Progetto_SAD

Livio Vona

2023-02-14

Contents

1 INTRODUZIONE

1.1 SET DI DATI

Il set di dati è stato reperito dal sito dell'istat, ed è relativo all'indagine statistica dell'anno 2020 atta ad analizzare per ogni regione il tipo di istruzione della popolazione tra i 25-64 anni.

I dati sono stati esportati da un file excel e memorizzati in un data frame, un oggetto simile alla matrice, costituito da righe e colonne. Le colonne rappresentano le variabili e le righe le osservazioni. Ogni singola colonna deve avere elementi dello stesso tipo, mentre colonne differenti possono avere elementi differenti (a differenza delle matrici).

In particolare per ogni regione è stato indagato un campione di persone tra i 25 e i 64 anni, e quindi per ciascuna riga (regione) i valori contenuti nelle celle rappresentano le percentuali di invidui indagati della regione per ciasun titolo di studio.

```
##                                LSE/NT  LSM DIP2-3 DIP4-5 L/PL
## Piemonte                          3.4 33.2    9.8   33.8 19.8
## Valle d'Aosta / Vallée d'Aoste    3.5 35.1    7.7   34.8 18.9
## Liguria                           2.6 28.6    6.7   40.2 21.8
## Lombardia                         2.9 32.0   10.4   33.0 21.7
## Trentino Alto Adige / Südtirol    2.2 30.3   18.3   29.3 20.0
## Veneto                            2.7 31.9   11.7   34.0 19.7
## Friuli-Venezia Giulia             2.3 26.4   12.0   37.9 21.4
## Emilia-Romagna                    2.7 28.8    8.8   36.4 23.3
## Toscana                           3.9 31.6    5.5   37.9 21.1
## Umbria                            3.0 25.4    6.2   42.3 23.1
## Marche                            3.5 31.2    6.5   37.7 21.2
## Lazio                             3.2 26.2    2.9   40.7 27.0
## Abruzzo                           3.8 28.4    3.9   42.7 21.2
## Molise                            4.1 33.9    3.2   40.3 18.5
## Campania                         10.0 36.1    3.2   34.5 16.3
## Puglia                            9.5 38.9    2.8   33.4 15.3
## Basilicata                        5.6 30.7    4.0   42.2 17.6
## Calabria                         10.5 34.7    2.1   36.8 15.9
## Sicilia                           9.2 38.2    2.1   35.6 14.9
## Sardegna                          5.6 41.4    2.4   32.9 17.7
```

1.2 IDENTIFICATIVI

- LSE/NT: licenza di scuola elementare, nessun titolo di studio;
- LSM: licenza di scuola media;
- DIP 2-3: diploma 2-3 anni (qualifica professionale);
- DIP 4-5: diploma 4-5 anni (maturità);
- L/PL: laurea e post-laurea.

1.3 STATISTICA DESCRITTIVA

La statistica descrittiva è utilizzata per analizzare il comportamento dei fenomeni oggetto di studio. Ogni fenomeno può essere descritto tramite opportune categorie di dati di tipo qualitativo oppure di tipo quantitativo discreti o continui. I dati sono utilizzati per ricavare misure di sintesi che consentano di comprendere il comportamento del fenomeno in esame. Sulla base dell'analisi dei dati è spesso possibile formulare opportune ipotesi statistiche da sottoporre successivamente a procedimenti di verifica mediante gli strumenti tipici

dell'inferenza statistica. Prima di iniziare una qualsiasi elaborazione dei dati è necessario avere informazioni generali sul fenomeno riguardanti:

- la natura del fenomeno in esame: istruzione e formazione;
- il numero di osservazioni disponibili (ampiezza del campione): 20 osservazioni;
- il numero di variabili utilizzate per rappresentare i diversi aspetti del fenomeno in esame (numero di caratteristiche): 5 variabili;
- il tipo di informazione disponibile per ciascuna variabile (qualitativa o quantitativa): variabili quantitative;
- lo scopo che l'analisi esplorativa dei dati si propone di raggiungere: analizzare nel dettaglio le varie tecniche di statistica descrittiva.

L'indagine statistica è sempre effettuata su un insieme di entità (individui, oggetti,...) in cui si manifesta il fenomeno che si studia. Questo insieme è detto popolazione o universo e può essere costituito da un numero finito oppure infinito di unità; nel primo caso si parla di popolazione finita e nel secondo caso di popolazione illimitata (non infinita ma molto grande). Se la popolazione risulta essere finita ma grande, oppure illimitata, non mi è possibile esaminare tutti gli individui che costituiscono la popolazione. Quindi non ho altro metodo che estrarre un campione dalla popolazione. Per ottenere campioni rappresentativi di una popolazione occorre scegliere gli elementi in modo completamente casuale poiché ogni criterio di selezione non casuale rischia di produrre campioni sbilanciati verso particolari valori.

2 RAPPRESENTAZIONI GRAFICHE DEI DATI

2.1 ISTOGRAMMI

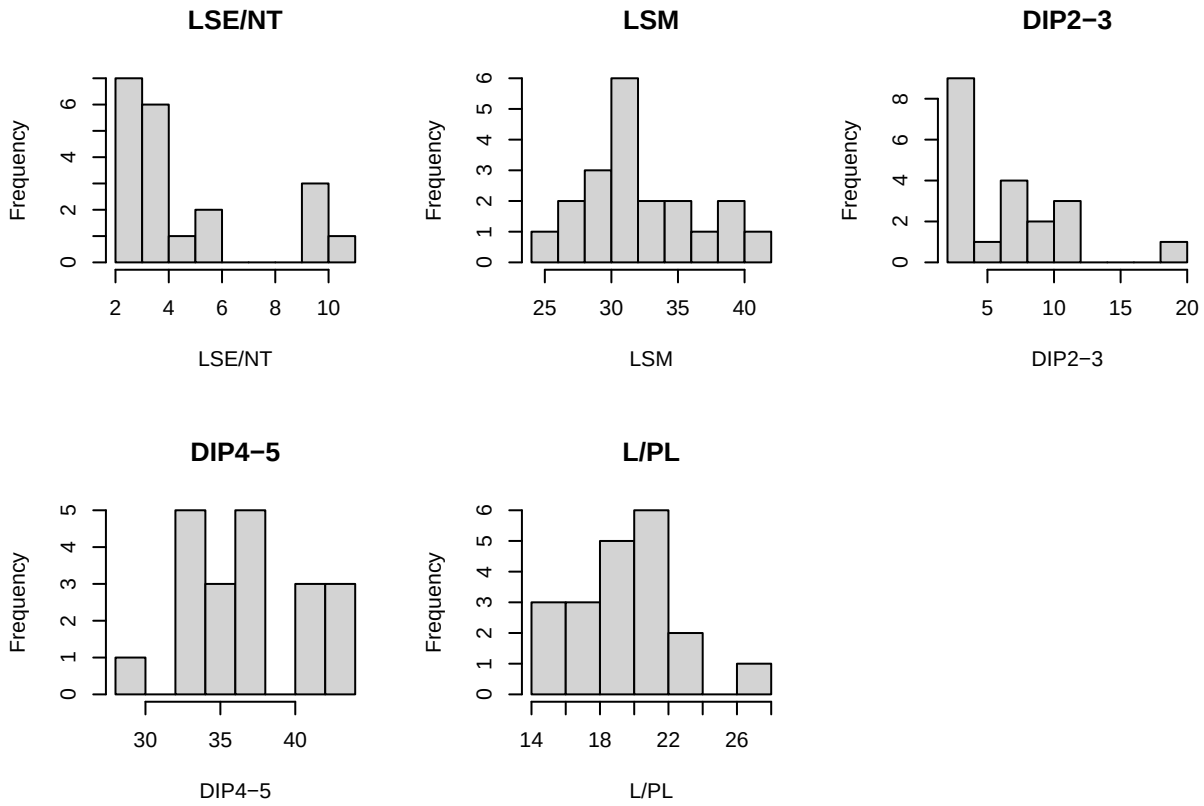
Gli istogrammi, utilizzati per variabili quantitative, sono una particolare rappresentazione grafica di una distribuzione di frequenza in classi. Si possono utilizzare anche i grafici a barre per rappresentare le frequenze delle classi, con la differenza che le barre sono centrate sulle singole classi e non adiacenti come negli istogrammi. In R funzione hist() per generare istogramma. Se parliamo di frequenza assoluta, l'area di ogni rettangolo deve rappresentare la frequenza assoluta della classe. Quindi la somma di tutte le aree è uguale alla numerosità del campione. Mentre per le frequenze relative la somma delle aree è uguale a 1. Nell'istogramma delle frequenze relative sull'ordinata deve esserci la density, e la base è l'ampiezza effettiva della classe. Mentre nell'istogramma delle frequenze assolute sull'ordinata abbiamo la frequenza assoluta e le basi sono unitarie, quindi l'altezza del rettangolo corrisponde proprio alla frequenza. Gli istogrammi ci permettono di avere informazioni sui nostri dati. L'istogramma si utilizza per variabili quantitative continue e pertanto, proprio per dare l'idea di continuità, le barre sono tra loro adiacenti. Il grafico a barre si utilizza invece per variabili discrete o qualitative e pertanto ci sarà uno spazio vuoto tra le barre. Inoltre, nell'istogramma l'asse orizzontale ha una unità di misura (quella della variabile) e la base dei rettangoli corrisponde all'ampiezza delle classi della variabile. Nel grafico a barre invece l'ampiezza delle basi dei rettangoli non ha un significato numerico.

La funzione hist() è in grado di generare oltre al grafico, anche una serie di informazioni sulla sua natura che possono essere salvate in un variabile h di tipo list. Queste informazioni possono essere visualizzate utilizzando la funzione str(h) applicata alla variabile generata dalla funzione hist().

La funzione str(h) fornisce i punti di suddivisione in classi (breaks), le frequenze assolute delle classi (counts), la densità delle classi (density) e i punti centrali delle classi (mids).

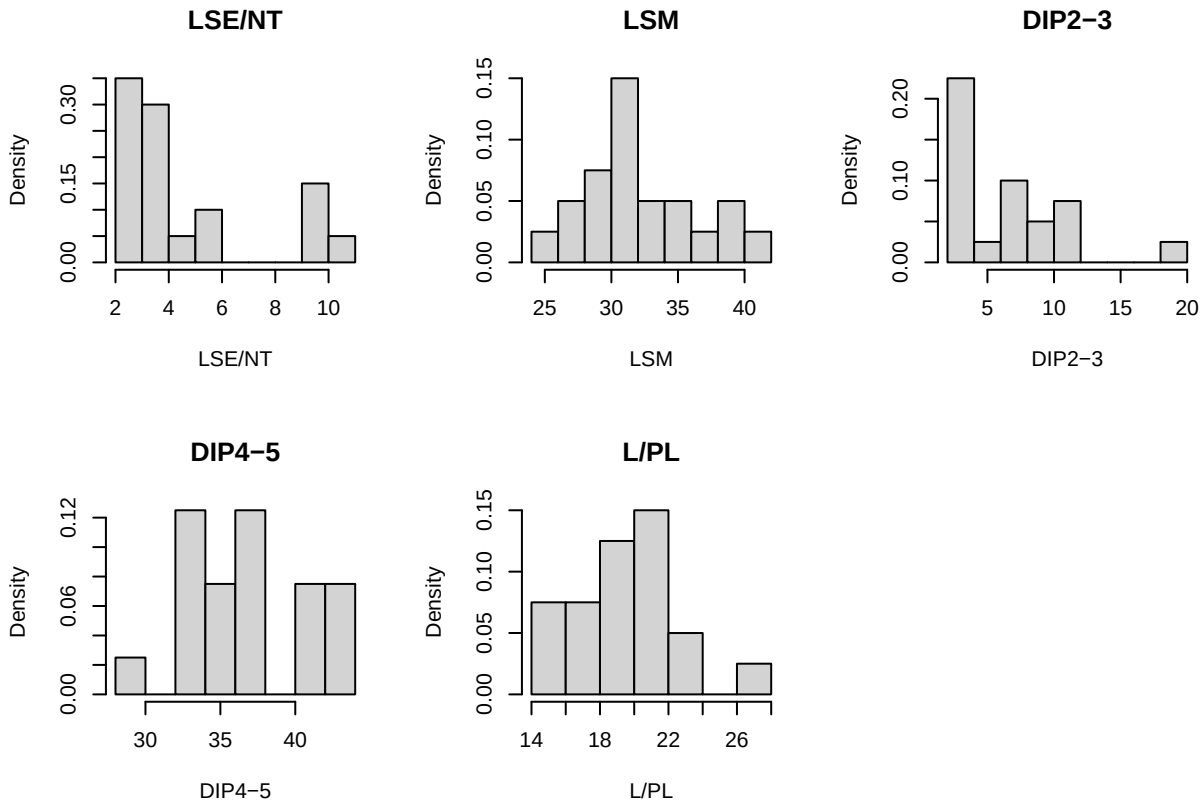
```
par(mfrow=c(2,3))

for(col in 1:ncol(df)) {
  hAbs <- hist(df[, col], freq=TRUE, main = names(df[col]), xlab=names(df[col]))
  str(hAbs)
}
```

```
## List of 6
##  $ breaks  : int [1:10] 2 3 4 5 6 7 8 9 10 11
##  $ counts  : int [1:9] 7 6 1 2 0 0 0 3 1
##  $ density : num [1:9] 0.35 0.3 0.05 0.1 0 0 0 0.15 0.05
##  $ mids    : num [1:9] 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5 10.5
##  $ xname   : chr "df[, col]"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"

## List of 6
##  $ breaks  : int [1:10] 24 26 28 30 32 34 36 38 40 42
##  $ counts  : int [1:9] 1 2 3 6 2 2 1 2 1
##  $ density : num [1:9] 0.025 0.05 0.075 0.15 0.05 0.05 0.025 0.05 0.025
##  $ mids    : num [1:9] 25 27 29 31 33 35 37 39 41
##  $ xname   : chr "df[, col]"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"

## List of 6
##  $ breaks  : int [1:10] 2 4 6 8 10 12 14 16 18 20
##  $ counts  : int [1:9] 9 1 4 2 3 0 0 0 1
##  $ density : num [1:9] 0.225 0.025 0.1 0.05 0.075 0 0 0 0.025
##  $ mids    : num [1:9] 3 5 7 9 11 13 15 17 19
##  $ xname   : chr "df[, col]"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"

## List of 6
##  $ breaks  : int [1:9] 28 30 32 34 36 38 40 42 44
##  $ counts  : int [1:8] 1 0 5 3 5 0 3 3
##  $ density : num [1:8] 0.025 0 0.125 0.075 0.125 0 0.075 0.075
##  $ mids    : num [1:8] 29 31 33 35 37 39 41 43
##  $ xname   : chr "df[, col]"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"

## List of 6
##  $ breaks  : int [1:8] 14 16 18 20 22 24 26 28
##  $ counts  : int [1:7] 3 3 5 6 2 0 1
##  $ density : num [1:7] 0.075 0.075 0.125 0.15 0.05 0 0.025
##  $ mids    : num [1:7] 15 17 19 21 23 25 27
##  $ xname   : chr "df[, col]"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"
```



Come possiamo vedere dagli istogrammi delle frequenze assolute/relative,in base alla suddivisione in classi effettuata da R si ottenie che:

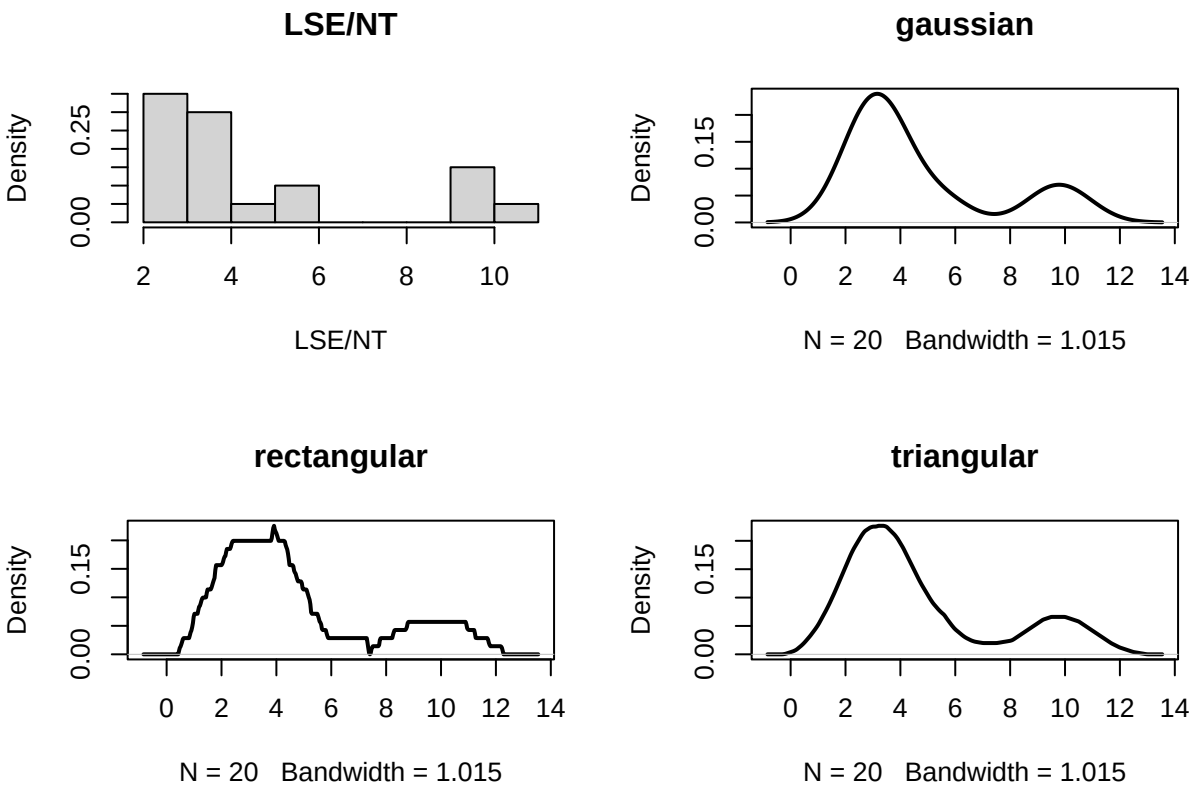
- i valori più frequenti per “licenza di scuola elementare, nessun titolo di studio” variano tra il 2% e 3%;
- i valori più frequenti per “licenza di scuola media” variano tra il 30% e il 32%;
- i valori più frequenti per “diploma 2-3 anni (qualifica professionale)” vanno dal 2% al 4%;
- i valori più frequenti per “diploma 4-5 anni (diploma)” vanno dal 32% al 34% e dal 36% al 38%;
- i valori più frequenti per “laurea e post-laurea” vanno dal 20% al 22%.

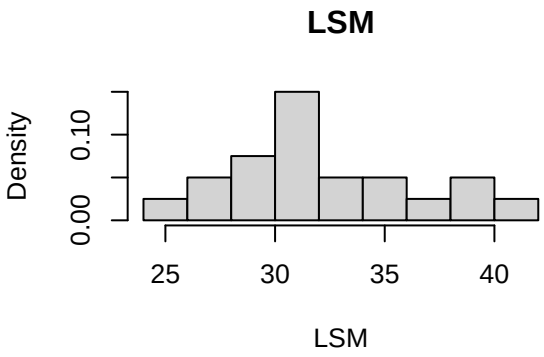


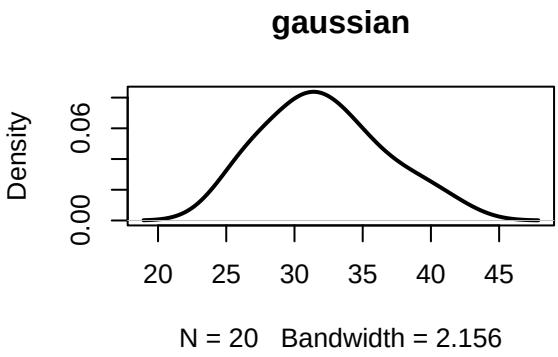
2.2 KERNEL DENSITY PLOT

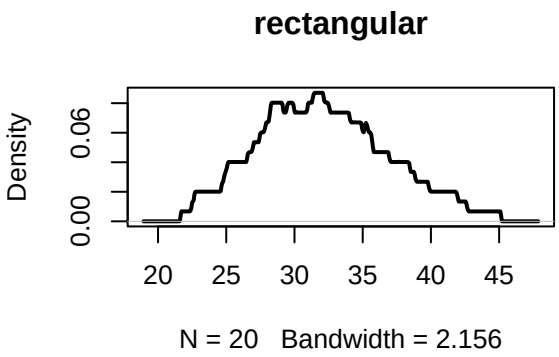
Gli istogrammi sono un importante strumento per la rappresentazione di una distribuzione di frequenza in classi per variabili quantitative univariate. La scelta degli intervalli delle classi è cruciale per l'aspetto finale del grafico dell'istogramma. Un modo più elegante per affrontare questo problema consiste nello smussare l'istogramma utilizzando grafici che utilizzano la stima di densità basata su kernel. Con tale metodo, invece di raccogliere le osservazioni in barre, come negli istogrammi, si traccia una curva continua determinata da un fattore K, detto kernel, e da un parametro h, detto ampiezza della banda (bandwidth). La scelta del kernel K(x) può influenzare l'aspetto generale del grafico. Inoltre, la scelta del parametro h è uno degli aspetti più delicati del metodo: un valore troppo vicino a zero rende la stima irregolare e con varianza troppo elevata; invece, un valore troppo elevato comporta problemi di distorsione. Il kernel che si basa sulla densità normale standard è chiamato "gaussian" ed è l'impostazione di default in R. La scelta del kernel dipende dal campione, ma la scelta di default è spesso quella da preferire.

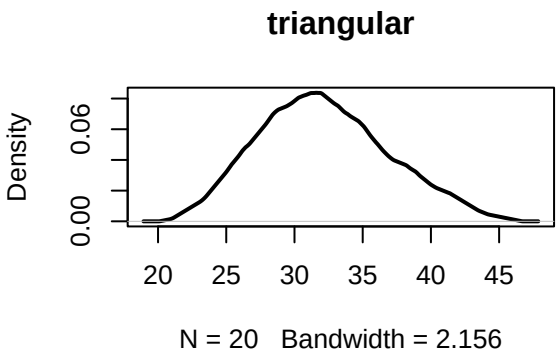
```
for(col in 1:ncol(df)) {
  par(mfrow=c(2,2))
  hist(df[, col], freq=FALSE, main = names(df[col]), xlab=names(df[col]))
  for(k in c("gaussian","rectangular","triangular")){
    d <- density(df[, col], kernel = k)
    plot(d,lwd=2, main = k)
  }
}
```

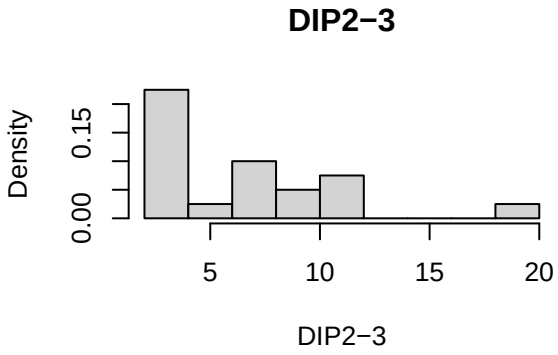

LSE/NT
Density
0.00
0.25
2
4
6
8
10
LSE/NT
gaussian
Density
0.00
0.15
0
2
4
6
8
10
12
14
N = 20 Bandwidth = 1.015
rectangular
Density
0.00
0.15
0
2
4
6
8
10
12
14
N = 20 Bandwidth = 1.015
triangular
Density
0.00
0.15
0
2
4
6
8
10
12
14
N = 20 Bandwidth = 1.015

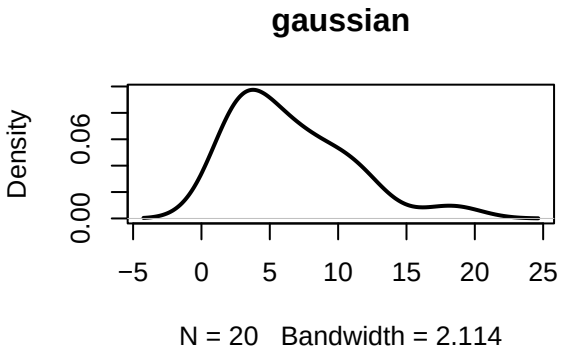

LSM
Density
0.00
0.10
25
30
35
40
LSM

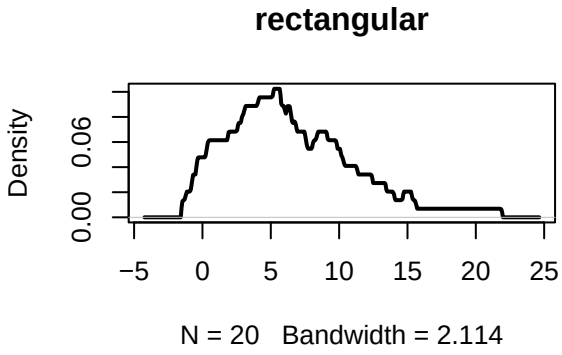

gaussian
Density
0.00
0.06
20
25
30
35
40
45
N = 20 Bandwidth = 2.156

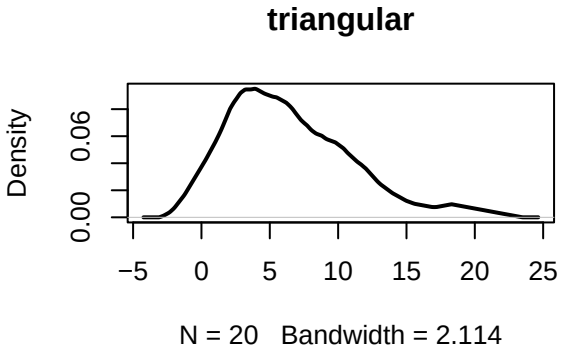

rectangular
Density
0.00
0.06
20
25
30
35
40
45
N = 20 Bandwidth = 2.156

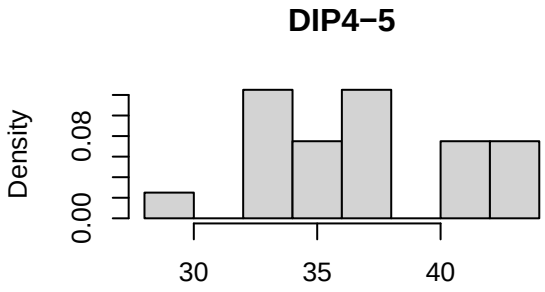

triangular
Density
0.00
0.06
20
25
30
35
40
45
N = 20 Bandwidth = 2.156

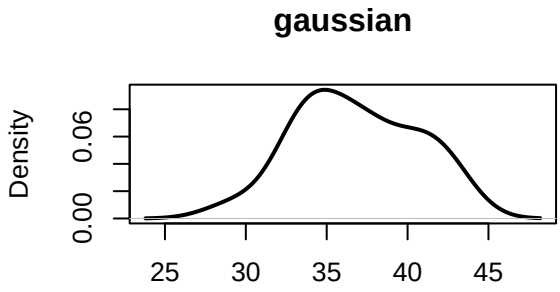

DIP2−3
Density
0.00
0.15
5
10
15
20
DIP2−3

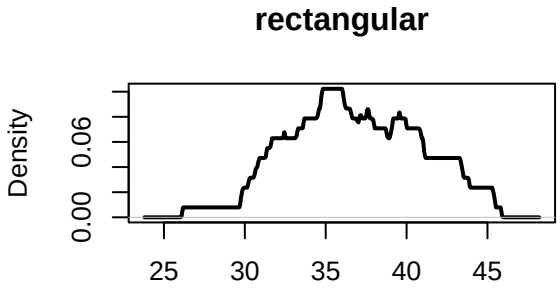

gaussian
Density
0.00
0.06
−5
0
5
10
15
20
25
N = 20 Bandwidth = 2.114

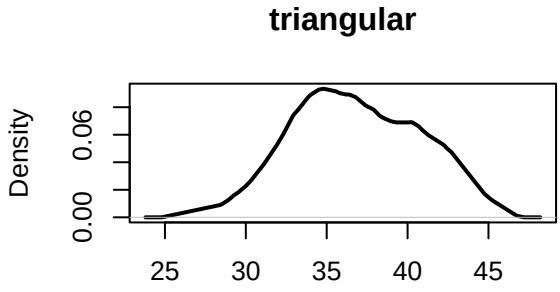

rectangular
Density
0.00
0.06
−5
0
5
10
15
20
25
N = 20 Bandwidth = 2.114

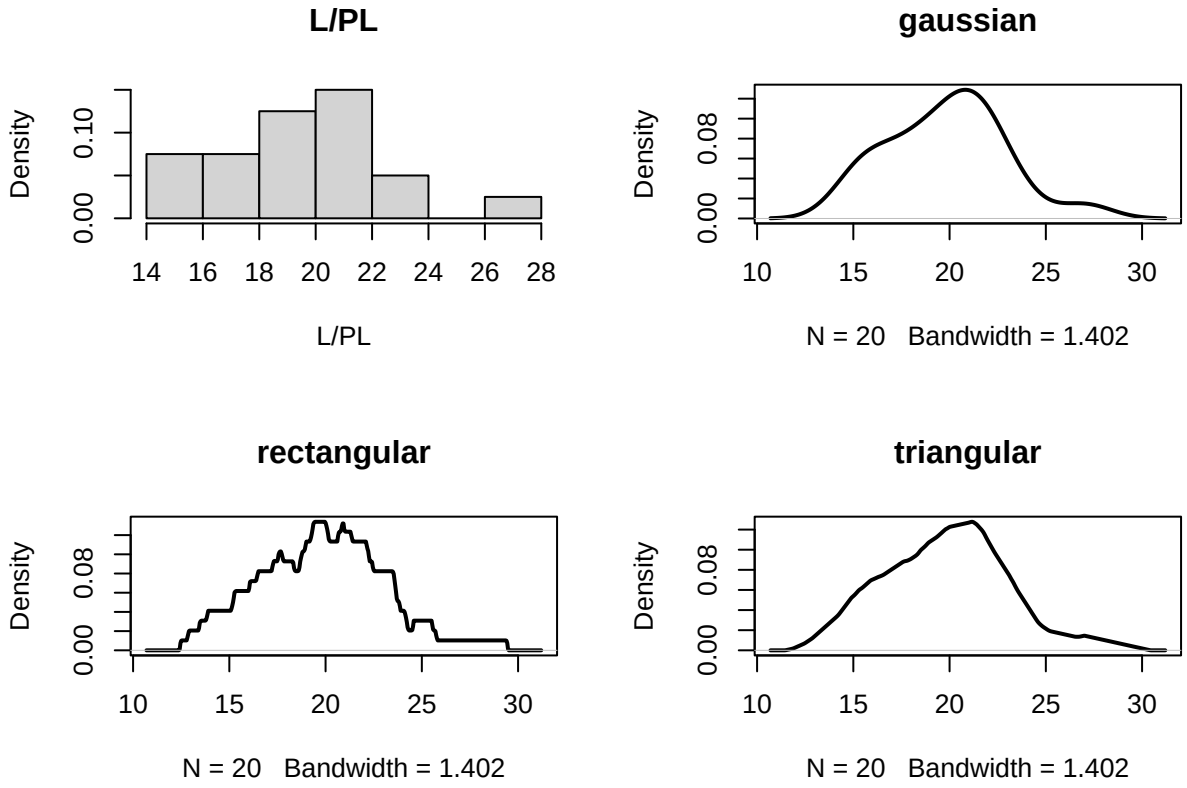

triangular
Density
0.00
0.06
−5
0
5
10
15
20
25
N = 20 Bandwidth = 2.114


DIP4−5
Density
0.00
0.08
30
35
40
DIP4−5


gaussian
Density
0.00
0.06
25
30
35
40
45
N = 20 Bandwidth = 1.835


rectangular
Density
0.00
0.06
25
30
35
40
45
N = 20 Bandwidth = 1.835


triangular
Density
0.00
0.06
25
30
35
40
45
N = 20 Bandwidth = 1.835

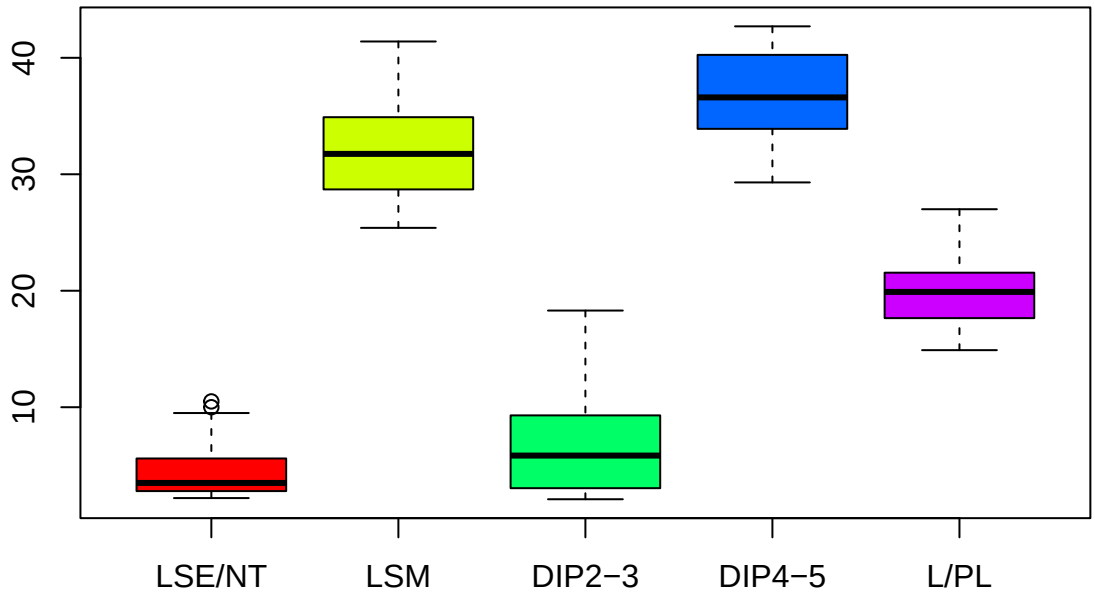


2.3 BOXPLOT

Un boxplot è una rappresentazione grafica utilizzata per descrivere la distribuzione di un campione, non è altro che una scatola che risulta avere dei baffi. Il boxplot può essere utilizzato solo per variabili quantitative. Il boxplot viene costruito autoamaticamente da R, ma se dovessimo crearlo a mano dovremmo prima di tutto ordinare i dati in ordine crescente. Poi dividiamo i dati in 4 parti, creando il 1' quartile Q1 (25% dei dati sono a sinistra e 75% a destra), il 2' quartile (50% dei dati sono a sinistra e 50% a destra, corrispone alla mediana), il 3' quartile (75% dei dati sono a sinistra e 25% a destra). Q0 e Q4 sono il minimo e il massimo dei valori del campione. La linea inferiore della scatola corrisponde a Q1, la linea centrale corrisponde a Q2 e la linea superiore corrisponde a Q3. Se la mediana risulta spostata verso l'alto o il basso vuol dire che nei dati non c'è simmetria. Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale di Q1 - 1.5 $\cdot$ (Q3 - Q1), mentre il baffo superiore corrisponde al volore più grande delle osservazioni che risulta minore o uguale a Q3 + 1.5 $\cdot$ (Q3 - Q1). La distanza tra Q1 e Q3 è detto intervalo interquartile o scarto interquartile. Se tutti i dati rientrano nell'intervallo (Q1 - 1.5 $\cdot$ (Q3 - Q1), Q3 + 1.5 $\cdot$ (Q3 - Q1)) i baffi sono posti in corrispondenza del minimo e del massimo valore del campione. Gli eventuali valori al di fuori di questo intervallo sono rappresentati sotto forma di punti e sono detti valori anomali o outlier. Outliers importanti anche successivamente con l'analisi dei cluster.

Per confrontare differenti variabili che descrivono insiemi di dati numerici di uno stesso fenomeno quantitativo, è possibile costruire un grafico che contiene i diversi boxplot delle distribuzioni associate alle diverse variabili. Col boxplot non vediamo la media, vediamo solo informazioni legate alla mediana e ai quartili.

```
boxplot(df, notch = FALSE, col = rainbow(5))
```



Con il boxplot possiamo capire varie cose, tra cui:

- centralità: espressa dalla mediana;
- simmetria dei dati: la si può dedurre esaminando le distanze del primo e del terzo quartile dalla mediana. Dal grafico si si può notare che la classe LSE/NT risulta fortemente asimmetrica, mentre le restanti classi sembrano essere più simmetriche;
- dispersione dei dati: guardando fra Q1 e baffo, Q3 e baffo, se c'è grossa variazione (dispersione) nei dati (se sono lunghi i baffi significa che c'è dispersione nei dati). La si può notare maggiormente nelle classi LSE/NT e DIP2-3 ;
- ed inoltre vediamo se ci sono dei valori anomali, e bisogna scoprire quali sono questi valori.

```
summary(df)
```

```
##      LSE/NT           LSM            DIP2-3           DIP4-5     
##  Min.   : 2.20   Min.   :25.40   Min.   : 2.100   Min.   :29.30  
##  1st Qu.: 2.85   1st Qu.:28.75   1st Qu.: 3.125   1st Qu.:33.95  
##  Median : 3.50   Median :31.75   Median : 5.850   Median :36.60  
##  Mean   : 4.71   Mean   :32.15   Mean   : 6.510   Mean   :36.82  
##  3rd Qu.: 5.60   3rd Qu.:34.80   3rd Qu.: 9.050   3rd Qu.:40.23  
##  Max.   :10.50   Max.   :41.40   Max.   :18.300   Max.   :42.70  
##       L/PL       
##  Min.   :14.90  
##  1st Qu.:17.68  
##  Median :19.90  
##  Mean   :19.82
```

```
##  3rd Qu.:21.48
##  Max.   :27.00
```

```
calcolaIntervallo <- function(x){
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  intervallo <- c(Q1 - 1.5 * IQR, Q3 + 1.5 * IQR)
  return (intervallo)
}

intervalli <- apply(df,2,calcolaIntervallo)
rownames(intervalli) <- c("Q1 - 1.5·(Q3 - Q1) ","Q3 + 1.5·(Q3 - Q1) ")
# round(intervalli,1)
intervalli
```

```
##                      LSE/NT    LSM  DIP2-3  DIP4-5   L/PL
## Q1 - 1.5·(Q3 - Q1)  -1.275 19.675 -5.7625 24.5375 11.975
## Q3 + 1.5·(Q3 - Q1)   9.725 43.875 17.9375 49.6375 27.175
```

```
out <- boxplot.stats(df$`LSE/NT`)$out
out_ind <- which(df$`LSE/NT` %in% c(out))
out_ind
```

```
## [1] 15 18
```

```
df[out_ind,]
```

```
##          LSE/NT  LSM DIP2-3 DIP4-5 L/PL
## Campania   10.0 36.1    3.2   34.5 16.3
## Calabria   10.5 34.7    2.1   36.8 15.9
```

Quindi i valori anomali per la variabile LSE/NT sono rappresentati dalle regioni Campania e Calabria.

3 STATISTICA DESCRITTIVA UNIVARIATA

La statistica descrittiva univariata consiste nell'analisi delle singole variabili (caratteristiche) della popolazione. Per i fenomeni quantitativi è spesso utile definire la funzione di distribuzione empirica. A seconda del fenomeno posso avere una funzione di distribuzione empirica discreta o continua.

3.1 FUNZIONE DI DISTRIBUZIONE EMPIRICA DISCRETA

Nel caso discreto questa funzione è definita a partire dalle frequenze relative cumulative, dove la generica F_i rappresenta la proporzione dei dati del campione minori o uguali di z_i .

Se supponiamo che i k valori distinti assunti dalla variabile quantitativa X siano ordinati in ordine crescente, ossia $z_1 < z_2 < \ldots < z_k$, allora la funzione di distribuzione empirica F(x) è così definita:

$$F(x) = \frac{\#\{x_i \leq x, i = 1, 2, \ldots, n\}}{n} = \begin{cases} 0, & x < z_1 \\ F_1, & z_1 \leq x < z_2 \\ \ldots & \\ F_i, & z_i \leq x < z_{i+1} \\ \ldots & \\ 1, & x \geq z_k \end{cases}$$

La funzione di distribuzione è definita per ogni x reale . In questo caso deve essere una funzione a gradini in cui ogni gradino indica quale proporzione di dati presenta un valore minore o uguale di quello indicato sull'asse delle ascisse e gode delle seguenti proprietà:

- non decrescente, poiche andiamo man mano a sommare elementi, ed in alcuni casi è costante;
- la funzione assume il valore a sinistra in corrispondenza ad ogni punto di salto;
- la funzione vale 0 per ogni valore minore dell'osservazione minima e vale 1 per ogni valore maggiore o uguale all'osservazione massima

Funzione di distribuzione usata anche per trovare la mediana.

3.2 FUNZIONE DI DISTRIBUZIONE EMPIRICA CONTINUA

Per fenomeni quantitativi continui occorre considerare la funzione di distribuzione empirica continua, ossia una funzione di distribuzione empirica strutturata in classi. Supponiamo di organizzare i dati numerici in k distinte classi $C_1 = [z_0, z_1), C_2 = [z_1, z_2), ..., C_k = [z_{k-1}, z_k]$, con $z_0 < z_1 < ... < z_k$, dove z_0 corrisponde al minimo delle osservazioni e z_k al massimo delle osservazioni. La funzione di distribuzione empirica continua è cosi definita:

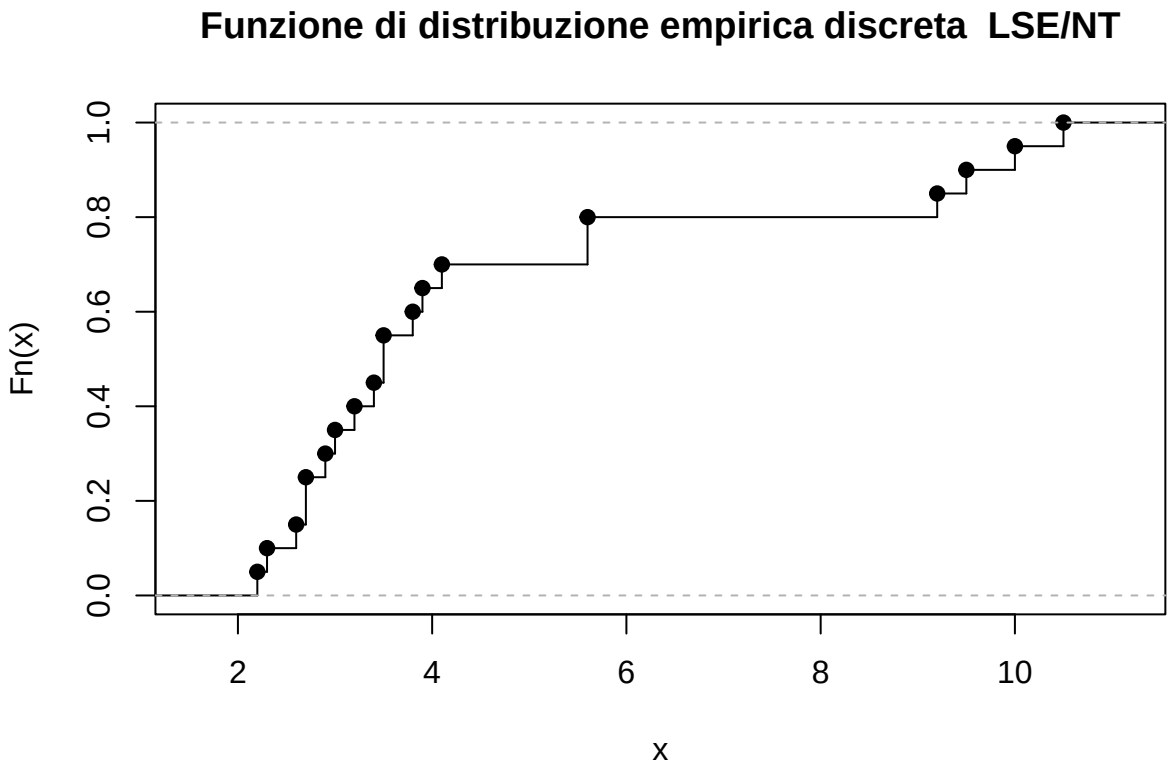
$$F(x) = \begin{cases} 0, & x < z_0 \\ \ldots\ldots & \\ F_{i-1}, & x = z_{i-1} \\ \dfrac{F_i - F_{i-1}}{z_i - z_{i-1}}\, x + \dfrac{z_i\, F_{i-1} - z_{i-1}\, F_i}{z_i - z_{i-1}}, & z_{i-1} < x < z_i \\ F_i, & x = z_i \\ \ldots\ldots & \\ 1, & x \geq z_k, \end{cases}$$

dove $F_0 = 0$ e F_i denota la frequenza relativa cumulativa della classe $C_i (i = 1, 2, ..., k)$. Questa funzione è non decrescente poiché andiamo a sommare delle frequenze cumulate, inoltre si nota che $F(x) = 0$ per $x < z_0$, $F(x) = 1$ per $x >= z_k$, mentre se $z_{i-1} < x < z_i$ la funzione di distribuzione empirica continua coincide con il segmento che passa per i punti (z_{i-1}, F_{i-1}) e (z_i, F_i), ossia

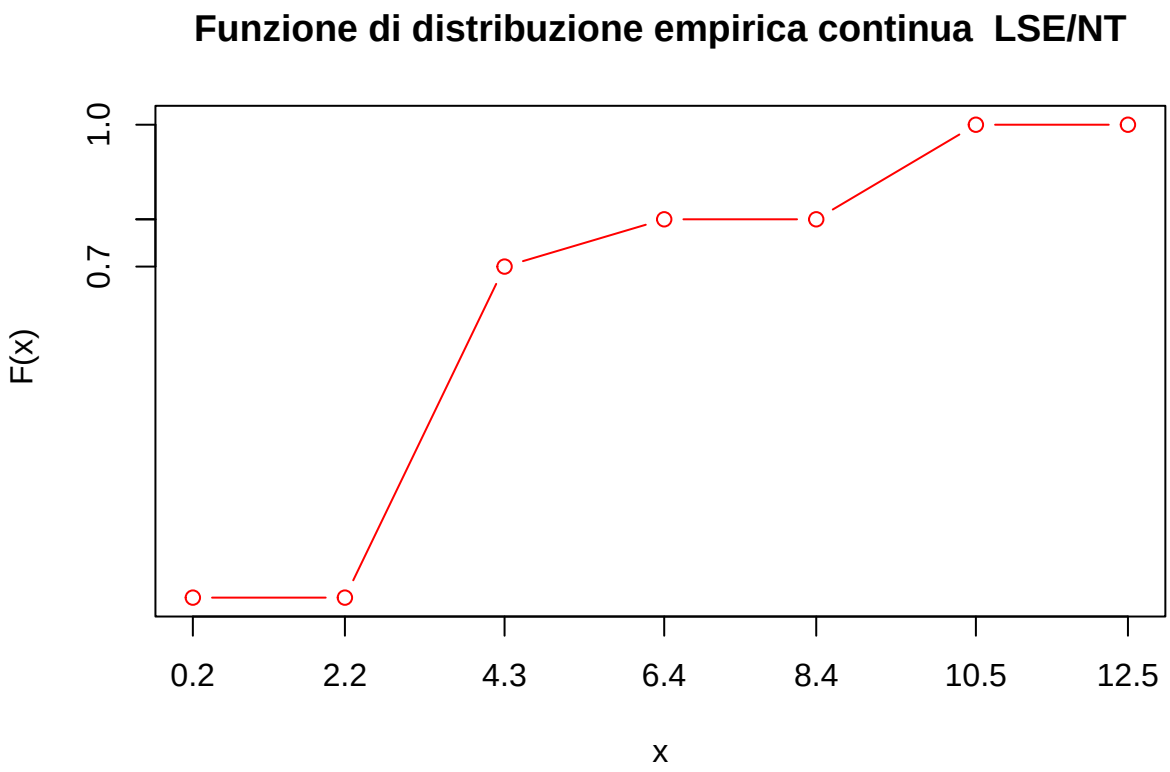
$$\frac{y - F_{i-1}}{x - z_{i-1}} = \frac{F_i - F_{i-1}}{z_i - z_{i-1}} \quad (i = 1, 2, .., k)$$

Andiamo ora a visualizzare entrambi le funzioni di distribuzione per le nostre variabili.

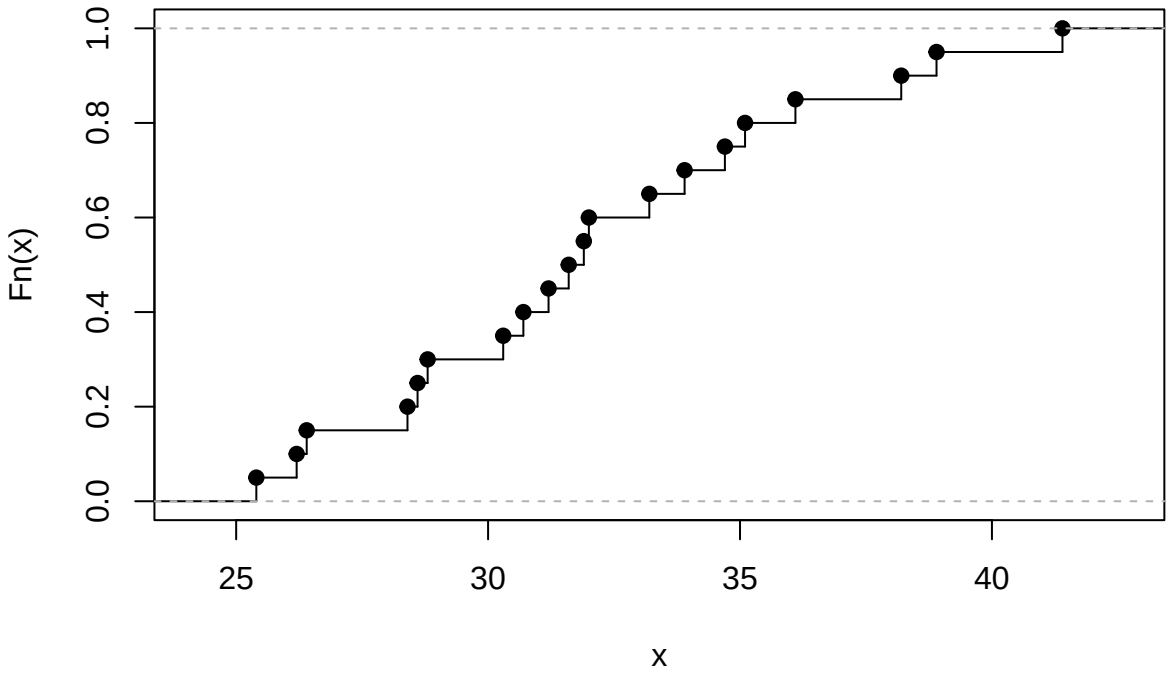
```r
for(i in 1:ncol(df)){

  plot(ecdf(df[, i]), main = paste("Funzione di distribuzione empirica discreta ",
                                   names(df[i])), verticals = TRUE)
  fRelCum <- cumsum(table(df[, i]))/length(df[, i])
  print(fRelCum)

  classi <- round(seq(min(df[, i]), max(df[, i]), by=(max(df[, i]) - min(df[, i]))/4),1)
  frelClassi <- table(cut(df[, i], breaks = classi, right = FALSE))/length(df[, i])
  frelClassi
  freqRel <- table(df[, i])/length(df[, i])
  m <- length(freqRel)
  Fcum <- cumsum(frelClassi)
  Fcum[4] <- Fcum[4] + freqRel[m]
  ascisse <- c(min(df[, i]) - 2, classi,max(df[, i]) + 2)
  ordinate <- c(0,0,Fcum[1:4],1)
  plot ( ascisse , ordinate , type = "b" , axes = FALSE ,
         main = paste("Funzione di distribuzione empirica continua ", names(df[i])),
         col =" red " , ylim = c(0,1) , xlab = "x", ylab = " F(x)")
  axis (1 , ascisse )
  axis (2 , format ( Fcum , digits = 2) )
  box ()
  print(Fcum)

}
```



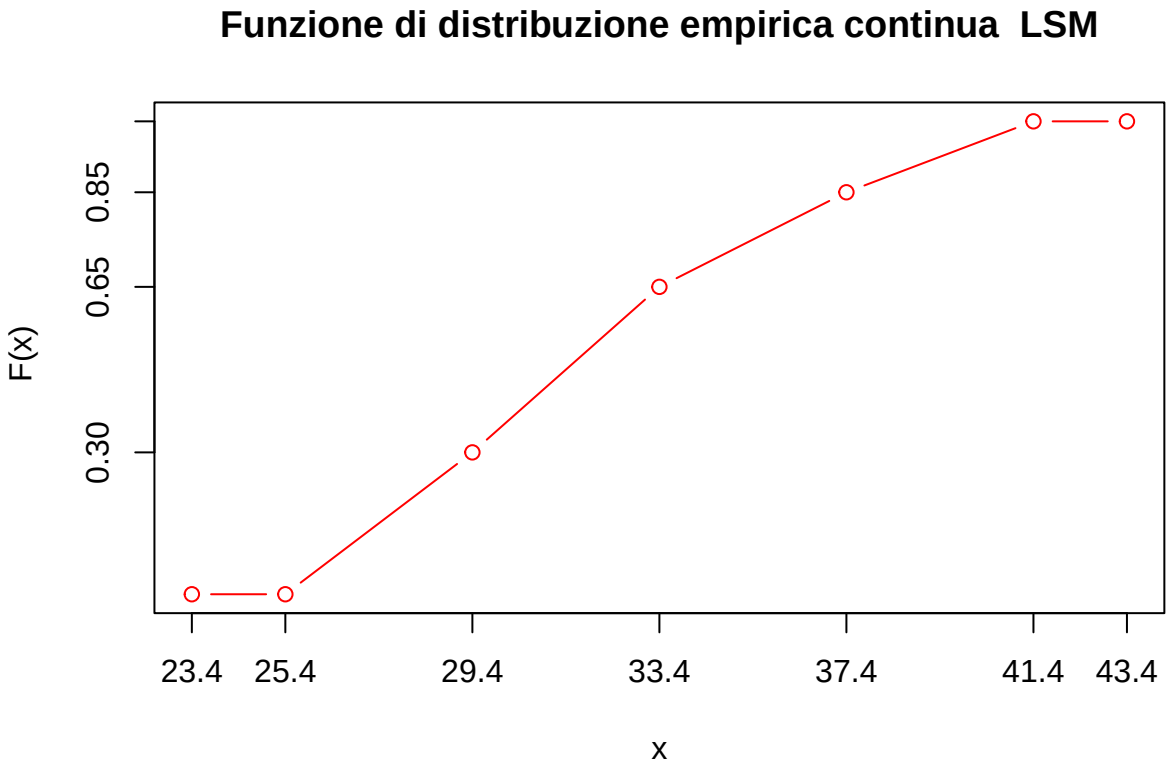
```
##  2.2  2.3  2.6  2.7  2.9    3  3.2  3.4  3.5  3.8  3.9  4.1  5.6  9.2  9.5   10
## 0.05 0.10 0.15 0.25 0.30 0.35 0.40 0.45 0.55 0.60 0.65 0.70 0.80 0.85 0.90 0.95
## 10.5
## 1.00
```



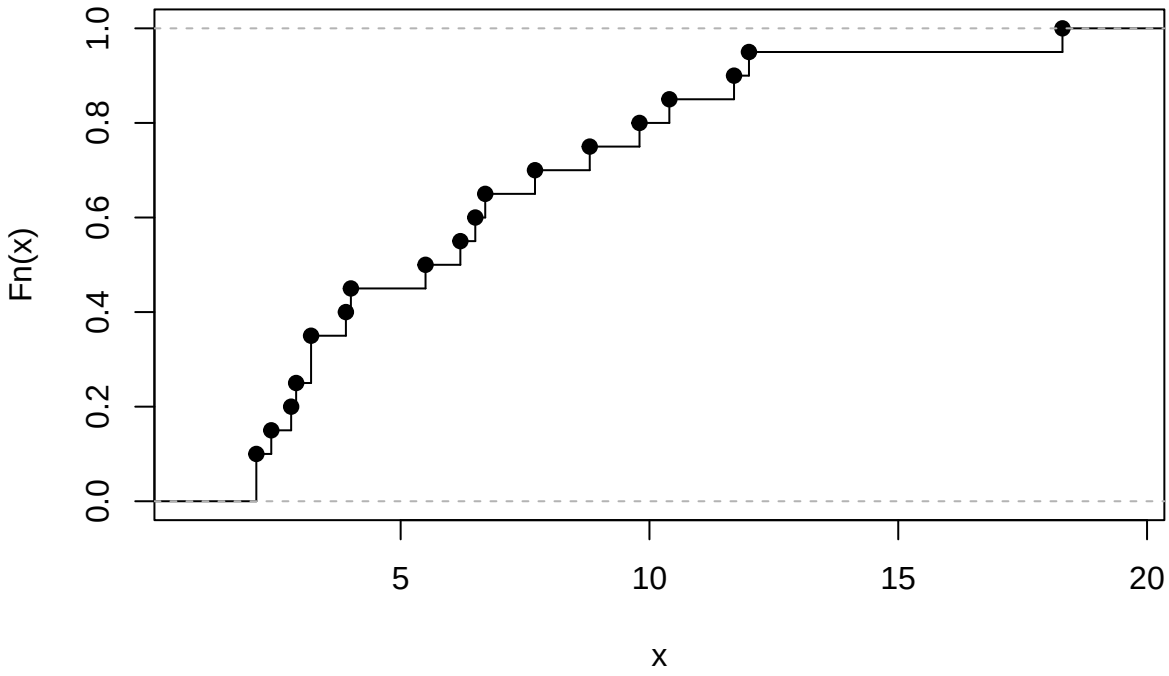

```
##  [2.2,4.3)  [4.3,6.4)  [6.4,8.4) [8.4,10.5)
##        0.7        0.8        0.8        1.0
```



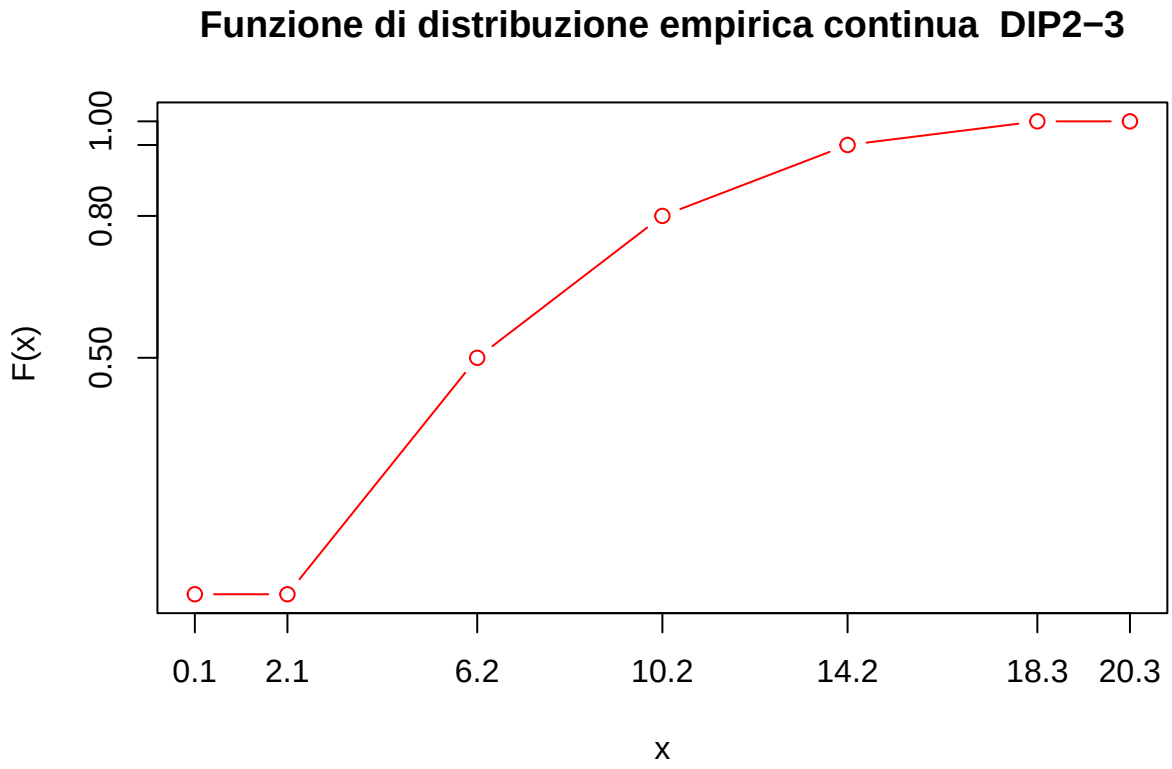

```
## 25.4 26.2 26.4 28.4 28.6 28.8 30.3 30.7 31.2 31.6 31.9   32 33.2 33.9 34.7 35.1
## 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80
## 36.1 38.2 38.9 41.4
## 0.85 0.90 0.95 1.00
```



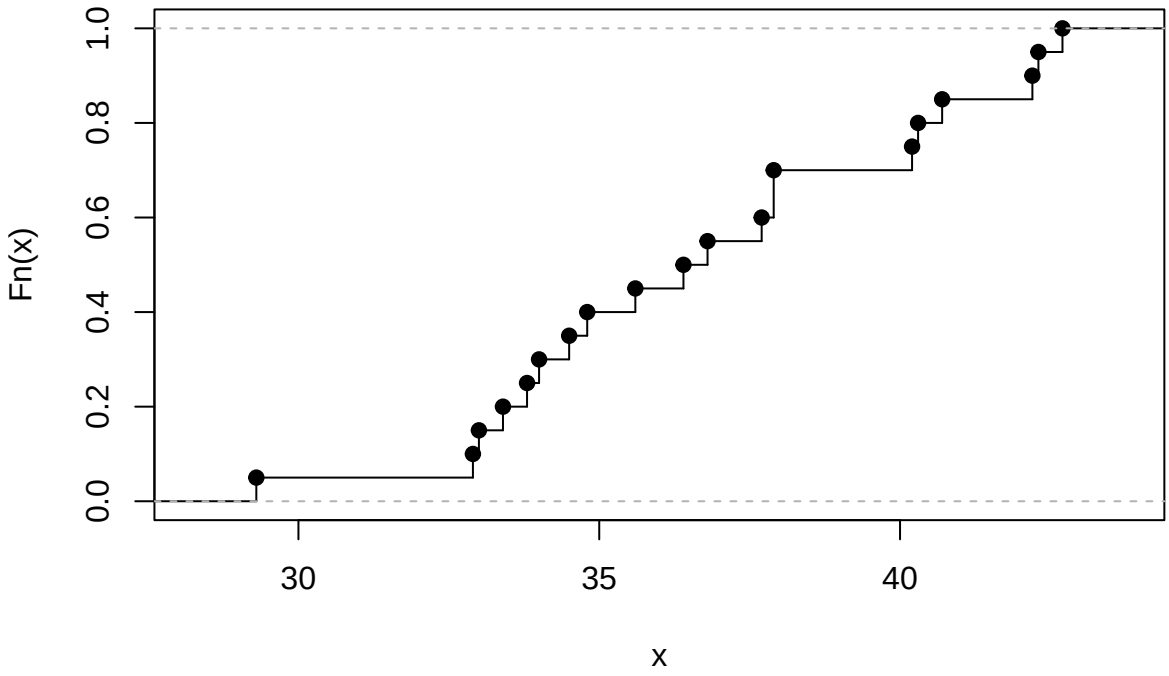

```
## [25.4,29.4) [29.4,33.4) [33.4,37.4) [37.4,41.4)
##        0.30        0.65        0.85        1.00
```



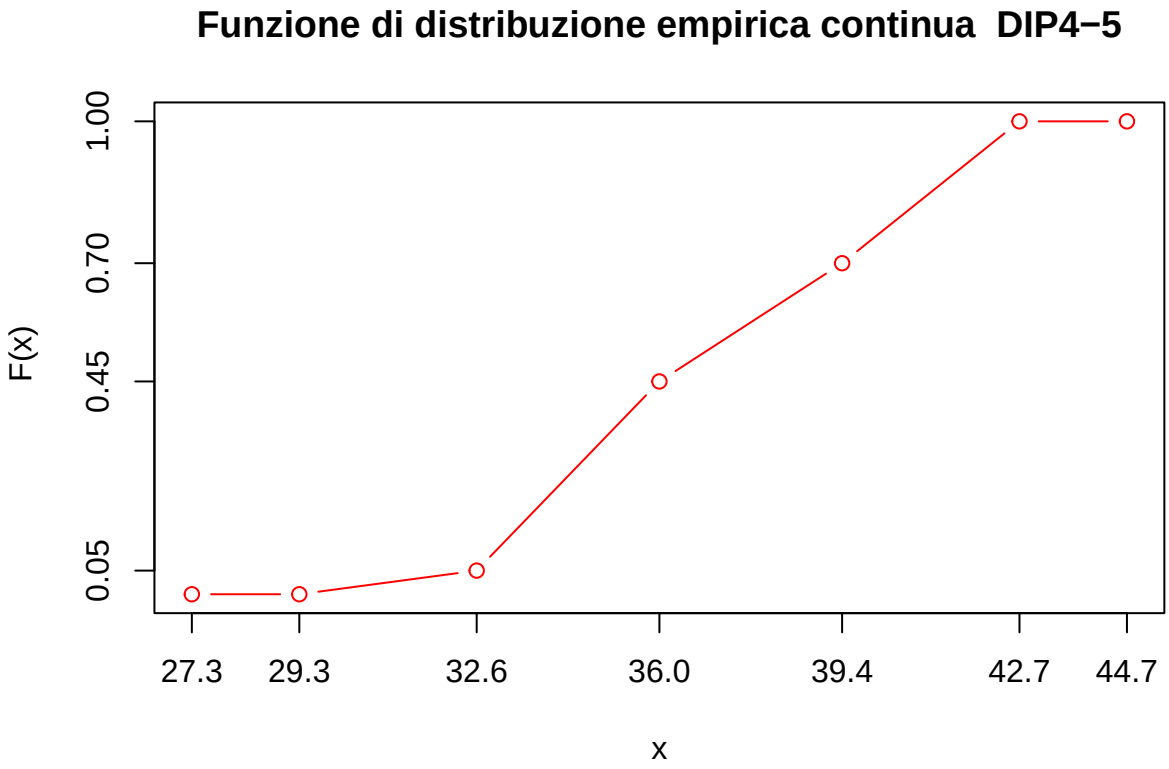

```
##  2.1  2.4  2.8  2.9  3.2  3.9    4  5.5  6.2  6.5  6.7  7.7  8.8  9.8 10.4 11.7
## 0.10 0.15 0.20 0.25 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90
##   12 18.3
## 0.95 1.00
```



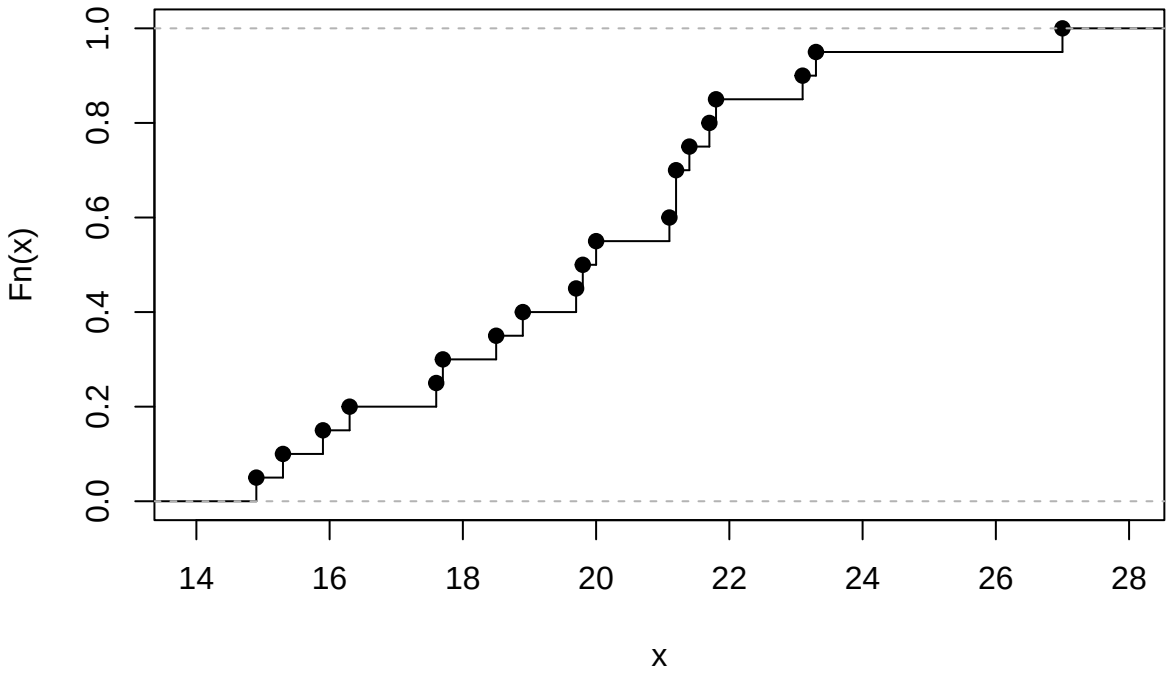

```
##   [2.1,6.2)  [6.2,10.2) [10.2,14.2) [14.2,18.3)
##        0.50        0.80        0.95        1.00
```



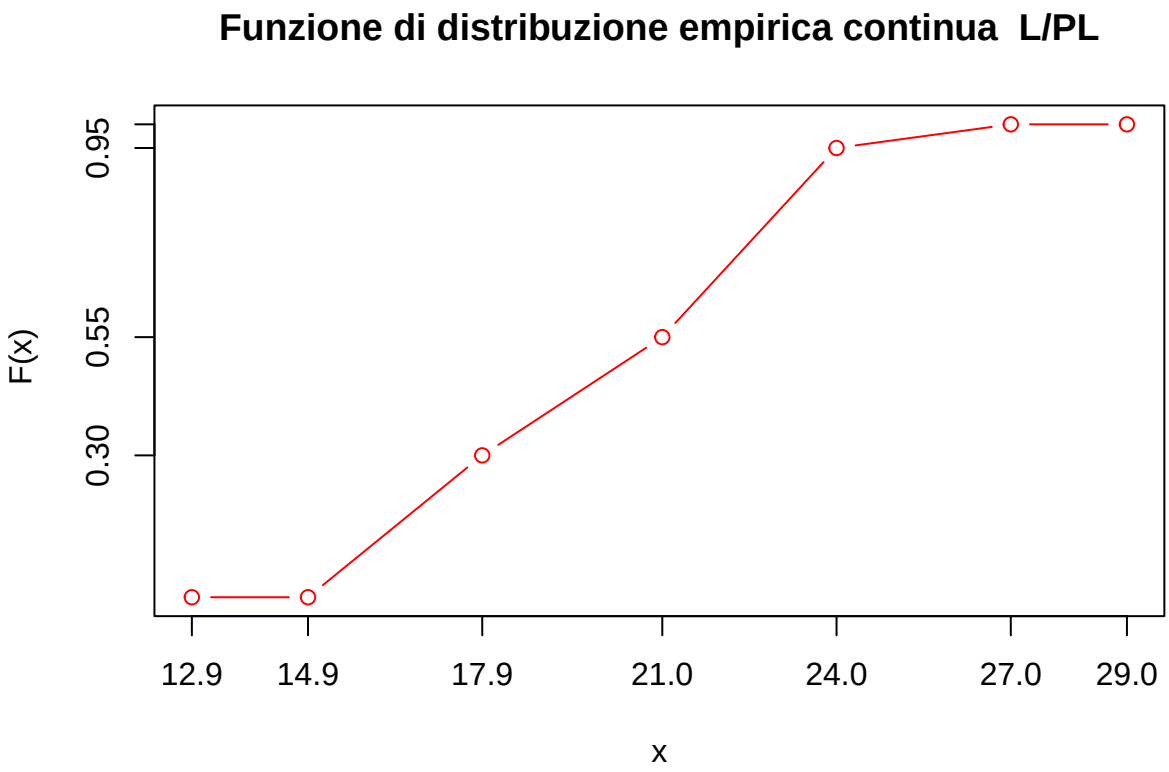

```
## 29.3 32.9   33 33.4 33.8   34 34.5 34.8 35.6 36.4 36.8 37.7 37.9 40.2 40.3 40.7
## 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.70 0.75 0.80 0.85
## 42.2 42.3 42.7
## 0.90 0.95 1.00
```




```
## [29.3,32.6)   [32.6,36)   [36,39.4) [39.4,42.7)
##        0.05        0.45        0.70        1.00
```




```
## 14.9 15.3 15.9 16.3 17.6 17.7 18.5 18.9 19.7 19.8   20 21.1 21.2 21.4 21.7 21.8
## 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.70 0.75 0.80 0.85
## 23.1 23.3   27
## 0.90 0.95 1.00
```




```
## [14.9,17.9)    [17.9,21)      [21,24)      [24,27)
##        0.30         0.55         0.95         1.00
```

3.3 INDICI DI SINTESI

Alcuni indici di sintesi, detti anche statistiche, utili a descrivere dei dati numerici, sono media, mediana, moda, quantili, varianza, deviazione standard e coefficiente di variazione. Media, mediana, moda, quantili sono indici di posizione. Media, mediana e moda sono indici di posizioni centrale, perché descrivono attorno a quali valori è centrato l'insieme dei dati, mentre i quartili sono indici di posizione non centrali. Varianza, deviazione standard e coefficiente di variazione sono invece indici di dispersione, servono quindi per cercare di misurare la dispersione dei dati intorno alla media. Tutti questi indici servono a misurare quantitativamente alcune delle caratteristiche osservate qualitativamente nei grafici delle distribuzioni di frequenze e nei boxplot.

3.3.1 MEDIA CAMPIONARIA

La media campionaria di un insieme di dati numerici non è altro che la media aritmetica di questi dati, ossia il rapporto tra la somma di tutti i dati per il numero di dati.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media campionaria gode della proprietà di linearità, ossia se consideriamo un nuovo insieme di dati $y_i = ax_i + b$ espresso come relazione lineare, allora la media campionaria di $y_1, y_2, ..., y_n$ è legata alla media

campionaria dei dati iniziali $x_1, x_2, ..., x_n$ dalla stessa relazione lineare.

$$\bar{y} = a\bar{x} + b$$

Tutti i valori di un campione contribuiscono al calcolo della media. Quindi la media è influenzata da valori troppo piccoli o troppo grandi. Cosa che non accade invece per la mediana.

Per ogni valore x_i si definisce lo scarto dalla media campionaria la quantità

$$s_i = x_i - \bar{x}$$

che indica il grado di scostamento del singolo valora dalla media campionaria. La somma di tutti gli scarti è nulla.

3.3.2 MEDIANA CAMPIONARIA

Consideriamo un insieme di dati numerici di ampiezza n, ordiniamo in ordine crescente. Se n è dispari la mediana campionaria corrisponde al valori in posizione $\frac{n+1}{2}$, se n è pari invece la mediana corrisponde alla media aritmetica dei valori in posizione $\frac{n}{2}$ e $\frac{n}{2} + 1$

La mediana campionaria a differenza della media dipende solo da uno o da due valori centrali dei dati e non risente dei valori estremi. Inoltre, l'uso della mediana come indice per descrivere le caratteristiche dei dati ha lo svantaggio di dover prima riordinare i dati in ordine crescente, il che non è richiesto per il calcolo della media.

La mediana è preferibile alla media quando si desidera eliminare gli effetti di valori estremi molto diversi dagli altri dati (valori anomali) poiché la mediana non utilizza tutti i dati, ma solo il valore centrale o i due valori centrali. Tuttavia occorre sottolineare che l'utilizzazione dei soli dati centrali rende la mediana poco sensibile a tutti gli altri valori dei dati e questo può costituire un limite per questo indice. Per descrivere la forma di una distribuzione si può confrontare la media campionaria e la mediana campionaria. Se queste due misure sono uguali la distribuzione di frequenze tende ad essere simmetrica; se la media campionaria è sensibilmente maggiore della mediana campionaria, la distribuzione di frequenze è più sbilanciata verso destra se invece la media campionaria è sensibilmente minore mediana campionaria la distribuzione di frequenze è più sbilanciata verso sinistra.

```
apply(df,2,mean)
```

```
## LSE/NT    LSM DIP2-3 DIP4-5   L/PL
##   4.71  32.15   6.51  36.82  19.82
```

```
apply(df,2,median)
```

```
## LSE/NT    LSM DIP2-3 DIP4-5   L/PL
##   3.50  31.75   5.85  36.60  19.90
```

I valori delli media e mediana ci confermano ciò che avevamo supposto guardando i poxplot. Infatti per la variabile LSE/NT la media è sensibilmente maggiore della mediana, mentre per le altre la media è leggermente superiore della mediana. La terza statistica utilizzata per descrivere la centralità di una distribuzione di dati è la moda campionaria.

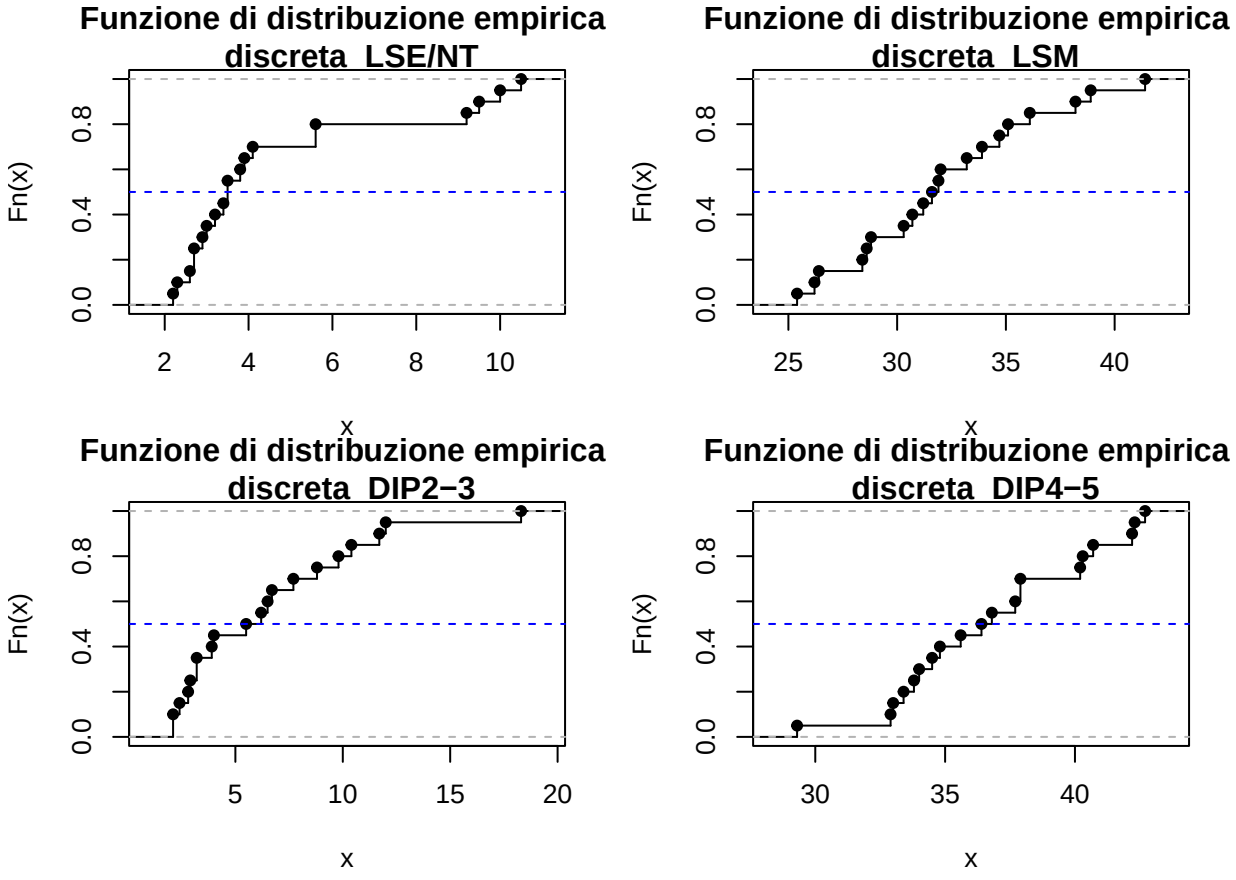
3.3.3 MEDIANA PER UNA DISTRIBUZIONE DI FREQUENZE

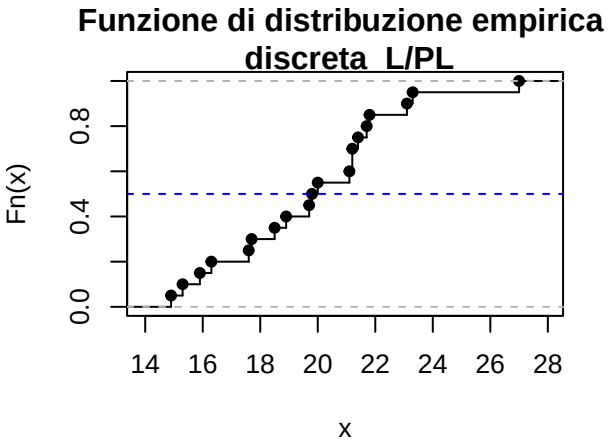
Un altro modo per definire la mediana consiste nel considerare le frequenze relative cumulate. La mediana per una distribuzione di frequenze è definita come la modalità i-esima (i=1,2,..,k) per cui vale la doppia disuguaglianza

$$F_{i-1} < 0.5 \qquad F_i >= 0.5$$

La mediana per una distribuzione di frequenze rappresenta sempre un valore realmente osservato, a differenza della mediana campionaria vista in precedenza.

La mediana di una distribuzione di frequenze può essere individuata graficamente a partire dalla funzione di distribuzione empirica discreta. Si traccia la funzione di distribuzione empirica e sull'asse delle ordinate si individua il punto 0.5 e da questo si traccia una linea orizzontale. Il minimo valore osservato (sulle ascisse) la cui funzione di distribuzione empirica supera 0.5 è proprio la mediana per una distribuzione di frequenze.





Utilizzando questa definizione si ottengono le seguenti mediane:

```
## LSE/NT    LSM DIP2-3 DIP4-5   L/PL
##    3.5   31.6    5.5   36.4   19.8
```

3.3.4 QUANTILI

I quantili sono indici che dividono un insieme ordinato di dati in numero fissato di gruppi di uguale grandezza. Supponiamo di avere un campione numerico di n osservazioni, si chiamano quantili di ordine α gli α - 1 numeri che servono per generare α gruppi di uguale grandezza. I quantili prendono il nome di quartili se suddiviamo in 4 parti, decili in 8, percentili in 100, ecc... Esistono 9 algoritmi diversi in R per calcolare i quantili attraverso la funzione

$$quantile(v, probs =, type = j), \qquad j = 1, 2, .., 9.$$

Quello di default è il 7. I risultati dipendono dall'algoritmo utilizzato. Il 2 e il 7 forniscono lo stesso valore della mediana. Il 2 generalizza la definizione di mediana campionaria, il 7 si basa su interpolazione di dati, mentre l'1 generalizza la definizione di mediana per una distribuzione di frequenze. Per grandi campioni i risultati sono pressapoco gli stessi.

3.3.4.1 QUANTILI CON ALGORITMO DI TIPO 2 Per calcolare il percentile k-esimo (k = 0, 1, ... , 100) P_k con l'algoritmo di tipo 2 si utilizza la seguente procedura.

- STEP 1: Ordinare i dati del campione di ampiezza n in ordine crescente (dal valore più piccolo al valore più grande) e sia v il vettore ordinato;

- STEP 2: Calcolare l'indice h
$$h = np = n\frac{k}{100},$$
in cui P_k è il percentile di interesse e n è il numero di osservazioni (ampiezza del campione);
- STEP 3: Se h = np è un intero, il percentile k-esimo si ottiene effettuando la media aritmetica dei valori nelle posizioni h e h+1 nell'insieme dei dati ordinati, ossia P_k = (v[h] + v[h + 1])/2; Se h = np non è un intero, si arrotonda h = np per eccesso al primo intero successivo ottenendo
$$h^* = ceiling(h).$$
Il percentile k-esimo è quello che corrisponde alla posizione h^*, ossia $P_k = v[h^*]$.

Questo algoritmo generalizza il concetto di mediana, ottenibile ponendo p = 0.5 e k = 50. Infatti, in questo caso se n è pari (h = n/2 è intero) si effettua la media aritmetica dei valori in posizione n/2 e n/2 + 1, mentre se n è dispari (h = n/2 non è intero) si arrotonda h = n/2 per eccesso e la mediana corrisponde al valore in posizione (n + 1)/2. L'algoritmo è implementato in R scegliendo j = 2 e usando la funzione
$$quantile(v, probs =, type = 2)$$

```
apply(df, 2, quantile, probs = c(0,0.25,0.5,0.75,1), type = 2)
```

```
##      LSE/NT   LSM DIP2-3 DIP4-5  L/PL
## 0%      2.2 25.40   2.10  29.30 14.90
## 25%     2.8 28.70   3.05  33.90 17.65
## 50%     3.5 31.75   5.85  36.60 19.90
## 75%     5.6 34.90   9.30  40.25 21.55
## 100%   10.5 41.40  18.30  42.70 27.00
```

3.3.5 VARIANZA, DEVIAZIONE STANDARD E COEFFICIENTE DI VARIAZIONE

Gli indici di posizione non tengono conto della variabilità dei dati; infatti esistono distribuzioni di frequenze che sono molto diverse tra loro, pur avendo la stessa media campionaria.

Indici significativi per misurare la variabilità di una distribuzione di frequenze sono la varianza campionaria e la deviazione standard campionaria, detta anche scarto quadratico medio campionario.

Assegnato un insieme di dati numerici $x_1, x_2, ..., x_n$, si definisce varianza campionaria e si denota con s^2 la quantità
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (n = 2, 3, ...)$$
mentre la deviazione standard campionara non è altro che la radice quadrata della varianza
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (n = 2, 3, ...)$$

Varianza campionaria e deviazione standard campionaria sono detti indici di dispersione o indici di variabilità poiché misurano la dispersione dei dati intorno alla media, più sono grandi più c'è dispersione.

I valori di queste due misure dipendono dalla misura dei dati. Inoltre la varianza è una misura al quadrato, mentre la deviazione standard misura la dispersione nella stessa unità di misura dei dati.

La varianza non gode della stessa proprietà di linearità della media campionaria, ma viene persa la costante b. Cioè se ho $y_i = ax_i + b$ (i=1,2,..,n) allora la varianza di y è ${s_y}^2 = a^2{s_x}^2$.

Per confrontare le variazioni esistenti tra diversi campioni di dati è utile introdurre il coefficiente di variazione.

$$CV = \frac{s}{|\bar{x}|}$$

Il coefficiente di variazione è una misura che mette insieme media e deviazione standard, perché hanno le stesso unità di misura. Quindi è un numero puro, senza unità di misura.

Il coefficiente di variazione è utilizzato quando è necessario confrontare tra loro le dispersioni (variabilità) di insiemi di dati espressi in differenti unità di misura (peso, altezza, redditi, . . .) oppure insiemi di dati aventi differenti range di variazione (il range di variazione è dato dalla differenza tra il massimo e il minimo dei dati). Ad esempio, la deviazione standard di un campione di redditi espressi in Dollari è completamente diversa della deviazione standard degli stessi redditi espressi in Euro, mentre il coefficiente di variazione è lo stesso in entrambi i casi.

Se il coeff. di variazione è molto piccolo (< 1) la deviazione standard è molto più piccola rispetto alla media, per cui i dati sono più raggruppati intorno alla media. Se il coeff. è grande (> 1), la dev. standard è grande rispetto alla media, quindi la media è poco rappresentativa dei dati (Quindi potrebbe essere meglio la mediana ad esempio).

Per confrontare piu insiemi di dati con medie differenti non mi basta solo la dev. standard ma mi serve il coeff. di variazione, che tiene conto sia della dev. std che della media.

```
var <- apply(df, 2, var)
sd <- apply(df, 2, sd)

cv <- function(x){
  sd(x)/abs(mean(x))
}

coeffVar <- apply(df,2,cv)

table <- rbind(var,sd,coeffVar)
table
```

```
##             LSE/NT        LSM     DIP2-3     DIP4-5      L/PL
## var      7.6746316 19.0226316 18.2872632 13.7711579 9.1722105
## sd       2.7703125  4.3614942  4.2763610  3.7109511 3.0285658
## coeffVar 0.5881768  0.1356608  0.6568911  0.1007863 0.1528035
```

Possiamo vedere che la variabili DIP2-3 LSE/NT sono quelle con coefficiente di variazione più alto, quindi c'è una dispersione maggiore dei dati intorno alla media rispetto alle altre variabili, a conferma di quanto avevamo visto con i boxplot.

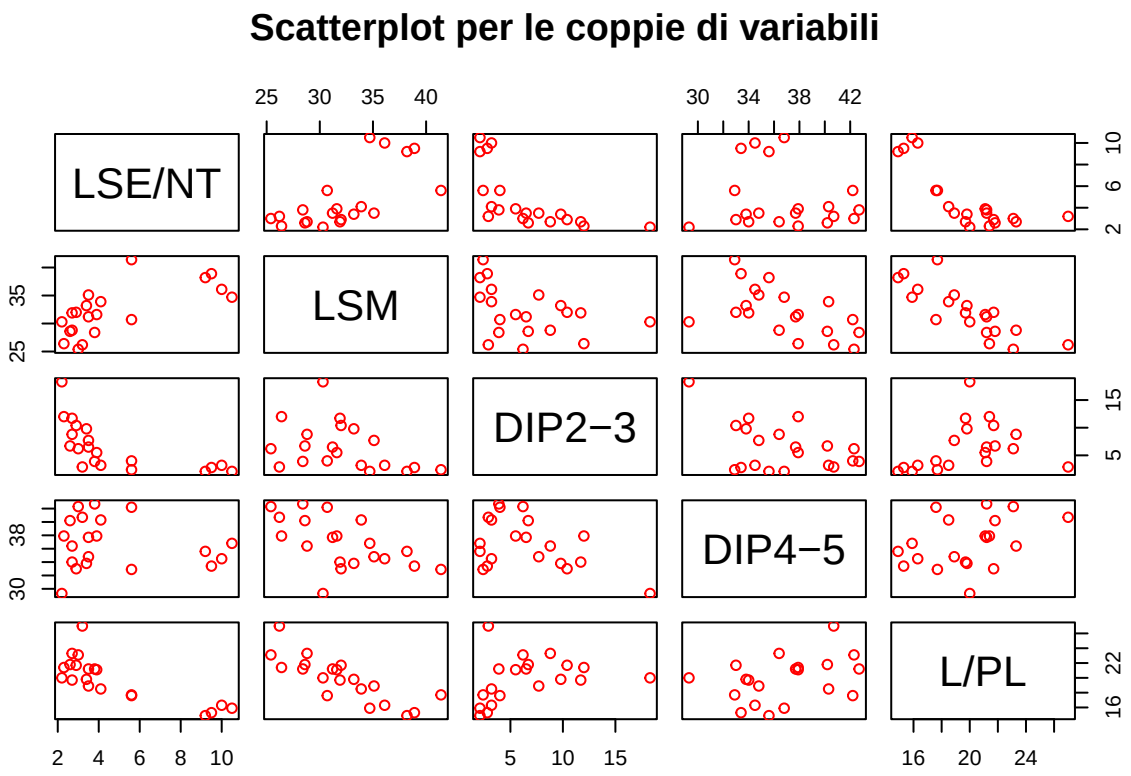
4 STATISTICA DESCRITTIVA BIVARIATA

La statistica descrittiva bivariata è il ramo della statistica che si occupa dei metodi grafici e statistici atti descrivere le relazioni che intercorrono tra due variabili quantitative. Le relazioni tra variabili quantitative possono essere rappresentate graficamente mediante diagrammi di dispersione (scatterplot) in cui ogni coppia di osservazioni viene rappresentata sotto forma di un punto o di un cerchietto in un piano euclideo. Dopo aver scelto la variabile da porre sulle ascisse (variabile indipendente) e la variabile da porre sulle ordinate (variabile dipendente), si disegnano dei punti in corrispondenza delle coppie (x_i, y_i). Il risultato finale è una nuvola di punti che può essere ottenuto con la funzione plot(x, y), dove x è il vettore contenente i valori $(x_1, x_2, ..., x_n)$ e y è il vettore contenente i valori $(y_1, y_2, ..., y_n)$. Il grafico che si ottiene mira ad evidenziare

se le coppie di punti presentano qualche forma di regolarità. Inoltre, il grafico di dispersione mostra se esiste una relazione tra le variabili e di quale tipo è tale relazione (lineare, quadratica, . . .).

Per prima cosa quindi realizziamo lo scatterplot per indagare l'eventuale relazione tra due variabili. Poi per ottenere una misura quantitativa di tale relazione ci sono la covarianza e la correlazione campionaria.

La funzione pairs() è in grado di visualizzare in un'unica finestra grafica una pluralità di scatterplot ottenuti mettendo in relazione tutte le coppie di variabili quantitative definite all'interno di un data frame.



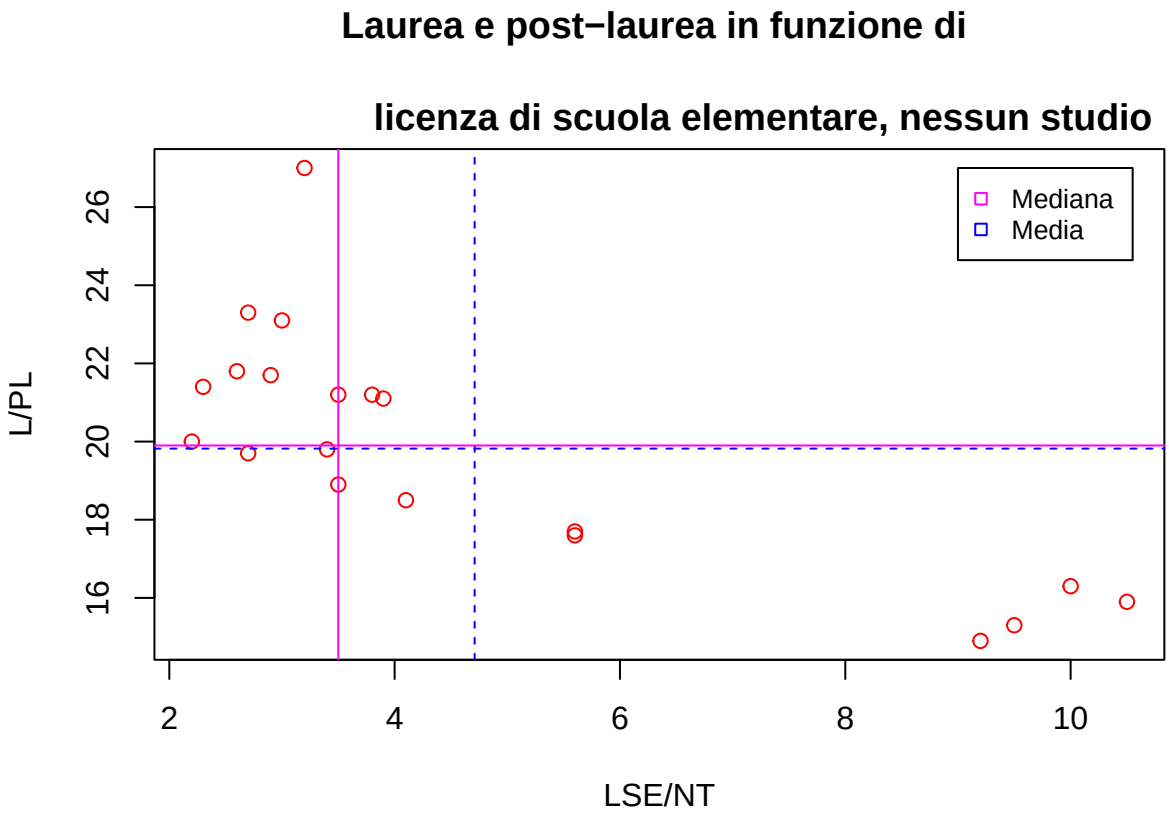
Dagli scatterplot ottenuti ho notato che sembra esserci una maggiore relazione tra la varibile L/PL e LSE/NT, e tra L/PL e LSM. Ho quindi proceduto ad analizzare nel dettaglio la relazione tra:

- il titolo di studio "laurea e post-laurea" (L/PL), la variabile dipendente, e il titolo di studio "Licenza di scuola elementare, nessun titolo" (LSE/NT).
- il titolo di studio "laurea e post-laurea" (L/PL), la variabile dipendente, e il titolo di studio "Licenza di scuola media" (LSM).

Ho quindi realizzato gli scatterplot in cui vengono tracciate delle linee orizzontali e verticali in corrsipondenza delle medie e mediane campionare delle due variabili.

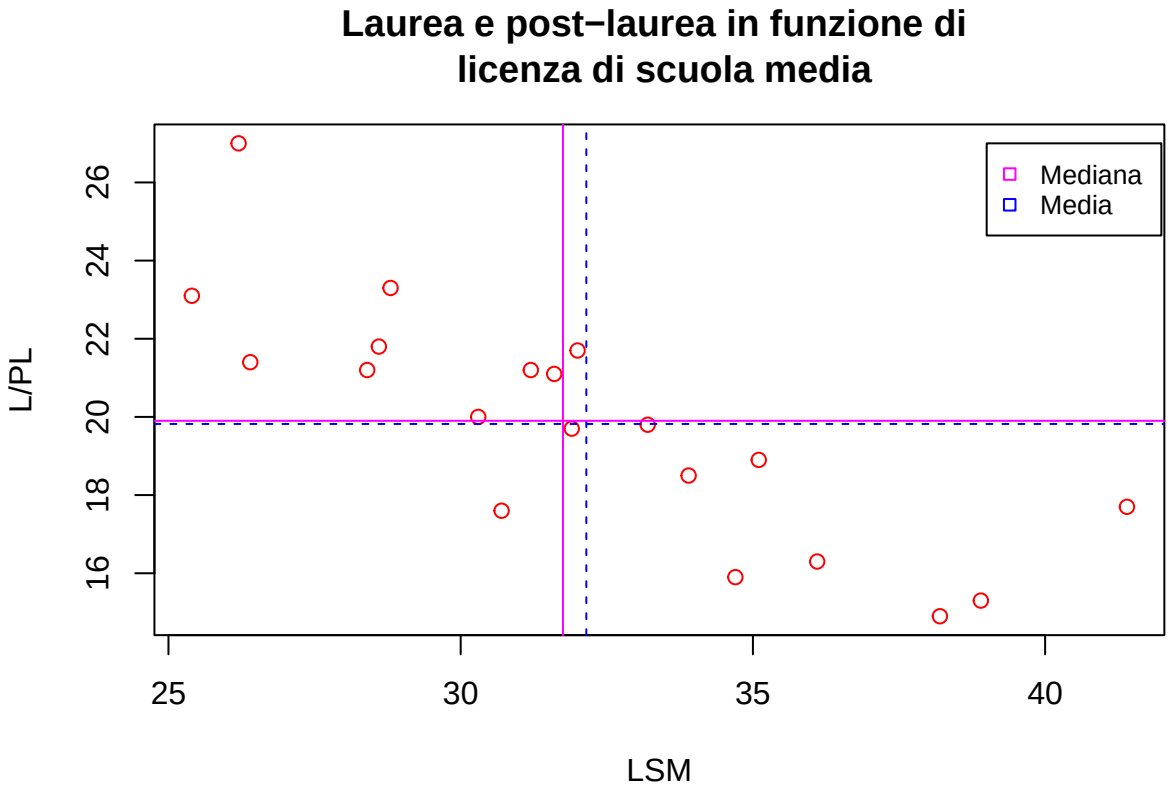
```
plot(df$`LSE/NT`, df$`L/PL`, xlab = "LSE/NT", ylab = "L/PL", col = "red",
     main = paste("Laurea e post-laurea in funzione di \n
                  licenza di scuola elementare, nessun studio" ))
abline (v = median (df$`LSE/NT`) , lty =1 , col =" magenta ")
abline (v = mean (df$`LSE/NT`) , lty =2 , col =" blue ")
abline (h = median ( df$`L/PL` ) , lty =1 , col =" magenta ")
abline (h = mean ( df$`L/PL` ) , lty =2 , col =" blue ")
```

```
legend (9 ,27 , c (" Mediana " ," Media " ) , pch =0 ,
        col =c( " magenta " ," blue ") , cex =0.8)
```



Si nota che i dati sembrano posizionarsi intorno ad una retta discendente e ciò induce a pensare che esista una correlazione lineare negativa tra le variabili.

```
plot(df$`LSM`, df$`L/PL`, xlab = "LSM", ylab = "L/PL", col = "red",
     main = paste("Laurea e post-laurea in funzione di \n licenza di scuola media" ))
abline (v = median (df$`LSM`) , lty =1 , col =" magenta ")
abline (v = mean (df$`LSM`) , lty =2 , col =" blue ")
abline (h = median ( df$`L/PL` ) , lty =1 , col =" magenta ")
abline (h = mean ( df$`L/PL` ) , lty =2 , col =" blue ")
legend (39 ,27 , c (" Mediana " ," Media " ) , pch =0 ,
        col =c( " magenta " ," blue ") , cex =0.8)
```



Anche in questo caso i dati sembrano posizionarsi intorno ad una retta discendente e ciò induce a pensare che esista una correlazione lineare negativa tra le variabili.

4.1 COVARIANZA E CORRELAZIONE CAMPIONARIA

Covarianza e correlazione campionaria sono indici che servono a definire relazioni di tipo lineare tra variabili. In particolare sono misure quantitative della correlazione tra due variabili.

Prese due variabili quantitavie X e Y, e dato un campione $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$ costituito da n osservazioni di (X,Y), e siano $\bar{x}$ e $\bar{y}$ le medie campionarie di X e Y, si definisce covarianza campionaria tra X e Y la quantità :

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad (n = 2, 3, ...)$$

Di norma si usa normalizzare tale sommatoria dividendo per n - 1, in maniera tale da ottenere la varianza campionaria nel caso in cui $x_i = y_i$ per ogni i = 1, 2, . . . , n. La covarianza campionaria può avere segno positivo, negativo o nullo. Quando $C_{xy} > 0$ si dice che le variabili sono correlate positivamente, se $C_{xy} < 0$ le variabili sono correlate negativamente e, infine, se $C_{xy} = 0$ le variabili sono non correlate.

Un'altra misura quantitativa della correlazione tra variabili si può ottenere utilizzando la correlazione campionaria, che è definita nel seguente modo:

$$r_{xy} = \frac{C_{xy}}{s_x s_y}$$

Questo coefficiente è adimensionale; non fa distinzioni tra variabile dipendente e indipendente, ossia $r_{xy} = r_{yx}$ (lo stesso vale per la covarianza); è indipendente dall'unità di misura dei dati; è influenzata dai valori anomali.

Il coeff. di correlazione ha lo stesso segno della covarianza. In particolare gode delle seguenti proprietà:

- $-1 <= r_{xy} <= 1$
- se esistono due numeri reali a e b, con a > 0, tali che $y_i = ax_i + b$ per ogni i = 1, 2, . . . , n, allora $r_{xy} = 1$;
- se esistono due numeri reali a e b, con a < 0, tali che $y_i = ax_i + b$ per ogni i = 1, 2, . . . , n, allora r_{xy} = -1;
- se esistono quattro numeri reali a,b,c,d e se risulta $z_i = ax_i + b$ e $w_i = cy_i + d$ per i = 1,2,. . . ,n , allora $r_{zw} = r_{xy}$ se ac > 0 e $r_{zw} = -r_{xy}$ se ac < 0

La (2) e (3) ci dicono che la correlazione assume i valori estremi -1 e 1 quando esiste una relazione lineare tra le variabili X e Y, ossia quando i punti dello scatter plot giacciono tutti sulla stessa linea.

Occorre ricordare che il coefficiente di correlazione campionario r_{xy} misura la forza del legame di natura lineare esistente tra due variabili quantitative. Eventuali relazioni tra le variabili che assumono una forma curvilinea non possono pertanto essere individuati con tale coefficiente. Riassumendo si può dire che il segno di r_{xy} indica la direzione della retta interpolante e indica la presenza di una tra le seguenti situazioni:

- $r_{xy} = 1$ (correlazione perfetta positiva): tutti i punti sono allineati su una linea retta ascendente;
- r_{xy} compreso tra 0 e 1 estremi esclusi (correlazione positiva): i punti (x_i, y_i) sono posizionati in una nuvola attorno ad una linea retta interpolante ascendente (in tal caso x_i e y_i tendono ad essere grandi e piccoli insieme);
- $r_{xy} = 0$ (nessuna correlazione): i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare;
- r_{xy} compreso tra -1 e 0 estremi esclusi (correlazione negativa): i punti (x_i, y_i) sono posizionati in una nuvola attorno ad una linea retta interpolante discendente (in tal caso x_i è grande y_i è piccolo e viceversa);
- r_{xy} = -1 (correlazione perfetta negativa): tutti i punti sono allineati su una linea retta discendente;

Di seguito sono riportate la matrice delle covarianze, in cui sono riportate le covarianze tra tutte le possibili coppie di variabili (sulla diagonale principale abbiamo le varianze), e la matrice delle correlazioni, in cui sono riportati i coefficienti di correlazione tra tutte le possibili coppie di variabili(sulla diagonale principale tutti 1).

```
cov(df)
```

```
##           LSE/NT        LSM    DIP2-3    DIP4-5       L/PL
## LSE/NT  7.674632   8.164211 -7.702737 -1.506526  -6.630737
## LSM     8.164211  19.022632 -7.312105 -9.050000 -10.843158
## DIP2-3 -7.702737  -7.312105 18.287263 -7.374421   4.151895
## DIP4-5 -1.506526  -9.050000 -7.374421 13.771158   4.141158
## L/PL   -6.630737 -10.843158  4.151895  4.141158   9.172211
```

```
cor(df)
```

```
##            LSE/NT        LSM     DIP2-3     DIP4-5       L/PL
## LSE/NT  1.0000000  0.6756941 -0.6501925 -0.1465422 -0.7903073
## LSM     0.6756941  1.0000000 -0.3920422 -0.5591496 -0.8208872
## DIP2-3 -0.6501925 -0.3920422  1.0000000 -0.4646954  0.3205789
## DIP4-5 -0.1465422 -0.5591496 -0.4646954  1.0000000  0.3684678
## L/PL   -0.7903073 -0.8208872  0.3205789  0.3684678  1.0000000
```

possiamo notare che il coefficiente di correlazione tra L/PL e LSE/NT, e tra L/PL e LSM è compreso tra 0 e -1, quindi c'è una correlazione negativa in entrambi i casi. Mentre sia tra L/PL e DIP2-3 che tra L/PL e DIP4-5 il coefficiente di correlazione è compreso tra 0 e 1, quindi ci sono delle correlazioni positive.

Gli scatterplot sono dei potenti mezzi per visualizzare le eventuali relazioni che possono intercorrere tra variabili quantitative, a volte osservando il grafico si nota che i punti si dispongono attorno a qualche linea orientata in qualche direzione.

Tuttavia per avere un'idea più accurata del fenomeno, è necessario utilizzare altre tecniche statistiche in grado di misurare con maggiore precisione questo legame, come la regressione.

4.2 REGRESSIONE LINEARE SEMPLICE

Il modello lineare viene di solito utilizzato per spiegare, descrivere, o anche prevedere un andamento futuro sulla base della relazione che si instaura tra una variabile Y, chiamata variabile dipendente, e una o più altre variabili che assumono il significato di variabili indipendenti X1,X2, . . . ,Xp. Nel caso in cui p = 1, l'analisi prende il nome di regressione semplice, mentre se p = 2, 3 . . . si parla di regressione multipla. Per poter utilizzare un modello di regressione è fondamentale individuare in primo luogo quali sono le variabili indipendenti e quale è la variabile dipendente. La regressione è una tecnica statistica che si utilizza per modellare le relazione tra una variabile dipendente ed una o più variabili indipendenti. Il modello di regressione lineare semplice è esprimibile attraverso l'equazione di una retta che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte e altre possibili rette.

$$Y = \alpha + \beta X$$

α è l'intercetta e β è il coefficiente angolare della retta.

Il coefficiente angolare β esprime quantitativamente la pendenza (inclinazione) della retta: un coefficiente angolare positivo ($\beta > 0$) indica una retta di regressione crescente, un coefficiente angolare negativo ($\beta < 0$) indica una retta decrescente; un coefficiente angolare nullo ($\beta = 0$) indica una retta orizzontale. L'intercetta α invece corrisponde all'ordinata del punto di intersezione della retta interpolante (di regressione) con l'asse delle ordinate.

L'identificazione di questa retta viene ottenuta applicando il metodo dei minimi quadrati. I coefficienti di regressione sono i valori α e β per i quali la somma Q dei quadrati degli errori

$$Q = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

sia minima. Il minimo si ottiene calcolando le derivate parziali rispetto ad α e rispetto a β e ponendole a 0, da cui si ottiene poi un sistema di equazioni da cui andiamo a ricavare poi i valori di α e β.

Dal metodo dei minimi quadrati si ottiene che

$$\beta = \frac{s_y}{s_x} r_{xy} \qquad \alpha = \bar{y} - \beta \bar{x}$$

In R funzione lm(y∼x), linear model, per eseguire analisi di regressione e fornisce in automatico i valori di α e β. y∼x sta ad indicare y in funzione di x, ossia y è la variabile dipendente e x quella indipendente.

- **"Laurea e post-laurea" in funzione di "Licenza scuola elementare, nessun titolo di studio"**

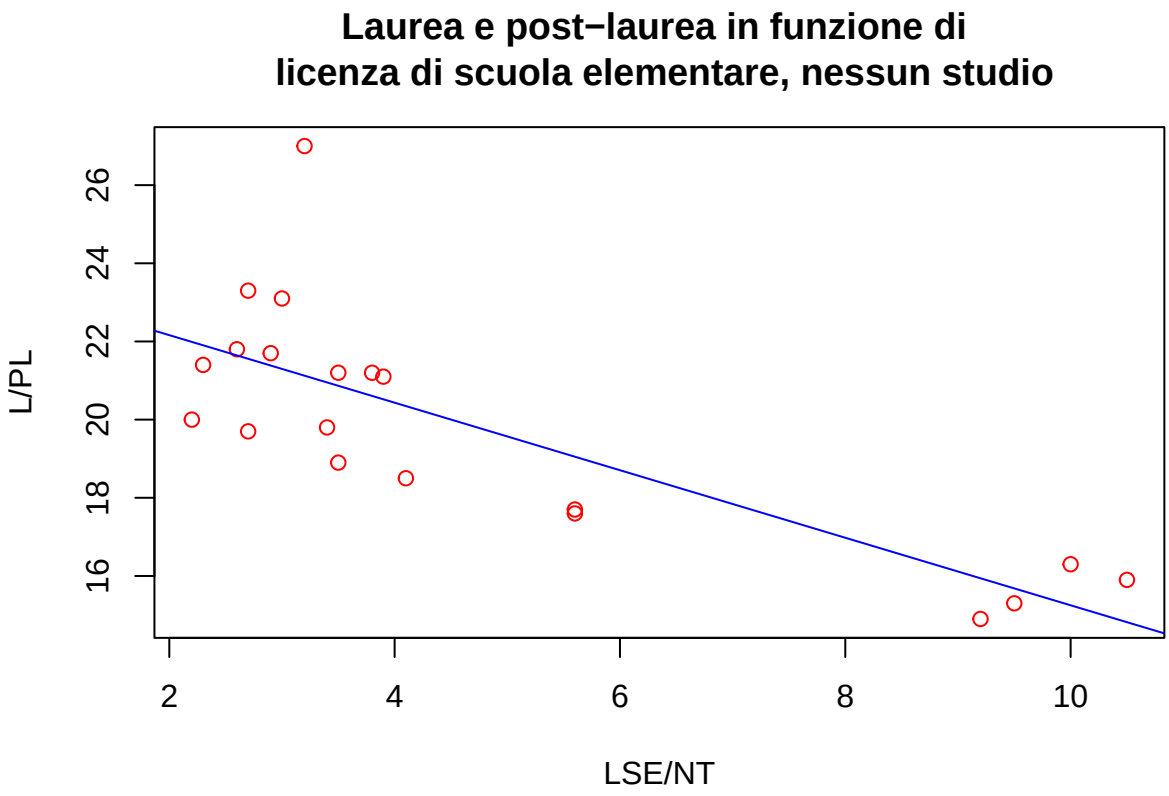
```
linearModel1 <- lm(df$`L/PL` ~ df$`LSE/NT`)
linearModel1
```

```
##
## Call:
## lm(formula = df$`L/PL` ~ df$`LSE/NT`)
##
## Coefficients:
## (Intercept)  df$`LSE/NT`
##      23.889       -0.864
```

La retta di regressiona ha quindi equazione

$$y = 23.889 + -0.864x$$

La rappresentazione della retta così calcolata può essere aggiunta allo scatterplot facendo uso della funzione abline(lm(df$L/PL ~ df$LSE/NT))



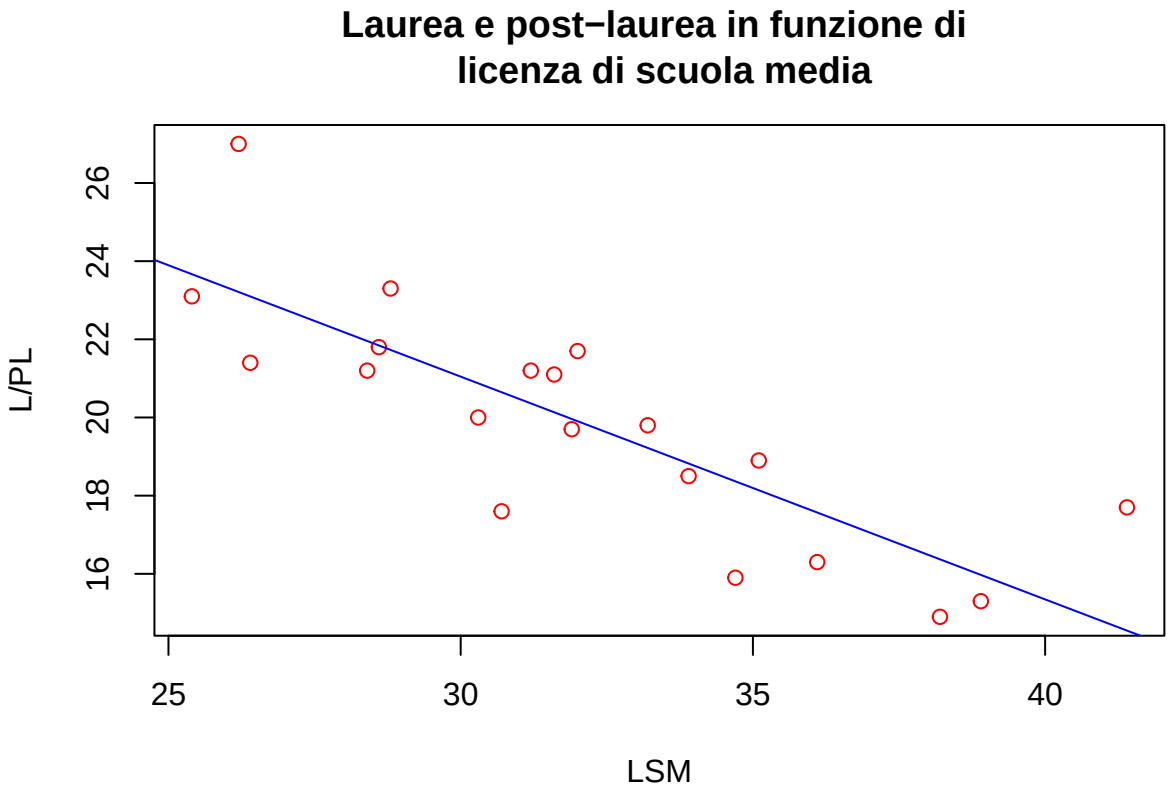
- **"Laurea e post-laurea" in funzione di "Licenza scuola media"**

```
linearModel2 <- lm(df$`L/PL` ~ df$`LSM`)
linearModel2
```

```
##
## Call:
## lm(formula = df$`L/PL` ~ df$LSM)
##
## Coefficients:
## (Intercept)       df$LSM
##       38.15        -0.57
```

La retta di regressione ha equazione

$$y = 38.15 - 0.57x$$



Residui

Una volta calcolati i valori dei coefficienti α e β e disegnata la retta di regressione che interpola la nuvola dei punti nel corrispondente scatterplot,è possibile osservare quanto questa retta si adatta ai punti che individuano le osservazioni. In particolare i valori stimati mediante la retta di regressione sono

$$\hat{y}_i = \alpha + \beta x_i \qquad (i = 1, 2, ..., n)$$

La media campionaria dei valori stimati $(\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ è uguale alla media campionaria $\bar{y}$ dei valori osservati $(y_1, y_2, ..., y_n)$

I residui sono invece la differenza tra i valori osservati e quelli stimati con la retta di regressione, e sono cosi definiti

$$E_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i) \qquad (i = 1, 2, ..., n)$$

La media campionaria dei residui è nulla, ossia in media gli scostamenti positivi e negativi si compensano.

La varianza campionaria dei residui invece è

$$s_E^2 = \frac{1}{n-1}\sum_{i=1}^{n}(E_i - \bar{E})^2 = \frac{1}{n-1}\sum_{i=1}^{n}E_i^2$$

In R per calcolare il vettore dei valori stimati si usa la funzione fitted(lm(y∼x)) con y che dipende da x, mentre per il vettore dei valori residui si usa la funzione resid(lm(y∼x))

La media dei residui è nulla. Invece, la mediana, la varianza campionaria e la deviazione standard campionaria dei residui sono:

```
##                             median      var       sd
## linearModel1$residuals -0.1122653 3.443379 1.855634
## linearModel2$residuals -0.2924898 2.991464 1.729585
```

Di seguito sono riportati i vettori dei valori stimati ed i vettore dei residui per i due modelli

```
stime1 <- fitted(lm(df$`L/PL` ~ df$`LSE/NT`))
stime1
```

```
##        1        2        3        4        5        6        7        8
## 20.95182 20.86542 21.64300 21.38381 21.98859 21.55660 21.90219 21.55660
##        9       10       11       12       13       14       15       16
## 20.51982 21.29741 20.86542 21.12461 20.60622 20.34703 15.24954 15.68153
##       17       18       19       20
## 19.05106 14.81755 15.94072 19.05106
```

```
residui1 <- resid(lm(df$`L/PL` ~ df$`LSE/NT`))
residui1
```

```
##          1          2          3          4          5          6          7
## -1.1518153 -1.9654172  0.1569998  0.3161942 -1.9885926 -1.8566021 -0.5021945
##          8          9         10         11         12         13         14
##  1.7433979  0.5801753  1.8025923  0.3345828  5.8753885  0.5937772 -1.8470285
##         15         16         17         18         19         20
##  1.0504602 -0.3815304 -1.4510568  1.0824507 -1.0407247 -1.3510568
```

```
stime2 <- fitted(lm(df$`L/PL` ~ df$`LSM`))
stime2
```

```
##        1        2        3        4        5        6        7        8
## 19.22149 18.13846 21.84355 19.90550 20.87453 19.96250 23.09758 21.72955
##        9       10       11       12       13       14       15       16
## 20.13351 23.66759 20.36151 23.21158 21.95755 18.82248 17.56845 15.97241
##       17       18       19       20
## 20.64652 18.36647 16.37142 14.54737
```

```
residui2 <- resid(lm(df$`L/PL` ~ df$`LSM`))
residui2
```

```
##           1           2           3           4           5           6
##  0.57851424  0.76153999 -0.04354813  1.79449797 -0.87452508 -0.26250339
##           7           8           9          10          11          12
## -1.69757795  1.57045458  0.96649254 -0.56759151  0.83848712  3.78841933
##          13          14          15          16          17          18
## -0.75755084 -0.32247627 -1.26844645 -0.67240849 -3.04651966 -2.46646543
##          19          20
## -1.47141798  3.15262540
```

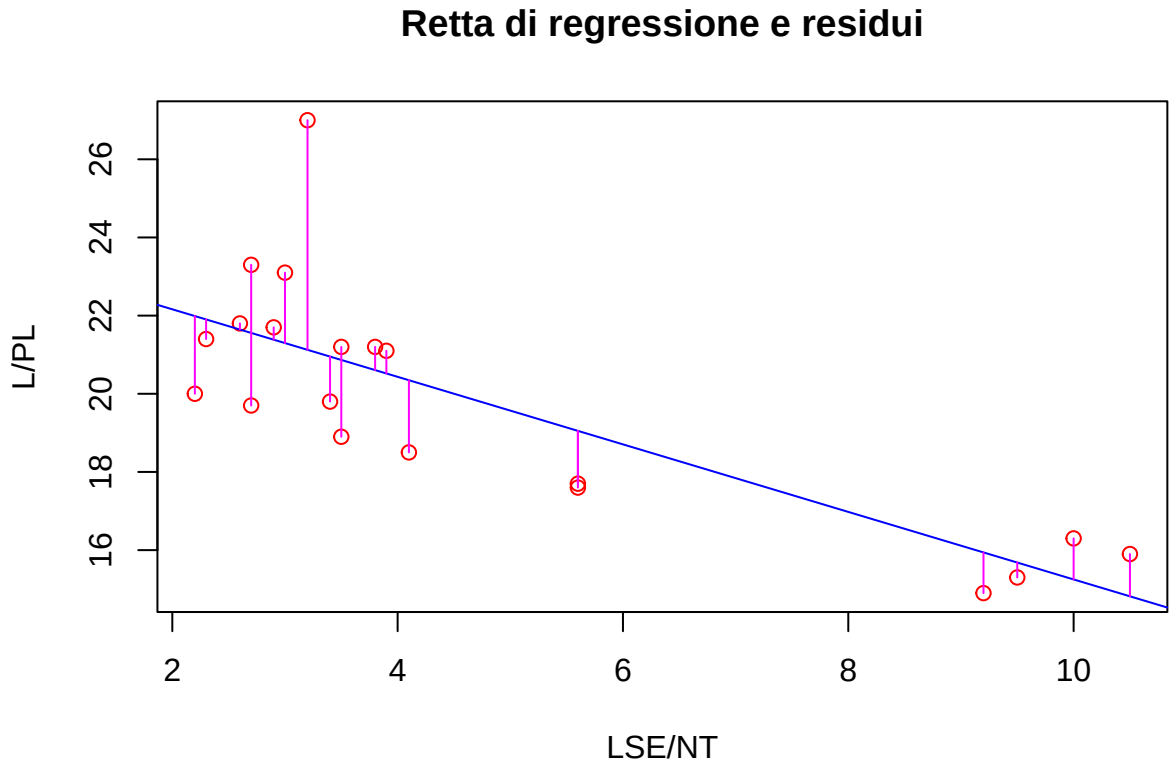
E possibile rappresentare graficamente i residui:

- tracciando dei segmenti verticali che congiungono i valori stimati $\hat{y}_i$ (sulla retta di regressione) e i valori osservati y_i (i = 1, 2,. . . , n);

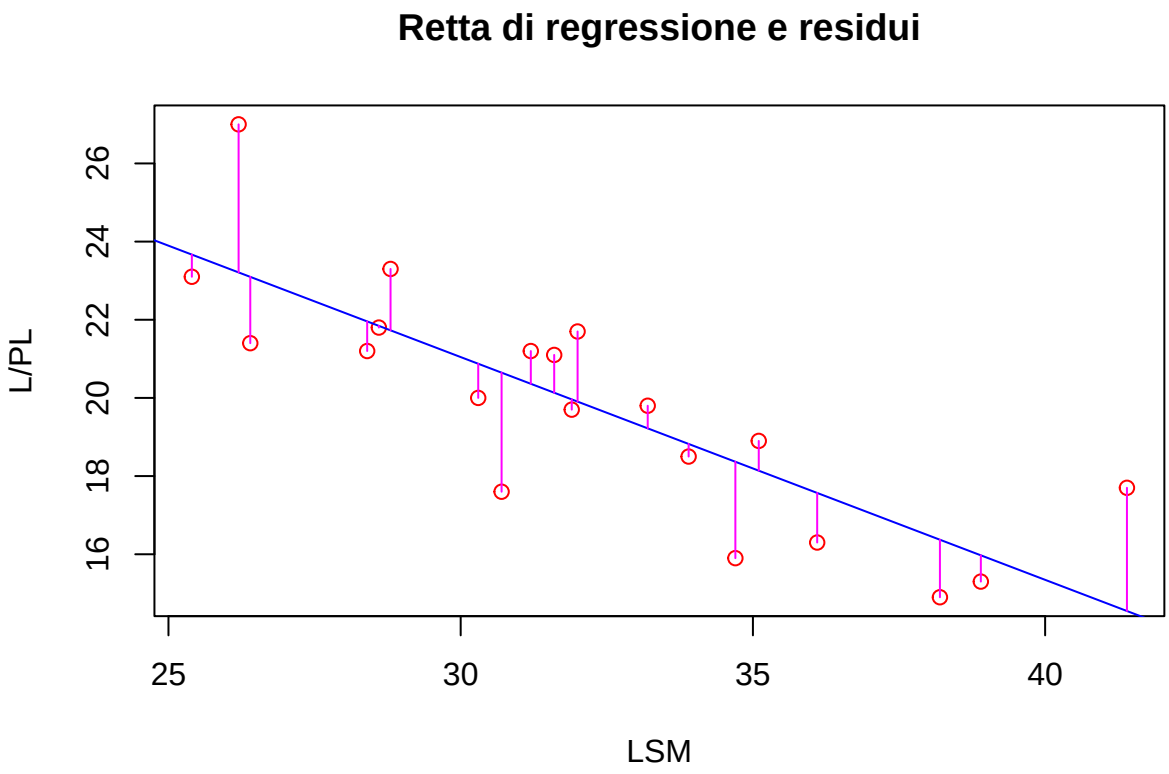
- rappresentando i valori dei residui E_i rispetto alle osservazioni x_i (variabile indipendente) (i = 1,2,. . . , n);
- rappresentando i residui standardizzati $E_i^{(s)} = E_i/s_E$ rispetto ai valori stimati $\hat{y}_i$ (i = 1, 2,. . . , n).

Segmenti che congiungono i valori stimati e i valori osservati

```
plot ( df$`LSE/NT`, df$`L/PL` , main =" Retta di regressione e residui " ,
       xlab = "LSE/NT" , ylab ="L/PL" , col =" red ")
abline ( lm ( df$`L/PL`~ df$`LSE/NT` ) , col =" blue ")
segments ( df$`LSE/NT` , stime1 , df$`LSE/NT` , df$`L/PL` , col =" magenta " )
```




```
plot ( df$`LSM`, df$`L/PL` , main =" Retta di regressione e residui " ,
       xlab = "LSM" , ylab ="L/PL" , col =" red ")
abline ( lm ( df$`L/PL`~ df$`LSM` ) , col =" blue ")
segments ( df$`LSM` , stime2 , df$`LSM` , df$`L/PL` , col =" magenta " )
```

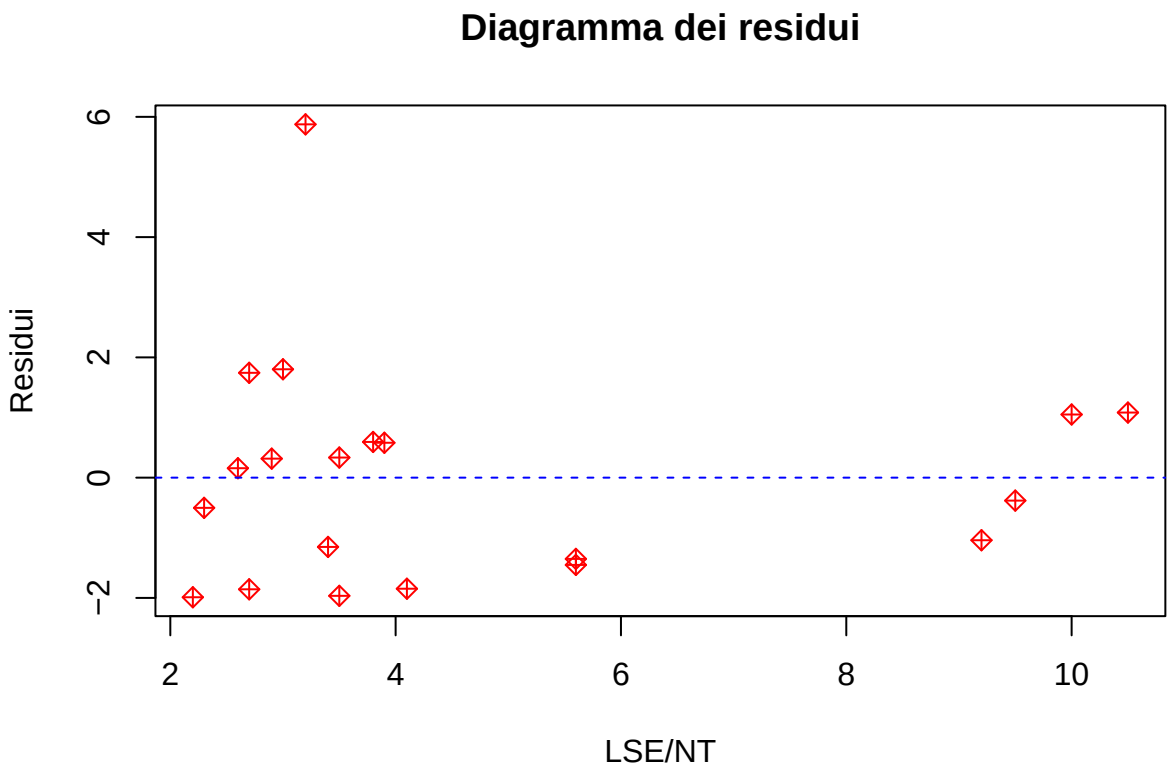


Valori dei residui rispetto alle osservazioni della variabile indipendente

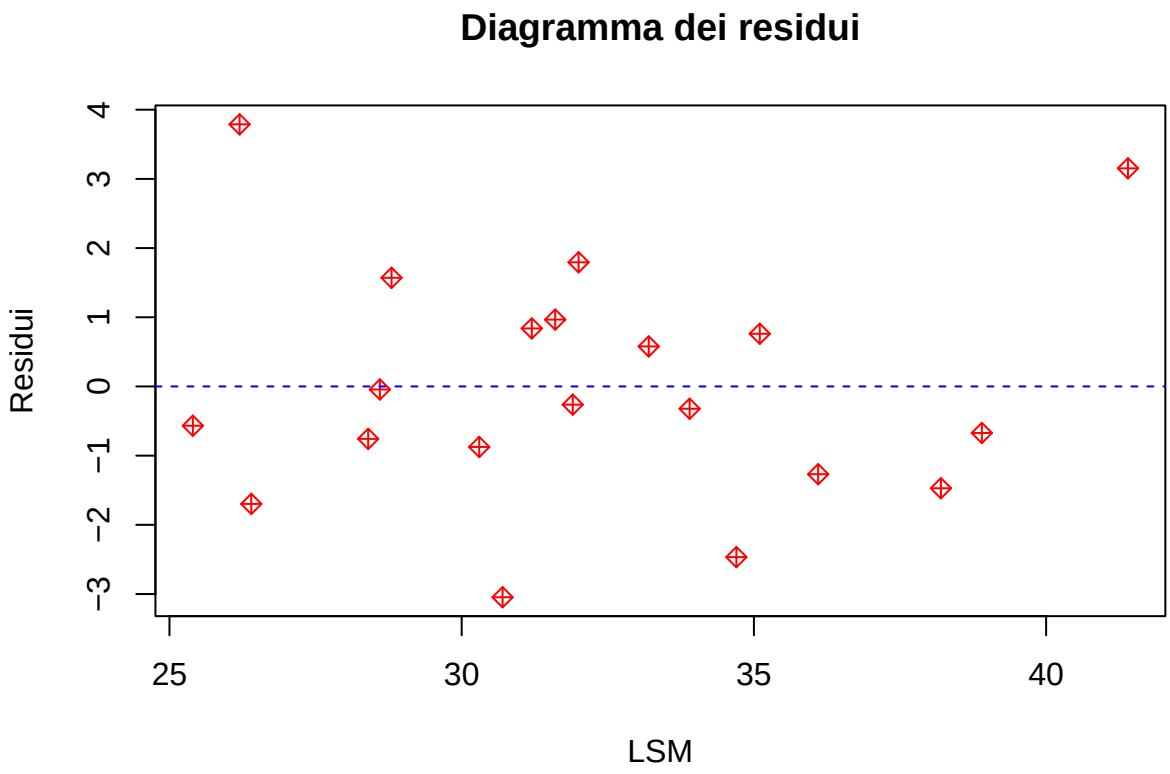
Un esame più accurato del modo con cui la retta di regressione interpola i dati e di come i residui si dispongano intorno alla retta interpolante influenzandone la posizione, può essere ottenuto attraverso il diagramma dei residui che è un grafico in cui i valori dei residui sono posti sull'asse delle ordinate e quelli della variabile indipendente sull'asse delle ascisse.

I punti indicano la posizione dove si collocano i residui rispetto ai valori della variabile indipendente. La retta orizzontale è posizionata nello zero e corrisponde alla media campionaria dei residui, che è nulla. Il diagramma dei residui aiuta a comprendere quale è l'adattamento della retta di regressione rispetto ai dati, consentendo di identificare quali sono le informazioni che hanno una forte influenza sulla collocazione e direzione della retta di regressione. Occorre notare che la posizione della retta di regressione è fortemente influenzata dalla presenza di eventuali valori anomali che si discostano in modo significativo dagli altri. L'analisi dei residui aiuta ad individuare eventuali punti isolati (valori anomali) che possono essere dovuti ad errori nella stima. Tali valori possono perturbare significativamente la stima dei parametri di regressione e influenzare l'interpretazione dei residui. Eliminando i valori anomali la varianza campionaria dei residui diminuisce.

```
plot (df$`LSE/NT`, residui1 , main =" Diagramma dei residui " ,
      xlab = "LSE/NT" , ylab =" Residui " , pch =9 , col =" red ")
abline ( h =0 , col =" blue " , lty=2)
```




```
plot (df$LSM, residui2 , main =" Diagramma dei residui " ,
      xlab = "LSM" , ylab =" Residui " , pch =9 , col =" red ")
abline ( h =0 , col =" blue " , lty=2)
```



Si nota che i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia nessun comportamento particolare nella distribuzione dei punti. Per il primo diagramma si nota un valore anomalo in alto e un gruppo di valori isolati a destra rispetto agli altri.

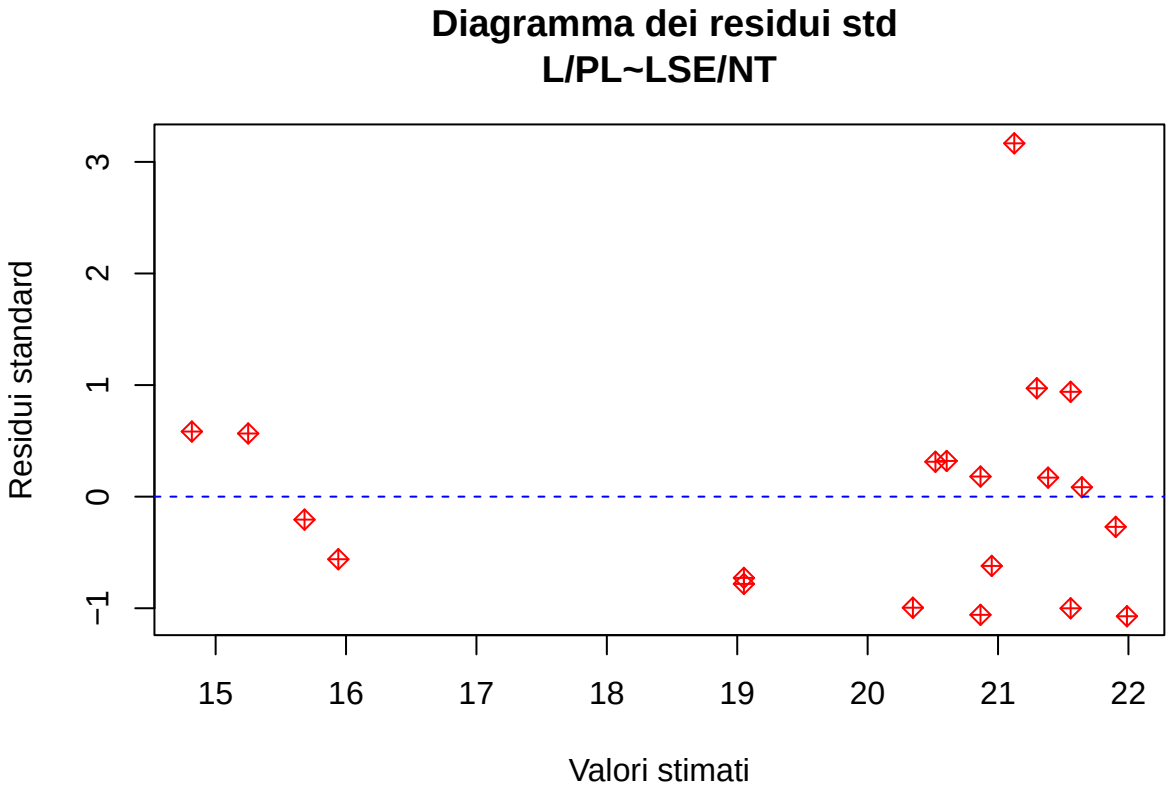
Valori dei residui standardizzati rispetto ai valori stimati

E spesso interessante calcolare i residui standardizzati così definiti:

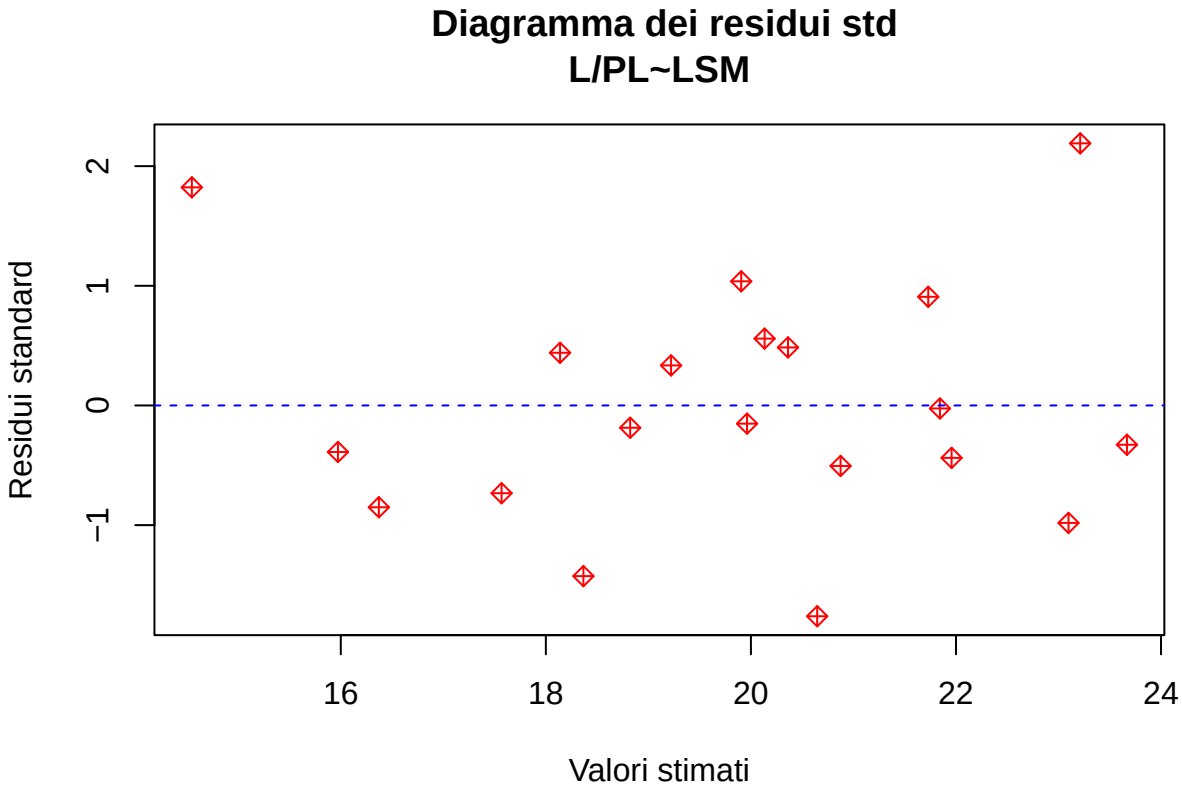
$$E_i^{(s)} = E_i - \bar{E}/s_E = E_i/s_E$$

che risultano essere caratterizzati da media campionaria nulla e varianza unitaria. E poi possibile realizzare un grafico in cui i residui standardizzati (ordinate) vengono disegnati in funzione dei valori stimati (ascisse) mediante la retta di regressione.

```
residuiStandard1 <- residui1/sd(residui1)
residuiStandard2 <- residui2/sd(residui2)

plot (stime1, residuiStandard1 , main =" Diagramma dei residui std\nL/PL~LSE/NT" ,
      xlab = "Valori stimati" , ylab =" Residui standard " , pch =9 , col =" red ")
abline ( h =0 , col =" blue " , lty=2)
```




```
plot (stime2, residuiStandard2 , main =" Diagramma dei residui std\nL/PL~LSM" ,
      xlab = "Valori stimati" , ylab =" Residui standard " , pch =9 , col =" red ")
abline ( h =0 , col =" blue " , lty=2)
```



I punti indicano la posizione dove si collocano i residui standardizzati rispetto ai valori stimati con la retta di regressione. La retta orizzontale è posizionata nello zero, che corrisponde alla media campionaria dei residui standardizzati. Anche in questo caso i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia nessuna tendenza particolare nella distribuzione dei punti.

4.2.1 COEFFICIENTE DI DETERMINAZIONE

Il coefficiente di determinazione è una misura di quanto la retta di regressione si adatta bene ai dati. Il coefficiente di determinazione (detto anche r-square) per la regressione semplice è il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati.

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Nel caso di regressione lineare semplice, il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione, ossia

$$D^2 = r_{xy}^2$$

Quindi r_{xy}^2 molto vicino a 1 indicherà che tutti i punti tenderanno ad allinearsi lungo la retta di regressione, mentre r_{xy}^2 molto vicino a 0 vuol dire che la retta è incapace di rappresentare i dati correttamente.

In R, nel caso di regressione lineare semplice, il coefficiente di determinazione D^2 si può calcolare utilizzando il quadrato del coefficiente di correlazione oppure la funzione summary(lm(y~x))$r.square.

```
summary(lm (df$`L/PL`~ df$`LSE/NT`))$r.square
```

```
## [1] 0.6245857
```

```
summary(lm (df$`L/PL`~ df$`LSM`))$r.square
```

```
## [1] 0.6738558
```

Quindi possiamo notare che per le due analisi di regressione abbiamo un coefficiente abbastanza buono, per cui le rette di regressione si adattano discretamente bene ai dati.

4.3 REGRESSIONE LINEARE MULTIPLA

Il modello di regressione lineare multipla viene utilizzato per spiegare la relazione tra una variabile quantitativa Y , detta variabile dipendente, e le variabili quantitative indipendenti $X_1, X_2, ..., X_p$.

Il modello di regressione lineare multipla con p variabili indipipendenti è esprimibile mediante l'equazione

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

dove:

- α è l'intercetta, ossia il valore di Y quando $X_1 = X_2 = ... = X_p = 0$;
- $\beta_1, \beta_2, ..., \beta_p$ sono i regressori. In particolare β_1 rappresenta l'inclinazione di Y rispetto alla variabile X_1 tenendo costanti le variabili $X_2, X_3, ..., X_p$, discorso analogo per gli altri regressori.

Per determinare le stime di $\alpha, \beta_1, \beta_2, ..., \beta_p$ utilizziamo sempre il metodo dei minimi quadrati. Minimizzando quindi la quantità

$$Q = \sum_{i=1}^{n} [y_i - (\alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p})]^2$$

Derivando rispetto ai parametri $\alpha, \beta_1, \beta_2, ..., \beta_p$ si ottiene questa volta un sistema di p + 1 equazioni, da cui si ricavano poi i coefficienti.

In R per eseguire l'analisi di regressione lineare multipla si usa $lm(y \sim x_1 + x_2 + ... + x_p)$. L'argomento passato ad lm indica che y dipende da $x_1, x_2, ..., x_p$.

Nel mio caso ho eseguito una regressione lineare multipla con Y = "L/PL" in funzione di x_1="LSE/NT" e x_2="LSM" visto che "DIP2-3" e "DIP4-5" hanno una bassa relazione lineare con "L/PL" (come evidenziato dagli scatterplot e dai coefficienti di correlazione).

```
multLinearModel <- lm( df$`L/PL` ~ df$`LSE/NT` + df$LSM)
multLinearModel
```

```
##
## Call:
## lm(formula = df$`L/PL` ~ df$`LSE/NT` + df$LSM)
##
## Coefficients:
## (Intercept)  df$`LSE/NT`       df$LSM
##     33.8378      -0.4740      -0.3666
```

Pertanto l'intercetta è α=33.8378 e i due regressori sono β_1=-0.4740 e β_2=-0.3666. Quindi il modello di regressione lineare multiplo stimato è

$$y = 33.8378 - 0.4740 x_1 - 0.3666 x_2$$

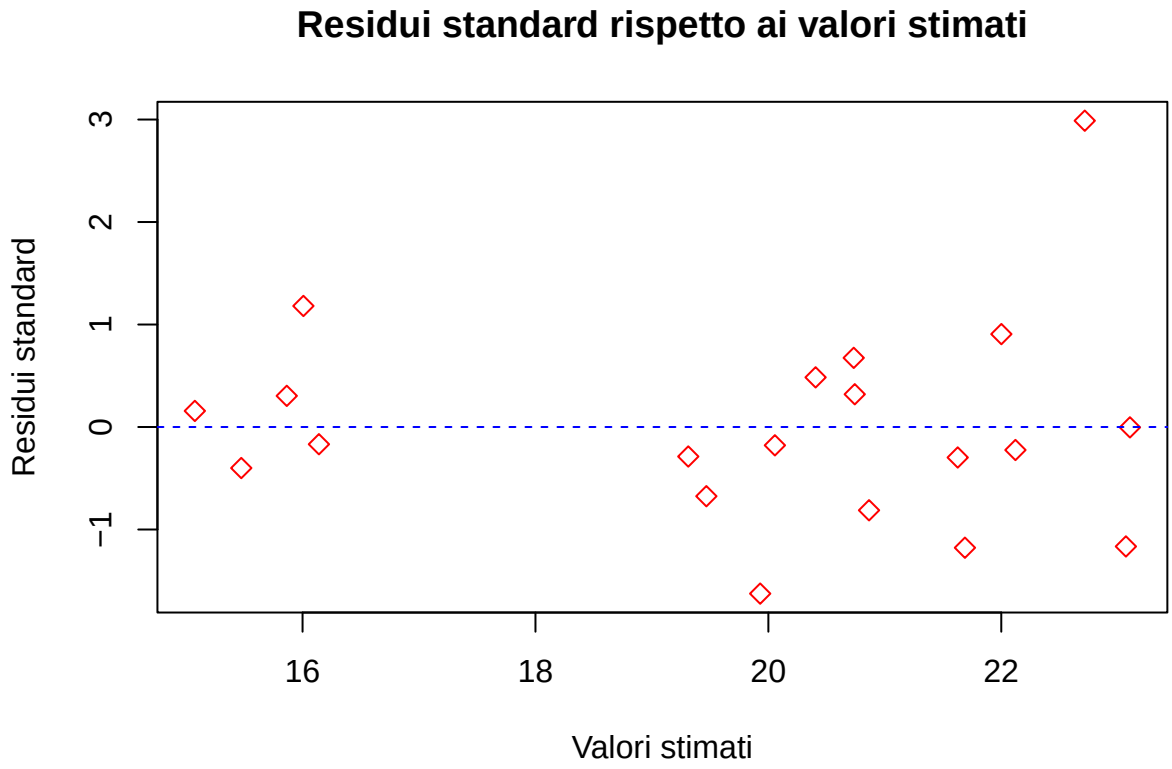
Possiamo notare che entrambi i regressori hanno segno negativo, quindi all'aumentare delle percentuali di "Licenza scuola elementare, nessun titolo di studio" e di "Licenza di scuola media" si abbassa la parcentuale di "laurea e post-laurea".

Anche per la regressione multipla la media dei valori stimati coincide con la media dei valori osservati, e la media dei residui uguale a 0. La varianza campionaria dei residui si calcola allo stesso del modo del caso semplice.

La media dei residui è nulla.

Anche nel caso multivariato è interessante calcolare i residui standardizzati che risultano essere caratterizzati da media nulla e varianza unitaria.

E' poi possibile realizzare un grafico in cui i residui standardizzati (ordinate) vengono disegnati in funzione dei valori stimati (ascisse) con il metodo dei minimi quadrati, come nel caso singolo.



Anche in questo caso i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia nessuna tendenza particolare nella distribuzione dei punti.

4.3.1 COEFFICIENTE DI DETERMINAZIONE

Il coefficiente di determinazione D^2 si calcola allo stesso modo del caso semplice, ossia il rapporto tra la varianza dei valori stimati tramite la funzione di regressione lineare multipla e la varianza dei valori osservati della variabile dipendente. In questo caso non coincide con il coefficiente di correlazione lineare al quadrato.

Quindi D^2 è un indice adimensionale, e si ha che $0<=D^2<=1$. Quindi come prima quando $D^2=0$ il modello di regressione lineare multipla non spiega per nulla i dati, mentre quando $D^2=1$ il modello spiega perfettamenti i dati.

```
summary(lm( df$`L/PL` ~ df$`LSE/NT`+ df$LSM ))$r.square
```

```
## [1] 0.7760305
```

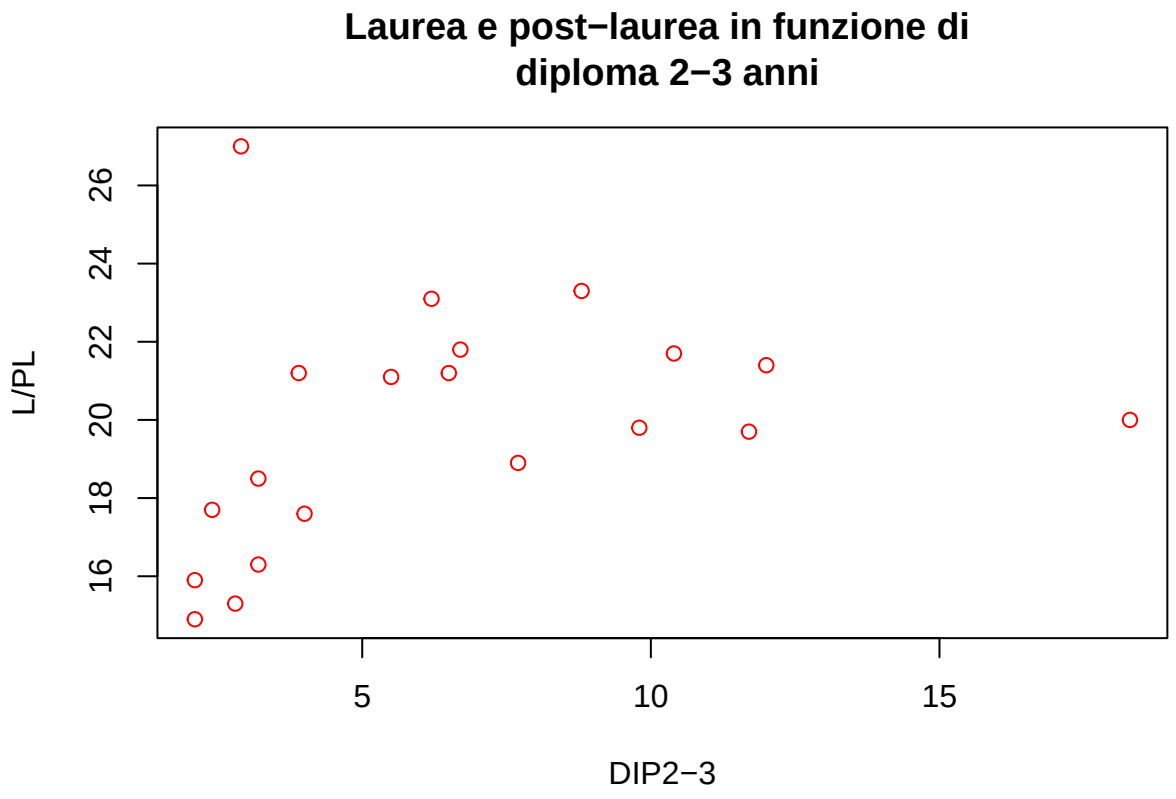
Si ottiene un coefficiente di determinazione pari a 0.7760305, quindi il modello di regressione lineare multipla spiega abbastanza bene i dati.

4.4 REGRESSIONE NON LINEARE

Spesso, osservando uno scatterplot, si nota che l'ipotesi di linearità di un modello non è accettabile poiché i dati sperimentali non evidenziano una correlazione di tipo lineare. In questo caso occorre ricorrere a modelli di regressione non lineare. Ho deciso di provare il modello polinomiale del secondo ordine:

$$Y = \alpha + \beta X + \gamma X^2$$

che può essere linearizzato tramite opportune trasformazioni. Prendiamo ad esempio lo scatterplot di "L/PL" in funzione di "DIP2-3".



In questo caso si nota che un modello lineare non può approssimare efficacemente i dati. Infatti il coefficiente di correlazione

```
cor(df$`DIP2-3`,df$`L/PL`)
```

```
## [1] 0.3205789
```

fornisce un valore basso ed il coefficiente di determinazione è $D^2 = (0.3205789)^2 = 0.1027708$ Ho deciso quindi di utilizzare il modello di regressione lineare polinomiale del secondo ordine. Per la stima di parametri α, β e γ si può ricorrere alla regressione multipla

$$Y = \alpha + \beta X_1 + \gamma X_2$$

con intercetta α e regressori β e γ per le variabili $X_1 = X$ e $X_2 = X^2$

Stimiamo poi i parametri α, β e γ tramite la funzione

$$lm(\tilde{y}x + I(x^2))$$

dove I() è un identificatore di variabile e viene inserito quando si debbono effettuare operazioni matematiche (divisione, elevamento a potenza) nelle variabili della regressione.

```
pol2 <- lm(df$`L/PL` ~ df$`DIP2-3` + I((df$`DIP2-3`)^2) )
pol2
```

```
##
## Call:
## lm(formula = df$`L/PL` ~ df$`DIP2-3` + I((df$`DIP2-3`)^2))
##
## Coefficients:
##        (Intercept)         df$`DIP2-3`  I((df$`DIP2-3`)^2)
##           15.57366             1.11345            -0.05024
```

Otteniam quindi $\alpha = 15.57366$, $\beta = 1.11345$ e $\gamma =$ -0.05024. Il modello polinomiale del secondo ordine è qundi:

$$Y = 15.57366 + 1.11345X - 0.05024X^2$$

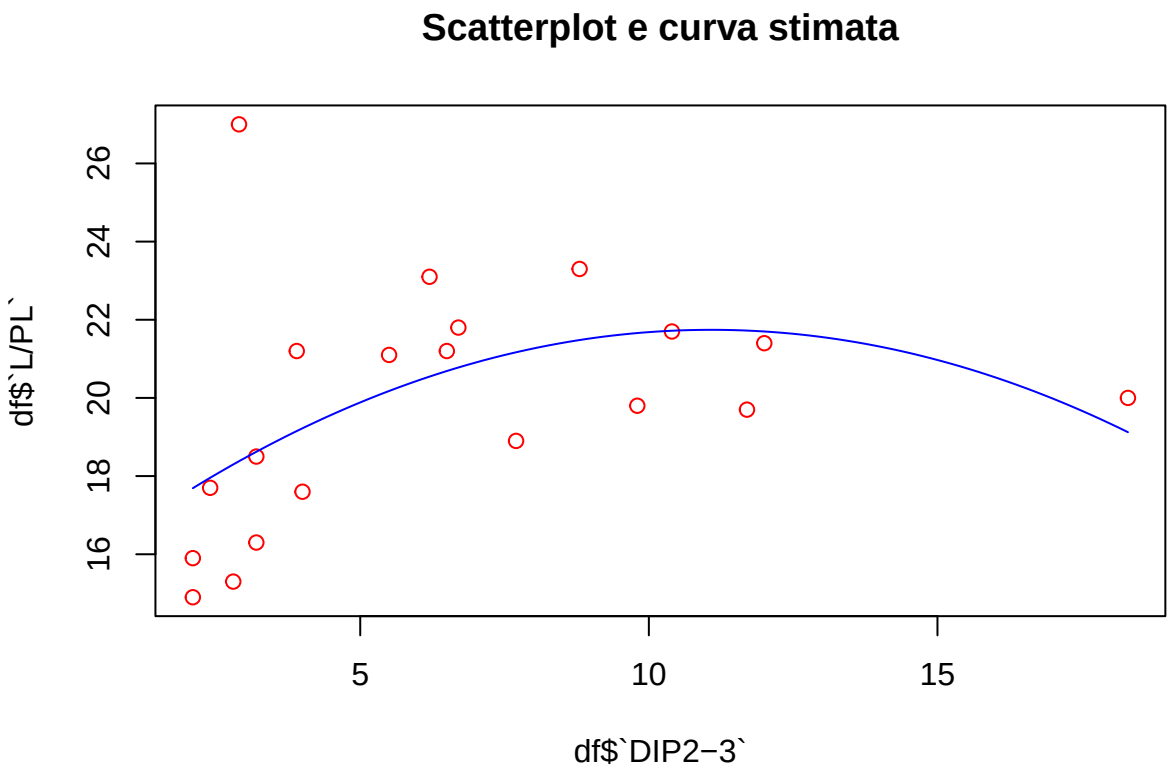
Calcoliamo poi il coefficiente di determinazione

```
summary(lm(df$`L/PL` ~ df$`DIP2-3` + I((df$`DIP2-3`)^2)))$r.square
```

```
## [1] 0.2418321
```

Si è ottenuto un coefficiente di determinazione migliore rispetto a quello del modello lineare $D^2 = 0.1027708$, Di seguito viene disegnata la curva stimata sullo scatterplot

```
plot(df$`DIP2-3`, df$`L/PL`, col = " red " , main =" Scatterplot e curva stimata ")
curve ( alpha + beta *x+ gamma * x ^2 , add = TRUE, col = "blue")
```



5 ANALISI DEI CLUSTER

L'analisi dei cluster è una metodologia che permette di raggruppare in sottoinsiemi, detti cluster, entità (unità) appartenenti ad un insieme più ampio. Abbiamo quindi un insieme di n individui $I = I_1, I_2, ..., I_n$, e per questo insieme andiamo a misurare p caratteristiche. Vogliamo creare dei raggruppamenti in modo tale che gli individui di uno stesso cluster siano tra loro il più possibile simili e gli individui di cluster distinti siano il più possibili diversi tra loro. All'insieme di individui è associata una matrice di misure X di cardinalità n x p, dove x_{ij} denota il valore della misura della caratteristica j-esima relativa all'individuo I_i. La matrice la possiamo vedere come n vettori di cardinalità 1 x p contenenti le misure delle p caratteristiche $X_1, X_2, ..., X_n$, dove ogni vettore rappresenta un individuo. In questo caso gli individui corrispondono alle regioni italiane.

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

Uno dei problemi che si presenta nell'analisi dei cluster riguarda la standardizzazione o meno delle variabili poiché attribuire un peso diverso a ciascuna caratteristica potrebbe condurre a risultati differenti circa la classificazione a seconda delle tecniche di clustering utilizzate. In molti metodi di clustering si raccomanda la standardizzazione di ogni variabile (caratteristica) usando la media campionaria e la deviazione standard campionaria entrambe derivate dall'insieme completo di individui della popolazione. Se abbiamo delle

percentuali non abbiamo bisogno di standardizzare. In primo luogo abbiamo bisogno di misurare quantitativamente quali individui sono più simili o meno distanti tra loro. A tale scopo ci sono due misure , una misura di distanza e una misura di similarità.

5.1 FUNZIONE DISTANZA

Una funzione a valori reali $d(X_i, X_j)$ è detta funzione distanza se e soltanto se soddisfa le seguenti proprietà:

$$(i)\ d(X_i, X_j) = 0 \ se\ e\ solo\ se\ X_i = X_j,\ con\ X_i\ e\ X_j\ in\ E_p;$$

$$(ii)\ d(X_i, X_j) \geq 0\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$

$$(iii)\ d(X_i, X_j) = d(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$

$$(iv)\ d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)\ per\ ogni\ X_i,\ X_j\ e\ X_k\ in\ E_p.$$

La prima proprietà implica che X_i è a distanza zero da se stesso e che ogni due punti a distanza nulla debbono essere identici. La seconda proprietà afferma che la funzione distanza è non negativa. La terza proprietà afferma che la funzione distanza è non negativa. La quarta proprietà, detta disuguaglianza triangolare, richiede che la distanza tra X_i e X_j debba essere sempre minore o uguale della somma delle distanze di ognuno dei due vettori considerati da qualunque altro terzo vettore X_k, ed è una proprietà molto potente. Le distanze tra tutte le possibili coppie di unità sono inserite in una matrice simmetrica D di cardinalità nxn, che risulta avere tutti 0 sulla diagonale principale. Inoltre visto anche che i valori al di sopra dell diagonale principale sono uguali a quella al di sotto, in totale il numero di distanze da calcolare è $\frac{n(n-1)}{2}$. E' sufficiente, pertanto, considerare la matrice triangolare al di sopra o al di sotto della diagonale principale di D.

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}$$

Non esiste un'unica funzione distanza, ma c'è un'intera famiglia di distanze che soddisfa almeno quattro delle proprietà sopra citate. Infatti, si possono facilmente dimostrare le seguenti proprieà:

- se d e d' sono due metrice anche d + d' è una metrica;
- se d è una metrica e c un numero reale positivo allora anche cd è una metrica;
- se d è una metrica e c un numero reale positivo allora anche d' = d/(c + d) è una metrica.

Il prodotto di due metriche (in particolare il quadrato di una metrica) non necessariamente soddisfa la disuguaglianza triangolare e quindi può non essere una misura di distanza.

5.1.1 METRICA EUCLIDEA

La più familiare misura di distanza è la metrica euclidea, così definita:

$$d_2(X_i, X_j) = [\sum_{k=1}^{p}(x_{ik} - x_{jk})^2]^{1/2}$$

dove x_{ik} rappresenta la k-esima caratteristica dell'i-esimo individuo. La distanza Euclidea usata su tutti i dati è fortemente influenzata dall'unità di misura in base alla quale è valutata ciascuna delle p caratteristiche. Occorre quindi effettuare una standardizzazione prima di calcolare le distanze. Ci sono anche altre metriche di distanza molto utilizzate, ne cito alcune senza entrare nel dettaglio (visto che in questo progetto verrà utilizzata la metrica euclidea): metrica del valore assoluto (di manhattan); metrica del massimo (metrica di Chebychev); metrica di Minkowski; metrica di Canberra; metrica di Jaccard.

5.2 FUNZIONE DI SIMILARITA'

Le misure di similarità sono più deboli di quelle di distanza, poiché non godono della proprietà della disuguaglianza triangolare.

Una funzione a valori reali $s(X_i, X_j)$ è detta misura di similarità se e soltanto se soddisfa le seguenti proprietà:

$$\begin{aligned}
&(i)\ s(X_i, X_i) = 1;\\
&(ii)\ 0 \leq s(X_i, X_j) \leq 1;\\
&(iii)\ s(X_i, X_j) = s(X_j, X_i)\ \textit{per ogni}\ X_i\ \textit{e}\ X_j.
\end{aligned}$$

Da una misura di distanza posso sempre passare a una misura di similarità, mentre non è detto che da una misura di similarità possa passare a una di distanza.

5.3 MISURE DI NON OMOGENEITA'

Le misure di non omogeneità vanno a misurare la bontà delle suddivisioni in cluster. La misura di non omogeneità totale è definita sull'insieme totale dei dati. Poi andremo a definire delle misura di non omogeneità interne ai cluster (within), e tra i cluster (between).

Quello che vogliamo è che l'omogeneità tra gli individui nello stesso cluster sia alta, mentre tra individui di cluster diversi sia bassa.

Per definire questa misura dobbiamo partire col definire la matrice W_x delle varianze e covarianze, di ordine pxp, associata alla matrice X dei dati.

$$W_X = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \dots & w_{pp} \end{pmatrix},$$

Tutti gli elementi della matrice sono le covarianze tra le varie caratterische, che sono p. Inoltra sulla diagonale principale ci sono le varianze delle singole caratteristiche.

Possiamo ora definire la matrice statistica di non omogeneità (statistical scatter matrix) per l'insieme I di individui, di cardinalità pxp, è cosi definita:

$$H_I = (n-1)W_I = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1p} \\ h_{21} & h_{22} & \dots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1} & h_{p2} & \dots & h_{pp} \end{pmatrix}$$

Di questa matrice siamo interessati principalmente agli elementi della diagonale principale, ossia le varianze moltiplicate per n-1.

Moltiplico per n - 1 cosi da liberarmi dell'n - 1 al denominatore delle varianze.

La misura di non omogeneità statistica dell'insieme di individui è definita come la traccia della matrice H_I di non omogeneità, ossia la somma degli elementi della diagonale principale.

$$trH_I = \sum_{r=1}^{p} h_{rr} = (n-1)\sum_{r=1}^{p} s_r^2$$

Se abbiamo un solo individuo, questa traccia è 0.

Di seguito viene calcolata in R la misura di non omogeneità statistica totale dei dati.

```
n <- nrow(df)
trT <-(n-1) * sum(apply(df,2,var))
trT
```

```
## [1] 1290.63
```

La misura di non omogeneità statistica dell'insieme I di individui si può anche scrivere:

$$trH_I = \sum_{i=1}^{n} d_2^2(X_i, \bar{X})$$

La distanza euclidea gioca quindi un ruolo rilevante nel calcolo della misura di non omogeneità statistica.

La traccia di una matrice di non omogeneità di un insieme di individui fornisce una misura della dispersione dei dati intorno al valore medio dell'insieme dal quale è stata ricavata. E' intuitivo pensare che più un insieme di dati è addensato più piccola è la traccia della matrice di non omogeneità.

Quindi, denotando con T la matrice di non omogeneità statistica relativa all'insieme totale I degli n individui, con S la somma delle matrici di non omogeneità dei singoli cluster e con B la matrice di non omogeneità tra i cluster, si ha che:

$$T = S + B$$

da cui segue

$$trT = trS + trB$$

La misura di non omogeneità totale è uguale alla misura di non omogeneità dell'unione dei singoli cluster, la misura di non omogeneità interna (within) invece è uguale alla somma delle misure di non om. dei singoli cluster e quella tra i cluster (between) invece me la calcolo per differenza delle altre due.

Quindi vado a considerare la misura di non omogeneità totale, poi mi calcolo la misura di non omogeneità dei singoli cluster, e mi calcolo la trB come trT - trS.

La matrice T è fissata per ogni matrice X di dati, quindi anche la traccia relativa. Mentre le matrici S e B, e le relative traccie, dipendono dalle partizioni dei dati.

Quindi visto che la trT è fissata, se la trS diminuisce allora la trB aumenta, e viceversa.

Poiché la traccia trT è univocamente determinata per ogni matrice X dei dati nxp, fissato il numero di cluster, questi dovrebbero essere individuati in modo tale da minimizzare trS (la misura di non omogeneità interna ai cluster, within) o equivalentemente massimizzare trB (la misura di non omogeneità tra i cluster, between).

Una volta scelta la misura di distanza (o di similarità) si pone il problema di procedere alla scelta di un idoneo algoritmo di raggruppamento delle unità osservate. I metodi di raggruppamento si distinguono in tre tipi:

- enumerazione completa;
- metodi gerarchici;
- metodi non gerarchici.

Le misure di non omogeneità statistiche sono utilizzate per valutare, fissato il numero di cluster, la bont'a della suddivisione in cluster ottenuta con i vari metodi (di enumerazione completa, gerarchici, non gerarchici).

5.4 INTRODUZIONE METODI GERARCHICI E NON GERARCHICI

I metodi gerarchici e non gerarchici non considerano tutte le possibili partizioni, ma solo un sottoinsieme.

Obiettivo dei metodi gerarchi non è quello di determinare la migliore partizione possibile ma cercare di costruire struttura ad albero detto dendrogramma che permette di visualizzare le possibili agglomerazioni che si sono man mano venute a creare. Spetterà poi al ricercatore tagliare questo albero e decidere qual è la partizione. Questi metodi si basano su misure di distanza o misure di similarità. Noi utilizzeremo le misure di distanza poiché sono più potenti.

Metodi gerarchici si dividono in agglomerativi e divisivi. Metodi agglomerativi che procedono per aggregazioni successive delle unità partendo da n gruppi formati da un singolo individuo e divisivi che partono da un solo gruppo formato da tutte le unità e procedono a divisioni successive fino a giungere a gruppi formati da una sola unità.

- Vantaggio: permettono di avere una visione completa e nessuna scelta a priori del numero di cluster;
- Svantaggio: non permettono di riallocare individui che sono stati gia riallocati in un passo precedente dell'analisi, cosa possibile invece nei metodi non gerarchici.

Nei metodi non gerarchici dobbiamo però specificare inizialmente il numero di cluster, e l'obiettivo è quello di individuare un'unica partizione degli n individui in cluster.

5.5 METODI NON GERARCHICI

Il metodo maggiormente utilizzato è il metodo k-means, il quale richiede che venga fissato il numero di cluster all'inizio e fornisce in output un'unica partizione. Esso consiste dei passi descritti nel seguente algoritmo:

- Step 1: fissare a priori il numero k di cluster specificando dei punti di riferimento iniziali che inducono una prima partizione provvisoria;

- Step 2: considerare tutti gli individui e attribuire ciascuno di essi al cluster individuato dal punto di riferimento da cui ha distanza minore;
- Step 3: calcolare il baricentro (centroide) di ognuno dei k gruppi. Tali cenrtoidi costituiscono i punti di riferimento per i nuovi cluster;
- Step 4: Valutare la distanza di ogni individuo da ogni centroide ottenuto al passo precedente. Se la distanza minima non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, allora si procede a spostare l'individuo presso il cluster che ha il centroide più vicino;
- Step 5: Ricalcolare i centroidi dei k gruppi e ripetere la procedura;
- Step 6: Ripetere il procedimento a partire dal punto (4) fino a che i centroidi non subiscono ulteriori modifiche rispetto all'iterazione precedente. Si procede così iterativamente a spostamenti successivi fino a raggiungere una configurazione stabile, ossia gli individui all'interno di ogni cluster non cambiano al ripetersi del procedimento.

Nel metodo k-means non posso applicare tutti i tipi di distanze, ma come misura di distanza tra i vettori delle caretteristiche (ossia gli individui) e i centroidi viene utilizzata la distanza euclidea e, come nel metodo del centroide, viene utilizzata la matrice contenente i qudrati delle distanze euclidee.

I vantaggi del metodo k-means sono la velocità di esecuzione dei calcoli e l'estrema libertà che viene lasciata agli individui di raggrupparsi e allontanarsi. Uno svantaggio è invece che la classificazione finale può essere influenzata dalla scelta iniziale dei k vettori delle caratteristiche come punti di riferimento, dall'ordine in cui sono presi tali vettori e naturalmente dalle proprietà geometriche dei vettori delle misure. Infatti, l'algoritmo potrebbe convergere ad un ottimo locale e non globale, il che significa che se si inizia con un diverso insieme di punti di riferimento si può giungere ad una differente partizione finale.

Ho provato k-means con valori di k crescenti fino ad arrivare ad avere un k che mi forniva un rapporto tra la misura di non omogeneità between e la misura totale superiore al 70%. Inoltre ho visto che andando ad aumentare k si otteneva una misura di non omogeneità migliore ma si andavano a creare cluster con un solo individuo, quindi ho preferito optare per k = 4. Inoltre ho scelto nstart = 20 (ossia il numero di ripetizioni della la scelta casuale iniziale dei centroidi) empiricamente andando ad aumentare il parametro finchè il risultato non cambiava.

```
km <- kmeans (df , centers = 4, nstart = 20)
km
```

```
## K-means clustering with 4 clusters of sizes 5, 5, 5, 5
##
## Cluster means:
##   LSE/NT   LSM DIP2-3 DIP4-5  L/PL
## 1   2.94 32.50  11.58  32.98 20.02
## 2   2.76 27.08   7.32  39.50 23.32
## 3   4.18 31.16   4.62  40.16 19.92
## 4   8.96 37.86   2.52  34.64 16.02
##
## Clustering vector:
##                       Piemonte Valle d'Aosta / Vallée d'Aoste
##                              1                               1
##                        Liguria                       Lombardia
##                              2                               1
## Trentino Alto Adige / Südtirol                          Veneto
##                              1                               1
##          Friuli-Venezia Giulia                 Emilia-Romagna
##                              2                               2
##                        Toscana                          Umbria
##                              3                               2
```

```
##                         Marche                         Lazio
##                              3                             2
##                        Abruzzo                       Molise
##                              3                             3
##                       Campania                       Puglia
##                              4                             4
##                     Basilicata                      Calabria
##                              3                             4
##                        Sicilia                      Sardegna
##                              4                             4
##
## Within cluster sum of squares by cluster:
## [1] 101.416  96.616  59.328  57.672
##  (between_SS / total_SS =  75.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
km$cluster
```

```
##                       Piemonte Valle d'Aosta / Vallée d'Aoste
##                              1                               1
##                         Liguria                       Lombardia
##                              2                               1
## Trentino Alto Adige / Südtirol                          Veneto
##                              1                               1
##          Friuli-Venezia Giulia                 Emilia-Romagna
##                              2                               2
##                        Toscana                         Umbria
##                              3                               2
##                         Marche                          Lazio
##                              3                               2
##                        Abruzzo                         Molise
##                              3                               3
##                       Campania                         Puglia
##                              4                               4
##                     Basilicata                       Calabria
##                              3                               4
##                        Sicilia                       Sardegna
##                              4                               4
```

Il metodo k-means individua la seguente partizione in 4 cluster: G1 = {Piemonte, Veneto, Valle d'Aosta, Lombardia, Trentino Alto Adige}, G2 = {Calabria, Campania, Sicilia, Puglia, Sardegna}, G3 = {Marche, Basilicata, Abruzzo, Molise, Toscana} , G4 = {Friuli, Lazio, Liguria, Emilia-Romagna, Umbria}

Di seguito ci sono le coordinate dei centroidi dei 4 cluster

```
km$centers
```

```
##   LSE/NT   LSM DIP2-3 DIP4-5  L/PL
## 1   2.94 32.50  11.58  32.98 20.02
```

```
## 2   2.76 27.08   7.32  39.50 23.32
## 3   4.18 31.16   4.62  40.16 19.92
## 4   8.96 37.86   2.52  34.64 16.02
```

Qui abbiamo la misura di non omogeneità totale

```
km$totss
```

```
## [1] 1290.63
```

La misura di non omogeneità statistica dei singoli cluster

```
km$withinss
```

```
## [1] 101.416  96.616  59.328  57.672
```

La misura di non omogeneità statistica within, ossia la somma delle misure dei singoli cluster

```
km$tot.withinss
```

```
## [1] 315.032
```

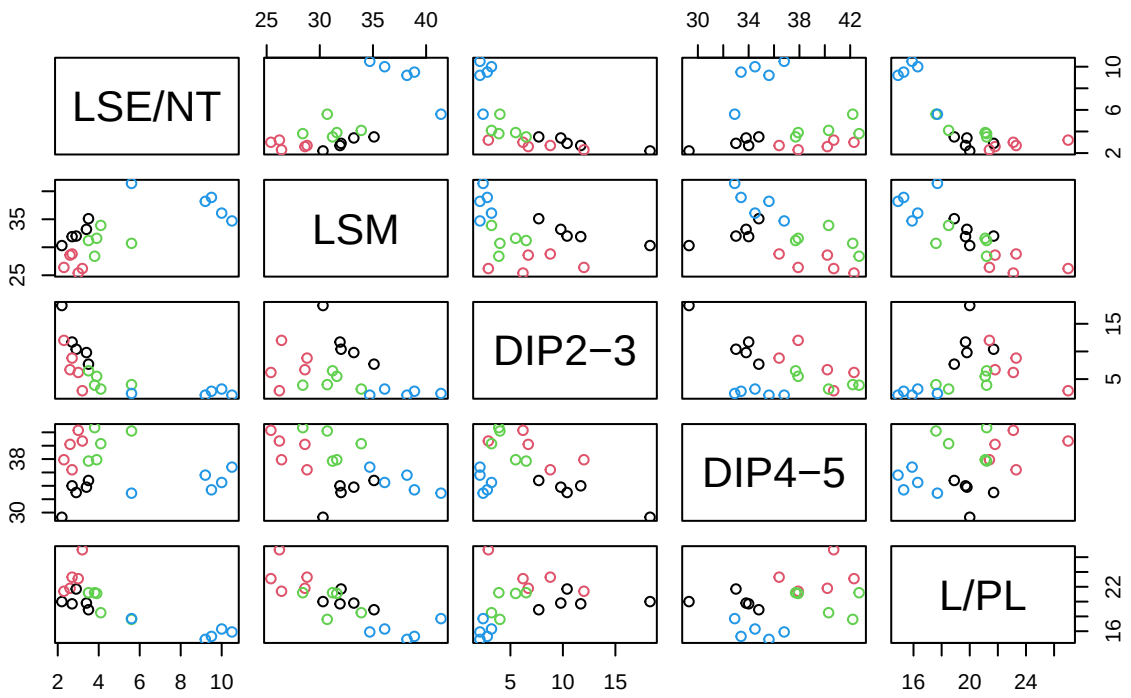
La misura di non omogeneità statistica tra i cluster, ossia la totale meno la within

```
km$betweenss
```

```
## [1] 975.598
```

Quindi si ha $\frac{trB}{trT} = \frac{975.598}{1290.63} = 0.7559 = 75,6\%$ e $\frac{trS}{trT} = \frac{315.032}{1290.63} = 0,1332 = 24,4\%$. Per cui per k = 4 abbiamo ottenuto una buona suddivisione poiché la misura di non omogeneità statistica between supera il 70%.

Ora, fissato il numero di cluster pari a 4, vediamo se i metodi gerarchici riescono a fornirci una partizione degli individui in 4 cluster migliore.



5.6 METODI GERARCHICI

Nei metodi gerarchici agglomerativi partiamo da n individui, ogni individuo viene inserito in un cluster, quindi abbiamo n cluster. Al passo successivo agglomeriamo 2 individui, e cosi abbiamo n-1 cluster, proseguendo fino ad avere un solo cluster.

Algoritmo generale dei metodi gerarchici agglomerativi:

- Step 1: a partire dalla matrice dei dati X originaria o dalla matrice scalata (nel caso in cui ci sono colonne con misure diverse o valori molto grandi) considerare la matrice delle distanze nxn D tra gli individui (visti come singoli cluster di un solo individuo). In questo step bisogna quindi stabilire quale funzione distanza utilizzare in questo primo passo;
- Step 2: cercare la distanza più piccola, e unire gli individui associati a questa distanza in un unico cluster. Calcolare la distanza tra questo nuovo cluster e gli altri cluster. In questo step bisogna quindi stabilire come calcolare questa distanza;
- Step 3: costruire una nuova matrice delle distanze, ridotta di una riga e di una colonna rispetto a quella precedente;
- Step 4: operare sulla matrice cosi ottenuta a partire dal passo 2 fino ad esaurire tutte le possibilità di raggruppamento, raggiungendo alla fine una matrice 2x2. La procedura richiede n-1 iterazioni;
- Step 5: rappresentare graficamente il processo di agglomerazione attraverso un dendrogramma che riporta sull'asse verticale il livello di distanza a cui avviene l'agglomerazione e sull'asse orizzontale riporta gli individui. Ad ogni livello di distanza corrisponde una partizione.

I passi salienti sono il passo 1 e il passo 2. Il passo 1 che mi permette di decidere la distanza da andare da utilizzare, e il 2 che mi permette di decidere come calcolare la distanza tra il nuovo cluster agglomerato e

i rimanenti. Il passo 2 è proprio quello che fornisce il nome al metodo che andiamo ad utilizzare. Infatti, ciascun metodo si differenzia dagli altri per il modo in cui si individuano i due cluster meno distanti (o più somiglianti) e per il modo in cui si determina la distanza (o similarità) che intercorre tra il nuovo cluster ottenuto e i rimanenti. Di seguito sono riportati i vari metodi gerarchici applicati al mio dataset.

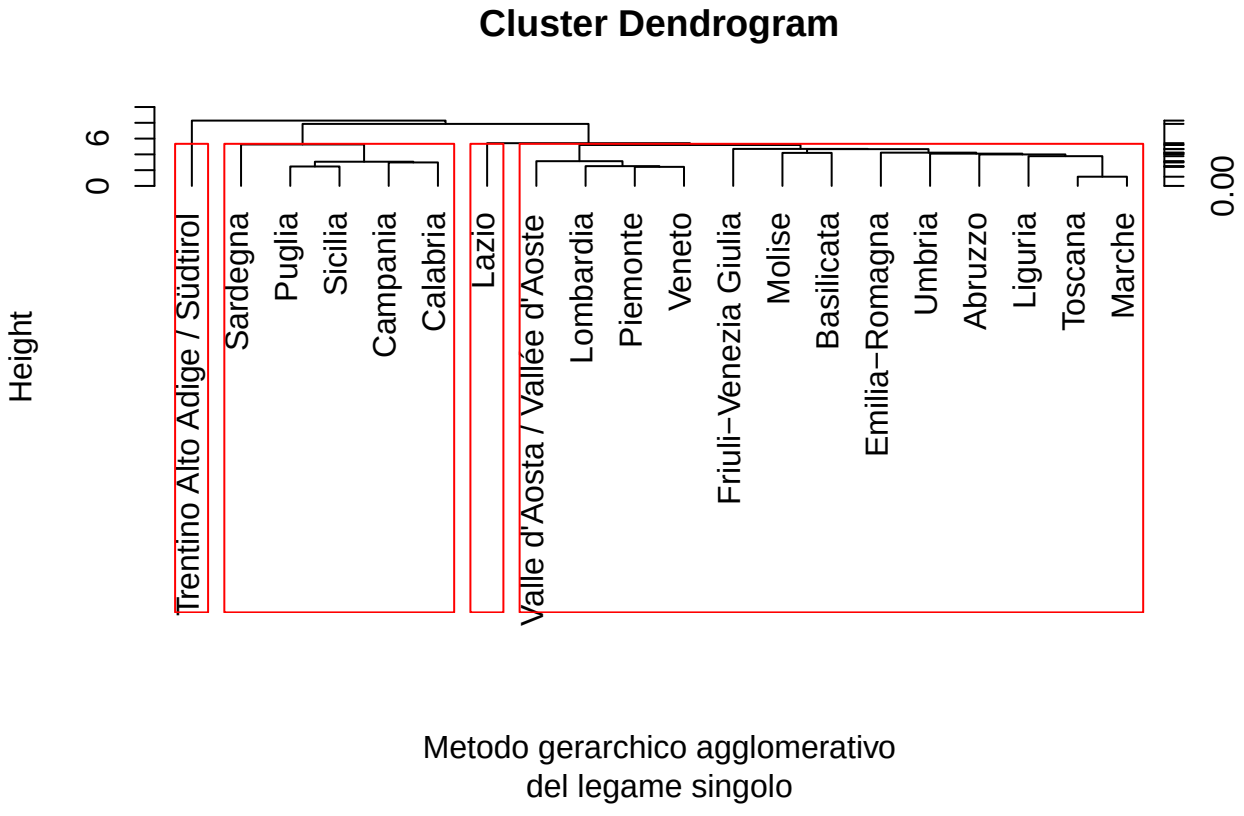
5.6.1 METODO LEGAME SINGOLO

Supponiamo di avere due cluster, G_1 (con n_1 individui) e G_2 (con n_2 individui). Quando dobbiamo agglomerare G_1 e G_2 ho $n_1 n_2$ possibili distanze, e con questo metodo la distanza tra G_1 e G_2 è la minima delle $n_1 n_2$ distanze.

Vantaggio: applicabile sempre; Svantaggio: se ho due cluster molto addensati e dei punti intermedi sparsi tra questi cluster, questo metodo agglomera tutto basandosi sulle singole distanze, quindi si possono avere dei cluster con individui dissimili.

```
d <- dist(df,method = "euclidean", diag = TRUE, upper = TRUE)
```

```
tree1 <- hclust (d , method = "single")
plot( tree1 , hang = -1 , xlab ="Metodo gerarchico agglomerativo" , sub ="del legame singolo")
axis ( side =4 , at = round(c (0 , tree1$height ) ,2))
rect.hclust ( tree1 , k = 4, border = " red ")
```




```
misureNonOmogeneita <- function(taglio){
  num <- table(taglio) #numero di elementi dei gruppi
  tagliolist <- list (taglio) # lista di indici per i gruppi
```

```
  agvar <- aggregate (df , tagliolist , var ) [, -1]
  misure <- cbind(0,0,0,0)
  for(i in 1:nrow(agvar)){
    if(num[[i]]>1){
      misure[i] <- (num[[i]]-1) * sum(agvar[i, ])
    }
    else
       misure[i] <- 0
  }
  colnames(misure) <- c("G1","G2","G3","G4")
  return (misure)

}
```

Misure di non omogeneità stastistica dei singoli cluster

```
misureNonOmogeneita(taglio1)
```

```
##            G1 G2 G3     G4
## [1,] 388.1338  0  0 57.672
```

Misura di non omogeneità statistica interna (within) ai cluster

```
trS = sum(misureNonOmogeneita(taglio1))
trS
```

```
## [1] 445.8058
```

Misura di non omogeneità statistica tra i cluster (between)

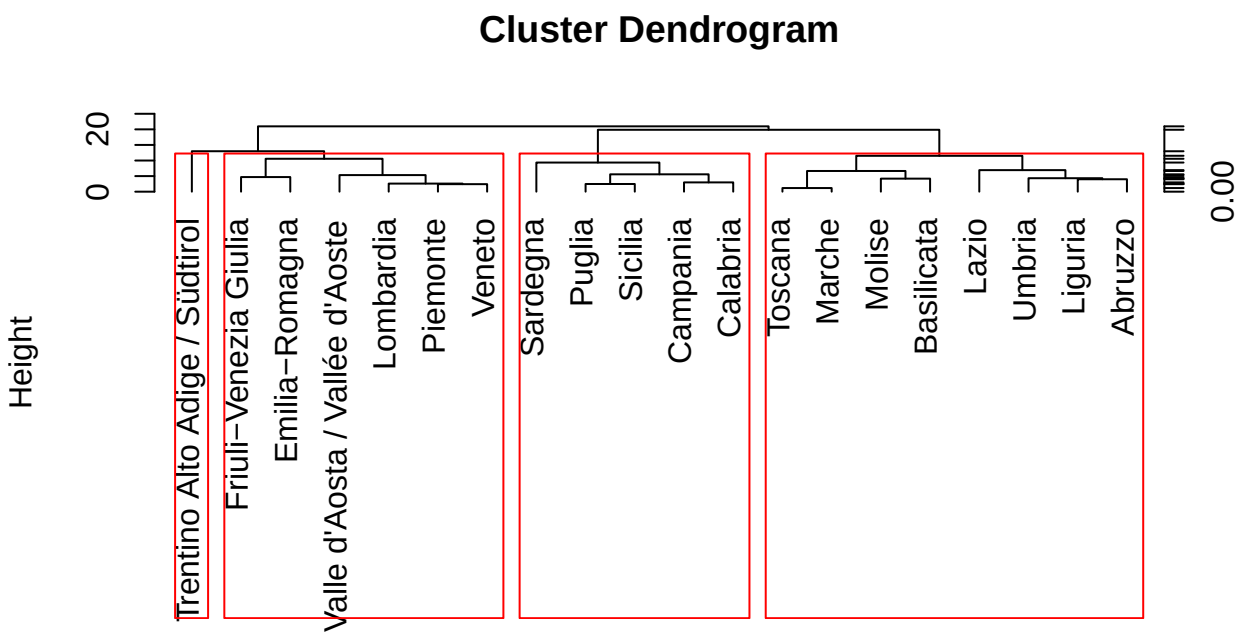
```
trB = trT - trS
trB
```

```
## [1] 844.8242
```

$\frac{trB}{trT} = \frac{844.8242}{1290.63} = 0.6545 = 65,4\%$

5.6.2 METODO DEL LEGAME COMPLETO

Prendiamo la massima delle possibili distanze (invece della minima come nel legame singolo). Il metodo del legame completo identifica soprattutto gruppi di forma ellissoidale, ossia una serie di punti che si addensano intorno ad un nucleo centrale. Questo algoritmo privilegia l'omogeneità tra gli elementi del gruppo a scapito della differenziazione tra i gruppi. Il dendrogramma costruito con questo metodo ha i rami molto più lunghi rispetto al dendrogramma ottenuto con il metodo del legame singolo poiché i gruppi si formano a livelli di distanza maggiori.



Misure di non omogeneità stastistica dei singoli cluster

```
misureNonOmogeneita(taglio2)
```

```
##            G1       G2 G3     G4
## [1,] 94.06333 163.3263  0 57.672
```

Misura di non omogeneità statistica interna (within) ai cluster

```
trS2 = sum(misureNonOmogeneita(taglio2))
trS2
```

```
## [1] 315.0616
```

Misura di non omogeneità statistica tra i cluster (between)

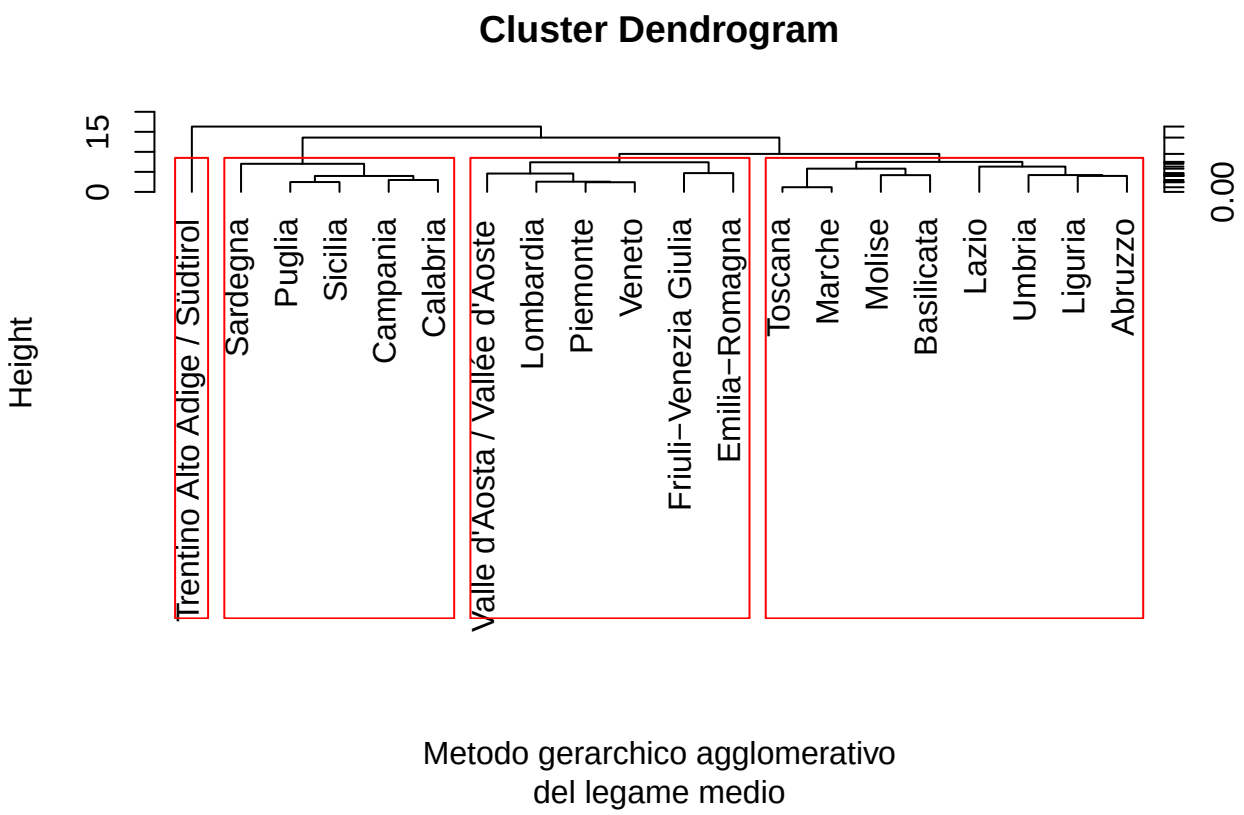
```
trB2 = trT - trS2
trB2
```

```
## [1] 975.5684
```

$\frac{trB}{trT} = \frac{975.5684}{1290.63} = 0.7558 = 75.6\%$

5.6.3 METODO DEL LEGAME MEDIO

In questo metodo consideriamo invece come distanza tra due cluster G_1 e G_2 la media delle distanze tra tutte le coppie di individui che compongono i due gruppi. Nel metodo del legame medio risulta fondamentale il numero di individui dei cluster che uniamo, a differenza degli altri due metodi visti prima. Se abbiamo un cluster numeroso G_u e l'altro piccolo G_v e li dobbiamo unire, il cluster piccolo non influenza nel calcolo della distanza dagli altri cluster. Ossia la distanza $d_{(uv),z}$ del nuovo cluster G_{uv} da un altro generico cluster G_z sarà molto vicina alla distanza $d_{u,z}$ del cluster numeroso G_u da G_z. Nel dendogramma ottenuto con questo metodo ci aspettiamo che le altezze a cui avvengono le aggregazioni siano intermedie rispetto a quelle ottenute con il legame singolo e il legame completo. Questo metodo spesso si comporta come quello del legame singolo, ossia se ci sono valori anomali, dei punti sparsi, e si è creata uan catena nel legame singolo allora si crea una catena anche nel legame medio. Essendo una media è influenzata dai valori anomali.



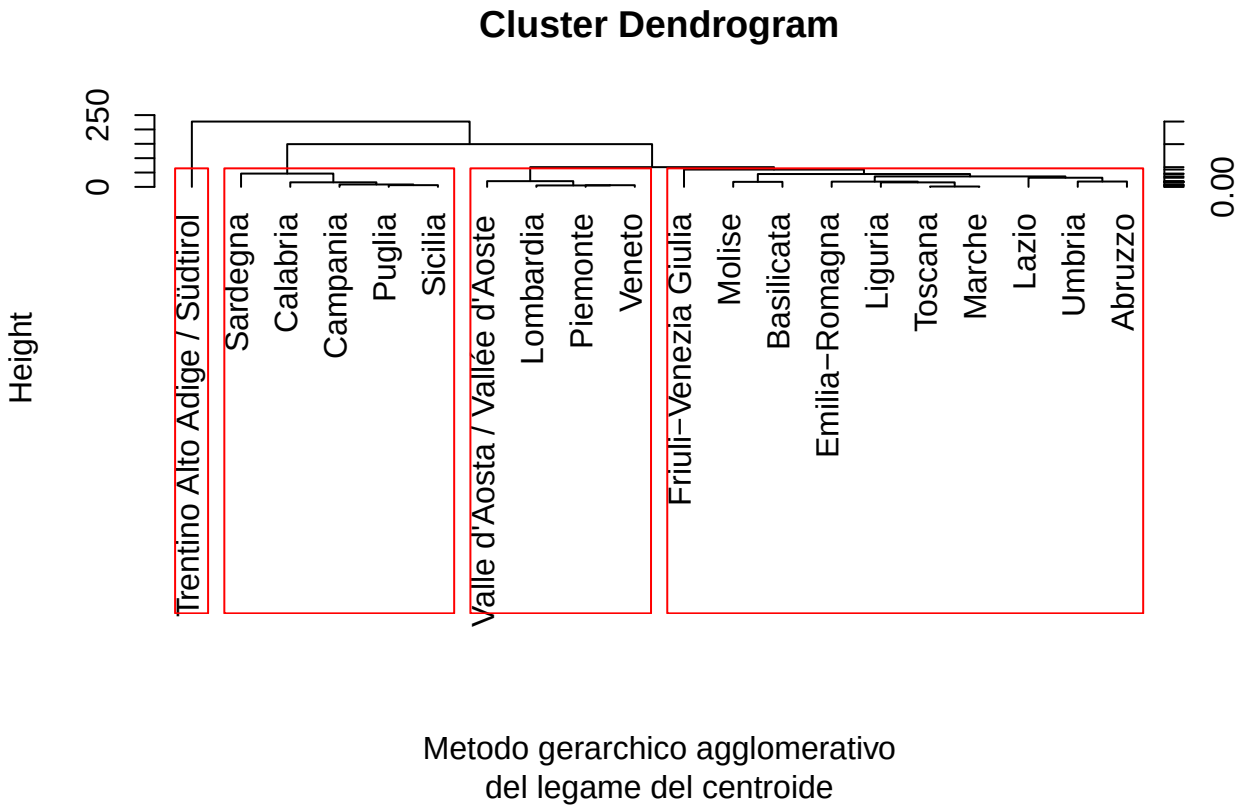
Metodo gerarchico agglomerativo
del legame medio

In questo caso si nota che i cluster individuati sono gli stessi di quelli individuati con il metodo del legame completo, quindi è inutile calcolare le misure di non omogeneità poiché sono le stesse.

5.6.4 METODO DEL CENTROIDE

I primi 3 metodi posso utilizzarli per qualsiasi tipo di distanza. Il metodo del centroide e quello della mediana si differenziano dagli altri perché in questi si considera quadrati delle distanze euclidee. Nel metodo del centroide consideriamo invece come distanza tra due cluster G_1 e G_2 la distanza tra i centroidi, ossie tra le medie campionarie calcolate sugli individui dei due cluster. Il metodo del centroide può dare origine a fenomeni gravitazionali, per cui i gruppi grandi tendono ad attrarre al loro interno i piccoli gruppi. Inoltre, le distanze in cui si verificano le successive agglomerazioni possono essere non crescenti, poiché abbiamo delle distanze al quadrato. Uno svantaggio del metodo del centroide è che se le misure dei due cluster da unire

sono molto differenti il centroide del nuovo cluster sarà molto vicino a quello del cluster più numeroso, come accade per il metodo del legame medio.



Misure di non omogeneità stastistica dei singoli cluster

```
##          G1      G2 G3     G4
## [1,] 21.305 250.815  0 57.672
```

Misura di non omogeneità statistica interna (within) ai cluster

```
## [1] 329.792
```

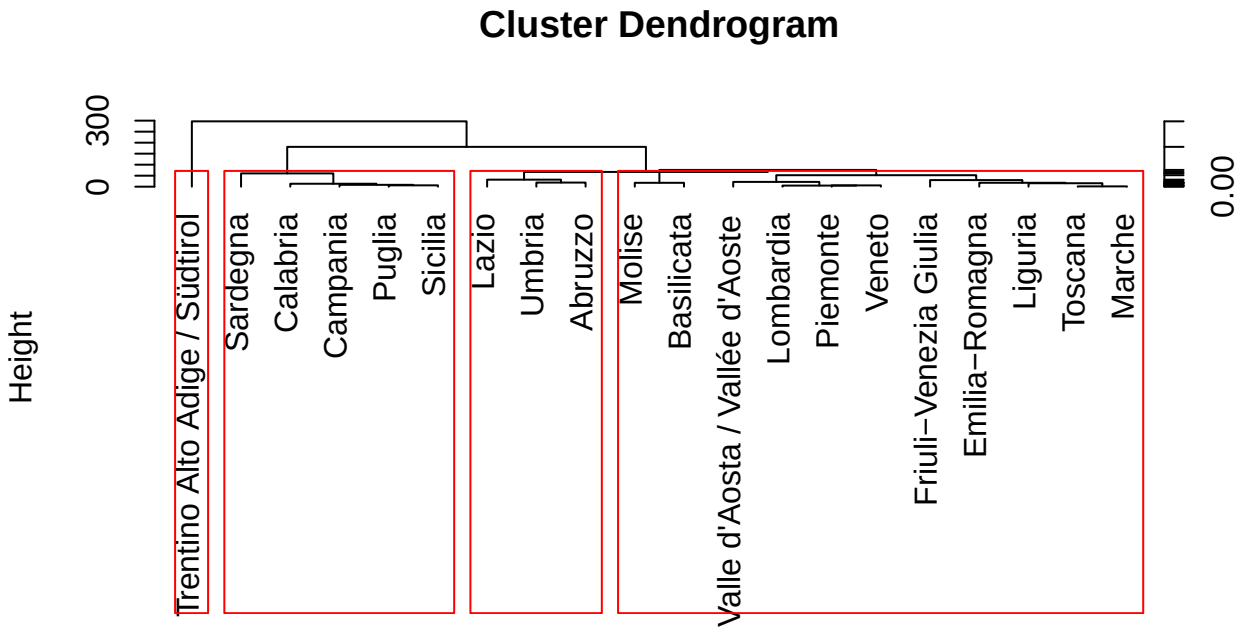
Misura di non omogeneità statistica tra i cluster (between)

```
## [1] 960.838
```

$\frac{trB}{trT} = \frac{960.838}{1290.63} = 0,7444 = 74,4\%$

5.6.5 METODO DELLA MEDIANA

Il metodo della mediana è simile a quello del centroide, con la differenza che la procedura è indipendente dalla numerosità dei cluster. Il metodo della mediana, così come il metodo del legame singolo, può dare origine alla formazione di una catena tra gli individui. Non risente dei valori anomali come quello dei centroidi, essendo una mediana, e i gruppi più numerosi non attraggono i più piccoli come prima.



Metodo gerarchico agglomerativo
della mediana

Misure di non omogeneità stastistica dei singoli cluster

```
##            G1 G2       G3     G4
## [1,] 279.9564  0 30.62667 57.672
```

Misura di non omogeneità statistica interna (within) ai cluster: 368.2550303

Misura di non omogeneità statistica tra i cluster (between): 922.3749697

$\frac{trB}{trT} = \frac{922.375}{1290.63} = 0.7146 = 71.5\%$

5.6.6 CONCLUSIONI ANALISI CLUSTER

Quindi fissato il numero di cluster pari a k=4, i metodi che forniscono la miglior partizione (ossia una misura di non omogeneità tra i cluster massima) sono k-means, il metodo del legame completo ed il metodo del legame medio, con una misura $\frac{trB}{trT} = 75.6\%$. Di seguito sono riportati tutti i risultati ottenuti con i vari metodi

	between/total
k-means	0.756
legame singolo	0.654
legame completo	0.756
legame medio	0.756
centroide	0.744
mediana	0.715

6 INTRODUZIONE: INFERENZA STATISTICA

Di particolare importanza in statistica è l'inferenza statistica. Essa ha lo scopo di estendere le misure ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto. Uno dei problemi centrali dell'inferenza statistica è il seguente: si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro $\vartheta \in \theta$ non noto (o più parametri non noti). Il termine osservabile significa che si possono osservare i valori assunti dalla variabile aleatoria X (ad esempio, eseguendo un esperimento casuale) e quindi il parametro non noto è presente soltanto nella legge di probabilità (funzione di distribuzione, funzione di probabilità, densità di probabilità). Ovviamente se ϑ è noto la legge di probabilità è completamente specificata. Per ottenere informazioni sul parametro non noto ϑ della popolazione, si può fare uso dell'inferenza statistica considerando un campione (x1, x2, . . . , xn) estratto dalla popolazione e effettuando su tale campione delle opportune misure. Affinché le conclusioni dell'inferenza statistica siano valide il campione deve essere scelto in modo tale da essere rappresentativo della popolazione. L'inferenza statistica si basa su due metodi fondamentali di indagine: la stima dei parametri e la verifica delle ipotesi, che vedremo più avanti.

7 DISTRIBUZIONE ESPONENZIALE

Per questa seconda parte del progetto ho scelto di approfondire la distribuzione esponenziale. In particolare consideriamo di avere una popolazione descritta da una variabile aleatoria esponenziale, ossia il tempo espresso in minuti tra successivi arrivi di clienti in un negozio. Estraiamo un campione casuale di ampiezza 50 contenente i tempi di interarrivo in minuti dei clienti.

```
set.seed(1)
camp1 <- rexp(n = 50, rate = 0.5)
camp1
```

```
##  [1] 1.51036367 2.36328556 0.29141345 0.27959052 0.87213725 5.78993707
##  [7] 2.45912411 1.07936568 1.91313499 0.29409198 2.78147026 1.52405971
## [13] 2.47520710 8.84786844 2.10908633 2.07048789 3.75207034 1.30949327
## [19] 0.67386695 1.17695944 4.72903051 1.28378518 0.58824078 1.13173105
## [25] 0.21214525 0.11887832 1.15742493 7.91786570 2.34662421 1.99362591
## [31] 2.87057069 0.07453705 0.64802031 2.64093586 0.40702070 2.04545175
## [37] 0.60348187 1.45042861 1.50308538 0.47005490 2.15976227 2.05649381
## [43] 2.58452330 2.50621071 1.10928280 0.60256599 2.58624931 1.98911158
## [49] 1.02834859 4.01566480
```

Su tale campione andrò ad effettuare delle misure per ricavare i parametri non noti della popolazione. Prima però vediamo nel dettaglio la distribuzione esponenziale.

La densità di probabilità esponenziale si può interpretare come l'analogo nel continuo della funzione di probabilità geometrica nel senso che una variabile aleatoria caratterizzata da densità di probabilità esponenziale può immaginarsi idonea a descrivere un tempo di attesa nel continuo.

Definizione 9.2 *Sia $\lambda > 0$. Una variabile aleatoria X di funzione di distribuzione*

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases} \tag{9.4}$$

e corrispondente densità di probabilità

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & altrimenti \end{cases} \tag{9.5}$$

si dice esponenzialmente distribuita con parametro λ.

Per una variabile aleatoria esponenziale si ha:

$$E(X) = \frac{1}{\lambda}, \quad E(X^2) = 2(\frac{1}{\lambda})^2, \quad Var(X) = E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2}$$

Il significato del parametro λ della distribuzione esponenziale può essere chiarito osservando che E(X) = $1/\lambda$. Se X descrive un tempo, misurato in minuti, allora λ è una frequenza, misurata in 1/min. Ad esempio, se gli arrivi avvengono in media ogni mezzo minuto, allora E(X) = 0.5 min e $\lambda = 2$, ossia gli arrivi si verificano con una frequenza (tasso di arrivo) di 2 arrivi al minuto. Il parametro λ ha lo stesso significato del parametro λ della distribuzione di Poisson, che descrive invece il numero di arrivi.

La distribuzione esponenziale, così come la distribuzione geometrica, gode della proprietà di "assenza di memoria". Infatti, per ogni s, t reali positivi risulta:

$$P(X > s + t | X > s) = P(X > t)$$

Se si interpreta X come un tempo di attesa, la probabilità condizionata che il tempo di attesa X sia maggiore di t + s dato che essa è maggiore di s non dipende da quanto si è già atteso, ossia da s.

Quando il numero di eventi (ad esempio, numero di arrivi ad un centralino telefonico, numero di arrivi ad un sistema di servizio, ...) è descritto da una distribuzione di Poisson, il tempo tra successivi eventi (ad esempio, tempi tra successivi arrivi, ...) è distribuito esponenzialmente.

8 STIMA PUNTUALE

Inziamo ora a vedere nel dettaglio la stima puntuale, che consiste nello stimare un unico valore per il parametro non noto della popolazione che in questa caso è il parametro λ, avendo una popolazione esponenziale.

Nei metodi di indagine dell'inferenza statistica si considera un campione casuale $X_1, X_2, ..., X_n$ di ampiezza n estratto dalla popolazione e si cerca di ottenere informazioni sui parametri non noti facendo uso di alcune variabili aleatorie, che sono funzioni misurabili del campione casuale, dette statistiche e stimatori.

Una statistica $t(X_1, X_2, ..., X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, ..., X_n$. Essendo la statistica osservabile, i valori da essa assunti dipendono solo dal campione osservato $(x_1, x_2, ..., x_n)$ estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

Uno stimatore $\hat{\theta} = t(X_1, X_2, ..., X_n)$ è quindi funzione misurabile e osservabile del campione casuale $X_1, X_2, ..., X_n$, i cui valori possono essere usati per stimare un parametro ϑ della popolazione. I valori $\hat{\vartheta}$ assunti da tale stimatore sono detti stime del parametro non noto ϑ.

Statistiche tipiche sono la media campionaria e la varianza campionaria.

Sia $X_1, X_2, ..., X_n$ un campione casuale. La statistica

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

è detta media campionaria, mentre la statistica

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

è detta varianza campionaria.

Sia $X_1, X_2, ..., X_n$ un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio E(X) = μ finito e varianza Var(X) = σ^2 finita. Risulta:

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

mentre per la varianza campionaria

$$E(S^2) = \sigma^2 \qquad Var(S^2) = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$$

dove $\mu_4 = E(X^4)$.

Detto ciò vediamo quelli che sono i principali metodi di stima puntuale dei parametri dei parametri non noti, ossia il metodo dei momenti e il metodo della massima verosimiglianza.

8.1 METODO DEI MOMENTI

Per definire il metodo dei momenti occorre definire in primo luogo i momenti campionari. Si definisce momenti campionario r-esimo relativo ai valori osservati $(x_1, x_2, ..., x_n)$ del campione casuale il valore

$$M_r(x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r \qquad (r = 1, 2, ...)$$

Quindi il momento campionario r-esimo è la media aritmetica delle potenze r-esime dell n osservazioni effettuate sulla popolazione. In particolare se r = 1 il momento campionario $M_1(x_1, x_2, ..., x_n)$ coincide con il valore osservato della media campionaria $\bar{X}$, ossia $M_1 = (x_1 + x_2 + ... + x_n)/n$ Se esistono k parametri da stimare, il metodo dei momenti consiste nell'uguagliare i primi k momenti della popolazione in esame con i corrispondenti momenti del campione casuale. Quindi, se i primi k momenti esistono e sono finiti, tale metodo consiste nel risolvere il sistema di k equazioni

$$E(X^r) = M_r(x_1, x_2, ..., x_n) \qquad (r = 1, 2, ..., k)$$

I termini alla sinistra di questo sistema di equazioni dipendono dalla legge di probabilità considerata e contengono i parametri non noti della popolazione. Invece, i termini alla destra possono essere calcolati a partire dai dati osservati del campione estratto dalla popolazione. Le incognite del sistema sono i parametri $\vartheta_1, \vartheta_2, ..., \vartheta_k$ e sono presenti alla sinistra di questo sistema di equazioni. Affinché il metodo sia utilizzabile occorre che il sistema di equazioni ammetta un'unica soluzione.

Le stime dei parametri ottenute con tale metodo, indicate con $\hat{\vartheta}_1, \hat{\vartheta}_2, ..., \hat{\vartheta}_k$ dipendono dal campione osservato $(x_1, x_2, ..., x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, ..., \vartheta_k$ della popolazione, detti stimatori del metodo dei momenti.

8.1.1 APPLICAZIONE DEL METODO DEI MOMENTI

Nel mio caso ho a che fare con una popolazione esponenziale, quindi vado ad applicare il metodo dei momenti per determinare lo stimatore del valore medio $1/\lambda$ della popolazione. In questo caso visto che abbiamo un solo parametro non noto nella legge di probabilità, il metodo dei momenti consiste nell'eguagliare la media della popolazione (ossia E(X) = $1/\lambda$) con il valore osservato della media campionaria. Ponendo quindi $\vartheta = 1/\lambda$, dalla definizione precedente del metodo dei momenti segue:

$$\hat{\vartheta} = \frac{(x_1 + x_2 + ... + x_n)}{n}, \qquad ossia \quad \hat{\lambda} = \frac{1}{\bar{x}}$$

Il metodo dei momenti fornisce quindi come stimatore del parametro $\vartheta = 1/\lambda$ la media campionaria $\bar{X}$, e quindi di conseguenza si ha come stima del parametro λ la quantità $\hat{\lambda} = \frac{1}{\bar{x}}$.

Vado quindi a calcolare la stima del valore medio e del parametro λ in R:

```
stimaMedia <- mean(camp1)
stimaMedia
```

```
## [1] 1.968083
```

```
stimaLambda <- 1/mean(camp1)
stimaLambda
```

```
## [1] 0.5081086
```

8.2 METODO DELLA MASSIMA VEROSIMIGLIANZA

Il metodo della massima verosimiglianza è il più importante metodo per la stima dei parametri non noti di una popolazione. Prima però bisogna introdurre la funzione di verosimiglianza.

Sia $X_1, X_2, ..., X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\vartheta_1, \vartheta_2, ..., \vartheta_k) = L(\vartheta_1, \vartheta_2, ..., \vartheta_k; x_1, x_2, ..., x_n)$ del campione osservato $(x_1, x_2, ..., x_n)$ è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione di densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale $X_1, X_2, ..., X_n$. Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, ..., \vartheta_k$. Tale metodo cerca quindi di determinare da quale funzione di probabilità congiunta (caso discreto) o di densità di probabilità congiunta (caso continuo) è più verosimile che provenga il campione osservato. Pertanto si cercano di determinare i valori di $\vartheta_1, \vartheta_2, ..., \vartheta_k$ che rendono massima la funzione di verosimiglianza e che quindi offrano la migliore spiegazione del campione osservato. In poche parole cerchiamo di trovare i valori dei parametri che possono aver prodotto con la maggiore probabilità i dati osservati.

I valori di $\vartheta_1, \vartheta_2, ..., \vartheta_k$ che rendono massima la funzione sono indicati con $\hat{\vartheta}_1, \hat{\vartheta}_2, ..., \hat{\vartheta}_k$; essi costituiscono le stime di massima verosimiglianza dei parametri non noti $\vartheta_1, \vartheta_2, ..., \vartheta_k$ della popolazione. Tali stime dipendono dal campione osservato $(x_1, x_2, ..., x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, ..., \vartheta_k$ della popolazione, detti stimatori di massima verosimiglianza.

8.2.1 APPLICAZIONE DEL METODO DELLA MASSIMA VEROSIMIGLIANZA

Determiniamo quindi lo stimatore di massima verosimiglianza del valore medio della nostra popolazione. Essendo E(X) = $1/\lambda$, ponendo $\vartheta = 1/\lambda$ si ha :

$$L(\vartheta) = (\frac{1}{\vartheta})^n e^{(-\frac{1}{\vartheta}\sum_{i=1}^{n} x_i)} \qquad (\vartheta > 0)$$

dove le x_i sono positive. Si nota che

$$logL(\vartheta) = -nlog\vartheta - \frac{1}{\vartheta}\sum_{i=1}^{n} x_i \qquad (\vartheta > 0)$$

e quindi per ottenere il massimo di questa funzione andiamo a vedere per quale valore del parametro non noto la derivata è uguale a 0, ottenendo così

$$\frac{dlogL(\vartheta)}{d\vartheta} = -\frac{n}{\vartheta} + \frac{1}{\vartheta^2}\sum_{i=1}^{n} x_i = \frac{n}{\vartheta^2}(\frac{1}{n}\sum_{i=1}^{n} x_i - \vartheta)$$

Possiamo vedere che la derivata vale 0 quando $\vartheta = \frac{1}{n}\sum_{i=1}^{n} x_i$, ossia il valore osservato della media campionaria. La stima di massima verosimiglianza del parametro $\vartheta = 1/\lambda$ è quindi:

$$\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Per cui lo stimatore di massima verosimiglianza e dei momenti del valore medio E(X) = $1/\lambda$ è la media campionaria $\bar{X}$.

8.3 PROPRIETA' DEGLI STIMATORI

In generale esistono molti stimatori che possono essere utilizzati per stimare il parametro non noto di una popolazione. Occorre quindi definire delle proprietà di cui può o meno godere uno stimatore. Uno stimatore può essere:

- corretto (o equivalentemente non distorto),
- più efficiente di un altro,
- corretto e con varianza uniformemente minima,
- asintoticamente corretto,
- consistente.

In questo caso la $\overline{X}$ è per $1/\lambda$:

- uno stimatore corretto (o non distorto): poiché il valore medio coincide con il corrispondente parametro non noto della popolazione;
- con varianza minina: poiché se ci mettiamo nella classe degli stimatori corretti non ci sono altri stimatori con varianza inferiore;
- consistente: al crescere di n la varianza tende a 0.

9 STIMA INTERVALLARE: INTERVALLI DI CONFIDENZA

Alla stima puntuale di un parametro non noto di una popolazione (costituita da un singolo valore) spesso si preferisce sostituire un intervallo di valori, detto intervallo di confidenza (o intervallo di fiducia), ossia si cerca di determinare in base ai dati del campione, due limiti (uno inferiore e uno superiore) entro i quali sia compreso il parametro non noto con un certo coefficiente di confidenza (detto anche grado di fiducia).

Sia $X_1, X_2, ..., X_n$ un campione casuale di ampiezza n estratto da una popolazione con funzione di probabilità (caso discreto) o densità di probabilità (caso continuo) $f(x;\vartheta)$. Denotiamo con $\underline{C}_n = g_1(X_1, X_2, ..., X_n)$ e con $\bar{C}_n = g_2(X_1, X_2, ..., X_n)$ due statistiche (funzioni osservabili del campione casuale) che soddisfano la condizione $\underline{C}_n < \bar{C}_n$, cioè godono delle proprietà che per ogni possibile fissato campione osservato x = $(x_1, x_2, ..., x_n)$ risulti $g_1(x) < g_2(x)$.

Fissato un coefficiente di confidenza $1-\alpha$ $(0 < \alpha < 1)$, se è possibile scegliere le statistiche $\underline{C}_n$ e $\bar{C}_n$ in modo tale che

$$P(\underline{C}_n < \vartheta < \bar{C}_n) = 1 - \alpha$$

allora si dice che $(\underline{C}_n, \bar{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per ϑ. Inoltre le statistiche $\underline{C}_n$ e $\bar{C}_n$ sono dette limite inferiore e superiore dell'intervallo di condifenza.

Mentre l'intervallo $(g_1(x), g_2(x))$ è detto stima dell'intervallo di confidenza di grado $1-\alpha$ In generale esistono diversi intervalli di confidenza dello stesso grado $1-\alpha$ per un parametro non noto, per cui il decisore, fissato un coefficiente di confidenza 1 - α, ad esempio decide l'intervallo che ha lunghezza più piccola possibile o lunghezza media più piccola possibile.

9.1 METODO PIVOTALE

Un metodo per la costruzione degli intervalli di confidenza è il metodo pivotale. Esso consiste nel determinare una variabile aleatoria di pivot $\gamma(X_1, X_2, ..., X_n; \vartheta)$ che

- dipende dal campione casuale $X_1, X_2, ..., X_n$;
- dipende dal parametro non noto ϑ;
- la sua funzione di distribuzione non contiene il parametro ϑ da stimare

Quindi la variabile aleatoria di pivot non è una statistica poiché dipende dal parametro non noto e quindi non è osservabile.

Per ogni fissato coefficiente $\alpha \quad (0 < \alpha < 1)$ siano α_1 e $\alpha_2 \quad (\alpha_1 < \alpha_2)$ due valori dipendenti soltanto dal coefficiente fissato α tali che per ogni $\vartheta \in \theta$ si abbia:

$$P(\alpha_1 < \gamma(X_1, X_2, ..., X_n; \vartheta) < \alpha_2) = 1 - \alpha$$

Se per ogni possibile campione osservato x = $(x_1, x_2, ..., x_n)$ e per ogni $\vartheta \in \theta$, si riesce a dimostrare che

$$\alpha_1 < \gamma(x; \vartheta) < \alpha_2 \iff g_1(x) < \vartheta < g_2(x)$$

con $g_1(x)$ e $g_2(x)$ dipendenti soltanto dal campione osservato, allora è equivalente a richiedere che

$$P(g_1(X_1, X_2, ..., X_n) < \vartheta < g_2(X_1, X_2, ..., X_n)) = 1 - \alpha$$

Quindi segue che $(\underline{C}_n, \bar{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per il parametro non noto ϑ della popolazione.

Quindi in poche parole per costruirci l'intervallo dobbiamo prima individuare una variabile aleatoria di pivot, poi bisogna determinare i valori α_1 e α_2 tali che la probabiltà che la variabile di pivot sia compresa tra questi due valori sia 1 - α. Poi manipoliamo la doppia disugualianza in modo da avere il parametro non noto al centro e ai due estremi due statistiche che costituiscono proprio gli estremi dell'intervallo di confidenza di grado $1-\alpha$.

9.2 INTERVALLI DI FIDUCIA APPROSSIMATI

I metodi per la ricerca degli intervalli di confidenza per una popolazione normale, non dipendono dalla dimensione del campione osservato. Invece per popolazioni che non siano normali, se la dimensione del campione è elevata (n > 30) è possibile utilizzare il teorema centrale di convergenza per determinare un intervallo di confidenza di grado $1-\alpha$ per il parametro non noto ϑ di una popolazione. Infatti, se X denota la variabile aleatoria che descrive la popolazione con E(X) = μ e Var(X) = σ^2 (supposti entrambi finiti) e con $(X_1, X_2, ..., X_n)$ il campione casuale, il teorema centrale di convergenza afferma che la variabile aleatoria

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard.

Si nota che questa variabile può essere interpretata come una variabile aleatoria di pivot poiché:

- dipende dal campione casuale $X_1, X_2, ..., X_n$ attraverso la media campionaria;
- dipende dal parametro non noto ϑ della popolazione attraverso il medio E(X) = μ e la varianza Var(X) = σ^2;
- per grandi campioni la sua funzione di distibuzione è approsimativamente normale standard e quindi non contiene il parametro ϑ da stimare.

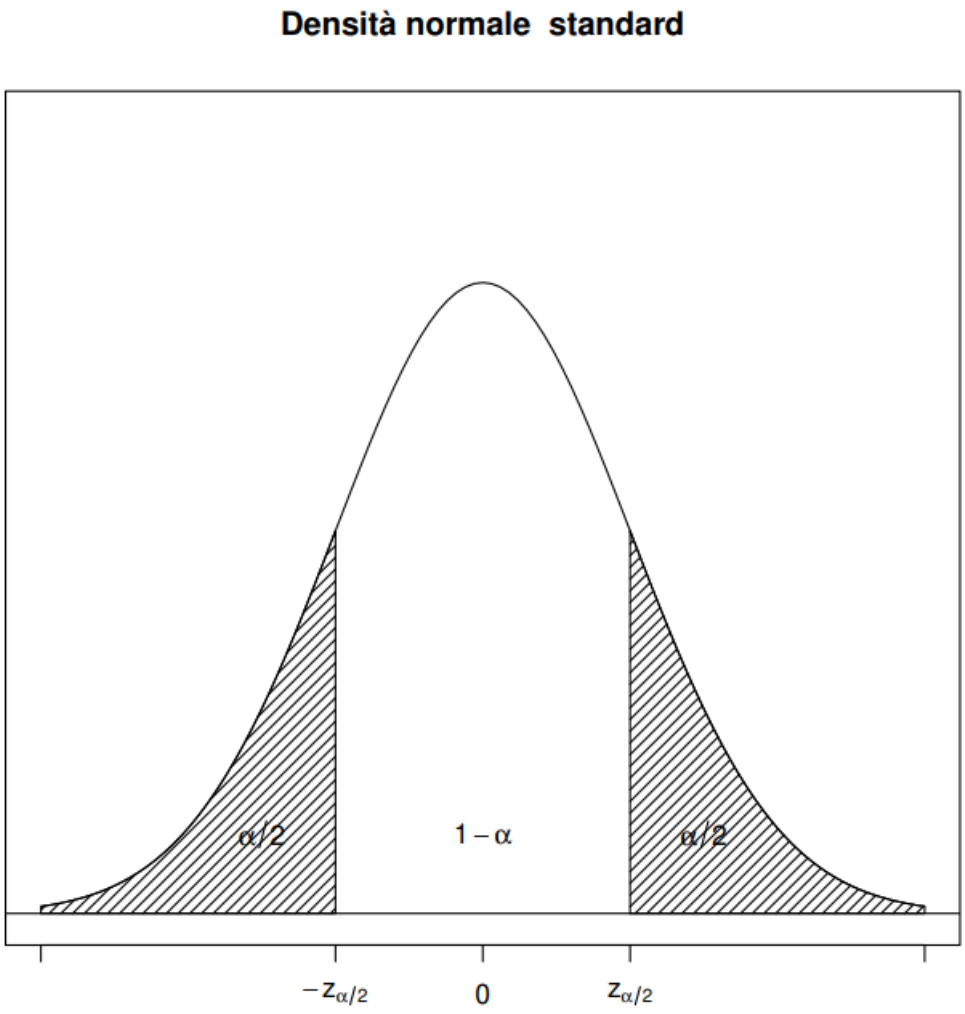
Quindi per campioni di ampiezza elevata possiamo applicare il metodo pivotale in forma approssimata, avendo

$$P(-z_{\alpha/2} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) \approx 1 - \alpha$$

In particolare sfruttando la simmetria della densità normale scegliamo α_1 e α_2 simmetrici intorno allo 0. Quindi abbiamo $\alpha_1 = -z_{\alpha/2}$ e $\alpha_2 = z_{\alpha/2}$, dove $z_{\alpha/2}$ è tale che

$$P(Z_n < -z_{\alpha/2}) = P(Z_n > z_{\alpha/2}) = \frac{\alpha}{2},$$

come possiamo vedere in figura.



9.3 APPLICAZIONE DEL METODO PIVOTALE APPROSSIMATO

Il valore medio di una variabile aleatoria esponenziale abbiamo visto all'inizio che è E(X) = $1/\lambda$ e la varianza è Var(X) = $1/\lambda^2$ ed entrambi dipendono dal parametro non noto λ. Ricaviamo che

$$E(\bar{X}_n) = \frac{1}{\lambda}, \qquad Var(\bar{X}_n) = \frac{1}{n\lambda^2}$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\bar{X}_n - 1/\lambda}{1/(\lambda\sqrt{n})} = \sqrt{n}\frac{\bar{X}_n - 1/\lambda}{1/\lambda} = \sqrt{n}(\lambda\bar{X}_n - 1)$$

converge in distribuzione a una variabile aleatoria normale standard. Per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1-\alpha$ per il parametro $1/\lambda$ può essere determinato richiedendo che

$$P(-z_{\alpha/2} < \sqrt{n}(\lambda\bar{X}_n - 1) < z_{\alpha/2}) \approx 1-\alpha,$$

ossia

$$P\Big\{\bar{X}_n\Big(1+\frac{z_{\alpha/2}}{\sqrt{n}}\Big)^{-1} < \frac{1}{\lambda} < \bar{X}_n\Big(1-\frac{z_{\alpha/2}}{\sqrt{n}}\Big)^{-1}\Big\} \approx 1-\alpha.$$

Quindi sia $(x_1, x_2, ..., x_n)$ un campione osservato di ampiezza n estratto da una popolazione esponenziale di parametro λ. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado $1-\alpha$ per $1/\lambda$ è

$$\bar{x}_n\Big(1+\frac{z_{\alpha/2}}{\sqrt{n}}\Big)^{-1} < \frac{1}{\lambda} < \bar{x}_n\Big(1-\frac{z_{\alpha/2}}{\sqrt{n}}\Big)^{-1}$$

dove $\bar{x}_n$ denota la media campionaria.

Andiamo ora ad applicare il metodo pivatole approssimato per stimare l'intervallo di confidenza di grado $1-\alpha = 0.95$ per il tempo medio tra arrivi successivi di clienti in un negozio. In particolare dal campione estratto di ampiezza 50 abbiamo visto in precedenza che si riscontra un tempo medio di circa 1.96 minuti. Quindi abbiamo che n = 50 (ampiezza del campione), $\bar{x}_{50} = 1.96$ (stima puntuale di $1/\lambda$) e $\alpha = 0.05$. Quindi $\alpha/2 = 0.025$ e $1-\alpha/2 = 0.975$.

```
alpha <- 0.05
n <- 50
alpha2 <- qnorm(1 - alpha/2, mean = 0, sd = 1) #z_alpha/2
alpha2
```

```
## [1] 1.959964
```

```
meanCamp <- 1.96
meanCamp/(1 + alpha2/sqrt(n))
```

```
## [1] 1.53463
```

```
meanCamp/(1 - alpha2/sqrt(n))
```

```
## [1] 2.711605
```

Segue che $z_{\alpha/2} = z_{0.05} = 1.95$ e quindi l'intervallo di confidenza approssimato di grado $1-\alpha = 0.95$ per il parametro $1/\lambda$, ossia il tempo medio tra arrivi successivi, è (1.534,2.711). Si nota che il tempo medio delle 50 osservazioni (1.96), ossia la stima puntuale del valore medio, è contenuto nell'intervallo.

9.4 CONFRONTO TRA DUE POPOLAZIONI

Spesso si è interessati a stimare la differenza tra le medie di due distinte popolazioni. In questo caso occorre quindi costruire un intervallo di confidenza di grado $1-\alpha$ per la differenza tra le due medie.

Considerate due popolazioni, descritte dalle variabili aleatore X e Y indipendenti aventi valori medi E(X) = μ_1 e E(Y) = μ_2 finiti e varianze Var(X) = σ_1^2 e Var(Y) = σ_2^2 finite, la distribuzione della differenza X - Y avrà valore medio e varianza :

$$E(X-Y) = E(X) - E(Y) = \mu_1 - \mu_2, \qquad Var(X-Y) = Var(X) + Var(Y) = \sigma_1^2 + \sigma_2^2$$

Un intervallo di confidenza $(\underline{C}_n, \overline{C}_n)$ per la differenza tra le due medie deve essere tale che

$$P(\underline{C}_n < \mu_1 - \mu_2 < \overline{C}_n) = 1-\alpha$$

dove $\underline{C}_n$ e $\overline{C}_n$ sono due statistiche dipendenti dai campioni estratti dalle due popolazioni. L'intervallo di confidenza stimato $(\underline{c}_n, \overline{c}_n)$ può essere così interpretato:

- se il limite inferiore e il limite superiore sono entrambi negativi allora $\mu_1 - \mu_2 < 0$; ciò implica che la media della prima popolazione è inferiore alla media della seconda popolazione con un grado di confidenza $1-\alpha$;
- se il limite inferiore e il limite superiore sono entrambi positivi allora $\mu_1 - \mu_2 > 0$; ciò implica che la media della prima popolazione è superiore alla media della seconda popolazione con un grado di confidenza $1-\alpha$;
- se l'intervallo contiene lo zero, ossia il limite inferiore risulta negativo e il limite superiore positivo, allora con un grado di confidenza $1-\alpha$ non si può affermare che la media di una popolazione sia superiore alla media dell'altra popolazione.

9.4.1 CONFRONTO TRA DUE POPOLAZIONI ESPONENZIALI

Consideriamo una prima popolazione esponenziale descritta da una variabile aleatoria $X \sim P(\lambda_1)$ con densità di probabilità

$$p_x(x) = \lambda_1 e^{-\lambda_1 x}, \qquad x > 0 \quad (\lambda_1 > 0)$$

ed una seconda popolazione esponenziale descritta da una variabile aleatoria $X \sim P(\lambda_2)$ con densità di probabilità

$$p_x(x) = \lambda_2 e^{-\lambda_2 x}, \qquad x > 0 \quad (\lambda_2 > 0)$$

e siano $X_1, X_2, ..., X_{n1}$ e $Y_1, Y_2, ..., Y_{n2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti dalle due popolazioni esponenziali. Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $\frac{1}{\lambda_1} - \frac{1}{\lambda_2}$ tra i parametri delle due popolazioni per grandi valori di n_1 e n_2. Denotiamo con $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ rispettivamente le medie campionarie delle due popolazioni. Dal teorema centrale di convergenza segue che la variabile aleatoria

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\frac{1}{\lambda_1} - \frac{1}{\lambda_2})}{\sqrt{\frac{1}{n_1\lambda_1^2} + \frac{1}{n_2\lambda_2^2}}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard. Poiché

$$E(\overline{X}_{n_1}) = 1/\lambda_1, \quad \lim_{n_1 \to +\infty} Var(\overline{X}_{n_1}) = 0, \quad E(\overline{Y}_{n_2}) = 1/\lambda_2, \quad \lim_{n_2 \to +\infty} Var(\overline{Y}_{n_2}) = 0$$

ossia le medie campionarie $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ sono stimatori corretti e consistenti per i parametri $1/\lambda_1$ e $1/\lambda_2$, per campioni sufficientemente numerosi possiamo sostituire $1/\lambda_1$ e $1/\lambda_2$ al denominatore con le medie campionarie e quindi l'intervallo di confidenza di grado $1-\alpha$ per la differenza $1/\lambda_1 - 1/\lambda_2$ può essere determinato supponendo che

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\frac{1}{\lambda_1} - \frac{1}{\lambda_2})}{\sqrt{\frac{\overline{X}_{n_1}^2}{n_1} + \frac{\overline{Y}_{n_2}^2}{n_2}}} < z_{\alpha/2}\right) \approx 1-\alpha$$

ossia

$$P\left(\overline{X}_{n_1} - \overline{Y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\overline{X}^2_{n_1}}{n_1} + \frac{\overline{Y}^2_{n_2}}{n_2}} < \frac{1}{\lambda_1} - \frac{1}{\lambda_2} < \overline{X}_{n_1} - \overline{Y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\overline{X}^2_{n_1}}{n_1} + \frac{\overline{Y}^2_{n_2}}{n_2}}\right) \approx 1 - \alpha$$

Quindi siano $(x_1, x_2, ..., x_{n_1})$ e $(y_1, y_2, ..., y_{n_2})$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni esponenziale di parametro λ_1 e λ_2 . Una stima approssimata dell'intervallo di confidenza di grado $1 - \alpha$ per la differenza $1/\lambda_1 - 1/\lambda_2$ è

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\overline{x}^2_{n_1}}{n_1} + \frac{\overline{y}^2_{n_2}}{n_2}} < \frac{1}{\lambda_1} - \frac{1}{\lambda_2} < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\overline{x}^2_{n_1}}{n_1} + \frac{\overline{y}^2_{n_2}}{n_2}}$$

dove $\overline{x}_{n_1}$ e $\overline{y}_{n_2}$ denotano rispettivamente le medie campionarie delle due osservazioni.

Supponiamo quindi di avere due popolazioni descritte da una variabile aleatoria esponenziale. Ad esempio ci sono due negozi A e B analizzati in base ai tempi di interarrivo dei clienti, dove il negozio A è quello che abbiamo considerato fin ora e di cui abbiamo estratto un campione di 50 osservazioni. Mentre per il negozio B sono state effettuate 40 osservazioni. Di seguito sono riportati i campioni delle rispettive popolazioni

```
set.seed(1)
camp1 <- rexp(n = 50, rate = 0.5)
camp1
```

```
##  [1] 1.51036367 2.36328556 0.29141345 0.27959052 0.87213725 5.78993707
##  [7] 2.45912411 1.07936568 1.91313499 0.29409198 2.78147026 1.52405971
## [13] 2.47520710 8.84786844 2.10908633 2.07048789 3.75207034 1.30949327
## [19] 0.67386695 1.17695944 4.72903051 1.28378518 0.58824078 1.13173105
## [25] 0.21214525 0.11887832 1.15742493 7.91786570 2.34662421 1.99362591
## [31] 2.87057069 0.07453705 0.64802031 2.64093586 0.40702070 2.04545175
## [37] 0.60348187 1.45042861 1.50308538 0.47005490 2.15976227 2.05649381
## [43] 2.58452330 2.50621071 1.10928280 0.60256599 2.58624931 1.98911158
## [49] 1.02834859 4.01566480
```

```
mean(camp1)
```

```
## [1] 1.968083
```

```
camp2 <- rexp(n = 40, rate = 0.5)
camp2
```

```
##  [1] 0.8444849 4.3575451 6.4355780 1.1156587 1.1892353 1.9547916 0.4197332
##  [8] 0.6188957 2.2118725 1.5483755 0.1793482 2.2163533 0.4945285 3.1439737
## [15] 9.6656255 0.8622643 5.4607786 2.2736628 1.6267365 1.6740130 3.5695308
## [22] 4.6249433 5.8197745 0.5711820 0.7775735 0.1041109 0.7037410 3.1304827
## [29] 1.6290716 5.5184876 0.7723871 2.0165129 1.6370284 0.1185224 4.5677069
## [36] 1.6083418 3.1673922 2.4675830 2.6912880 4.2007545
```

```
mean(camp2)
```

```
## [1] 2.449747
```

Supponendo che i tempi tra arrivi successivi dei clienti al negozio A siano descritti da una variabile aleatoria esponenziale $X \sim P(\lambda_1)$ e i tempi tra arrivi successivi dei clienti al negozio B siano descritti da una variabile aleatoria esponenziale $Y \sim P(\lambda_2)$, andiamo a determinare l'intervallo di confidenza per $1/\lambda_1 - 1/\lambda_2$ di grado $1 - \alpha = 0.95$.

```
alpha <- 0.05
alpha2 <- qnorm(1 - alpha/2, mean = 0, sd = 1) #z_alpha/2
n1 <- length(camp1)
n2 <- length(camp2)
m1 <- mean(camp1)
m2 <- mean(camp2)
rad <- sqrt(m1^2/n1 + m2^2/n2)

m1 - m2 - alpha2*rad #stima estremo inferiore intervallo
```

```
## [1] -1.416504
```

```
m1 - m2 + alpha2*rad #stima estremo superiore intervallo
```

```
## [1] 0.4531768
```

Quindi una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per $1/\lambda_1 - 1/\lambda_2$ è (-1.4165, 0.4532). Visto che l'intervallo include anche lo zero, ossia la possibilità che $1/\lambda_1 = 1/\lambda_2$, non si può concludere che il tempo medio tra arrivi successivi dei clienti in uno dei due negozi sia superiore a quello dell'altro con un grado di fiducia del 95%.

10 VERIFICA DELLE IPOTESI

Le aree più importanti dell'inferenza statistica sono la stima dei parametri e la verifica delle ipotesi. La verifica delle ipotesi interviene spesso nelle ricerche di mercato, nelle indagini sperimentali e industriali, nei sondaggi di opinione, nelle indagini sulle condizioni sociali degli abitanti di una città o di una nazione.

In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x; \vartheta)$, un'ipotesi su di un parametro non noto ϑ della popolazione ed un campione casuale $X_1, X_2, ..., X_n$ estratto dalla popolazione. Occorre in primo luogo precisare il significato di ipotesi statistica.

Un'ipotesi statistica è una congettura o un'affermazione sul parametro non noto ϑ. Se l'ipotesi statistica specifica completamente $f(x; \vartheta)$ è detta ipotesi semplice, altrimenti è chiamata ipotesi composita.

L'ipotesi soggetta a verifica è denotata con H_0 ed è chiamata ipotesi nulla. Si chiama test di ipotesi il procedimento o regola con cui si decide, sulla base dei dati del campione, se accettare o rifiutare H_0. La costruzione del test richiede la formulazione, in contrapposizione all'ipotesi nulla, di una proposizione alternativa. Questa proposizione prende il nome di ipotesi alternativa ed è di solito indicata con H_1. L'ipotesi nulla, cioè l'ipotesi soggetta a verifica, si ha quando $\vartheta \in \theta_0$ e l'ipotesi alternativa si ha quando $\vartheta \in \theta_0$ e si scrive

$$H_0 : \vartheta \in \theta_0, \qquad H_1 : \vartheta \in \theta_1,$$

avendo denotato con θ_0 e θ_1 due sottoinsiemi disgiunti dello spazio θ dei parametri.

Il problema della verifica delle ipotesi consiste nel determinare un test ψ che permetta di suddividere, mediante opportuni criteri, l'insieme dei possibili campioni in due sottoinsiemi: una regione di accettazione A dell'ipotesi nulla ed una regione di rifiuto R dell'ipotesi nulla. Il test ψ può allora essere così formulato:

accettare come valida l'ipotesi nulla se il campione osservato $(x_1, x_2, ..., x_n) \in A$ e rifiutare l'ipotesi nulla se $(x_1, x_2, ..., x_n) \in R$. Nel caso si verifichi che l'ipotesi nulla sia falsa, l'ipotesi alternativa sarà vera e viceversa. Nel seguire questo tipo di ragionamento si può incorrere in due tipi di errori:

- rifiutare l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia vera; si dice allora che si commette un errore di tipo 1 e si denota la probabilità di commettere tale errore con

$$\alpha(\vartheta) = P(rifiutare H_0|\vartheta), \qquad \vartheta \in \theta_0$$

- accettare l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia falsa; si dice allora che si commette un errore di tipo 2 e si denota la probabilità di commettere tale errore con

$$\beta(\vartheta) = P(accettare H_0|\vartheta), \qquad \vartheta \in \theta_1$$

La probabiltà massima di commettere un errore del primo tipo viene detta misura della regione critica del test ψ (o livello di significatività del test ψ). In generale per campioni casuali di fissata ampiezza, se si diminuisce la probabilità di commettere un errore di tipo 1 aumenta la probabilità di commettere un errore di tipo 2 e viceversa. Nella costruzione del test conviene quindi fissare la probabilità di commettere un errore di tipo 1 e cercare un test ψ che minimizzi la probabilità di commettere un errore di tipo 2. La giustificazione del fissare la probabilità di commettere un errore di 1 tipo (che solitamente si sceglie piccola) deriva dal fatto che di solito le ipotesi vengono formulate in maniera tale che l'errore di tipo 1 sia più grave e quindi il decisore desidera imporre che la probabilità di commettere tale errore sia piccola. Ad esempio, nell'ambito giudiziario scegliere come ipotesi nulla "l'imputato è innocente" significa ritenere che condannare un innocente sia un errore più grave che assolvere un colpevole.

Solitamente la probabilità di commettere un errore di tipo 1 si sceglie uguale a 0.05, 0.01, 0.001 ed il test viene rispettivamente detto statisticamente significativo, statisticamente molto significativo e statisticamente estremamente significativo. Infatti, quanto minore è il valore di α tanto maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla. I test statistici sono di due tipi:

- test bilaterali (detti anche test bidirezionali);
- test unilaterali (detti anche test unidirezionali).

Un test bilaterale è il seguente

$$H_0 : \vartheta = \vartheta_0 \qquad H_1 : \vartheta \neq \vartheta_0$$

mentre il test unilaterale sinistro e test unilaterale destro sono rispettivamente i seguenti

$$H_0 : \vartheta \leq \vartheta_0 \qquad H_1 : \vartheta > \vartheta_0$$

$$H_0 : \vartheta \geq \vartheta_0 \qquad H_1 : \vartheta < \vartheta_0$$

avendo fissato a priori il livello di significatività α. Le conclusioni dei test statistici unilaterali e bilaterali dipendono dal livello di significatività α, scelto a priori dal decisore per verificare l'ipotesi nulla H_0. Spesso, nei test statistici si calcola anche il livello di significatività osservato, noto come p-value. Il p-value si basa su una statistica del test, che dipende dal campione osservato e dal test statistico considerato. Il p-value è definito come la probabilità, supposta vera l'ipotesi H_0, che la statistica del test assuma un valore uguale o più estremo di quello effettivamente osservato. Essendo una probabilità il p-value è un numero compreso tra 0 e 1. Calcolando il p-value è possibile comportarsi come segue:

- se p $> \alpha$, l'ipotesi H_0 non può essere rifiutata;
- se $p \leq \alpha$, l'ipotesi H_0 deve essere rifiutata.

Nel condurre un test statistico è importante fissare il livello di significatività α prima di calcolare il p-value. Se si calcola prima il p-value, il decisore potrebbe scegliere il livello di significatività α in funzione del risultato desiderato in modo da accettare o rigettare l'ipotesi nulla H_0.

10.1 TEST STATISTICI PER IL VALORE MEDIO PER GRANDI CAMPIONI

Quando l'ampiezza del campione è grande, per una popolazione descritta da una variabile aleatoria X caratterizzata da valore medio μ e varianza σ^2, entrambi finiti, si può utilizzare il teorema centrale di convegenza ricordando che la variabile aleatoria

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$$

converge in distribuzione a una variabile aleatoria normale standard.

Test bilaterale approssimato: Per campioni numerosi, il test bilaterale di misura α per le ipotesi

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

considera come variabile aleatoria

$$\frac{\overline{X}_n - \mu_0}{\sigma_0/\sqrt{n}},$$

dove σ_0 è la deviazione standard della popolazione quando $\mu = \mu_0$. Tale variabile aleatoria deve dipendere soltanto dal campione casuale e costituisce la statistica del test. Il test bilaterale di misura α è il seguente:

- si accetti H_0 se $\quad -z_{\alpha/2} < \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_{\alpha/2}$
- si rifiuti H_0 se $\quad \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < -z_{\alpha/2} \quad \textit{oppure} \quad \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_{\alpha/2}$

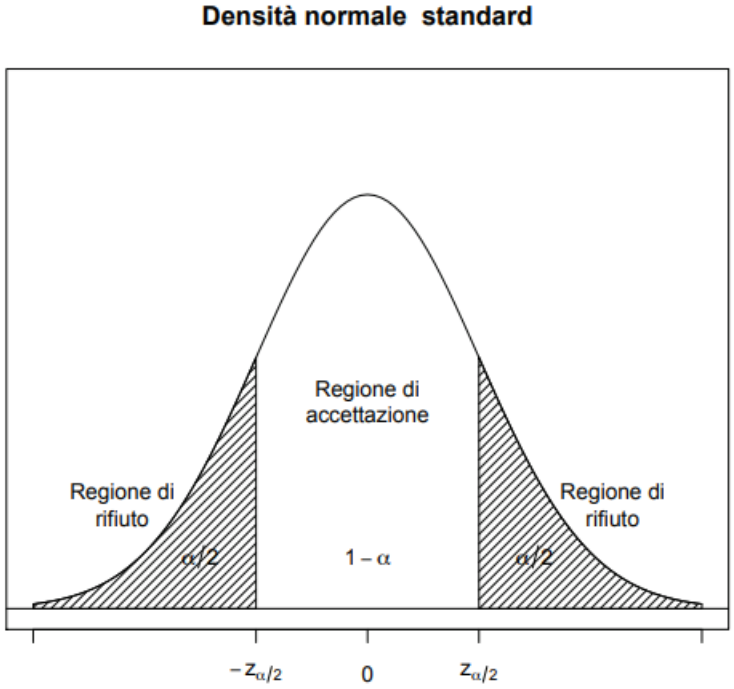


Figura 13.13: Test bilaterale

dove $z_{\alpha/2}$ rappresenta sempre il valore x assunto dalla variabile aleatoria X tale che $P(X \leq x) = 1 - \alpha/2$, calcolato in R con qnorm($1 - \alpha/2$, mean = 0, sd = 1).

Mentre il p-value per il test bilaterale approssimato corrisponde a:

$$p - value \approx P(Z_n < -|z_{os}|) + P(Z_n > |z_{os}|) = 2P(Z_n > |z_{os}|) = 2[1 - P(Z_n \leq |z_{os}|)]$$

Quindi in R può essere cosi calcolato:

$$2 * (1 - pnorm(abs(z_{os}), mean = 0, sd = 1))$$

Test unilaterale sinistro approssimato:

- si accetti H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_\alpha$
- si rifiuti H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha$

dove z_α viene calcolato in R con qnorm($1 - \alpha$, mean $= 0$, sd $= 1$). Il p-value per il test unilaterale sinistro approssimato corrisponde a:

$$p - value \approx P(Z_n > z_{os}) = (1 - P(Z_n \leq z_{os})$$

Quindi in R può essere cosi calcolato:

$$1 - pnorm(z_{os}, mean = 0, sd = 1)$$

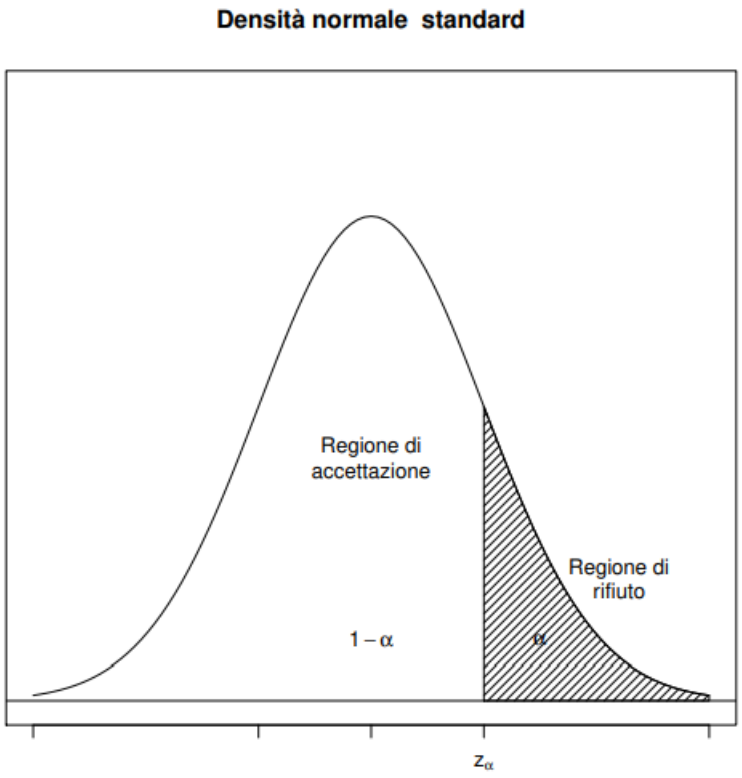


Figura 13.14: Test unilaterale sinistro

Test unilaterale destro approssimato:

- si accetti H_0 se $\quad \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > -z_\alpha$
- si rifiuti H_0 se $\quad \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < -z_\alpha$

dove -z_α viene calcolato in R con qnorm(α, mean = 0, sd = 1).

Il p-value per il test unilaterale destro approssimato corrisponde a:

$$p - value \approx P(Z_n \leq z_{os})$$

Quindi in R può essere cosi calcolato:

$$pnorm(z_{os}, mean = 0, sd = 1)$$

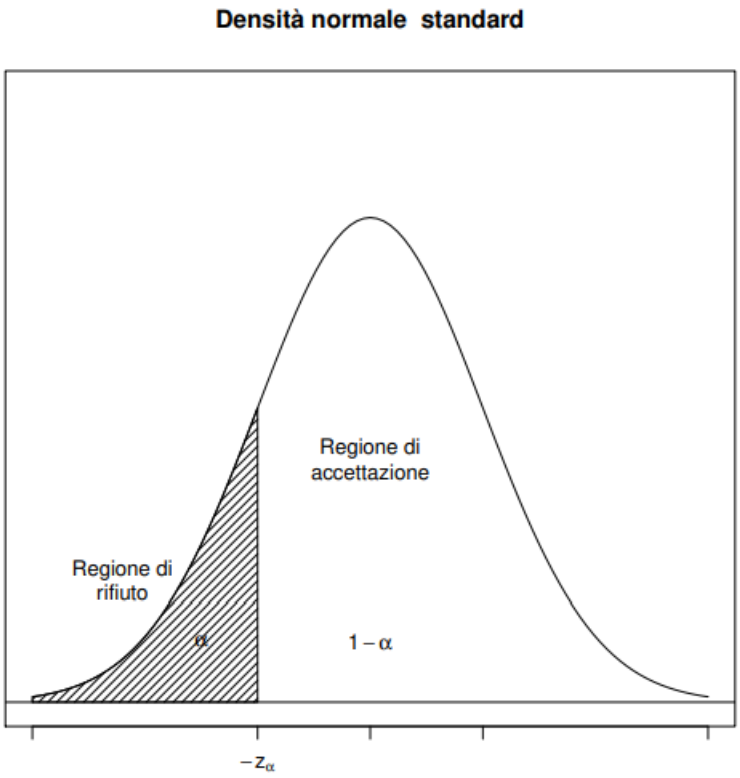


Figura 13.15: Test unilaterale destro

10.2 VERIFICA IPOTESI PER POPOLAZIONE ESPONENZIALE

Andiamo ora a costruire dei test unilaterali e bilaterali per il valore medio E(X) = $1/\lambda$ della nostra popolazione esponenziale.

Il test bilaterale può essere cosi formulato:

$$H_0 : \frac{1}{\lambda} = \frac{1}{\lambda_0} \qquad H_1 : \frac{1}{\lambda} \neq \frac{1}{\lambda_0}$$

,

mentre il test unilaterale sinistro e quello destro sono rispettivamente i seguenti:

$$H_0 : \frac{1}{\lambda} \leq \frac{1}{\lambda_0} \qquad H_1 : \frac{1}{\lambda} > \frac{1}{\lambda_0}$$

$$H_0 : \frac{1}{\lambda} \geq \frac{1}{\lambda_0} \qquad H_1 : \frac{1}{\lambda} < \frac{1}{\lambda_0}$$

avendo fissato a priori un livello di significatività α. Essendo $\mu_0 = 1/\lambda_0$ e $\sigma_0^2 = 1/\lambda_0^2$, nei test unilaterali e bilaterali occorre considerare come statistica del test la seguente variabile aleatoria

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\overline{x}_n - \frac{1}{\lambda_0}}{\sqrt{\frac{1}{n\lambda_0^2}}} = \sqrt{n}(\lambda_0 \overline{x}_n - 1)$$

Nelle 50 osservazioni della popolazione esponenziale considerata abbiamo visto che in media il tempo tra gli arrivi successivi dei clienti è di circa 2 minuti. Abbiamo poi visto che una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro $\frac{1}{\lambda}$ è (1.534,2.711). Detto ciò andiamo a costruire i test bilaterali e unilaterali per il paremetro $\frac{1}{\lambda}$.

10.2.1 TEST BILATERALE

Verifichiamo l'ipotesi $H_0 : 1/\lambda = 2.2$ in alternativa a $H_1 : 1/\lambda \neq 2.2$ con un livello di significatività $\alpha = 0.05$.

```
lambda0 <- 1/2.2
alpha <- 0.05
n <- 50
qnorm(1 - alpha/2, mean = 0, sd = 1) #z_alpha/2
```

```
## [1] 1.959964
```

```
zos <- sqrt(n)*(lambda0*m1- 1)
zos
```

```
## [1] -0.7454084
```

Quindi $z_{\alpha/2} = 1.959964$ e la stima della statistica del test $z_{os} = -0.7454084$ cade nella regione di accettazione. Occorre quindi accettare l'ipotesi nulla che $1/\lambda = 2.2$ con un livello di significatività del 5%. Mentre il p-value per il test bilaterale considerato corrisponde a:

$$p - value = P(Z_n < -|z_{os}|) + P(Z_n > |z_{os}|) = 2P(Z_n > |z_{os}|) = 2[1 - P(Z_n \leq |z_{os}|)]$$

Quindi in R può essere cosi calcolato:

```
2*(1 - pnorm(abs(zos), mean = 0, sd = 1))
```

```
## [1] 0.4560248
```

Si nota che p-value $> \alpha$ e quindi anche il criterio del p-value consiglia di accettare l'ipotesi nulla.

10.2.2 TEST UNILATERALE SINISTRO

Verifichiamo ora l'ipotesi $H_0 : 1/\lambda \leq 1.45$ in alternativa a $H_1 : 1/\lambda > 1.45$ con un livello di significatività $\alpha = 0.05$

```r
lambda0_2 <- 1/1.45
alpha <- 0.05
qnorm(1 - alpha, mean = 0, sd = 1) #z_alpha
```

```
## [1] 1.644854
```

```r
zos_2 <- sqrt(n)*(lambda0_2*m1- 1)
zos_2
```

```
## [1] 2.526484
```

Si ha che $z_\alpha = 1.644854$ e $z_{os} = 2.526484$ cade nella regione di rifiuto. Occorre quindi rifiutare l'ipotesi nulla che $1/\lambda \leq 1.45$.

```r
1 - pnorm(zos_2, mean = 0 ,sd = 1)
```

```
## [1] 0.005760526
```

Ovviamente si nota anche pvalue $< \alpha$ e quindi anche il criterio del p-value consiglia di rifiutare l'ipotesi nulla.

10.2.3 TEST UNILATERALE DESTRO

Verifichiamo ora l'ipotesi $H_0 : 1/\lambda \geq 1.8$ in alternativa a $H_1 : 1/\lambda < 1.8$ con un livello di significatività $\alpha = 0.05$

```r
lambda0_3 <- 1/1.8
alpha <- 0.05
qnorm(alpha, mean = 0, sd = 1) #-z_alpha
```

```
## [1] -1.644854
```

```r
zos_3 <- sqrt(n)*(lambda0_3*m1- 1)
zos_3
```

```
## [1] 0.6602937
```

Si ha che z_α = -1.644854 e z_{os} = 0.6602937 cade nella regione di accettazione. Occorre quindi accettare l'ipotesi nulla che $1/\lambda \geq 1.80$.

```r
pnorm(zos_3, mean = 0 ,sd = 1)
```

```
## [1] 0.7454673
```

Ovviamente si nota anche pvalue $> \alpha$ e quindi anche il criterio del p-value consiglia di accettare l'ipotesi nulla.

11 CRITERIO DEL CHI-QUADRATO

In molti problemi reali, si desidera verificare se il campione osservato può essere stato estratto da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione $F_X(x)$. A questo scopo, viene utilizzato il criterio di verifica delle ipotesi del chi–quadrato, detto anche test del chi–quadrato o test del buon adattamento.

Con il criterio del chi–quadrato si desidera verificare l'ipotesi che un certa popolazione, descritta da una variabile aleatoria X, sia caratterizzata da una funzione di distribuzione $F_X(x)$, con k parametri non noti da stimare. Il test chi–quadrato con livello di significatività α mira a verificare l'ipotesi nulla

H_0: X ha una funzione di distribuzione $F_X(x)$ (avendo stimato k parametri non noti in base al campione)

in alternativa all'ipotesi

H_1: X non ha una funzione di distribuzione $F_X(x)$.

Occorre determinare un test con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. Il test di verifica delle ipotesi considerato è bilaterale.

Inanzitutto bisogna suddividere l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi $I_1, I_2, ..., I_r$ in modo che risulti essere uguale a p_i la probabilità che, secondo la distribuzione ipotizzata, la variabile aleatoria assuma un valore appartente ad I_i, ossia

$$p_i = P(X \in I_i) \qquad (i = 1, 2, ..., r)$$

Si estrae poi un campione $x_1, x_2, ..., x_n$ di ampiezza n e si osservano le frequenze assolute $n_1, n_2, ..., n_r$ con cui gli n elementi si distribuiscono nei rispettivi sottoinsiemi.

Il numero medio di elementi che cadono in I_i è np_i. Si calcola poi la quantità

$$X^2 = \sum_{i=1}^{r} \left(\frac{n_i - np_i}{\sqrt{np_i}} \right)^2$$

Il criterio del chi-quadrato si basa sulla statistica

$$Q = \sum_{i=1}^{r} \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)^2,$$

dove N_i è la variabile aleatoria che descrive il numero di elementi del campione casuale $X_1, X_2, ..., X_n$ che cadono nell'i-esimo intervallo.

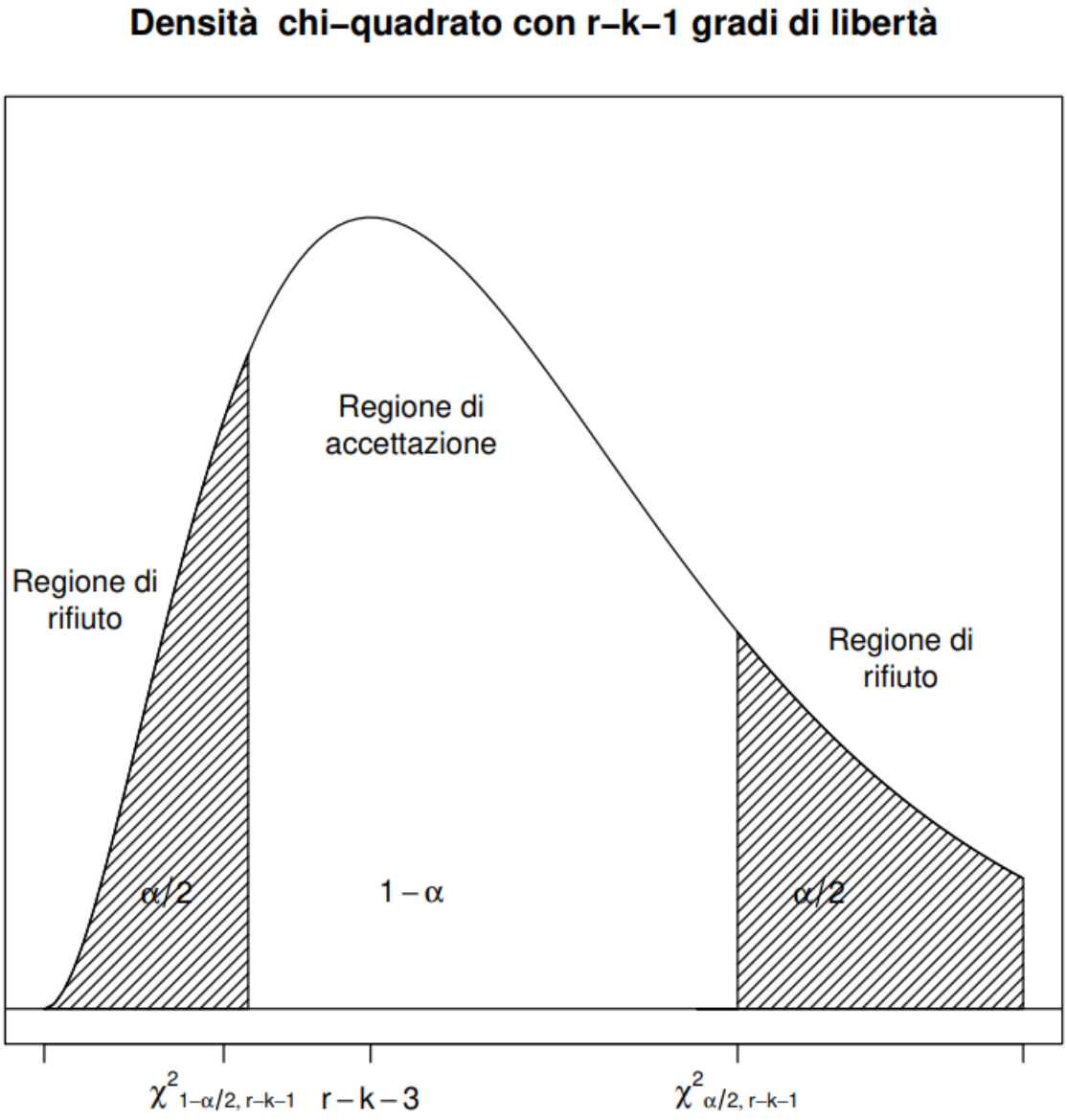
Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti, si può dimostrare che per n sufficientemente grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con r-k-1 gradi di libertà. Si sottrae 1 da r poiché se conosciamo r - 1 delle probabilità p_i la rimanente probabilità può essere univocamente determinata e si sottrae k poiché si suppone che siano k i parametri indipendenti non noti sostituiti da stime.

Inoltre la suddivisione in classi deve essere tale che ogni classe deve contenere in media almeno 5 elementi, quindi deve risultare che

$$min(np_1, np_2, ..., np_r) \geq 5$$

Quindi per un campione sufficientemente numeroso di ampiezza n, il test chi–quadrato bilaterale di misura α è il seguente:

- si accetti H_0 se $X^2_{\alpha/2,r-k-1} < X^2 < X^2_{1-\alpha/2,r-k-1}$
- si rifiuti H_0 se $X^2 < X^2_{\alpha/2,r-k-1} \quad oppure \quad X^2 > X^2_{1-\alpha/2,r-k-1}$



Andiamo quindi ad applicare il test chi-quadrato di misura $\alpha = 0.01$ per verificare se la popolazione da cui proviene il nostro campione può essere descritta da una variabile aleatoria X di densità esponenziale

Suddividiamo i valori assumibili da X in 4 sottoinsiemi di uguale area, ossia tale che $p_i = 0.25$ per i = 1,2,3,4. Abbiamo almeno 5 elementi per ogni sottoinsieme essendo $np_i = 50 * 0.25 = 12.5 \geq 5$. Utilizziamo i quantili della distribuzione esponenziale per determinare i 4 sottoinsiemi, andando a sostituire il parametro non noto λ con il valore stimato tramite la stima puntuale.

```
a <- numeric(3)
for (i in 1:3)
  a[i] <- qexp(0.25*i,rate = stimaLambda)
a
```

```
## [1] 0.5661823 1.3641714 2.7283428
```

I sottoinsiemi $I_1, I_2, I_3, I4$ sono dunque i seguenti:

$$I_1 = (0, 0.57), \quad I_2 = [0.57, 1.36), \quad I_3 = [1.36, 2.73), \quad I_4 = [2.73, +\infty)$$

Andiamo a calcolare le frequenze assolute delle varie classi

```
r <-4
nint <- numeric(r)
nint[1] <- length(which(camp1 < a[1]))
nint[2] <- length(which((camp1 >= a[1]) & (camp1 <a[2])))
nint[3] <- length(which((camp1 >= a[2]) & (camp1 <a[3])))
nint[4] <- length(which(camp1 >= a[3]))
nint
```

```
## [1]  8 14 20  8
```

Calcoliamo poi X^2

```
n <- 50
chi2 <- sum(((nint - n*0.25)/sqrt(n*0.25))^2)
chi2
```

```
## [1] 7.92
```

La distribuzione esponenziale ha un solo parametro non noto (λ) e quindi k = 1. Quindi la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con r-k-1 = 2 gradi di libertà. Bisogna quindi calcolare $X^2_{1-\alpha/2,2}$ e $X^2_{\alpha/2,2}$

```
k <- 1
alpha <- 0.01
qchisq(alpha/2, df = r-k-1)
```

```
## [1] 0.01002508
```

```
qchisq(1 - alpha/2, df = r-k-1)
```

```
## [1] 10.59663
```

Quindi si ha $X^2_{\alpha/2,2} = 0.010$ e $X^2_{1-\alpha/2,2} = 10.597$. Essendo $0.010 < X^2 < 10.597$ l'ipotesi H_0 di popolazione esponenziale può essere accettata.