Capitolo 3

Giovanni Marchetti

6 marzo 2019

Sommario

- Misure di tendenza centrale
- Misure di variabilità
- Misura della relazione tra variabili quantitative
- Retta dei minimi quadrati

Misure di tendenza centrale

Misure di tendenza centrale

Sono indici che rappresentano un **valore tipico** attorno a cui si addensano i dati

Tale valore si può prendere come un sommario approssimato di tutti i dati

- Media aritmetica
- Moda
- Mediana

Notazione

I dati su una variabile quantitativa X sono indicati con

$$x_1, x_2, \ldots, x_n$$

La loro **somma** la indichiamo con

$$\text{somma}(X) = x_1 + \cdots + x_n = \sum_{i=1}^{n} x_i$$

```
somma = 0
for i = 1:n
    somma = somma + x[i]
end
```

Media aritmetica

- La **media aritmetica** della variabikle X è definita da

$$\text{media}(X) = (x_1 + \cdots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Ha un'interpretazione semplice: è ottenuta **equiripartendo** il totale della variabile tra le unità.

Esempio

- Il voto medio alla maturità: media(Voto) $=$ 77.2.
- Il voto medio al punteggio al test: media (test) =13.3.
- Nota: **non si calcolano medie di dati qualitativi**

Media su un campione e su una popolazione

Un **campione** è un insieme di dati raccolti da un insieme più grande detto **popolazione**

La media si denota diversamente:

- **in un campione**: $\bar{x}$
- **nella popolazione**: μ

Ovviamente $\bar{x}$ è una **stima** di μ

Alcune proprietà utili

- È **sempre compresa tra il minimo e il massimo**

$$60 \leq \text{ Voto medio Maturità } \leq 100$$

- La **somma degli scarti dalla media è sempre zero**:

Voti :	70	60	70	100	80	$Media = 76$
Scarti :	-6	-16	-6	24	4	$Somma = 0$

x_i :	x_1	x_2	x_3	x_4	x_5	
Scarti :	$x_1 - \bar{x}$	$x_2 - \bar{x}$	$x_3 - \bar{x}$	$x_4 - \bar{x}$	$x_5 - \bar{x}$	$Somma = 0$

- Infatti

$$Somma\ scarti = (x_1 + \cdots + x_5) - 5\bar{x} = 5\bar{x} - 5\bar{x} = 0.$$

La media è il valore più vicino ai dati

- Soddisfa il **criterio dei minimi quadrati**
- Consideriamo i dati precedenti

$$\textit{Voti}: \quad 70 \quad 60 \quad 70 \quad 100 \quad 80 \quad \textit{Media} = 76$$

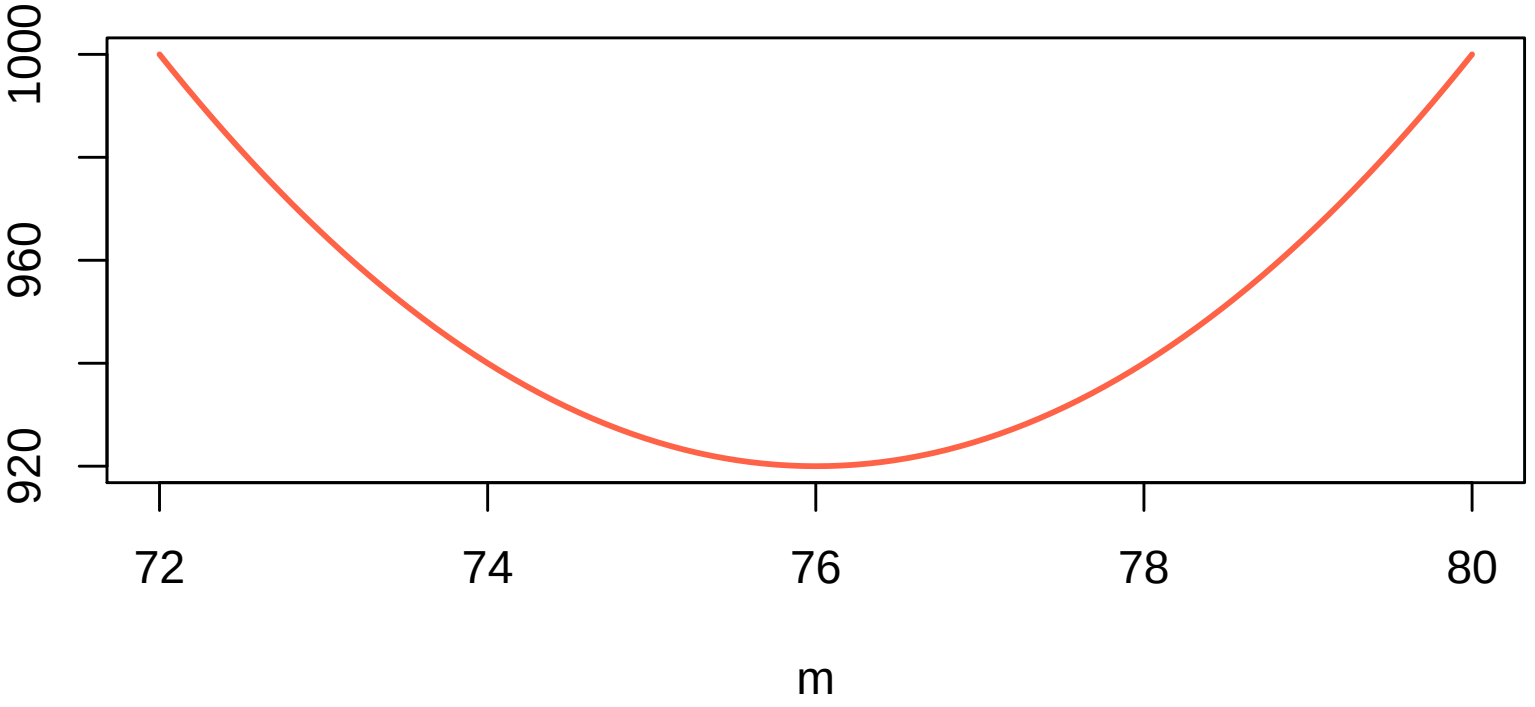
- Il valore m che minimizza

$$\textit{dist}(m) = (70-m)^2+(60-m)^2+(70-m)^2+(100-m)^2+(80-m)^2$$

è proprio la media aritmetica $m = 76$!

Dimostrazione

Infatti $dist(m) = 29800 - 760m + 5m^2$ è una parabola col vertice in $m = 76$



Media calcolata da una distribuzione di frequenze

Esercizio 3.30 Distribuzione di polizze assicurative secondo il numero di richieste di indennizzo nell'ultimo anno

Richieste indennizzo	Numero polizze
0	21
1	13
2	5
3	4
4	2
5	3
6	2
Totale	

Qual è il numero medio annuo di richieste di indennizzo ?

Soluzione

n. medio di richieste = numero richieste / numero di polizze

Richieste indennizzo	Numero polizze	Totale richieste
0	21	0
1	13	13
2	5	10
3	4	12
4	2	8
5	3	15
6	2	12
Totale	50	70

Quindi: media(richieste)$= \frac{70}{50} = 1.4$

Formula della media per distribuzioni di frequenza

X : modalità	*frequenza*
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k
Totale	n

Quindi

$$\text{media}(X) = \frac{x_1 n_1 + x_2 n_2 + \cdots x_k n_k}{n}.$$

Media di tassi di interesse

Supponiamo di avere un capitale di 100 Euro, investito con rendimenti

Anno	*Tasso*	*Capitale*
0	–	100
1	0.25	$100 \times 1.25 = 125$
2	0.20	$125 \times 1.20 = 150$
3	0.02	$150 \times 1.02 = 153$

- Qual è il tasso medio?
- È giusto calcolare $(0.25 + 0.20 + 0.02)/3 = 0.16$?
- **No perché i tassi non sono additivi**

Soluzione

- Qual è il tasso costante che applicato al capitale di 100 lo fa diventare 153 alla fine dei tre anni?
- Se g è questo tasso, dopo 3 anni a interesse composto il capitale è

$$100 \cdot (1+g)^3$$

- Quindi per trovare g bisogna risolvere l'equazione

$$100 \cdot (1+g)^3 = 153$$

- Soluzione

$$1+g = \sqrt[3]{153/100} = 1.152 \qquad \therefore \; g = 0.152$$

- Il tasso medio ottenuto con una **media geometrica**.

Media geometrica

- La **media geometrica** è un indice di tendenza centrale per dati **positivi che hanno natura moltiplicativa**
- Formula

$$\text{media-geometrica}(X) = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}.$$

- I montanti sono moltiplicative

$$1.25 \times 1.20 \times 1.02 = 1.53$$

- quindi

$$1 + g = \sqrt[3]{1.25 \times 1.20 \times 1.02} = 1.152$$

Mediana

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

Come si calcola

- Supponiamo di avere 9 individui adulti di altezze diverse

$$x_i : 183, 170, 175, 172, 169, 185, 170, 190, 165 \qquad \bar{x} = 175.4$$

- I dati ordinati sono

$$165, 169, 170, 170, 172, 175, 183, 185, 190$$

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

$$165, 169, 170, 170, \mathbf{172}, 175, 183, 185, 190$$

Mediana = 172 cm

Pari e dispari

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

$$183, 170, 175, 172, 169, 185, 170, 190 \qquad (n = 8)$$

- I dati ordinati

$$169, 170, 170, \mathbf{172}, \mathbf{175}, 183, 185, 190$$

- La mediana è

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

Che cos'è la mediana?

- È un indice di tendenza centrale
- **Interpretazione**: il 50% dei dati dei dati ha valore inferiori alla mediana, l'altro 50% ha valore superiore
- Per quanto riguarda il calcolo **la mediana va calcolata dopo aver ordinato i dati** e selezionando il valore alla massima profondità

Mediana da una distribuzione di frequenza

Calcolare la mediana dell'altezza

Altezza	Frequenza	Frequenza cumulata
160	10	10
165	45	55
170	65	120
175	50	170
180	30	200
Totale	200	-

- I valori centrali sono 100 e 101
- Appartengono entrambi al gruppo degli individui alti 170 cm
- Mediana = 170 cm

Quando media e mediana coincidono?

Se i dati hanno una distribuzione **simmetrica**, la media e la mediana sono uguali

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 classe delle persone alte 170 cm

È meglio la media o la mediana?

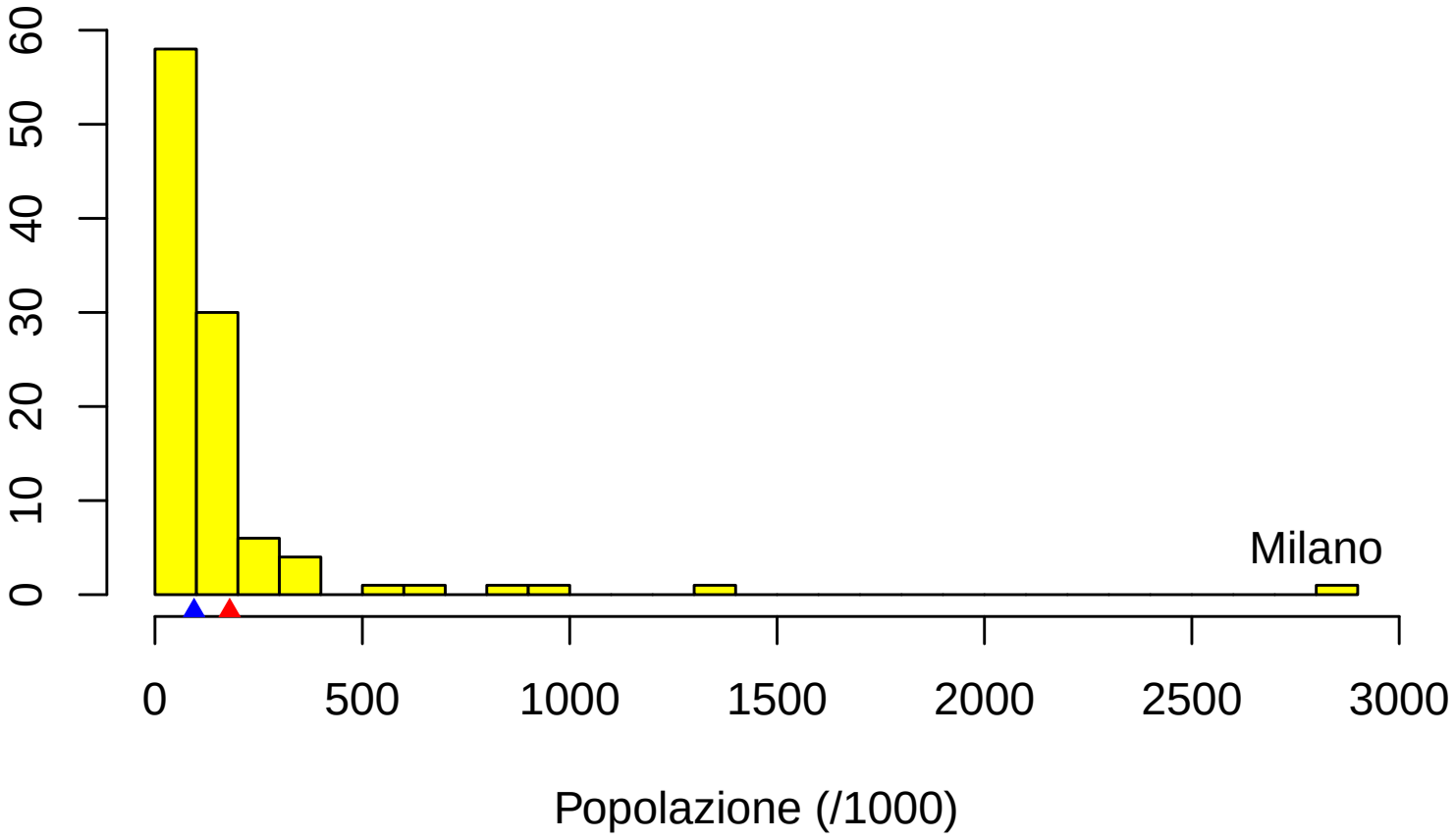
- Se ci **outlier** (valori atipici) è meglio la mediana.
- Esempio: in un asilo

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 atipici
- La mediana è più **resistente** e quindi fornisce un valore tipico migliore

Outlier in realtà

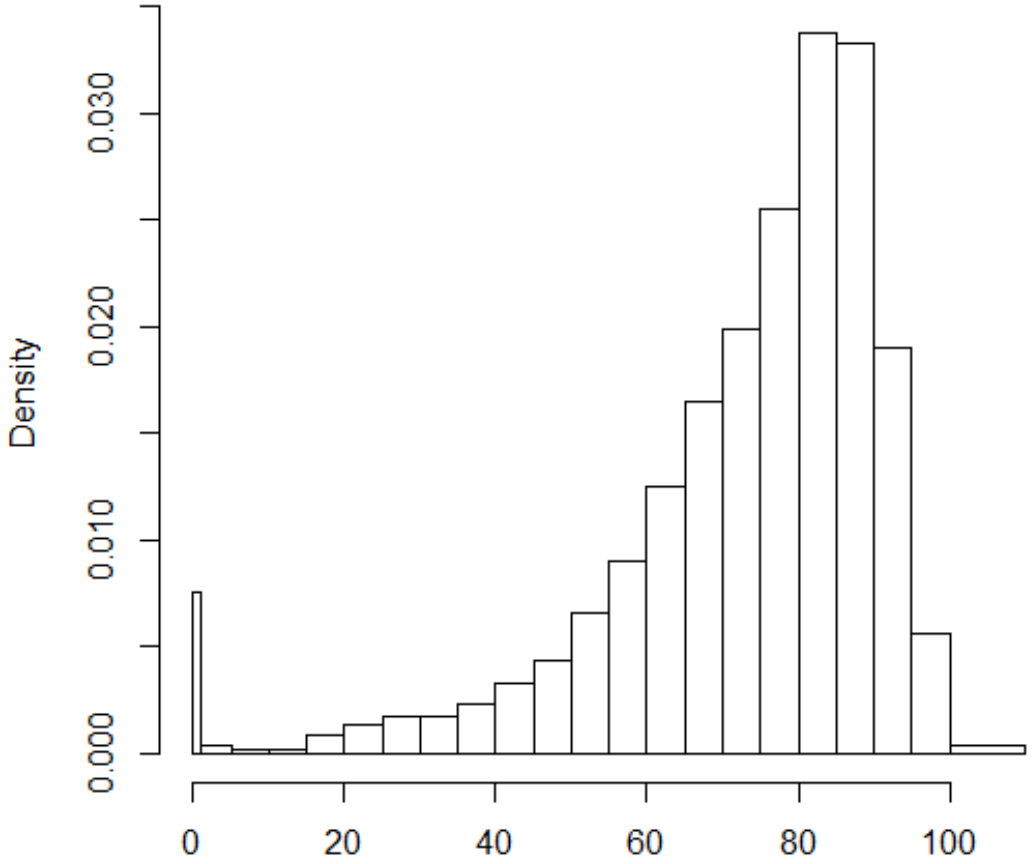
Un outlier è un dato che sta molto fuori rispetto al grosso dei dati



Media = 180.000 abitanti. Mediana = 94.270 abitanti

Moda

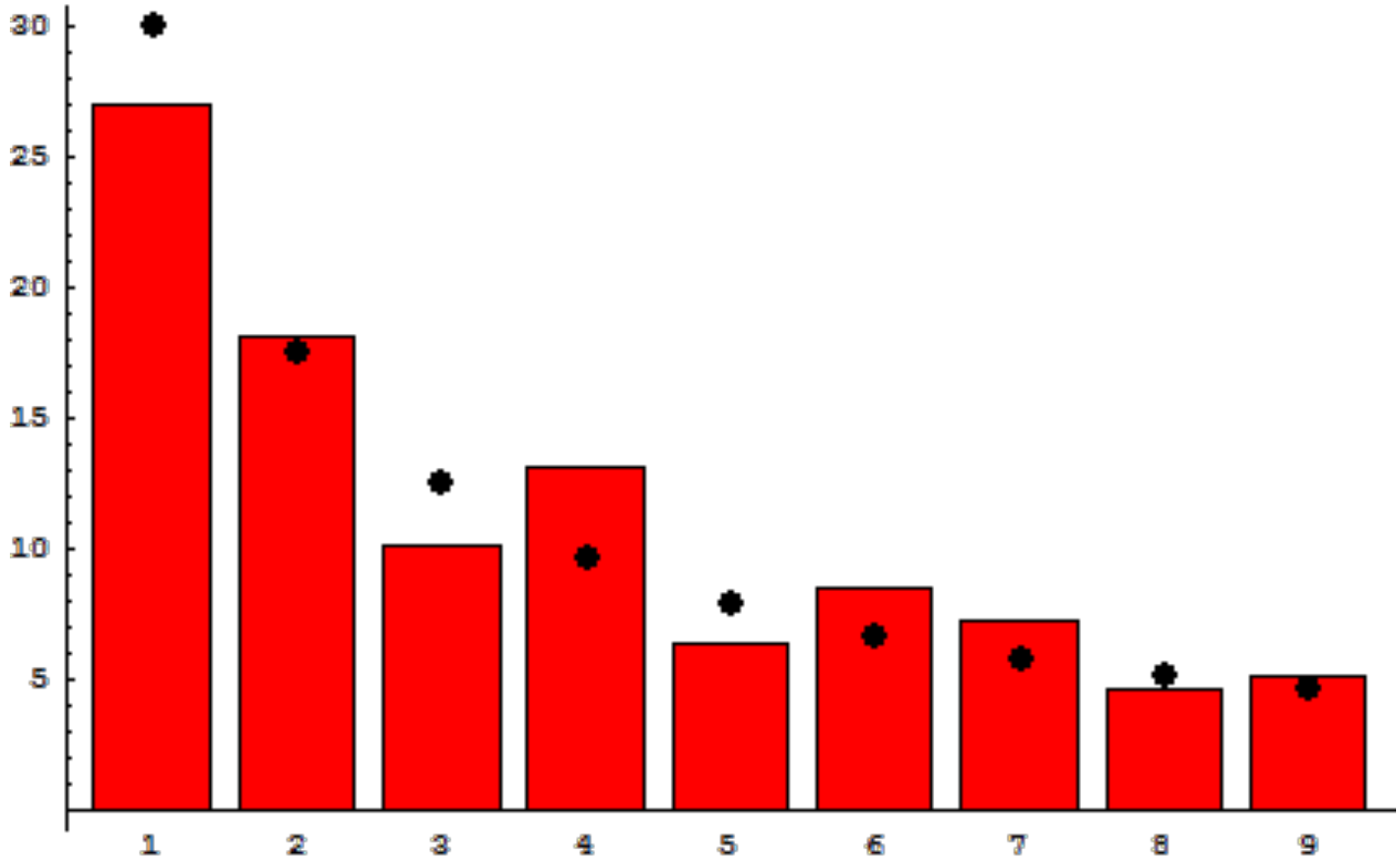
La **moda** è la modalità della variabile che ha **maggiore frequenza** o **maggiore densità**.



Età alla morte dei maschi (Australia, 2012). Moda 80-85 anni

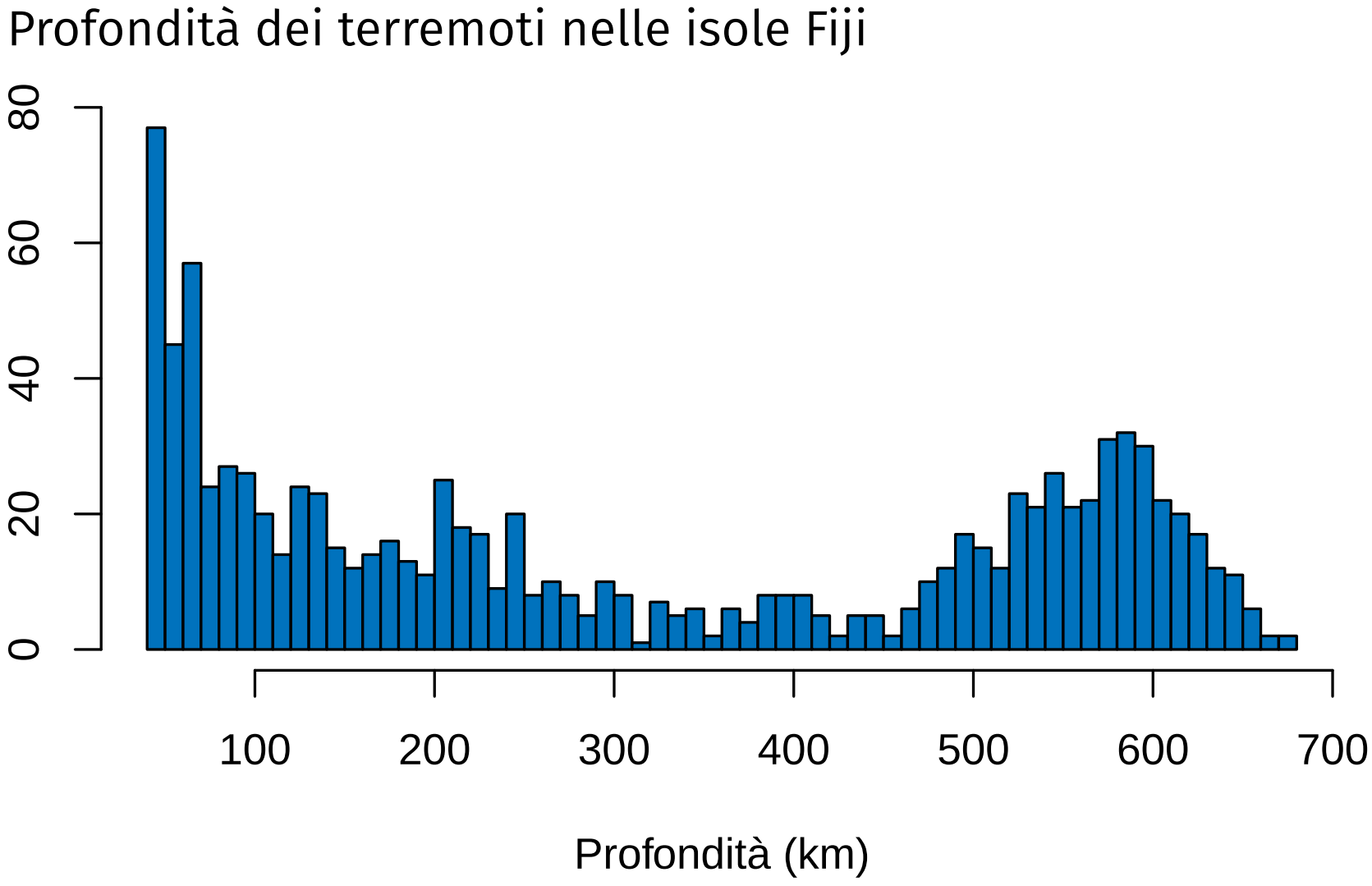
Prima cifra in collezioni di numeri

Legge di **Benford**



Moda = 1

Distribuzioni bimodali



Misure di variabilità

Variabilità

- Senza variabilità non c'è statistica
- Misuriamo l'eterogeneità dei dati quantitativi

Esempio. 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

- **Campo di variazione**: $x_{max} - x_{min}$
- **Differenza interquartile** basato sulla lunghezza dell'intervallo che contiene il 50% dei dati (maggiore la lunghezza maggiore la variabilità)
- **Varianza**: basata sulla media degli scarti al quadrato dalla media (maggiore lo scarto, maggiore la variabilità)

Importanza della variabilità

- Finanza: **rischio, volatilità**
- Ingegneria: precisione delle misure (l'inverso della variabilità è la **precisione**)
- Economia: **bontà delle previsioni**
- Politica: **incertezza** dei sondaggi

Differenza interquartile

IQR

La differenza interquartile (IQR) è la lunghezza dell'intervallo che contiene il 50% centrale dei dati.

- Il **primo quartile** Q_1 è il valore che ha prima di sé il 25% dei dati
- Il **terzo di quartile** Q_3 è il valore che ha prima di sé il 75% dei dati

$$IQR = Q_3 - Q_1.$$

Esempio

Dati dell'**Esercizio 3.15**: Tempo impiegato da 24 persone a svolgere un compito

```
23 35 14 37 28 45 12 40 27 13 26 25
37 20 29 49 40 13 27 16 40 20 13 66
```

Ramo-foglia normale e **ordinato**

```
1 | 423363
2 | 387650970
3 | 577
4 | 50900
5 |
6 | 6
```

```
1 | 233346
2 | 003567789
3 | 577
4 | 00059
5 |
6 | 6
```

I quartili sono $Q_1 = x_{(n+1)0.25}$ $Q_3 = x_{(n+1)0.75}$

Calcolo dei quartili

- **Ordinare i dati**
- Q_1 è il dato della osservazione numero $(n+1)0.25$ arrotondato
- Q_3 è il dato della osservazione numero $(n+1)0.75$ arrotondato
- $IQR = Q_3 - Q1$

Arrotondare all'intero più vicino

Il numero di osservazioni è $n = 24$

$$posto(Q_1) = (25)(0.25) = 6.25 \approx 6$$
$$posto(Q_3) = (25)(0.75) = 18.75 \approx 19.$$

```
1 | 233346          (6)    Q1 = 16
2 | 003567789       (9)
3 | 577             (3)
4 | 00059           (5)    Q3 = 40
5 |                 (0)
6 | 6               (1)
```

IQR = $40 - 16 = 24$ sec. $\quad Me = 26.5$ sec.

5 numeri di sintesi

- $\min, Q_1, Me, Q_3, \max$

$$Min = 12,\ Q_1 = 16,\ Me = 26.5,\ Q_3 = 40,\ max = 66$$

- **Campo di variazione**: $range = 66 - 12 = 54$ sec
- **Differenza interquartile**: IQR = 24 sec

Varianza e deviazione standard

Varianza

La varianza è basata su una media degli scarti^2

Esempio 3.5. Voti a un test

$$x_i : 50, 60, 70, 80, 90, \qquad \bar{x} = 70$$

(**la tabella seguente non è una distribuzione di frequenze**)

Unità	Dati	Scarti dalla media	Scarti^2
1	50	-20	400
2	60	-10	100
3	70	0	0
4	80	10	100
5	90	20	400
Totale	350	0	1000

$$varianza = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \frac{1000}{4} = 250.$$

Simbologia e avvertenze

La varianza si usa denotare con

- s^2 nel **campione**
- σ^2 nella **popolazione**

Nota: Quando si calcola nel campione — per un motivo che verrà spiegato più avanti — non si divide per n ma per $n - 1$.

Nella popolazione n è molto grande e non fa differenza dividere per n o $n - 1$.

Deviazione standard

- La varianza è espressa in una **unità di misura al quadrato**
- Per le interpretazioni è opportuno usare la **deviazione standard** = $\sqrt{\text{varianza}}$

La deviazione standard (SD) — detta anche **scarto quadratico medio** — è misurata nella stessa unità di misura

$$SD = \sqrt{250} = 15.8.$$

Esempio precedente

Esercizio 3.15 Tempi per completare un lavoro

```
23 35 14 37 28 45 12 40 27 13 26 25
37 20 29 49 40 13 27 16 40 20 13 66
```

- Varianza = $\frac{4234.958}{24-1} = 184.1\,\text{sec}^2$
- $SD = \sqrt{184.1} = 13.6$ sec
- Confronta

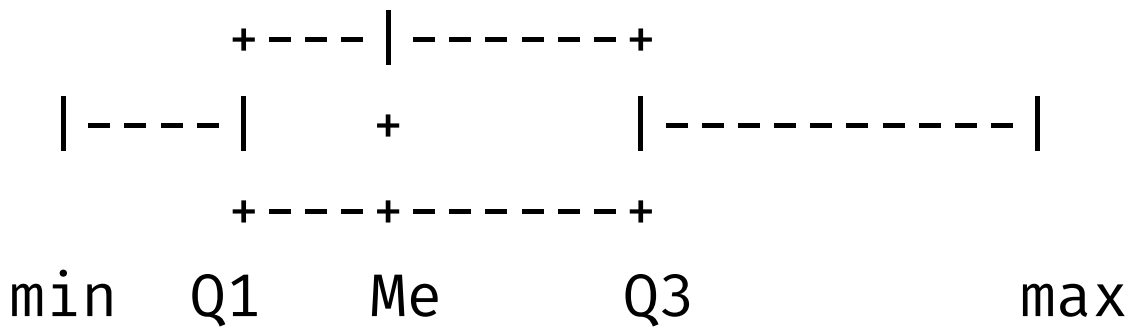
$$media(X) = 29\,sec \qquad SD(X) = 13.6\,sec$$
$$mediana(X) = 26.5\,sec \qquad IQR = 24\,sec$$

Box-plot

- Un **box-plot** è una rappresentazione grafica di un gruppo di dati
- È utile per verificare la simmetria o la asimmetria della distribuzione
- e per confrontare due o più gruppi di dati

Come si costruisce il box-plot

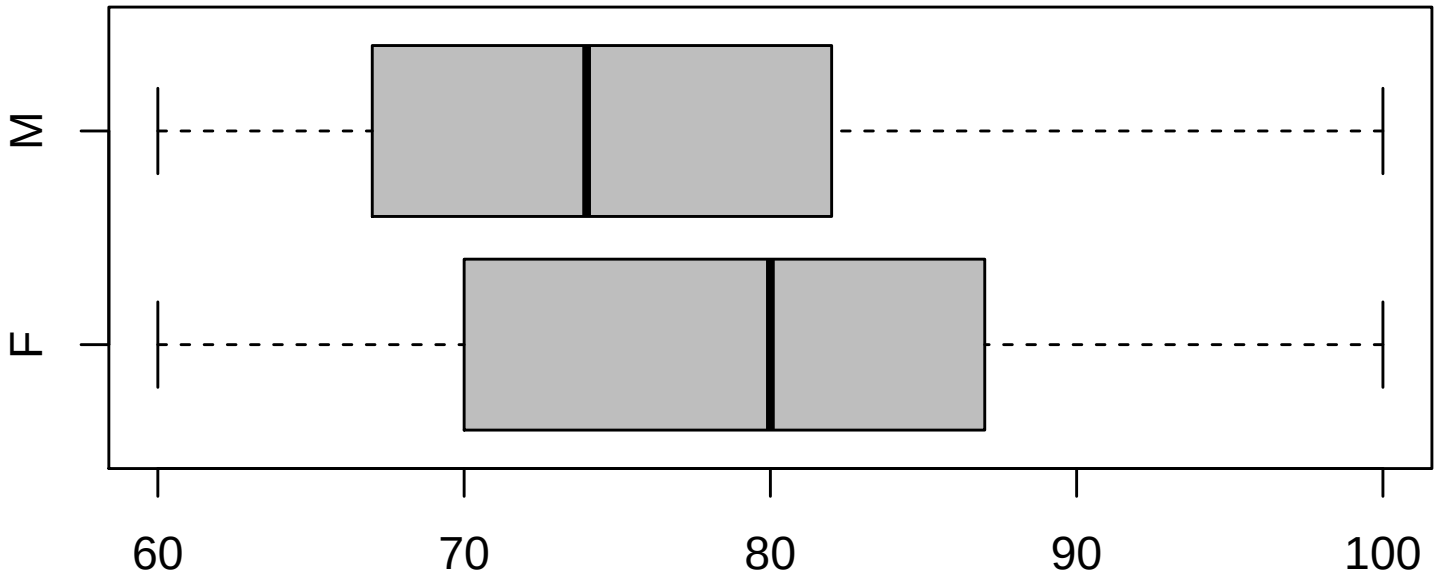
- Si calcolano i 5 numeri di sintesi
- Si disegna una **scatola** e due **code**



- Nella scatola ci sta il 50% centrale dei dati
- I box-plot si possono affiancare per confrontare 2 gruppi di dati

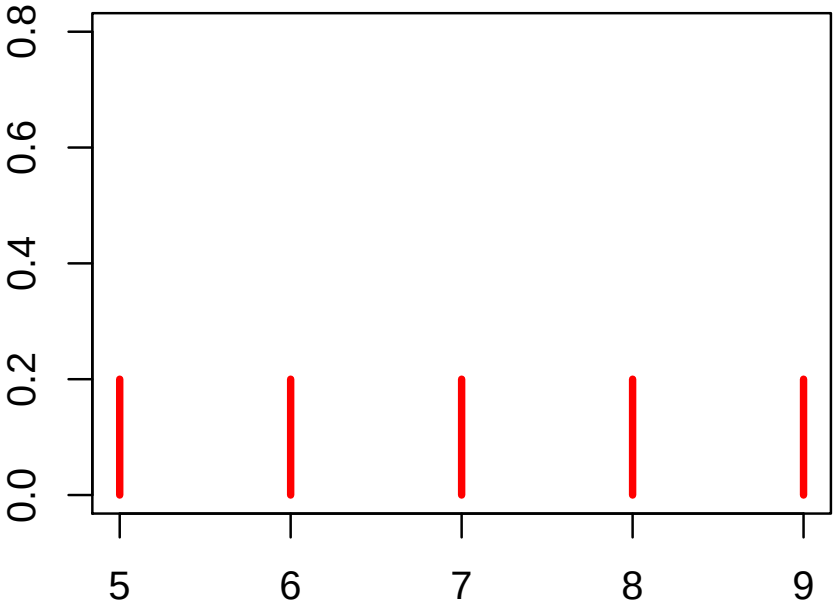
Esempio

Box-plot appaiati del voto alla maturità separatamente per i maschi e per le femmine

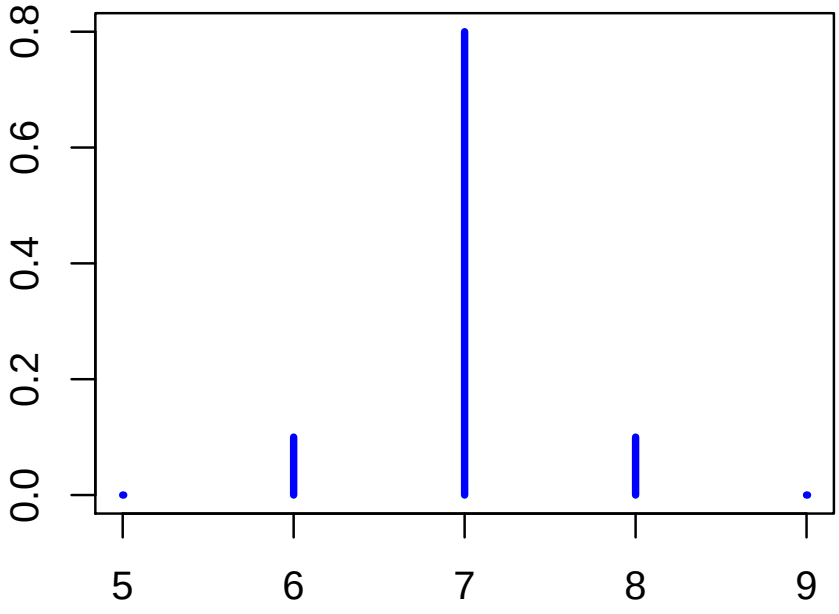


Confronti di variabilità

La variabilità si valuta **in orizzontale**



Gennaio: Più variabile



Giugno: Meno variabile

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

Varianza con dati raggruppati

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
- che si pesano con pesi uguali alle frequenze

(segue)

Fatturato	Frequenza	Scarti	Scarti^2	Scarti^2 $\times$ Frequenza
6	10	6-7	1	10
7	80	0	0	0
8	10	8-7	1	10
Totale	100			20

Quindi

$$varianza = 20/99 \approx 0.2$$

Coefficiente di variazione

- È una misura di variabilità relativa (numero puro)
- Si calcola in genere quando la variabile assume **solo valori positivi**:

$$CV = \frac{deviazione\ standard}{media}$$

e si esprime spesso in percentuale.

Esempio

Vendite giornaliere

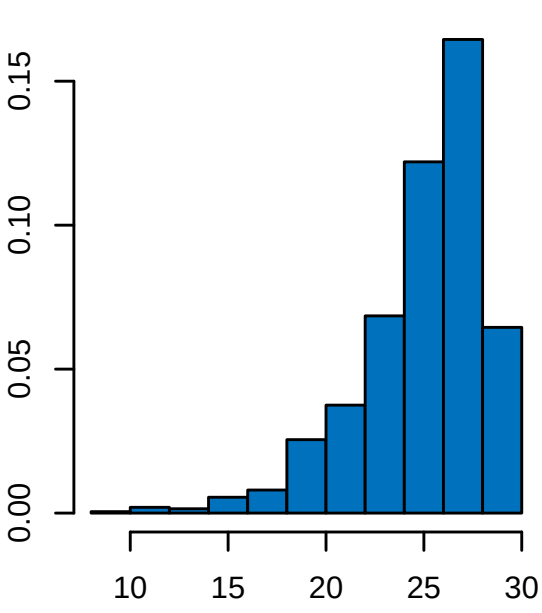
	Media	Deviazione standard
Negozio grande	20 000 Euro	1000 Euro
Negozio piccolo	2000 Euro	200 Euro

$$CV_{grande} = 5\% \quad CV_{picc} = 10\%$$

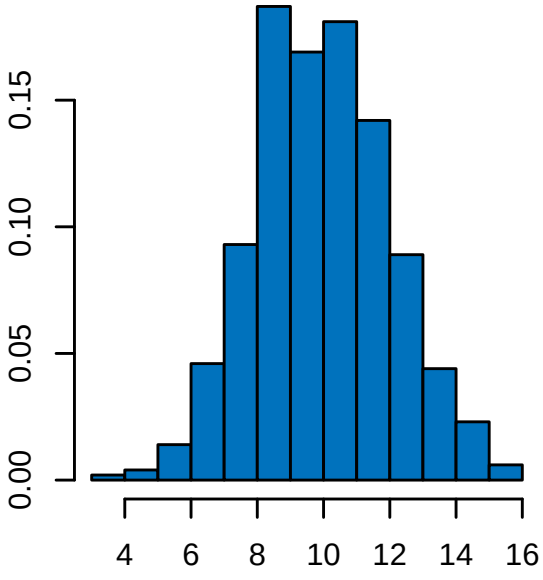
Il negozio grande ha una variabilità relativa minore.

Regola empirica

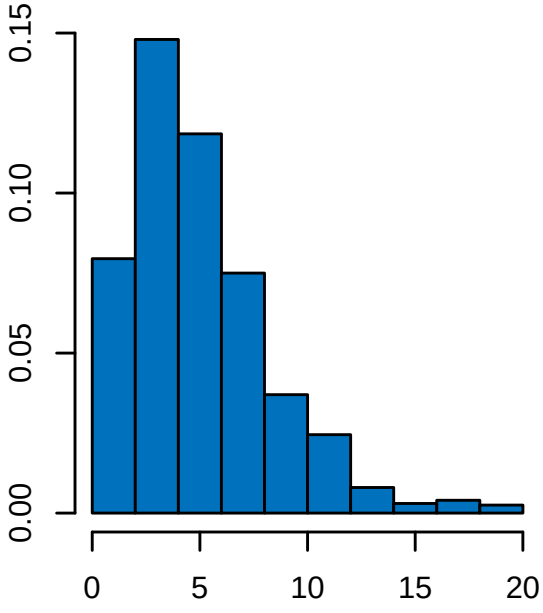
Asimmetria nelle distribuzioni unimodali



Asimmetria negativa



Simmetria



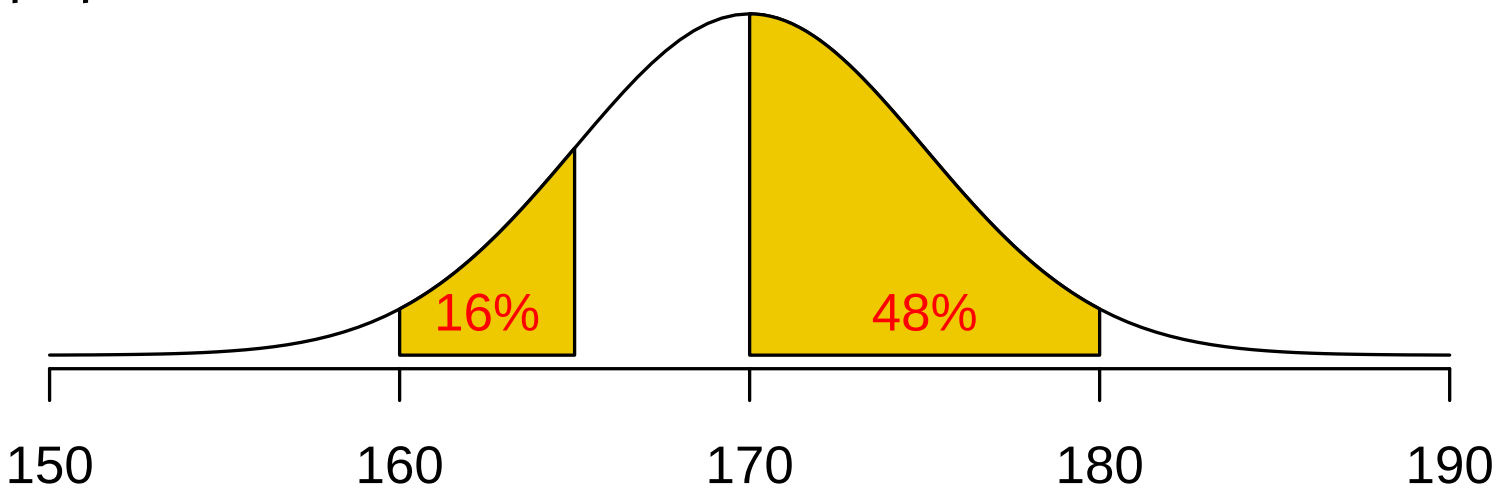
Asimmetria positiva

- **Asimmetria negativa**: Media minore della mediana
- **Simmetria**
- **Asimmetria positiva**: Media maggiore della mediana

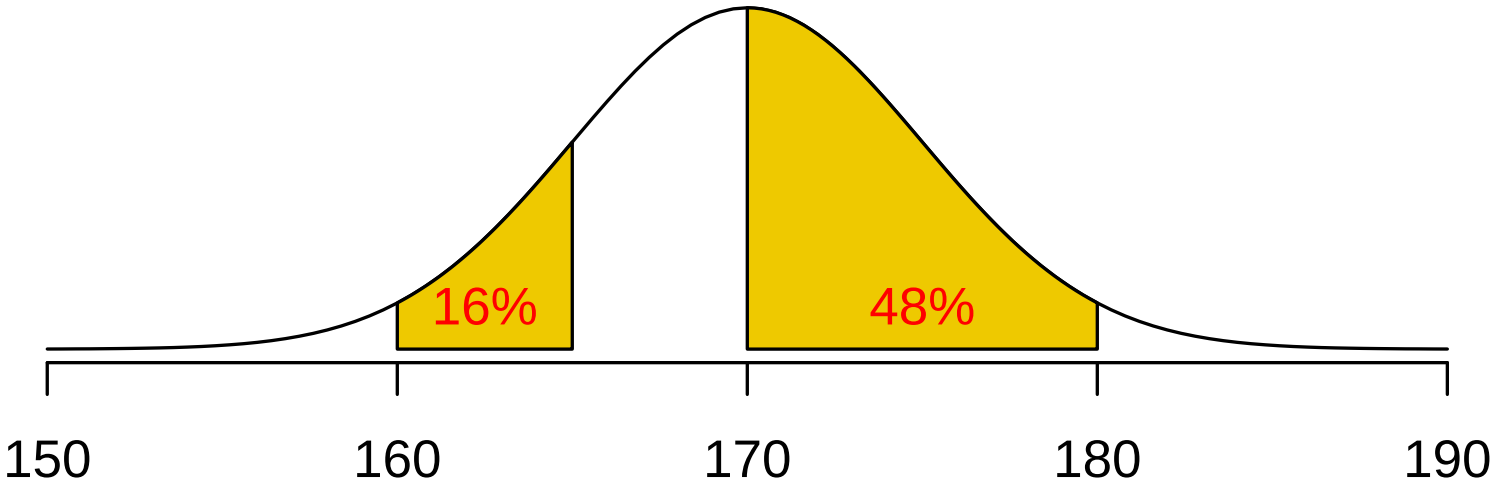
Proporzione di individui che appartengono a un intervallo

Normalizziamo l'area degli istogrammi (basta dividere le densità per il numero di osservazioni n)

- 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 $a \leq X \leq b$
- Esempio: Istogramma della statura (cm) in una popolazione



(segue)



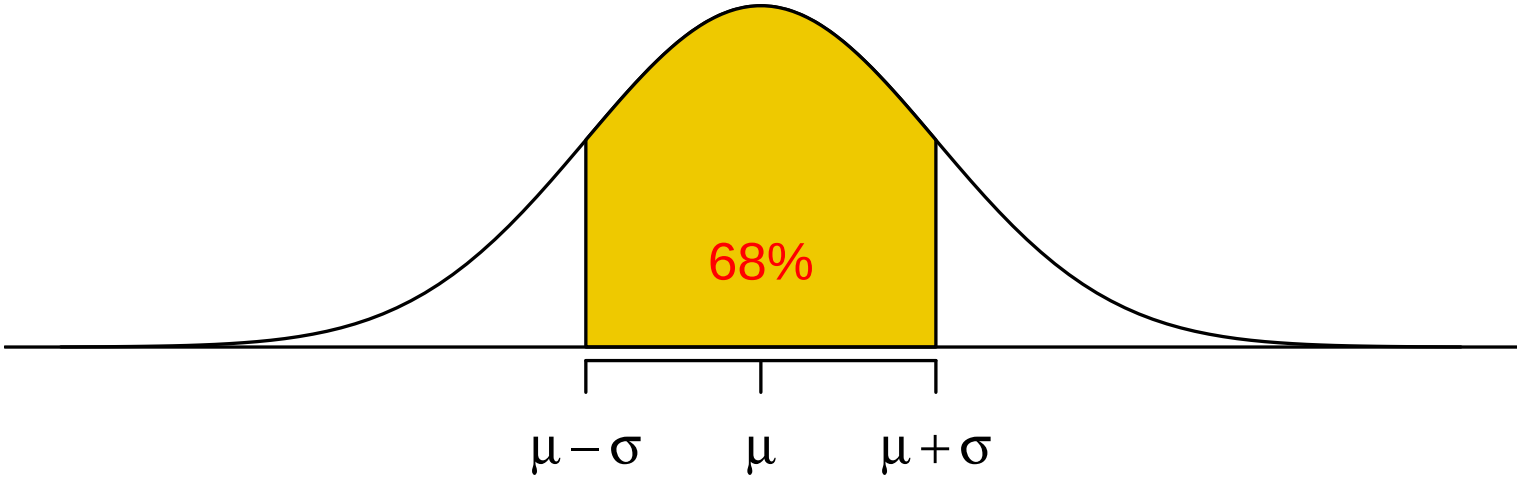
- La percentuale della popolazione con statura tra 160 e 165 cm è il 16%
- La percentale della popolazione con statura compresa tra 170 e 180 cm è il 48%
- Questo calcolo deve essere fatto **valutando tutti gli individui della popolazione**

Regola empirica (1)

- Se l'istogramma ha una forma come quella delle altezze si dice **normale**
- La forma della normale è **campanulare simmetrica** e si sa quindi che la **media = moda**
- Nel caso della normale si sa che la percentuale di popolazione che cade nell'**intervallo centrato sulla media**

$$\textit{Media} \pm \textit{deviazione standard}$$

contiene sempre il 68% della popolazione

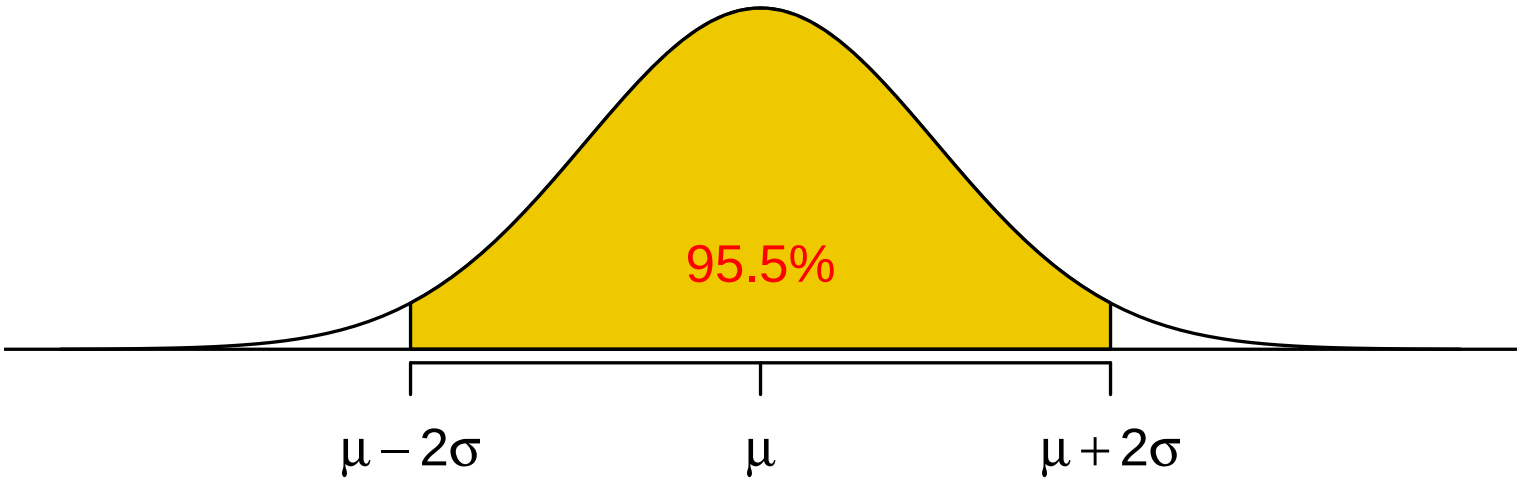


Regola empirica (2)

Nel caso della normale si sa che la percentuale di popolazione che cade nell'**intervallo centrato sulla media**

$$\textit{Media} \pm 2 \times \textit{deviazione standard}$$

contiene sempre il 95% (circa) della popolazione

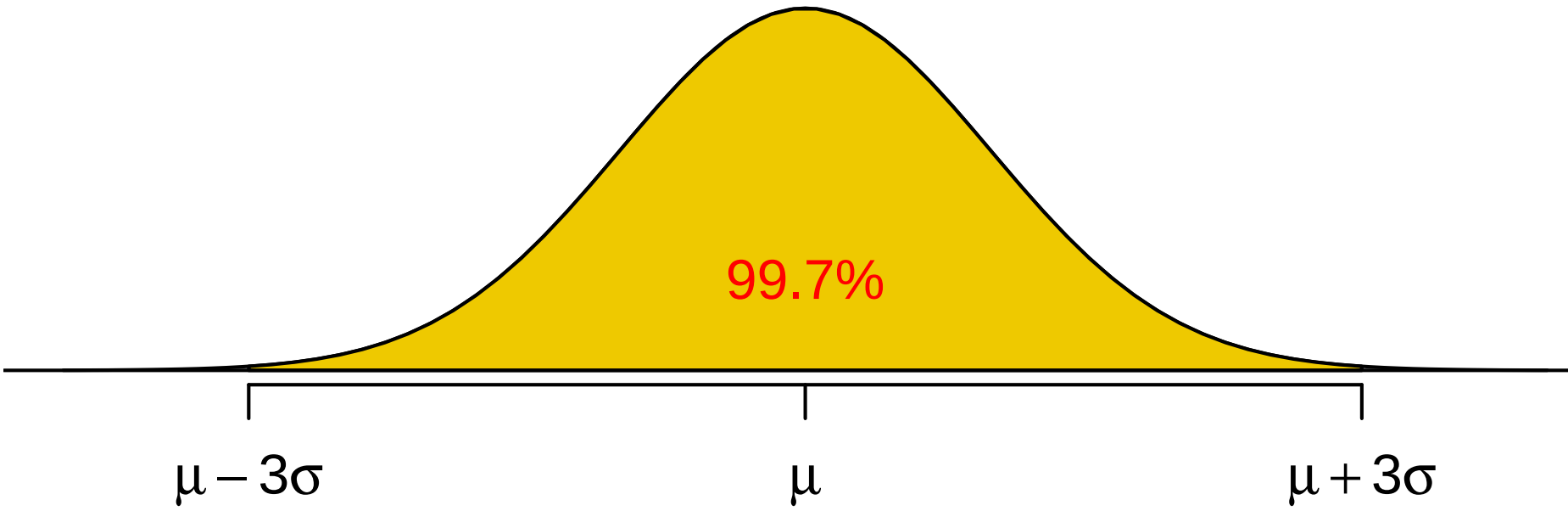


Regola empirica (3)

Nel caso della normale si sa che la percentuale di popolazione che cade nell'**intervallo centrato sulla media**

$$\textit{Media} \pm 3 \times \textit{deviazione standard}$$

contiene sempre il 99% (circa) della popolazione



Utilità della regola empirica

- La devizione standard è un **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%.

La disuguaglianza di Chebyshev la facciamo dopo.

Associazione tra due variabili quantitative

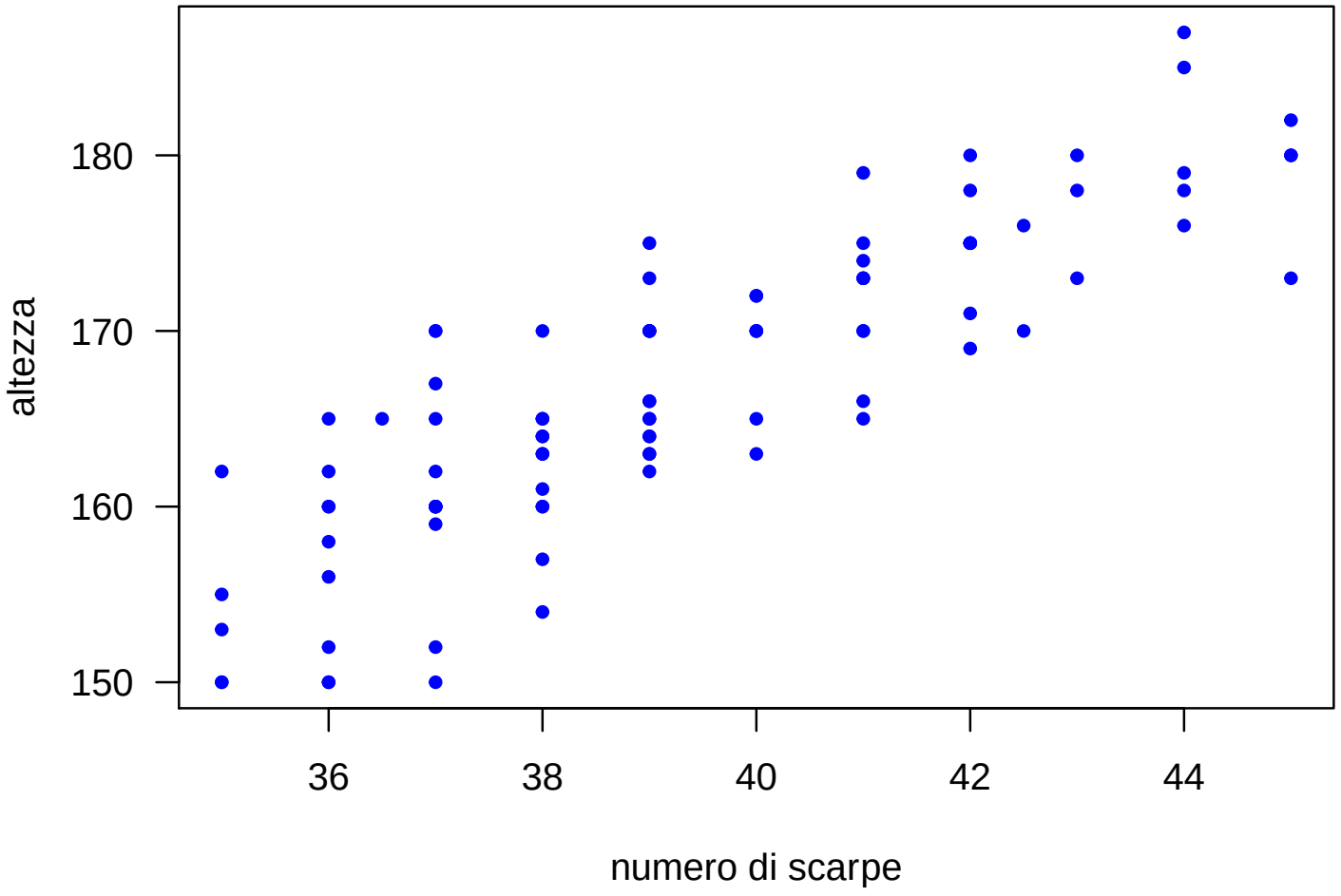
C’è associazione tra

- 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...

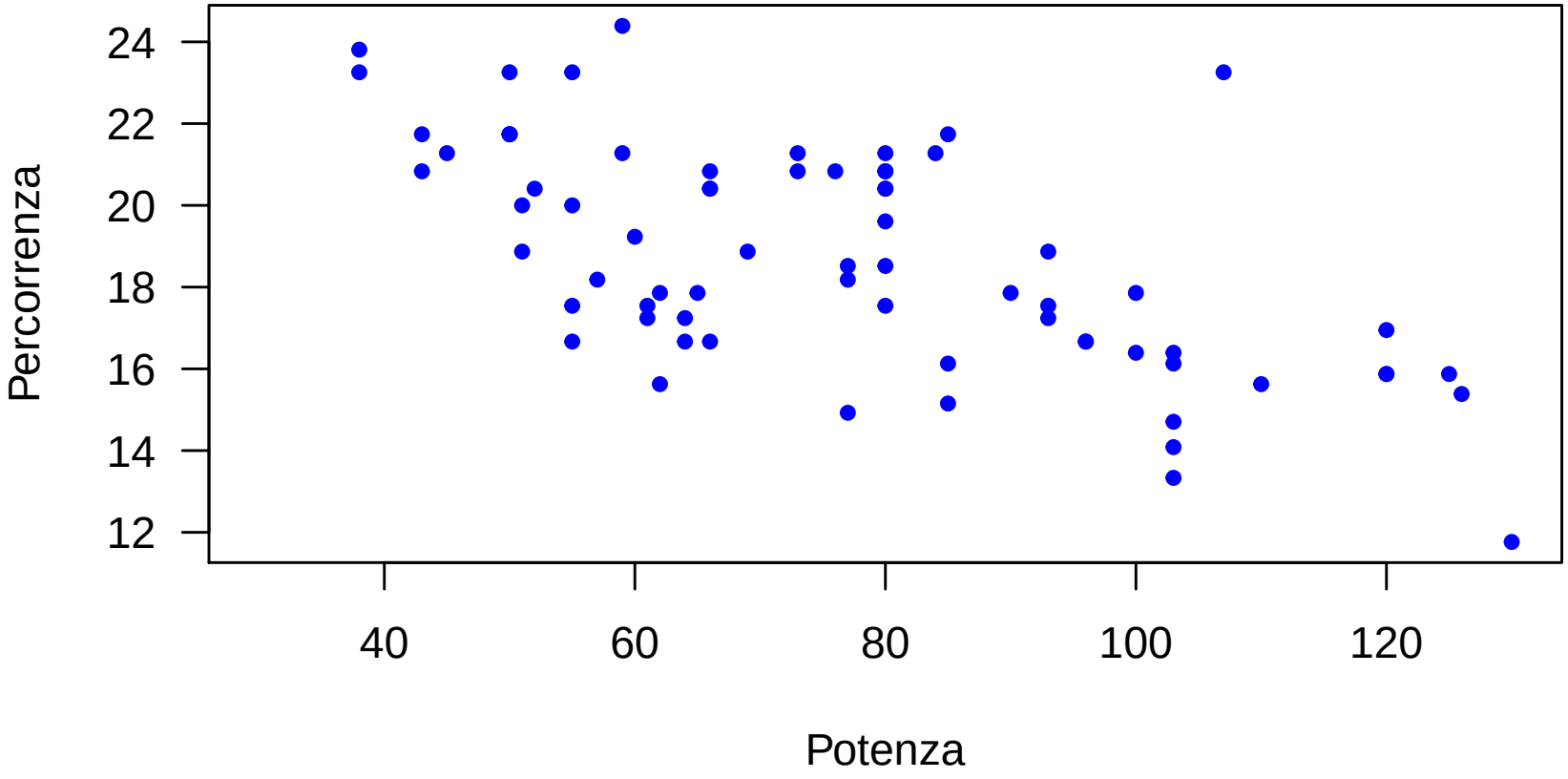
Scatter



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

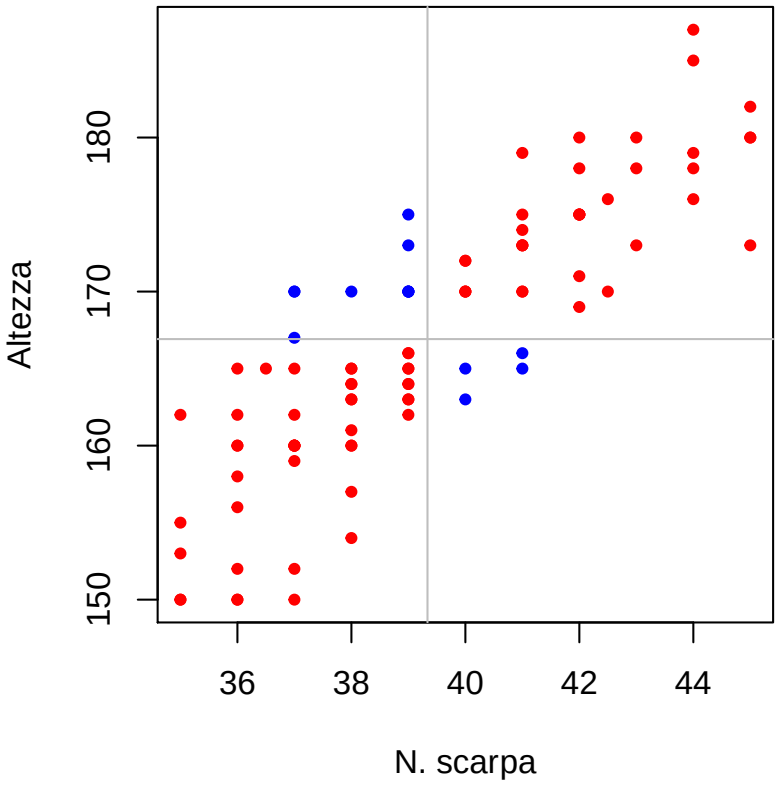
Associazione positiva o negativa?

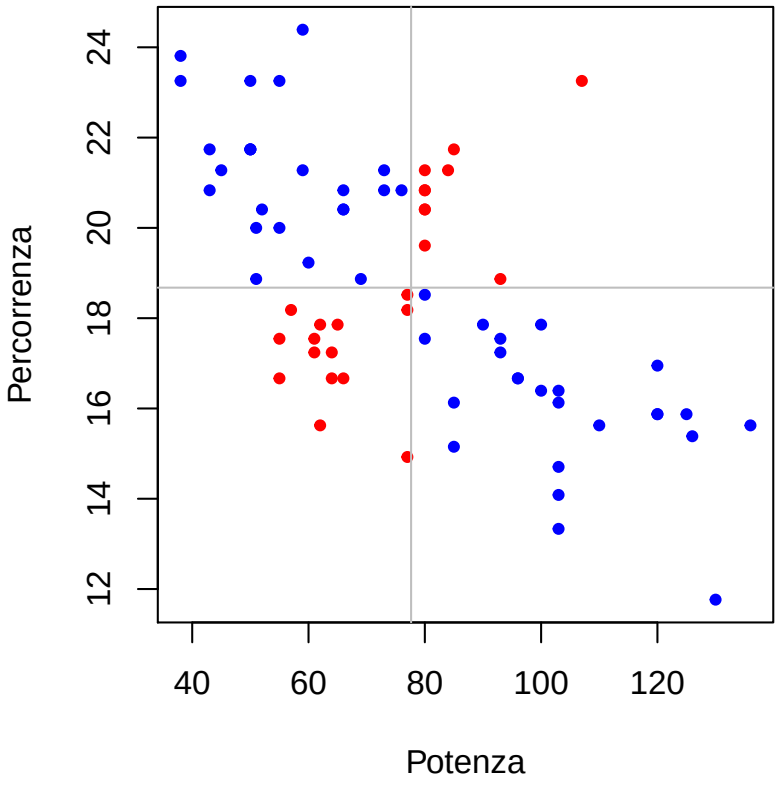
C’è associazione tra percorrenza di un auto (km/l) e potenza (kW = 1.36 CV)?



C’è evidenza di una relazione inversa (negativa)

Come si fa a valutare?





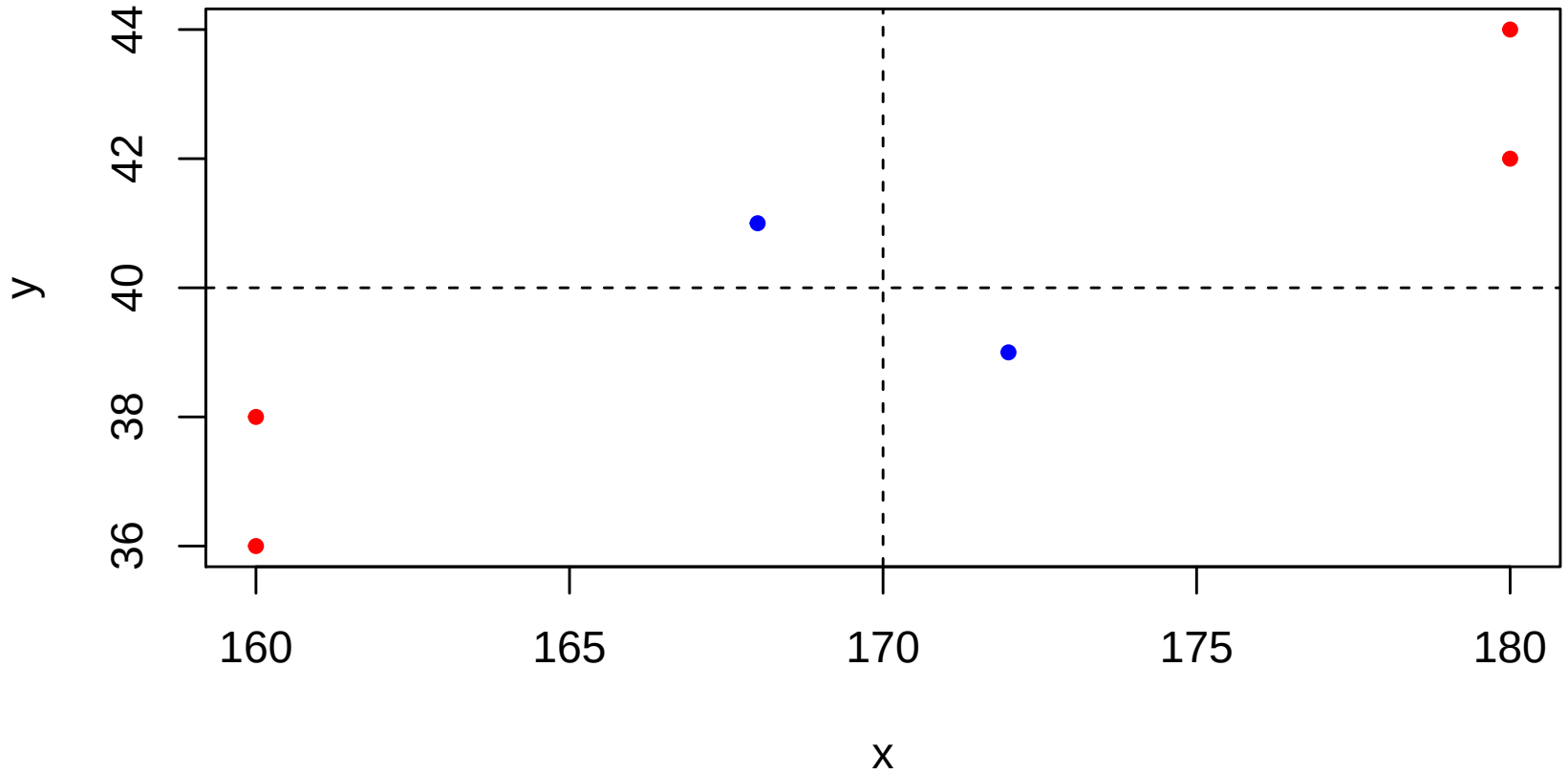
- Associazione positiva se vince il I e III quadrante
- Associazione negativa se vince il II e IV quadrante

Indici fondamentali

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

Covarianza: esempio

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



Calcolo

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

	1	2	3	4	5	6	Somma
X	160	160	168	172	180	180	
Y	36	38	41	39	42	44	
x-170	-10	-10	-2	2	10	10	0
y -40	-4	-2	1	-1	2	4	0
Prodotto	40	20	-2	-2	20	40	116

$$covarianza = 116/5 = 23.2$$

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

Coefficiente di correlazione

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

$$\text{cov}(\textit{altezza}\ (\textit{cm}), \textit{n.scarpe}) = 23.2$$

$$\text{cov}(\textit{altezza}\ (\textit{metri}), \textit{n.scarpe}) = 0.232$$

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

Formula generale

$$\mathrm{corr}(X, Y) = \frac{cov(X, Y)}{dev.\ st(X), dev.st(Y)}$$

- Qualsiasi denominatore si usi, si ottiene la formula seguente

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

Correlazione tra n. scarpe e altezza

Abbiamo

$$\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = 116$$

$$\sum_{i=1}^{6}(x_i - \bar{x})^2 = 408$$

$$\sum_{i=1}^{6}(y_i - \bar{y})^2 = 42$$

dunque

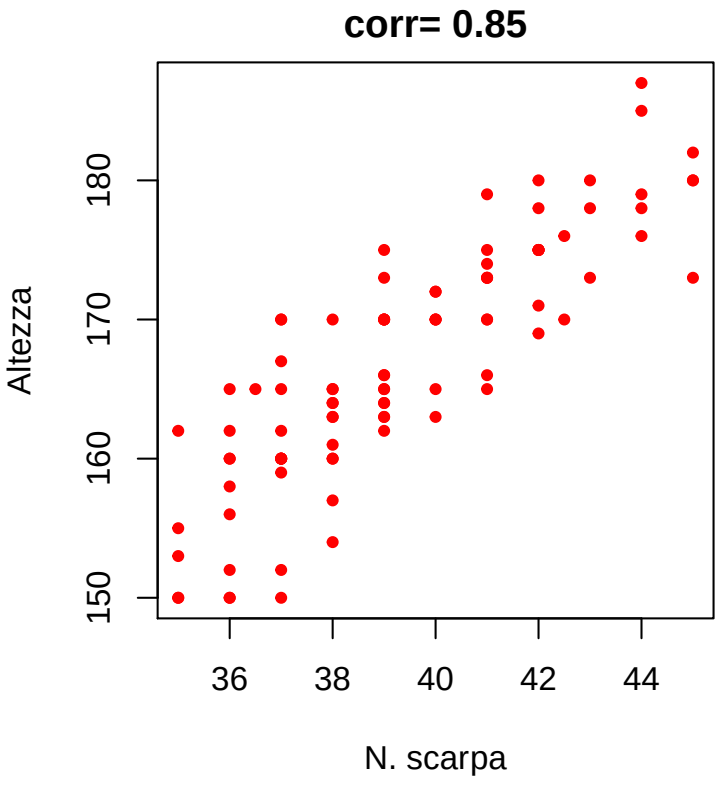
$$correlazione = \frac{116}{\sqrt{408 \cdot 42}} = 0.89$$

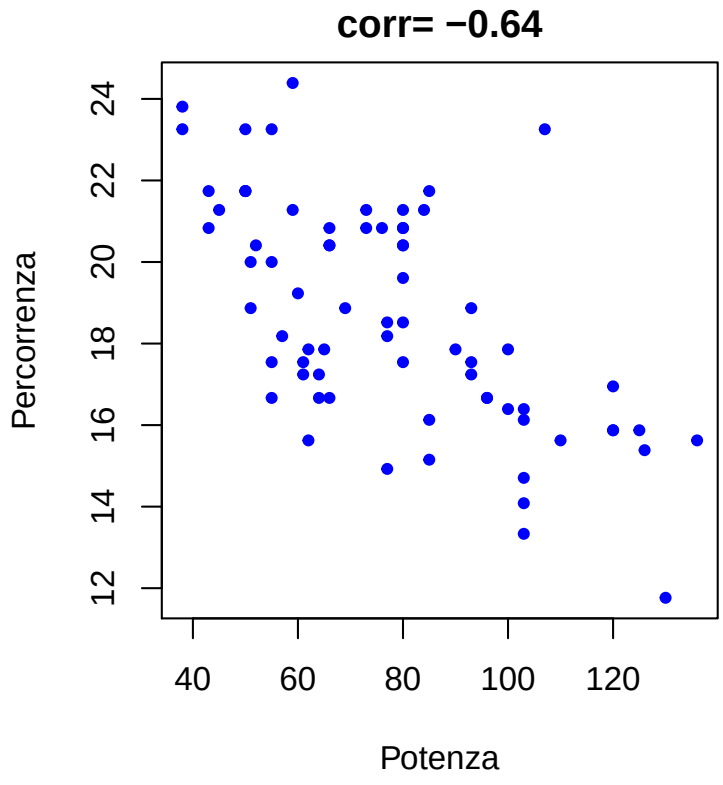
Proprietà del coefficiente di correlazione

- Introdotto da Karl Pearson (1900)
- 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

Altri link interessanti

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

Previsioni

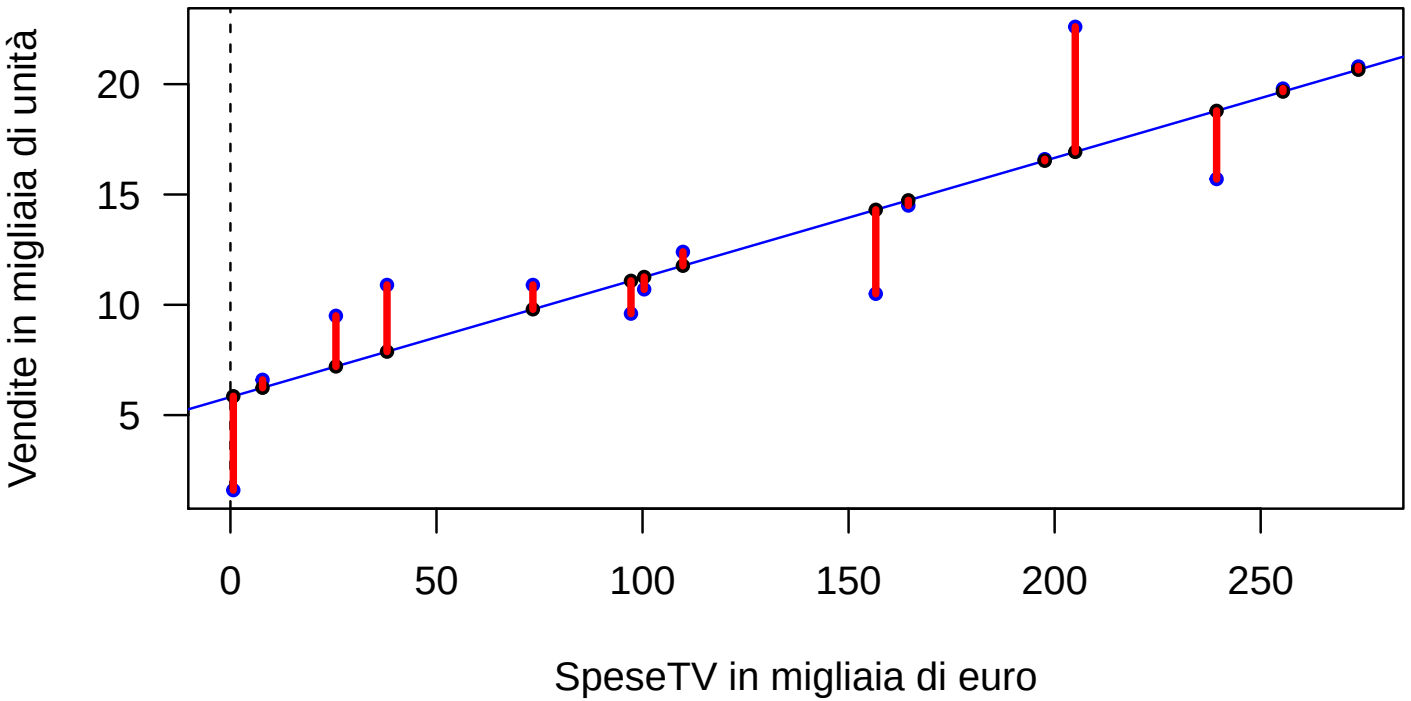
- Date due variabili quantitative talvolta si usa X per **prevedere** Y
- La variabile X è detta variabile **esplicativa**
- La variabile Y è detta **dipendente**
- Y = livello delle vendite
- X = spese di pubblicità TV

Relazioni lineari

- Talvolta la relazione tra X e Y è circa lineare

$$Y = \beta_0 + \beta_1 X + \textit{residuo}$$

- β_0 è l'**intercetta**, β_1 è la **pendenza**
- il residuo è la **differenza** tra il valore vero e quello teorico



Retta dei minimi quadrati

- Se sapessimo la vera relazione lineare potremmo usarla per fare previsioni
- Usiamo come approssimazione della retta vera, **la retta dei minimi quadrati** che rende minima la **distanza tra i dati e la retta**

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Retta dei MQ

La pendenza si trova calcolando

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r \cdot \frac{st.dev\ Y}{st.dev\ X}$$

La retta si trova calcolando

$$y = \bar{y} + m(x - \bar{x})$$

La pendenza è

- **positiva** se $r > 0$
- **negativa** se $r < 0$
- **nulla** se $r = 0$

Un esempio

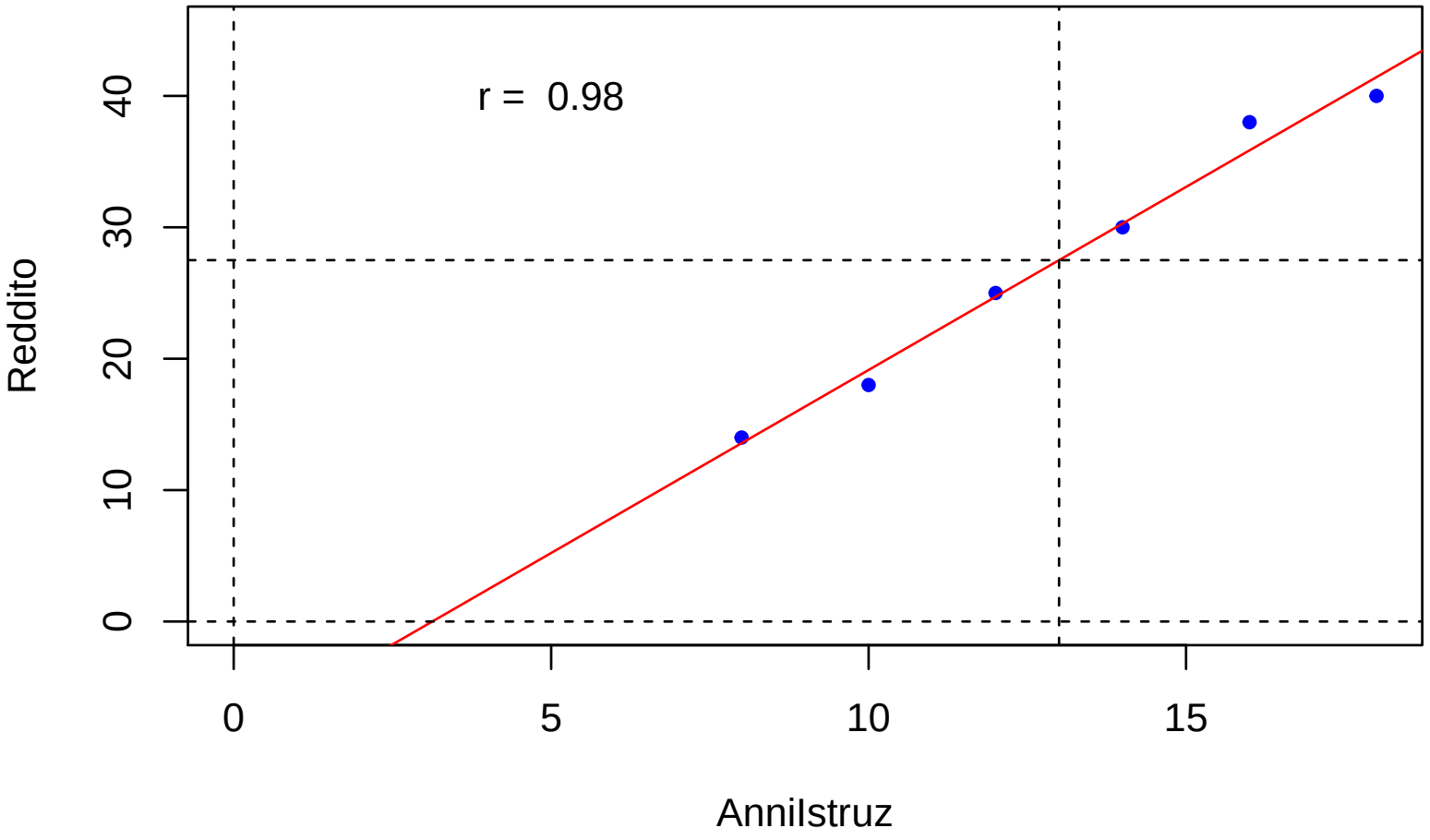
Relazione tra **reddito** Y e **anni di istruzione**

Unità	1	2	3	4	5	6	Media
Anni	8	10	12	14	16	18	13.0
Reddito	14	18	25	30	38	40	25.5
							Somma
$x - 13$	-5	-3	-1	1	3	5	0
$y - 25.5$	-13.5	-9.5	-2.5	2.5	10.5	12.5	0
Prodotto	67.5	28.5	2.5	2.5	31.5	62.5	195
Scarti x^2	25	9	1	1	9	25	70

Pendenza: $m = 195/70 = 2.79$

Retta dei MQ: $\hat{y} = 27.5 + 2.79(x - 13)$

Grafico e interpretazione



Per ogni anno di istruzione in più il reddito atteso aumenta di 2790 Euro all'anno.