Slide Set 6
Modello Normale Standard e Simmetria

Pietro Coretto
pcoretto@unisa.it

Corso di
Analisi e Visualizzazione dei Dati (Parte I)

Corso di Laurea in "Statistica per i Big Data" (L-41)
Università degli Studi di Salerno

Versione: 22 febbraio 2022 (h10:21)

Notes

Distribuzione al limite quando $n \to \infty$

Consideriamo un campione di enormi dimensioni. Trattiamo n come se $n \to \infty$, tuttavia quando diciamo n *sufficientemente grande*, immaginiamo che n sia enorme ma non infinito

Per n *sufficientemente grande*

- il campionamento produrrà un grande numero di livelli distinti x_k. Poiché la X è continua ci aspettiamo che $n_k = 1$ per ogni x_k
- $\varepsilon = \frac{1}{n}$ sarà piccolissimo ma positivo ($\varepsilon > 0$)
- ogni osservazione tenderà ad avere una frequenza relativa $\frac{1}{n} \approx \varepsilon$

Quale effetto mi aspetto nella distribuzione di X? Pensiamo a cosa dovrebbe accadere in termini di istogramma e ECDF

Notes

Istogramma

- se $(x_{k-1}, x_k]$ è un intervallo dove la X è densa di dati, prendendo $\Delta_k = x_k - x_{k-1}$ troppo largo mi aspetto troppe osservazioni
- assumendo un windowing uniforme, se $n \uparrow$ devo prendere $\Delta \downarrow$. Quindi devo assicurarmi che $\Delta \to 0$ per $n \to \infty$ (ovvero il numero di classi $K \to \infty$)

Con Δ sufficientemente piccolo, $(x_{k-1}, x_k]$ sarà strettissimo, e per $x' \in (x_{k-1}, x_k]$ avrò $h(x') =$ densità in un intorno strettissimo di x'

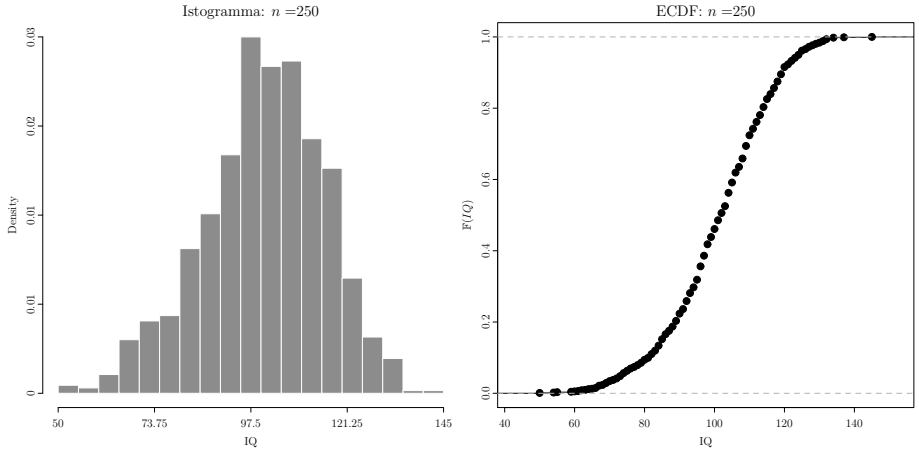
ECDF

- se le osservazioni sono tutte distinte $\mathbb{F}(x_{(j)}) = \frac{j}{n}$, inoltre la ECDF salterà su ogni $x_{(j)}$
- se $n \uparrow$ mi aspetto che $(x_{(j)} - x_{(j-1)}) \downarrow \implies$ i salti/scalini di $\mathbb{F}$ diventano sempre più vicini
- Quanto sono alti i gradini? Poichè la $\mathbb{F}$ salterà su $x_{(j-1)}$ e poi subito dopo su $x_{(j)}$

$$\mathbb{F}(x_{(j)}) - \mathbb{F}(x_{(j-1)}) = \left(\frac{j}{n} - \frac{j-1}{n}\right) = \frac{1}{n} = \varepsilon$$

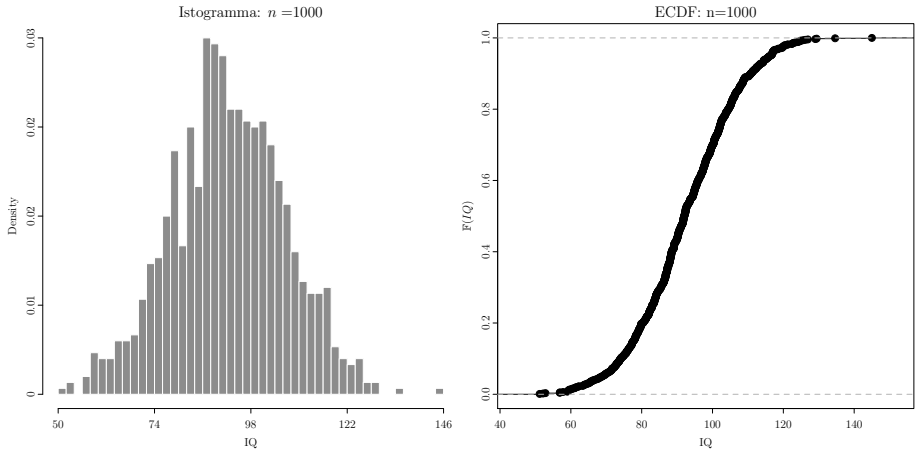
Notes

Esempio: NLSY survey (US, 1980) si rileva l'IQ, prendiamo $n = 250$ osservazioni a caso



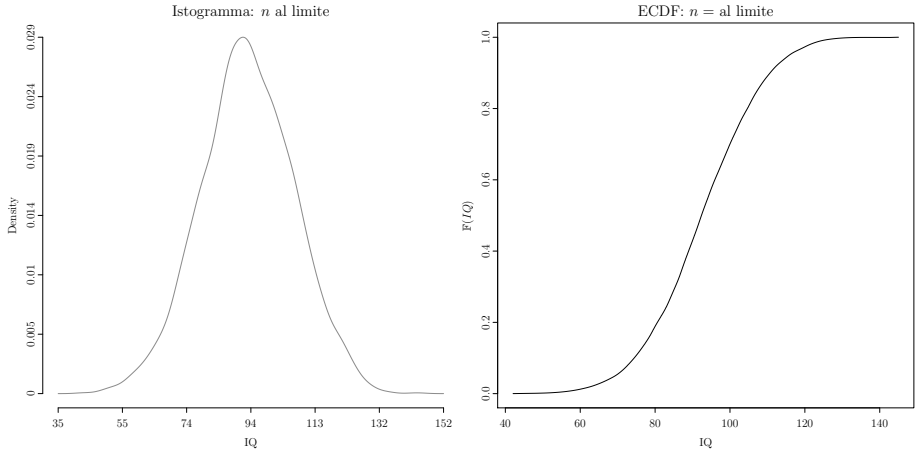
Notes

Esempio: NLSY survey (US, 1980) si rileva l'IQ, prendiamo $n = 1000$ osservazioni a caso eva



Notes

. . . al limite



Notes

Modello di densità normale standard

Adesso introduciamo una curva di densità che assumeremo come un *modello di riferimento*, pensiamo ad un'*istogramma ideale*

Immaginiamo n =grandissimo, tanto da poter costruire un istogramma con classi uniformi di ampiezza Δ=sufficientemente piccolo

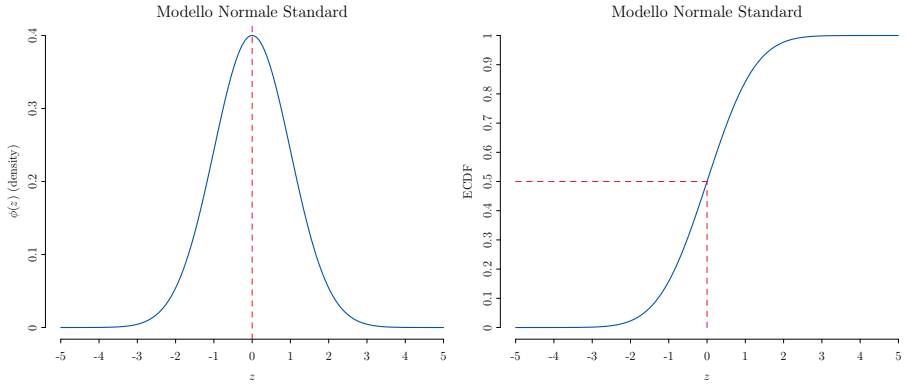
Definizione (Modello normale standard)

Si Z una variabile continua con media=0 e varianza=1, la distribuzione di Z segue quella del modello *normale standard* se la densità in un intorno (piccolo a piacere) del punto x è descritta dalla seguente *funzione di densità*

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Nota: il modello normale standard è anche detto "*modello Gaussiano*"

Notes



Notes

Premessa

- immaginiamo la densità normale come un ideale istogramma *finissimo* (tante barre strette in corrispondenza di tantissime classi di valori)
- l'area di una barra dell'istogramma è pari alla frequenza relativa delle osservazioni contenute in esso. Quindi la frequenza f dei dati in $(a, b]$, è pari alla somma delle aree delle barre che coprono l'intervallo $(a, b]$.

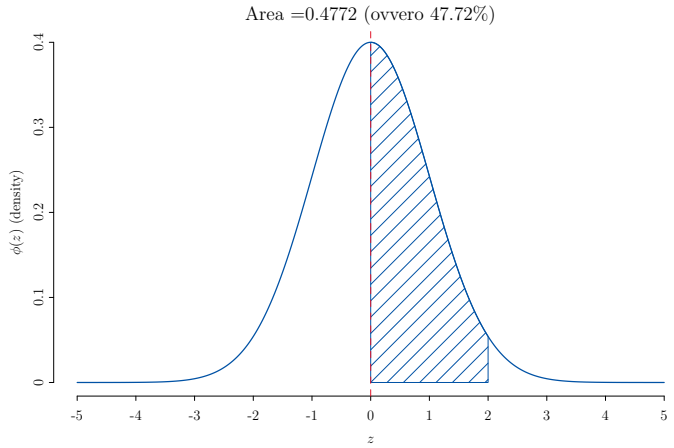
Quindi

- l'area sotto ϕ deve essere pari ad uno $\implies \int_{-\infty}^{+\infty} \phi(z)dz = 1$
- se il mio data set avesse una distribuzione *approssimativamente* simile al riferimento normale standard la frequenza f dei dati contenuti in $(a, b]$ sarebbe *approssimativamente*

$$f = \int_a^b \phi(z)dz$$

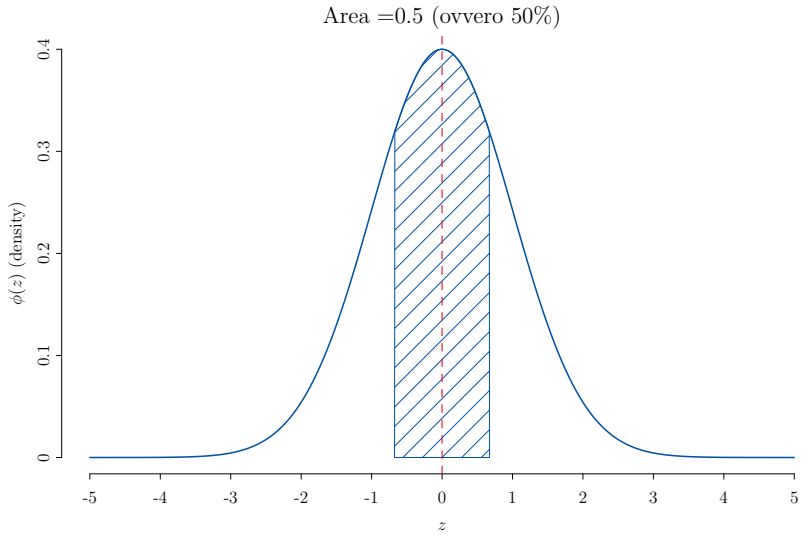
Notes

Domanda: se il mio data set avesse una distribuzione approssimativamente simile a quella normale standard, quante osservazioni mi devo aspettare nell'intervallo $[0, 2]$?



Notes

Quale sarebbero il bulk dei dati? Ovvero il 50% dei dati più centrali?



Notes

Proprietà rilevanti del modello normale

Unimodalità

la densità del modello normale standard ha un unico massimo, ovvero

$$\text{Mod}(z) = 0 = \bar{z}$$

Simmetria

esiste un punto $z = 0$ $(= \text{Mod}(z) = \bar{z})$ tale che la funzione di densità si comporta in modo speculare rispetto ad esso

Decadimento esponenziale delle code

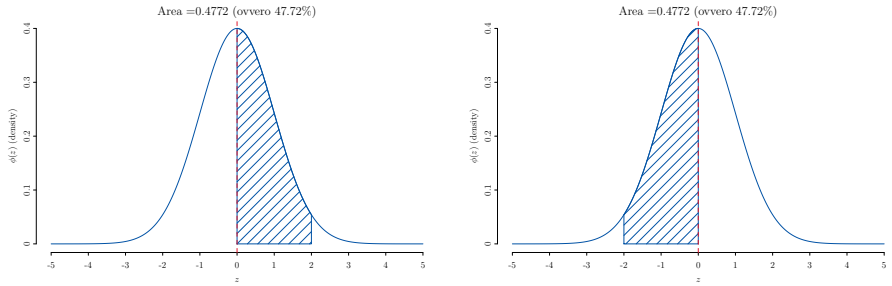
nel modello normale la densità dei dati nella code della distribuzione si annulla piuttosto velocemente $\Longrightarrow$ non ci aspettiamo dati estremamente lontani dal centro della distribuzione. Su questa nozione ci torneremo presto.

Nota: collocazione della moda, simmetria, e comportamento delle code qualificano quello che in letteratura si chiama *analisi della forma di una distribuzione* (dall'inglese: *shape*)

Notes

Conseguenze della simmetria

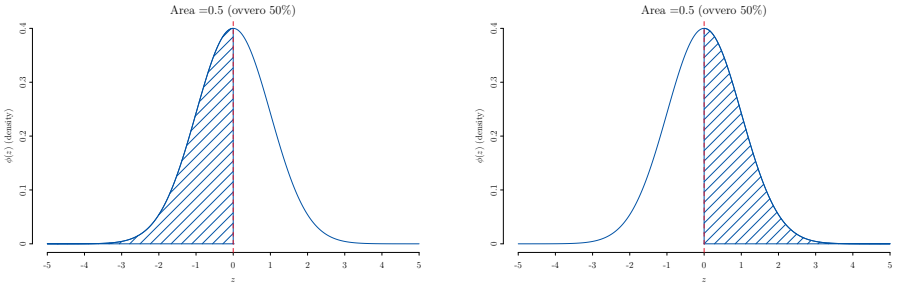
Ci aspettiamo la stessa massa di dati in intervalli *simmetrici* intorno alla media=0



Notes

Per un modello di densità unimodale e simmetrico come quello normale standard, ci aspettiamo che

$$\text{moda} = \text{media} = \text{mediana}$$



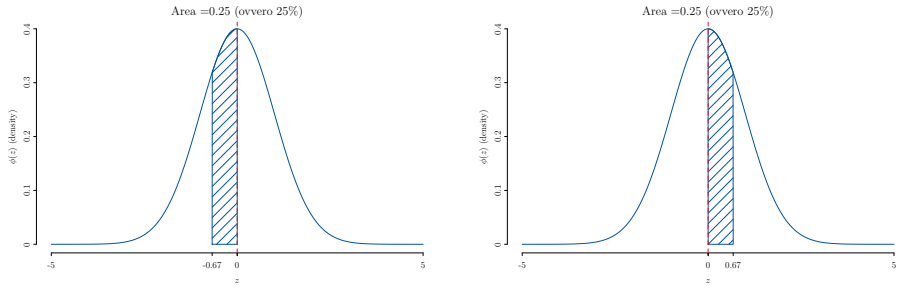
Notes

Per un modello di densità simmetrico come quello normale standard, ci aspettiamo che fissato $\alpha \in (0, 1)$

$$\text{Med}(Z) - Q_{\alpha\%} \;=\; Q_{(1-\alpha)\%} - \text{Med}(Z)$$

In particolare, per $\alpha = 25\%$

$$\text{Med}(Z) - Q_{25\%} = \; Q_{75\%} - \text{Med}(Z) \;\;\Longrightarrow\;\; Q_2 - Q_1 \;= Q_3 - Q_2$$



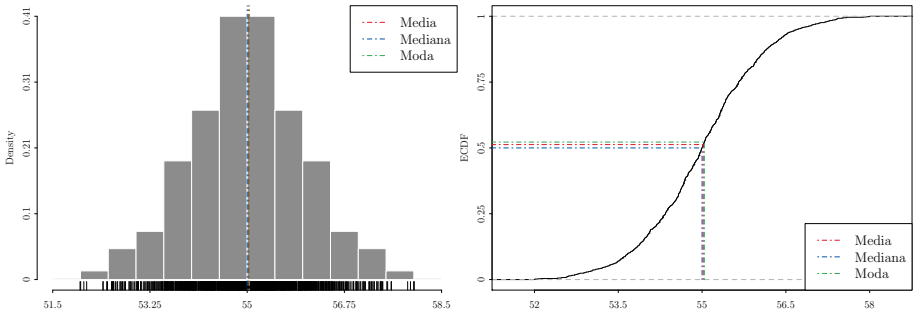
$$Q_{0.25\%} = -0.67, \quad \text{Med}(Z) = 0, \quad Q_{0.75\%} = 0.67$$

Notes

Simmetria, asimmetria, e centralità

Distribuzioni unimodali simmetriche

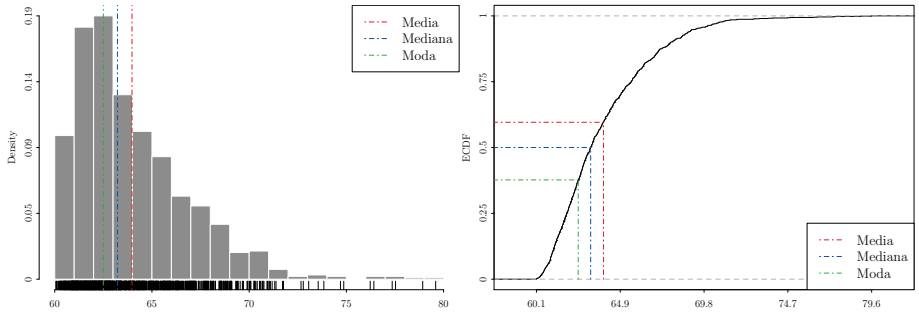
- Moda = Mediana = Media
  le tre misure di centralità sono equivalenti
- valgono le osservazioni a proposito della simmetria del modello normale standard



Notes

Distribuzioni unimodali con asimmetria positiva

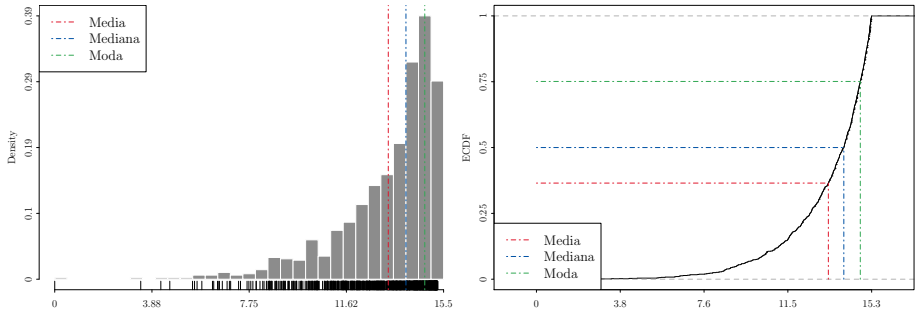
- Moda $<$ Mediana $<$ Media
  le tre misure di centralità non sono equivalenti
- la densità dei dati a sinistra della mediana è più alta che a destra
- le due code si comportano in modo molto diverso



Notes

Distribuzioni unimodali con asimmetria negativa

- Media $<$ Mediana $<$ Moda
  le tre misure di centralità non sono equivalenti
- la densità dei dati a destra della mediana è più alta che a sinistra
- le due code si comportano in modo molto diverso



Notes

Un modo per caratterizzare la simmetria, è il confronto con un caso ideale di simmetria come quello normale standard

Supponiamo di osservare $\{x_1, x_2, \ldots, x_n\}$, e definiamo il coefficiente di asimmetria di Pearson

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)^3 \tag{6.1}$$

Si noti che γ_1 può essere visto come il momento terzo dei dati standardizzati

- $\gamma_1 \approx 0$: evidenza di simmetria nei dati
- $\gamma_1 > 0$: evidenza di asimmetria positiva nei dati
- $\gamma_1 < 0$: evidenza di asimmetria negativa nei dati nei dati

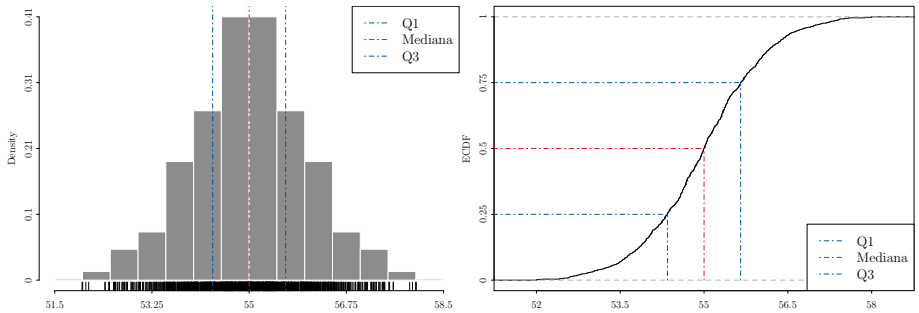
⚠ Nei dati possiamo trovare $\gamma_1 = 0.3$, è abbastanza piccolo/grande?

Notes

Simmetria, asimmetria, quantili

Distribuzioni simmetriche

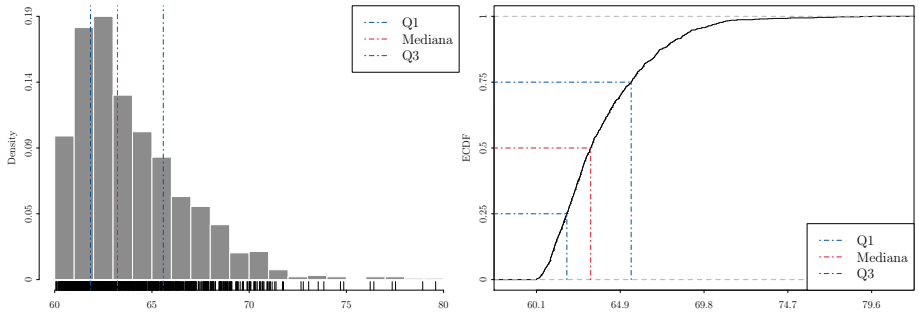
- alla stessa distanza a destra oppure a sinistra della mediana i dati sono ugualmente densi
- Mediana$-Q_{25\%}$ $=$ $Q_{75\%}-$Mediana



Notes

Distribuzioni con asimmetria positiva

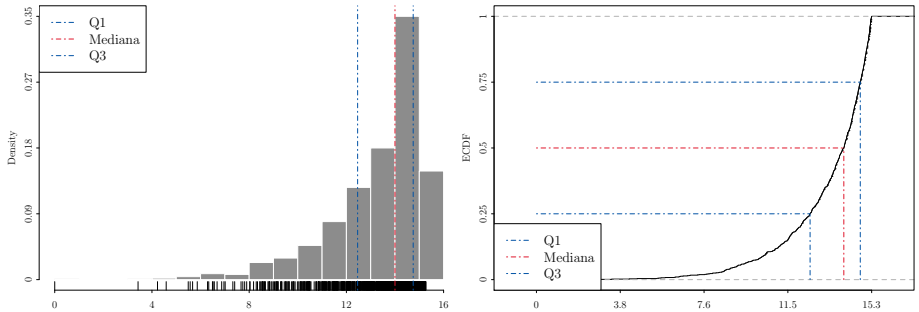
- mettendosi alla stessa distanza dalla mediana, i dati alla sua sinistra sono più densi che alla sua destra
- Mediana$-Q_{25\%} \ < \ Q_{75\%}-$Mediana



Notes

Distribuzioni con asimmetria negativa

- mettendosi alla stessa distanza dalla mediana, i dati alla sua sinistra sono meno densi che alla sua destra
- Mediana$-Q_{25\%} \ > \ Q_{75\%}-$Mediana



Notes

Queste considerazioni ci suggeriscono un altro strumento numerico per misurare la simmetria/asimmetria. Coefficiente di Yule-Bowley

$$B_1 = \frac{(Q_3 - \text{Med}(X)) \; - \; (\text{Med}(X) - Q_2)}{Q_3 - Q_2} = \frac{Q_1 + Q_3 - 2Q_2}{\text{IQR}(X)} \tag{6.2}$$

Nota: γ_1 non dipende dall'unità di misura

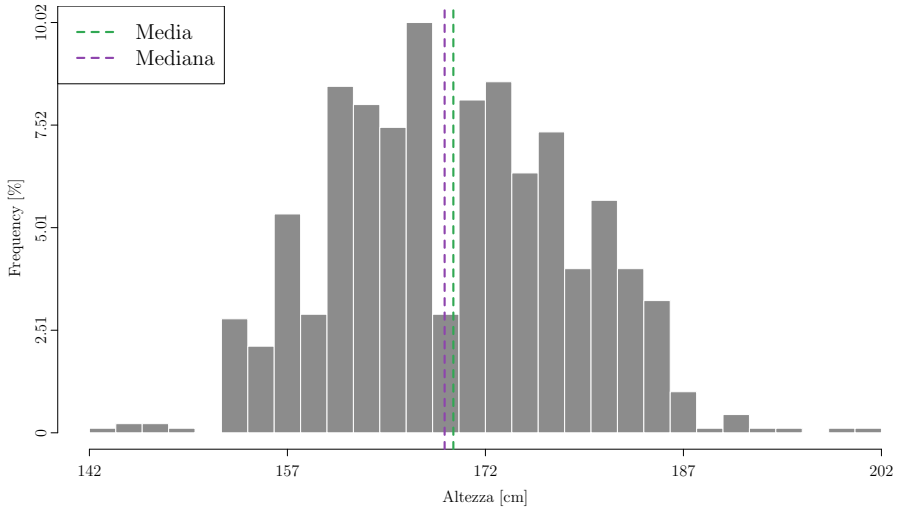
- $B_1 \approx 0$: evidenza di simmetria nei dati
- $B_1 > 0$: evidenza di asimmetria positiva nei dati
- $B_1 < 0$: evidenza di asimmetria negativa nei dati

Esempi/Applicazioni $\longrightarrow$ `R` **script file**

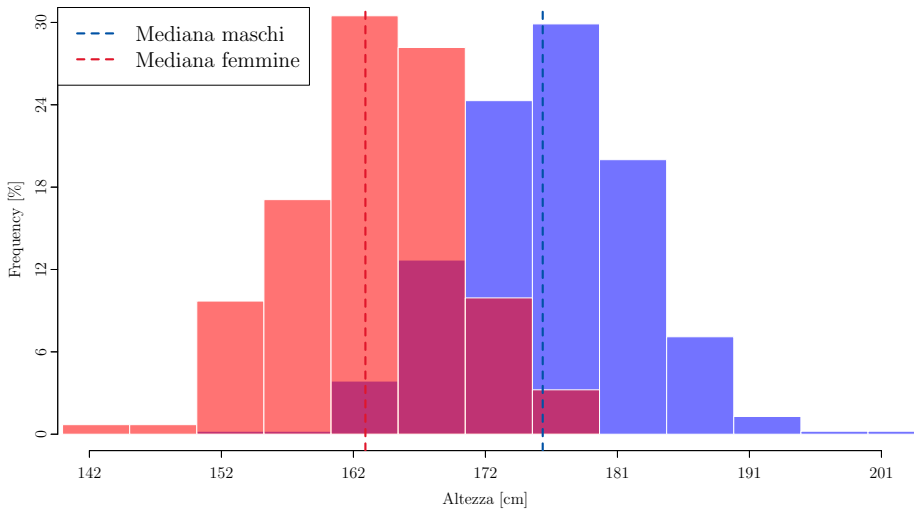
Notes

Centralità e distribuzioni multimodali

Dati: "*Galton's height data*", $n = 898$ soggetti



Notes


Mediana maschi
Mediana femmine
Frequency [%]
0
6
12
18
24
30
142
152
162
172
181
191
201
Altezza [cm]

Notes

Notes