LEZIONE_2

2022-10-11

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**
  - **Sy = {y1, ... , yj} con j<=N**
    - per qualsiasi j != i –> yj != yi
  - **N = numero di unità statistiche considerate**

TIPOLOGIE

- **VARIABILI QUALITATIVE (CATEGORIALI) –> 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
  - [yo,y1) –> y1 escluso
  - [y1,y2) –> y2 escluso

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

```
## [1] 25
```

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

```
## [1] 23.5
```

```r
(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
```

```r
(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
```

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

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

FREQUENZE RELATIVE

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

```r
# p1 = f1 / sum([f1,f2,...,fj]) j<=N --> N = n° unità
#    = f1 / n
```

```r
(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
```

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

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

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

FREQUENZE CUMULATE

- **VARIABILI ORDINABILI, QUALITATIVE E QUANTITATIVE**
- Fi = Frequenza assoluta con cui si presentano modalità con ordini <= i-esimo ordine
  - Dato F = {f1,...,fj} j = | Sy | <= N n° unità
  - F1 = f1 –> Fj = N
- Pi = Frequenza relativa cumulata, analoga a Fi
  - Dato P = {p1, ... , pj} –> | P | = | F |
  - IPOTESI: P1 = p1 –> Pj = N

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

```
##  4  6  8
## 11 18 32
```

```
ll = nrow(cylFreqTable) # = nrow(cylFreqTable)
if (   (cylCumFreq[[1]] == cylFreqTable[1,"frequency"]) &&
       (cylCumFreq[[ll]] == 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
```

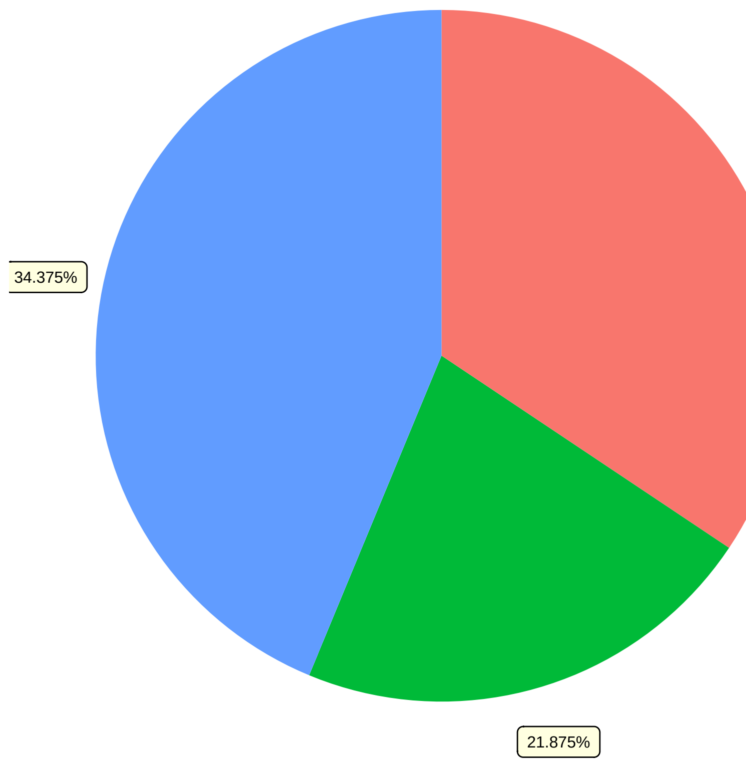
```
##        Pi
## 4  34.375
## 6  56.250
## 8 100.000
```

RAPPRESENTAZIONI GRAFICHE

DATI QUALITATIVI

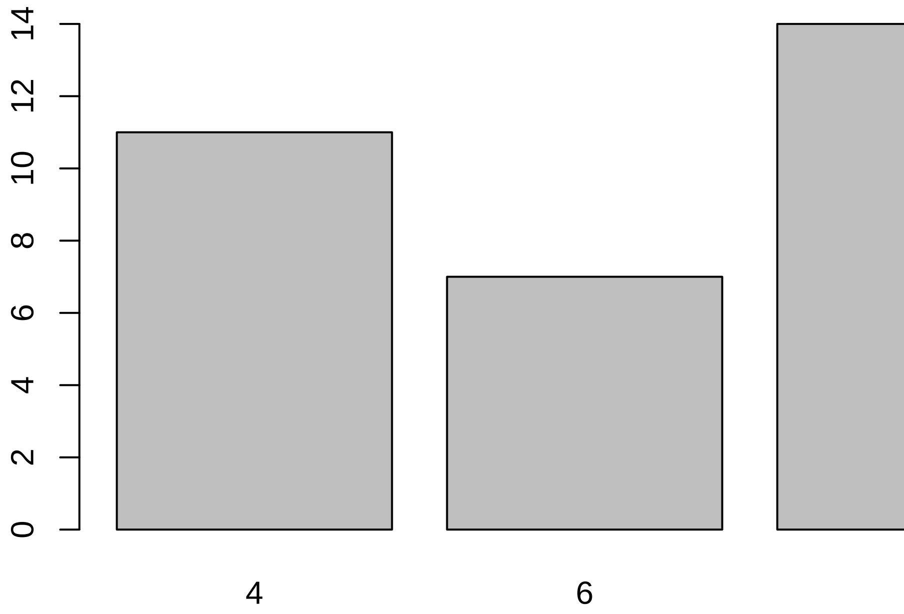
```
tab1 <- as.data.frame(table(mtcars$cyl))
colnames(tab1) = c("CYL","FREQ")
```

```r
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)
```



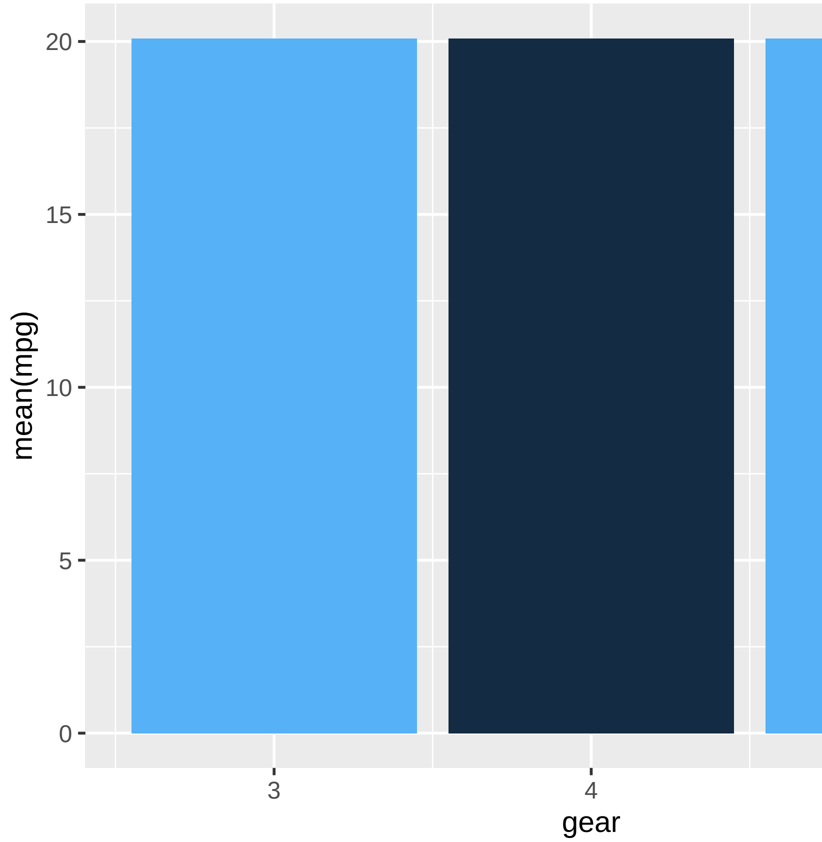
DIAGRAMMI CIRCOLARI(A TORTA)

```r
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

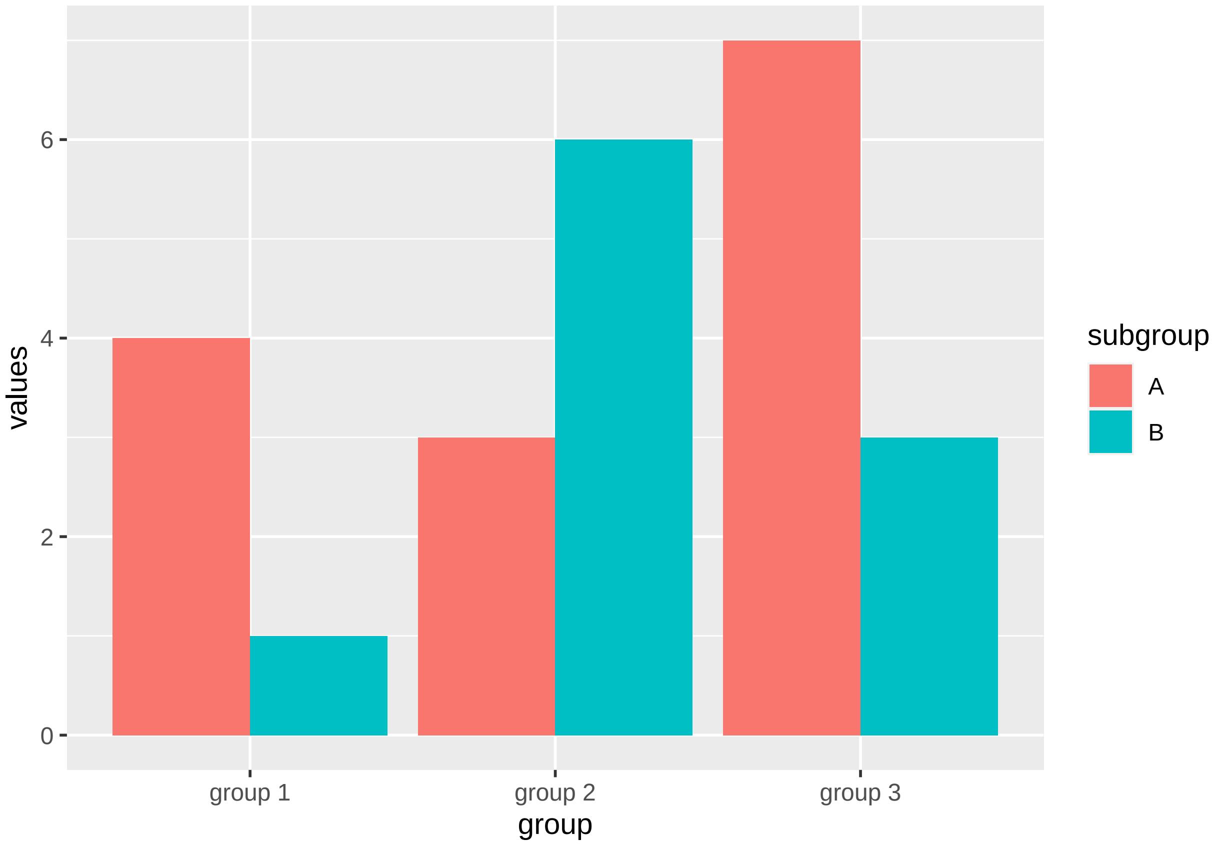
DATI QUANTITATIVI

```
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
```

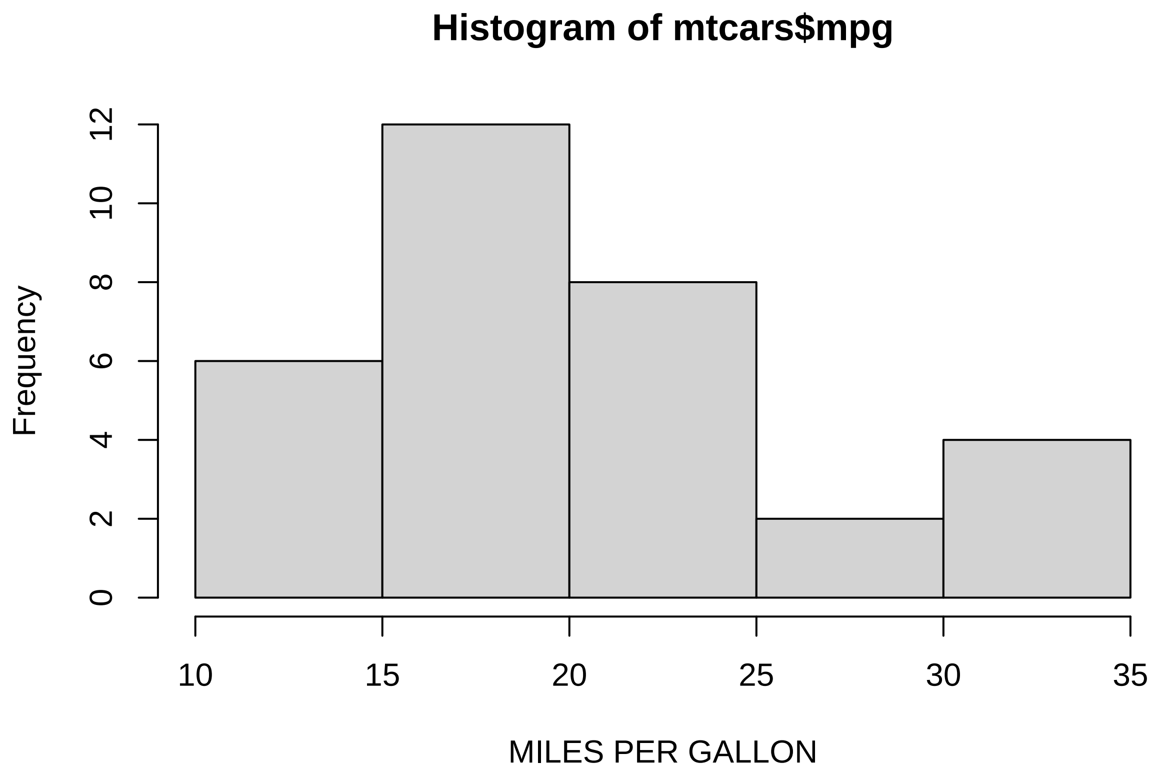
```
ggplot(data,                                      # Grouped barplot using ggplot2
       aes(x = group,
           y = values,
           fill = subgroup)) +
  geom_bar(stat = "identity",
           position = "dodge")
```



ISTOGRAMMI

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

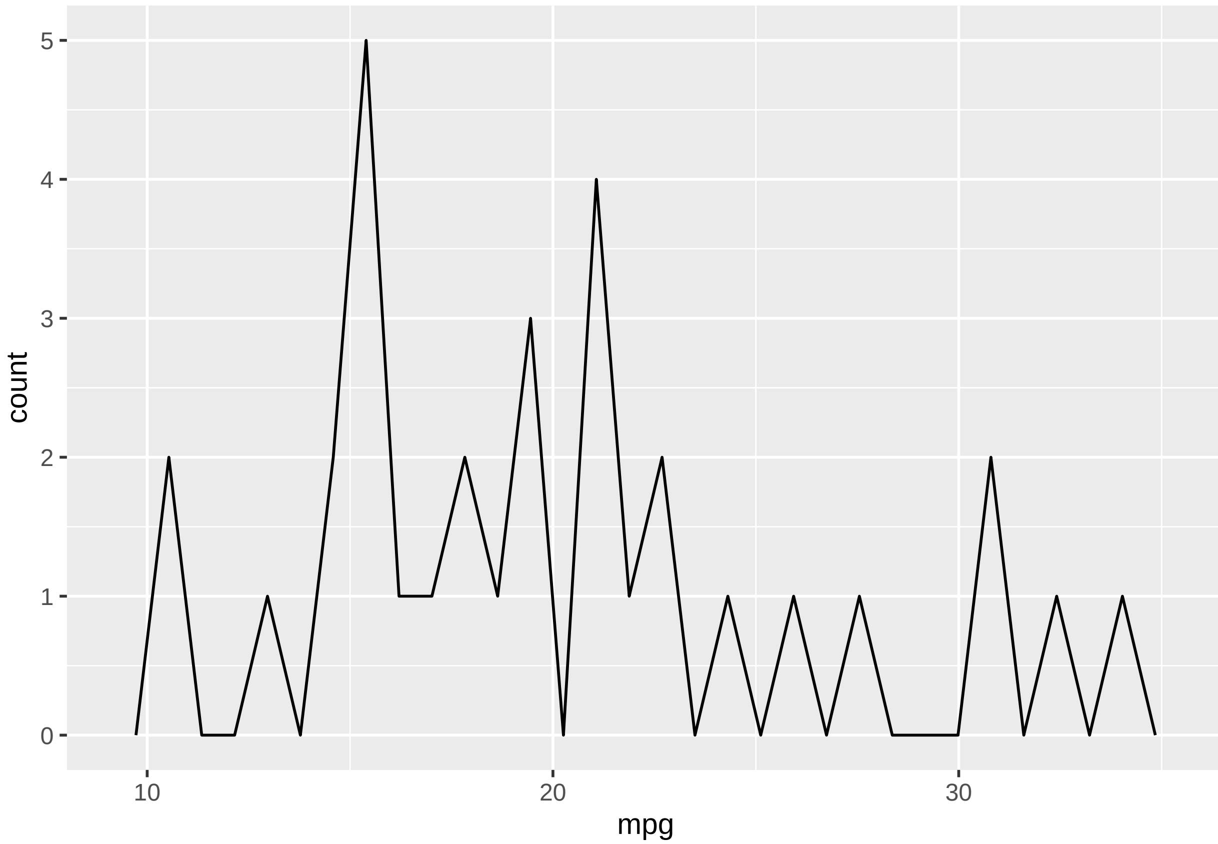
```
# hist(x = dati$nominees,xlab = "Nominees",main = "Nominees Frequency")
hist(mtcars$mpg,xlab="MILES PER GALLON")
```



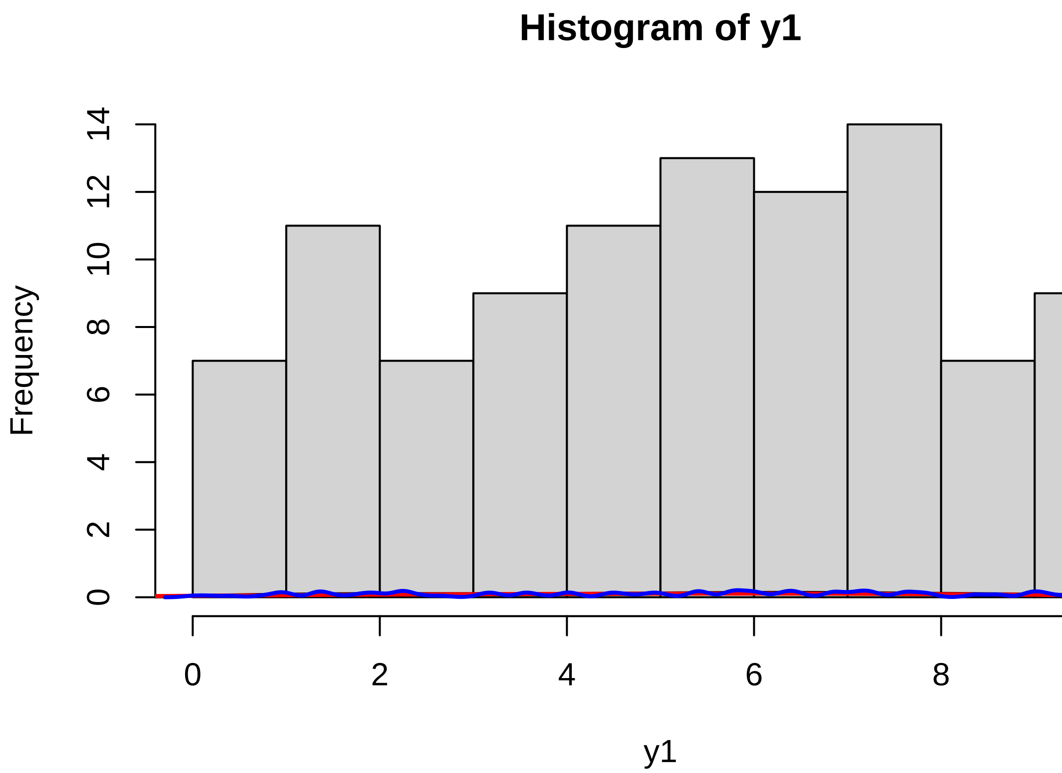

```
ggplot(mtcars) + geom_freqpoly(aes(x=mpg))
```

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
```

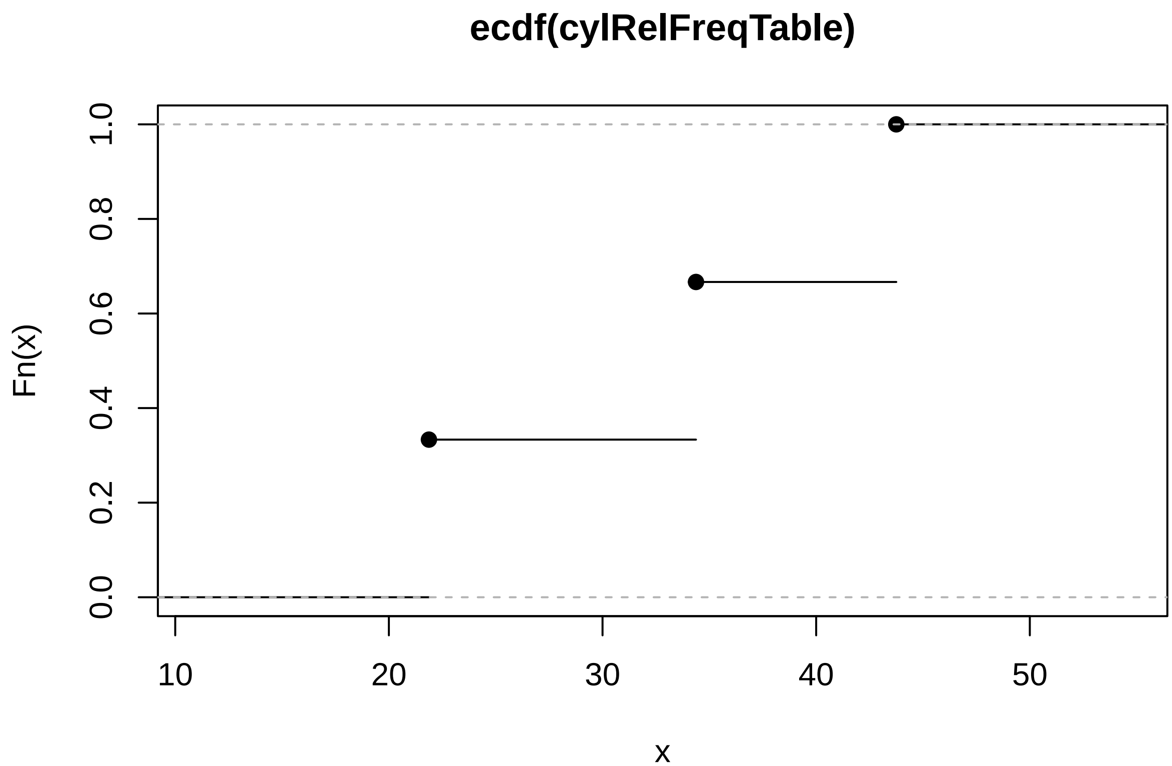


STIMA DELLA DENSITÀ

FUNZIONI DI RIPARTIZIONE EMPIRICA

- Rappresenta graficamente l'andamento delle frequenza cumulate

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



-<u>BOXPLOT</u>

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 = VARIABILE –> E(Y) = MEDIA ARITMETICA DI TUTTI I VALORI DI Y

- Dato Y = {y1, .. , yn} n = totale unità statistiche
- E(Y) = (y1 + .. + yn) / n –> se si dispone dei dati grezzi
- E(Y) = 1/n * $\sum_{i=1}^{J} fi * yi$ –> J = | Sy |
  - FREQUENZE ASSOLUTE
- E(Y) = 1/n * $\sum_{i=1}^{J} pi * yi$ –> J = | Sy |
  - FREQUENZE RELATIVE
- E(Y) = 1/n * $\sum_{i=1}^{J} yci * yi$ –> J = | Sy |

  - yci = valore centrale dell'intervallo [yi-1;yi)
  - 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 assoulute 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[1,2]
# }
```

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

PROPRIETÀ DI CAUCHY:

- Dato il supporto Sy = {y1, .. , yj} con yi < yi+1 per qualsiasi i
- IPOTESI y1 <= E(Y) <= yj
- y1 <= yi <= yj con $i \in [1; J]$
- y1*pi <= yi*pi <= yj*pi con $i \in [1; J]$
- $\sum_{k=1}^{J} y1 * pk$ <= $\sum_{k=1}^{J} yk * pk$ <= $\sum_{k=1}^{J} yJ * pk$
- y1* $\sum_{k=1}^{J} pk$ <= $\sum_{k=1}^{J} yk * pk$ <= yJ* $\sum_{k=1}^{J} pk$
  - E(Y) = $\sum_{k=1}^{J} yk * pk$
  - $\sum_{k=1}^{J} pk$ = 1 –> la somma di tutte le frequenze relative da il 100%
- y1 <= $\sum_{k=1}^{J} yk * pk$ <= yJ
- y1 <= E(Y) <= yJ –> IPOTESI CONFERMATA

PROPRIETÀ DEL BARICENTRO

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

    – 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 * $\sum_{i=1}^{n}(a * yi + b)$ =
- 1/n * [$\sum_{i=1}^{n} a * yi$ + $\sum_{i=1}^{n} b$]=
- a * 1/n * $\sum_{i=1}^{n} yi$ + 1/n *n * b =
- a* E(Y) + b –> 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
- y0.5 = MEDIANA –> quantile di livello 0.5

GREZZI

- y0.5 divide a metà Y in modo che i valori precedenti siano <= e successivi <= y0.5
- se n è dispari allora y0.5 si trova nell'indice i = (n+1)/2
- se n è pari allora y0.5 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
```

```r
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

- Sy = {y1, … , yJ} -> J <= n
- F = {f1, … , fJ} -> sum(F) = n
- n dispari
  - y0.5 = Fj tale che Fj >= (n+1)/2
  - y0.5 = Fj e Fi tali che Fj >=n/2 e Fi >=n/2+1

RELATIVE

- Y0.5 = Pj >= 0.5

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

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

```r
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.

$y\alpha = yi, Y[1, \alpha * N] < yi < Y[\alpha * N, N]$

$\alpha \in [0; 1]$

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

NOTAZIONI

- **QUARTILI**
  - $\alpha \in [0.25, 0.5, 0.75]$
- **DECILI**
  - $\alpha \in [0.1, 0.2, .., 0.9]$
- **PERCENTILI**
  - $\alpha \in [0.01, 0.02, .., 0.99]$

GREZZI

$Y = [y1, ..., yn]$

- $y\alpha = yi$
- $i = \alpha * (n + 1)$
  - **i intero** : yi è il quantile di livello α
  - **i non intero** : si considerano gli indici interi prima e dopo i
    - yα 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
## 0.06250 0.03125 0.03125 0.03125 0.03125 0.06250 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
```

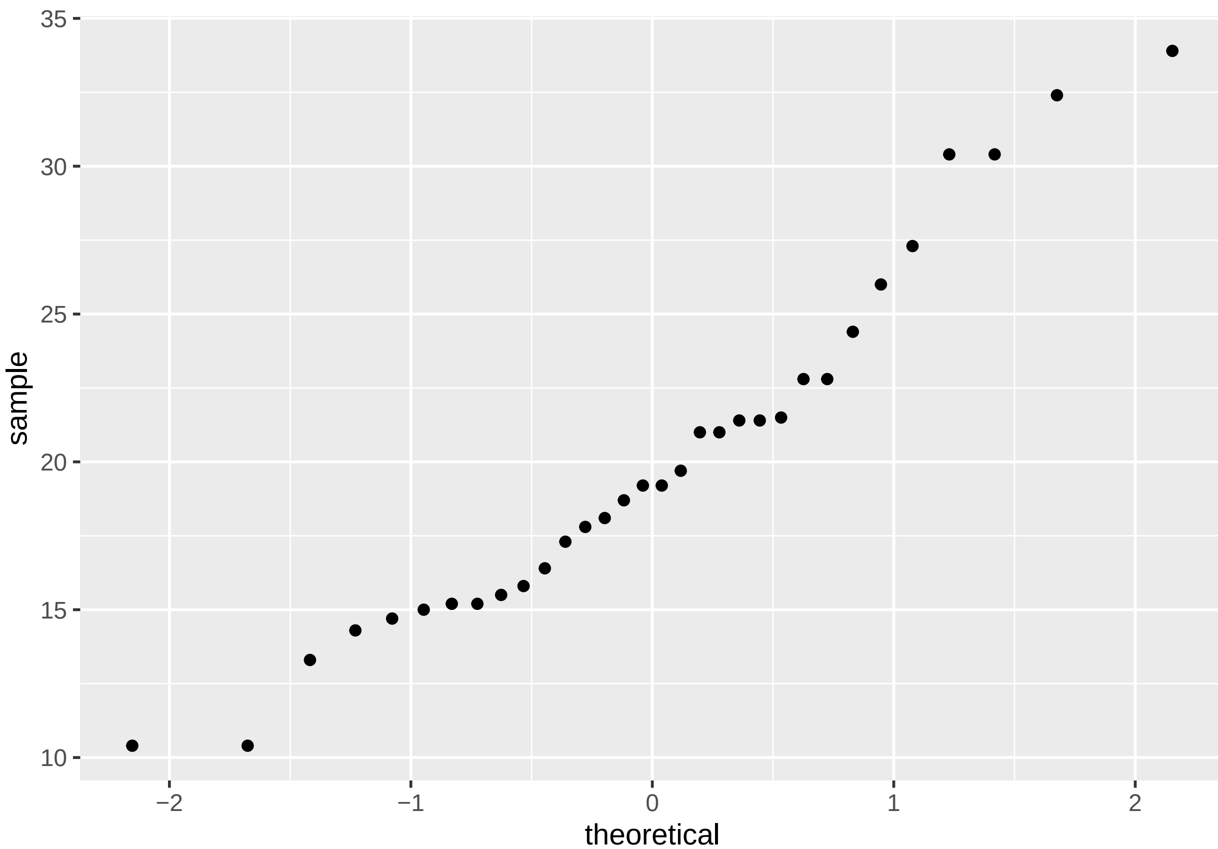
```
# frequenze relative cumulate
relCumFreq = data.frame(y = sort(unique(Y)),P = cumsum(table(Y)/length(Y)))
# QUARTILI
quartili = data.frame(q = quantile(Y,probs = c(0.25,0.5,0.75)))

median(Y)
```

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

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

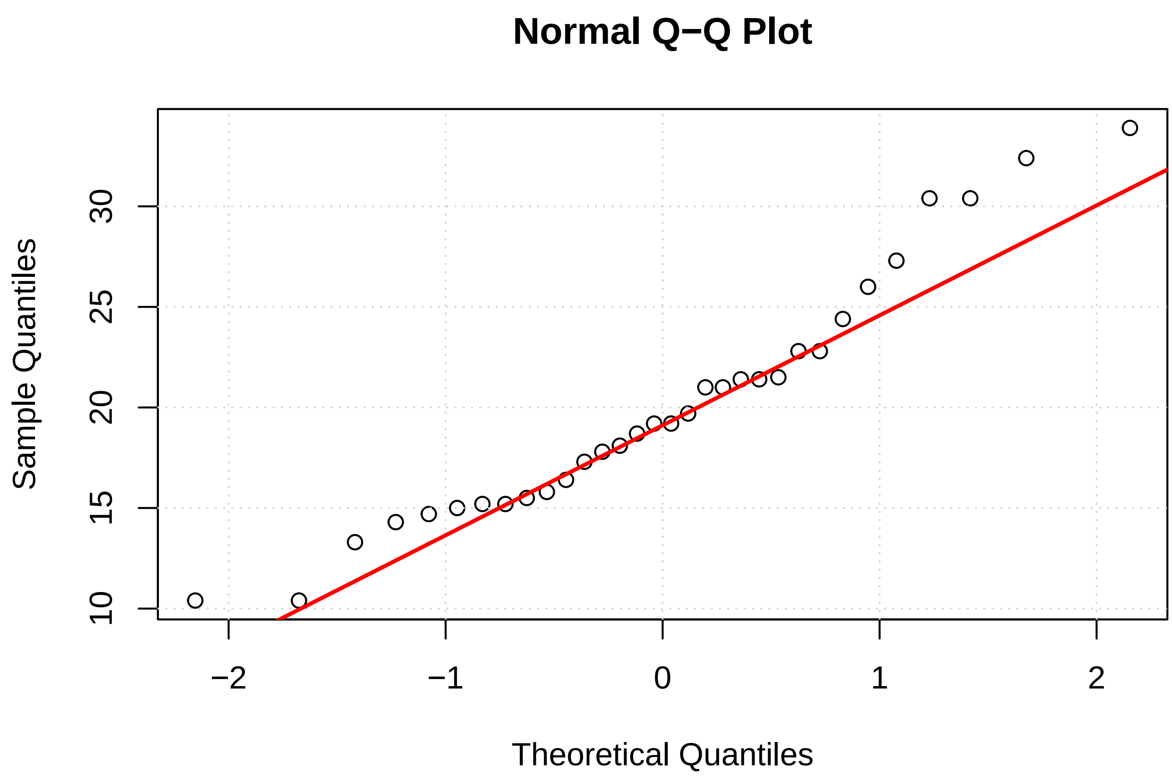
```
##  50%
## 19.2
```

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

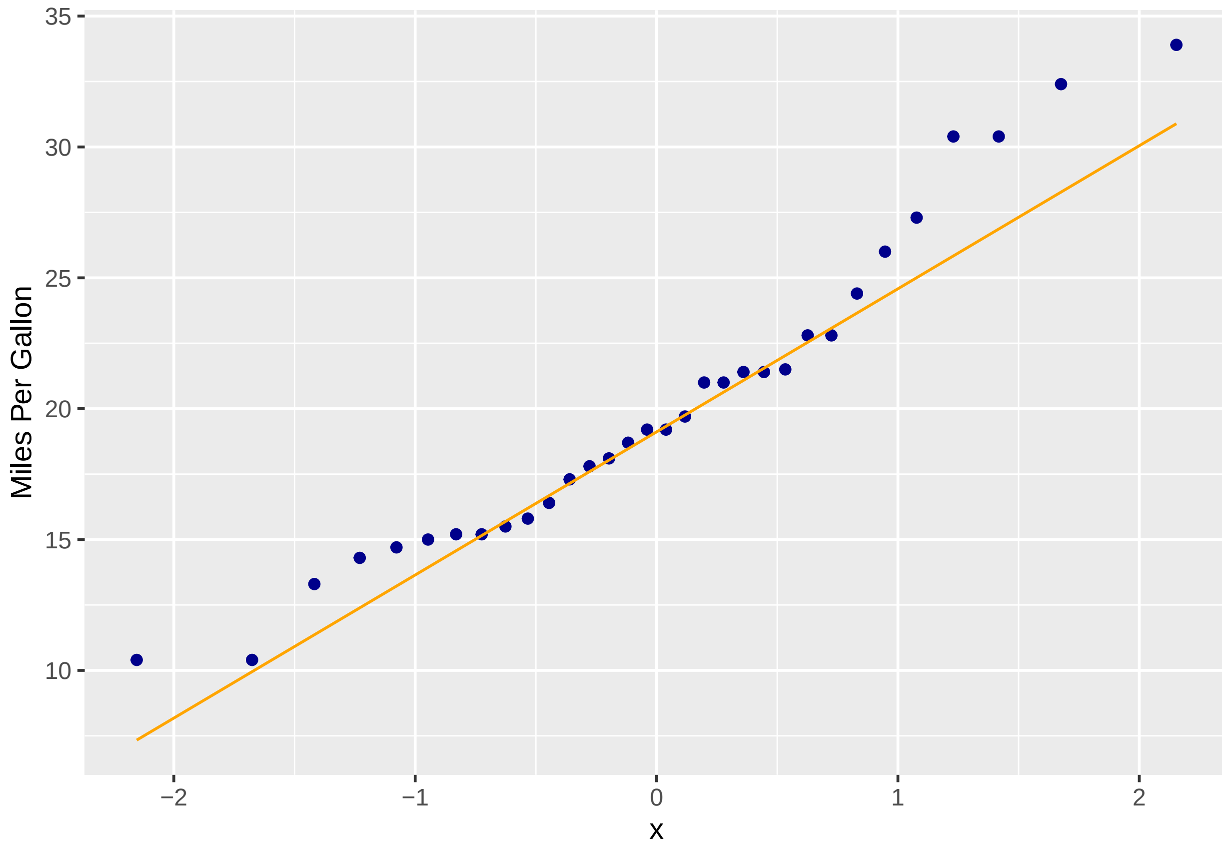



```
# ggplot(relCumFreq,aes(P, y)) +  geom_point() + geom_segment(aes(xend = P, yend = y)) + geom_line(aes(


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

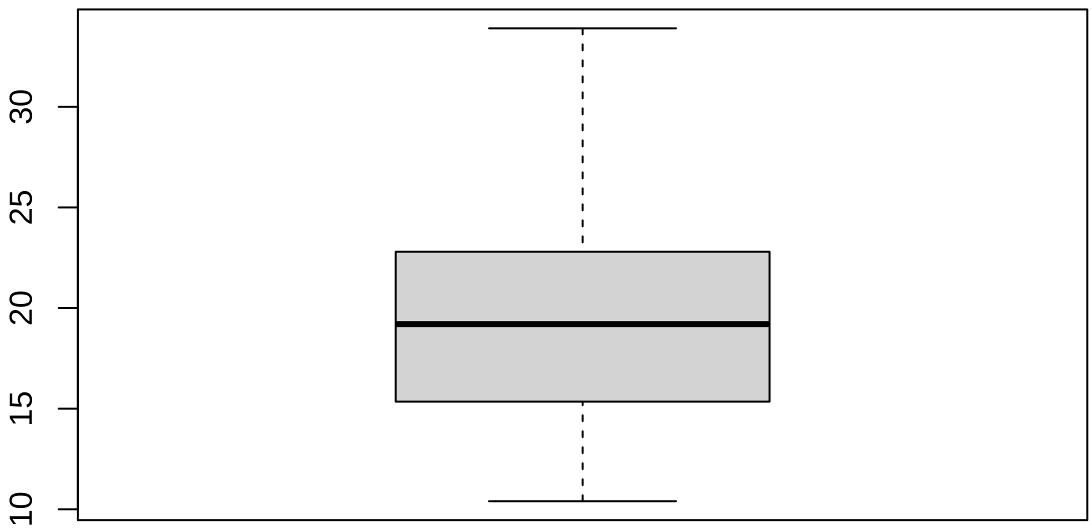
```
       col = "red"    # colore
       )
```




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

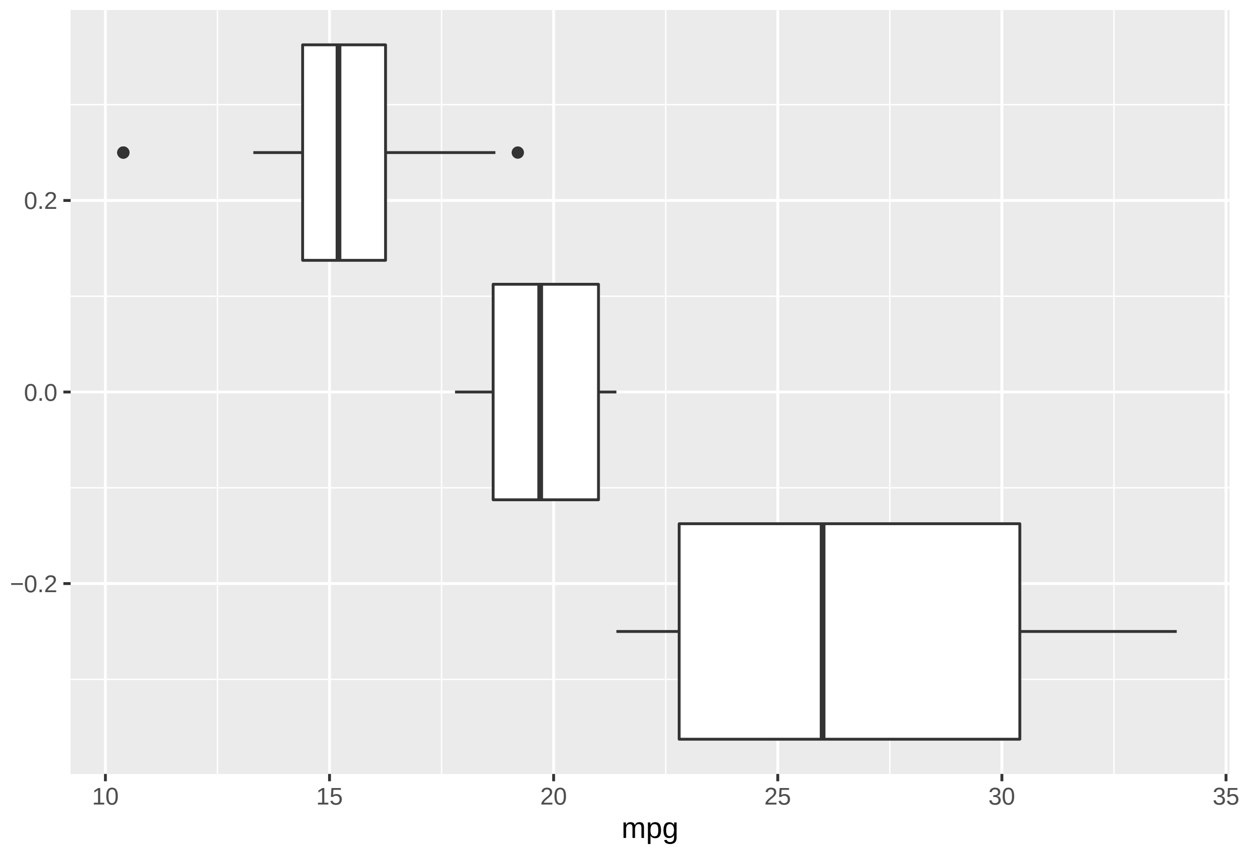



```
boxplot(mtcars$mpg)
```



BOXPLOT

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



MODA

Ymo = 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 **Sy={y1}**

La sua variabilità periciò dovrà essere nulla Vy=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={y1, … ,yj} j<=N –> N = | Y |
- y1 < y2 … < yj
- Range(Y) = Ry = yj - y1 = max(Sy) - min(Sy) > 0
  - Sy={y1} -> min(Y) = max(Y)
  - Ry = max(Y)-min(Y)=0

SCARTO INTERQUANTILICO

Data Y quantitativa

SIy = y0.75 - y0.25

Si tratta della dimensione della scatola nel grafico boxplot

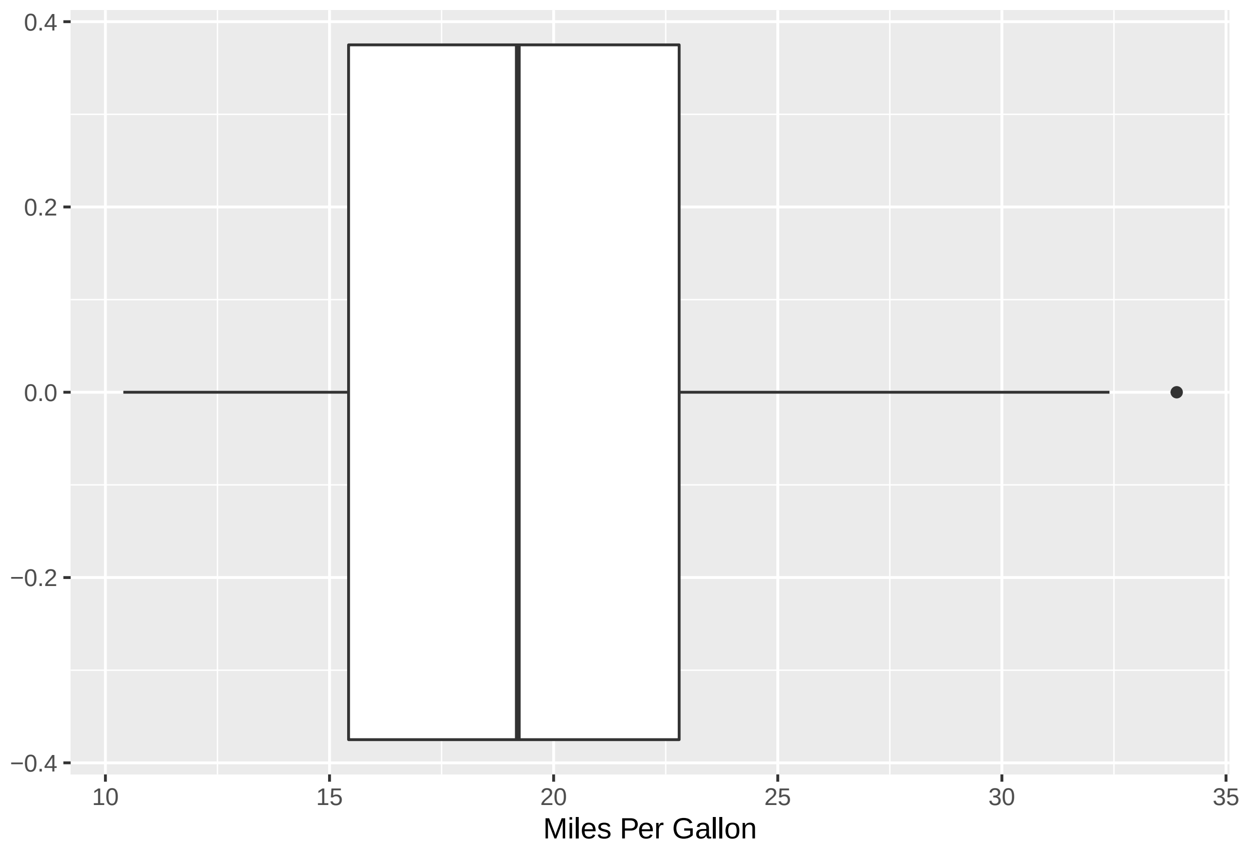
```
Y = mtcars$mpg
quantile(Y,0.75)
```

```
##  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)

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

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

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

$ScY = Y - E(Y)$ −> Variabile Scarto

Possiede unità di misura al quadrato rispetto alla variabile originale Y

$U.D.M V(Y) = [U.D.M(Y)]^2 $

$$\sigma y = V(Y)^{0.5}$$

TIPI DI DATI

GREZZI

- Y = {y1, . . . , yn} n = | Y |
- E(Y) = media aritmetica di Y

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

FREQUENZE

$$V(Y) = \frac{1}{n} * \sum_{j=1}^{J} (yj - E(Y))^2 * fi = \sum_{j=1}^{J} (yj - E(Y))^2 * pi$$

CLASSI

$$V(Y) = \frac{1}{n} \sum_{j=1}^{J} (Ycj - E(Y))^2 * fi = \sum_{j=1}^{J} (Ycj - E(Y))^2 * pi$$

PROPRIETÀ

NON NEGATIVITÀ

V(Y) >= 0

V(Y) = 0 <—> Sy = {y1}

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 –>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 grandeza del fenomeno.
- Si tratta di un numero puro, perciò permette il confronto con altri dati di categoria differente.

$$CVy = \frac{\sigma y}{|E(Y)|}$$

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

- Trasformazione lineare delle modalità osservate
  - Data Y variabile generica {y1, ... , yn}
    - $Z = \frac{(Y-E(Y))}{\sigma y}$ –> nuova Variabile Z standardizzata
    - $zi = \frac{(yi-\mu y)}{\sigma y}, i \in [1, n]$
  - Data Z variabile standardizzata {z1, ... , zn}
    - $Y = \sigma Z + \mu$
      - $\mu = E(Y)$
      - $\sigma^2 = V(Y)$

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

- $\beta y = 3$