Slide Set 8
Boxplot e Kernel Smoothing

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

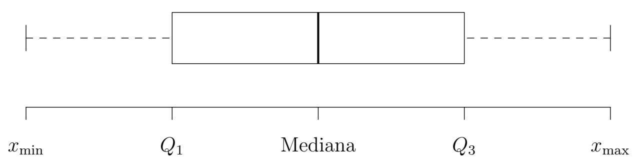
`Versione: 22 febbraio 2022 (h10:23)`

Notes

Box and whiskers plot (boxplot)

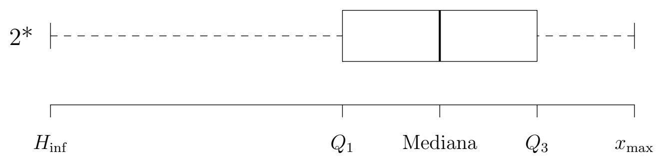
- Una rappresentazione semplice e potente di una distribuzione univariata (proposta da J. W. Tukey)
- Si basa su $x_{\min}, Q_1, \mathrm{Med}, Q_3, x_{\max}$ e $H_{\inf}, H_{\sup}$ calcolati con $k = 1.5$
- Mostra tutto o quasi: posizione, dispersione, simmetria, code, e sospetti outliers

In assenza di sospetti outliers rappresentiamo la distribuzione con: (i) due scatole detti *Boxes* (il bulk); (ii) due baffi (le code)



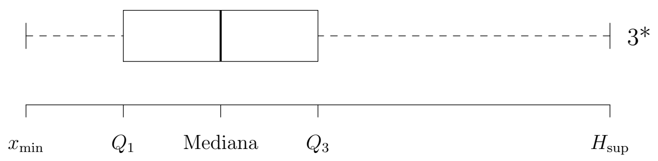
Notes

Se ci fossero stati solo valori eccezionalmente piccoli ($< H_{\text{inf}}$)



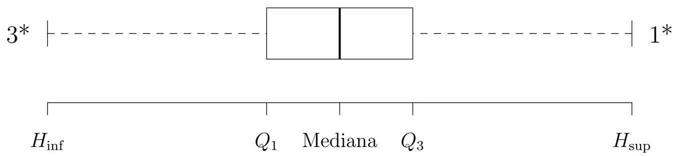
Notes

Se ci fossero stati solo valori eccezionalmente grandi ($> H_{\text{sup}}$)



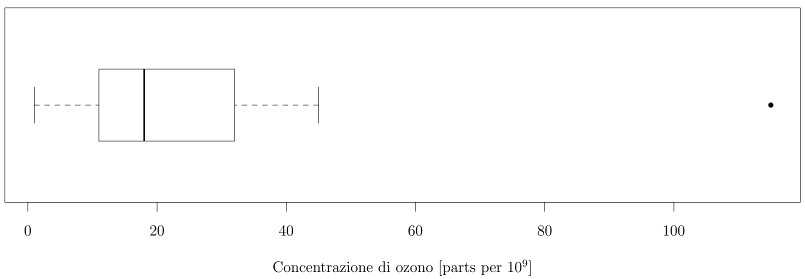
Notes

Ed infine con valori eccezionalmente grandi/piccoli



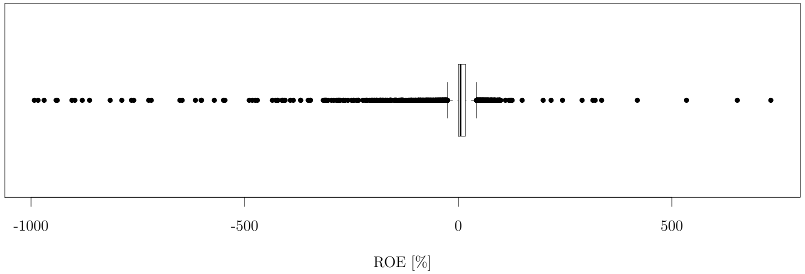
Notes

Questo sarebbe il boxplot "*ortodosso*" così come pensato da Tukey. Spesso il range non è così enorme e si rappresenta tutto. Ad esempio per il data set "*Air Quality*" R produce



Notes

. . . Tuttavia se le code sono “molto pesanti”, come nel caso dei dati sul ROE



Lo scaling del boxplot “non ortodosso” non ci consentirebbe di vedere la parte centrale dei dati

Notes

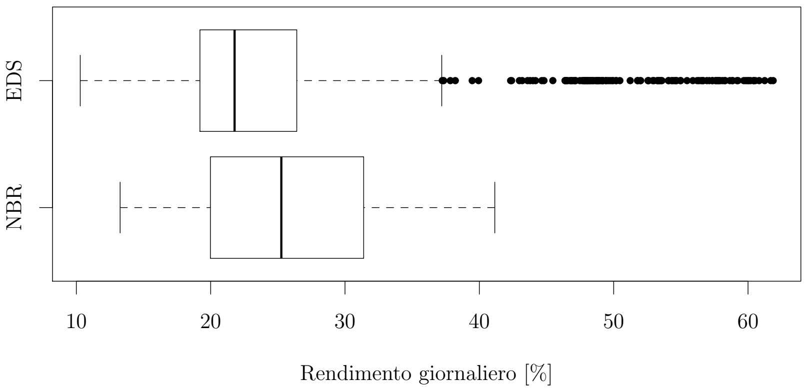
Il boxplot consente di vedere tanto in un solo grafico

- Posizione $\longleftarrow$ mediana
- Dispersione $\longleftarrow$ IQR
- Asimmetria $\longleftarrow$ segmenti $(Q_3 - \mathrm{Med})$ vs $(\mathrm{Med} - Q_1)$
  - Simmetria: Box SX = Box DX
  - Asimmetria positiva: Box SX più corto del Box DX
  - Asimmetria negativa: Box SX più lungo del Box DX
- Outliers e lunghezza delle code $\longleftarrow$ baffi

I boxplot paralleli consentono di fare confronti utilissimi

Notes

Dati: rendimenti giornalieri

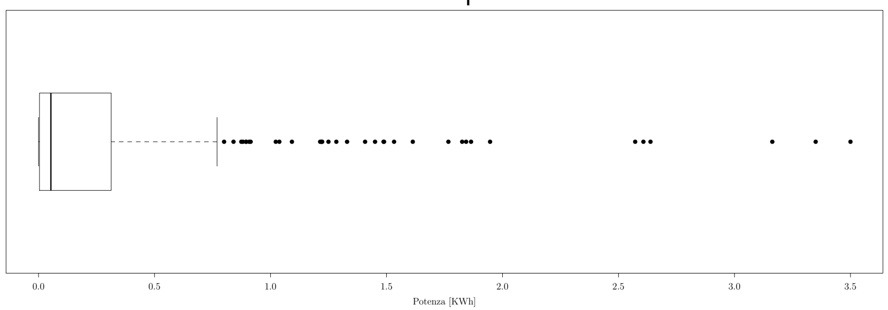


Notes

⚠ Classificazione degli outliers basata sui "*fences*": si basa sull'ipotesi che i dati *regolari* al centro sono approssimativamente simmetrici come nel caso normale standard.

In caso di forti asimmetrie, i "*sospetti outliers*" del boxplot sono spesso il normale prodotto del processo generatore dei dati

Dati: assorbimento di potenza elettrica



Esempi/Applicazioni ⟶ R script file

Notes

Bandwidth ottimale nell'istogramma

Ritorniamo al problema dell'approssimazione della densità dei dati in un intorno di un punto t

$$\text{densità}(t) \;=\; \frac{\text{massa (peso) dei dati intorno a } t}{\text{ampiezza dell'intorno } t} \;=\; \frac{\mathcal{K}}{\Delta}$$

Consideriamo una classe di valori $[t_k, t_{k+1}]$ di ampiezza Δ. Prendiamo il midpoint t della classe e riscriviamo l'intervallo come

$$\left[t - \frac{\Delta}{2}\ ,\ t + \frac{\Delta}{2}\right] \quad = \quad [t - b\ ,\ t + b] \quad = \quad [t_k, t_{k+1}]$$

dove adesso $\Delta = 2b$. La quantità b prende il nome di bandwidth = larghezza di banda. b ci dice quanto siamo vicino/locali rispetto a t (da Dx/Sx)

Notes

Quanto vale l'istogramma nell'intervallo $[t_k, t_{k+1}]$? Vale

$$\frac{f_t}{\Delta} = \frac{f_t}{2b} = \frac{1}{n}\,\frac{1}{b}\,\frac{n_t}{2}$$

dove f_t e n_t sono le frequenze assolute e relative delle osservazioni in $[t - b, t + b]$.

Definiamo la funzione indicatrice

$$\mathbf{1}\{x_i \in [t_k, t_{k+1}]\} := \begin{cases} 1 & \text{se } x_i \in [t_k, t_{k+1}] \\ 0 & \text{altrimenti} \end{cases}$$

Ovviamente sommando sulle x_i otteniamo il conteggio delle x_i che cadono in $[t_k, t_{k+1}]$, quindi

$$n_t = \sum_{i=1}^{n} \mathbf{1}\{x_i \in [t_k, t_{k+1}]\}$$

Notes

Su tutto $[t_k, t_{k+1}]$, il valore dell'istogramma è costante e uguale al valore che assume al centro della classe (ovvero nel punto t)

$$h(t) = \frac{1}{n}\frac{1}{b}\frac{n_t}{2} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\underbrace{\frac{\mathbf{1}\{x_i \in [t_k, t_{k+1}]\}}{2}}_{\mathcal{K}} \tag{8.1}$$

Dove

- b = bandwidth (= larghezza di banda = ampiezza della finestra)
  definisce l'*ampiezza dello zooming* che facciamo intorno al punto t, fissa **quanto localmente** guardiamo ai dati **intorno** a t (a Dx/Sx)
- Funzione Kernel
  corrisponde $\mathcal{K}$ nella (8.1). Questa è la **bilancia** che usiamo per **pesare** i **dati** che cadono in $[t_k\ ,\ t_{k+1}]$

Osserva
Nell'istogramma *pesiamo* i dati con un *Kernel rettangolare*, ovvero tutti i dati che cadono in un bin $[t-b, t+b]$ pesano tutti 1/2, quelli fuori pesano 0.

Notes

In assenza di ogni informazione sui dati, costruiamo l'istogramma con classi uniformi di ampiezza $\Delta = 2b$. La scelta *cruciale* è l'ampiezza di banda b.

- si è visto che regole come quella di Sturges, spesso tendono a fornire un numero di classi, e quindi un'ampiezza di banda b, non opportuni
- ragionando su $n \to \infty$ abbiamo visto che è necessario che $\Delta = 2b$ sia via via più piccolo al crescere di n.

Esistono *metodi di scelta di b data-driven* (ovvero basati esclusivamente sui dati osservati), che consento di ottimizzare un qualche aspetto dell'approssimazione della densità (i perché/come saranno chiari quando diventerete grandi!)

Notes

Banda ottimale di Scott (1979)

$$b_S = \frac{1}{2} 3.5 \frac{s_X^2}{n^{\frac{1}{3}}}, \quad \text{quindi} \quad \Delta_S = 3.5 \frac{s_X^2}{n^{\frac{1}{3}}}$$

quindi facciamo un *windowing* dei dati con finestre di ampiezza Δ_S

Banda ottimale di Freedman–Diaconis (1981)

$$b_{FD} = \frac{\text{IQR}(x)}{n^{\frac{1}{3}}}, \quad \text{quindi} \quad \Delta_{FD} = 2\frac{\text{IQR}(x)}{n^{\frac{1}{3}}},$$

quindi facciamo un *windowing* dei dati con finestre di ampiezza Δ_{FD}

⚠ In assenza di informazione su variabili/dati, si consiglia come primo tentativo di usare sempre la banda ottimale di Freedman–Diaconis.

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

Notes

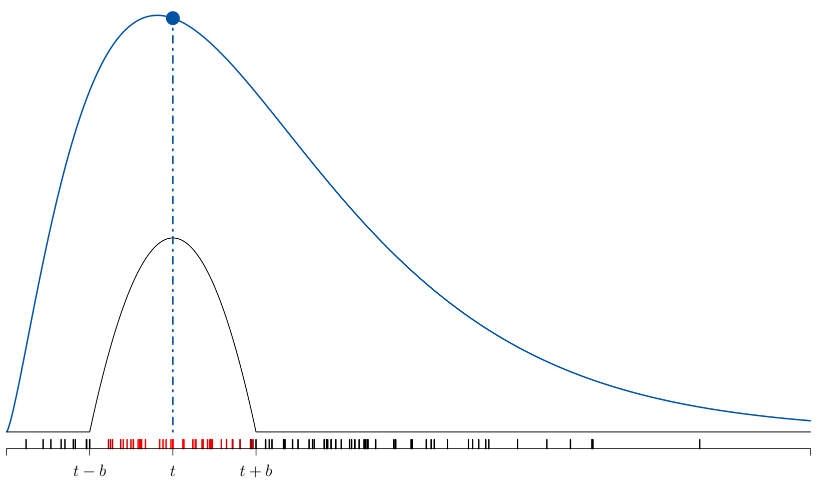
Kernel density

Nell'istogramma il *kernel rettangolare* pesa ogni osservazione con peso 1/2 oppure 0.

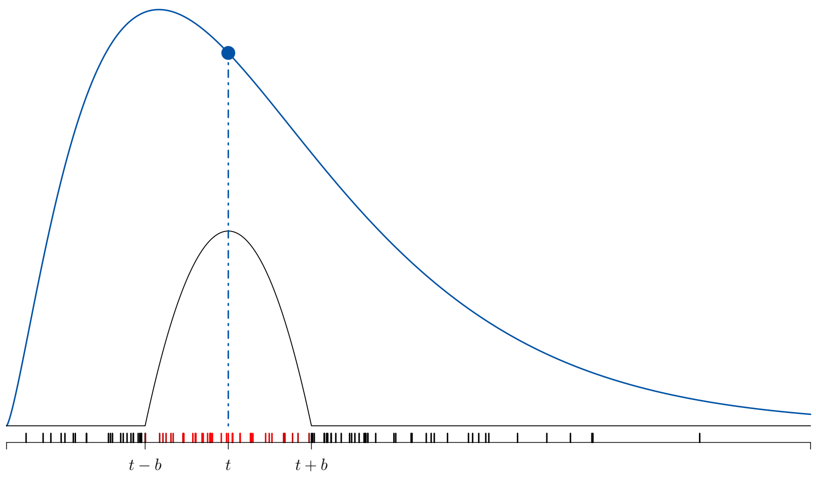
Immaginiamo invece di pesare le osservazioni intorno a t con un peso che decresce simmetricamente spostandosi a da t in direzione Dx/Sx

In pratica muovendo t lungo il range $[x_{\min}, x_{\max}]$, pesiamo ogni volta le osservazioni con una funzione kernel (la bilancia) $\mathcal{K}$ che assegna un peso relativamente maggiore alle osservazioni in prossimità di t

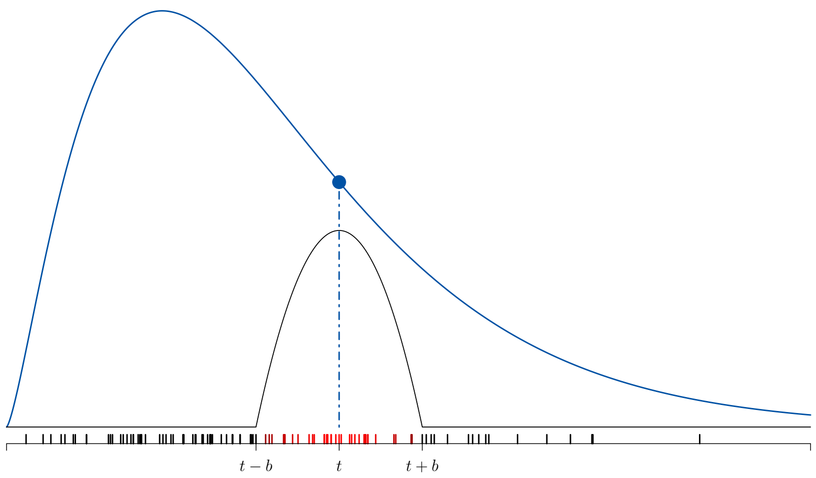
Notes



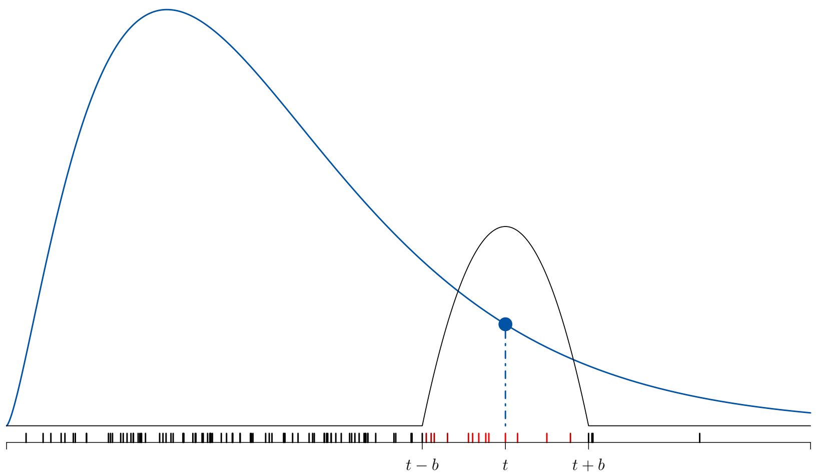
Notes



Notes



Notes



Notes

Smoothing

fissato un punto t, ed un'ampiezza di banda b, definiamo una funzione kernel $\mathcal{K}$ (una bilancia) tale che

$$2\,\mathcal{K}(x_i; t, b)$$

pesa il contributo dell'osservazione x_i alla massa di dati intorno t

La massa di dati intorno a t la misuriamo come media empirica dei contributi

$$\text{massa dei dati intorno a } t \;=\; \frac{1}{n}\sum_{i=1}^{n} 2\mathcal{K}(x_i; t, b)$$

dove x_i pesa sempre meno man mano che ci allontaniamo da t, con un peso nullo all'esterno di $[t-b\ ,\ t+b]$

Notes

Fissata un'ampiezza di banda b l'approssimazione kernel della densità in t è data da

$$\begin{aligned}\hat{f}(t) &= \frac{\text{massa dei dati intorno a } t}{\text{ampiezza dell'intorno di } t}\\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} 2\mathcal{K}(x_i; t, b)}{2b}\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\,\mathcal{K}(x_i; t, b) \qquad (8.2)\end{aligned}$$

Osserva

- la (8.2) coincide con la riscrittura dell'istogramma (8.1), cambia solo la forma della *bilancia* $\mathcal{K}$
- il vantaggio è che, facendo variare continuamente $t \in [x_{\min}, x_{\max}]$ e calcolando ogni volta $\hat{f}(t)$, eliminiamo le discontinuità dell'istogramma perché le osservazioni entrano ed escono dalle finestre in modo graduale

Notes

L'approssimazione kernel della densità richiede due scelte: $\mathcal{K}$ e b

Scelta di $\mathcal{K}$ = funzione kernel = la bilancia

- nella maggior parte dei casi il suo impatto non è enorme
- in letteratura esistono tantissime funzioni kernel
- deve obbedire a certe condizioni di regolarià (per adesso non entriamo nel merito)

Nell'esempio grafico abbiamo usato la funzione kernel di Epanechnikov:

$$\mathcal{K}(x_i; t, b) := \frac{3}{4}\left(1 - \left(\frac{t - x_i}{b}\right)^2\right) \tag{8.3}$$

Fissato b, la densità nel punto t si ottiene sostituendo la (8.3) in (8.2):

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\frac{3}{4}\left(1 - \left(\frac{t - x_i}{b}\right)^2\right)$$

Notes

Scelta di b = bandwidth = ampiezza di banda

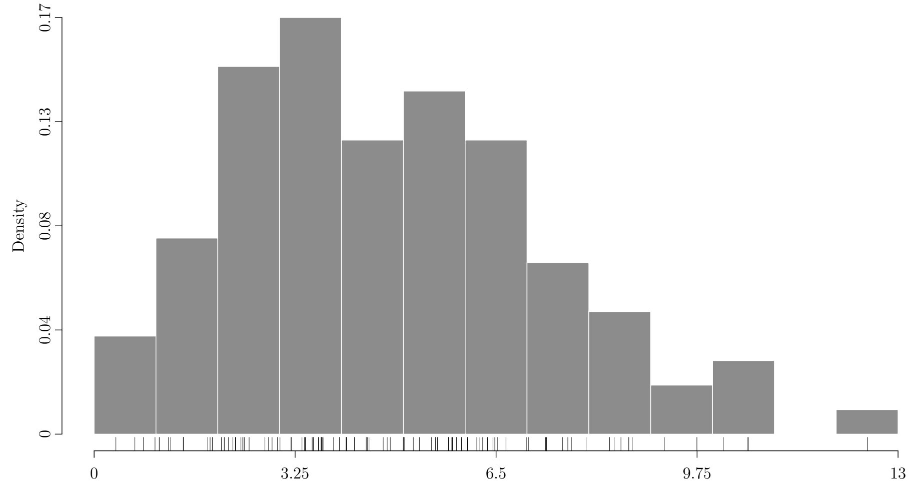
- ha un grande impatto sull'approssimazione finale
- come per l'istogramma deve diventare piccola mano mano che n cresce. Deve adattarsi alla scala dei dati e quindi alla variabilità.
- difficilmente si ha un'idea del livello ottimale di b (livello di smoothing). Ci affidiamo quasi sempre a regole ottimali (come nell'istogramma)

⚠ b ha un impatto enorme perché stabilisce "*quanto localmente*" guardiamo la distribuzione dei dati

⚠ Non dimentichiamo che la densità è una nozione che riguarda i soli dati continui

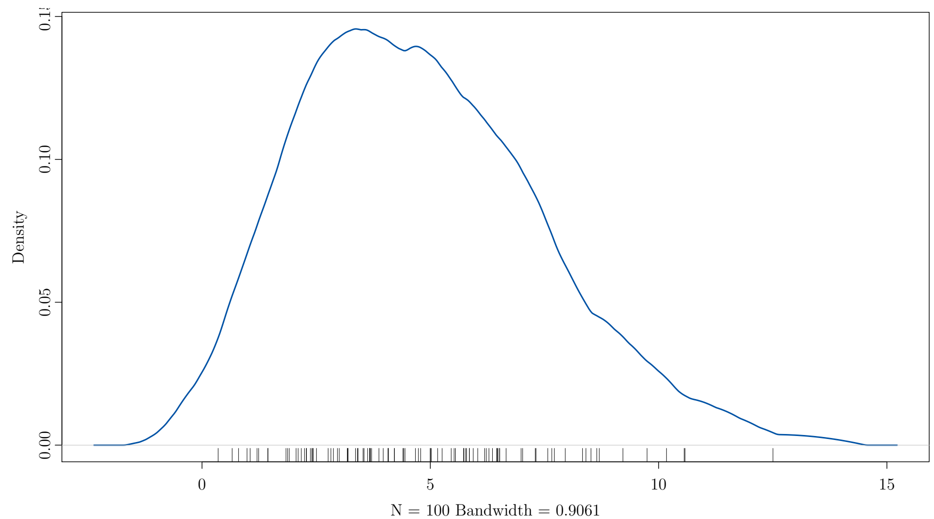
Notes

Esempio: istogramma con banda ottimale di Freedman-Diaconis



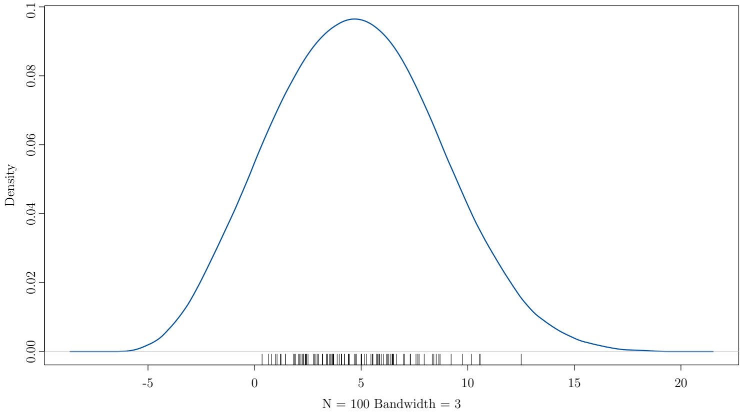
Notes

Kernel density con kernel di Epanechnikov, e "*bandwidth ottimale*" derivata con il metodo *default* di R. Si seleziona $b = 0.9061$



Notes

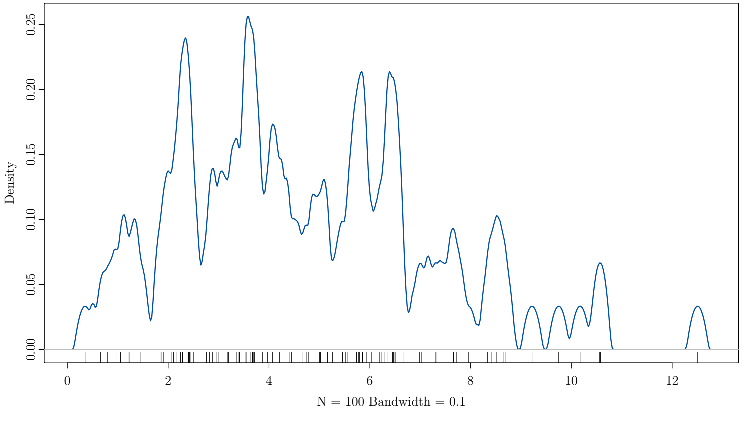
Funzione kernel=Epanechnikov, fissiamo $b = 3$, questo è circa 3 volte il valore ottimale (0.9061)



Osserva: b è troppo grande. Questo effetto è detto "*over-smoothing*": si includono troppi punti in ogni finestra omogenizzando il risultato. Nota che adesso la distribuzione appare addirittura simmetrica

Notes

Funzione kernel=Epanechnikov, fissiamo $b = 0.1$, questo è circa un decimo del valore ottimale (0.9061)



Osserva: b è troppo piccolo! Questo effetto è detto "*under-smoothing*": kernel pesa i punti troppo localmente, in alcuni casi cattura intorni che contengono una sola osservazione!

Notes