Statistica descrittiva con R

Silvia Parolo

21 Novembre 2014

Sintesi dei dati

Le votazioni in matematica di 20 studenti della Yale University sono state le seguenti: 68 84 75 82 68 90 62 88 76 93 73 79 88 73 60 93 71 59 85 75

Inseriamo i dati in un vettore

```
voti <- c(68, 84, 75, 82, 68, 73, 62, 88, 76, 93,73, 79,
          88, 73, 60, 93, 71, 59, 85, 75)
```

Calcoliamo le seguenti misure di posizione:

- moda

In R non esiste una funzione che calcoli la moda ma la si può estrarre utilizzando il comando `table`

```
table(voti)
```

```
## voti
## 59 60 62 68 71 73 75 76 79 82 84 85 88 93
##  1  1  1  2  1  3  2  1  1  1  1  1  2  2
```

- mediana e quartili

```
median(voti)
```

```
## [1] 75
```

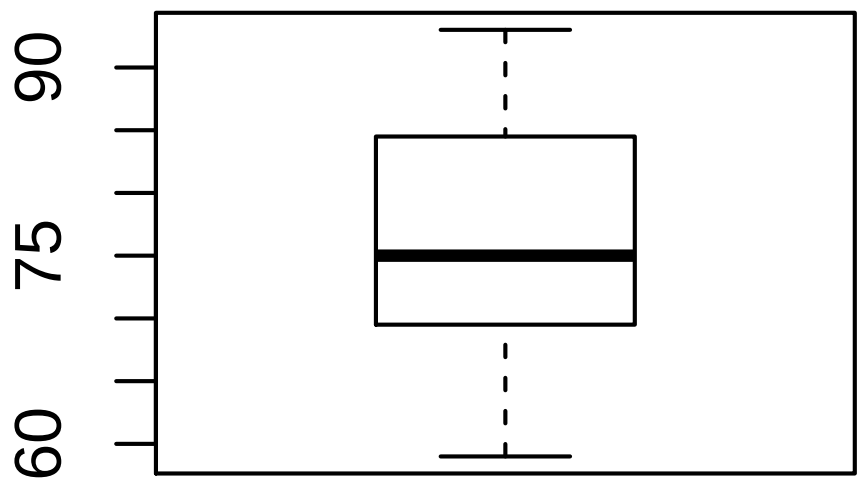
La funzione `quantile` applicata a un vettore numerico restituisce il minimo, il massimo e i tre quartili principali. Oppure quantili specifici di interesse.

```
quantile(voti)
```

```
##    0%   25%   50%   75%  100%
## 59.00 70.25 75.00 84.25 93.00
```

I quartili si possono rappresentare anche graficamente attraverso un grafico chiamato `boxplot`. Il boxplot è il disegno di una scatola i cui estremi sono il primo (Q1) e il terzo quartile (Q3) e la riga in mezzo è la mediana (Q2). In alto e in basso ci sono altre due righe orizzonatli che sono poste a 1.5 volte la distanza tra Q3 e Q1. I valori restanti sono indicati con dei punti.

```
boxplot(voti)
```



- media

```
mean(voti)
```

```
## [1] 76.25
```

Indici di dispersione

- varianza

```
var(voti)
```

```
## [1] 104.1
```

- deviazione standard

```
sd(voti)
```

```
## [1] 10.2
```

- coefficiente di variazione

```
sd(voti)/mean(voti)
```

```
## [1] 0.1338
```

Trovate ora :

- La votazione più bassa
- La votazione più alta
- La votazione dei cinque migliori studenti
- La votazione del decimo studente, cominciando dal migliore
- Quanti studenti hanno avuto la votazione di 75 o migliore
- Che percentuale di studenti ha avuto una votazione più alta di 65 ma inferiore a 85

La votazione più bassa

```
min(voti)
```

```
## [1] 59
```

La votazione più alta

```
max(voti)
```

```
## [1] 93
```

La votazione dei cinque migliori studenti

```
voti_sorted <- sort(voti,decreasing=TRUE)
best5 <- voti_sorted[1:5]
best5
```

```
## [1] 93 93 88 88 85
```

La votazione del decimo studente, cominciando dal migliore

```
best10 <- voti_sorted[10]
best10
```

```
## [1] 75
```

Quanti studenti hanno avuto la votazione di 75 o migliore

```
voti_sorted >= 75 #vettore di TRUE e FALSE
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TR
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FAL
```

```
#poi faccio il subset del vettore
n <- voti_sorted[voti_sorted >= 75 ]
#infine ottengo la lunghezza del vettore
length(n)
```

```
## [1] 11
```

Che percentuale di studenti ha avuto una votazione più alta di 65 ma inferiore a 85

```
#numero di voti esaminati
length(voti_sorted)
```

```
## [1] 20
```

```
#subset >65 & < 85
s <- voti_sorted[voti_sorted > 65 & voti_sorted < 85]
#percentuale
(length(s)/length(voti_sorted))*100
```

```
## [1] 60
```

Distribuzione di frequenza

Calcoliamo la distribuzione di frequenze assolute

```
table(voti_sorted)
```

```
## voti_sorted
## 59 60 62 68 71 73 75 76 79 82 84 85 88 93
##  1  1  1  2  1  3  2  1  1  1  1  1  2  2
```

Calcoliamo la distribuzione di frequenze relative

```
table(voti_sorted)/length(voti_sorted)
```

```
## voti_sorted
##    59    60    62    68    71    73    75    76    79    82    84
## 0.05 0.05 0.05 0.10 0.05 0.15 0.10 0.05 0.05 0.05 0.05 (
```

Calcoliamo la distribuzione di frequenze percentuali

```
table(voti_sorted)/length(voti_sorted)*100
```

```
## voti_sorted
## 59 60 62 68 71 73 75 76 79 82 84 85 88 93
##  5  5  5 10  5 15 10  5  5  5  5  5 10 10
```

Calcoliamo le frequenze cumulate

```
#assolute
cumsum(table(voti_sorted))
```

```
## 59 60 62 68 71 73 75 76 79 82 84 85 88 93
##  1  2  3  5  6  9 11 12 13 14 15 16 18 20
```

```
#relative
cumsum(table(voti_sorted)/length(voti_sorted))
```

```
##   59   60   62   68   71   73   75   76   79   82   84
## 0.05 0.10 0.15 0.25 0.30 0.45 0.55 0.60 0.65 0.70 0.75 (
```

La distribuzione in classi

Quando si vogliono riassumere grandi quantità di dati spesso è utile distribuire i dati in classi e determinare il numero di individui appartenenti a ciascuna classe.

```
#Definiamo i punti estremi delle classi
estremi_classi <- c(50,60,70,80,90,100)
#con la funzione cut divido la distribzione in classi
#classi chiuse a sinistra
classi <- cut(voti_sorted, breaks=estremi_classi,
              right=FALSE)
fr.ass <- table(classi)
fr.ass
```

```
## classi
##  [50,60)  [60,70)  [70,80)  [80,90) [90,100)
##        1        4        8        5        2
```

frequenze cumulate

```
fr.cum <- cumsum(table(classi))
fr.cum
```

```
##   [50,60)   [60,70)   [70,80)   [80,90) [90,100)
##         1         5        13        18        20
```

frequenza relativa e frequenza relativa cumulata

```
#frequenza relativa
fr.rel <- table(classi) / (sum(table(classi)))
fr.rel
```

```
## classi
##  [50,60)  [60,70)  [70,80)  [80,90) [90,100)
##     0.05     0.20     0.40     0.25     0.10
```

```
#frequenza relativa cumulata
#applico la funzione cumsum
fr.rel.cum <- cumsum(table(classi)) / sum(table(classi))
fr.rel.cum
```

```
##  [50,60)  [60,70)  [70,80)  [80,90) [90,100)
##     0.05     0.25     0.65     0.90     1.00
```

Creiamo una tabella con i risultati

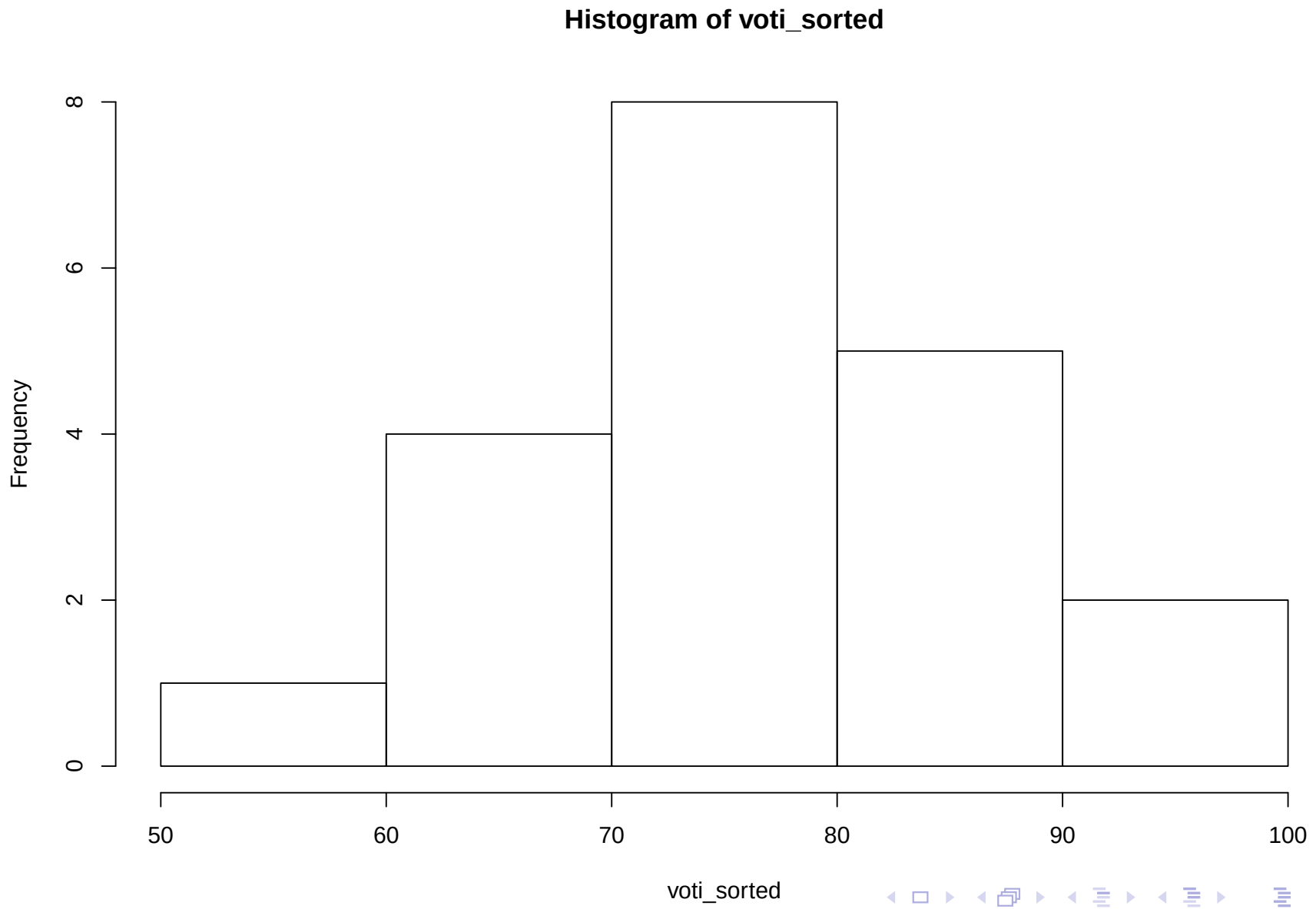
attraverso la funzione `cbind` possiamo mettere le varie frequenze che abbiamo calcolato in colonne parallele

```
results <- cbind(fr.ass,fr.cum,fr.rel,fr.rel.cum)
results
```

```
##           fr.ass fr.cum fr.rel fr.rel.cum
## [50,60)        1      1   0.05        0.05
## [60,70)        4      5   0.20        0.25
## [70,80)        8     13   0.40        0.65
## [80,90)        5     18   0.25        0.90
## [90,100)       2     20   0.10        1.00
```

Rappresentazione grafica delle distribuzioni di frequenza

```
hist(voti_sorted, breaks=estremi_classi,right=F)
```



Esercizio riassuntivo

Il dataset nel file *peso_nascita.csv* contiene l'informazione relativa al peso alla nascita di 15 bambini.

- Importa il dataset in R
- Quante variabili sono riportate?
- Quanti individui?
- Calcola media e deviazione standard del peso.
- Calcola i quartili e rappresentali graficamente

- Se il peso alla nascita è inferiore a 2.5 kg il neonato è definito sottopeso.

Crea una nuova variabile che divida i neonati in due gruppi: normopeso e sottopeso. Quanti neonati sono normopeso? Quanti sottopeso? Qual è l'età media delle madri dei bambini normopeso e di quelli sottopeso?

- Dividi il peso alla nascita in 4 classi. Calcola distribuzione di frequenza assoluta, relativa e cumulata delle classi.