

LEZIONE_2

2022-10-19

INDICE

- INTRODUZIONE
- VARIABILI STATISTICHE
- DISTRIBUZIONI DI FREQUENZA
- [RAPPRESENTAZIONI GRAFICHE]
- [INDICI SINTETICI]

INTRODUZIONE

- DISTINZIONE tra ANALISI
 - UNIVARIATE: 1 CARATTERISTICA
 - MULTIVARIATE: 2+ CARATTERISTICHE
- PRIMA di un'analisi INFERENZIALE è opportuno svolgerne una DESCRITTIVA iniziale, per comprendere il CONTESTO TEORICO

ANALISI ESPLORATIVA

- TIPOLOGIA DI DATI
 - OSSERVAZIONALI / SPERIMENTALI
 - CAMPIONARI / CENSITI
 - VARIABILI STATISTICHE
- INDIVIDUARE UNITÀ STATISTICHE
 - DATI MANCANTI
 - DATI SPORCHI
- PULIZIA DATI
 - CODIFICA
 - ORGANIZZAZIONE
- METODI GRAFICI

VARIABILI STATISTICHE

IDENTIFICANO LE PROPRIETÀ DELLE UNITÀ STATISTICHE

MODALITÀ

SONO I VALORI CHE UNA VARIABILE PUÒ ASSUMERE

NOTAZIONI

- Y = variabile generica
- y = modalità generica
- Y' = dominio dei valori ammessi da Y
 - $S_y = \{y_1, \dots, y_j\}$ con $j \leq N$
 - * per qualsiasi $j \neq i \rightarrow y_j \neq y_i$
 - N = numero di unità statistiche considerate

TIPOLOGIE

- **VARIABILI QUALITATIVE (CATEGORIALI) \rightarrow STRINGHE TESTUALI:**
 - **SCONNESSE (NOMINALI):** non è possibile individuare un'ordine "naturale"
 - * religione
 - * colore degli occhi
 - * genere
 - **ORDINALI:** è possibile identificare un ordine
 - * livello di istruzione
 - * gerarchie
 - **DICOTOMICHE:** se $|Y'| = 2$
 - * i due valori ammessi possono essere codificati come 0 e 1, risparmiando memoria e preservando la quantità di informazione
- **VARIABILI QUANTITATIVE (NUMERICHE):**
 - **DISCRETE:** se Y' è un insieme FINITO
 - **CONTINUE:** se Y' possiede un range continuo di valori, e ogni valore è valido
 - **INTERVALLI:** non esiste uno 0 arbitrario
 - **RAPPORTI:** esiste uno 0 arbitrario

GERARCHIA

1. QUANTITATIVE CONTINUE
2. QUANTITATIVE DISCRETE
3. QUALITATIVE ORDINALI
4. QUALITATIVE NOMINALI

SALENDO DI LIVELLO GERARCHICO SI AUMENTA LA QUANTITÀ DI INFORMAZIONE

DISTRIBUZIONI DI FREQUENZA

FREQUENZE ASSOLUTE

```
head(mtcars)

##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710      22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive  21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant        18.1   6  225 105 2.76 3.460 20.22  1  0    3    1

Y = mtcars$cyl # SCEGLIAMO LA VARIABILE Y = mtcars$cyl
(Sy = unique(Y)) # SUPPORTO DELLA VARIABILE Y

## [1] 6 4 8

(J = length(Sy)) # LUNGHEZZA DEL SUPPORTO

## [1] 3

(N = length(Y)) # NUMERO DI UNITÀ STATISTICHE

## [1] 32

(cylFreqTable = group_by(mtcars,cyl) %>% summarise(frequency = n()) ) # ANALOGO A --> (cylFreqTable =

## # A tibble: 3 x 2
##   cyl frequency
##   <dbl>     <int>
## 1     4         11
## 2     6          7
## 3     8         14

# LA SOMMA DI TUTTE LE FREQUENZE ASSOLUTE = NUMERO DI UNITÀ STATISTICHE
sum(cylFreqTable$frequency) == N

## [1] TRUE

VARIABILE QUANTITATIVA CONTINUA -> J QUASI UGUALE A N
È OPPORTUNO CREARE DELLE CLASSI DI MODALITÀ E CONTARE LE OCCORRENZE
IN ESSE
```

CLASSI DI MODALITÀ

- RAPPRESENTANO DEI SOTTOINSIEMI DEL RANGE DI UNA VARIABILE
- DEVONO ESSERE NE TROPPE NE TROPPO POCHE
 - NUMERO OTTIMALE CIRCA $N^{(1/2)}$
- DEVONO RAPPRESENTARE DEGLI INTERVALLI DISGIUNTI
 - $[y_0, y_1)$ -> y_1 escluso
 - $[y_1, y_2)$ -> y_2 escluso

```
Sy = levels(factor(mtcars$mpg))
length(Sy)
```

```
## [1] 25
# ci sono tanti possibili valori ammessi da mpg (Miles Per Gallon)
(mpgRange = max(mtcars$mpg)-min(mtcars$mpg) )# ==> range(mtcars$mpg)
```

```
## [1] 23.5
(classList = split(mtcars$mpg, cut(mtcars$mpg, length(Sy)^0.5)) )
```

```
## $(10.4,15.1]`
## [1] 14.3 10.4 10.4 14.7 13.3 15.0
##
## $(15.1,19.8]`
## [1] 18.7 18.1 19.2 17.8 16.4 17.3 15.2 15.5 15.2 19.2 15.8 19.7
##
## $(19.8,24.5]`
## [1] 21.0 21.0 22.8 21.4 24.4 22.8 21.5 21.4
##
## $(24.5,29.2]`
## [1] 27.3 26.0
##
## $(29.2,33.9]`
## [1] 32.4 30.4 33.9 30.4
```

```
(classFreqTable = data.frame(
  "freq"= unlist(lapply(classList, length))
))
```

```
##          freq
## (10.4,15.1]    6
## (15.1,19.8]   12
## (19.8,24.5]    8
## (24.5,29.2]    2
## (29.2,33.9]    4
```

```
sum(classFreqTable$freq) == length(mtcars$mpg)
```

```
## [1] TRUE
```

FREQUENZE RELATIVE

INDICA IL RAPPORTO TRA LA FREQUENZA ASSOLUTA E IL NUMERO TOTALE DI UNITÀ STATISTICHE

```
#  $p_1 = f_1 / \sum([f_1, f_2, \dots, f_j]) \quad j \leq N \rightarrow N = n^\circ \text{ unità}$ 
#      =  $f_1 / n$ 
```

```
(cylRelFreqTable = table(mtcars$cyl)/length(mtcars$cyl)*100) # *100 finale serve a mostrare i valori p
```

```
##
##          4          6          8
## 34.375 21.875 43.750
```

```
cylRelFreqTable[[1]]/100 * length(mtcars$cyl) == cylFreqTable[cylFreqTable$cyl==4,]$frequency
```

```
## [1] TRUE
```

```
# calcola la frequenza assoluta a partire da quella relativa e la confronta con quella assoluta effettiva
```

FREQUENZE CUMULATE

- **VARIABILI ORDINABILI, QUALITATIVE E QUANTITATIVE**
- F_i = Frequenza assoluta con cui si presentano modalità con ordini $\leq i$ -esimo ordine
 - Dato $F = \{f_1, \dots, f_j\}$ $j = |S_y| \leq N$ n° unità
 - $F_1 = f_1 \rightarrow F_j = N$
- P_i = Frequenza relativa cumulata, analoga a F_i
 - Dato $P = \{p_1, \dots, p_j\} \rightarrow |P| = |F|$
 - IPOTESI: $P_1 = p_1 \rightarrow P_j = N$

```
(cylCumFreq = cumsum(table(mtcars$cyl)))      # FREQUENZE CUMULATE

##      4      6      8
##    11    18    32

l1 = nrow(cylFreqTable) # = nrow(cylFreqTable)
if ( (cylCumFreq[[1]] == cylFreqTable[1,"frequency"]) &&
     (cylCumFreq[[l1]] == N)      # N definito in precedenza come n° di unità
) {
  print("IPOTESI CONFERMATA")
}

## [1] "IPOTESI CONFERMATA"

ok=TRUE
for(i in 1:nrow(cylFreqTable)){
  if(cylCumFreq[[names(cylCumFreq)[i]]] != sum(cylFreqTable[1:i,"frequency"])){
    ok = FALSE;
    break;
  }
}
if (ok) {
  print("cumsum WORKS!!!")
} else {
  print("SOMETHING IS WRONG!!!")
}

## [1] "cumsum WORKS!!!"

(cylRelCumFreq = data.frame(Pi = cumsum(cylRelFreqTable)))      # frequenze relative cumulate

##      Pi
## 4  34.375
## 6  56.250
## 8 100.000
```

RAPPRESENTAZIONI GRAFICHE

DATI QUALITATIVI

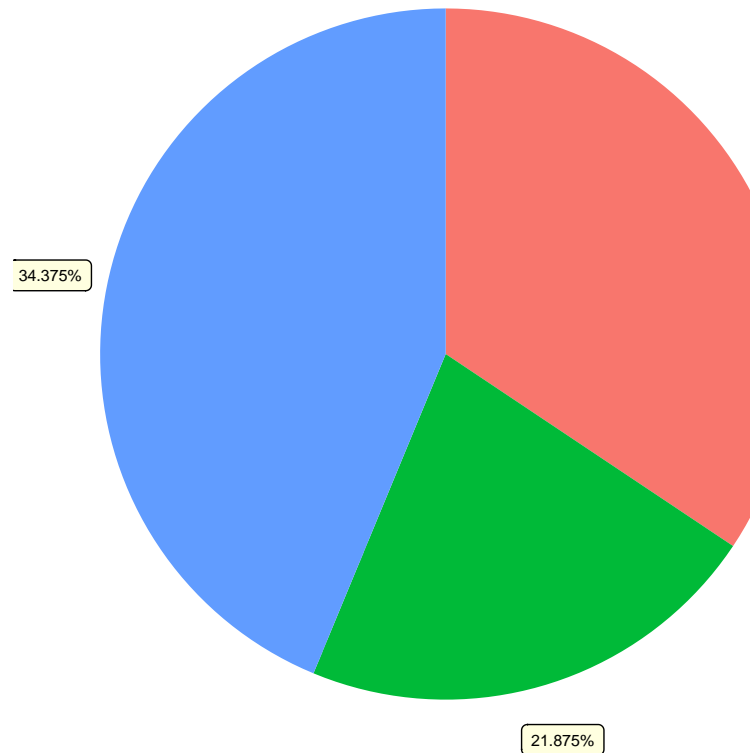
```
tab1 <- as.data.frame(table(mtcars$cyl))
colnames(tab1) = c("CYL","FREQ")
```

```

pos <- cumsum(rev(tab1$FREQ)) - rev(tab1$FREQ)/2
relFreq <- tab1$FREQ / length(mtcars$cyl) *100
freqLabels <- paste(relFreq,"%",sep = "")      # etichette

tab1 %>% ggplot(aes(x = factor(1), y = FREQ,
                    fill = CYL)) +
geom_col() +
coord_polar(theta = "y",
            direction = -1) +
theme_void() +
# geom_label
geom_label(x = 1.6,                                # etichette all'esterno
          y = pos,
          aes(label = freqLabels),
          fill = "lightyellow",
          size = 2)

```

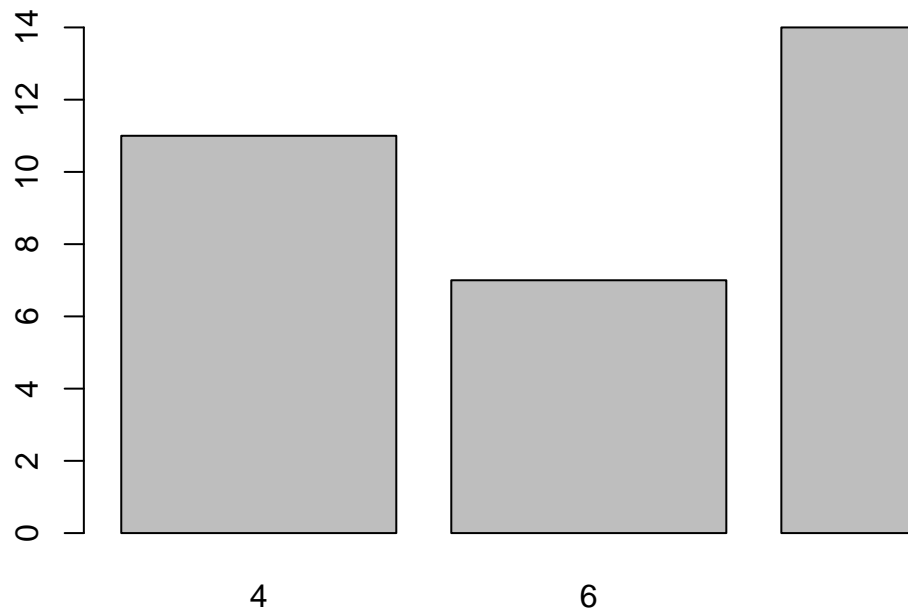


DIAGRAMMI CIRCOLARI(A TORTA)

```

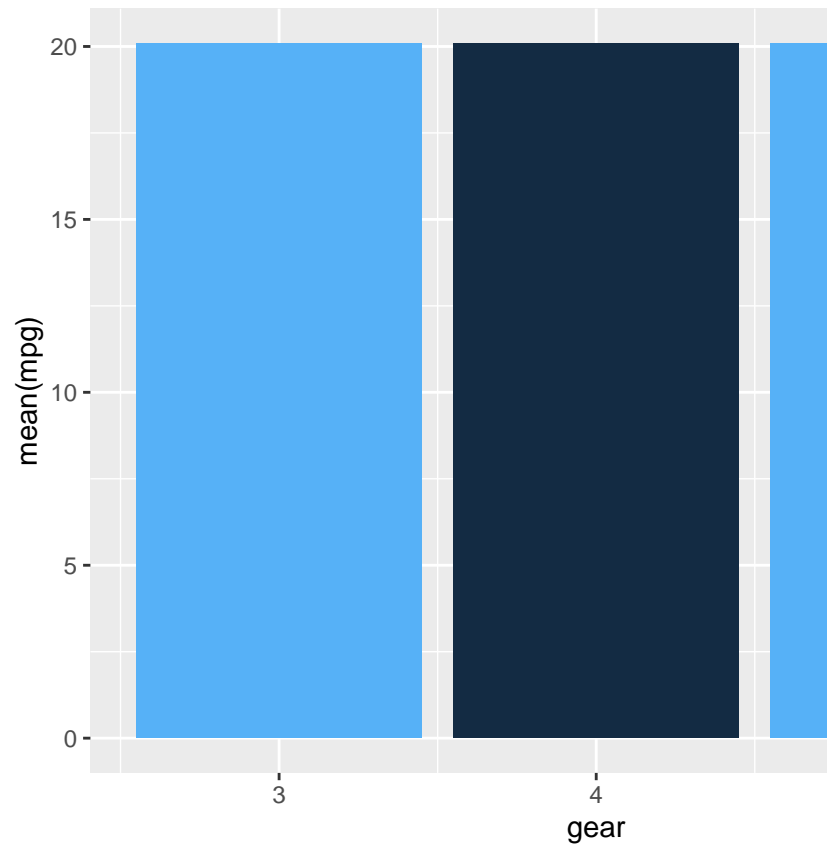
barplot(height = table(mtcars$cyl))

```



DIAGRAMMI A RETTANGOLI

```
ggplot(mtcars, aes(x=gear,y=mean(mpg), fill=cyl)) + geom_bar(stat = "identity", position = "dodge")
```



DIAGRAMMI A RETTANGOLI MULTIPLI

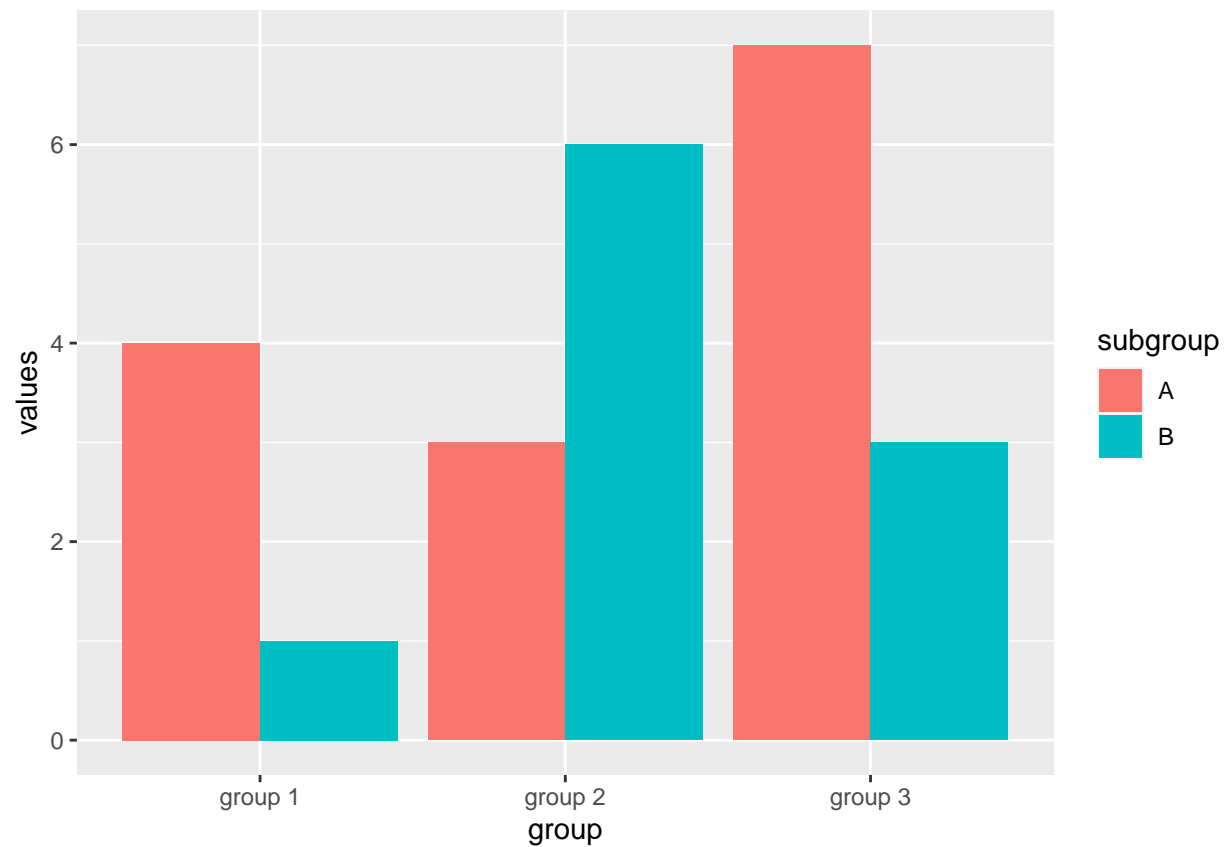
DATI QUANTITATIVI

```
data <- data.frame(values = c(4, 1, 3, 6, 7, 3), # Create example data
  group = rep(c("group 1",
    "group 2",
    "group 3"),
    each = 2),
  subgroup = LETTERS[1:2])
head(data)
```

DIAGRAMMI A BASTONCINI

```
##   values  group subgroup
## 1     4 group 1      A
## 2     1 group 1      B
## 3     3 group 2      A
## 4     6 group 2      B
## 5     7 group 3      A
## 6     3 group 3      B
```

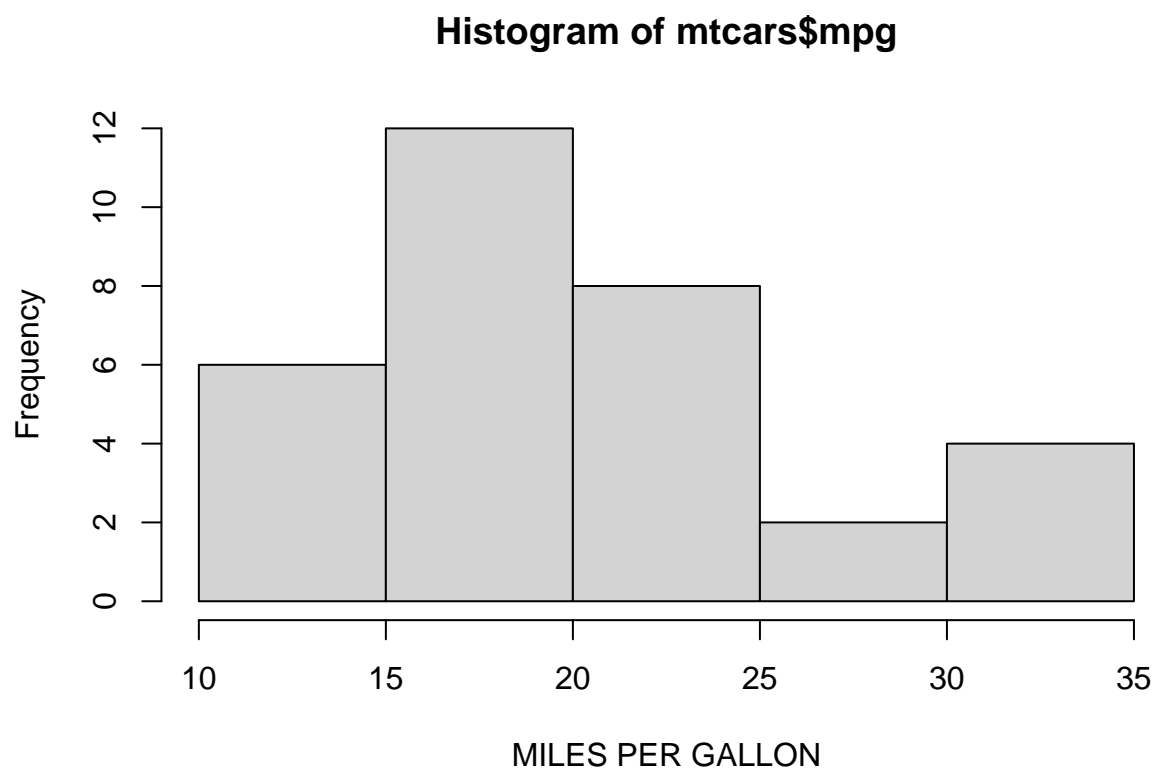
```
ggplot(data, # Grouped barplot using ggplot2
  aes(x = group,
    y = values,
    fill = subgroup)) +
  geom_bar(stat = "identity",
    position = "dodge")
```

ISTOGRAMMI

- Le basi dei rettangoli sono proporzionali alle classi definite per suddividere il range continuo dei valori dell'asse X
- In questo esempio

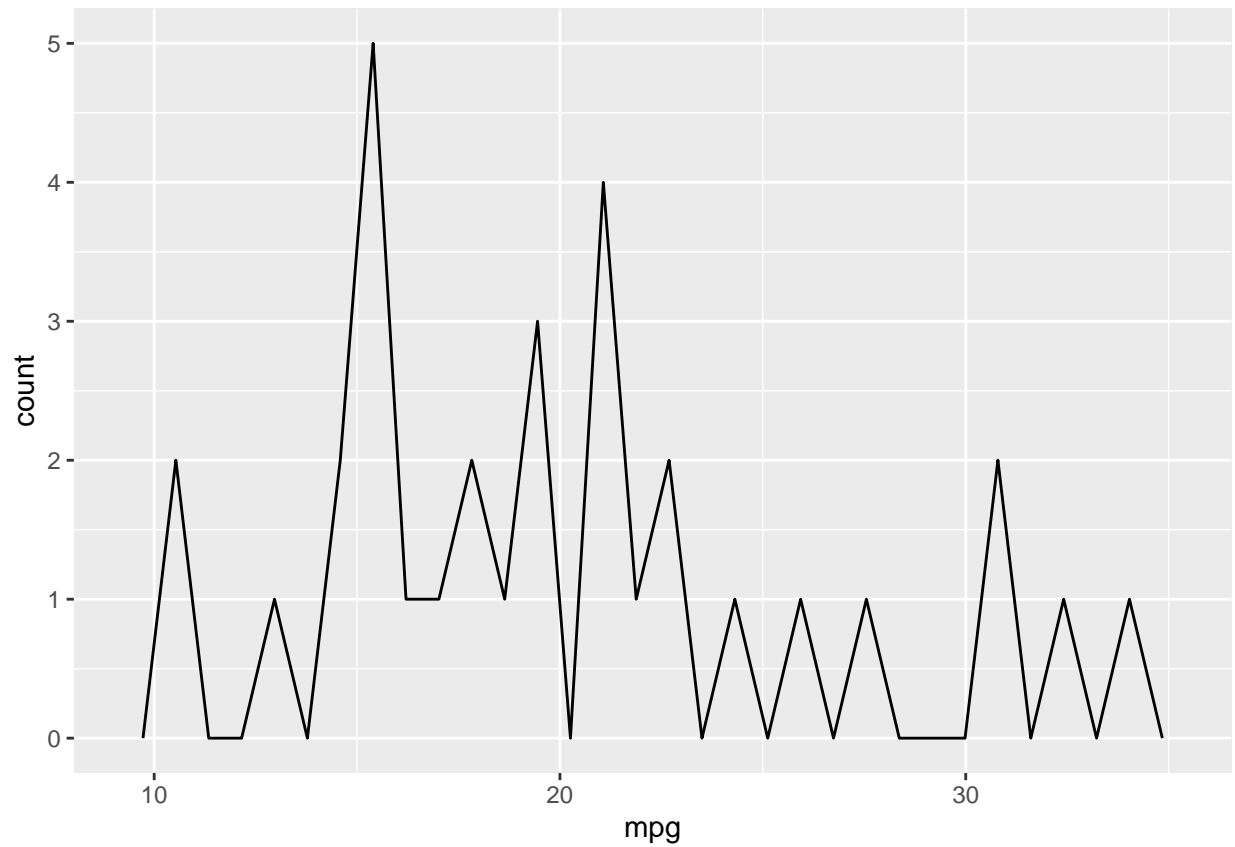
```
# hist(x = dati$nominees, xlab = "Nominees", main = "Nominees Frequency")
hist(mtcars$mpg, xlab="MILES PER GALLON")
```



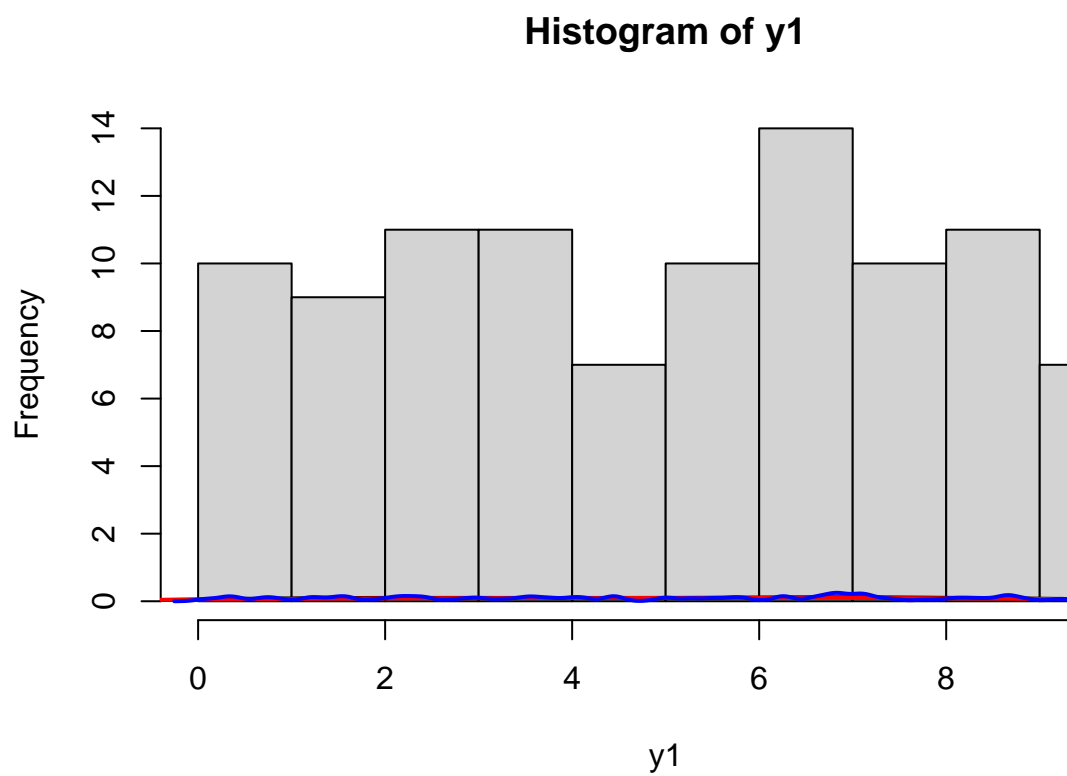
```
ggplot(mtcars) + geom_freqpoly(aes(x=mpg))
```

POLIGONI DI FREQUENZA

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
y1 = runif(100,0,10)
hist(y1)
lines(density(y1),lwd=2)      #scelta ottimale della banda
lines(density(y1,bw=2),lwd=2,col="red")    #scelta ottimale della banda
lines(density(y1,bw=0.1),lwd=2,col="blue")  #scelta ottimale della banda
```

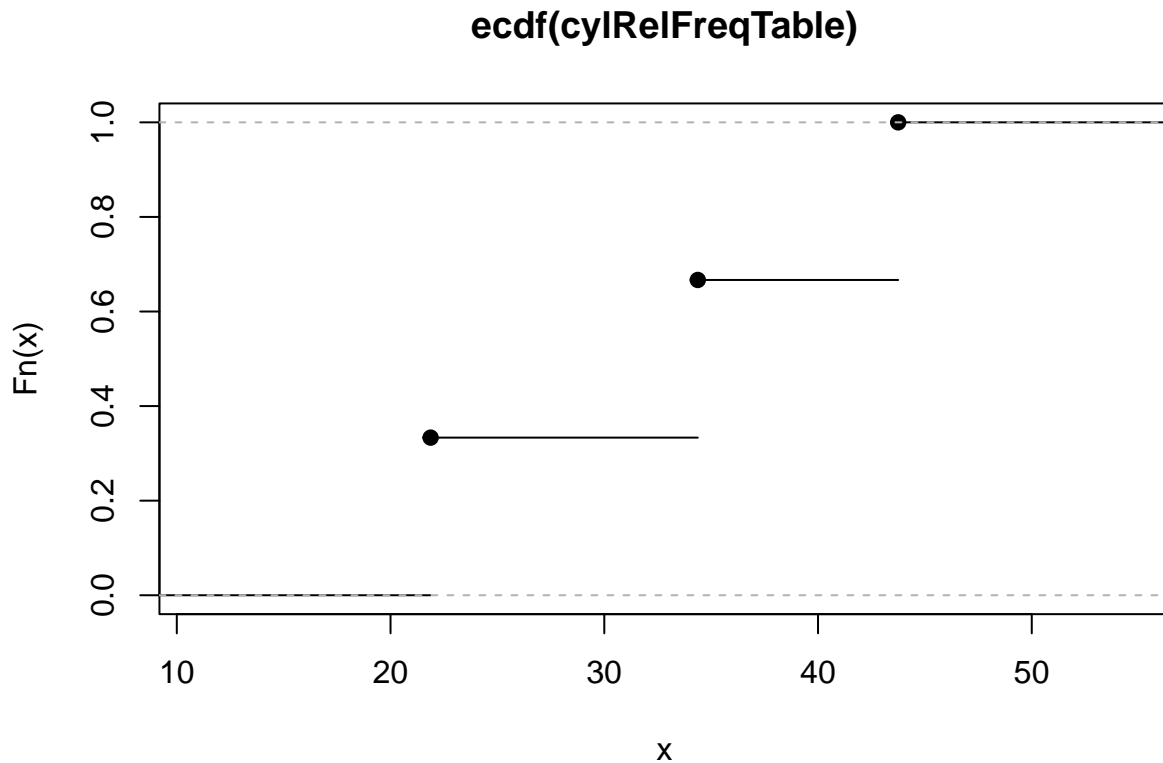


STIMA DELLA DENSITÀ

FUNZIONI DI RIPARTIZIONE EMPIRICA

- Rappresenta graficamente l'andamento delle frequenza cumulate

```
plot(ecdf(cylRelFreqTable))
```



-BOXPLOT

INDICI SINTETICI

POSIZIONE

ESPRESSO NELLA STESSA UNITÀ DI MISURA DELLA VARIABILE Y DI RIFERIMENTO

MEDIA ARITMETICA

È CALCOLABILE per le VARIABILI QUANTITATIVE e QUALITATIVE DICOTOMICHE, dopo opportuna codifica numerica in 0 e 1

$Y = \text{VARIABILE} \rightarrow E(Y) = \text{MEDIA ARITMETICA DI TUTTI I VALORI DI } Y$

- Dato $Y = \{y_1, \dots, y_n\}$ $n = \text{totale unità statistiche}$
- $E(Y) = (y_1 + \dots + y_n) / n \rightarrow \text{se si dispone dei dati grezzi}$
- $E(Y) = 1/n * \sum_{i=1}^J f_i * y_i \rightarrow J = |S_y|$
 - FREQUENZE ASSOLUTE
- $E(Y) = 1/n * \sum_{i=1}^J p_i * y_i \rightarrow J = |S_y|$
 - FREQUENZE RELATIVE
- $E(Y) = 1/n * \sum_{i=1}^J y_{ci} * y_i \rightarrow J = |S_y|$

- y_i = valore centrale dell'intervallo $[y_{i-1}; y_i]$
- si usa quando l'asse X è suddiviso in classi

```
Y = mtcars$mpg
mean(Y)
```

```
## [1] 20.09062
```

```
mean(Y) == sum(Y)/length(Y)
```

```
## [1] TRUE
```

```
FY = data.frame(table(Y))
```

```
head(FY) #tabella delle frequenze assolute fi, i in [1,|Sy|]
```

```
##      Y Freq
## 1 10.4    2
## 2 13.3    1
## 3 14.3    1
## 4 14.7    1
## 5  15    1
## 6 15.2    2
```

```
# for (i in 1:nrow(FY)){
#   *FY[i,2]
# }
```

ROBUSTEZZA: in caso di valori molto discostanti dal “centro” il valore della media aritmetica può venire sballato

PROPRIETÀ DI CAUCHY:

- Dato il supporto $S_y = \{y_1, \dots, y_J\}$ con $y_i < y_{i+1}$ per qualsiasi i
- IPOTESI $y_1 \leq E(Y) \leq y_J$
- $y_1 \leq y_i \leq y_J$ con $i \in [1; J]$
- $y_1 * p_i \leq y_i * p_i \leq y_J * p_i$ con $i \in [1; J]$
- $\sum_{k=1}^J y_1 * p_k \leq \sum_{k=1}^J y_k * p_k \leq \sum_{k=1}^J y_J * p_k$
- $y_1 * \sum_{k=1}^J p_k \leq \sum_{k=1}^J y_k * p_k \leq y_J * \sum_{k=1}^J p_k$
 - $E(Y) = \sum_{k=1}^J y_k * p_k$
 - $\sum_{k=1}^J p_k = 1 \rightarrow$ la somma di tutte le frequenze relative da il 100%
- $y_1 \leq \sum_{k=1}^J y_k * p_k \leq y_J$
- $y_1 \leq E(Y) \leq y_J \rightarrow$ IPOTESI CONFERMATA

PROPRIETÀ DEL BARICENTRO

- Data la variabile scarto $Sc = Y - E(Y)$
- $E(Sc) = 0$
 - $E(Y - E(Y)) = 1/n * \sum_{i=1}^n (y_i - E(Y)) =$
 - $1/n * [\sum_{i=1}^n y_i - \sum_{i=1}^n E(Y)]$

$$- E(Y) - 1/n * n * E(Y) = E(Y) - E(Y) = 0$$

PROPRIETÀ DI LINEARITÀ

- DATA LA VARIABILE Y
 - $aY + b$ = trasformazione lineare di Y con $a, b \in R$
- IPOTESI: $E(aY+b) = a * E(Y) + b$
- $E(aY+b) = 1/n * \sum_{i=1}^n (a * y_i + b) =$
- $1/n * [\sum_{i=1}^n a * y_i + \sum_{i=1}^n b] =$
- $a * 1/n * \sum_{i=1}^n y_i + 1/n * n * b =$
- $a * E(Y) + b \rightarrow$ IPOTESI CONFERMATA

```
#IPOTESI CAUCHY min(Y) < E(Y) < max(Y)
if (min(mtcars$mpg) < mean(mtcars$mpg) && mean(mtcars$mpg) < max(mtcars$mpg)){
  print("IPOTESI DI CAUCHY CONFERMATA")
}
```

ESEMPIO

```
## [1] "IPOTESI DI CAUCHY CONFERMATA"

#IPOTESI TEOREMA BARICENTRO E(Y-E(Y))=0
Y = mtcars$mpg
mean(Y) # E(Y)

## [1] 20.09062

# Y-mean(Y) = variabile scaro
if (mean(Y-mean(Y))==0) {
  print("TEOREMA BARICENTRO CONFERMATO")
}
```

MEDIANA

- CALCOLABILE PER VARIABILI QUALITATIVE ORDINALE O QUANTITATIVA
- $y_{0.5}$ = MEDIANA \rightarrow quantile di livello 0.5

GREZZI

- $y_{0.5}$ divide a metà Y in modo che i valori precedenti siano \leq e successivi $\leq y_{0.5}$
- se n è dispari allora $y_{0.5}$ si trova nell'indice $i = (n+1)/2$
- se n è pari allora $y_{0.5}$ possiede due indici $n/2$ e $n/2+1$. i due valori negli indici possono essere uguali o diversi
 - MEDIANA POSSIEDE 2 VALORI
 - dato l'intervallo di valori tra i due indici precedenti, la mediana viene rappresentata come VALORE INTERMEDIO tra i due ESTREMI dell'intervallo

```
Y = nycflights13::airports$lat
Y=sort(Y)
(n = length(Y))
```

```
## [1] 1458
if (n%%2==0){
  print("N PARI, MEDIANA POSSIEDE DUE INDICI")
  i=n/2
  j=n/2 + 1
  print(Y[i])
  print(Y[j])

  print(median(Y))
  # confronto con valore medio tra i due estremi
  if((Y[i]+Y[j])/2 == median(Y)){
    print("la mediana corrisponde al valore medio tra i due estremi individuati dai due indici")
  }
  print("MEDIANA = QUANTILE LIVELLO 0.5")
  quantile(Y,0.5)
} else {
  print("N DISPARI, MEDIANA UNICO VALORE")
  i=(n+1)/2
  print(Y[i])
  print(median(Y))
  print("MEDIANA = QUANTILE LIVELLO 0.5")
  quantile(Y,0.5)
}
```

```
## [1] "N PARI, MEDIANA POSSIEDE DUE INDICI"
## [1] 40.08194
## [1] 40.0935
## [1] 40.08772
## [1] "la mediana corrisponde al valore medio tra i due estremi individuati dai due indici"
## [1] "MEDIANA = QUANTILE LIVELLO 0.5"

##      50%
## 40.08772
```

ASSOLUTE

- $S_y = \{y_1, \dots, y_J\} \rightarrow J \leq n$
- $F = \{f_1, \dots, f_J\} \rightarrow \sum(F) = n$
- n dispari
 - $y_{0.5} = F_j$ tale che $F_j \geq (n+1)/2$
 - $y_{0.5} = F_j$ e F_i tali che $F_j \geq n/2$ e $F_i \geq n/2+1$

RELATIVE

- $Y_{0.5} = P_j \geq 0.5$

```
unique(mtcars$cyl)
```

```
## [1] 6 4 8
```

```
median(mtcars$cyl) #frequenza relativa percentuale cumulata >= 50%
```

```
## [1] 6
```



```
(cylRelCumFreq )
```

```
##          Pi
## 4   34.375
## 6   56.250
## 8  100.000
```

QUANTILI

Y = VARIABILE QUALITATIVE ORDINALE O QUANTITATIVA

LIVELLO (α) = valore percentuale rispetto al 100% N totale delle osservazioni.

$y_\alpha = y_i, Y[1, \alpha * N] < y_i < Y[\alpha * N, N]$

$\alpha \in [0; 1]$

rappresenta il valore che è preceduto da $\alpha\%$ delle osservazioni totali N e seguito da $(1-\alpha)\%$

NOTAZIONI

- **QUARTILI**

– $\alpha \in [0.25, 0.5, 0.75]$

- **DECILI**

– $\alpha \in [0.1, 0.2, \dots, 0.9]$

- **PERCENTILI**

– $\alpha \in [0.01, 0.02, \dots, 0.99]$

GREZZI

$Y = [y_1, \dots, y_n]$

- $y_\alpha = y_i$

- $i = \alpha * (n + 1)$

– **i intero** : y_i è il quantile di livello α

– **i non intero** : si considerano gli indici interi prima e dopo i

* y_α possiede i due valori identificati dai due indici interi

```
Y = mtcars$mpg
# frequenze assolute fi per ogni yi in Sy
table(Y)
```

```
## Y
## 10.4 13.3 14.3 14.7   15 15.2 15.5 15.8 16.4 17.3 17.8 18.1 18.7 19.2 19.7   21
##    2    1    1    1    1    2    1    1    1    1    1    1    1    2    1    2
## 21.4 21.5 22.8 24.4   26 27.3 30.4 32.4 33.9
##    2    1    2    1    1    1    2    1    1
```

```
# frequenze relative pi in Sy
table(Y)/length(Y)
```

```
## Y
## 10.4 13.3 14.3 14.7   15 15.2 15.5 15.8 16.4 17.3
## 0.0625 0.03125 0.03125 0.03125 0.03125 0.06250 0.03125 0.03125 0.03125 0.03125 0.03125
```

```
##      17.8      18.1      18.7      19.2      19.7      21      21.4      21.5      22.8      24.4
## 0.03125 0.03125 0.03125 0.06250 0.03125 0.06250 0.06250 0.03125 0.06250 0.03125
##      26      27.3      30.4      32.4      33.9
## 0.03125 0.03125 0.06250 0.03125 0.03125
```

```
# frequenze relative cumulate
relCumFreq = data.frame(y = sort(unique(Y)), P = cumsum(table(Y)/length(Y)))
# QUARTILI
quartili = data.frame(q = quantile(Y, probs = c(0.25, 0.5, 0.75)))

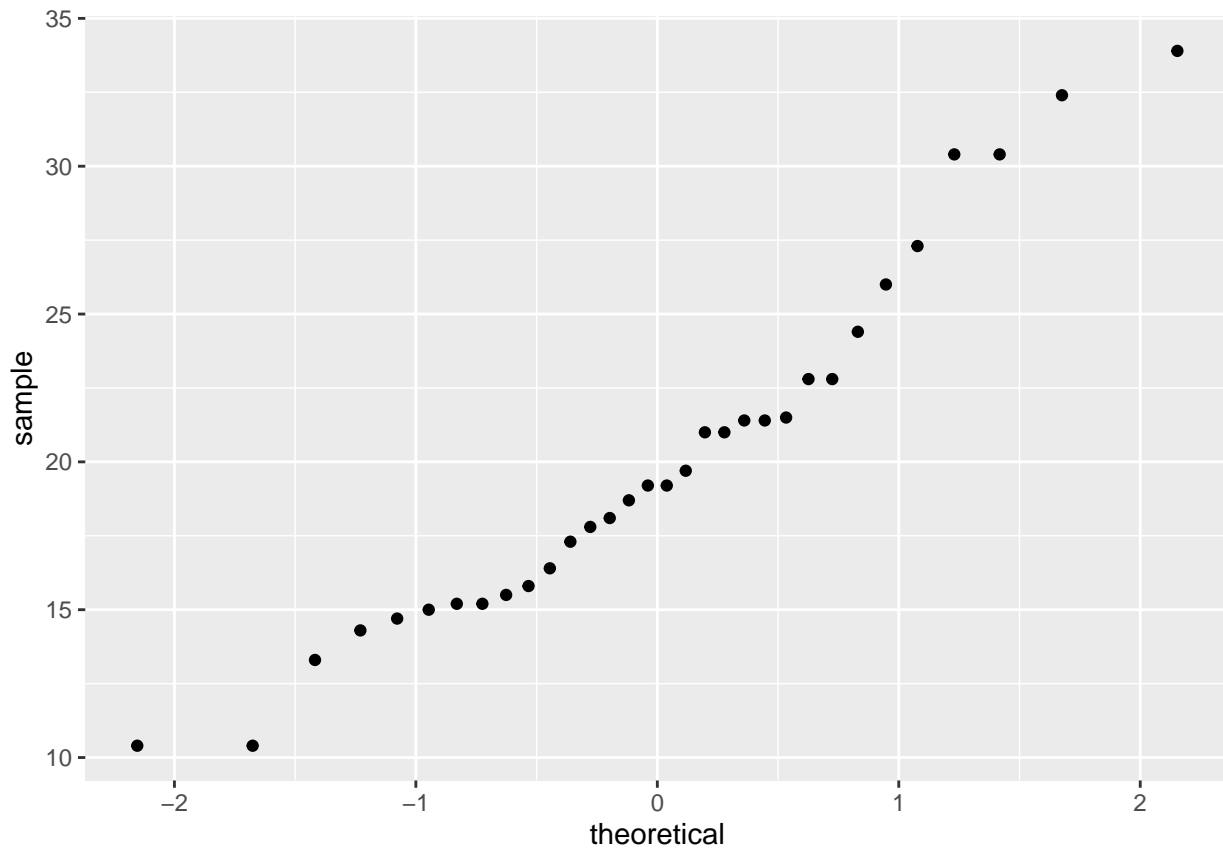
median(Y)
```

```
## [1] 19.2
```

```
# ==
quantile(Y, 0.5)
```

```
## 50%
## 19.2
```

```
ggplot(mtcars, aes(sample = mpg)) + stat_qq()
```

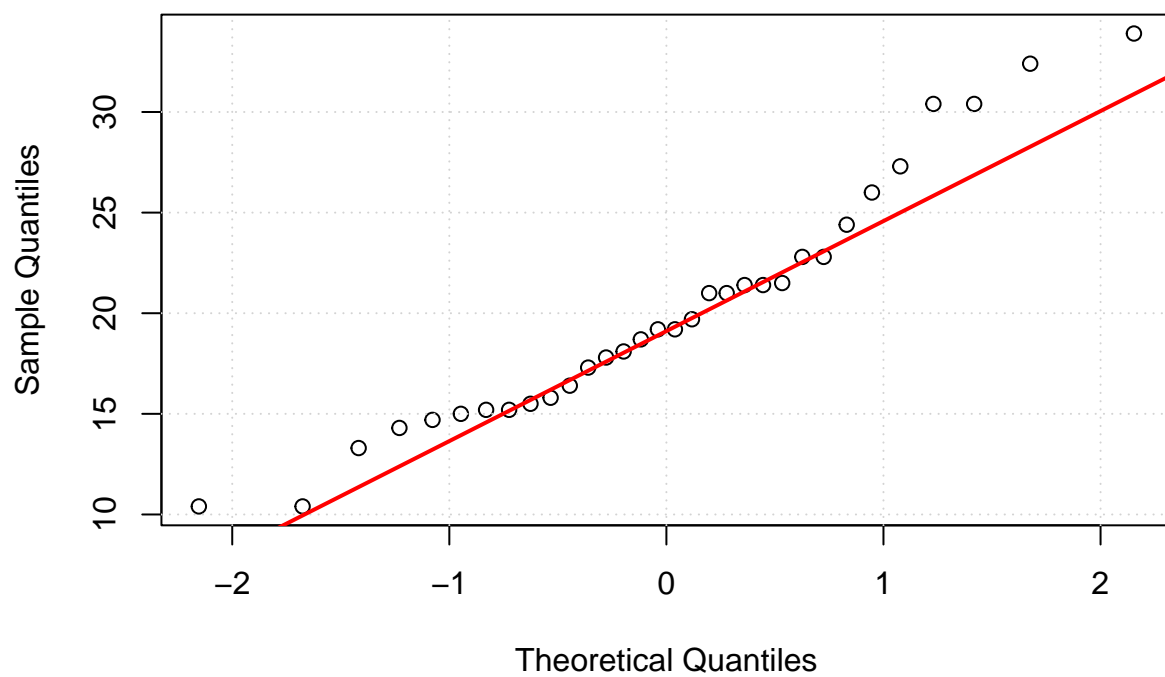


```
# ggplot(relCumFreq, aes(P, y)) + geom_point() + geom_segment(aes(xend = P, yend = y)) + geom_line(aes(P, y))
```

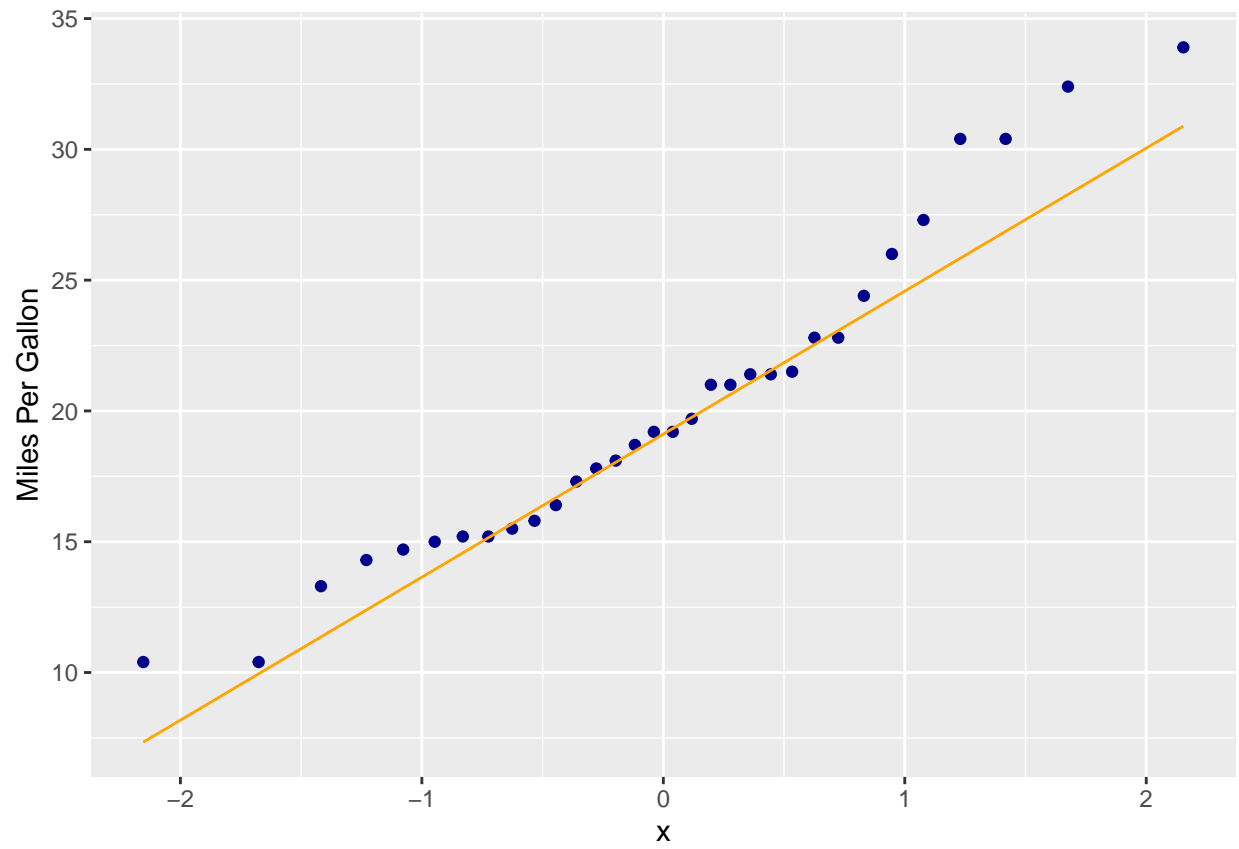
```
qqnorm(Y)
grid() # griglia
qqline(Y,
      lwd = 2, # spessore
```

```
col = "red"    # colore  
)
```

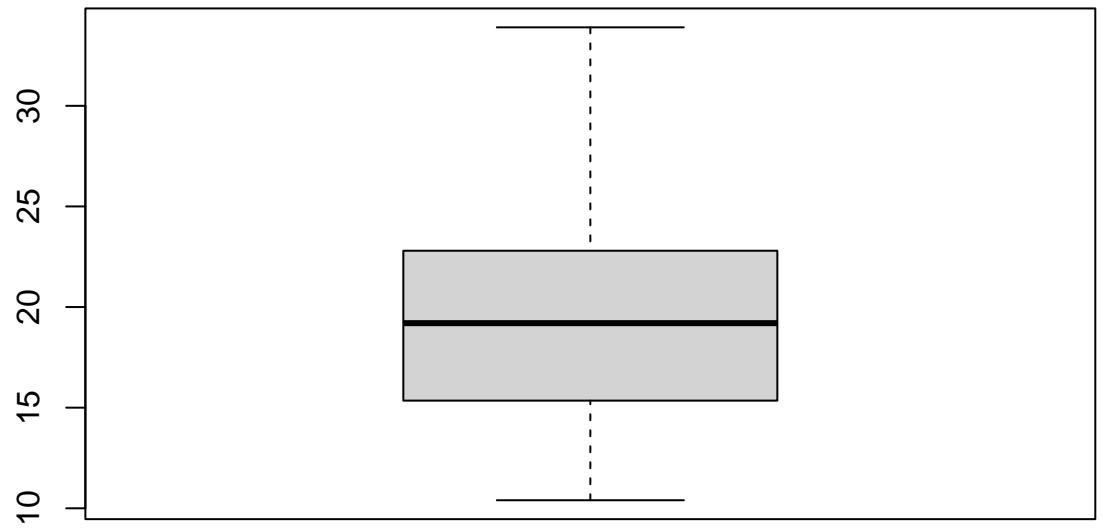
Normal Q-Q Plot



```
ggplot(data = mtcars, aes(sample = mpg)) +  
  geom_qq(color = "dark blue") +  
  geom_qq_line(color = "orange") +  
  labs(y = "Miles Per Gallon")
```

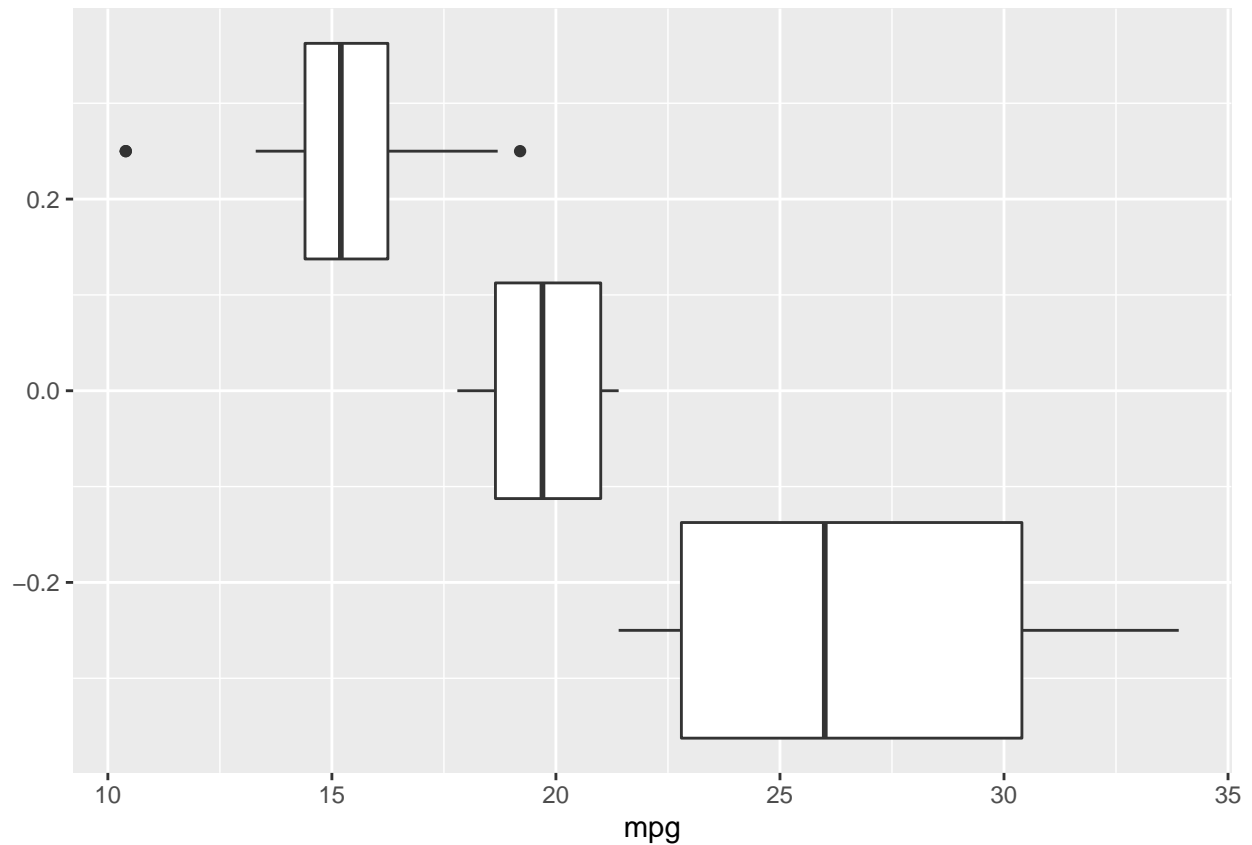


```
boxplot(mtcars$mpg)
```



BOXPLOT

```
# BOXPLOT SEPARATI PER OGNI CATEGORIA DI CILINDRI  
ggplot(mtcars)+ geom_boxplot(aes(mpg,group=cyl))
```



MODA

Y_{mo} = valore più occorrente all'interno di Y

```
Y = mtcars$cyl
```

```
cylFreqTable
```

```
## # A tibble: 3 x 2
##   cyl frequency
##   <dbl>     <int>
## 1     4         11
## 2     6          7
## 3     8         14
```

```
Ymo = (cylFreqTable %>% filter(frequency== max(cylFreqTable$frequency)))$cyl
# == cylFreqTable[cylFreqTable$frequency==max(cylFreqTable$frequency), "cyl"]
```

VARIABILITÀ

È un indice in grado di quantificare la variabilità della variabile osservata

Se Y è una variabile statistica **degenere** il suo supporto è composto da un unico elemento

$S_y = \{y_1\}$

La sua variabilità perciò dovrà essere nulla $V_y = 0$

CAMPO DI VARIAZIONE

Data Y variabile statistica quantitativa essa possiede un Range di valori ammessi.

È sensibile alla presenza di eventuali valori anomali, troppo alti o bassi

- $Sy = \{y_1, \dots, y_j\} \ j \leq N \rightarrow N = |Y|$
- $y_1 < y_2 \dots < y_j$
- $Range(Y) = Ry = y_j - y_1 = \max(Sy) - \min(Sy) > 0$
 - $Sy = \{y_1\} \rightarrow \min(Y) = \max(Y)$
 - $Ry = \max(Y) - \min(Y) = 0$

SCARTO INTERQUANTILICO

Data Y quantitativa

$SI_y = y_{0.75} - y_{0.25}$

Si tratta della dimensione della scatola nel grafico boxplot

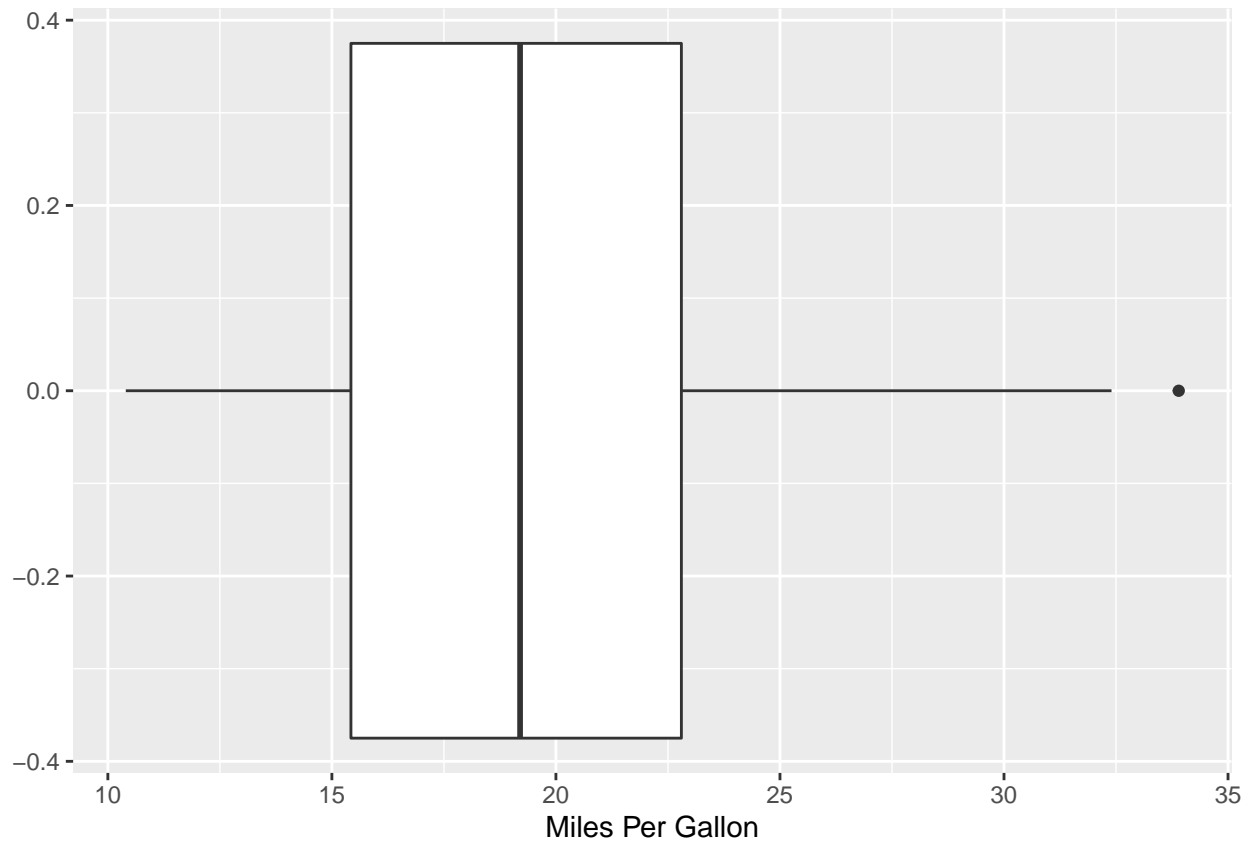
```
Y = mtcars$mpg
quantile(Y,0.75)
```

```
## 75%
## 22.8
```

```
quantile(Y,0.25)
```

```
## 25%
## 15.425
```

```
ggplot(mtcars) + geom_boxplot(aes(x=mpg)) + labs(x="Miles Per Gallon")
```



VARIANZA

Data Y quantitativa con media aritmetica $E(Y)$

$$V(Y) = \sigma_y^2 = \sigma^2$$

$$V(Y) = E[(Y - E(Y))^2]$$

$$V(Y) = E(Y^2) - (E(Y))^2$$

$ScY = Y - E(Y) \rightarrow$ Variabile Scarto

Possiede unità di misura al quadrato rispetto alla variabile originale Y

$$\text{\$U.D.M } V(Y) = [\text{\$U.D.M}(Y)]^2$$

$$\sigma_y = V(Y)^{0.5}$$

TIPI DI DATI

GREZZI

- $Y = \{y_1, \dots, y_n\}$ $n = |Y|$
- $E(Y)$ = media aritmetica di Y

$$V(Y) = \frac{1}{n} * \sum_{i=1}^n (y_i - E(Y))^2$$

FREQUENZE

$$V(Y) = \frac{1}{n} * \sum_{j=1}^J (y_j - E(Y))^2 * f_j = \sum_{j=1}^J (y_j - E(Y))^2 * p_j$$

CLASSI

$$V(Y) = \frac{1}{n} \sum_{j=1}^J (Y_{cj} - E(Y))^2 * f_j = \sum_{j=1}^J (Y_{cj} - E(Y))^2 * p_j$$

PROPRIETÀ

NON NEGATIVITÀ

$$V(Y) \geq 0$$

$$V(Y) = 0 \iff S_Y = \{y_1\}$$

CALCOLO

\$\$

$$V(Y) = E(Y^2) - (E(Y))^2$$

\$\$

$$V(Y) = E[(Y - E(Y))^2] = E[Y^2 + (E(Y))^2 - 2Y * E(Y)] = E(Y^2) + (E(Y))^2 - 2E(Y)E(Y) = E(Y^2) - (E(Y))^2$$

INVARIANZA PER TRASLAZIONI

$$V(Y + b) = V(Y); b \in R$$

$$V(Y + b) = E[(Y + b - E(Y + b))^2] = E[(Y + b - E(Y) - b)^2] = E[(Y - E(Y))^2] = V(Y)$$

OMOGENEITÀ DI SECONDO GRADO

$$V(a * Y) = a^2 V(Y), a \in R$$

$$V(aY) = E[(aY - E(aY))^2] = E[(aY - aE(Y))^2] = E[a^2(Y - E(Y))^2] = a^2 E[(Y - E(Y))^2] = a^2 V(Y)$$

- $E(Y)=0 \rightarrow V(Y) = E(Y^2)$
- $V(aY + b) = a^2 V(Y)$

COEFFICIENTE DI VARIAZIONE

- Si tratta di un indice **adimensionale** che misura la variabilità dei dati tenendo conto dell'ordine di grandezza del fenomeno.
- Si tratta di un numero puro, perciò permette il confronto con altri dati di categoria differente.

$$CV_y = \frac{\sigma_y}{|E(Y)|}$$

- Data Y con $E(Y)=0$ e $V(Y)=1$ si dice **standardizzata**

- Trasformazione lineare delle modalità osservate
 - Data Y variabile generica $\{y_1, \dots, y_n\}$
 - * $Z = \frac{(Y - E(Y))}{\sigma_y}$ -> nuova Variabile Z standardizzata
 - * $z_i = \frac{(y_i - \mu_y)}{\sigma_y}, i \in [1, n]$
 - Data Z variabile standardizzata $\{z_1, \dots, z_n\}$
 - * $Y = \sigma Z + \mu$
 - $\mu = E(Y)$
 - $\sigma^2 = V(Y)$

SIMMETRIA

Una distribuzione di frequenza si dice simmetrica se il suo grafico a istogramma o diagramma a bastoncini è simmetrico

Il grafico perciò è divisibile a metà uguali tramite un asse verticale identificato dal valore della mediana=**y0.5**

INDICE DI SIMMETRIA

$$\gamma_y = \frac{E[(Y - E(Y))^3]}{\sigma^3_y}$$

- $\sigma_y = V(Y)^{0.5}$ -> $\sigma^2_y = V(Y)$
- $\gamma_y = 0$ -> **SIMMETRIA**
- $\gamma_y < 0$ -> **ASIMMETRIA SX**
- $\gamma_y > 0$ -> **ASIMMETRIA DX**

CURTOSI

Rappresenta l'andamento delle frequenze nei valori più estremi del supporto

INDICE DI CURTOSI

$$\beta_y = \frac{E[(Y - E(Y))^4]}{\sigma^4_y}$$

PLATICURTICA (IPONORMALE)

- CODE LEGGERE
- $\beta_y < 3$

LEPTOCURTICA (IPERNORMALE)

- CODE PENSANTI
- $\beta_y > 3$

NORMOCURTICA

- $\beta_y = 3$