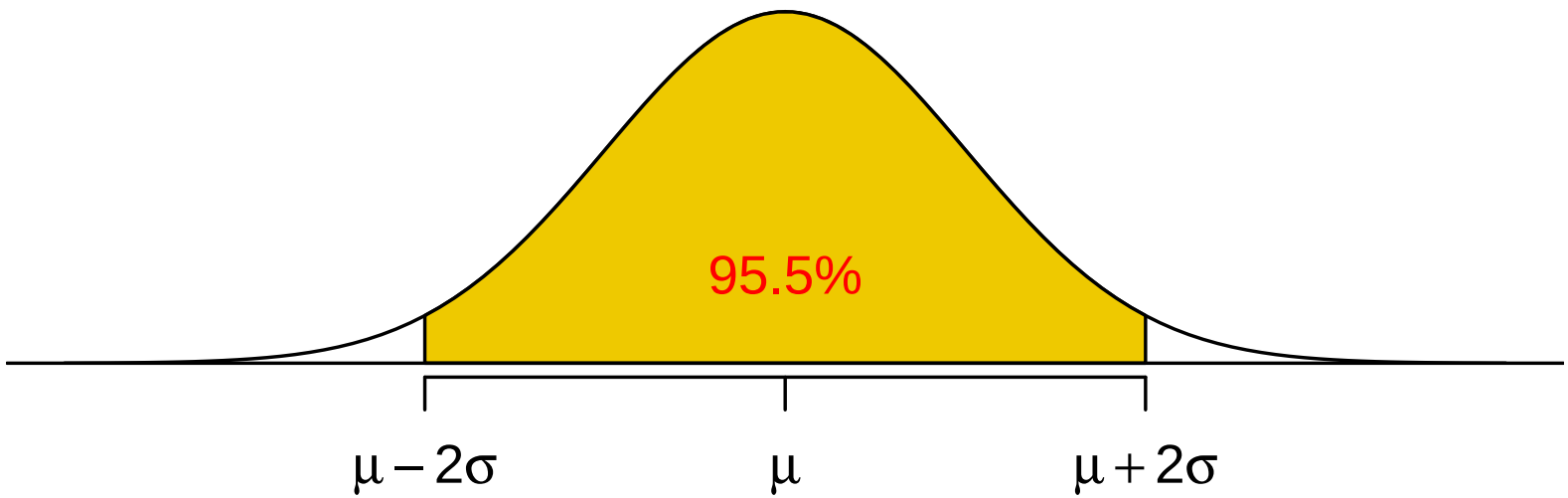
$$(\mu - \sigma, \quad \mu + \sigma) \quad \text{contiene il 68\% dei dati}$$



Regola empirica = Regola per la curva normale

- Se l'istogramma ha una forma "normale" (cioè Gaussiana) l'intervallo

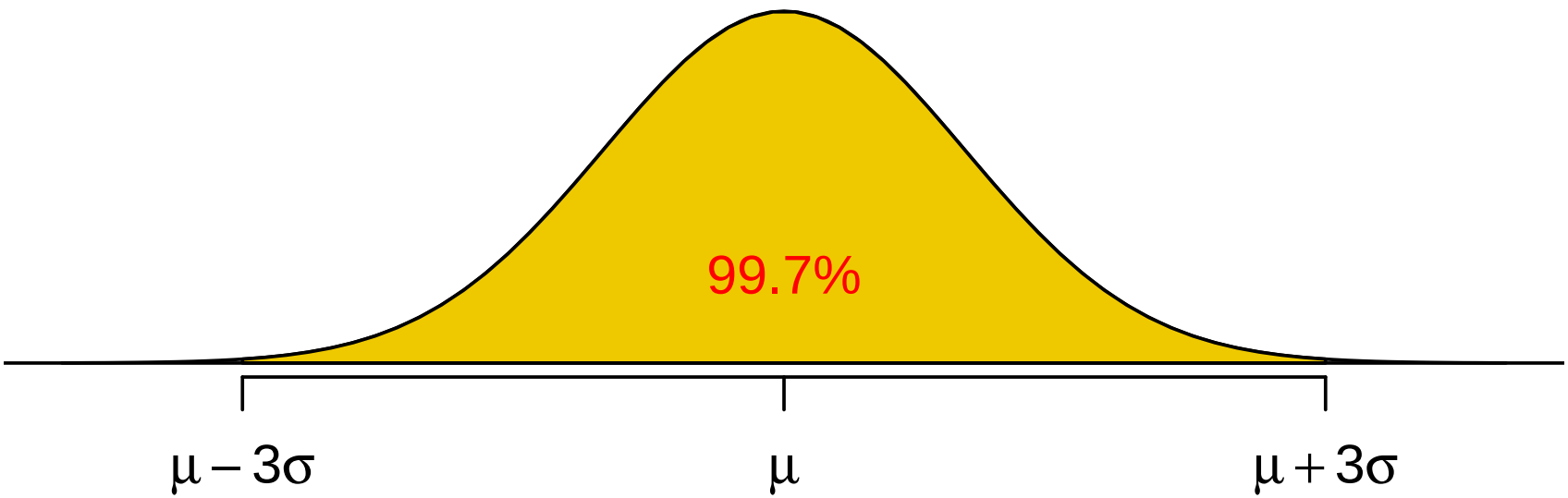
$$(\mu - 2\sigma, \quad \mu + 2\sigma) \quad \text{contiene il 95.5\% dei dati}$$



Regola empirica = Regola per la curva normale

- Se l'istogramma ha una forma "normale" (cioè Gaussiana) l'intervallo

$$(\mu - 3\sigma, \quad \mu + 3\sigma) \quad \text{contiene il } 99.7\% \text{ dei dati}$$



Morale della favola

- La varianza è una specie di metro per valutare la bontà della media.
- Se la variabile ha una distribuzione di frequenza di forma normale, dalla deviazione standard possiamo dedurre subito delle informazioni utili con la regola empirica
- Esempio: Il rendimento di un titolo ha media 5% e deviazione standard 2%
- Automaticamente sappiamo che il 68% dei rendimenti starà tra il 3% e il 7%
- Analogamente il 95% dei rendimenti starà tra l'1% e il 9%.

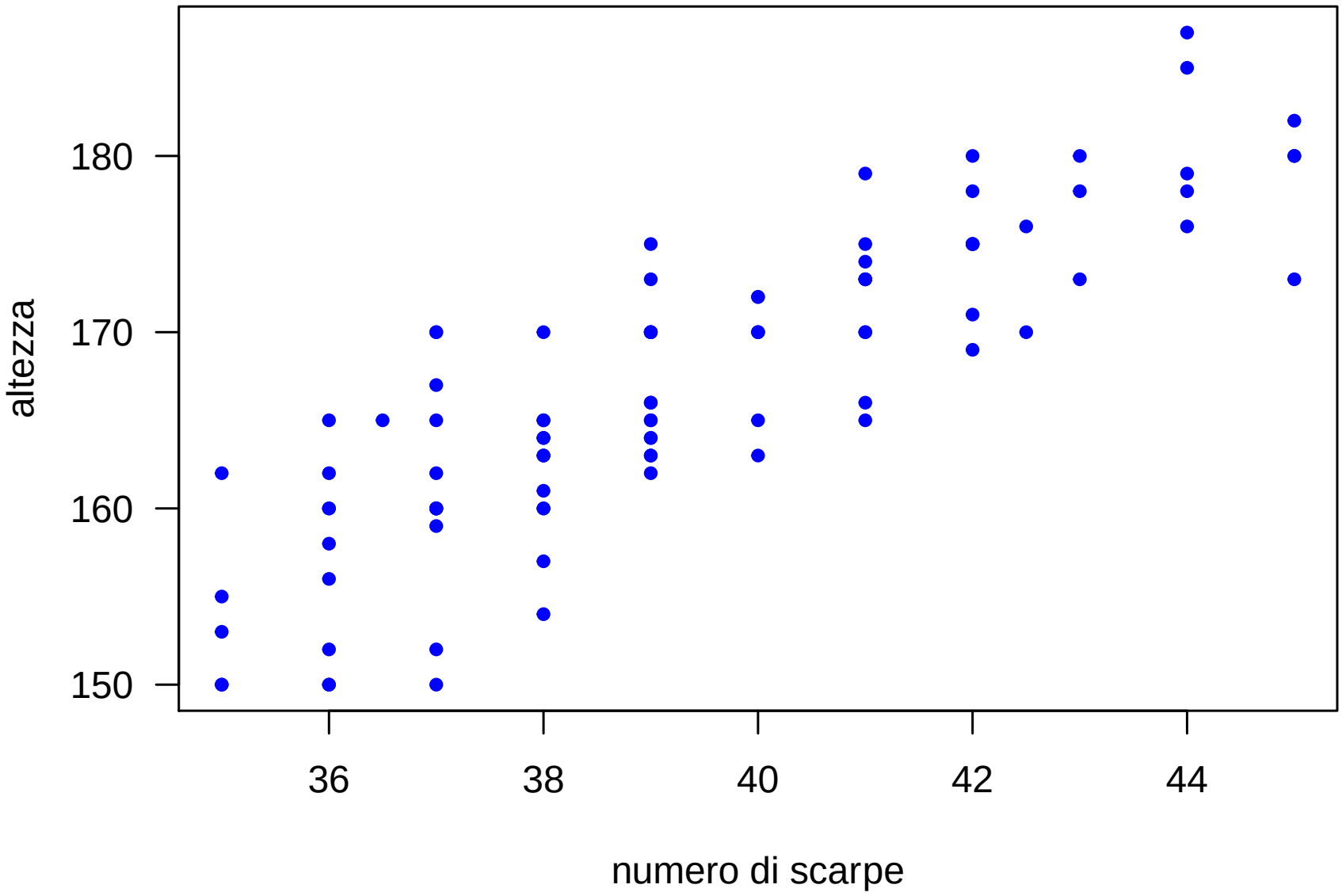
Associazione tra due variabili quantitative

- C'è associazione tra
  - Altezza (cm)
  - Numero di scarpe ?

	sesso	scarpe	altezza
1	m	39	170
2	f	40	170
3	f	37	162
4	f	38	160
5	f	38	157
6	m	42	169

etc. . .

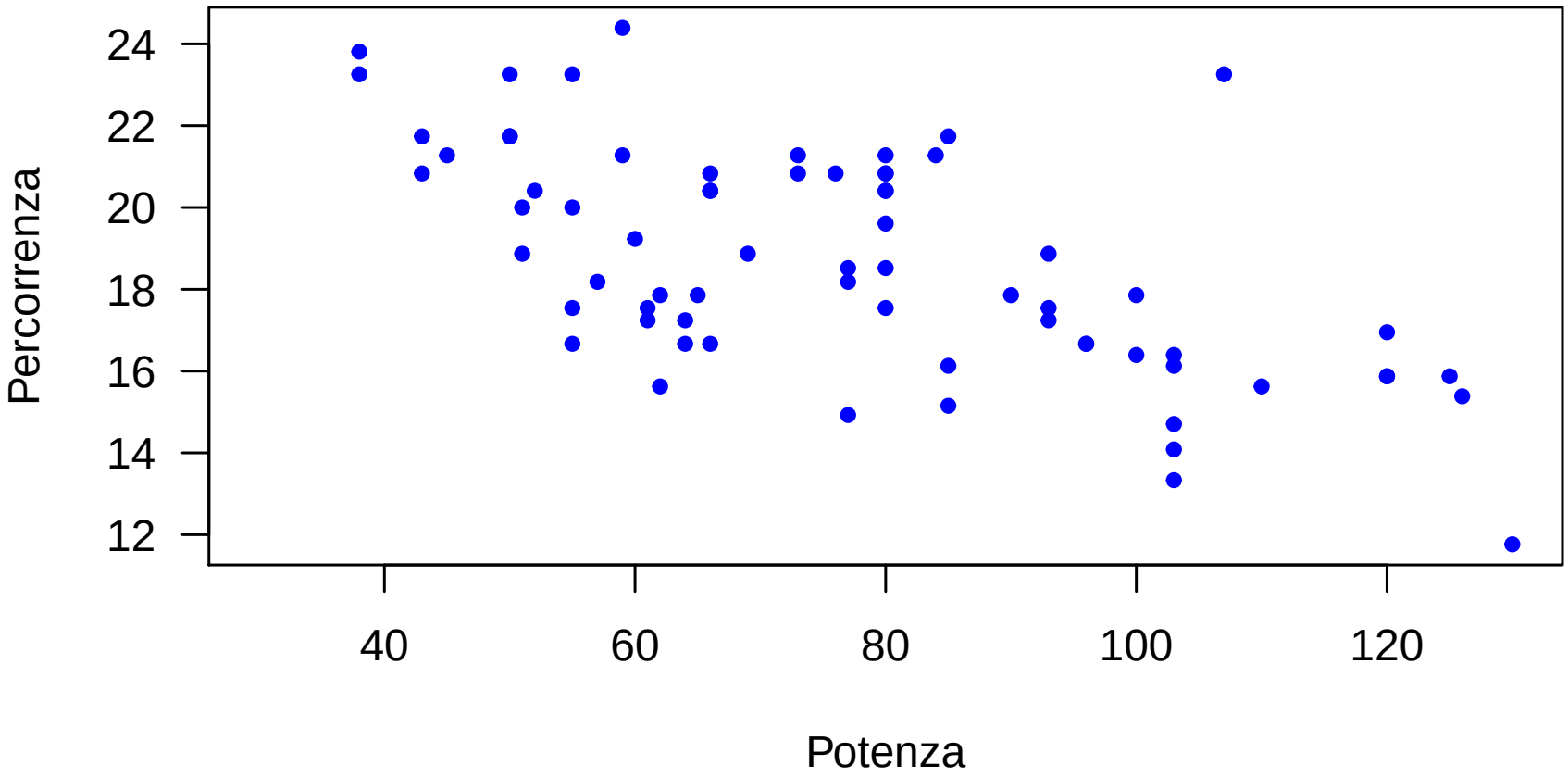
Scatter



- C'è evidenza di una relazione crescente tra altezza e numero di scarpa

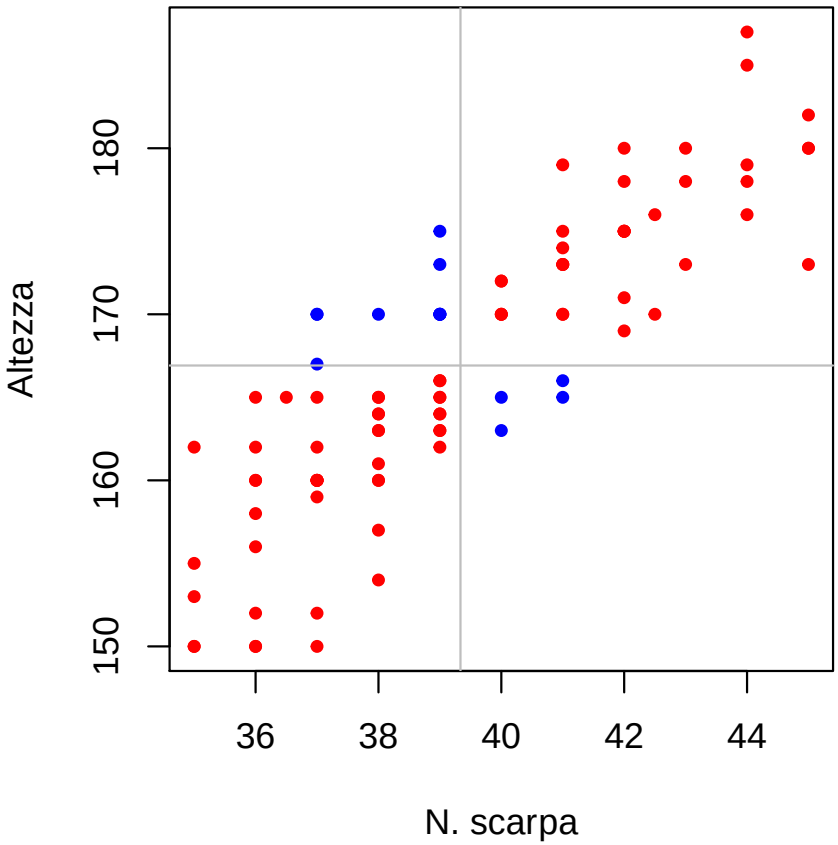
Associazione positiva o negativa?

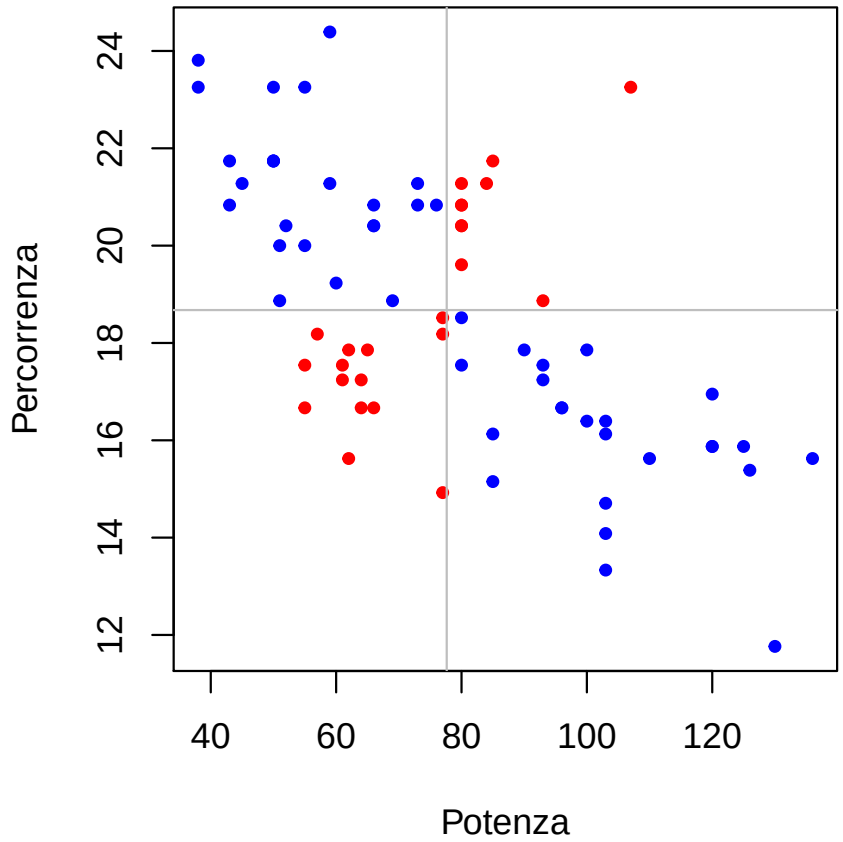
- C'è associazione tra percorrenza di un auto (km/l) e potenza (kW = 1.36 CV)



- C'è evidenza di una relazione inversa (negativa)

Come si fa a valutare?





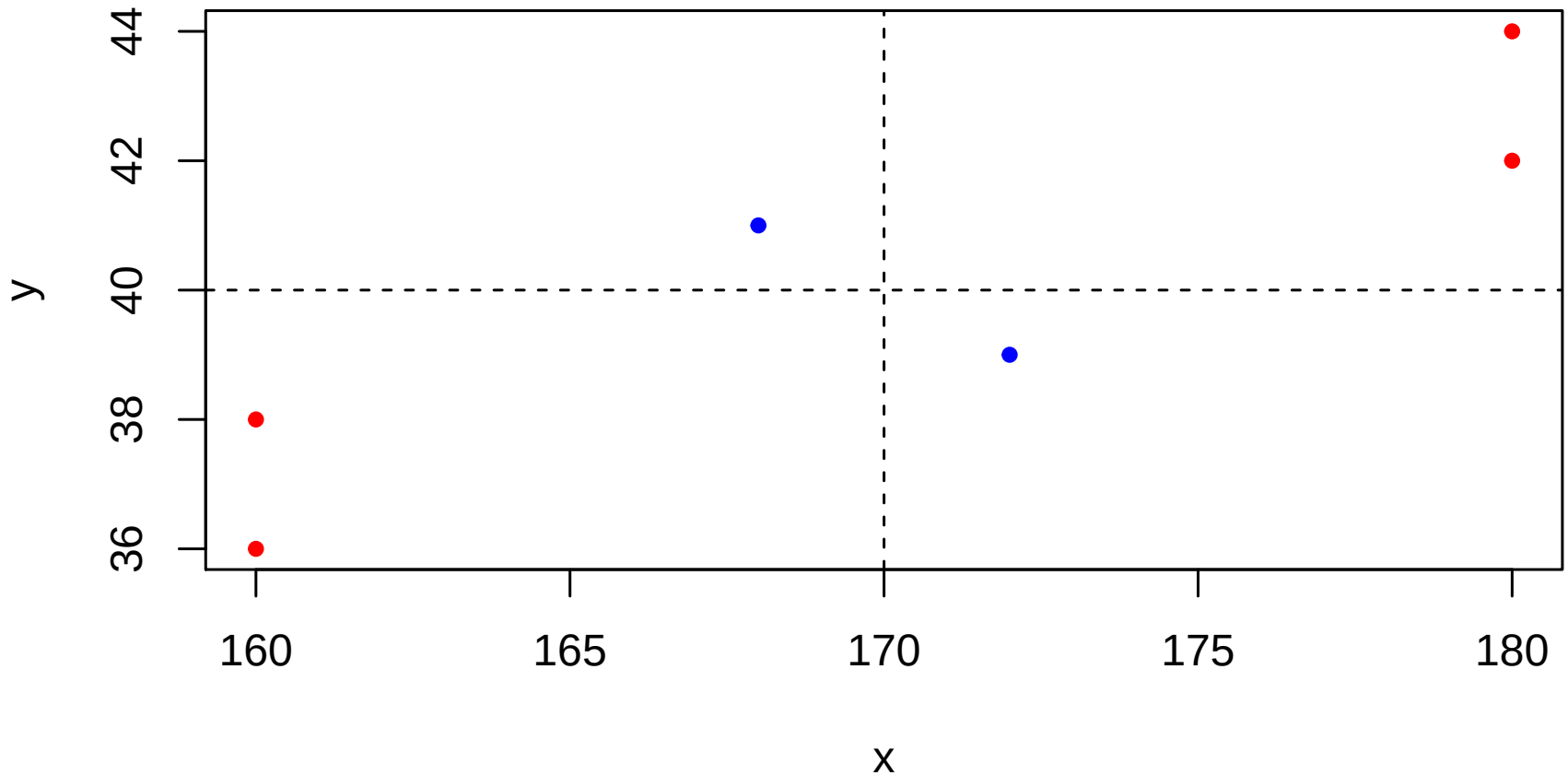
- Associazione positiva se vince il I e III quadrante
- Associazione negativa se vince il II e IV quadrante

Indici fondamentali

- La **covarianza**
- Il coefficiente di **correlazione lineare**

Covarianza: esempio

i	1	2	3	4	5	6	Media
X	160	160	168	172	180	180	170
Y	36	38	41	39	42	44	40



Calcolo

$$\sigma_{xy} = \frac{1}{N}[(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})]$$

i	1	2	3	4	5	6	Media
X	160	160	168	172	180	180	170
Y	36	38	41	39	42	44	40
scarti x	-10	-10	-2	2	10	10	0
scarti y	-4	-2	1	-1	2	4	0
Prodotto	40	20	-2	-2	20	40	19.3

Covarianza $= \sigma_{xy} = 116/6 = 19.3$: associazione positiva

Associazione positiva o negativa

- **Concordanza**
  - Scarti dalla media con lo stesso segno ($++$ o $--$)
  - Prodotto degli scarti positivo
- **Discordanza**
  - Scarti dalla media con segno opposto ($+-$ o $-+$)
  - Prodotto degli scarti negativo
- Covarianza = media dei prodotti degli scarti.
- Indice positivo se prevalgono i concordanti indice negativo se prevalgono i discordanti

Formula generale

- Se i dati sono un campione di dimensione n

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

- Sei dati sono una popolazione di dimensione N

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

- > 0 se a valori di X grandi corrispondono Y grandi e a X piccoli corrispondono Y piccoli.
- < 0 se in media a X grandi corrispondono Y piccoli e a X piccoli corrispondono Y grandi.

Coefficiente di correlazione

- La covarianza **dipende dall'unità di misura delle variabili**
- Esempio:

$$\text{cov}(\textit{altezza (cm)}, \textit{n.scarpe}) = 19.3$$

$$\text{cov}(\textit{altezza (metri)}, \textit{n.scarpe}) = 0.193$$

- Per misurare l'associazione in modo **invariante** si usa il coefficiente di correlazione

Formula generale

$$\text{corr}(X, Y) = \frac{\textit{covarianza}(X, Y)}{\textit{dev. st}(X), \textit{dev.st}(Y)} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Qualsiasi denominatore si usi, si ottiene la formula seguente

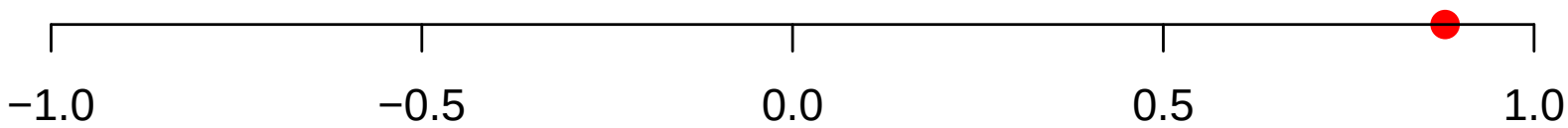
$$\text{corr}(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Correlazione tra n. scarpe e altezza

Usando denominatore $N = 6$

- covarianza $= 19.3$,
- varianza altezza $= 68$
- varianza n.scarpe $= 7$, dunque

$$\textit{correlazione} = \frac{19.3}{\sqrt{68 \cdot 7}} = 0.88$$



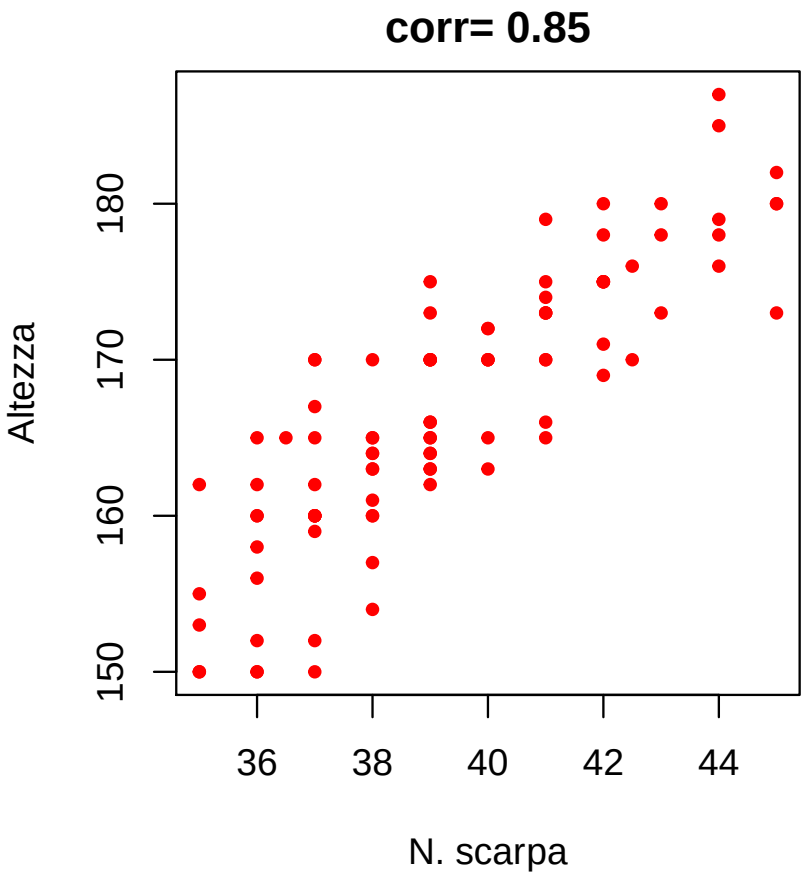
correlazione

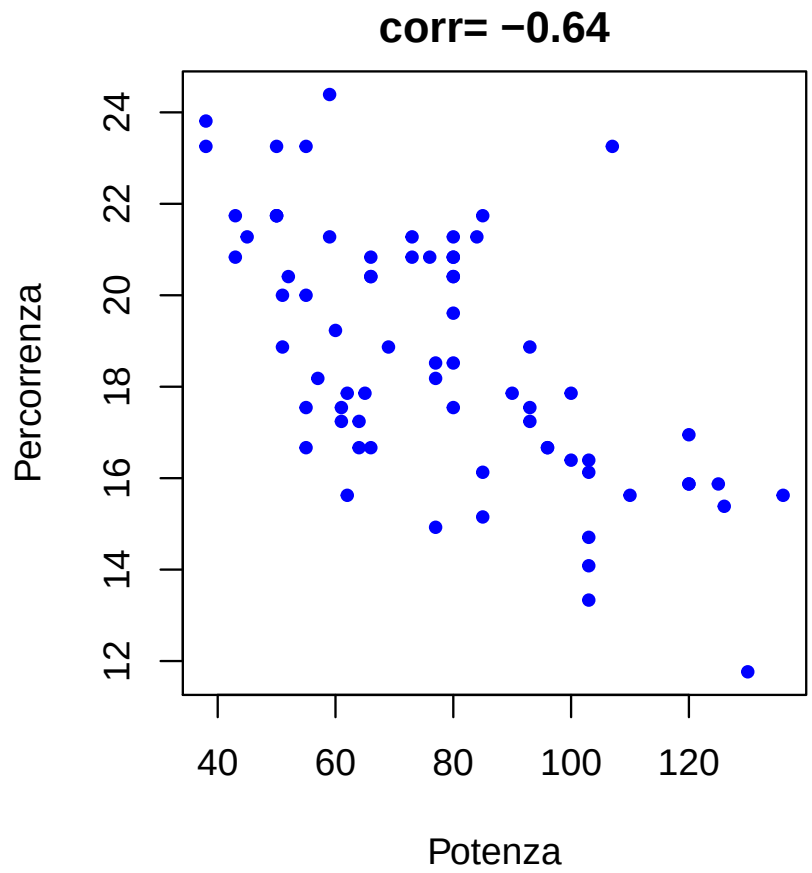
Proprietà del coefficiente di correlazione

- Introdotto da Karl Pearson
- È un indice simmetrico (le due variabili sullo stesso piano)
- È un numero puro, cioè non ha unità di misura
- È sempre compreso nell'intervallo [-1, 1]
- È tanto più vicino a 1 o a -1 quanto più le variabili X e Y tendono a essere allineate

Per questo si dice che il coefficiente di correlazione è un indice di **associazione lineare**

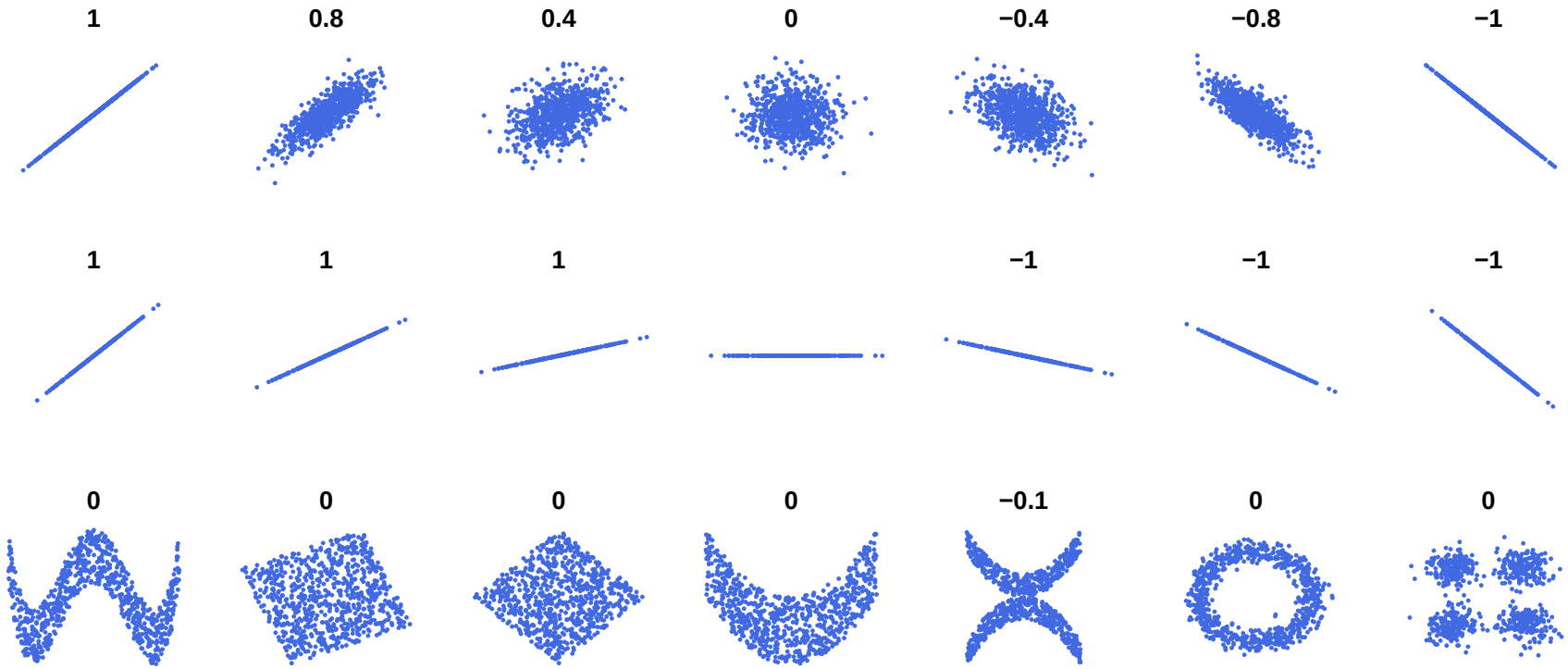
Esempi


corr= 0.85
Altezza
150
160
170
180
36
38
40
42
44
N. scarpa


corr= −0.64
Percorrenza
12
14
16
18
20
22
24
40
60
80
100
120
Potenza

Imparare a riconoscere il grado di correlazione

Il grafico è tratto da **Wikipedia**.



Altri link interessanti

- Disegnare i punti e vedere la correlazione **Applet**
- Indovina la correlazione (giochino virale!) **Guess the correlation**