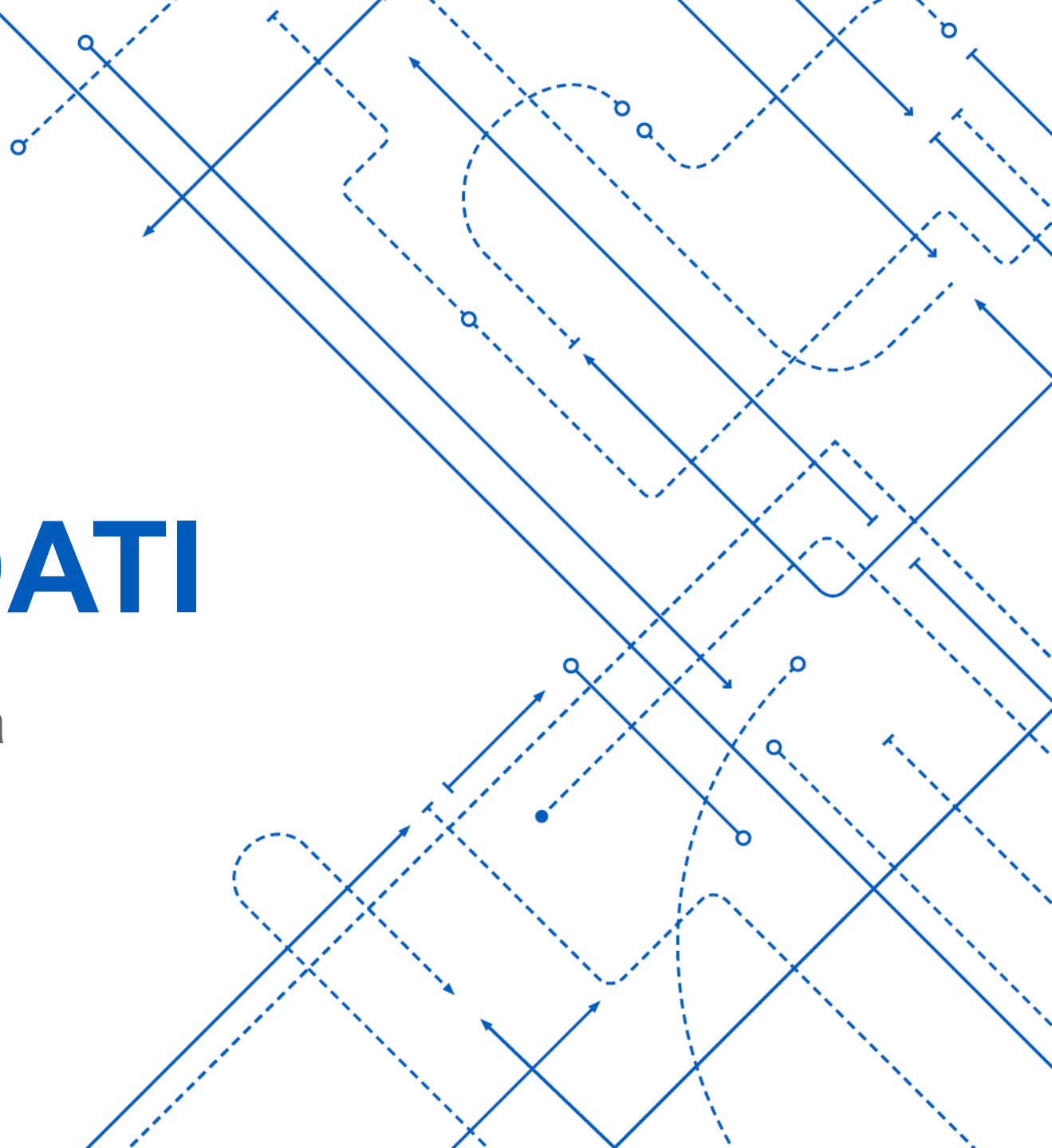
STATISTICA E ANALISI DEI DATI

Capitolo 3 – Da Istogrammi fino a Boxplot ad intaglio

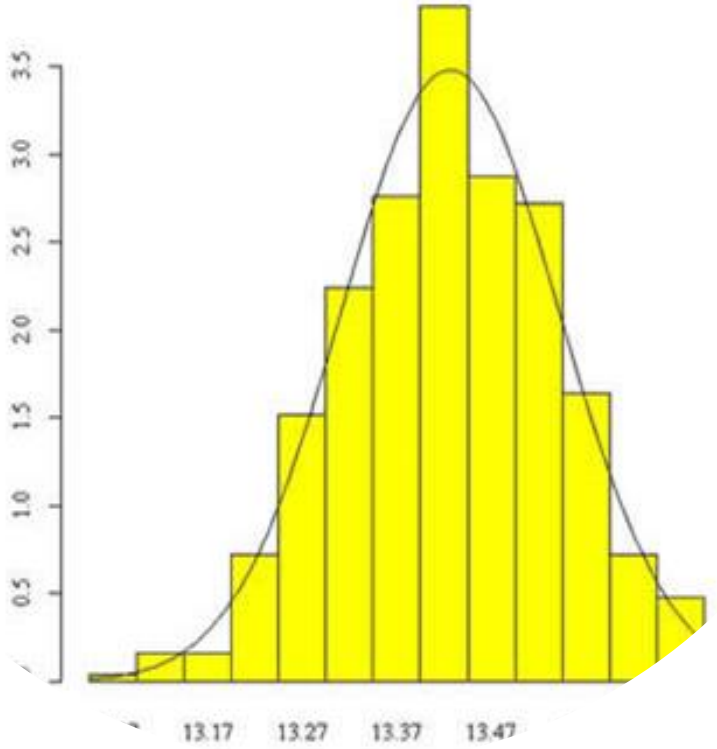
Dott. Stefano Cirillo
Dott. Luigi Di Biasi

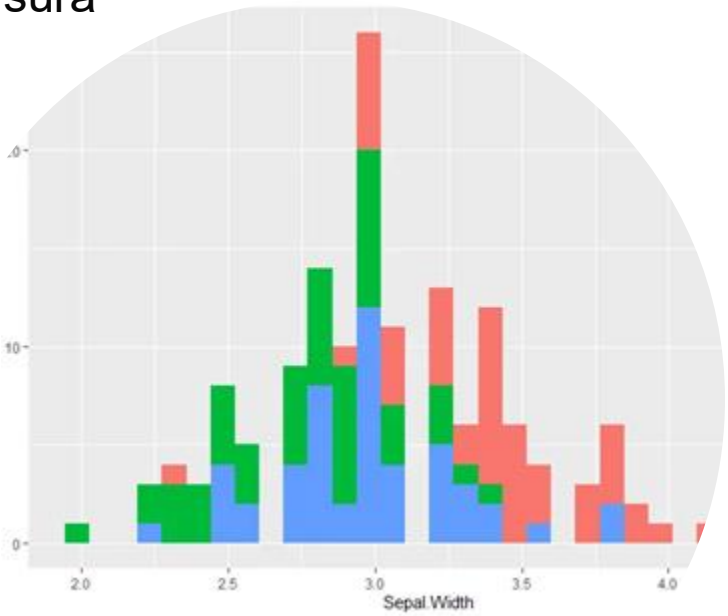
a.a. 2025-2026

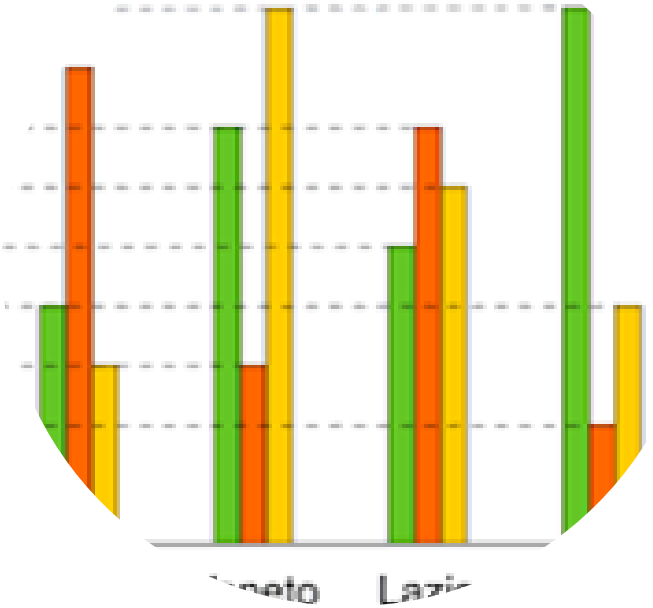


ISTOGRAMMI

- Si utilizzano per variabili quantitative
  - Sono una particolare rappresentazione grafica di una distribuzione di frequenza in classi
- Consideriamo un campione $(x_1, x_2, ..., x_n)$ costituito da *n* osservazioni, che suddividiamo in classi
  - ogni osservazione deve cadere in **una ed una** sola classe (o intervallo)
- L'asse delle ascisse di un istogramma è quindi sempre dotato di un'unità di misura





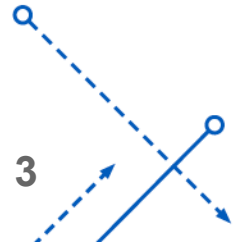


FUNZIONE HIST

- **hist** genera istogrammi utilizzando un vettore numerico

Esempi:

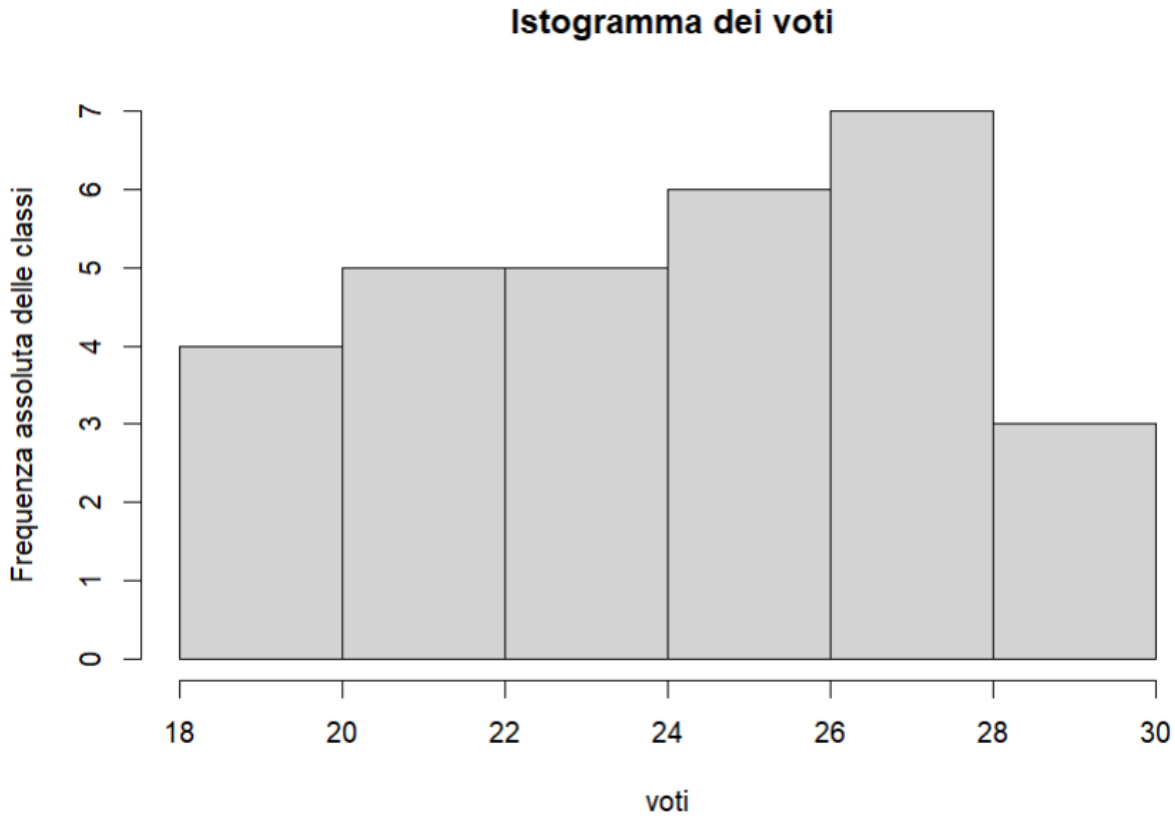
- **hist**(x):
  genera un istogramma utilizzando il vettore numerico x
- **hist**(x, nclass=n):
  genera un istogramma con un numero n di classi
- **hist**(x, breaks=b, ...):
  i punti di break degli intervalli dei valori di x che delimitano le classi sono esplicitamente elencati con il parametro breaks
- **hist**(x, probability=TRUE)
  le colonne rappresentano frequenze relative invece che assolute



ISTOGRAMMA CON LE FREQUENZA ASSOLUTE

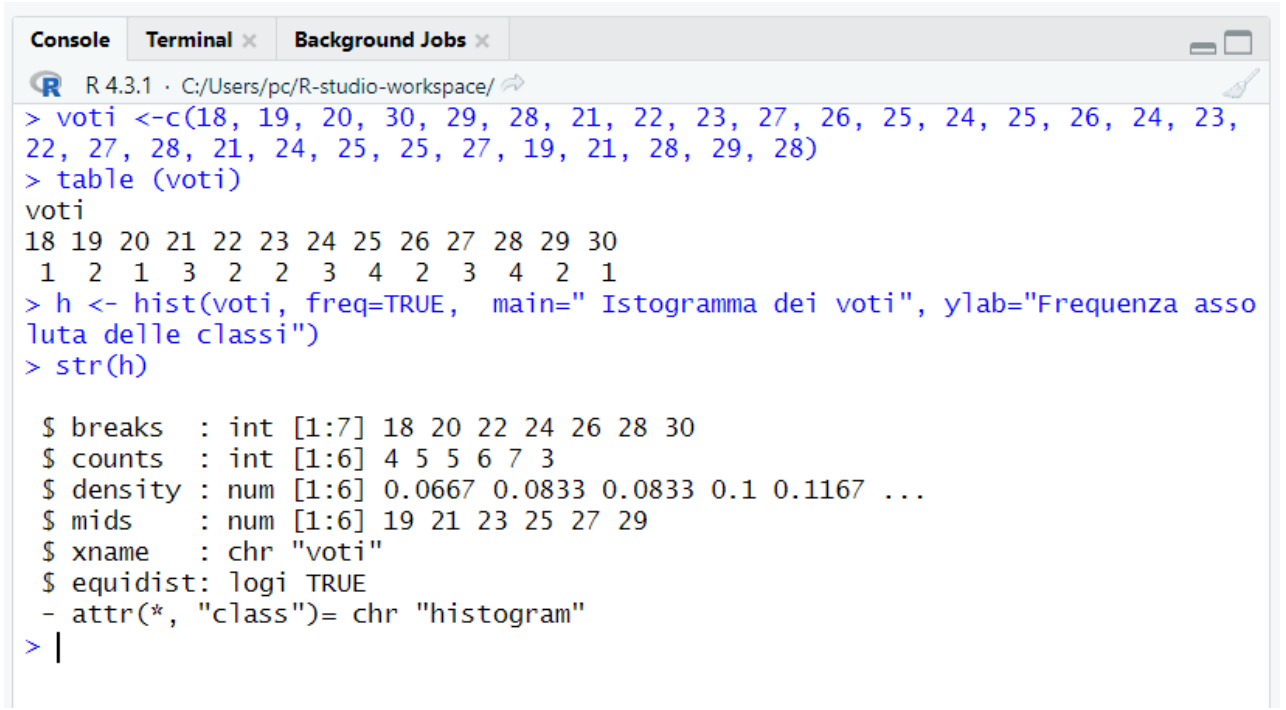
- Per realizzare un istogramma in base alle frequenze assolute, occorre impostare nella funzione **hist**() il parametro **freq = TRUE**

```
Console   Terminal ×   Background Jobs ×
R 4.3.1 · C:/Users/pc/R-studio-workspace/
> voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26,
25, 24, 25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27,
19, 21, 28, 29, 28)
> table (voti)
voti
18 19 20 21 22 23 24 25 26 27 28 29 30
 1  2  1  3  2  2  3  4  2  3  4  2  1
> hist(voti, freq =TRUE,  main=" Istogramma dei voti",
ylab=" Frequenza assoluta delle classi ")
>
```



FUNZIONE HIST SU INTERVALLI

- Se si salva l'istogramma in una variabile *h* è possibile visualizzare delle informazioni relative all'istogramma e ai dati
- Queste informazioni possono essere visualizzate utilizzando la funzione **str**(*h*) applicata alla variabile generata dalla funzione hist()
  - **breaks**: Punti di divisione delle classi
  - **counts**: Frequenze assolute delle classi
  - **density**: Densità delle classi
  - **mids**: Punti centrali delle classi

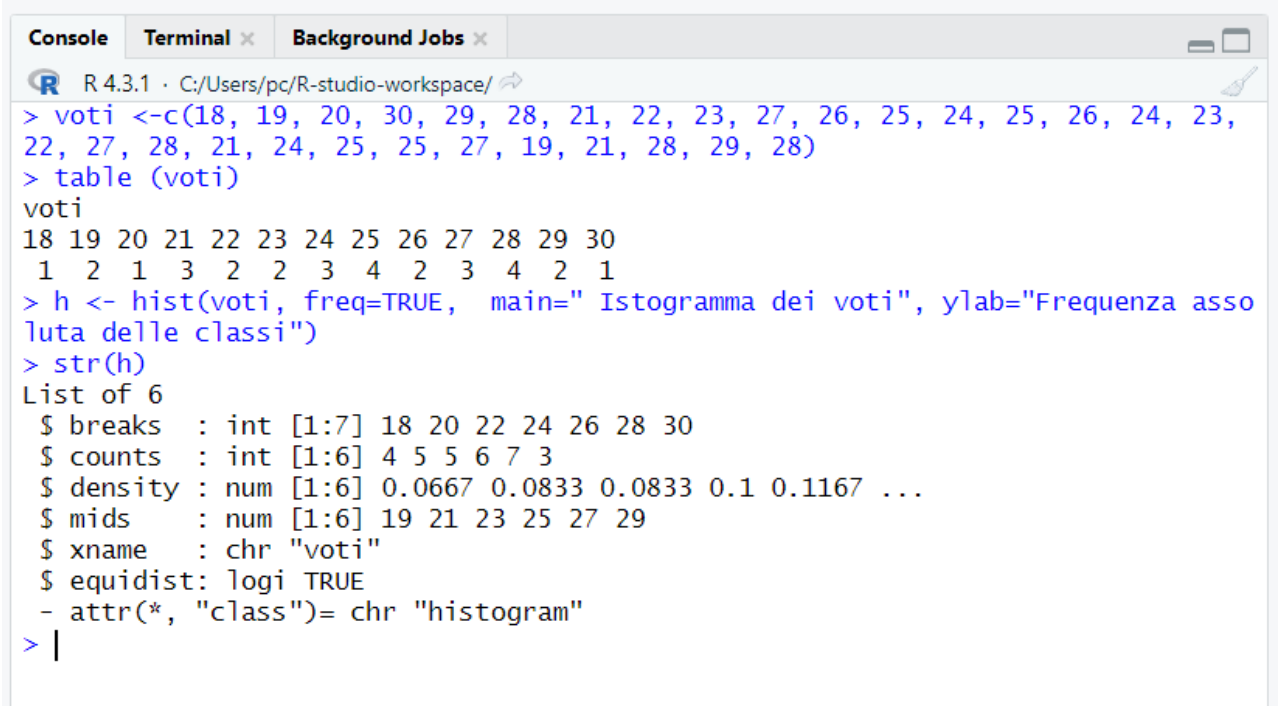


FUNZIONE HIST SU INTERVALLI

- Nell'esempio la suddivisione in classi scelta automaticamente da R è la seguente:

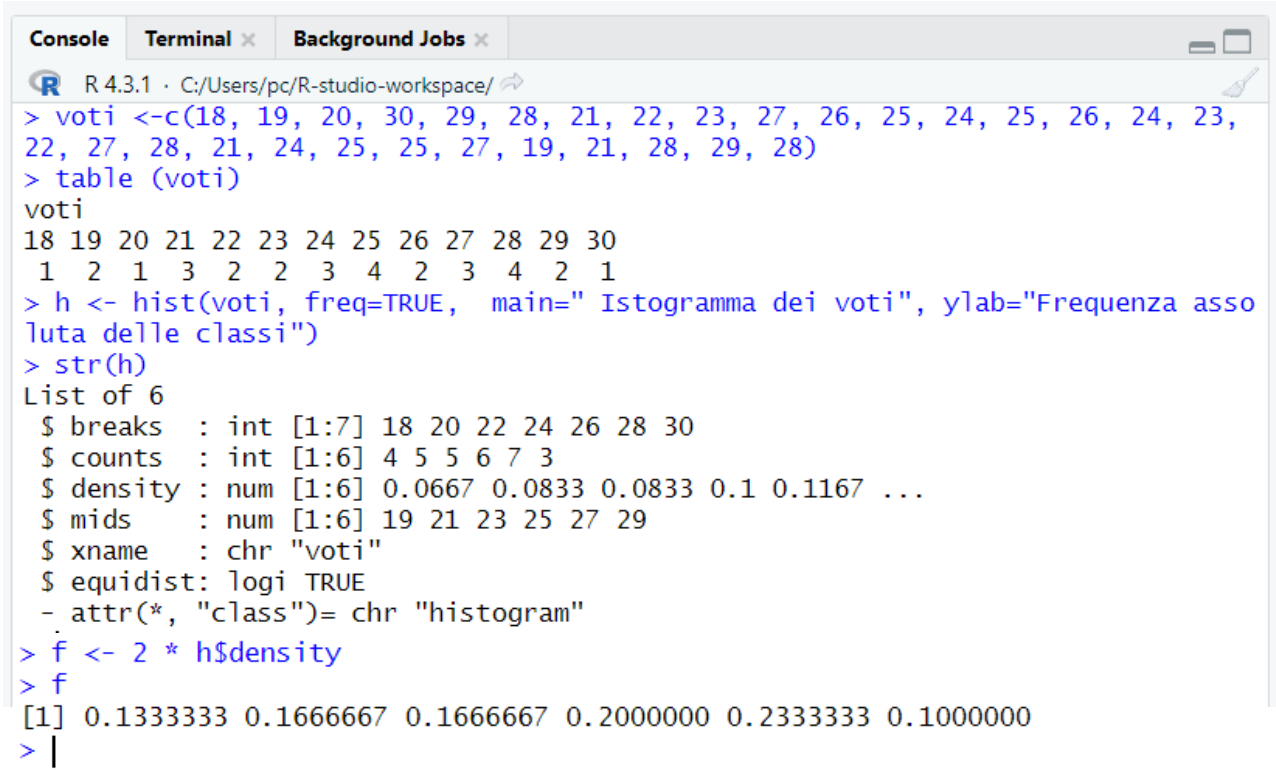
  (18, 20], (20, 22], (22, 24], (24, 26], (26, 28], (28, 30]

- Frequenza assoluta delle classi: 4 voti cadono nella prima classe, 5 nella seconda classe, 6 nella quarta classe, 7 nella quinta classe e 3 nella sesta classe



FUNZIONE HIST SU INTERVALLI

- Le frequenze relative associate alle sei classi possono essere ottenute moltiplicando gli elementi del vettore **h$density** per 2 (ampiezza effettiva di ogni classe)



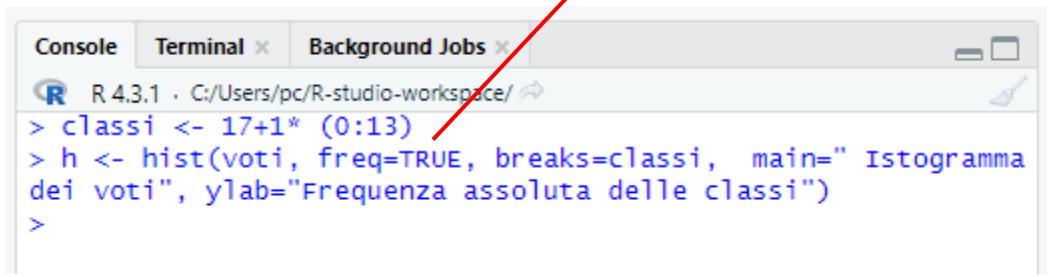
ossia $f_1 = 0.133$, $f_2 = 0.167$, $f_3 = 0.167$, $f_4 = 0.200$, $f_5 = 0.233$ e $f_6 = 0.100$

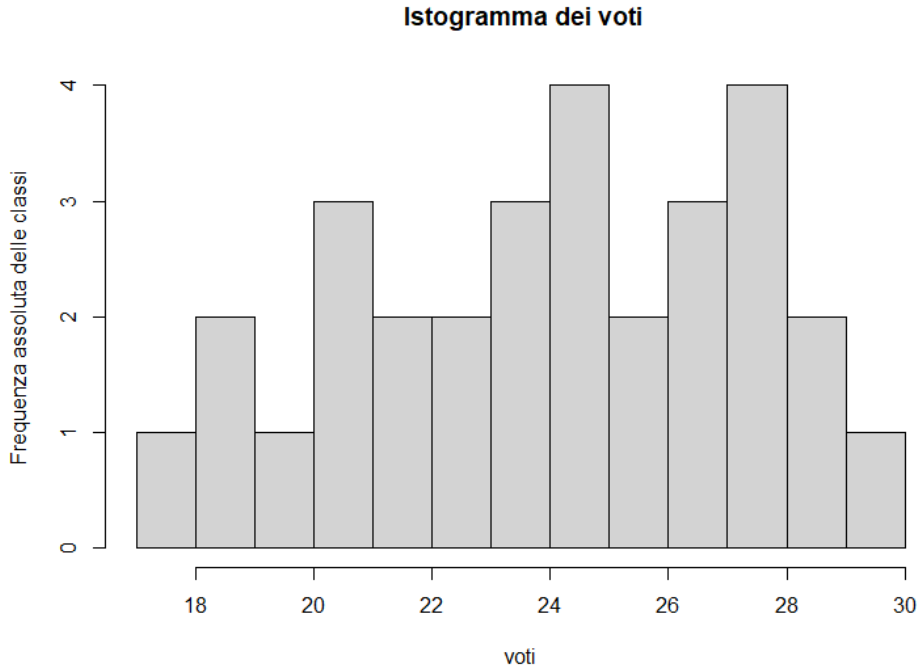
- **Nota**: la somma delle frequenze relative associate alle classi è unitaria

FUNZIONE HIST SU INTERVALLI

- La funzione **hist**() permette di fissare le classi dell'istogramma con **breaks**
  - Definizione di un vettore numerico per le classi

```
> 17+1*(0:13)
 [1] 17 18 19 20 21 22 23 24 25 26 27 28 29 30
```



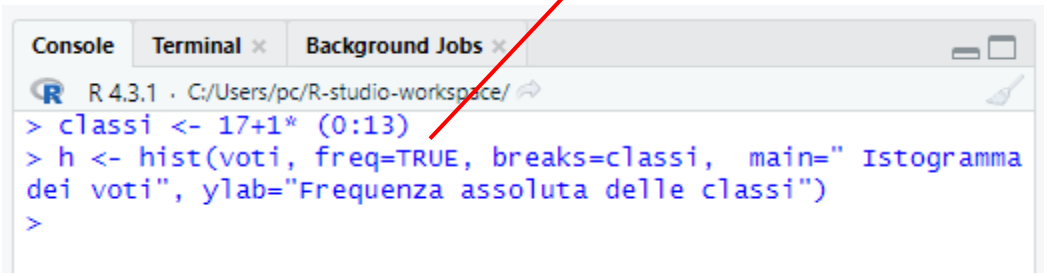


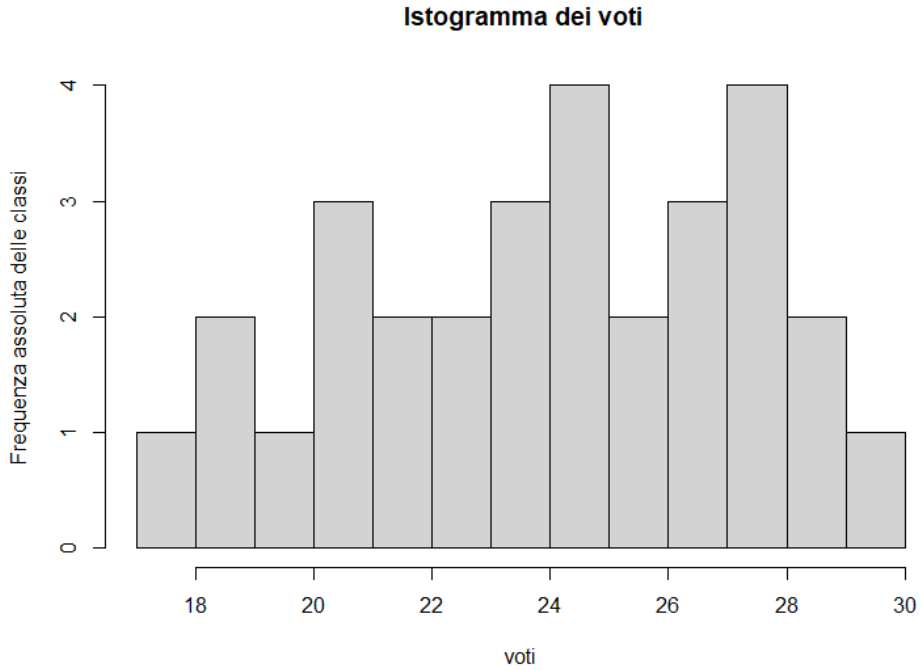
- Gli intervalli unitari sono aperti a sinistra e chiusi a destra, ossia del tipo $(k, k+1]$ dove k = 17,18, ..., 29
- **Nota**:
  - Frequenze Assolute: l'area totale è uguale all'ampiezza del campione

FUNZIONE HIST SU INTERVALLI

- La funzione **hist**() permette di fissare le classi dell'istogramma con **breaks**
  - Definizione di un vettore numerico per le classi

```
> 17+1*(0:13)
 [1] 17 18 19 20 21 22 23 24 25 26 27 28 29 30
```



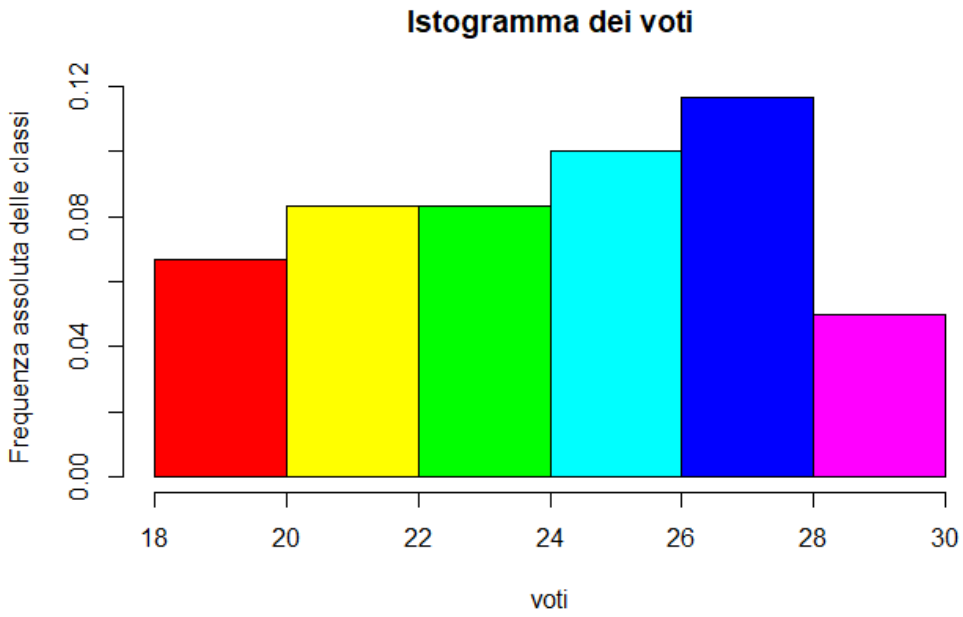


- **Nota**:
  - Frequenze Relative:
    - L'area del rettangolo *i*-esimo è uguale alla frequenza relativa della classe stessa: $f_i = \frac{n_i}{n} = b_i * n_i$
      - f_i è la frequenza relativa dei valori della classe *i*-esima;
      - b_i l'ampiezza e h_i l'altezza della classe i-esima .
    - L'altezza di ogni rettangolo esprime la **densità di frequenza** associata alla classe *i*-esima
    - L'area totale è uguale all'unità

ISTOGRAMMA CON LE FREQUENZE RELATIVE

- Per realizzare un istogramma in base alle frequenze relative, occorre impostare nella funzione **hist**() il parametro **freq = FALSE**

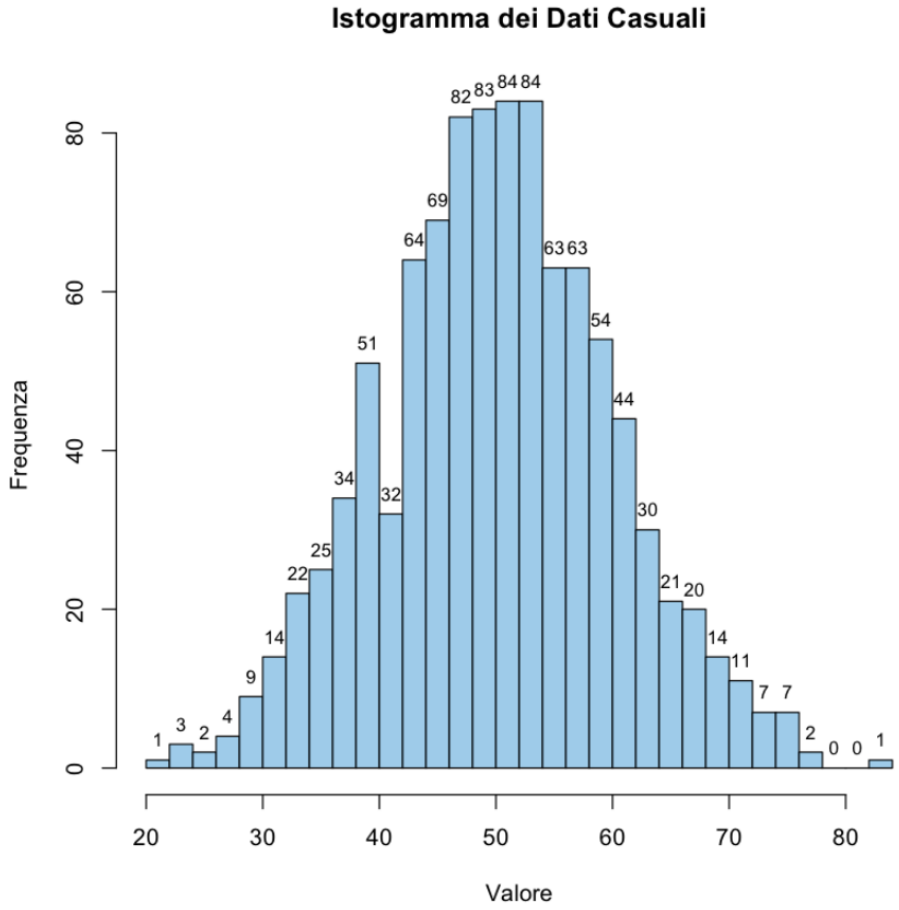
```
Console   Terminal ×   Background Jobs ×
R 4.3.1 · C:/Users/pc/R-studio-workspace/
> voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25,
24, 25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21,
28, 29, 28)
> table (voti)
voti
18 19 20 21 22 23 24 25 26 27 28 29 30
 1  2  1  3  2  2  3  4  2  3  4  2  1
> h <- hist(voti, freq=FALSE,  main="Istogramma dei voti",
ylab="Frequenza assoluta delle classi", col=rainbow(6))
```



ISTOGRAMMA

- Per realizzare un istogramma in base alle frequenze relative, occorre impostare nella funzione **hist**() il parametro **freq = FALSE**

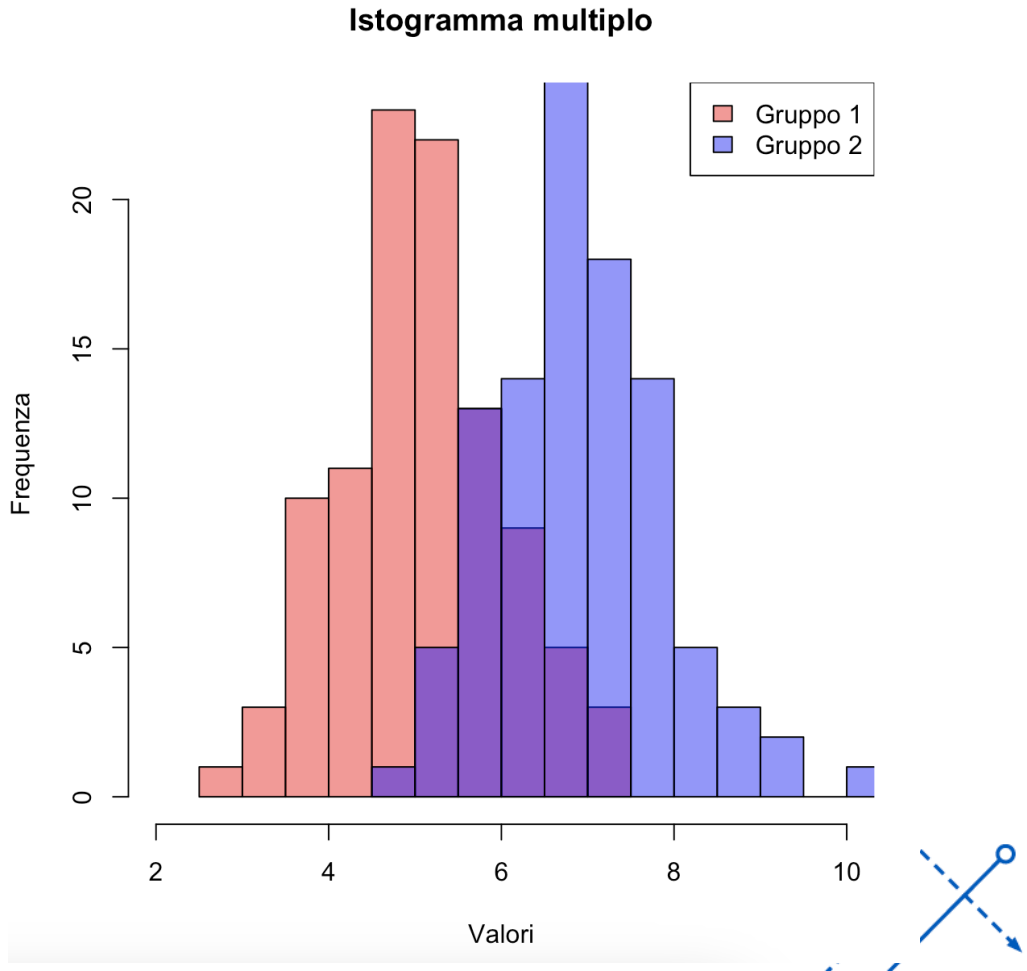
```r
# Impostiamo il seed per la riproducibilità
set.seed(123)

# Creiamo un vettore di dati casuali
dati <- rnorm(1000, mean = 50, sd = 10)  # 1000 dati con media 50 e deviazione standard 10

# Creiamo l'istogramma
hist(dati, breaks = 30, col = "skyblue", xlab = "Valore", ylab = "Frequenza",
     main = "Istogramma dei Dati Casuali", border = "black")

# Aggiungiamo i numeri sopra le barre
frequenze <- hist(dati, breaks = 30, plot = FALSE)$counts  # Otteniamo le frequenze
centro_barre <- hist(dati, breaks = 30, plot = FALSE)$mids  # Otteniamo i centri delle barre

# Aggiungiamo i numeri
text(centro_barre, frequenze, labels = frequenze, pos = 3, cex = 0.8, col = "black")
```



ISTOGRAMMA

- Gli **istogrammi multipli** permettono di confrontare **due o più distribuzioni** sullo stesso grafico.
- Per esempio, possiamo confrontare due gruppi di dati generati da distribuzioni normali con medie diverse.

```r
set.seed(123)
x1 <- rnorm(100, mean = 5, sd = 1)
x2 <- rnorm(100, mean = 7, sd = 1)

# Calcolo degli istogrammi senza plot
hist1 <- hist(x1, plot = FALSE)
hist2 <- hist(x2, plot = FALSE)

# Imposto il grafico del primo istogramma
plot(hist1, col = rgb(1, 0, 0, 0.5), xlim = c(2, 10),
     main = "Istogramma multiplo",
     xlab = "Valori", ylab = "Frequenza")

# Aggiungo il secondo istogramma
plot(hist2, col = rgb(0, 0, 1, 0.5), add = TRUE)

# Aggiungo la legenda
legend("topright", legend = c("Gruppo 1", "Gruppo 2"),
       fill = c(rgb(1, 0, 0, 0.5), rgb(0, 0, 1, 0.5)))
```

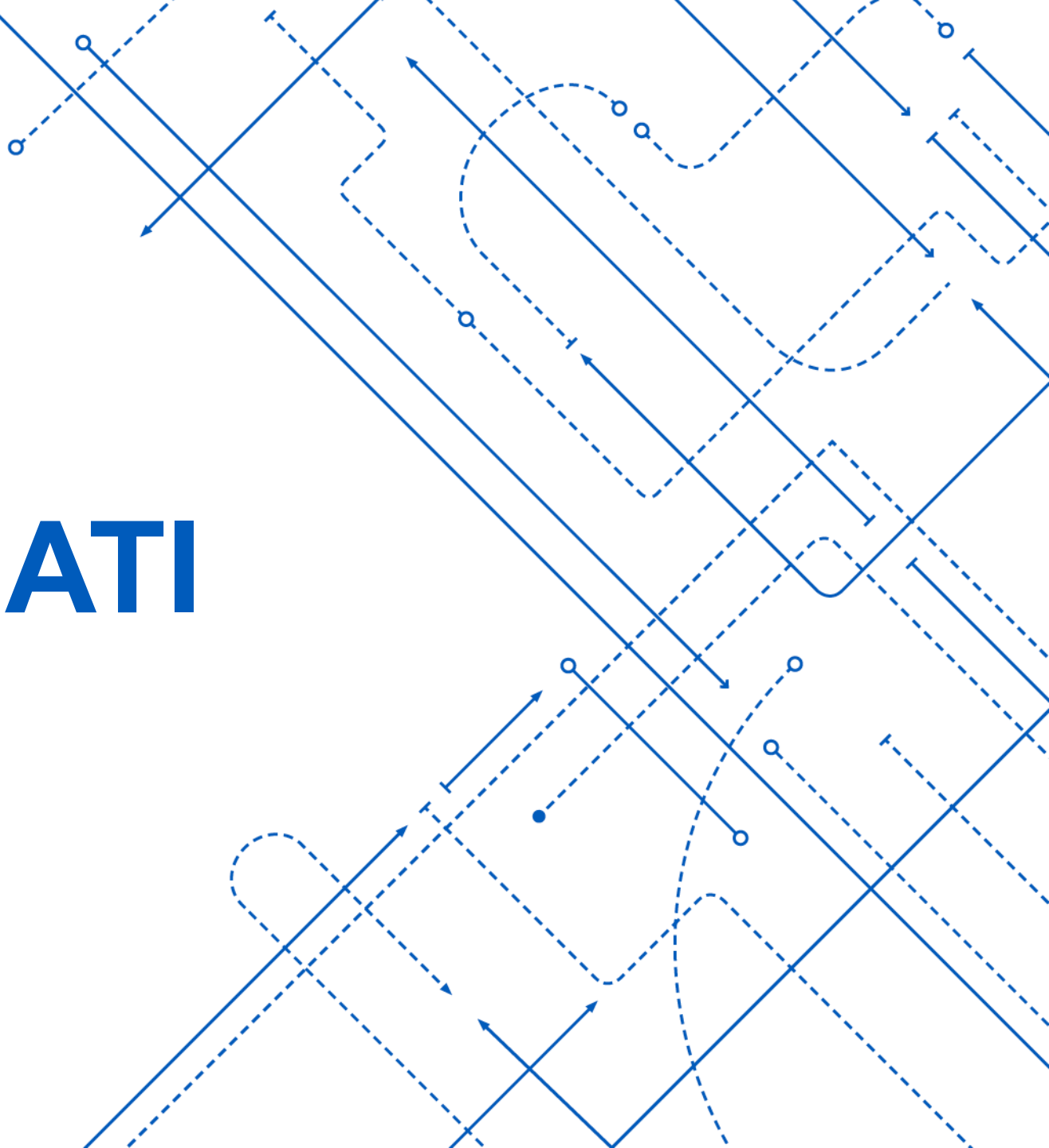


STATISTICA E ANALISI DEI DATI

Capitolo 3 – Kernel Density Plot

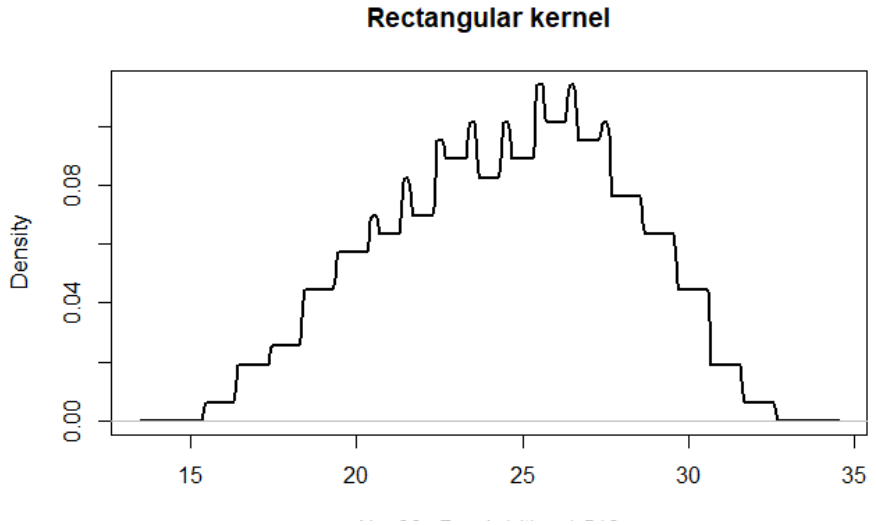
Dott. Stefano Cirillo
Dott. Luigi Di Biasi

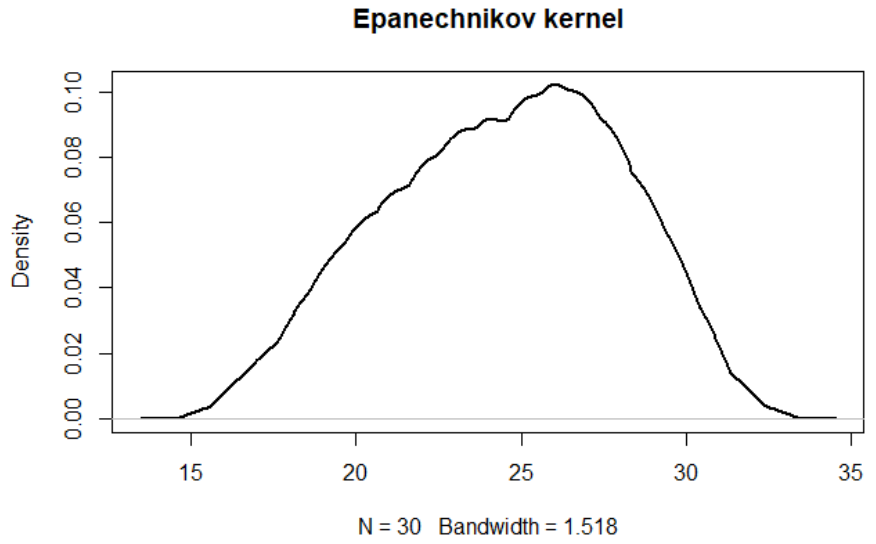
a.a. 2025-2026



KERNEL DENSITY PLOT

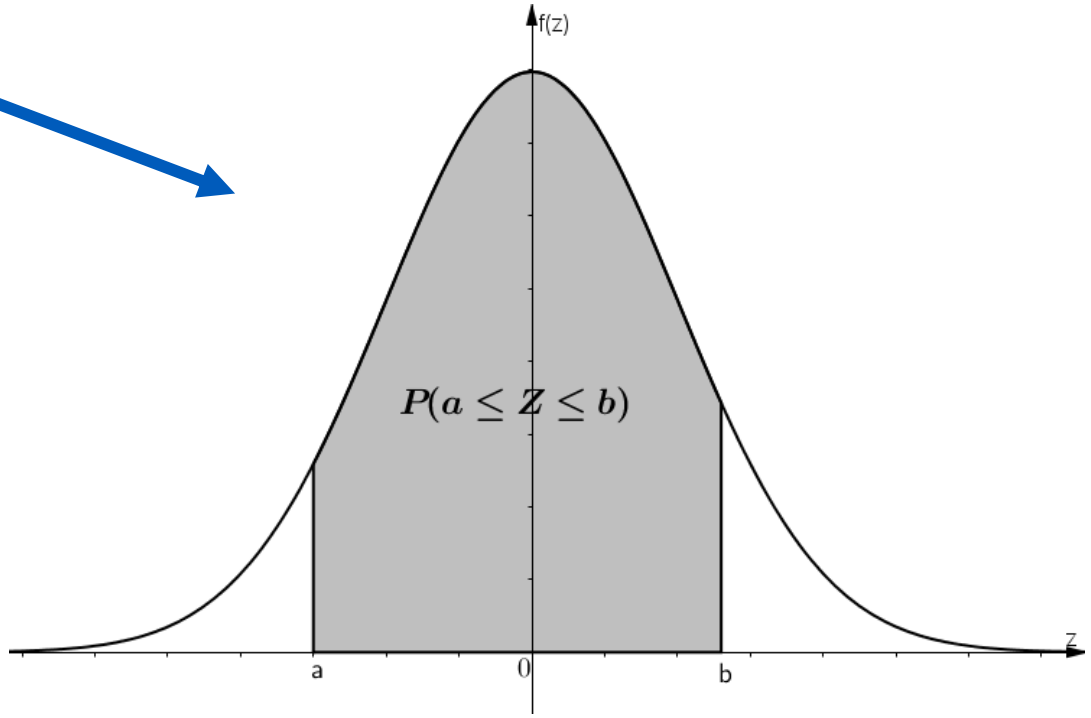
- Gli istogrammi sono un importante strumento per la rappresentazione di una distribuzione di frequenza in classi per variabili quantitative univariate
- La scelta degli intervalli delle classi è cruciale per l'aspetto finale del grafico dell'istogramma
- I **kernel density plot** possono essere utilizzati in alternativa al tradizionale istogramma
  - Sono grafici basati sulla stima kernel di densità
    - La **stima kernel di densità** è il metodo non parametrico impiegato per realizzare i kernel density plot
    - Si traccia una curva continua determinata da un fattore *K* (kernel), e da un parametro h (ampiezza della banda o bandwidth)





DENSITÀ NEI DATI

- La **densità nei dati** è un concetto statistico che descrive la distribuzione della probabilità di una variabile continua
  - La densità fornisce informazioni su quanto è probabile trovare un valore specifico in un intervallo di valori
    - La densità in una variabile continua è rappresentata l'area sotto la curva della funzione di densità in un intervallo specifico
- Stima della Densità:
  - I kernel density plot forniscono una stima continua della densità dei dati
  - A differenza degli istogrammi, che segmentano i dati in intervalli discreti, i kernel density plot utilizzano una funzione kernel per creare una rappresentazione più fluida della distribuzione dei dati



DENSITÀ NEI DATI

- Sia $(x_1, x_2, ..., x_n)$ un campione costituito da n osservazioni di una variabile quantitativa la cui densità di frequenza $f(x)$ non è nota in ogni punto x
- Vogliamo stimare la forma (shape) di $f(x)$ in base al campione di osservazioni
- Un grafico della densità basata sul kernel può essere realizzato utilizzando la seguente funzione stimata in base al campione

$$\widehat{f}_h(x) = \frac{1}{n\,h}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad x \in \mathbb{R},$$

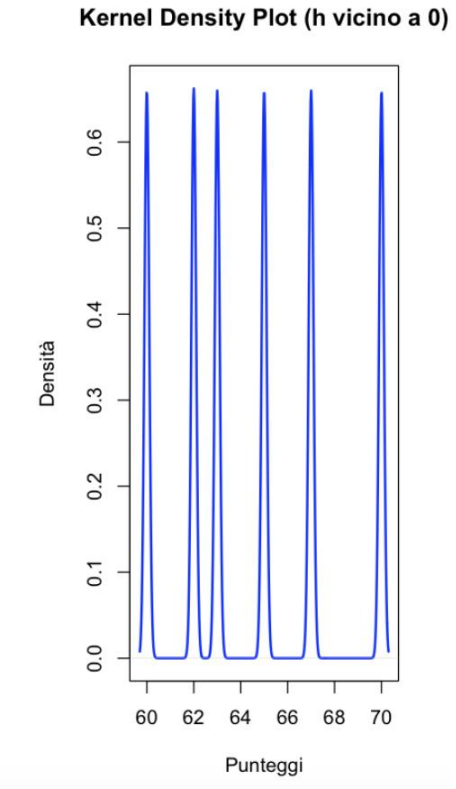
Dove:

- n è l'ampiezza del campione,
- $K(x)$ è il kernel, ossia una funzione densità di probabilità (non negativa) con media nulla
- $h > 0$ è un parametro di smoothing, detto bandwidth

DENSITÀ NEI DATI

- La scelta del:
  - Kernel *K(x)* può influenzare l'aspetto generale del grafico;
  - Parametro *h* è importante in quanto
    - **Un valore troppo vicino a zero** rende la stima irregolare e la funzione kernel applicata a ciascun punto dati ha un "picco" molto stretto.
      - **Stima Irregolare**: Ogni punto dati influisce sulla densità stimata in modo molto localizzato. La stima della densità può risultare "saltellante" o irregolare, con picchi e valli marcati. Ciò significa che anche piccole variazioni nei dati possono provocare grandi cambiamenti nella stima della densità
      - **Varianza Elevata**: Poiché la stima della densità è estremamente sensibile ai dati individuali, la varianza della stima aumenta. Questo comporta che la densità stimata non è affidabile e può non riflettere la vera distribuzione sottostante

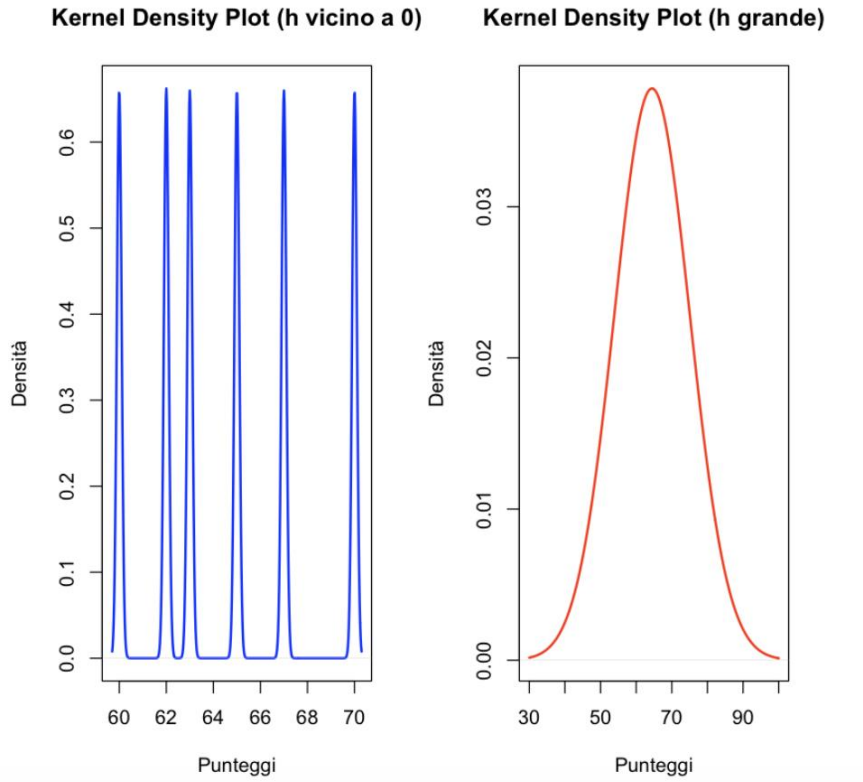
$$\widehat{f}_h(x) = \frac{1}{n\,h}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad x \in \mathbb{R},$$



DENSITÀ NEI DATI

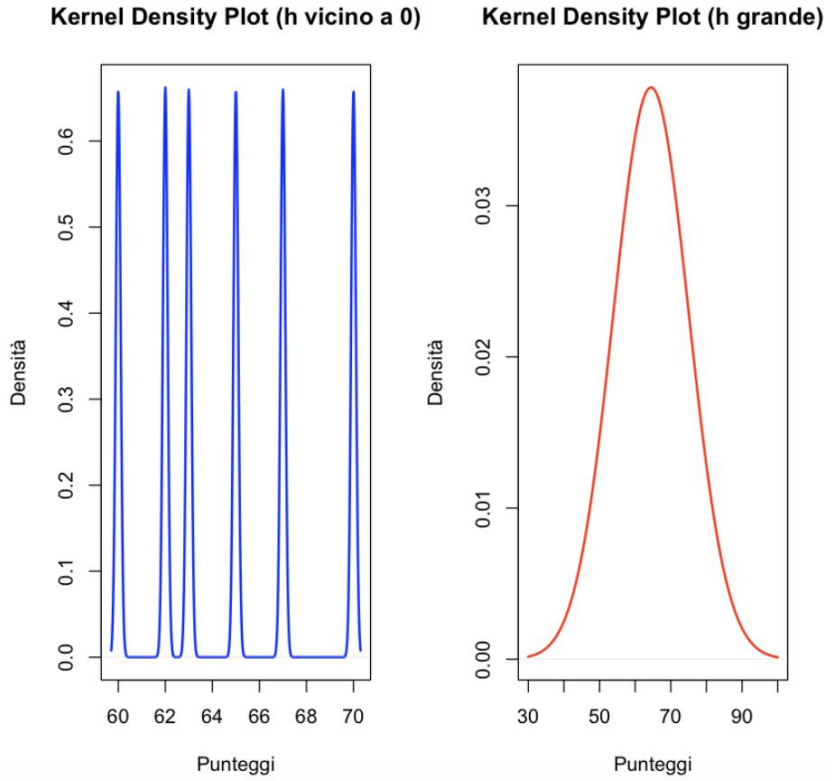
$$\widehat{f}_h(x) = \frac{1}{n\,h}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad x \in \mathbb{R},$$

- Parametro *h* è importante in quanto
  - **Un valore troppo grande** fa si che la funzione kernel applicata a ciascun punto dati è "spalmata" su un'area molto più ampia
    - **Distorsione della Stima**: Una larghezza di banda eccessiva provoca una smussatura eccessiva della stima della densità, con la perdita di dettagli importanti. I picchi nei dati possono essere "appiattiti", e le caratteristiche reali della distribuzione possono essere nascoste. La stima diventa così **generica** che potrebbe non rappresentare correttamente la variabilità nei dati.
    - **Perdita di Informazioni**: Le aree dove ci sono effettive concentrazioni di dati possono apparire come una densità uniforme, portando a conclusioni errate o fuorvianti.



DENSITÀ NEI DATI

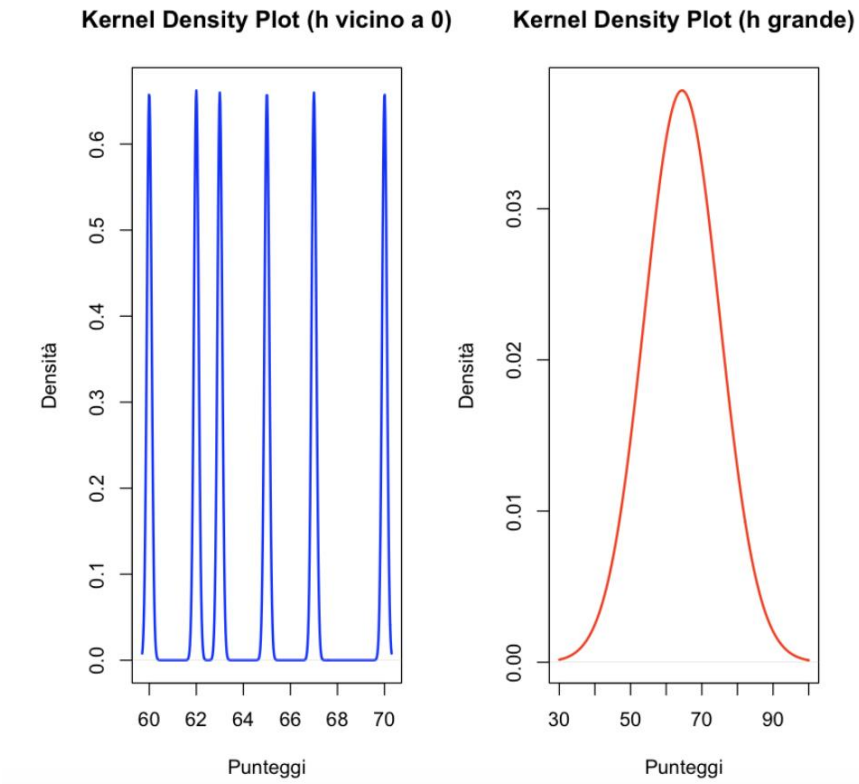
- Esempio: Supponiamo di avere un dataset con i seguenti punteggi: [60, 62, 63, 65, 67, 70].
  - ***h* troppo vicino a zero:** Se usiamo una larghezza di banda molto piccola (es. h=0.1), il kernel density plot apparirà come un grafico con picchi molto alti e stretti intorno a ciascun punteggio, dando un'impressione di irregolarità.
    - I picchi possono risultare in aree con densità apparentemente elevate anche dove ci sono pochi dati, creando l'illusione di una distribuzione complessa
  - ***h* troppo grande:** Usando lo stesso dataset [60,62,63,65,67,70], se impostiamo una larghezza di banda molto grande (es. h=10), il kernel density plot apparirà piatto e uniforme.
    - Non saremo in grado di distinguere tra le aree di alta e bassa densità, e potremmo erroneamente concludere che i dati sono distribuiti in modo omogeneo, quando in realtà ci sono picchi di densità in corrispondenza di determinati punteggi.



Densità Nei Dati

- Esempio: Supponiamo di avere un dataset con i seguenti punteggi: [60, 62, 63, 65, 67, 70].

```r
# Dataset di punteggi
punct_scores <- c(60, 62, 63, 65, 67, 70)

# Creazione della finestra grafica
par(mfrow=c(1, 2))  # Due grafici affiancati

# Esempio 1: Larghezza di banda vicina a zero
h_small <- 0.1  # Larghezza di banda piccola
density_small <- density(punct_scores, bw=h_small)  # Stima della densità
plot(density_small, main="Kernel Density Plot (h vicino a 0)", xlab="Punteggi",
     ylab="Densità", col="blue", lwd=2)

# Esempio 2: Larghezza di banda grande
h_large <- 10  # Larghezza di banda grande
density_large <- density(punct_scores, bw=h_large)  # Stima della densità
plot(density_large, main="Kernel Density Plot (h grande)", xlab="Punteggi",
     ylab="Densità", col="red", lwd=2)

# Ripristino della finestra grafica
par(mfrow=c(1, 1))  # Una finestra grafica
```

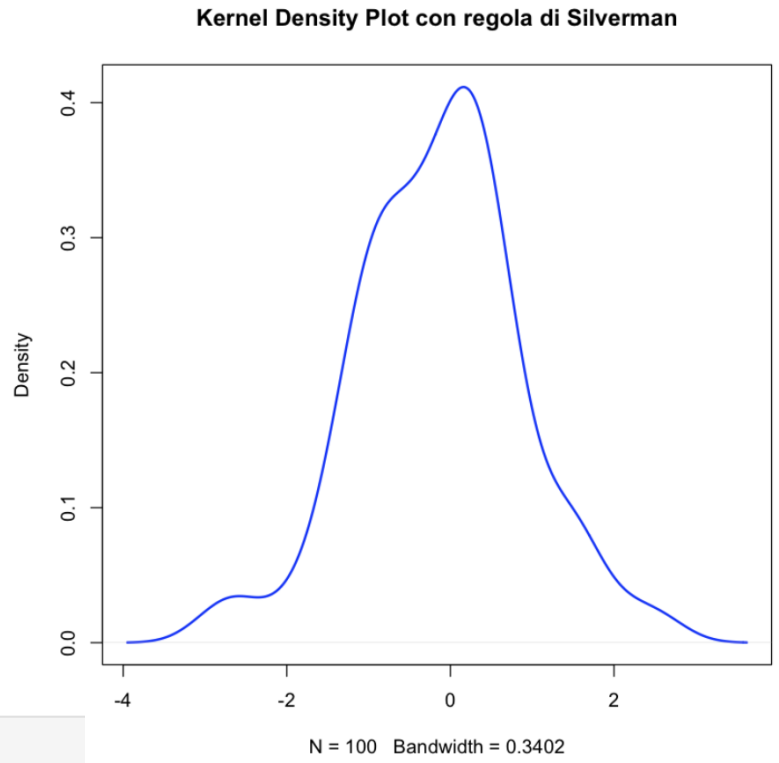


CALCOLARE L'*H* OTTIMALE

- Stimare h permette di bilanciare la regolarità e la precisione della stima della densità
- Come stimare un valore ottimale per *h*?
  1. Larghezza di Banda Ottimale di Silverman: tiene conto della deviazione standard delle osservazioni e può essere calcolata come:

$$h = 0.9\, min\left(\hat{\sigma}, \frac{IQ}{1.34}\right) n^{-\frac{1}{5}}$$

Dove $\hat{\sigma}$ è la deviazione standard dei dati, *IQ* è il range interquartile (differenza tra il 75° e il 25° percentile) e *n* è il numero di osservazioni




```
data <- rnorm(100)  # Genera 100 osservazioni casuali da una distribuzione normale
density_plot <- density(data)  # Usa la larghezza di banda di Silverman di default
plot(density_plot, main="Kernel Density Plot con regola di Silverman", col="blue", lwd=2)
```

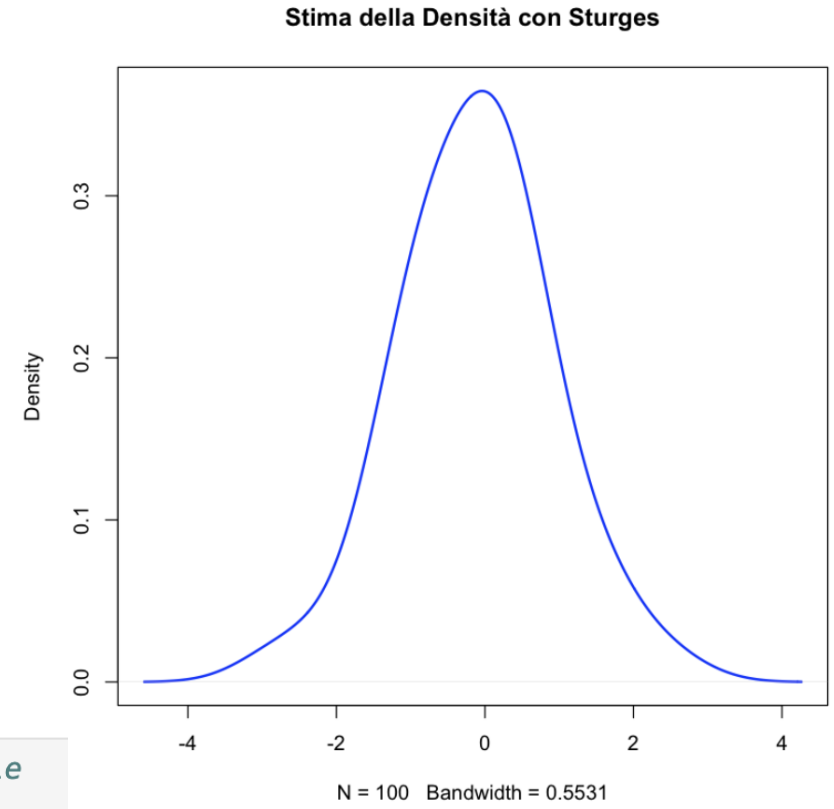
- In R, la regola di Silverman è implementata come valore di default nella funzione density()

CALCOLARE L'*H* OTTIMALE

- Stimare h permette di bilanciare la regolarità e la precisione della stima della densità
- Come stimare un valore ottimale per *h*?
  1. Metodi Basati sulla Regola di Sturges: La regola è un approccio semplice che suggerisce di calcolare la larghezza di banda in base al numero di osservazioni n:

$$h = \frac{\Delta}{\sqrt{n}}$$

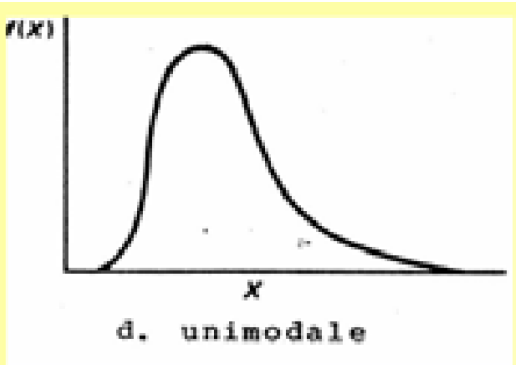
Dove Δ è la differenza tra il valore massimo ed il valore minimo delle osservazioni considerate

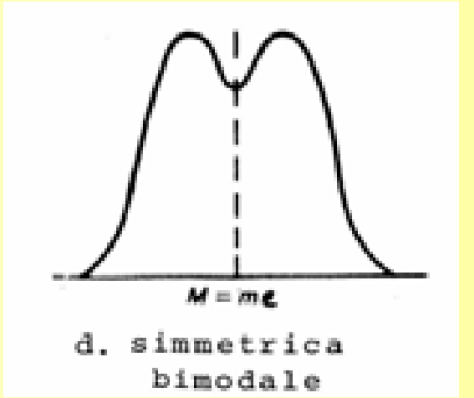


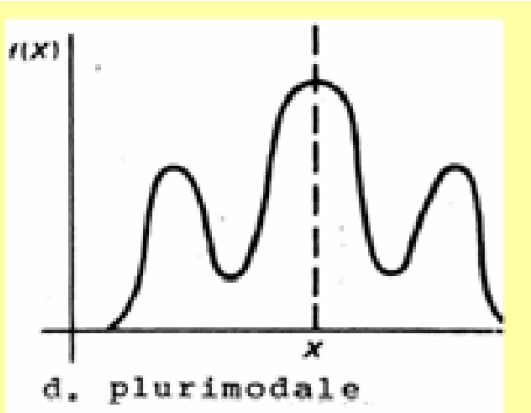

```r
data <- rnorm(100)  # Genera 100 osservazioni casuali da una distribuzione normale
# Calcolo larghezza di banda usando la Regola di Sturges
n <- length(data)
h_sturges <- (max(data) - min(data)) / sqrt(n)

# Stima della densità con h di Sturges
density_sturges <- density(data, bw=h_sturges)
plot(density_sturges, main="Stima della Densità con Sturges", col="blue", lwd=2)
```

CALCOLARE L'*H* OTTIMALE

Caratteristica	Regola di Sturges	Regola di Silverman
Obiettivo	Analisi esplorativa iniziale	Analisi dettagliata, dati complessi
Robustezza	Non ideale per dati complessi o multimodali	Più robusta per distribuzioni variabili
Outliers	Meno sensibile	Più sensibile
Precisione	Rapida, ma meno precisa	Più accurata
Dimensione del Campione	Piccoli/medi dataset	Grandi dataset
Tipo di Dati	Distribuzioni semplici, unimodali	Distribuzioni complesse, multimodali







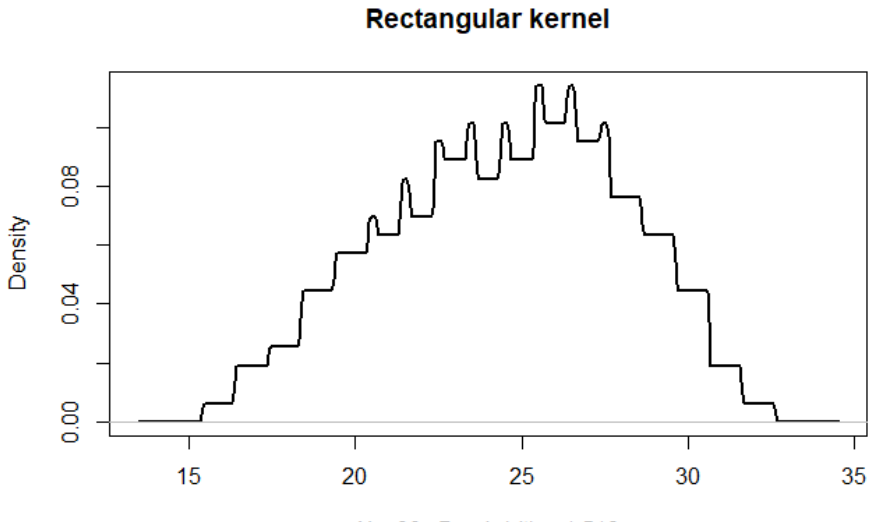
Source

ALCUNI TIPI DI KERNEL

rectangular (uniform):

$$K(x) = \begin{cases} 1/2, & -1 \le x \le 1, \\ 0, & \text{altrimenti}, \end{cases} \qquad \sigma^2 = \frac{1}{3}$$

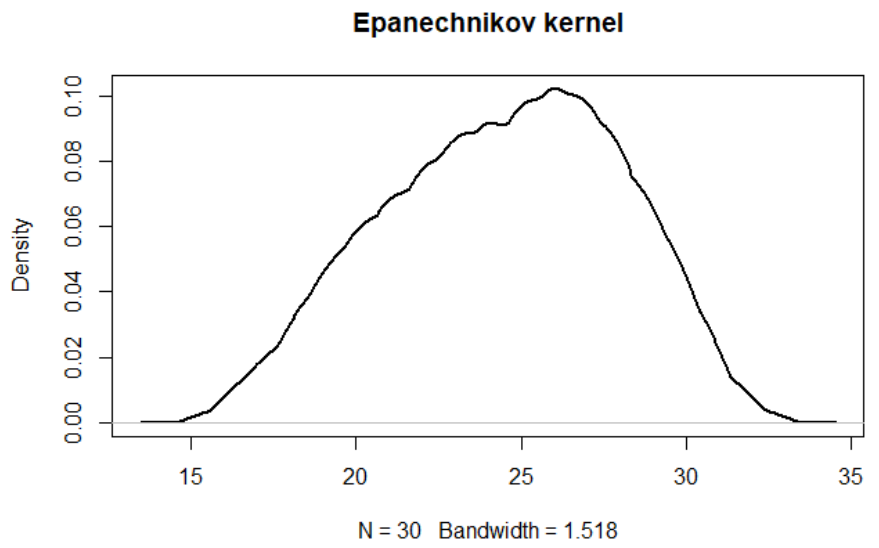
```
Console  Terminal  Background Jobs
R 4.3.1 · C:/Users/pc/R-studio-workspace/
> voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24,
25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29,
28)
> d <- density(voti, kernel = "rectangular")
> plot(d, lwd = 2, main = "Rectangular kernel")
```



epanechnikov (parabolic):

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2), & -1 \le x \le 1, \\ 0, & \text{altrimenti.} \end{cases} \qquad \sigma^2 = \frac{1}{5}$$

```
Console  Terminal  Background Jobs
R 4.3.1 · C:/Users/pc/R-studio-workspace/
> voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24,
25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29,
28)
> d <- density(voti, kernel = "epanechnikov")
> plot(d, lwd = 2, main = "Epanechnikov kernel")
```

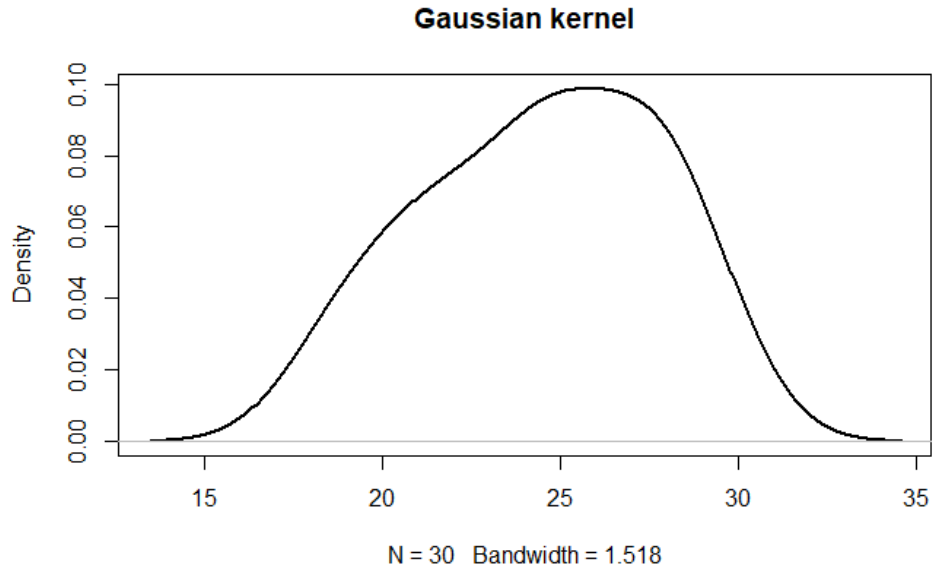


ALCUNI TIPI DI KERNEL

gaussian:

$$K(x) = \frac{1}{\sqrt{2\,\pi}} \exp\{-\frac{x^2}{2}\}, \quad x \in \mathbb{R} \qquad \sigma^2 = 1$$

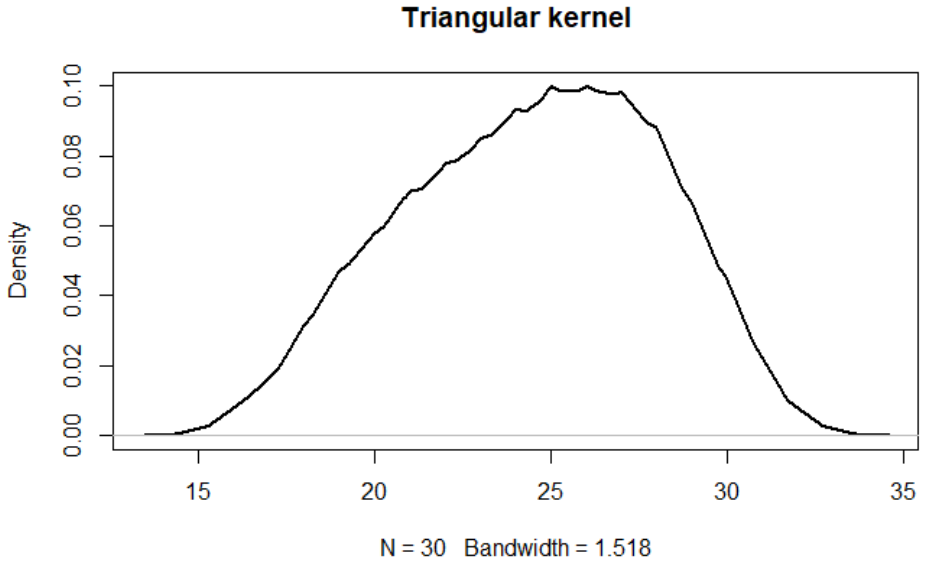
```
Console  Terminal  Background Jobs
R 4.3.1 · C:/Users/pc/R-studio-workspace/
> voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24,
25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29,
28)
> d <- density(voti, kernel = "gaussian")
> plot(d, lwd = 2, main = "Gaussian kernel")
>
```



triangular:

$$K(x) = \begin{cases} 1 - |x|, & -1 \le x \le 1, \\ 0, & \text{altrimenti.} \end{cases} \qquad \sigma^2 = 1/6$$

```
Console  Terminal  Background Jobs
R 4.3.1 · C:/Users/pc/R-studio-workspace/
> voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24,
25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29,
28)
> d <- density(voti, kernel = "triangular")
> plot(d, lwd = 2, main = "Triangular kernel")
```

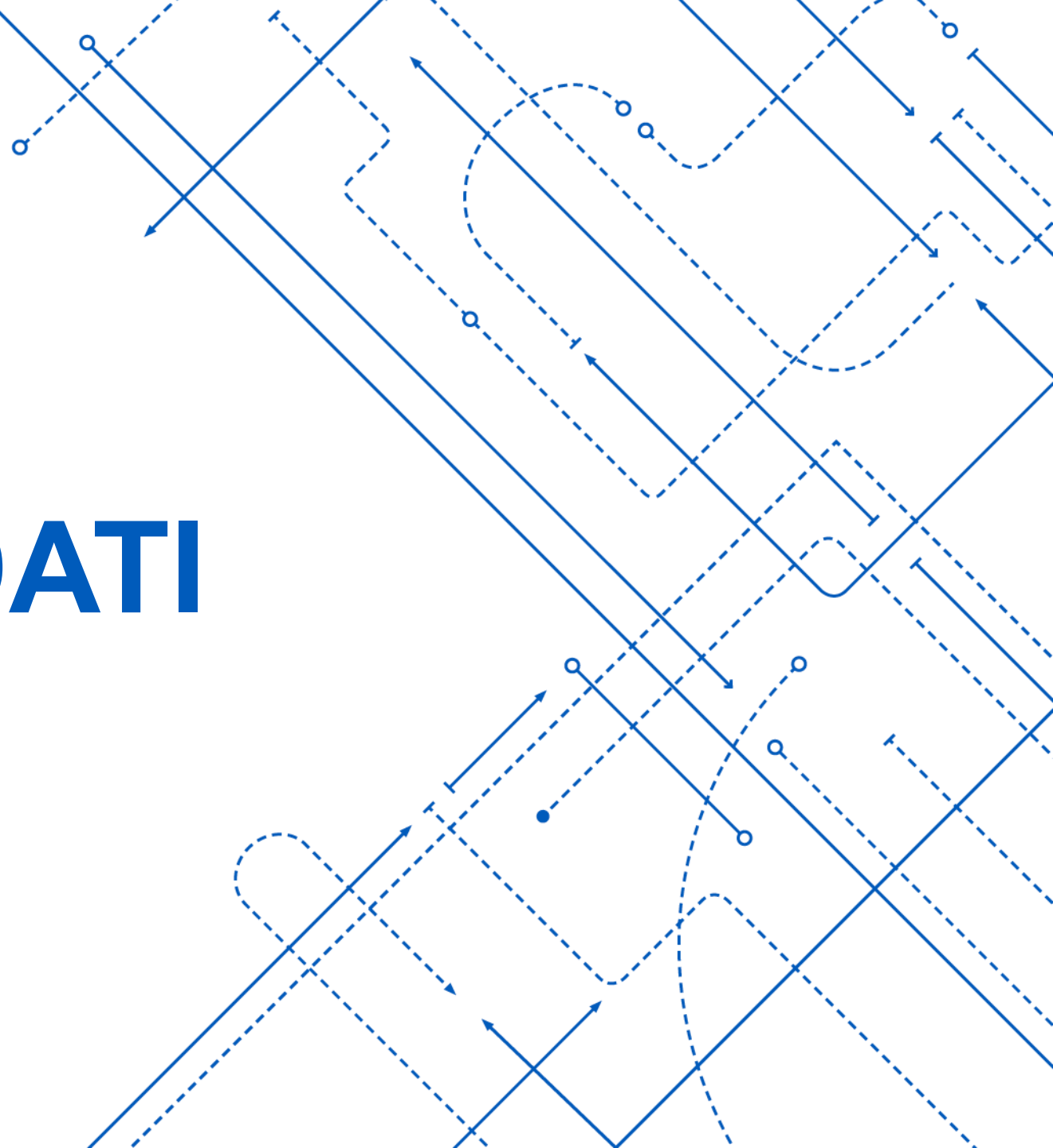


STATISTICA E ANALISI DEI DATI

Capitolo 3 – Quartili e BoxPlot

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

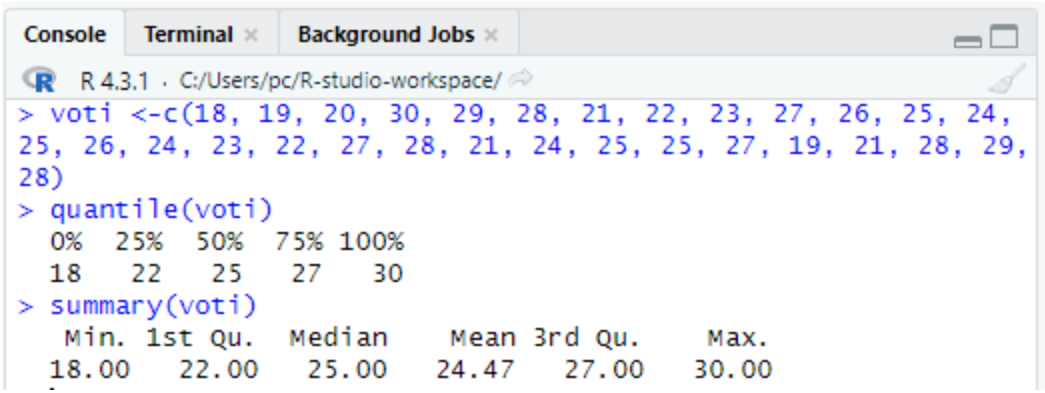


QUARTILI

- Consideriamo un campione $(x_1, x_2, ..., x_n)$ dei valori assunti da una variabile quantitativa *X*
  - Sono valori che dividono un insieme di dati ordinato in quattro parti uguali, in modo che ciascuna parte contenga il 25% dei dati
- Procediamo ad ordinare i valori del campione in ordine crescente
  - Primo quartile: il valore per il quale il 25% dei dati sono alla sua sinistra Q_1
  - Secondo quartile: il valore per il quale il 50% dei dati sono alla sua sinistra Q_2 (detto anche **mediana**)
  - Terzo quartile: il valore per il quale il 75% dei dati sono alla sua sinistra Q_3
- Q_0 e Q_4 forniscono in **minimo** ed il **massimo** dei valori del campione
- Per calcolare i quartili in un insieme di dati ordinato:
  - Q1: Si trova prendendo il punto a $\frac{n+1}{4} - esima$ posizione
  - Q2: (mediana): Si trova a $\frac{n+1}{2} - esima$ posizione
  - Q3: Si trova a $\frac{3(n+1)}{4} - esima$ posizione

QUARTILI

- Consideriamo un campione $(x_1, x_2, ..., x_n)$ dei valori assunti da una variabile quantitativa *X*
- Procediamo ad ordinare i valori del campione in ordine crescente
  - Primo quartile: il valore per il quale il 25% dei dati sono alla sua sinistra Q_1
  - Secondo quartile: il valore per il quale il 50% dei dati sono alla sua sinistra Q_2 (detto anche **mediana**)
  - Terzo quartile: il valore per il quale il 75% dei dati sono alla sua sinistra Q_3
- Q_0 e Q_4 forniscono in **minimo** ed il **massimo** dei valori del campione
- La funzione:
  - **quantile**(vettore) calcola i quartili di un campione
  - **summary**(vettore) calcola minimo, massimo, media e mediana di un campione



$Q_0 = 18$, $Q_1 = 22$, $Q_2 = 25$,
$Q_3 = 27$ e $Q_4 = 30$

QUARTILI

- I **quartili** sono strumenti fondamentali in **data science** e **analisi dei dati** perché aiutano a comprendere la distribuzione e la dispersione dei dati, specialmente quando si lavora con dataset grandi o complessi
- **Identificazione della distribuzione**:
    - I quartili permettono di suddividere i dati in parti uguali, fornendo una comprensione di come i dati sono distribuiti. Ciò aiuta a identificare se i dati sono concentrati in una parte specifica o se sono distribuiti uniformemente.
- **Rilevazione di outlier (valori anomali)**:
    - L'uso dei quartili consente di calcolare l'**intervallo interquartile (IQR)**, che è la differenza tra il terzo quartile (Q3) e il primo quartile (Q1).
        - I valori che si trovano **al di fuori dell'IQR** possono essere considerati **outlier**, poiché si distanziano dalla distribuzione centrale dei dati.
- **Misura della dispersione**:
    - L'IQR è una misura robusta della dispersione che non è influenzata dagli outlier come lo sono altre misure di variabilità, come la deviazione standard. Fornisce una visione più affidabile della variabilità dei dati

BoxPlot

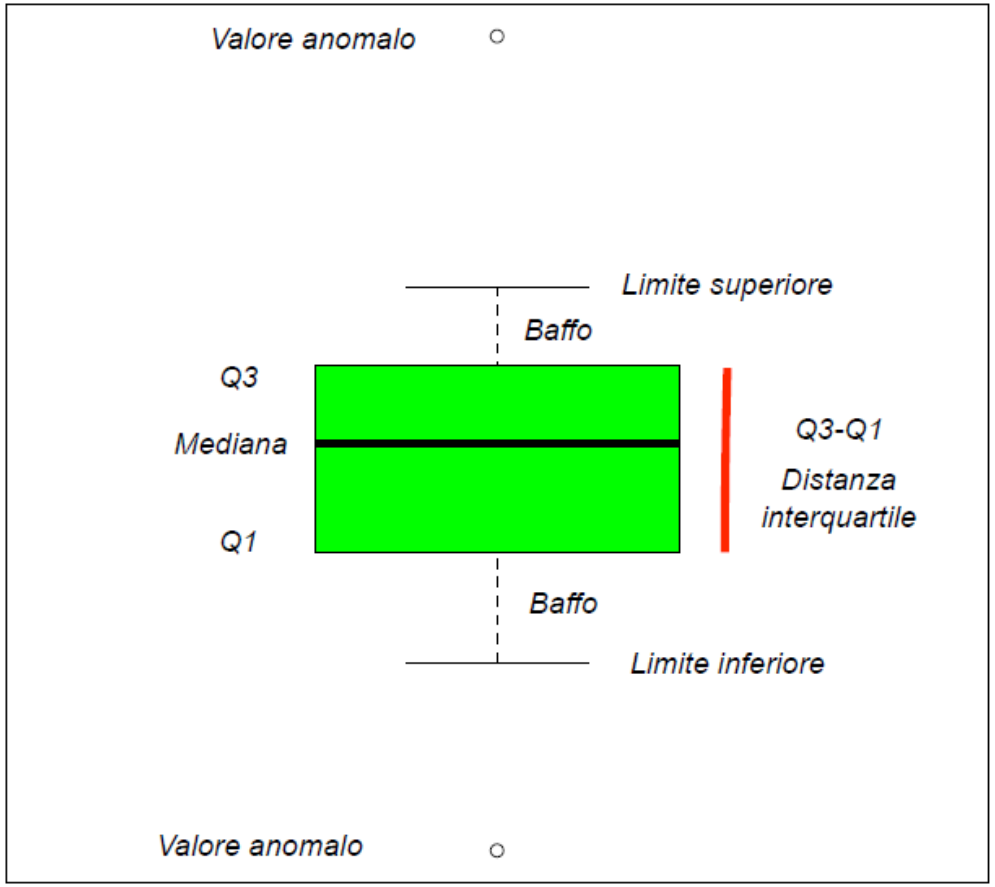
- Il **boxplot**, detto anche scatola con **baffi**, è il grafico di una 'scatola' i cui estremi sono Q_1 e Q_3, tagliata da una linea orizzontale in corrispondenza di Q_2
  - L'estremo del baffo inferiore rappresenta il valore più piccolo tra le osservazioni che risulta maggiore o uguale di

$$(Q_1 - 1.5 * (Q_3 - Q_1))$$

  - L'estremo del baffo superiore rappresenta il valore più grande tra le osservazioni che risulta minore o uguale di

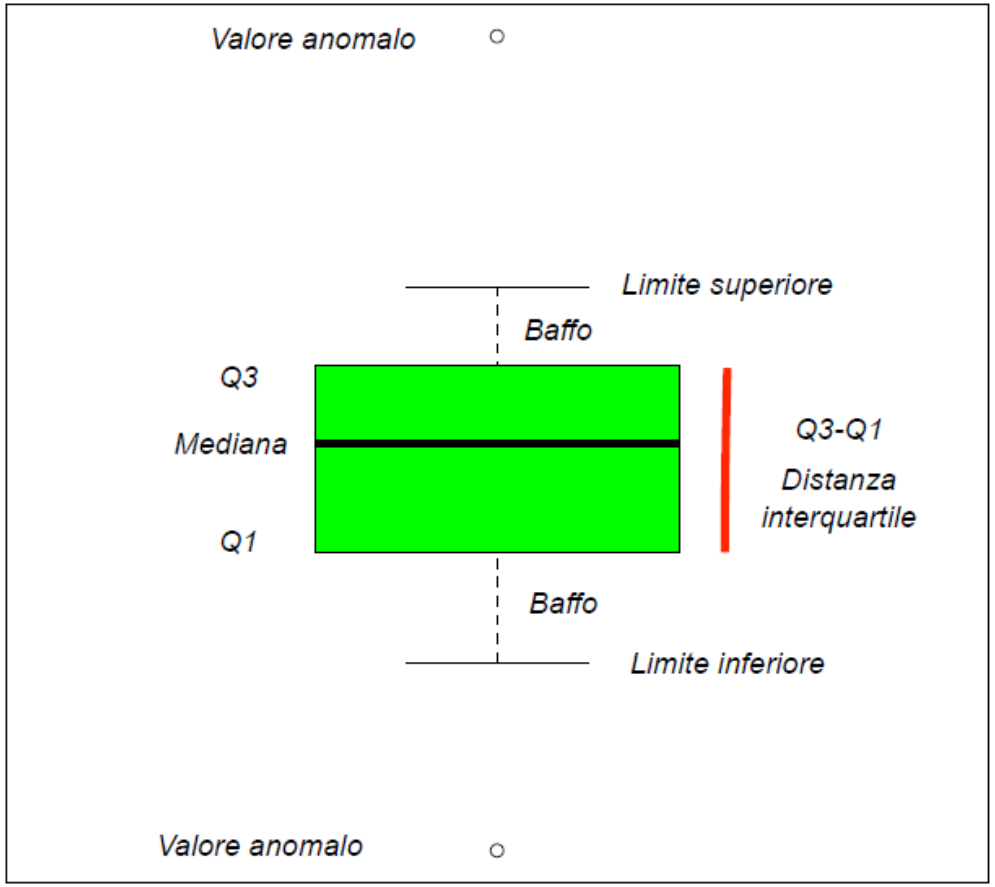
$$(Q_1 + 1.5 * (Q_3 - Q_1))$$

- Possono esserci valori al di fuori dell'intervallo superiore ed inferiore
  - Tali valori sono detti valori **anomali** o **outlier** e sono rappresentati con dei punti nel grafico



BoxPlot

- Il boxplot viene utilizzato per illustrare alcune caratteristiche di una distribuzione di frequenza:
  - Centralità: è espressa dalla mediana
  - Forma: può essere simmetrica o asimmetrica e si deduce esaminando le distanze tra Q_1 e Q_3
  - Dispersione: è deducibile esaminando le distanze dell'estremo del baffo superiore da Q_3 e dell'estremo del baffo inferiore da Q_1
  - Presenza di valori anomali
- La funzione **boxplot**(vettore) permette di disegnare un boxplot in R
  - Il parametro *horizontal* permette di specificare se si vuole un grafico in senso orizzontale o verticale



ESEMPIO

- Un grafico può contenere diversi elementi che permettono di confrontare più variabili
- Ogni variabile descrive insiemi di dati numerici di uno **stesso fenomeno quantitativo**
- Esempio:
  - Analizzare i voti riportati di certo esame da studenti (Uomini e Donne) appartenenti a due differenti classi:
    - **Classe A**: costituita da 30 studenti
    - **Classe B**:costituita da 24 studenti
  - I cui voti sono:

Classe A	18	19	20	30	29	28	21	22	23	27	26	25
Sesso	U	U	U	U	U	U	U	U	U	U	U	U
Classe A	24	25	26	24	23	22	27	28	21	24	25	25
Sesso	U	U	D	D	D	D	D	D	D	D	D	D
Classe A	27	19	21	28	29	28						
Sesso	D	D	D	D	D	D						

Classe B	19	19	20	27	19	28	21	22	23	26	27	25
Sesso	U	U	U	U	U	U	U	U	U	U	U	U
Classe B	24	25	26	24	23	22	28	29	21	24	21	19
Sesso	U	D	D	D	D	D	D	D	D	D	D	D

GRAFICO DEI VALORI

- La funzione **points**() permette di aggiungere punti al grafico rappresentato
  - Il parametro **pch** individua il tipo di carattere da utilizzare
  - Il parametro **bg** indica quale colore di sfondo utilizzare
  - Il parametro **cex** indica la grandezza del testo e dei simboli

```
votiClasseA <- c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24,
                 25, 26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19,
                 21, 28, 29, 28)
votiClasseB <- c(19, 19, 20, 27, 19, 28, 21, 22, 23, 26, 27, 25, 24,
                 25, 26, 24, 23, 22, 28, 29, 21, 24, 21, 19)
```

```
plot(votiClasseA, pch ="+",ylim=c(17 ,31), ylab="Voti delle due classi",
     col ="blue")
points(votiClasseB, pch ="x",col ="red")
legend(25 ,31 , c("Classe A","Classe B"),pch =c("+","x"),
       col=c("blue","red"),bg="gray",cex =0.6)
```

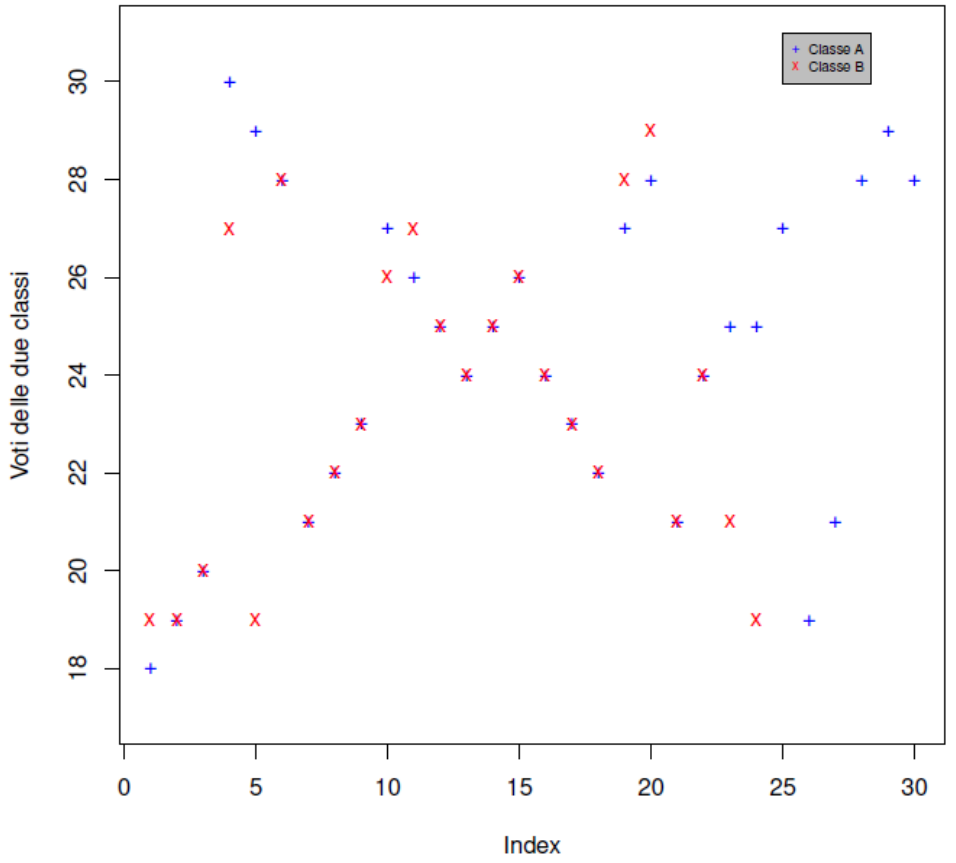


GRAFICO DEI VALORI

- Realizziamo ora un grafico che permetta di confrontare i boxplot delle distribuzioni dei voti delle due classi di studenti

```
boxplot(votiClasseA, votiClasseB, names=c("Classe A"," Classe B"),
        col=c("blue", "red"))
```

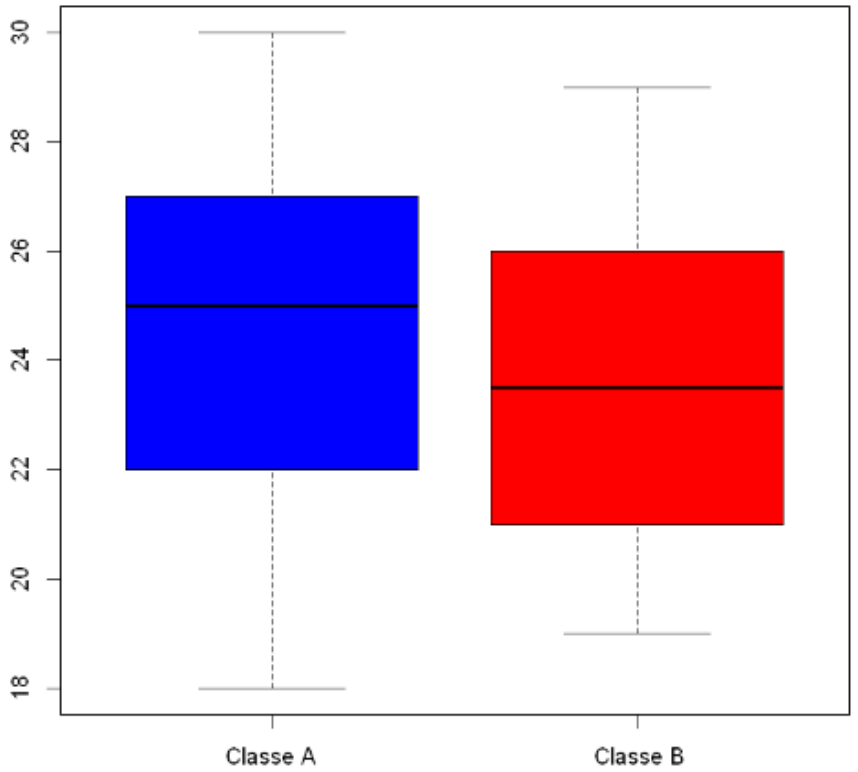


GRAFICO DEI VALORI

- Realizziamo ora un grafico che permetta di confrontare i boxplot delle distribuzioni dei voti delle due classi di studenti
- La funzione **summary**() permette di visualizzare le principali misure statistiche

```
boxplot(votiClasseA, votiClasseB, names=c("Classe A"," Classe B"),
        col=c("blue", "red"))
summary(votiClasseA)
summary(votiClasseB)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   22.00   25.00   24.47   27.00   30.00

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   21.00   23.50   23.42   26.00   29.00
```

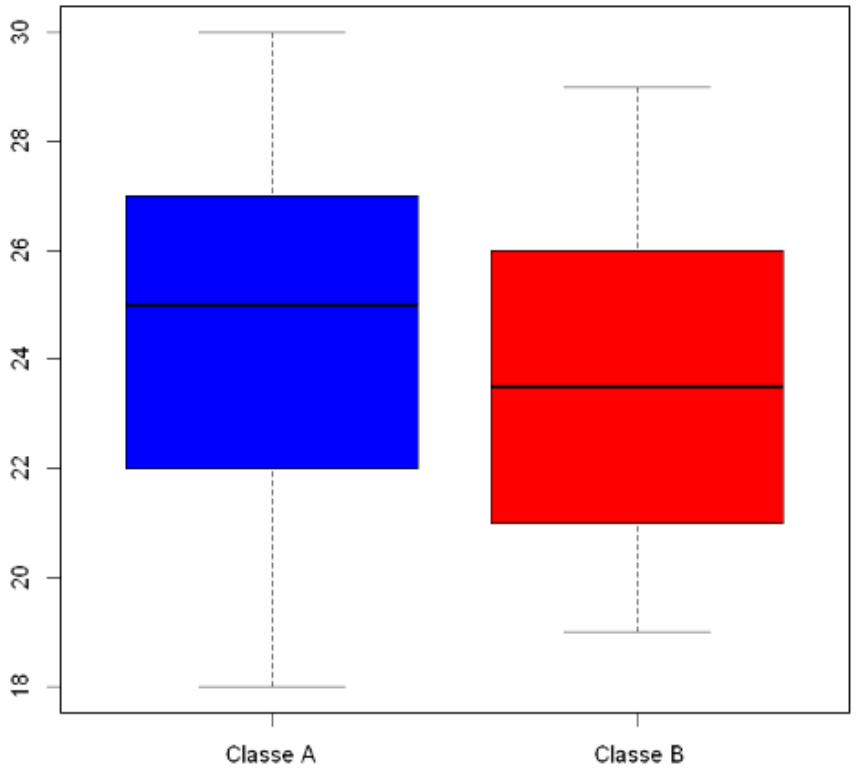
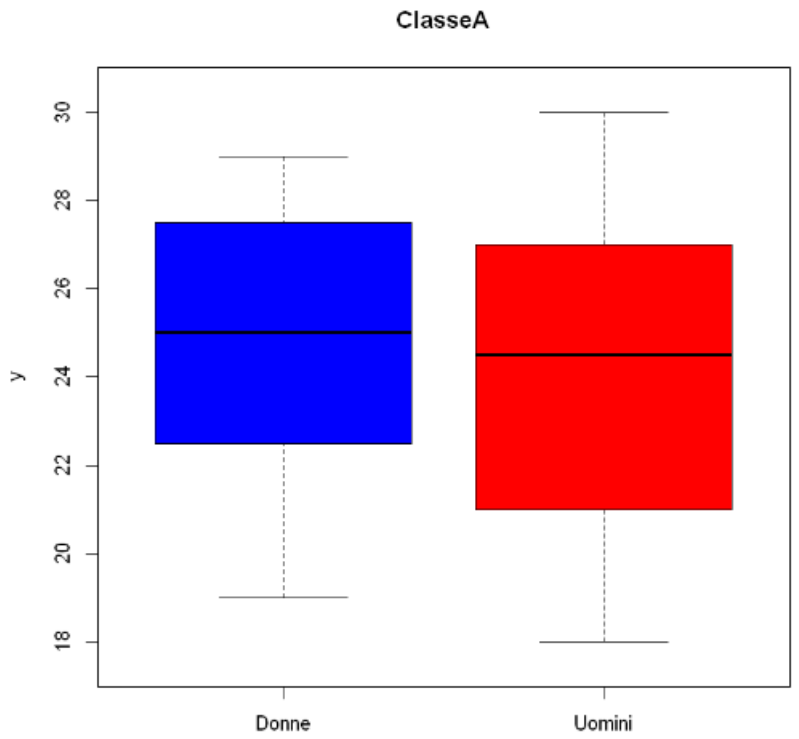


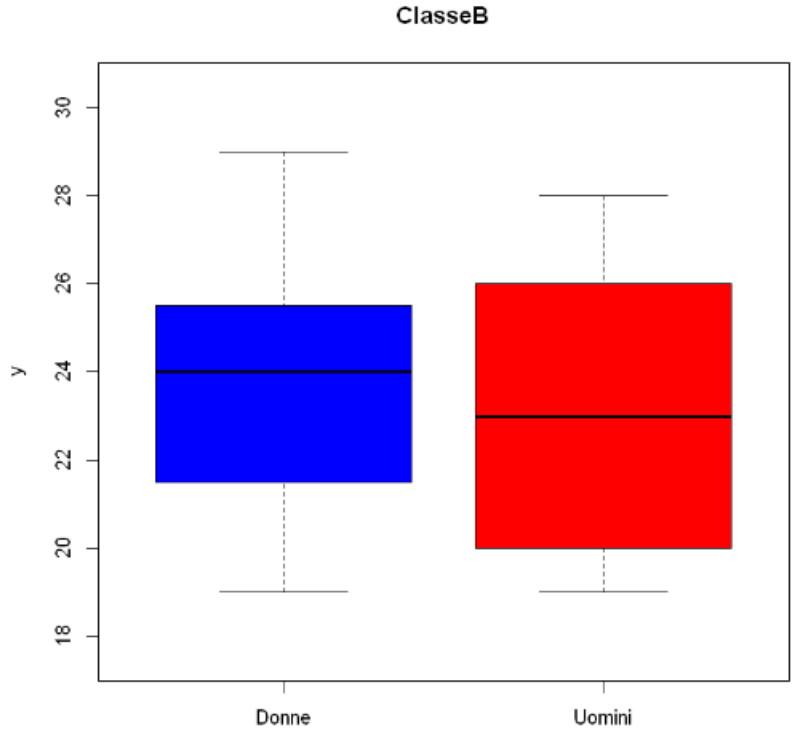
GRAFICO DI CONFRONTO DEI VALORI

- Possiamo estrarre le statistiche dei voti degli studenti uomini e donne della Classe A e della Classe B

```
sessoA <- c(rep("Uomini",14),rep("Donne",16))
sessoClasseA <- factor(sessoA)
plot(sessoClasseA,votiClasseA,main="ClasseA",
ylim=c(17.5,30.5),col=c("blue", "red"))
```




```
sessoB <- c(rep("Uomini",13),rep("Donne",11))
sessoClasseB <- factor(sessoB)
plot(sessoClasseB,votiClasseB,main="ClasseB",
ylim=c(17.5,30.5),col=c("blue", "red"))
```

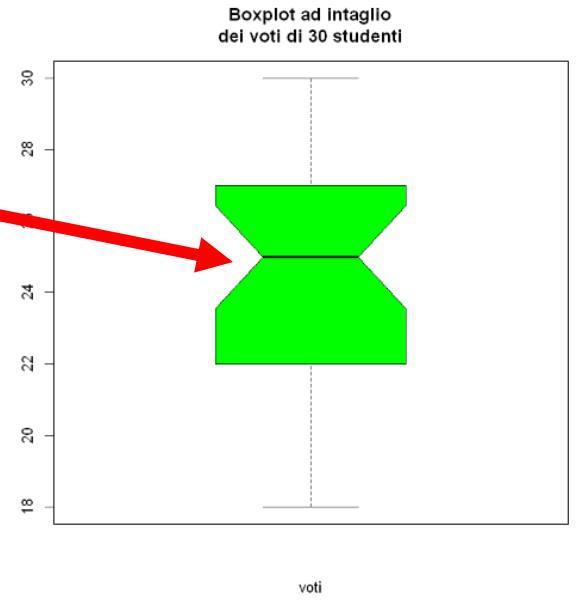


BoxPlot ad Intaglio

- I **boxplot ad intaglio** (notched boxplots) sono una alternativa grafica dei boxplot
  - Sono utilizzati per visualizzare la distribuzione dei dati, ma forniscono un'informazione aggiuntiva sulla **significatività statistica delle differenze tra mediane di gruppi**
- Le tacche si restringono attorno alla mediana
- Se le tacche di due boxplot non si sovrappongono, c'è una forte evidenza che le mediane dei gruppi sono significativamente differenti
- Gli altri elementi del boxplot rimangono invariati (quartili, estremi dei baffi, outlier)
- La tacca è un intervallo di confidenza basato sulla formula:

$$Mediana \pm 1.57 \times \frac{IQR}{\sqrt{n}}$$

  - *IQR* = $Q_1 - Q_3$ è lo scarto interquartile
  - *n* il numero di osservazioni nel campione.



BoxPlot ad Intaglio

- A differenza dei boxplot tradizionali forniscono un'informazione aggiuntiva sulla **significatività statistica delle differenze tra mediane di gruppi**
  - Permettono di visualizzare anche l'intervallo di confidenza (o fiducia) del 95% per la mediana
    - In pratica, mostra quanto siamo sicuri che il valore mediano del campione considerato sia accurato
  - Se si sceglie come grado di fiducia nella stima della mediana:

$$1 - \alpha = 0.95$$

  - Si dimostra che per campioni numerosi l'intervallo di confidenza approssimato per la mediana è:

$$(M_1, M_2) = \left(M - 1.57\frac{IQR}{\sqrt{n}}, M + 1.57\frac{IQR}{\sqrt{n}}\right)$$

  - M è la mediana
  - M_1 è l'estremo inferiore dell'intervallo di confidenza per la mediana
  - M_2 è l'estremo superiore dell'intervallo di confidenza per la mediana
  - $IQR = Q_1 - Q_3$ è lo scarto interquartile
  - n il numero di osservazioni nel campione.

McGill R., Tuket J.W., Larsen W.A. Variations of Box Plots. The American Statistician, 32, 11-16, 1978

1,57 è una costante derivata da un'approssimazione dell'intervallo di confidenza del 95% attorno alla mediana, sotto l'ipotesi di distribuzione **normale**

BOXPLOT AD INTAGLIO CON R

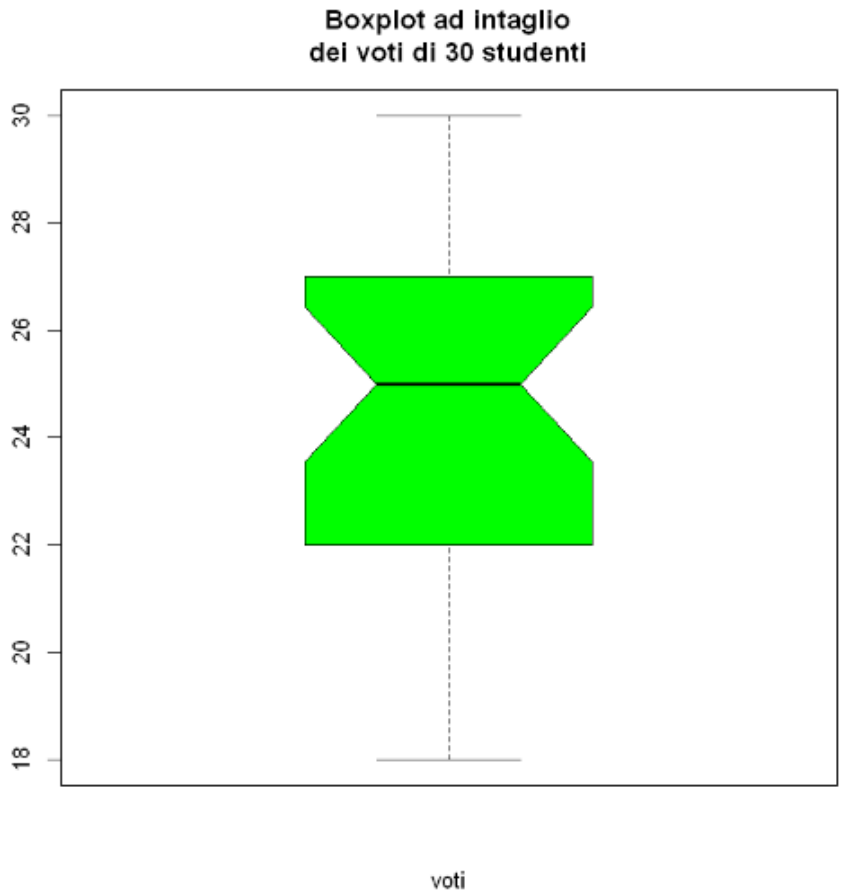
- In R un boxplot ad intaglio è ottenuto tramite la funzione **boxplot**(vettore) inserendo il parametro

notch = TRUE

- Esempio relativamente al vettore voti di 30 studenti universitari:

```
# boxplot ad intaglio
voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24, 25,
26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29, 28)
quantile(voti)
boxplot(voti, notch=TRUE, xlab ="voti", main="Boxplot ad intaglio
dei voti di 30 studenti", col ="green")
```

0%: 18 **25%:** 22 **50%:** 25 **75%:** 27 **100%:** 30



BoxPlot ad Intaglio con R

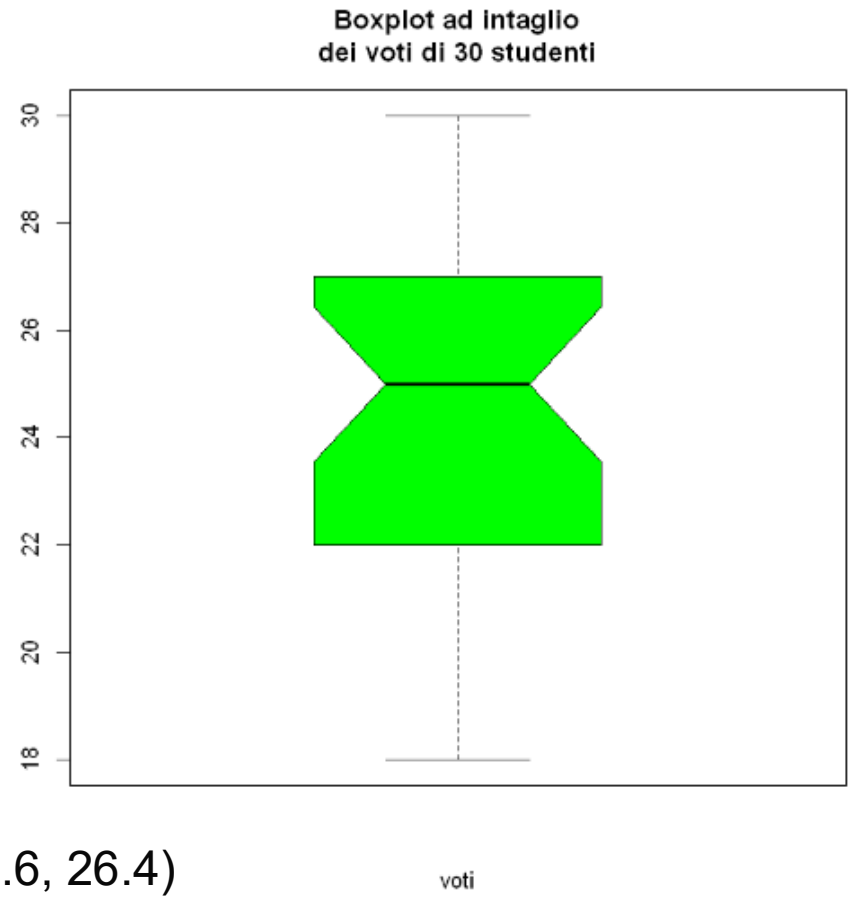
- In R un boxplot ad intaglio è ottenuto tramite la funzione **boxplot**(vettore) inserendo il parametro

notch = TRUE

- Esempio relativamente al vettore voti di 30 studenti universitari:

```
IQR <- quantile(voti,0.75)-quantile(voti,0.25)
M1 <- quantile(voti,0.5)-1.57*IQR/sqrt(length(voti))
M2 <- quantile(voti,0.5)+1.57*IQR/sqrt(length(voti))
c(M1 ,M2)
```

50%: 23.5667926411948 **50%:** 26.4332073588052



  - Lo scarto interquartile è IQR = 27 − 22 = 5
  - La mediana è M = 25
  - l'ampiezza del campione è 30
- Quindi, con un grado di fiducia del 95%:
  - L'intervallo di confidenza approssimato per la mediana è (23.6, 26.4)

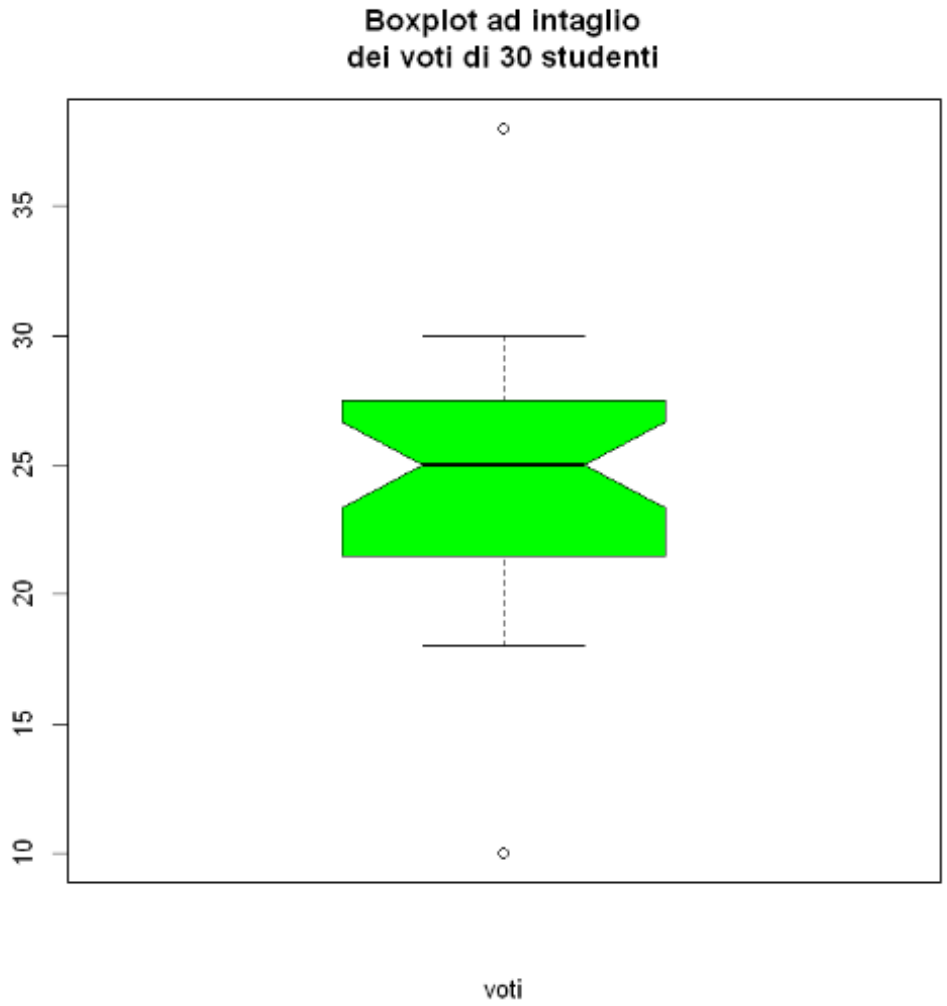
BoxPlot ad Intaglio

- Consideriamo sempre il vettore di voti utilizzato in precedenza

```
voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24, 25,
26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29, 28)
```

- Ed aggiungiamo due nuovi voti al vettore:

```
voti1<-c(voti,10,38)
```



BoxPlot ad Intaglio

- Consideriamo sempre il vettore di voti utilizzato in precedenza

```
voti <-c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24, 25,
26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29, 28)
```

- Ed aggiungiamo due nuovi voti al vettore:

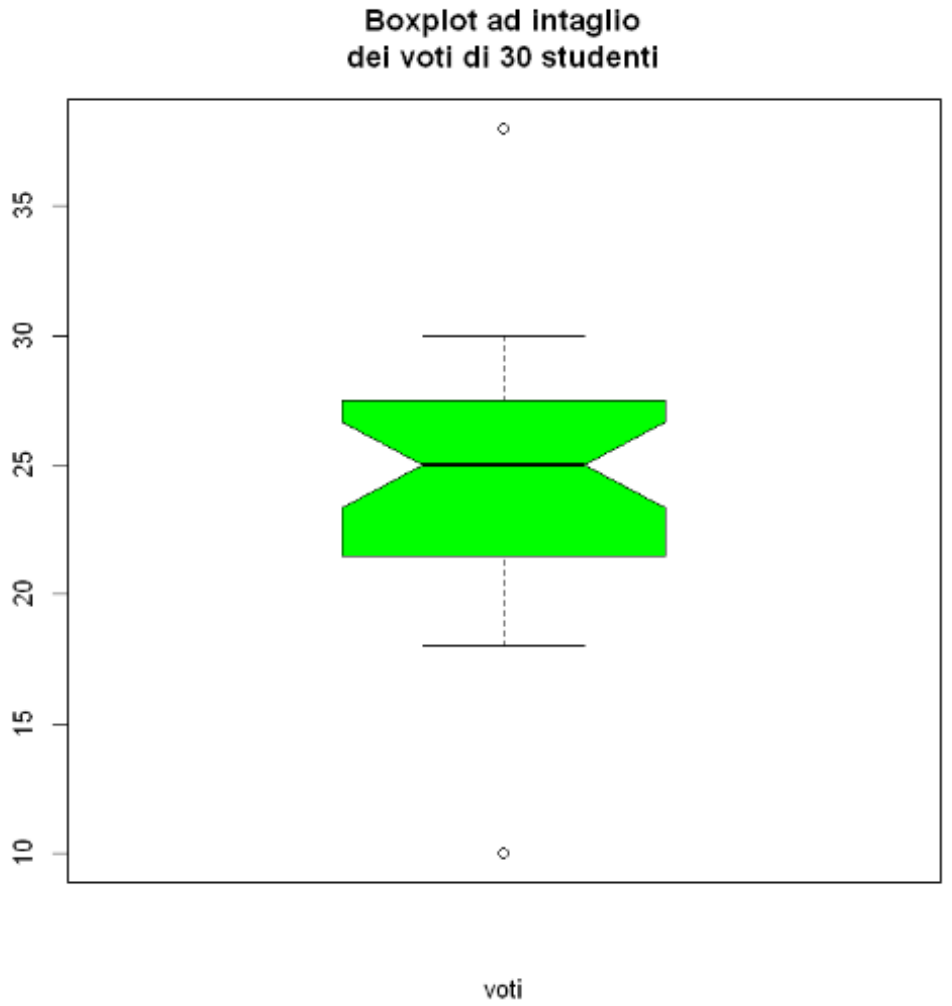
```
voti1<-c(voti,10,38)
```

- Calcoliamo i quartili e creiamo il grafico:

```
quantile(voti1)
```

0%: 10 **25%:** 21.75 **50%:** 25 **75%:** 27.25 **100%:** 38

```
boxplot(voti1, notch=TRUE, xlab ="voti", main="Boxplot ad intaglio
dei voti di 30 studenti", col ="green")
```

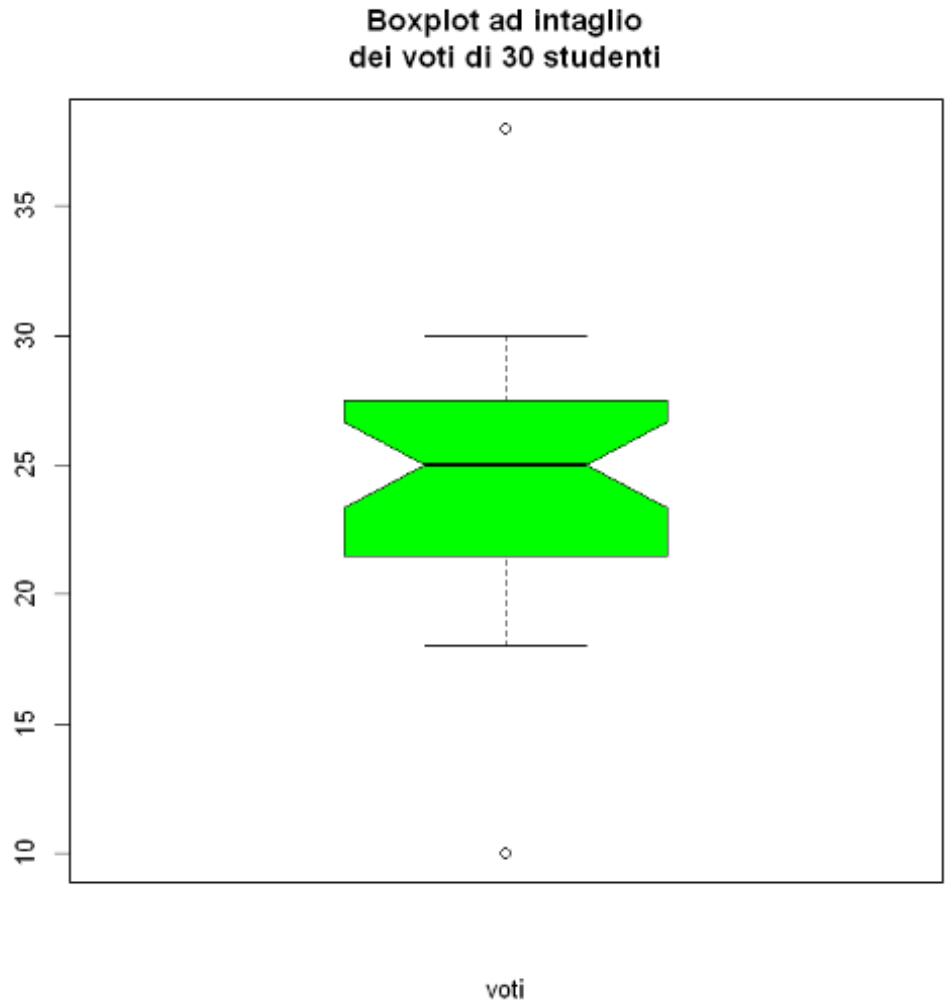


BoxPlot ad Intaglio

- Determiniamo ora l'intervallo di confidenza approssimato per la mediana del vettore aggiornato:
  - Lo scarto interquartile è IQR = 27.25 − 21.75 = 5.5
  - La mediana è M = 25
  - l'ampiezza del campione è 32
- Quindi, con un grado di fiducia del 95%:
  - L'intervallo di confidenza approssimato per la mediana è:

(23.5, 26.5)

- Verifichiamo in R se i conti tornano…



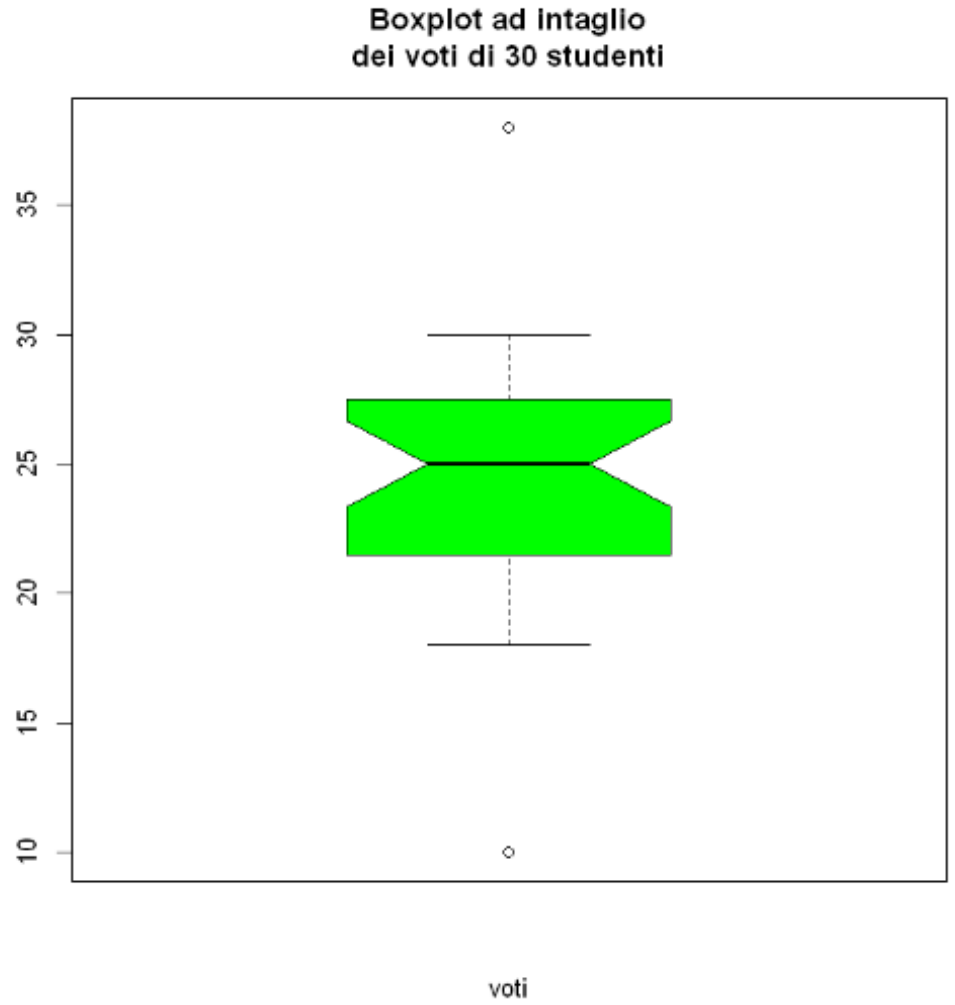
BoxPlot ad Intaglio

- Determiniamo ora l'intervallo di confidenza approssimato per la mediana del vettore aggiornato:

```
IQR <- quantile(voti1,0.75) - quantile(voti1,0.25)
M1 <- quantile(voti1,0.5) - 1.57*IQR/sqrt(length(voti1))
M2 <- quantile(voti1,0.5) + 1.57*IQR/sqrt(length(voti1))
print("IQR")
print(IQR)
print("M1")
print(M1)
print("M2")
print(M2)
print("Intervallo di confidenza approssimato:")
c(M1 ,M2)
```

```
[1] "IQR"
75%
5.5
[1] "M1"
     50%
23.47353
[1] "M2"
     50%
26.52647
[1] "Intervallo di confidenza approssimato:"
```

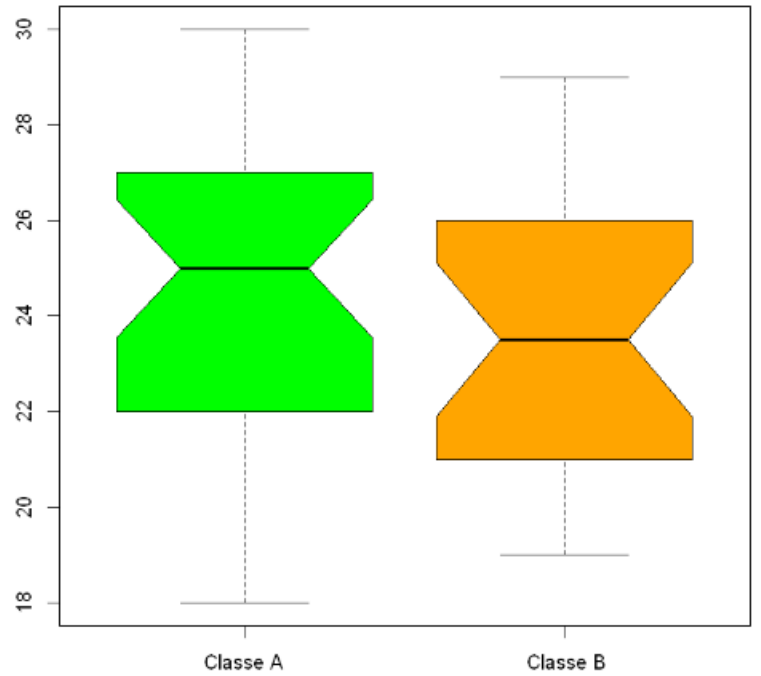
50%: 23.4735332361135 **50%:** 26.5264667638865



BoxPlot ad Intaglio

- I boxplot ad intaglio sono spesso utilizzati per confrontare gruppi
- Se si effettua un test statistico sulla differenza tra le mediane dei due gruppi con un livello di significatività del 5%, per grandi campioni le mediane dei due gruppi differiscono statisticamente se gli intervalli di confidenza dei due boxplot non si sovrappongono
- Esempio:
  - Riprendiamo i voti delle classi definiti in precedenza

```
votiClasseA <- c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27,
26, 25, 24, 25, 26, 24, 23, 22, 27, 28,
21, 24, 25, 25, 27, 19, 21, 28, 29, 28)
votiClasseB <- c(19, 19, 20, 27, 19, 28, 21, 22, 23, 26,
27, 25, 24, 25, 26, 24, 23, 22, 28, 29,
21, 24, 21, 19)

summary(votiClasseA)
summary(votiClasseB)

boxplot(votiClasseA, votiClasseB, notch=TRUE,
names=c("Classe A"," Classe B"), col=c("green","orange"))
```




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   22.00   25.00   24.47   27.00   30.00

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   21.00   23.50   23.42   26.00   29.00
```

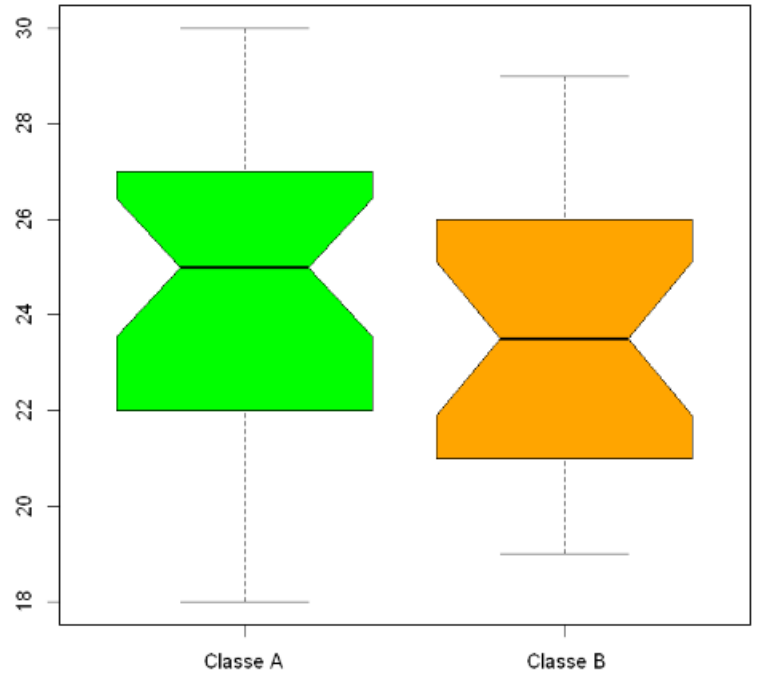
BoxPlot ad Intaglio

- Esempio:
  - Per la classe A:
    - Intervallo di confidenza approssimato per la mediana dei voti è (23.6, 26.4)
  - Per la classe B:

```
IQR <- quantile(votiClasseB,0.75) - quantile(votiClasseB,0.25)
M1 <- quantile(votiClasseB,0.5) - 1.57*IQR/sqrt(length(votiClasseB))
M2 <- quantile(votiClasseB,0.5) + 1.57*IQR/sqrt(length(votiClasseB))
c(M1 ,M2)
```

```
[1] "IQR"
75%
  5
[1] "M1"
     50%
21.89763
[1] "M2"
     50%
25.10237
[1] "Intervallo di confidenza approssimato:"
```

50%: 21.8976254599293 **50%:** 25.1023745400707




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00   22.00   25.00   24.47   27.00   30.00
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.00   21.00   23.50   23.42   26.00   29.00
```

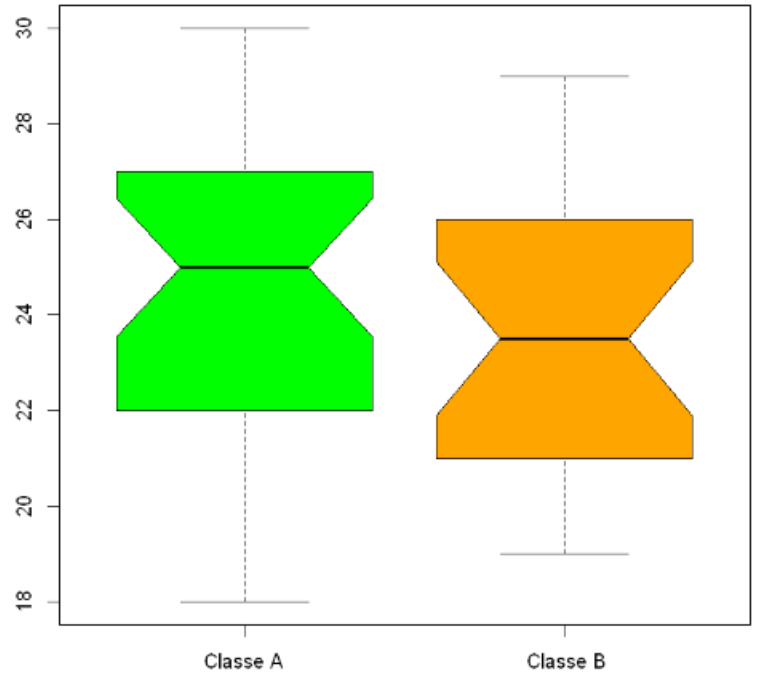
BoxPlot ad Intaglio

- Esempio:
  - Per la classe A:
    - Intervallo di confidenza approssimato per la mediana dei voti è (23.56679, 26.43321)
  - Per la classe B:
    - Intervallo di confidenza approssimato per la mediana dei voti è (21.89763, 25.1)
  - La differenza tra le mediane varia nell'intervallo

    (23.56679 - 25.10237, 26.43321 - 21.89763) =

    = **(−1.53558, 4.53558)**

    Che corrisponde a circa il **5%** di sovrapposizione delle mediane




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00   22.00   25.00   24.47   27.00   30.00

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.00   21.00   23.50   23.42   26.00   29.00
```

- Dato che si sovrappongono con un livello di significatività del 5% non si può affermare che le mediane dei voti delle due classi sono significativamente differenti

CONFRONTO DEGLI INTERVALLI

- Dal confronto degli **intervalli di confidenza** sulle mediane nei **boxplot ad intaglio**, si possono dedurre informazioni sulla **differenza statistica tra le mediane** dei gruppi confrontati

1. Sovrapposizione delle tacche
    - Se le tacche si sovrappongono tra due gruppi, significa che non ci sono prove sufficienti per affermare che le mediane dei gruppi sono significativamente diverse
        - Questo suggerisce che **non c'è evidenza di una differenza significativa tra le mediane**
2. Non sovrapposizione delle tacche
    - Se le tacche non si sovrappongono, c'è una forte indicazione che le mediane dei due gruppi sono significativamente diverse
        - Questo suggerisce che è molto probabile che la differenza osservata tra le mediane non sia dovuta al caso, ma **rappresenti una differenza reale** tra i gruppi

DOMANDE?