UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

CORSO DI LAUREA MAGISTRALE IN INFORMATICA

Statistica e Analisi dei Dati

Docente del corso:
Amelia Giuseppina Nobile

Studentesse:
Castaldo Giusy 0522501106
Cerullo Mary 0522501418

ANNO ACCADEMICO 2022/2023

Sommario

PRIMA PARTE:
Statistica e Analisi dei Dati

CAPITOLO 1: INTRODUZIONE

1.1 SET DI DATI

La prima parte dello studio è incentrata sulla scelta del set di dati da cui poter ricavare alcune informazioni. In particolare, il Set di Dati scelto è stato reperito dal sito dell'ISTAT.
L'indagine è rivolta a un campione di laureati e approfondisce la loro condizione lavorativa e il loro percorso occupazionale a distanza di alcuni anni dal conseguimento della laurea di I livello. L'indagine fa parte del sistema di rilevazioni sulla transizione istruzione-lavoro. I dati analizzati riferiscono all'anno 2015 relativamente ai laureati che hanno conseguito il titolo di I livello nell'anno 2011.

1.1.1 IDENTIFICATIVI E DATA FRAME

L'oggetto scelto per la rappresentazione dei dati è un Data Frame: un oggetto di tipo lista che si presenta in forma tabellare. È costituito da righe e colonne in cui ogni riga individua un'osservazione e ad ogni colonna corrisponde una variabile.
Le colonne di un Data Frame hanno tutte la stessa lunghezza e i valori contenuti in ogni singola colonna sono omogenei (dello stesso tipo).
Per una maggiore comprensione del Set di Dati saranno utilizzati degli identificativi associati ai gruppi di laurea e alle condizioni occupazionali.
Per i gruppi di laurea:

- SC: scientifico
- CF: chimico-farmaceutico
- GB: geo-biologico
- MD: medico
- ING: ingegneria
- ARC: architettura
- AGR: agrario
- ES: economico-statistico
- PS: politico-sociale
- GIU: giuridico
- LET: letterario
- LING: linguistico
- INS: insegnamento
- PSI: psicologico
- EF: educazione fisica
- DS: difesa e sicurezza

Per condizione occupazionale:

- LPT: lavora avendo iniziato prima del conseguimento del titolo
- LDT: lavora avendo iniziato dopo il conseguimento del titolo
- CL: cercano lavoro
- NCL: non cercano lavoro

Relativamente ai dati presi in esame, a questi ultimi sono associati valori in percentuale.
Di seguito viene riportato il Data Frame ottenuto:

```
> occupazione
      LPT  LDT   CL  NCL
SC   10.6 69.9  9.5 10.0
CF    8.3 65.3 20.1  6.3
GB    6.7 51.9 31.1 10.3
MD   12.8 72.8 11.7  2.8
ING   9.3 65.8 15.9  9.1
ARC  10.0 55.8 25.4  8.8
AGR  15.6 60.1 19.6  4.7
ES   12.4 66.0 16.1  5.4
PS   23.8 47.6 22.5  6.1
GIU  34.0 36.1 21.5  8.4
LET  13.6 48.1 27.2 11.1
LING  8.6 61.5 21.1  8.8
INS  32.4 42.0 19.3  6.4
PSI  18.0 36.5 29.3 16.3
EF   26.5 52.8 15.4  5.3
DS   51.8 40.8  7.3  0.0
```

1.1.2 STATISTICHE E TIPO DI DATI

La statistica descrittiva è costituita da un insieme di metodi di natura logica e matematica atti a raccogliere, elaborare, analizzare ed interpretare dati allo scopo di descrivere fenomeni collettivi e di estendere la descrizione di certi fenomeni osservati ad altri fenomeni dello stesso tipo non ancora osservati.
Questi fenomeni possono riguardare particolari aspetti del mondo reale di natura economica, industriale, sociale oppure descrivere comportamenti o situazioni riguardanti singoli individui o insiemi di individui.
È utilizzata per analizzare il comportamento dei fenomeni oggetto di studio.
Ogni fenomeno può essere descritto tramite opportune categorie di dati di tipo qualitativo (attraverso stringhe di caratteri oppure mediante opportune classi) oppure di tipo quantitativo (rappresentati tramite valori numerici) 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 che si vuole analizzare.
Nel nostro caso:

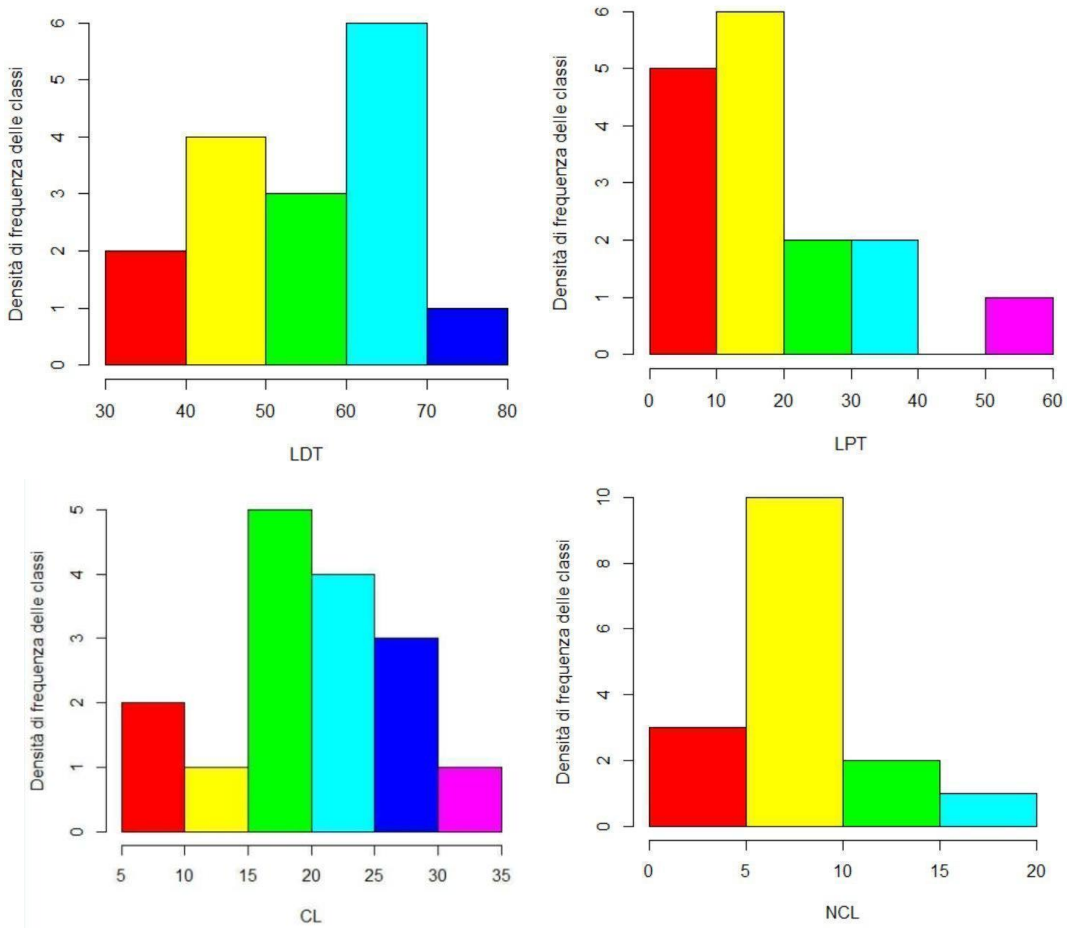
- la natura del fenomeno preso in esame corrisponde al tasso occupazionale in merito al possedimento o meno di un titolo di I livello in uno specifico gruppo di laurea;
- il numero di osservazioni disponibili (ampiezza del campione): 16 gruppi di laurea;
- il numero di variabili utilizzate per rappresentare i diversi aspetti del fenomeno in esame (numero di caratteristiche): 4;
- il tipo di informazione disponibile per ciascuna variabile: quantitative;
- lo scopo che l'analisi esplorativa dei dati si propone di raggiungere: comprendere la correlazione tra il possedimento di un titolo di laurea di primo I livello in uno specifico gruppo di laurea e il tasso occupazionale.

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.
La conoscenza delle caratteristiche di una popolazione finita può essere ottenuta osservando la totalità delle entità della popolazione oppure un sottoinsieme di questa, detto campione estratto dalla popolazione.
Una popolazione illimitata può, invece, essere studiata solo tramite un campione estratto dalla popolazione.
Poiché il campione deve contenere informazioni sulla popolazione complessiva, deve essere rappresentativo di quella popolazione.
In generale, 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.

CAPITOLO 2: RAPPRESENTAZIONE GRAFICA DEI DATI

2.1 ISTOGRAMMI

Gli istogrammi sono utilizzati per le variabili quantitative e sono una particolare rappresentazione grafica ottenuta mediante rettangoli adiacenti aventi per basi segmenti i cui estremi corrispondono agli estremi delle classi.
Fissate le basi, le altezze debbono essere tali che l'area di ogni rettangolo risultante sia uguale alla frequenza (relativa o assoluta) della classe stessa.
Se si utilizzano le frequenze assolute delle classi, l'area di ogni rettangolo è uguale alla frequenza assoluta della classe e l'area totale dei rettangoli è uguale all'ampiezza del campione.
Quindi, l'area del rettangolo i-esimo è uguale a $n_i=b_i \times h_i$, dove n_i è il numero di valori che cadono nella classe i-esima (frequenza assoluta della classe i-esima), b_i è la base e h_i è l'altezza della classe i-esima. Se si sceglie $b_i=1$ per ogni classe, le altezze hi delle classi dell'istogramma corrispondono alle frequenze assolute ni delle classi.
Riportiamo ora tali nozioni teoriche a livello pratico, applicandole al nostro set di dati.
Tramite R sono stati generati gli istogrammi per i dati relativi a: lavora avendo iniziato prima del conseguimento del titolo (LPT), lavora avendo iniziato dopo il conseguimento del titolo (LDT), cercano lavoro (CL), non cercano lavoro (NCL):



Ne consegue che:

- I valori più frequenti per "lavora avendo iniziato prima del conseguimento del titolo" variano dal 10% al 20%;
- I valori più frequenti "lavora avendo iniziato dopo il conseguimento del titolo" variano dal 60% al 70%;
- I valori più frequenti per "cercano lavoro" variano dal 15% al 20%;
- I valori più frequenti per "non cercano lavoro" variano dal 5% al 10%.

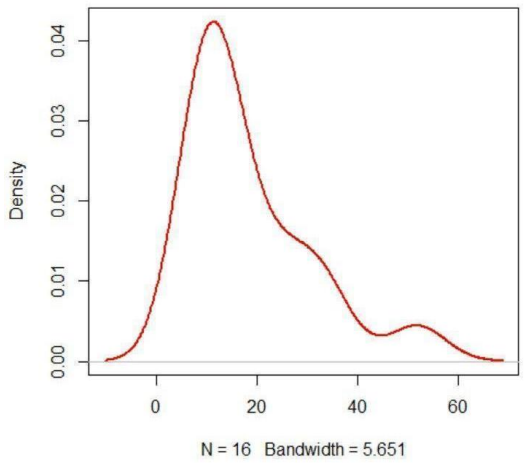
Da una prima analisi ne deriva che il tasso occupazionale è più alto per coloro che hanno conseguito il titolo di primo livello.

2.2 KERNEL DENSITY PLOT

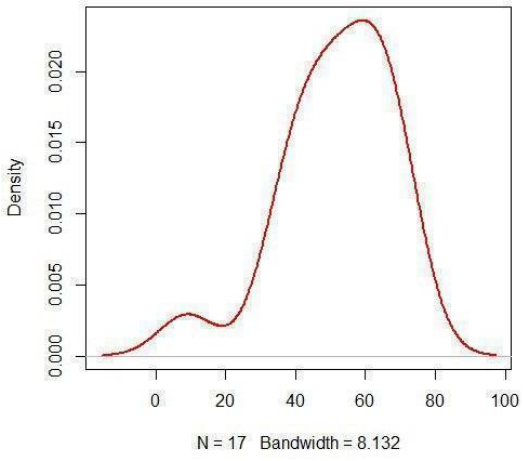
Un'altra metodologia per rappresentare una distribuzione di frequenza in classi per variabili quantitative univariate è utilizzare i Kernel density plot.
Questi ultimi utilizzano la stima di densità basata su Kernel. Con tale metodo, invece di raccogliere le osservazioni in barre, come negli istogrammi, 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 (un campione distorto fornisce, in generale, una stima falsata delle caratteristiche della popolazione oggetto dell'indagine statistica).
La scelta del kernel dipende dal campione, ma la scelta di default spesso si rivela essere quella da preferire.
R mette a disposizione vari tipi di kernel: "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" e "optcosine".
Di seguito sono riportati i grafici ottenuti mediante ogni kernel utilizzato.

2.2.1 GAUSSIAN

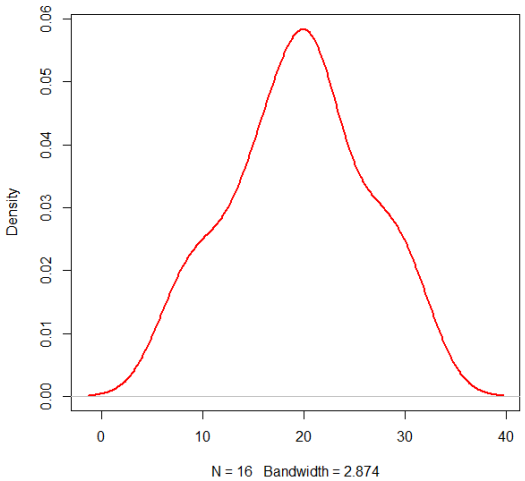
$$K(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{x^2}{2}\right\}, \quad x \in R \qquad \sigma^2 = 1$$



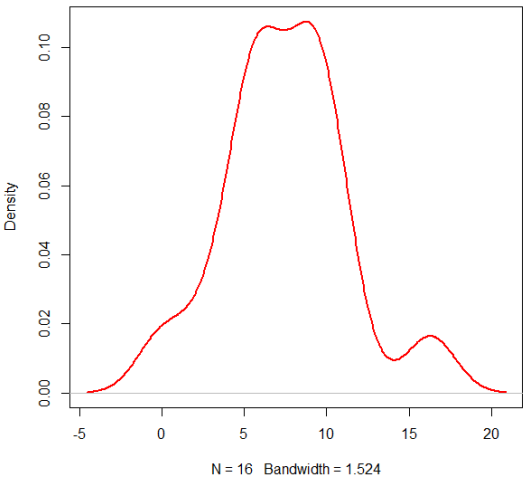
LPT



LDT



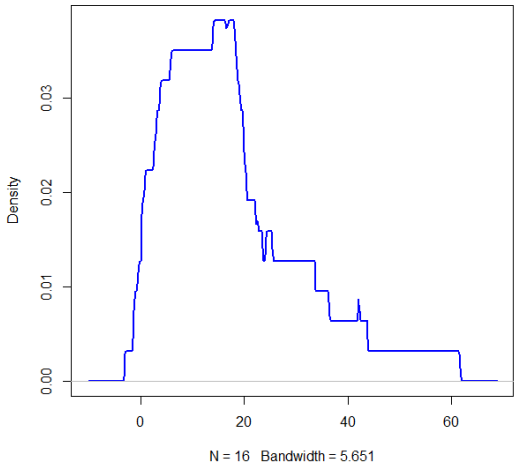
CL



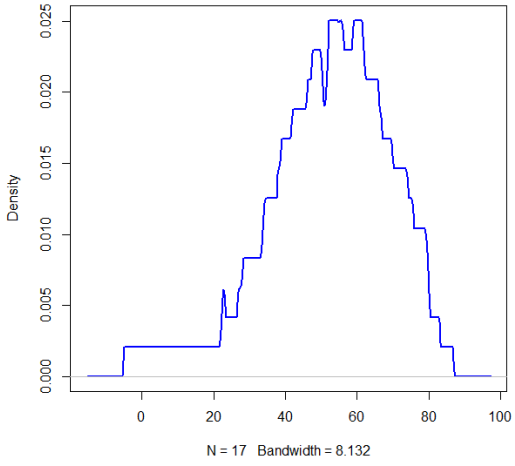
NC

2.2.2 RECTANGULAR

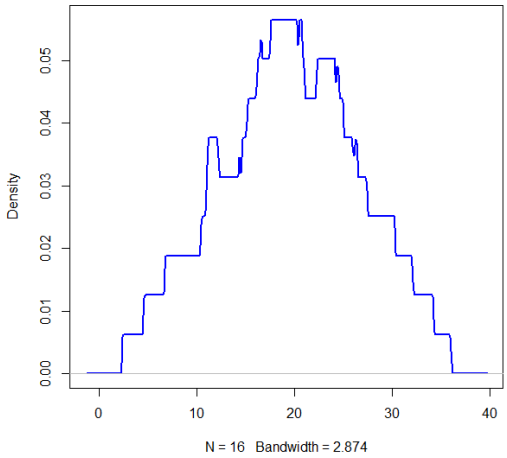
$$K(x)=\begin{cases}\frac{1}{2}, & -1 \le x \le 1,\\ 0, & altrimenti\end{cases} \qquad \sigma^2 = 1/3$$



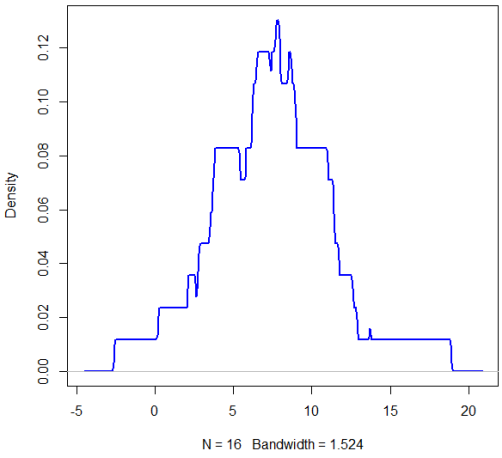
LPT



LDT



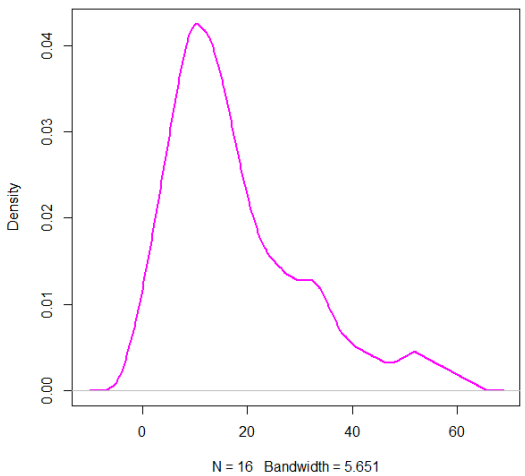
CL



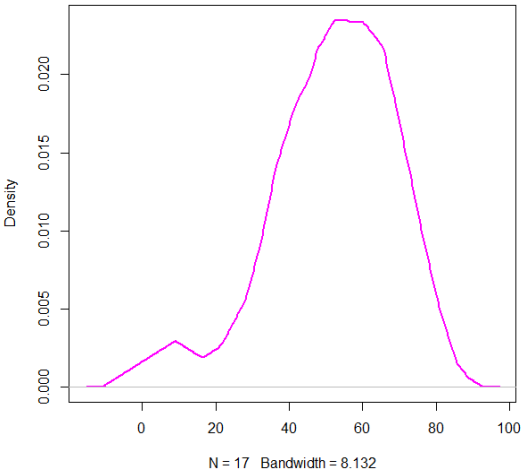
NCL

2.2.3 TRIANGULAR

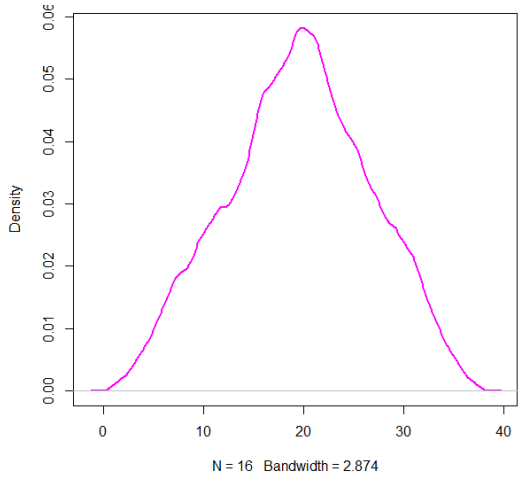
$$K(x)=\begin{cases}1-|x|, & -1 \le x \le 1,\\ 0, & altrimenti\end{cases} \qquad \sigma^2 = 1/6$$



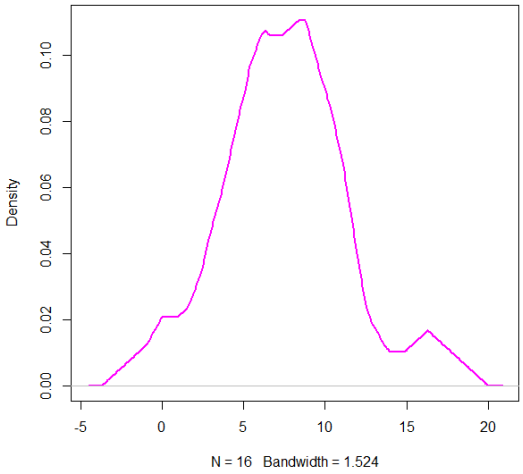
LPT



LDT



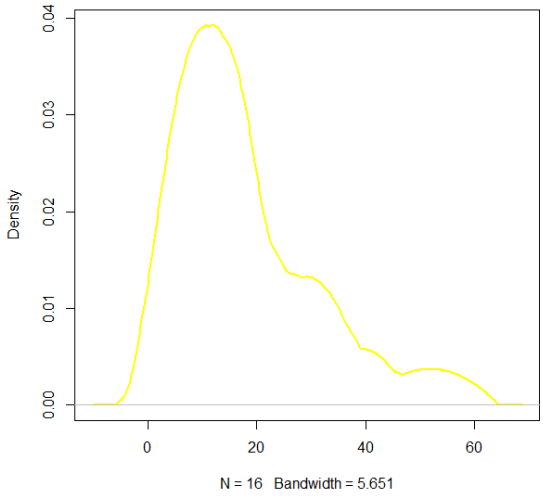
CL



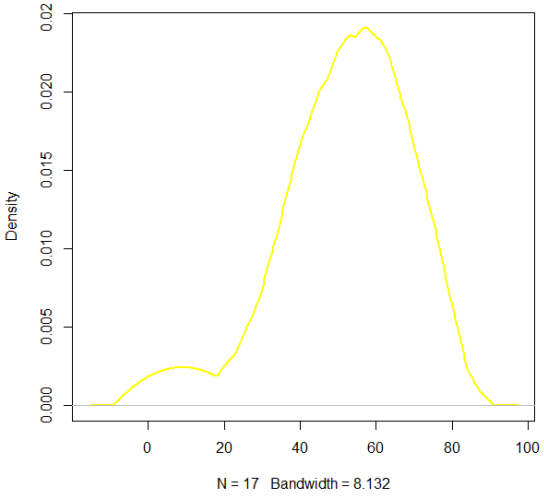
NCL

2.2.4 EPANECHNIKOV

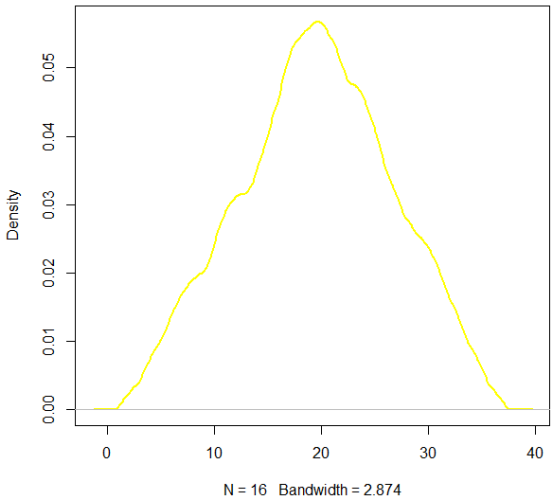
$$K(x)=\begin{cases}\frac{3}{4}(1-x^2), & -1 \le x \le 1, \\ 0, & altrimenti\end{cases} \qquad \sigma^2 = 1/5$$



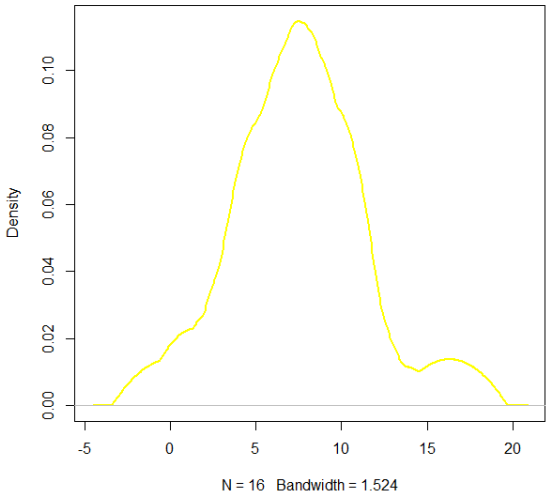
LPT



LDT



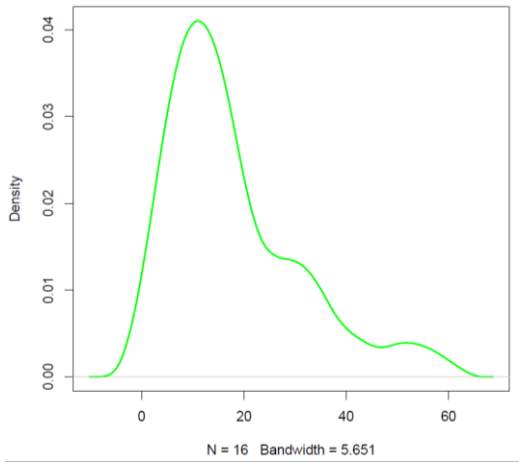
CL



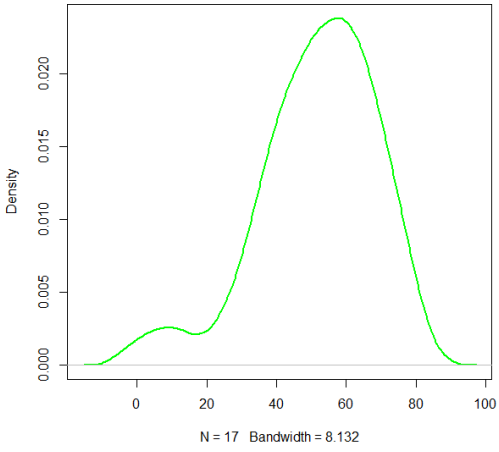
NCL

2.2.5 BIWEIGHT

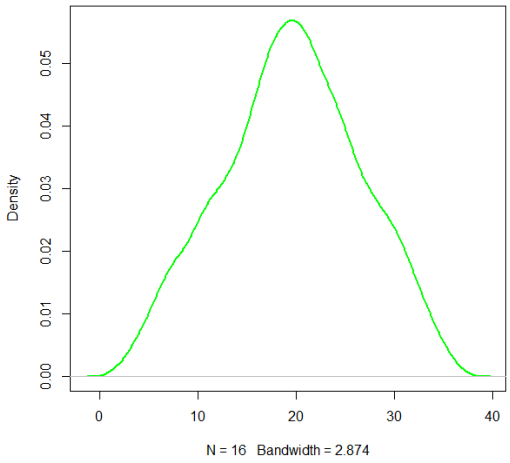
$$K(x)=\begin{cases}\frac{15}{16}(1-x^2)^2, & -1\le x\le 1,\\ 0, & altrimenti\end{cases} \qquad \sigma^2 = 1/7$$



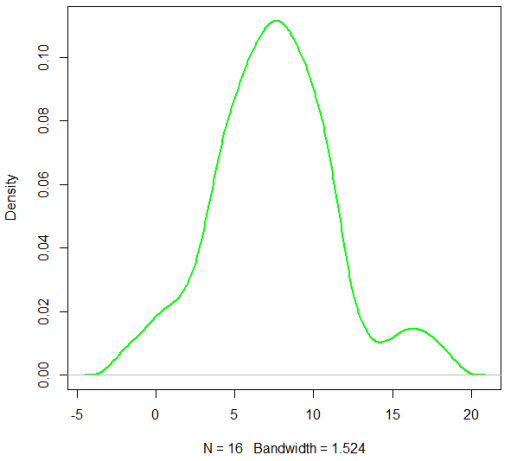
LPT



LDT



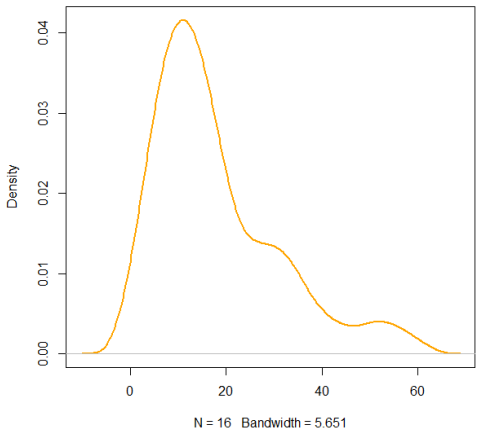
CL



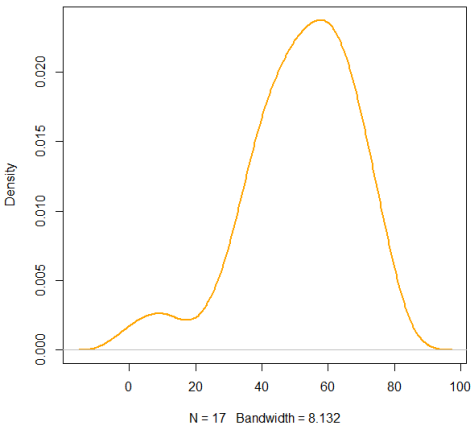
NCL

2.2.6 COSINE

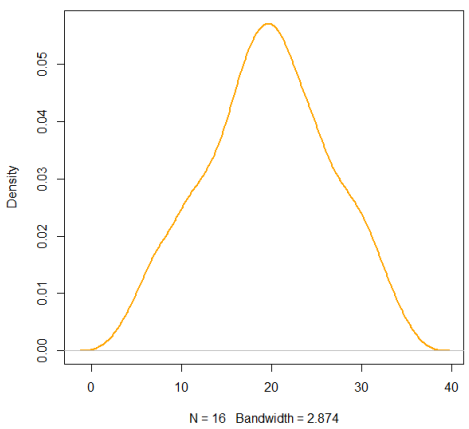
$$K(x)=\frac{\pi}{4}cos\left(\frac{\pi}{2}x\right),\ -1\le x\le 1,\ \ \sigma^2 = 1-\frac{8}{\pi^2}$$



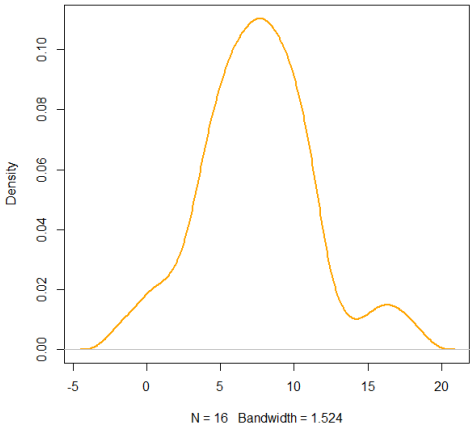
LPT



LDT



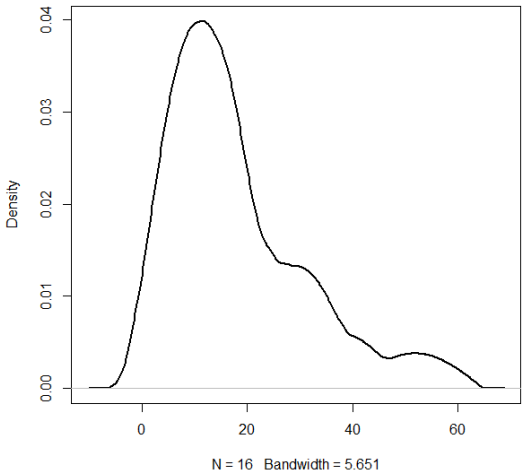
CL



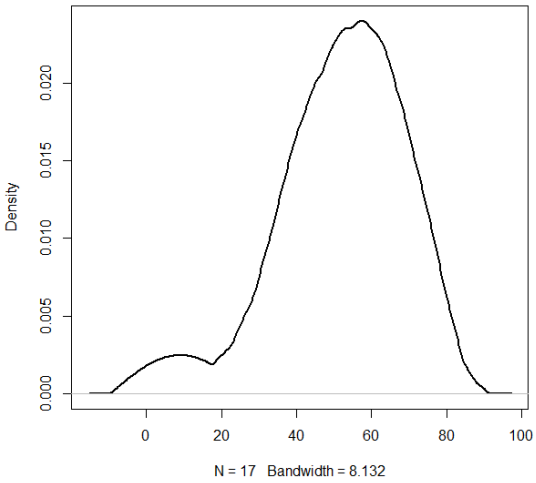
NCL

2.2.7 OPTCOSINE

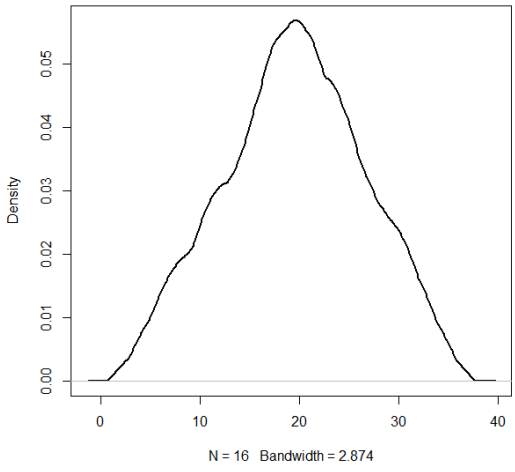
$$K(x)=\frac{1}{2}[1+cos(\pi x)],\ -1 \leq x \leq 1,\ \sigma^2 = 1/3 - \frac{2}{\pi^2}$$



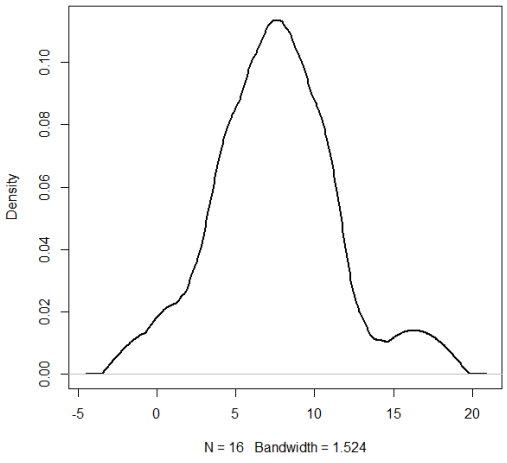
LPT



LDT



CL



NCL

Notiamo come tutte siano caratterizzate da valore medio nullo e varianza coincidente con il momento del secondo ordine.

2.3 BOXPLOT

Consideriamo un campione (x_1, x_2, …, x_n) dei valori assunti da una variabile quantitativa X.
Procediamo ad ordinare i valori del campione in ordine crescente.
Si chiama primo quartile, e si indica con Q_1, il valore per il quale il 25% dei dati sono alla sua sinistra e il restante 75% alla sua destra. Analogamente, si chiama terzo quartile, e si indica con Q_3, il valore per il quale il 75% dei dati sono alla sua sinistra e il restante 25% alla sua destra. Il secondo quartile Q_2, detto anche mediana, per il quale 50% dei dati sono alla sua sinistra e il restante 50% è alla sua destra.
Q_0 e Q_4 forniscono il minimo e il massimo dei valori del campione.
Relativamente al nostro caso analizziamo i seguenti quartili per le classi in esame:

```
> quantile(LPT)
    0%    25%    50%    75%   100%
 6.700  9.825 12.600 24.475 51.800
> quantile(LDT)
  0%  25%  50%  75% 100%
 9.0 42.0 52.8 65.3 72.8
> quantile(CL)
    0%    25%    50%    75%   100%
 7.300 15.775 19.850 23.225 31.100
> quantile(NCL)
    0%    25%    50%    75%   100%
 0.000  5.375  7.400  9.325 16.300
```

Inoltre, si ha la funzione summary(nomeVettore) che permette di determinare i valori precisi del minimo, del massimo, della media, della mediana, del primo e del terzo quantile.

```
> summary(LPTsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.700   9.825  13.200  18.400  24.475  51.800
> summary(LDTsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  36.10   46.20   54.30   54.56   65.42   72.80
> summary(CLsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.30   15.78   19.85   19.56   23.23   31.10
> summary(NCLsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   5.375   7.400   7.487   9.325  16.300
```

Il boxplot, detto anche scatola con baffi, è il disegno di una scatola i cui estremi sono Q_1 e Q_3, tagliata da una linea orizzontale in corrispondenza di Q_2, ossia della mediana.
In basso e in alto sono presenti altre due linee tratteggiate, dette baffi.
L'estremo del baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale di $Q_1 - 1.5 \cdot (Q_3 - Q_1)$, mentre l'estremo del baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a $Q_3 + 1.5 \cdot (Q_3 - Q_1)$.
La distanza tra il primo e il terzo quartile è detta intervallo interquartile o scarto interquartile. Se tutti i dati del campione rientrano nell'intervallo $(Q_1 - 1.5 \cdot (Q_3 - Q_1), Q_3 + 1.5 \cdot (Q_3 - Q_1))$ gli estremi dei baffi sono posti in corrispondenza del minimo valore e del massimo valore del campione.
Gli eventuali valori al di fuori dell'intervallo $(Q_1 - 1.5 \cdot (Q_3 - Q_1), Q_3 + 1.5 \cdot (Q_3 - Q_1))$ sono visualizzati nel grafico sotto forma di punti, detti valori anomali o outlier.
Questi valori infatti costituiscono una "anomalia" rispetto alla maggior parte dei valori osservati e pertanto è necessario identificarli per poterne analizzare le caratteristiche e le eventuali cause che li hanno determinati.
Nel nostro caso possiamo individuare:

- anomalie: relativamente alle classi LPT (51.8) e NCL (16.3).
  Per LPT:
  - L'estremo del baffo inferiore = valore minimo $\geq (Q_1 - 1.5 \cdot (Q_3 - Q_1)$
    $= 9{,}825 - 1.5 \cdot (24{,}475 - 9{,}825)$
    $= 9{,}825 - 21{,}975$
    $= -12{,}15$

    Il valore minimo che corrisponde a 6,700 risulta essere ≥ di -12,15;

Dunque, non è questo il valore che presenta anomalie.

- L'estremo del baffo superiore = valore massimo $\leq Q_3 + 1.5 \cdot (Q_3 - Q_1)$
  $= 24{,}475 + 1.5 \cdot (24{,}475 - 9{,}825)$
  $= 24{,}475 + 21{,}975$
  $= 46{,}45$

  Il valore massimo che corrisponde a 51,800 per il gruppo di laurea **DS** risulta essere > di 46,45.
  Relativamente al Boxplot questo valore risulta essere l'anomalia riscontrata.

Per NCL:

- L'estremo del baffo inferiore = valore minimo $\geq (Q_1 - 1.5 \cdot (Q_3 - Q_1)$
  $= 5{,}375 - 1.5 \cdot (9{,}325 - 5{,}375)$
  $= 5{,}375 - 5{,}925$
  $= -0{,}55$

  Il valore minimo che corrisponde a 0 risulta essere $\geq$ di $-0{,}55$;
  Dunque, non è questo il valore che presenta anomalie.

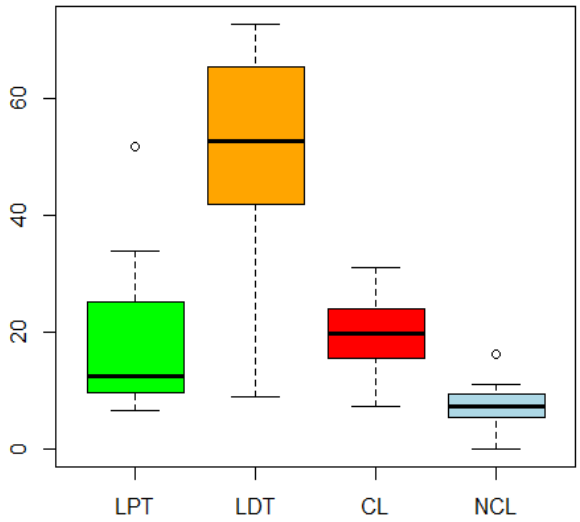
- L'estremo del baffo superiore = valore massimo $\leq Q_3 + 1.5 \cdot (Q_3 - Q_1)$
  $= 9{,}325 + 1.5 \cdot (9{,}325 - 5{,}375)$
  $= 9{,}325 + 5{,}925$
  $= 15{,}25$

  Il valore massimo che corrisponde a 16,300 per il gruppo di laurea **PSI** risulta essere > di 15,25.
  Relativamente al Boxplot questo valore risulta essere l'anomalia riscontrata;

- centralità: espressa dalla mediana;

- simmetria dei dati analizzati: la si può dedurre esaminando le distanze del primo e del terzo quartile dalla linea mediana. Come si può osservare in figura, la classe LPT risulta fortemente asimmetrica mentre LDT, CL e NCL risultano essere classi con maggiore simmetria;

- dispersione dei dati: questa misura è deducibile osservando i baffi. La dispersione è deducibile osservando le distanze dell'estremo del baffo superiore da Q3 e dall'estremo del baffo inferiore da Q1. Si concentra nelle classi LPT, LDT, NCL.



CAPITOLO 3: STATISTICA DESCRITTIVA UNIVARIATA

Per fenomeni quantitativi, come nel nostro caso, è utile definire la funzione di distribuzione empirica. Nel paragrafo che segue riportiamo una comparazione tra caso discreto e caso continuo.

3.1 FUNZIONE DI DISTRIBUZIONE EMPIRICA DISCRETA E CONTINUA

Questa funzione, nel caso discreto, è definita a partire dalle frequenze relative cumulative.
Consideriamo una variabile quantitativa X e indichiamo con $z_1, z_2, \ldots, z_k$ i valori distinti da essa assunti e assumiamo che essi siano ordinati in ordine crescente, ossia $z_1 < z_2 < \ldots < z_k$.
Consideriamo poi un campione $(x_1, x_2, \ldots, x_n)$ costituito da n osservazioni di X.
Denotiamo con n_i il numero di volte in cui ciascun valore z_i è presente nel campione, ossia la frequenza assoluta con cui esso appare nel campione, e con $f_i=n_i/n$ e frequenze relative.
Le frequenze relative cumulative sono così definite:

$$F_i=f_1+f_2\ldots+f_i=\frac{n_1+n_2+\cdots n_i}{n} \quad (i=1,2\ldots,k),$$

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:

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

dove # indica la cardinalità dell'insieme.
La funzione di distribuzione empirica F(x) è definita per ogni x reale ed è 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.
La funzione di distribuzione empirica F(x) gode delle seguenti proprietà:

- è una funzione non decrescente;
- 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 dell'osservazione massima.

Per fenomeni quantitativi continui, invece, 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:

$$C_1=[z_0, z_1),\ C_2=[z_1, z_2),\ \ldots,\ C_k=[z_{k-1}, z_k],\ \text{con } z_0 < z_1 < \ldots < z_{k-1} < z_k$$

Dove z_0 corrisponde al minimo delle osservazioni e z_k al massimo delle osservazioni.
La funzione di distribuzione empirica continua è così definita:

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

Dove $F_0=0$ e F_i denota la frequenza relativa cumulativa della classe C_i (i=1,2, …, k).
Si nota che $F(x)=0$ per $x<z_0$, $F(x)=1$ per $x\geq z_k$, 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

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

da cui segue

$$y = F_{i-1}+\frac{F_i-F_{i-1}}{z_i-z_{i-1}}(x-z_{i-1})=\frac{F_i-F_{i-1}}{z_i-z_{i-1}}x+\frac{F_{i-1}(z_i-z_{i-1})-z_{i-1}(F_i-F_{i-1})}{z_i-z_{i-1}}=\frac{F_i-F_{i-1}}{z_i-z_{i-1}}x+\frac{z_iF_{i-1}-z_{i-1}F_i}{z_i-z_{i-1}}.$$

Effettuiamo, a questo punto, un'analisi relativa al caso da noi trattato proponendo una comparazione tra le due tipologie di funzioni di distribuzioni empiriche:

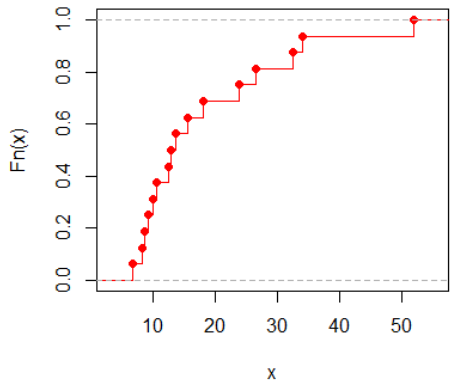
LPT

Funzione di distribuzione empirica discreta

Frequenze relative cumulative:

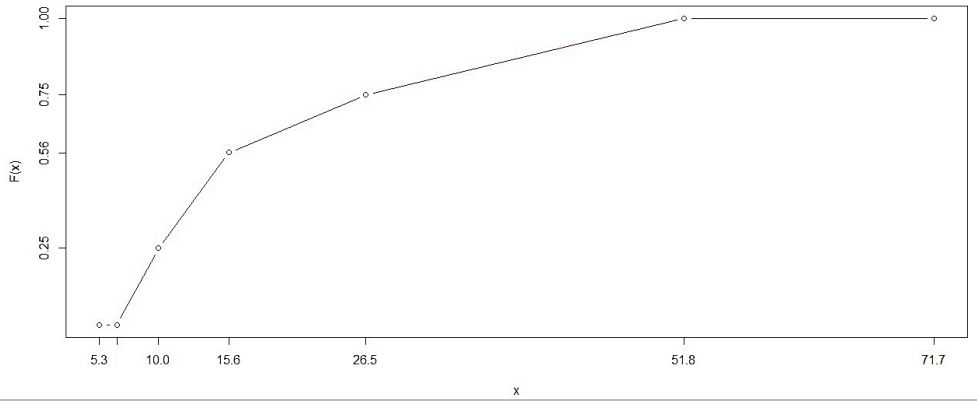
```
> LPTfcumrel<-round(cumsum(table(LPTsorted)/length(LPTsorted)),3)
> LPTfcumrel
  6.7   8.3   8.6   9.3    10  10.6  12.4  12.8  13.6  15.6    18  23.8
0.062 0.125 0.188 0.250 0.312 0.375 0.438 0.500 0.562 0.625 0.688 0.750
 26.5  32.4    34  51.8
0.812 0.875 0.938 1.000
```

Grafico della funzione di distribuzione empirica discreta:



Funzione di distribuzione empirica continua

```
> LPTfcum
   [6.7,10)   [10,15.6) [15.6,26.5) [26.5,51.8)
     0.2500      0.5625      0.7500      1.0000
> LPTfreqrel
LPT
   6.7    8.3    8.6    9.3     10   10.6   12.4   12.8   13.6   15.6     18
0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625
  23.8   26.5   32.4     34   51.8
0.0625 0.0625 0.0625 0.0625 0.0625
> LPTclassi
[1]  6.7 10.0 15.6 26.5 51.8
> LPTfreqrelclassi

   [6.7,10)   [10,15.6) [15.6,26.5) [26.5,51.8)
     0.2500      0.3125      0.1875      0.1875
> LPTfcum
   [6.7,10)   [10,15.6) [15.6,26.5) [26.5,51.8)
     0.2500      0.5625      0.7500      1.0000
> LPTascisse
[1]  5.3  6.7 10.0 15.6 26.5 51.8 71.7
> LPTordinate
                        [6.7,10)   [10,15.6) [15.6,26.5) [26.5,51.8)
     0.0000      0.0000      0.2500      0.5625      0.7500      1.0000

     1.0000
> plot(LPTascisse, LPTordinate, type="b", main="Funzione di distribuzione empirica continua", col="red", ylim=c(0,1), xlab="x", ylab="F(x)", axes=FALSE)
> axis(1,LPTascisse)
> axis(2, format(LPTfcum, digits=2))
> box()
```



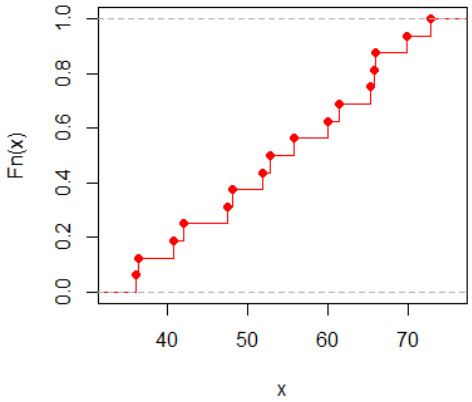
LDT

Funzione di distribuzione empirica discreta

Frequenze relative cumulative:

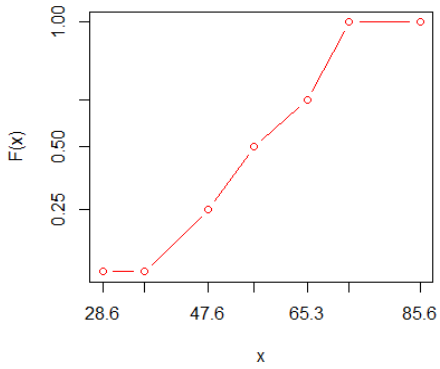
```
> LDTfcumrel<-round(cumsum(table(LDTsorted)/length(LDTsorted)),3)
> LDTfcumrel
 36.1  36.5  40.8    42  47.6  48.1  51.9  52.8  55.8  60.1  61.5  65.3
0.062 0.125 0.188 0.250 0.312 0.375 0.438 0.500 0.562 0.625 0.688 0.750
 65.8    66  69.9  72.8
0.812 0.875 0.938 1.000
```

Grafico della funzione di distribuzione empirica discreta:



Funzione di distribuzione empirica continua

```
> LDT
 [1] 69.9 65.3 51.9 72.8 65.8 55.8 60.1 66.0 47.6 36.1 48.1 61.5 42.0 36.5
[15] 52.8 40.8
> LDTsorted
 [1] 36.1 36.5 40.8 42.0 47.6 48.1 51.9 52.8 55.8 60.1 61.5 65.3 65.8 66.0
[15] 69.9 72.8
> LDTfreqrel
Errore: oggetto 'LDTfreqrel' non trovato
> LDTfreqrel <- table(LDT) / length(LDT)
> LDTfreqrel
LDT
  36.1   36.5   40.8     42   47.6   48.1   51.9   52.8   55.8   60.1   61.5
0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625
  65.3   65.8     66   69.9   72.8
0.0625 0.0625 0.0625 0.0625 0.0625
> LDTclassi <- c(36.1, 47.6, 55.8, 65.3, 72.8)
> LDTclassi
[1] 36.1 47.6 55.8 65.3 72.8
> LDTfreqrelclassi <- table(cut(LDTsorted, breaks=LDTclassi, right=FALSE))/length(LDTsorted)
> LDTfreqrelclassi

[36.1,47.6) [47.6,55.8) [55.8,65.3) [65.3,72.8)
     0.2500      0.2500      0.1875      0.2500
> LDTfcum <- cumsum(LDTfreqrelclassi)
> LDTfcum
[36.1,47.6) [47.6,55.8) [55.8,65.3) [65.3,72.8)
     0.2500      0.5000      0.6875      0.9375
> LDTfcum[4] <- LDTfcum[4]+LDTfreqrel[m]
> LDTfcum
[36.1,47.6) [47.6,55.8) [55.8,65.3) [65.3,72.8)
     0.2500      0.5000      0.6875      1.0000
> LDTascisse <- c(28.6, 36.1, 47.6, 55.8, 65.3, 72.8, 85.6)
> LDTordinate <- c(0,0, LDTfcum[1:4], 1)
> plot(LDTascisse, LDTordinate, type="b", main="Funzione di distribuzione empirica continua", col="red", ylim=c(0,1), xlab="x", ylab="F(x)", axes=FALSE)
> axis(1,LDTascisse)
> axis(2, format(LDTfcum, digits=2))
> box()
```



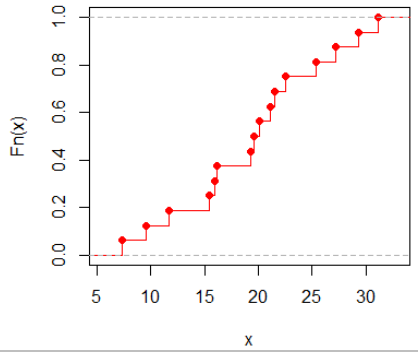
CL

Funzione di distribuzione empirica discreta

Frequenze relative cumulative:

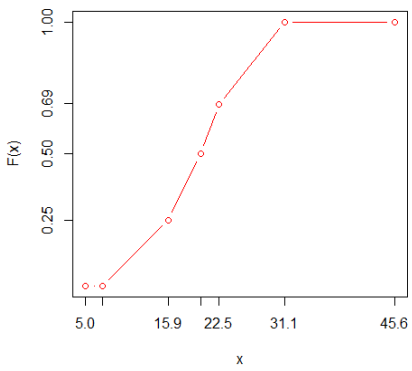
```
> CLfcumrel<-round(cumsum(table(CLsorted)/length(CLsorted)),3)
> CLfcumrel
  7.3   9.5  11.7  15.4  15.9  16.1  19.3  19.6  20.1  21.1  21.5  22.5
0.062 0.125 0.188 0.250 0.312 0.375 0.438 0.500 0.562 0.625 0.688 0.750
 25.4  27.2  29.3  31.1
0.812 0.875 0.938 1.000
```

Grafico della funzione di distribuzione empirica discreta:



Funzione di distribuzione empirica continua

```
> CL
 [1]  9.5 20.1 31.1 11.7 15.9 25.4 19.6 16.1 22.5 21.5 27.2 21.1 19.3 29.3
[15] 15.4  7.3
> CLsorted
 [1]  7.3  9.5 11.7 15.4 15.9 16.1 19.3 19.6 20.1 21.1 21.5 22.5 25.4 27.2
[15] 29.3 31.1
> CLfreqrel <- table(CL)/length(CL)
> CLfreqrel <- table(CL)/length(CL)
> CLfreqrel
CL
   7.3    9.5   11.7   15.4   15.9   16.1   19.3   19.6   20.1   21.1   21.5
0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625
  22.5   25.4   27.2   29.3   31.1
0.0625 0.0625 0.0625 0.0625 0.0625
> CLclassi <- c(7.3, 15.9, 20.1, 22.5, 31.1)
> CLclassi
[1]  7.3 15.9 20.1 22.5 31.1
> CLfreqrelclassi <- table(cut(CLsorted, breaks=CLclassi, right=FALSE))/length(CLsorted)
> CLfreqrelclassi

 [7.3,15.9) [15.9,20.1) [20.1,22.5) [22.5,31.1)
     0.2500      0.2500      0.1875      0.2500
> CLfcum <- cumsum(CLfreqrelclassi)
> CLfcum
 [7.3,15.9) [15.9,20.1) [20.1,22.5) [22.5,31.1)
     0.2500      0.5000      0.6875      0.9375
> CLfcum[4]<-CLfcum[4]+CLfreqrel[m]
> CLfcum
 [7.3,15.9) [15.9,20.1) [20.1,22.5) [22.5,31.1)
     0.2500      0.5000      0.6875      1.0000
> CLascisse<-c(5.0, 7.3, 15.9, 20.1, 22.5, 31.1, 45.6)
> CLordinate<-c(0,0,CLfcum[1:4],1)
> plot(CLascisse, CLordinate, type="b", main="Funzione di distribuzione empirica continua", col="red", ylim=c(0,1), xlab="x", ylab="F(x)", axes=FALSE)
> axis(1,CLascisse)
> axis(2, format(CLfcum, digits=2))
> box()
```



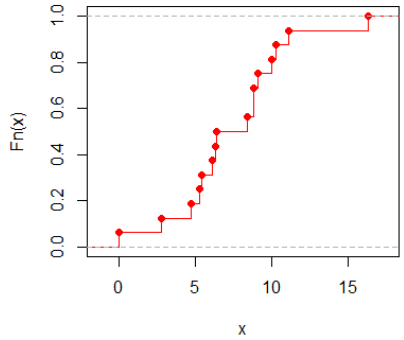
NCL

Funzione di distribuzione empirica discreta

Frequenze relative cumulative:

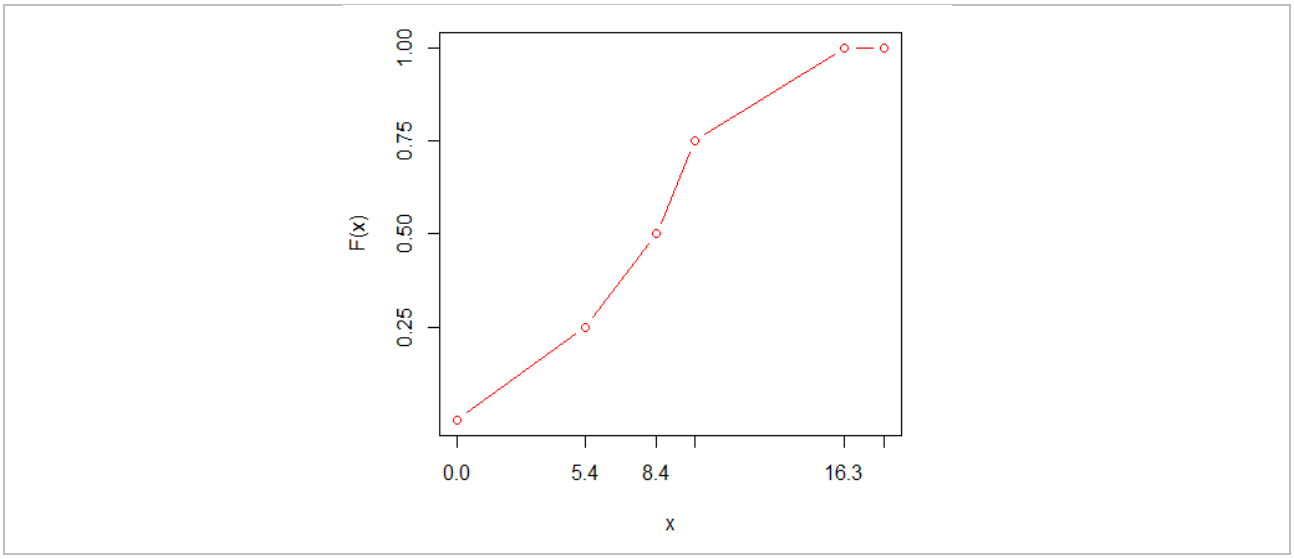
```
> NCLfcumrel<-round(cumsum(table(NCLsorted)/length(NCLsorted)),3)
> NCLfcumrel
    0   2.8   4.7   5.3   5.4   6.1   6.3   6.4   8.4   8.8   9.1    10
0.062 0.125 0.188 0.250 0.312 0.375 0.438 0.500 0.562 0.688 0.750 0.812
 10.3  11.1  16.3
0.875 0.938 1.000
```

Grafico della funzione di distribuzione empirica discreta:



Funzione di distribuzione empirica continua

```
> NCL
 [1] 10.0  6.3 10.3  2.8  9.1  8.8  4.7  5.4  6.1  8.4 11.1  8.8  6.4 16.3  5.3  0.0
> NCLsorted
 [1]  0.0  2.8  4.7  5.3  5.4  6.1  6.3  6.4  8.4  8.8  8.8  9.1 10.0 10.3 11.1 16.3
> NCLfreqrel<-table(NCL)/length(NCL)
> NCLfreqrel
NCL
     0    2.8    4.7    5.3    5.4    6.1    6.3    6.4    8.4    8.8    9.1     10   10.3
0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.0625 0.1250 0.0625 0.0625 0.0625
  11.1   16.3
0.0625 0.0625
> NCLclassi<-c(0.0, 5.4, 8.4, 10, 16.3)
> NCLclassi
[1]  0.0  5.4  8.4 10.0 16.3
> NCLfreqrelclassi <- table(cut(NCLsorted, breaks=NCLclassi, right=FALSE))/length(NCLsorted)
> NCLfreqrelclassi

  [0,5.4) [5.4,8.4)  [8.4,10) [10,16.3)
   0.2500    0.2500    0.2500    0.1875
> NCLfcum <- cumsum(NCLfreqrelclassi)
> NCLfcum
  [0,5.4) [5.4,8.4)  [8.4,10) [10,16.3)
   0.2500    0.5000    0.7500    0.9375
> NCLascisse<-c(0.0,0.0, 5.4, 8.4, 10.0, 16.3,18.0)
> NCLordinate<-c(0, 0, NCLfcum[1:4],1)
> plot(NCLascisse, NCLordinate, type="b", col="red", ylim=c(0,1), xlab="x", ylab="F(x)", axes=FALSE)
> axis(1,NCLascisse)
> axis(2, format(NCLfcum, digits=2))
```

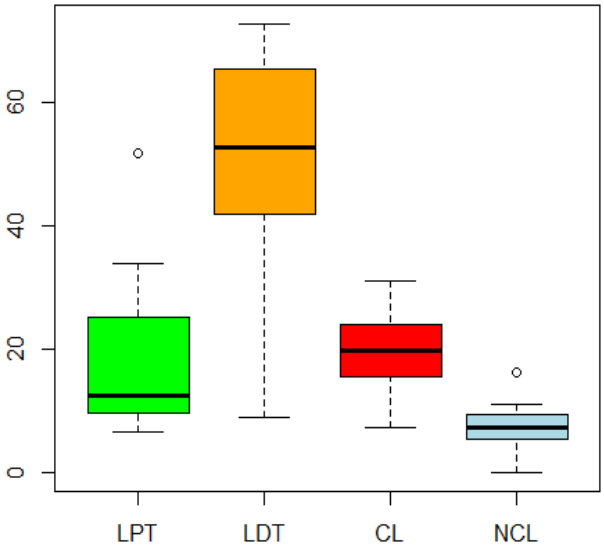


3.2 INDICI DI SINTESI

Gli indici di sintesi, detti anche statistiche, sono utili a descrivere dei dati numerici.
Sono:

- media e mediana
- moda
- varianza
- deviazione standard
- coefficiente di variazione

La media, la mediana e la moda sono misure di centralità, mentre la varianza e la deviazione standard misurano la dispersione dei dati.
Il boxplot presentato in figura ci ha consentito di poter effettuare delle osservazioni qualitative, analizzate nel capitolo precedente;




```
> summary(LPTsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.700   9.825  13.200  18.400  24.475  51.800
> summary(LDTsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  36.10   46.20   54.30   54.56   65.42   72.80
> summary(CLsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.30   15.78   19.85   19.56   23.23   31.10
> summary(NCLsorted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   5.375   7.400   7.487   9.325  16.300
```

gli indici di sintesi, invece, ci consentono di misurare quantitativamente alcune delle caratteristiche osservate qualitativamente nei grafici delle distribuzioni di frequenze e nei boxplot.

Poiché i dati oggetti dello studio sono di tipo quantitativo, non abbiamo incluso la moda campionaria utilizzata quando si trattano dati di tipo qualitativo.

3.2.1 MEDIA E MEDIANA

Supponiamo di avere un insieme $x_1, x_2, x_3, \ldots, x_n$ di n valori numerici (dati statistici quantitativi), detto campione di ampiezza o numerosità pari a n. La **media campionaria** è la media aritmetica di questi valori.

Si definisce media campionaria e si denota con $\bar{x}$, la quantità:

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

Si nota che se si considera un nuovo insieme di dati $y_i = ax_i+b$ (i=1, 2, …, n) con a, b ∈ R, allora la media campionaria di $y_1, y_2, y_3, \ldots, y_n$ è legata alla media campionaria dei dati iniziali $x_1, x_2, x_3, \ldots, x_n$ dalla stessa relazione lineare, ossia

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i+b)=a\bar{x}+b$$

Se si denotano con $z_1, z_2, z_3, \ldots, z_k$ i valori distinti assunti dai dati, con $n_1, n_2, n_3, \ldots, n_k$ le frequenze assolute e con $f_1, f_2, f_3, \ldots, f_k$ le frequenze relative, allora $\bar{x}$ può essere così riscritta:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i =\frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k}\frac{n_i}{n} z_i = \sum_{i=1}^{k} f_i z_i$$

che mostra che la media campionaria è una media pesata dei valori distinti assunti dai dati.

Ogni valore distinto usa come peso la sua frequenza relativa, ovvero la frazione dei dati uguale a tale valore numerico.

Per ogni valore x_i si definisce lo scarto dalla media campionaria la quantità $s_i=x_i-\bar{x}$ (i=1,2…, n) che indica il grado di scostamento del singolo valore xi dalla media campionaria $\bar{x}$.

Si nota immediatamente che la somma algebrica degli scarti dalla media campionaria è sempre nulla.

Infatti, risulta:

$$\sum_{i=1}^{n} s_i = \sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}(x_i-n\bar{x}) =n(\bar{x}-\bar{x})=0$$

Una seconda statistica che indica la centralità di un insieme di dati è la **mediana campionaria**.

Assegnato un insieme di dati di ampiezza n, lo si ordini in ordine crescente (dal valore più piccolo al valore più grande).

Se n è dispari, si definisce mediana campionaria il valore che è in posizione (n+1)/2, mentre se n è pari la mediana campionaria è invece definita come la media aritmetica dei valori che occupano le posizioni n/2 e n/2+1.

Questa definizione della mediana campionaria bipartisce le osservazioni (dopo aver effettuato l'ordinamento) in due gruppi di uguale numerosità, in maniera tale che lo stesso numero di valori cada sia a sinistra che a destra della mediana stessa.

Media campionaria e **mediana campionaria** sono entrambe statistiche utili per descrivere misure di centralità dei dati. La media campionaria utilizza tutti i dati ed è influenzata in maniera sensibile da valori eccezionalmente alti o bassi. La mediana campionaria invece 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.

```
> mean(LPTsorted)
[1] 18.4
> median(LPTsorted)
[1] 13.2
> mean(LDTsorted)
[1] 54.5625
> median(LDTsorted)
[1] 54.3
> mean(CLsorted)
[1] 19.5625
> median(CLsorted)
[1] 19.85
> mean(NCLsorted)
[1] 7.4875
> median(NCLsorted)
[1] 7.4
```

Se le due misure risultano essere uguali, allora la distribuzione di frequenze tende ad essere simmetrica; se invece la media campionaria è sensibilmente maggiore della mediana campionaria, allora la distribuzione di frequenze è più sbilanciata verso destra. Al contrario, risulterebbe sbilanciata verso sinistra.
Nel nostro caso ciò si verifica per il vettore LPT dove la media risulta essere maggiore rispetto alla mediana e questo è ben visibile anche nel boxplot visto precedentemente.

3.2.2 QUANTILI
Oltre la mediana, che è quel valore che divide a metà un insieme di dati ordinati, si possono definire altri indici di posizione, detti quantili, che dividono l'insieme dei dati ordinati in un fissato numero di parti uguali.
Sia X una variabile quantitativa e sia $x_1, x_2, x_3, \ldots, x_n$ un campione di n osservazioni disposte in ordine crescente.
Supponiamo di suddividere i dati ordinati in α gruppi, ognuno dei quali contenga (circa) lo stesso numero di osservazioni; gli $\alpha-1$ numeri che consentono tale suddivisione sono i quantili di ordine α.
Il sistema R mette a disposizione diversi algoritmi per il calcolo dei quantili. Per definizione questi dividono l'insieme in un determinato numero di parti uguali.
Per effettuarne il calcolo si è scelto di utilizzare l'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

$$h^* = \text{ceiling}(h).$$

Il percentile k-esimo è quello che corrisponde alla posizione h*, ossia $P_k = v[h^*]$.
Senza entrare troppo nel dettaglio di questo algoritmo, la particolarità consiste nel fatto che questo **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)

```
> #QUANTILI
> quantile(LPTsorted, c(0, 0.25, 0.5, 0.75, 1), type=2)
   0%   25%   50%   75%  100%
 6.70  9.65 13.20 25.15 51.80
> quantile(LDTsorted, c(0, 0.25, 0.5, 0.75, 1), type=2)
   0%   25%   50%   75%  100%
36.10 44.80 54.30 65.55 72.80
> quantile(CLsorted, c(0, 0.25, 0.5, 0.75, 1), type=2)
   0%   25%   50%   75%  100%
 7.30 15.65 19.85 23.95 31.10
> quantile(NCLsorted, c(0, 0.25, 0.5, 0.75, 1), type=2)
   0%   25%   50%   75%  100%
 0.00  5.35  7.40  9.55 16.30
```

Un modo di procedere diverso per definire i quantili consiste nel considerare le frequenze relative cumulative.
Sia X una variabile quantitativa e siano $z_1, z_2, \ldots, z_k$ le modalità distinte da essa assunte, con $z_1<z_2<\ldots<z_k$. Considerato un campione $(x_1, x_2, \ldots, x_n)$, siano $F_1, F_2, \ldots, F_k$ le frequenze relative cumulative. Assegnata una probabilità p, $0<p<1$, il quantile di ordine p, è definito come la modalità i-esima (i=1,2, ..., k) che soddisfa la doppia disuguaglianza:

$$F_{i-1}<p, F_i\geq p$$

Si nota che i quantili di una distribuzione di frequenze forniscono sempre valori realmente osservati nel campione.
Il caso che si presenta più di frequente è quello dei quartili, che suddividono la distribuzione in quattro parti. Il primo quartile di una distribuzione di frequenze è definito come la modalità i-esima (i=1,2, …, k) che soddisfa la doppia disuguaglianza:

$$F_{i-1}<0.25, F_i\geq 0.25$$

Il secondo quartile (mediana) di una distribuzione di frequenze è definita come la modalità i-esima (i=1,2, …, k) che soddisfa la doppia disuguaglianza:

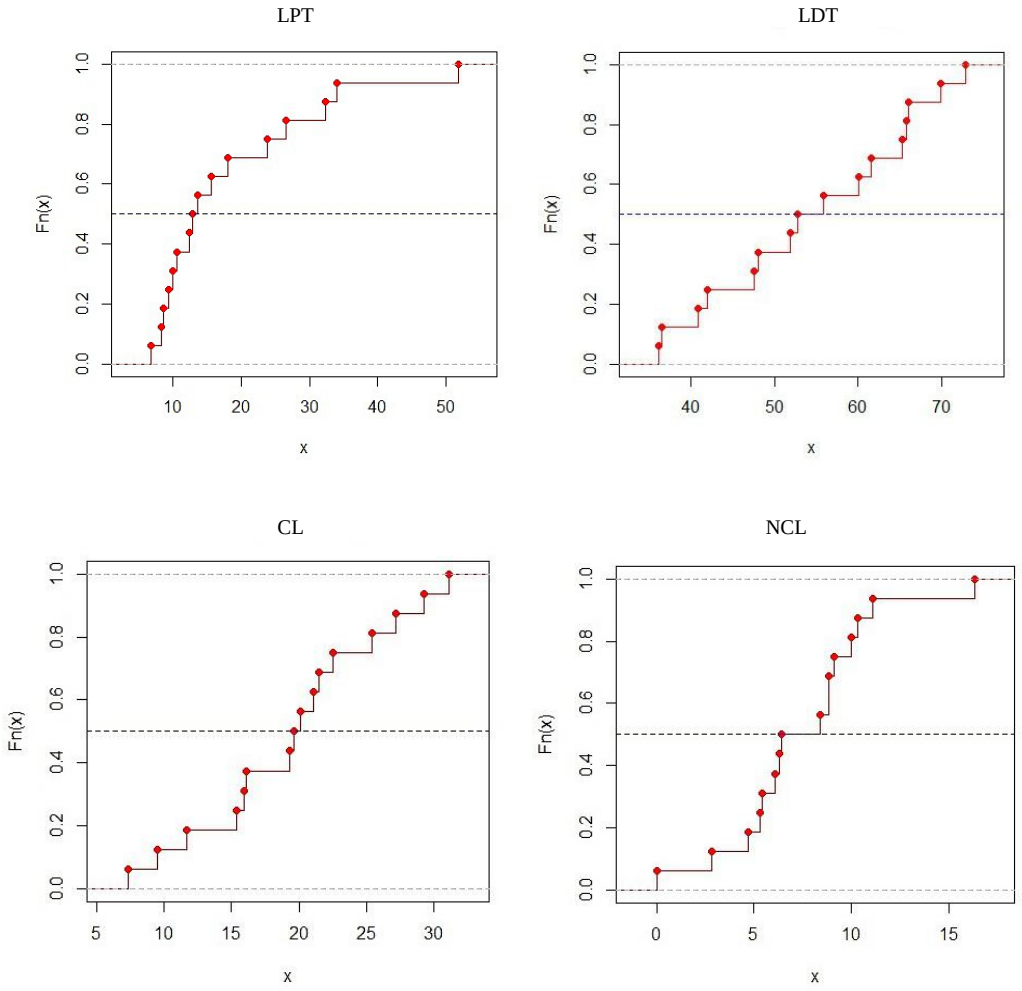
$$F_{i-1}<0.5, F_i\geq 0.5$$

Il terzo quartile di una distribuzione di frequenze è definito come la modalità i-esima (i = 1, 2, …, k) che soddisfa la doppia disuguaglianza:

$$F_{i-1}<0.75, F_i\geq 0.75$$

L'algoritmo è implementato in R scegliendo j = 1 e usando la funzione quantile (v, prob =, type = 1).
I quartili per una distribuzione di frequenze possono essere visualizzati graficamente a partire dalla funzione di distribuzione empirica discreta.
Si traccia la funzione di distribuzione empirica e sull'asse delle ordinate si individuano i punti 0.25, 0.5, 0.75 e da questi si tracciano delle linee orizzontali:

- il minimo valore osservato la cui funzione di distribuzione empirica supera 0.25 è il primo quartile Q1 per una distribuzione di frequenze;
- il minimo valore osservato la cui funzione di distribuzione empirica supera 0.5 è il secondo quartile Q2 (mediana) per una distribuzione di frequenze;
- il minimo valore osservato la cui funzione di distribuzione empirica supera 0.75 è il terzo quartile Q3 per una distribuzione di frequenze.



I grafici proposti evidenziano i valori delle mediane per le distribuzioni di frequenza delle quattro condizioni occupazionali considerate.

3.2.3 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 frequenza sono la varianza campionaria e la deviazione standard campionaria, detta anche scarto quadratico medio campionario.

Assegnato un insieme di dati numerici $x_1, x_2, \ldots, x_n$, si definisce **varianza campionaria**, e si denota con s^2, la quantità:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad (n=2,3,\ldots)$$

dove $\bar{x}$ denota la media campionaria dei dati. Inoltre, si definisce **deviazione standard campionaria** la radice quadrata della varianza campionaria, ossia:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (n=2,3,\ldots)$$

Varianza campionaria e **deviazione standard campionaria** sono detti indici di dispersione o indici di variabilità poiché misurano la dispersione dei dati intorno alla media.

Dalle due formule precedenti si nota che la varianza e la deviazione standard sono tanto più grandi quando più i dati si discostano dalla media.

I valori della varianza campionaria s^2 e della deviazione standard campionaria s dipendono dall'unità di misura dei dati.

In particolare, a differenza della varianza, la deviazione standard campionaria s misura la dispersione dei dati con la stessa unità di misura dei dati sperimentali e quindi con la stessa unità di misura della media campionaria.
Per un insieme di dati numerici $x_1, x_2, \ldots, x_n$ la varianza campionaria può anche essere calcolata nel seguente modo:

$$s^2=\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2-n\bar{x}^2\right]=\frac{n}{n-1}\left(M_2-\bar{x}^2\right)$$

Dove M_2 è il momento campionario del secondo ordine.
Infatti, la formula precedente può essere così dimostrata:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)=\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2-2\bar{x}\sum_{i=1}^{n} x_i+\sum_{i=1}^{n}\bar{x}^2\right)\Big]=$$
$$\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2-2n\bar{x}^2+n\bar{x}^2\right]=\frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2-\bar{x}^2\right]=\frac{n}{n-1}\left(M_2-\bar{x}^2\right).$$

Si nota inoltre che se si considera un nuovo insieme di dati $y_i=ax_i+b$ $(i=1,2,\ldots,n)$ con $a, b \in R$, allora la varianza campionaria $s^y{}_2$ di y_1,y_2, è legata alla varianza campionaria $s^x{}_2$ dei dati iniziali $x_1, x_2, \ldots, x_n$ dalla relazione:

$$s^y{}_2 = a^2\, s^x{}_2$$

Infatti, ricordando la proprietà di linearità della media campionaria, dalla seguente

$$\bar{y}=\frac{1}{n}\sum_{1=1}^{n} y_i=\frac{1}{n}\sum_{1=1}^{n}(ax_i+b)=a\bar{x}+b$$

segue che

$$s^y{}_2=\frac{1}{n-1}\sum_{1=1}^{n}(y_i-\bar{y})^2=\frac{1}{n-1}\sum_{1=1}^{n}(ax_i+b-a\bar{x}-b)^2=a^2\frac{1}{n-1}\sum_{1=1}^{n}(x_i-\bar{x})^2=a^2\,s^2{}_x$$

Pertanto, sommare una costante a ciascuno dei dati non fa cambiare la varianza campionaria, mentre moltiplicare ciascuno dei dati per un fattore costante fa sì che la varianza campionaria dell'insieme iniziale dei dati risulta moltiplicata per il quadrato di tale fattore. La media campionaria e la deviazione standard campionaria sono i due indici di posizione e di dispersione dei dati maggiormente utilizzati. Per confrontare le variazioni esistenti tra diversi campioni di dati è utile introdurre coefficiente di variazione.
Assegnato un insieme di dati numerici $x_1, x_2, \ldots, x_n$, si definisce **coefficiente di variazione** il rapporto tra la deviazione standard campionaria e il modulo della media campionaria, ossia:

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

Si nota che il coefficiente di variazione è un numero puro, ossia è un indice adimensionale che non dipende dall'unità di misura utilizzata, poiché la media campionaria e la deviazione standard campionaria sono espressi in identiche unità di misura.
Dalla precedente segue che il coefficiente di variazione è un indice di dispersione che ha senso soltanto per campioni aventi la media campionaria non nulla.
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).

```
> #VARIANZA CAMPIONARIA  > #DEVIAZIONE STANDARD CAMPIONARIA
> var(LPTsorted)         > sd(LPTsorted)
[1] 153.1493             [1] 12.37535
> var(LDTsorted)         > sd(LDTsorted)
[1] 142.3558             [1] 11.9313
> var(CLsorted)          > sd(CLsorted)
[1] 46.10783             [1] 6.790275
> var(NCLsorted)         > sd(NCLsorted)
[1] 14.08517             [1] 3.753021
```

Nel nostro caso, si nota come per tutti e quattro i vettori ed in particolare per il vettore LPT ed LDT i valori della varianza risultano essere molto elevati, ciò ci permette di notare che i valori si discostano

maggiormente dalla media che nel caso dei vettori LPT ed LDT corrisponde rispettivamente a 18,4 e 54,5.

```
> cv<-function(x){
+ sd(x)/abs(mean(x))}
> cv(LPTsorted)
[1] 0.6725735
> cv(LDTsorted)
[1] 0.2186721
> cv(CLsorted)
[1] 0.3471067
> cv(NCLsorted)
[1] 0.5012382
```

Il coefficiente di variazione più alto si ottiene proprio per i vettori LPT ed NCL i cui valori si discostano maggiormente dal valore della media campionaria.
Dunque, possiamo facilmente osservare che vi è una notevole variabilità dei valori per i vettori presi in esame. Questo conferma quanto osservato nel paragrafo 2.3, e mostrato nel boxplot.

CAPITOLO 4: STATISTICA DESCRITTIVA BIVARIATA

In questo capitolo sviluppiamo la statistica descrittiva bivariata, ossia il ramo della statistica che si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili.

Siano X e Y due variabili di tipo quantitativo. Considerato un campione (x_1, y_1), (x_2, y_2), …, (x_n, y_n) costituito da n osservazioni di (X, Y).
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 i punti in corrispondenza delle coppie (x_i, y_i).
Ricordiamo che una **variabile indipendente** è una grandezza che può assumere diversi valori, ma essi non dipendono da un'altra variabile; mentre, **una variabile dipendente** è una grandezza che può assumere diversi valori a seconda di quello assunto da altre variabili indipendenti.
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 $(x_1, x_2, \ldots, x_n)$ e y è il vettore contenente i valori $(y_1, y_2, \ldots, y_n)$.

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, …).

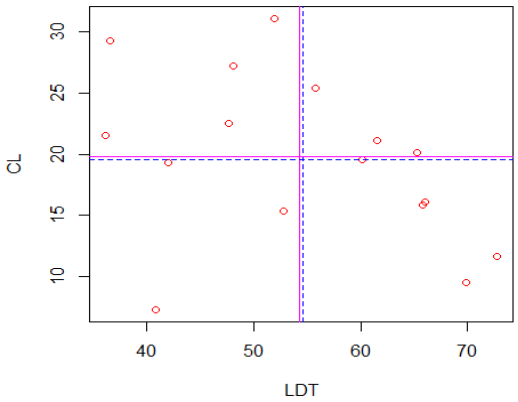
Nello scatterplot, inoltre, vengono tracciate delle **linee orizzontali e verticali** in corrispondenza delle mediane campionarie e delle medie campionarie dei vettori presi in esame.

Nello specifico, dunque, la funzione plot() costruisce il diagramma, abline() evidenzia la media e la mediana per le variabili, mentre legend() è la funzione che permette l'inserimento della legenda per comprendere ciò che è stato evidenziato nel grafico.

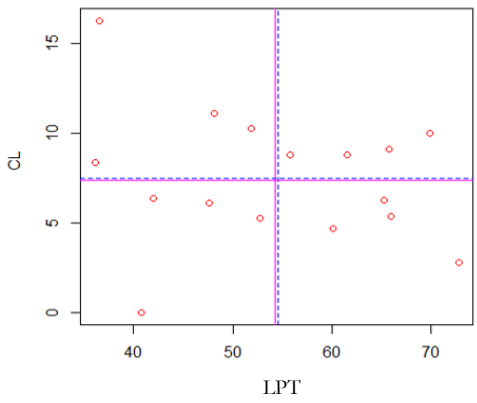
Nel nostro caso vogliamo occuparci di analizzare se vi sia qualche tipo di relazione tra:

- il cercare lavoro (CL), la variabile dipendente, e il lavorare dopo aver conseguito il titolo (LDT), la variabile indipendente.

```
> plot(occupazione$LDT, occupazione$CL, xlab="LDT", ylab="CL", col="red")
> abline(v=median(occupazione$LDT), lty=1, col="magenta")
> abline(v=mean(occupazione$LDT), lty=2, col="blue")
> abline(h=median(occupazione$CL), lty=1, col="magenta")
> abline(h=mean(occupazione$CL), lty=2, col="blue")
> legend(c("Mediana","Media"), pch=0, col=c("magenta", "blue"), cex=0.8)
```



- il cercare lavoro (CL), la variabile dipendente, e il lavorare prima di aver conseguito il titolo (LPT), la variabile indipendente.



I diagrammi di dispersione (scatterplot) risultanti mostrano come sembra esserci una correlazione negativa tra le variabili.

4.1 COVARIANZA E CORRELAZIONE CAMPIONARIA

Precedentemente abbiamo visto che un primo passo per indagare l'eventuale dipendenza tra due variabili quantitative X e Y consiste nel disegnare il diagramma di dispersione o scatterplot.

Per ottenere una misura quantitativa della correlazione tra le variabili si considera **la covarianza campionaria**:
assegnato un campione bivariato $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ di una variabile quantitativa bidimensionale (X, Y), siano $\bar{x}$ e $\bar{y}$ **rispettivamente le medie campionarie** di $x_1, x_2, \ldots, x_n$ e di $y_1, y_2, \ldots, y_n$.
La covarianza campionaria tra le due variabili X e Y è così definita:

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

Si nota che se un valore x_i della variabile quantitativa X è grande rispetto a $\bar{x}$, allora la differenza $x_i - \bar{x}$ sarà positiva, mentre se x_i è piccolo rispetto a $\bar{x}$, allora la differenza $x_i - \bar{x}$ sarà negativa. È possibile ragionare in modo analogo per i valori della variabile quantitativa Y.
Inoltre, se l'intero campione presenta una forte correlazione, c'`e da aspettarsi che $\sum_{1=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$ assuma un valore molto positivo o molto negativo. 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 $i = 1, 2, \ldots, n$.

La covarianza campionaria può avere segno positivo, negativo o nullo. Quando è > 0 si dice che le variabili sono correlate positivamente, se è < 0 variabili sono correlate negativamente e, infine, se $= 0$ le variabili sono non correlate.

Per ottenere una misura quantitativa della correlazione tra le variabili si può anche considerare il **coefficiente di correlazione campionario:**
assegnato un campione bivariato $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ di una variabile quantitativa bidimensionale (X, Y), siano $\bar{x}$ e $\mathbf{s_x}$ rispettivamente la **media campionaria** e la **deviazione standard campionaria** di $x_1, x_2, \ldots, x_n$ ed inoltre siano $\bar{y}$ e $\mathbf{s_y}$ la media campionaria e la deviazione standard campionaria di $y_1, y_2, \ldots, y_n$.
Il coefficiente di correlazione campionario tra due variabili X e Y è così definito:

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

Si nota che il coefficiente di correlazione campionario:
- è un indice adimensionale;
- non fa distinzione tra variabile dipendente e variabile indipendente, ossia $r_{xy} = r_{yx}$;
- può essere calcolato soltanto se entrambe le variabili sono quantitative;
- non cambia al variare dell'unità di misura con cui sono espresse le variabili;

- è fortemente influenzato dalla presenza di eventuali valori anomali, così come accade per la media campionaria e la varianza campionaria.

Il coefficiente di correlazione campionario ha lo stesso segno della covarianza.
Quando è > 0 si dice che le variabili sono correlate positivamente, se è < 0 le variabili sono correlate negativamente e, infine, se è = 0 le variabili sono non correlate.

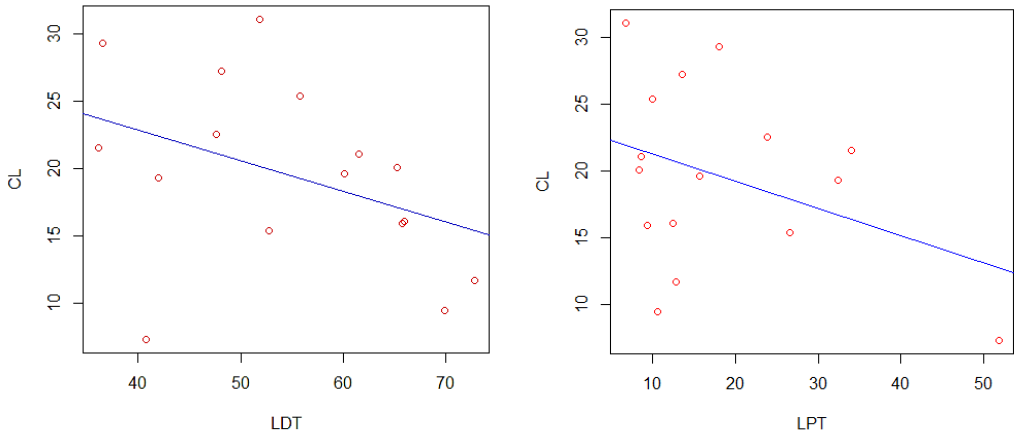
È importante ricordare che il coefficiente di correlazione campionario misura la forza del legame di natura lineare esistente tra due variabili quantitative, ciò significa che eventuali relazioni tra le variabili che assumono una forma curvilinea non possono essere individuati con tale coefficiente.

Riassumendo si può dire che il segno del coefficiente di correlazione campionario indica la direzione della retta interpolante e indica la presenza di una tra le seguenti situazioni:
- se = 1 (correlazione perfetta positiva) tutti i punti sono allineati su una linea retta ascendente;
- se 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, y_i tendono ad essere grandi e piccoli insieme);
- se = 0 (nessuna correlazione) i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare;
- se 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);
- se = −1 (correlazione perfetta negativa) tutti i punti sono allineati su una linea retta discendente.

Dopo aver compreso il tutto in linea teorica, vediamo come è possibile con il sistema R calcolare la covarianza e la correlazione campionaria fra una coppie di variabili numeriche X e Y: possono essere immediatamente ottenute con le rispettive funzioni cov(X, Y) e cor(X, Y).
Ne vengono mostrati, inoltre, i grafici della retta di regressione: la funzione utilizzata relativamente ai grafici riportati è abline che permette di aggiungere allo scatterplot la linea interpolante stimata (retta di regressione). In abline() è presente la funzione lm(), acronimo di linear model.

```
> #COVARIANZA CAMPIONARIA
> cov(occupazione$LDT, occupazione$CL)
[1] -32.4935
> cov(occupazione$LPT, occupazione$CL)
[1] -31.29533
```

```
> #COEFFICIENTE DI CORRELAZIONE CAMPIONARIO
> cor(occupazione$LDT, occupazione$CL)
[1] -0.4010712
> cor(occupazione$LPT, occupazione$CL)
[1] -0.3724214
```



Sia osservando i valori della correlazione campionaria che i grafici della retta di regressione, viene confermata l'osservazione precedentemente fatta mediante scatterplot, ossia che le variabili sono correlate negativamente.
Inoltre, il valore del coefficiente di correlazione, per entrambi i casi, è compreso tra -1 e 0, ciò dimostra che i punti sono posizionati in una nuvola attorno ad una linea retta interpolante discendente.

4.2 REGRESSIONE LINEARE SEMPLICE

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.
Consideriamo l'equazione della retta:

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

dove α è l'intercetta e β è il coefficiente angolare.

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**. Questa è una tecnica di regressione che permette di trovare una funzione, rappresentata da una curva ottima (o curva di regressione), che si avvicini il più possibile ad un insieme di dati.

I coefficienti di regressione sono i valori α e β per i quali la somma Q dei quadrati degli errori

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

sia minima, dove n è il numero di osservazioni, $x_1, x_2, \ldots, x_n$ sono i valori osservati della variabile X e $y_1, y_2, \ldots, y_n$ sono i valori osservati della variabile Y.
L'applicazione del metodo dei minimi quadrati, utilizzando le formule della varianza e della covarianza campionaria, conduce ai seguenti valori di β e α:

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

La funzione

lm(y~x)

è utilizzata per eseguire le analisi di regressione lineari e fornisce i valori dell'intercetta α e del coefficiente angolare β. Ricordiamo che il nome lm() della funzione rappresenta l'acronimo di linear model.
L'argomento y~x passato alla funzione lm() indica che y dipende da x, ossia che x è la variabile indipendente e y la variabile dipendente.

La funzione

abline(lm(y~x))

permette di aggiungere la retta di regressione al grafico dello scatterplot, con y che dipende da x.

È possibile visualizzare successivamente l'analisi di regressione lineare sulla console di R per ogni variabile indipendente. Notiamo che i valori dell'intercetta α sono tutti positivi, segno che il punto di intersezione della retta di regressione con l'asse delle ordinate risulta essere al di sopra dell'origine.
Considerando i coefficienti angolari β notiamo che, in entrambi i casi, essendo negativi; la retta di regressione è discendente e va dall'alto a sinistra verso il basso a destra.

```
> linearmodelLDTCL<-lm(occupazione$CL~occupazione$LDT)
> linearmodelLDTCL$coefficients
    (Intercept) occupazione$LDT
     32.0166900      -0.2282555
> linearmodelLPTCL<-lm(occupazione$CL~occupazione$LPT)
> linearmodelLPTCL$coefficients
    (Intercept) occupazione$LPT
     23.3224519      -0.2043452
```

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.**

Denotiamo con

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

i valori stimati ottenuti mediante la retta di regressione.

La media campionaria dei valori stimati $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ è uguale alla media campionaria $\bar{y}$ delle osservazioni $y_1, y_2, \ldots, y_n$. Infatti, risulta che

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} (\alpha + \beta x_i) = \alpha + \beta\bar{x} = (\bar{y} - \beta\bar{x}) + \beta\bar{x} = \bar{y}$$

I residui sono così definiti

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

e mostrano di quanto si discostano i valori osservati y_i dai valori stimati $\hat{y}_i$ con la retta di regressione.

La media campionaria dei residui $\bar{E}$ è nulla, ossia in media gli scostamenti positivi e negativi si compensano. Infatti, ricordando la formula precedente risulta:

$$\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = 0$$

La varianza campionaria dei residui è

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

essendo $\bar{E} = 0$.

I **valori stimati** sono calcolati utilizzando l'equazione di regressione e i valori per ogni variabile esplicativa. Idealmente, i valori stimati sarebbero uguali ai valori effettivi della variabile dipendente.

I **valori residui** sono la differenza tra i valori osservati e stimati in un'analisi di regressione. In base alla posizione che i valori osservati assumono rispettivamente alla curva di regressione, il valore residuo cambia.

Nel particolare:

- i valori osservati che si trovano al di sopra della curva di regressione hanno un valore residuo positivo;
- i valori osservati che scendono al di sotto della curva di regressione hanno un valore residuo negativo.

La **curva di regressione** deve trovarsi lungo il centro dei punti dati, da ciò ne consegue che la somma dei residui deve essere zero.

In R per calcolare **il vettore dei valori stimati** $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ si utilizza la funzione fitted(lm(y~x)) con y che dipende da x.

Invece, per calcolare **il vettore dei residui** $(E_1, E_2, \ldots, E_n)$ si utilizza la funzione resid(lm(y~x)).

```
> #RESIDUI
> stimeLDTCL<-fitted(lm(occupazione$CL~occupazione$LDT))
> stimeLDTCL
       1        2        3        4        5        6        7        8
16.06163 17.11161 20.17023 15.39969 16.99748 19.28003 18.29854 16.95183
       9       10       11       12       13       14       15       16
21.15173 23.77667 21.03760 17.97898 22.42996 23.68536 19.96480 22.70387
> stimeLPTCL<-fitted(lm(occupazione$CL~occupazione$LPT))
> stimeLPTCL
       1        2        3        4        5        6        7        8
21.15639 21.62639 21.95334 20.70683 21.42204 21.27900 20.13467 20.78857
       9       10       11       12       13       14       15       16
18.45904 16.37471 20.54336 21.56508 16.70167 19.64424 17.90730 12.73737
> residuiLDTCL<-resid(lm(occupazione$CL~occupazione$LDT))
> residuiLDTCL
          1           2           3           4           5           6
 -6.5616315   2.9883933  10.9297698  -3.6996906  -1.0974790   6.1199662
          7           8           9          10          11          12
  1.3014648  -0.8518279   1.3482712  -2.2766669   6.1623989   3.1210224
         13          14          15          16
 -3.1299595   5.6146353  -4.5648003 -15.4038661
> residuiLPTCL<-resid(lm(occupazione$CL~occupazione$LPT))
> residuiLPTCL
          1           2           3           4           5           6
-11.6563927  -1.5263867   9.1466610  -9.0068332  -5.5220414   4.1210002
          7           8           9          10          11          12
 -0.5346666  -4.6885713   4.0409642   5.1252853   6.6566430  -0.4650831
         13          14          15          16
  2.5983330   9.6557619  -2.5073038  -5.4373698
```

Ad esempio, il primo valore del vettore stimeLDTCL può essere cosi estratto: $y_1 = \alpha + \beta\, x_1 = 32.0166900 + (-0.2282555 * 69.9)$ |69.9 è il primo valore di LDT| = 16.0616.
Ad esempio, il primo valore del vettore residuiLDTCL può essere cosi estratto: $E_1 = y_1 - \hat{y}_1 = 9.5$ |primo valore di CL| - 16.06163 = - 6.5616315.

Successivamente sono presenti i calcoli in R per la **mediana, la varianza campionaria e la deviazione standard campionaria dei residui** per ogni variabile indipendente. Non si può invece calcolare il coefficiente di variazione, essendo la media campionaria dei residui nulla.

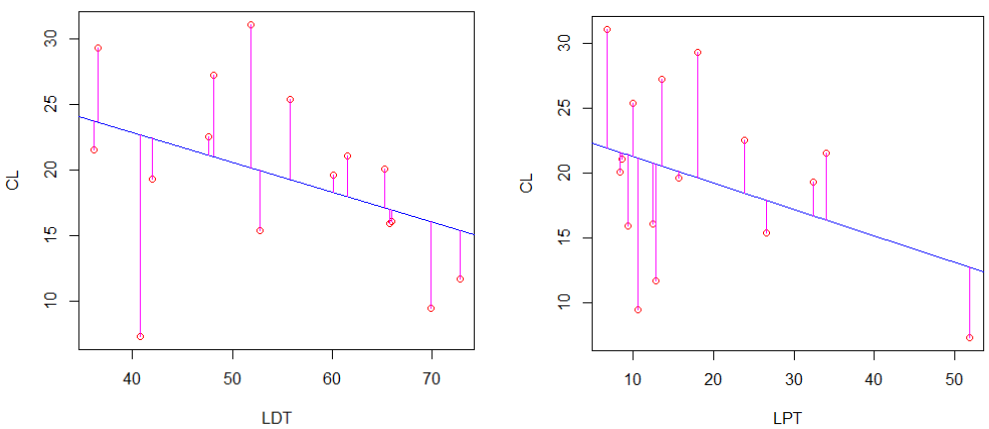
```
> median(linearmodelLDTCL$residuals)
[1] 0.2248184
> var(linearmodelLDTCL$residuals)
[1] 38.69101
> sd(linearmodelLDTCL$residuals)
[1] 6.22021
```

```
> median(linearmodelLPTCL$residuals)
[1] -0.4998748
> var(linearmodelLPTCL$residuals)
[1] 39.71278
> sd(linearmodelLPTCL$residuals)
[1] 6.301808
```

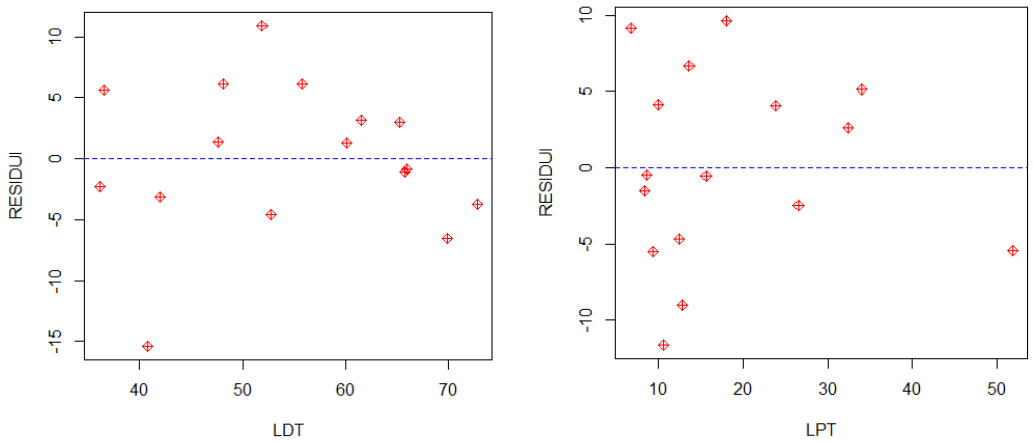
È possibile **rappresentare graficamente i residui**:

- 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 osservati:
Vogliamo ora realizzare i grafici dei residui ottenuti aggiungendo, ai grafici contenenti scatterplot e retta di regressione, dei segmenti verticali che visualizzano i residui.
Sulla base dei grafici di seguito riportati, i valori osservati che si trovano al di sopra della retta di regressione hanno un valore residuo positivo e i valori osservati che scendono al di sotto della retta di regressione hanno un valore residuo negativo.



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.
Quindi 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) 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.
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 delle variabili indipendenti. La retta orizzontale è posizionata nello zero e corrisponde alla media campionaria dei residui che è nulla.



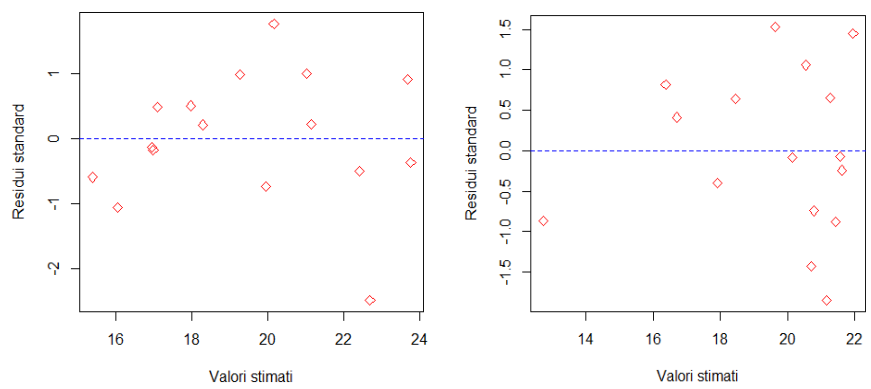
Possiamo notare la presenza di un valore anomalo nel grafico a destra, valore posizionato in basso a destra. Questo valore anomalo per LPT è lo stesso riscontrato anche nel boxplot.
Per entrambi i grafici è possibile notare che i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia nessuna tendenza particolare nella distribuzione dei punti.

Valori dei residui standardizzati rispetto ai valori stimati:
È spesso interessante calcolare i residui standardizzati così definiti:

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

che risultano essere caratterizzati da media campionaria nulla e varianza unitaria.

Per LDT: Per LPT:



Considerando il valore non elevato del coefficiente di correlazione ottenuto per i due casi considerati, i residui sono elevati (variano tra -15 e 10). Di conseguenza, i grafici finali con i residui standard non sono simili a quelli precedenti.
Inoltre, notiamo come anche in questo caso i punti sono disposti quasi casualmente intorno alla linea orizzontale senza evidenziare nessuna tendenza particolare nella distribuzione dei punti.

COEFFICIENTE DI DETERMINAZIONE

Poiché si è interessati a vedere quanto la retta si adatta ai dati, l'accento può essere posto sul quadrato del coefficiente di correlazione e su quanto esso si avvicini ad uno.
È chiaro che r^2_{xy} molto vicino ad 1 indicherà che tutti i punti tenderanno ad allinearsi lungo la retta di regressione, mentre r^2_{xy} prossimo a 0 esprime una completa incapacità della retta di rappresentare la distribuzione dei dati considerati.

Il coefficiente di determinazione per la regressione lineare semplice è il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati.
Pertanto, se si denota con $y_1, y_2, \ldots, y_n$ il vettore dei dati della variabile dipendente, con $\bar{y}$ la sua media campionaria e con $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ i valori stimati attraverso la retta di regressione (la cui media campionaria è $\bar{y}$), il coefficiente di determinazione (detto anche r-square) è così definito:

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

Nel caso di regressione lineare semplice, il coefficiente di determinazione coincide con il quadrato del coefficiente di correlazione, ossia

$$D^2=r^2_{xy}$$

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.

```
> (cor(occupazione$LPT, occupazione$CL))^2
[1] 0.1386977
> (cor(occupazione$LDT, occupazione$CL))^2
[1] 0.1608581
```

I risultati ottenuti confermano quanto mostrato nei grafici. I valori sono prossimi allo 0, ciò implica la completa incapacità della retta di rappresentare la distribuzione dei dati.

4.3 REGRESSIONE LINEARE MULTIPLA

La regressione lineare multipla è un'estensione dell'analisi della correlazione e della regressione lineare semplice. Nel modello di regressione lineare semplice che abbiamo già visto, si studia la relazione tra la variabile dipendente (che chiameremo Y) ed una sola variabile indipendente (che chiameremo X).
Nel modello di regressione lineare multipla invece si includono due o più variabili indipendenti per studiare contemporaneamente l'effetto di più X sulla Y.
La costruzione di un modello di regressione lineare multipla permette di quantificare la relazione esistente tra la variabile dipendente ed un insieme di variabili indipendenti. Inoltre, aiuta a predire quale sarà il valore della Y per determinati valori di X.

Quindi, per ogni coppia di variabili possiamo utilizzare il modello di regressione lineare semplice esprimibile attraverso l'equazione $Y = \alpha + \beta X$; mentre, il modello di regressione lineare multipla con p variabili indipendenti si esprime attraverso l'equazione:

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

Per determinare le stime di $\alpha, \beta_1, \ldots \beta_p$ ricorriamo al metodo dei minimi quadrati. Occorre quindi minimizzare la quantità

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

dove n è il numero di osservazioni, $x_{1,J}, x_{2,J} \ldots x_{n,J}$ sono i valori osservati della variabile $X_J (j=1,2\ldots,p)$ e $(y_1 y_2 \cdots y_n)$ i valori osservati della variabile Y.
Derivando rispetto ai parametri $\alpha, \beta_1, \beta_2 \ldots \beta_p$ si perviene ad un sistema di p + 1 equazioni.
Dalla prima equazione ricaviamo

$$\bar{y} - \left(\alpha + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \ldots + \beta_p \bar{x}_p \right) = 0$$

dove

$$\bar{x}_J = \frac{1}{n} \sum_{i=1}^{n} x_{i,j} \ (j=1,2,\ldots,p)$$

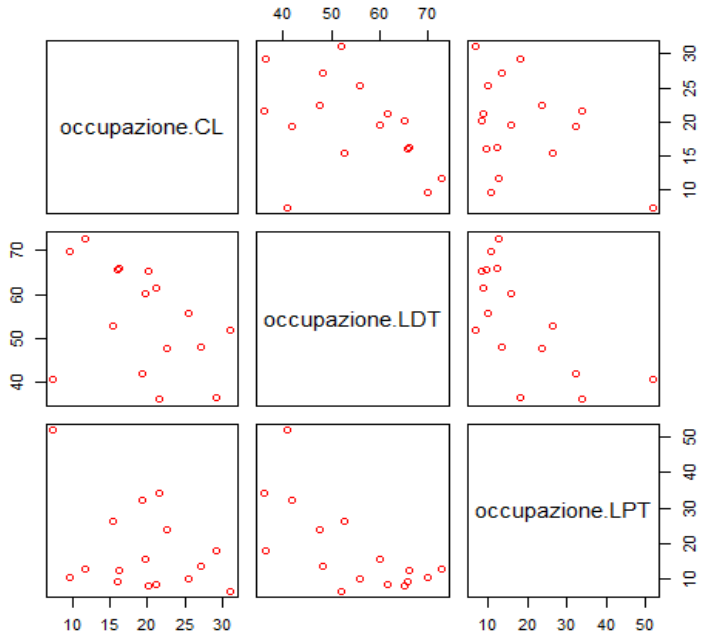
è la media campionaria relativa alla variabile X_J (effettuata sulla colonna j-esima).

In R, le due funzioni cov(dataframe) e cor(dataframe) forniscono due matrici simmetriche. **La matrice delle covarianze** contiene sulla diagonale principale la varianza delle singole colonne del data frame, mentre la **matrice delle correlazioni** contiene il numero 1 sulla diagonale principale.
La matrice di correlazione evidenzia tutte le correlazioni lineari tra le coppie di variabili, ossia misura la forza del legame di natura lineare esistente tra tutte le coppie di variabili quantitative.

```
> dfm
   occupazione.CL occupazione.LDT occupazione.LPT
1             9.5            69.9            10.6
2            20.1            65.3             8.3
3            31.1            51.9             6.7
4            11.7            72.8            12.8
5            15.9            65.8             9.3
6            25.4            55.8            10.0
7            19.6            60.1            15.6
8            16.1            66.0            12.4
9            22.5            47.6            23.8
10           21.5            36.1            34.0
11           27.2            48.1            13.6
12           21.1            61.5             8.6
13           19.3            42.0            32.4
14           29.3            36.5            18.0
15           15.4            52.8            26.5
16            7.3            40.8            51.8
> cov(dfm)
                occupazione.CL occupazione.LDT occupazione.LPT
occupazione.CL        46.10783        -32.4935       -31.29533
occupazione.LDT      -32.49350        142.3558      -100.06800
occupazione.LPT      -31.29533       -100.0680       153.14933
> cor(dfm)
                occupazione.CL occupazione.LDT occupazione.LPT
occupazione.CL       1.0000000      -0.4010712      -0.3724214
occupazione.LDT     -0.4010712       1.0000000      -0.6777196
occupazione.LPT     -0.3724214      -0.6777196       1.0000000
```

Nel nostro caso, consideriamo il valore di CL in relazione alle restanti variabili indipendenti. Possiamo notare che non vi è alcuna correlazione lineare fra le varie caratteristiche in quanto i valori sono tutti negativi.

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.
Tale grafico viene mostrato nella figura sottostante e conferma quanto detto in precedenza.



Mettendo in relazione le variabili indipendenti con la variabile dipendente CL otteniamo l'output mostrato:

```
> lm(dfm$occupazione.CL~dfm$occupazione.LDT+dfm$occupazione.LPT)

Call:
lm(formula = dfm$occupazione.CL ~ dfm$occupazione.LDT + dfm$occupazione.LPT)

Coefficients:
        (Intercept)  dfm$occupazione.LDT  dfm$occupazione.LPT
            69.1206              -0.6878              -0.6538
```

In tale figura l'intercetta ha un valore pari a 69.1206; i regressori hanno valori pari a -0.6878 e -0.6538. Dunque, il modello di regressione multipla stimato risulta essere uguale a:

$$y = 69.1206 + (-0.6878x1 - 0.6538x2)$$

Notiamo che l'intercetta ha un valore positivo. Entrambi i valori dei regressori sono negativi quindi la percentuale di persone che lavora prima del conseguimento del titolo e dopo il conseguimento dello stesso incidono negativamente sulla percentuale di persone che cercano lavoro.

RESIDUI

Una volta calcolati i valori dei coefficienti α e $\beta_1,\beta_2\ldots\beta_p$ è possibile osservare gli scostamenti (ovvero i residui) tra le ordinate dei punti (valori osservati) y_i e i corrispondenti valori stimati

$$\hat{y}_i=\alpha+\beta_1 x_{i,1} + \beta_2 x_{i,2} +\ldots+ \beta_p x_{i,p} \quad (i=1,2, \ldots, n)$$

ottenuti mediante la regressione lineare multipla.

Come abbiamo visto nella regressione lineare semplice, anche in questo caso si ha che **la media campionaria dei valori stimati coincide con la media campionaria dei valori osservati**, infatti abbiamo che:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}\left(\alpha+\beta_1 x_{i,1} +\beta_2 x_{i,2}+\ldots++\beta_p x_{i,p}\right)$$

$$= \alpha+\beta_1\bar{x}_1 +\beta_2\bar{x}_2+\ldots++\beta_p\bar{x}_p=\left(\bar{y}- \beta_1\bar{x}_1- \beta_2\bar{x}_2- \ldots- \beta_p\bar{x}_p\right) + \beta_1\bar{x}_1 +\beta_2\bar{x}_2+\ldots++\beta_p\bar{x}_p= \bar{y}$$

I residui

$$E_i=y_i-\hat{y}_i= y_i-\left(\alpha+\beta_1 x_{i,1} +\beta_2 x_{i,2}+\ldots++\beta_p x_{i,p}\right) \quad (i=1,2,\ldots, n)$$

mostrano di quanto si discostano i valori osservati dai valori stimati con la regressione lineare multipla.

La media campionaria dei residui $\bar{E}$ è nulla, ossia in media gli scostamenti positivi e negativi si compensano. Infatti, ricordando la formula precedente, risulta:

$$\bar{E}=\frac{1}{n}\sum_{i=1}^{n} E_i=\frac{1}{n}(y_i-\hat{y}_i)=\bar{y}-\frac{1}{n}\sum_{i=1}^{n}\hat{y}=0$$

La varianza campionaria dei residui è

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

Anche nel caso multivariato è interessante calcolare **i residui standardizzati** così definiti:

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

che risultano essere caratterizzati da media campionaria nulla e varianza unitaria.

Successivamente viene mostrata la console di R contenente i valori stimati e i valori residui. Per calcolare il **vettore dei valori stimati** tramite regressione lineare multipla $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ si utilizza la funzione

fitted(lm(y~x_1 + x_2 + . . . + x_p)

con y che dipende da x_1, x_2 … x_p.
Invece, per calcolare il **vettore dei residui** $(E_1, E_2, \ldots, E_n)$ si utilizza la funzione

resid(lm(y~x_1 + x_2 + . . . + x_p)

```
> stimemult<-fitted(multiplelinearmodel)
> stimemult
       1        2        3        4        5        6        7        8        9
14.112510 18.780112 29.042844 10.679568 17.782441 24.202947 17.584268 15.618210 20.821080
      10       11       12       13       14       15       16
22.062550 27.145566 21.197676 19.050469 32.247648 15.479284  7.192825

> residuimult<-resid(multiplelinearmodel)
> residuimult
         1          2          3          4          5          6          7
-4.61250979  1.31988804  2.05715627  1.02043196 -1.88244091  1.19705254  2.01573162
         8          9         10         11         12         13         14
 0.48179026  1.67891960 -0.56255032  0.05443404 -0.09767648  0.24953070 -2.94764839
        15         16
-0.07928415  0.10717500
```

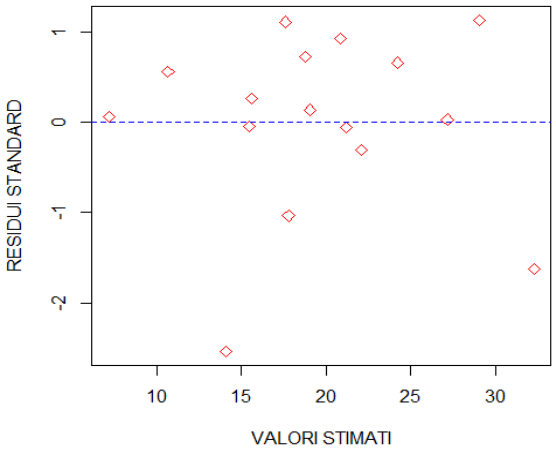
La mediana, la varianza e la deviazione standard campionaria dei residui vengono calcolate di seguito:

```
> median(multiplelinearmodel$residuals)
[1] 0.1783528
> var(multiplelinearmodel$residuals)
[1] 3.298583
> sd(multiplelinearmodel$residuals)
[1] 1.8162
```

Calcoliamo ora i residui standardizzati:

```
> residuimultstandard<-residuimult/sd(residuimult)
> residuimultstandard
         1          2          3          4          5          6          7
-2.53964838  0.72673049  1.13267046  0.56184995 -1.03647216  0.65909725  1.10986204
         8          9         10         11         12         13         14
 0.26527377  0.92441331 -0.30974027  0.02997139 -0.05378068  0.13739163 -1.62297551
        15         16
-0.04365386  0.05901057
```

Possiamo, dunque, realizzare il grafico in cui i residui standardizzati vengono disegnati in funzione dei valori stimati con il metodo dei minimi quadrati:



Dall'analisi del grafico 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. Come nella regressione lineare semplice, è possibile notare come i punti siano disposti senza un particolare ordine rispetto alla retta orizzontale. Inoltre, anche in questo caso non si evidenzia una particolare tendenza nella distribuzione degli stessi.

COEFFICIENTE DI DETERMINAZIONE

Il coefficiente di determinazione in un modello di regressione lineare multipla è il rapporto tra la varianza dei valori stimati tramite la funzione di regressione multipla e la varianza dei valori osservati della variabile dipendente.

Se si denota con $(y_1, y_2 \cdots y_n)$ il vettore dei dati della variabile dipendente, con $\bar{y}$ la sua media campionaria e con $(\hat{y}_1, \hat{y}_2 \cdots \hat{y}_n)$ i valori stimati attraverso la funzione di regressione (la cui media campionaria è $\bar{y}$, il coefficiente di determinazione è:

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_1 - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_1 - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_1 - \bar{y})^2}{\sum_{i=1}^{n}(y_1 - \bar{y})^2}$$

L'indice D^2 è adimensionale e risulta $0 \leq D^2 \leq 1$. Quando $D^2 = 0$ il modello di regressione multipla utilizzato non spiega per nulla i dati. Invece, quando $D^2 = 1$ il modello di regressione multipla utilizzato spiega perfettamente i dati.

In R per calcolare l'indice D^2 per la regressione lineare multipla basta utilizzare la funzione: summary(lm(y˜x1 + x2 + . . . + xp))$r.square.

Il coefficiente di determinazione è 0.9284594, ossia il modello di regressione multipla utilizzato può spiegare significativamente i dati:

```
> summary(lm(dfm$occupazione.CL~dfm$occupazione.LDT+dfm$occupazione.LPT))$r.square
[1] 0.9284594
```

Occorre poi notare che rispetto al modello di regressione semplice, si è ottenuto un notevole miglioramento di tale indice che risultava aggirarsi intorno al valore 0.13 e 0.16.

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, nel nostro caso la scelta è ricaduta su un modello di regressione quadratica.

4.4.1 REGRESSIONE QUADRATICA

Il modello non lineare più semplice è il modello polinomiale di secondo ordine:

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

Visualizziamo di seguito i due data frame con la variabile dipendente e la variabile indipendente:

```
> dfp
   occupazione.LDT occupazione.CL
1             69.9            9.5
2             65.3           20.1
3             51.9           31.1
4             72.8           11.7
5             65.8           15.9
6             55.8           25.4
7             60.1           19.6
8             66.0           16.1
9             47.6           22.5
10            36.1           21.5
11            48.1           27.2
12            61.5           21.1
13            42.0           19.3
14            36.5           29.3
15            52.8           15.4
16            40.8            7.3
```

```
> dfp2
   occupazione.LPT occupazione.CL
1             10.6            9.5
2              8.3           20.1
3              6.7           31.1
4             12.8           11.7
5              9.3           15.9
6             10.0           25.4
7             15.6           19.6
8             12.4           16.1
9             23.8           22.5
10            34.0           21.5
11            13.6           27.2
12             8.6           21.1
13            32.4           19.3
14            18.0           29.3
15            26.5           15.4
16            51.8            7.3
```

Per la stima dei 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 $X_2 = X^2$.

Con R è facile stimare i parametri α, β, γ tramite la funzione

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

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

Il modello polinomiale di secondo ordine per il primo caso considerato è:

```
> pol2

Call:
lm(formula = dfp$occupazione.CL ~ dfp$occupazione.LDT + I((dfp$occupazione.LDT)^2))

Coefficients:
                (Intercept)         dfp$occupazione.LDT
                  -25.55008                    2.01369
I((dfp$occupazione.LDT)^2)
                   -0.02082
```

$$Y = -25.55008 + 2.01369 - 0.02082$$

Il modello polinomiale di secondo ordine per il secondo caso considerato è:

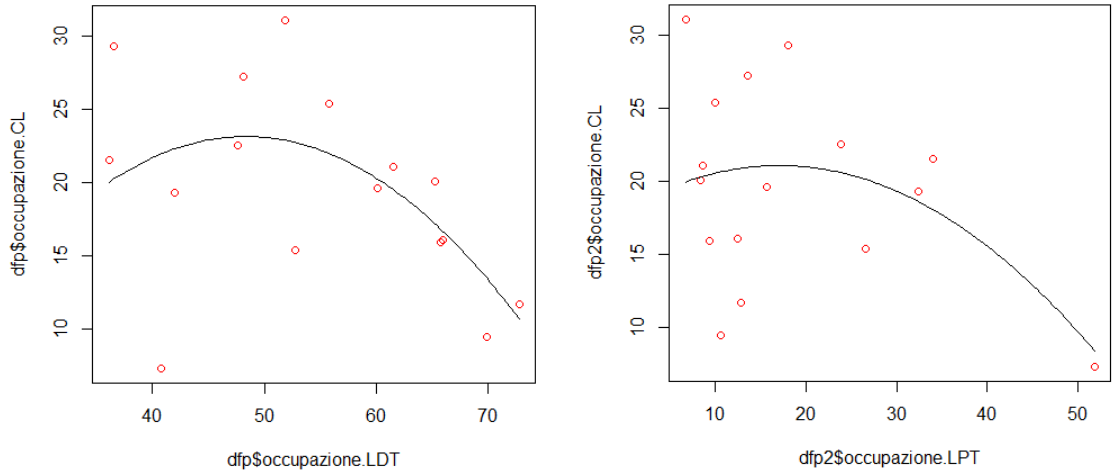
```
> pol2lpt

Call:
lm(formula = dfp2$occupazione.CL ~ dfp2$occupazione.LPT + I((dfp2$occupazione.LPT)^2))

Coefficients:
                (Intercept)         dfp2$occupazione.LPT  I((dfp2$occupazione.LPT)^2)
                    18.0302                       0.3585                      -0.0105
```

$$Y = 18.0302 + 0.3585 - 0.0105$$

Calcoliamo, ora, i coefficienti di determinazione per entrambi i casi di studio:

```
> summary(lm(dfp$occupazione.CL~dfp$occupazione.LDT+I((dfp$occupazione.LDT)^2)))$r.square
[1] 0.2948097

> summary(lm(dfp2$occupazione.CL~dfp2$occupazione.LPT+I((dfp2$occupazione.LPT)^2)))$r.square
[1] 0.206731
```

Infine, disegniamo la curva stimata sugli scatterplot:



Avendo a disposizione i risultati ottenuti per la regressione lineare semplice e regressione non lineare, è possibile analizzare quale si adatta meglio ai dati a nostra disposizione.
Nel primo caso, il valore del coefficiente di determinazione D^2 è pari a:

LDT - 0.1608581
LPT- 0.1386977

Nel caso secondo caso è pari a:

LDT - 0.2948097
LPT - 0.206731

I risultati ottenuti ci permettono di affermare con certezza che il valore di regressione non lineare rappresenta nel modo migliore i dati.

CAPITOLO 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.
L'insieme originario delle entità su cui si attua l'analisi per ricavare i cluster non è sottoposto ad alcuna restrizione. Può infatti contenere variabili, individui, osservazioni, dati, misure, …
I vari metodi attraverso cui si attua l'analisi dei cluster hanno in comune uno **scopo generale**: ottenere raggruppamenti in base alla somiglianza in modo che gli elementi di uno stesso gruppo siano tra loro il più possibile simili e gli elementi appartenenti a gruppi distinti siano tra loro il più possibile diversi.
In altre parole, tale analisi ha lo scopo di distribuire le osservazioni in gruppi in modo tale che il grado di naturale associazione sia alto tra i membri dello stesso gruppo e basso tra i membri di gruppi diversi.
In questo modo si otterrà quindi un'alta omogeneità all'interno dei gruppi e un'alta eterogeneità tra gruppi distinti.
Lo scopo di tale analisi è quello di organizzare in una struttura un insieme di dati e di scoprire quali siano i legami esistenti tra essi.
Sia $I=\{I_1,I_2,\ldots,I_n\}$ un insieme di n individui (entità o unità) appartenenti ad una popolazione.
Assumiamo che esista un insieme di caratteristiche (features) $C=\{C_1,C_2,\ldots,C_n\}$ che sono osservabili e sono possedute da ogni individuo in I. Il termine osservabile denota caratteristiche che danno origine a dati sia di tipo qualitativo che di tipo quantitativo (detti anche misure).
Denotiamo con il simbolo x_{ij} il valore della misura della caratteristica j– esima relativa all'individuo I_i e con $X_i=\left(x_{i1},x_{i2}\ldots,x_{ip}\right)$ (i=1,2…n) il vettore di cardinalità 1×p di tali misure.
Quindi, in generale l'iniziale naturale collezione dei dati su cui il ricercatore deve operare consiste di un insieme di n vettori di misure $\{X_1, X_2 \ldots X_n\}$ che descrive l'insieme I degli individui a cui è associata una matrice di misure X di cardinalità n × p, ossia

$$x=\begin{pmatrix} x_{11} & x_{12} & \ldots \\ x_{21} & x_{11} & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots \end{pmatrix}=\begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ \cdots \\ X_n \end{pmatrix}$$

dove $X_i=\left(x_{i1},x_{i2}\ldots,x_{ip}\right)$ (i=1,2…n).
Il problema dell'analisi dei cluster consiste nel determinare m sottoinsiemi (cluster), di individui.
Gli individui assegnati allo stesso cluster sono detti simili, altrimenti dissimili.

5.1 DISTANZA E SIMILARITA'

Le misure metriche di somiglianza sono soprattutto basate sulle funzioni distanza tra i vettori delle caratteristiche. Occorre quindi definire tale funzione.

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

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

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

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

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

La proprietà *(i)* implica che X_i è a distanza zero da se stesso e che ogni due punti a distanza nulla debbono essere identici.
La proprietà *(ii)* afferma che la funzione distanza è non negativa.
La proprietà *(iii)* impone la simmetria richiedendo che la distanza tra X_i e X_j deve essere la stessa della distanza tra X_je X_i.

La proprietà *(iv)*, nota come 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.
In generale, le distanze tra tutte le possibili coppie di unità sono inserite in una matrice simmetrica D di cardinalità n × n, ossia

$$D=\begin{pmatrix} 0 & d_{12} & \dots \\ d_{21} & 0 & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots \end{pmatrix}$$

dove $d_{ij}=d(X_i, X_j)$ (i,j=1,2,…,n). Segue che i termini sulla diagonale principale sono tutti uguali a zero mentre i termini simmetrici sono uguali a due a due.

In R la funzione:

>dist(X, method = ”euclidean”, diag = FALSE, upper = FALSE)

ritorna la matrice delle distanze D calcolata utilizzando le misure di distanza tra le righe della matrice X dei dati, dove:

- X rappresenta una matrice numerica o un data frame;
- method seleziona la misura di distanza da utilizzare (di default è euclidean);
- diag è posta uguale a TRUE se si desidera che la matrice delle distanze contenga anche i valori nulli sulla diagonale (di default è FALSE);
- upper è posta uguale a TRUE se si desidera che la matrice delle distanze contenga anche i valori al di sopra della diagonale principale (di default è FALSE).

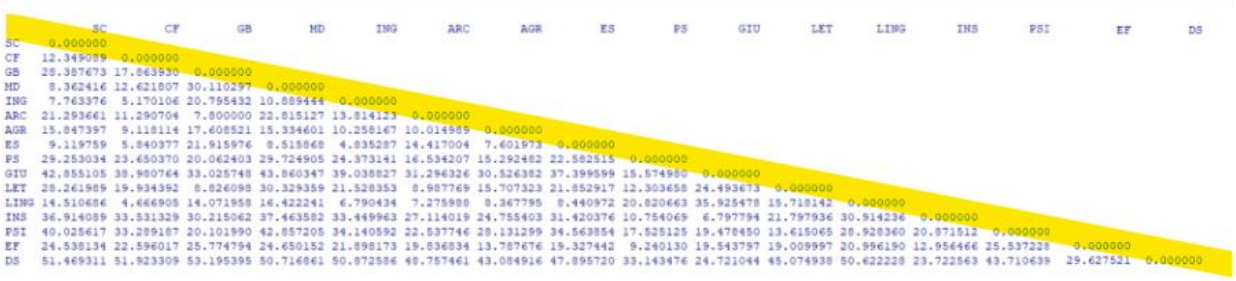
Le opzioni disponibili per il calcolo della matrice delle distanze sono:

(1) metrica euclidea (euclidean);
(2) metrica del valore assoluto o metrica di Manhattan (manhattan);
(3) metrica del massimo o metrica di Chebycev (maximum);
(4) metrica di Minkowski (minkowski);
(5) distanza di Canberra (canberra);
(6) distanza di Jaccard (binary).

L’output della funzione precedentemente citata è la matrice delle distanze per tutte le caratteristiche presenti nel set di dati:

	SC	CF	GB	MD	ING	ARC	AGR	ES	PS	GIU	LET	LING	INS	PSI	EF	DS
SC	0.000000	12.349089	28.387673	8.362416	7.763376	21.293661	15.847397	9.119759	29.253034	42.855105	28.261989	14.510686	36.914089	40.025617	24.538134	51.469311
CF	12.349089	0.000000	17.863930	12.621807	5.170106	11.290704	9.118114	5.840377	23.650370	38.980764	19.934392	4.666905	33.531329	33.289187	22.596017	51.923309
GB	28.387673	17.863930	0.000000	30.110297	20.795432	7.800000	17.608521	21.915976	20.062403	33.025748	8.826098	14.071958	30.215062	20.101990	25.774794	53.195395
MD	8.362416	12.621807	30.110297	0.000000	10.889444	22.815127	15.334601	8.515868	29.724905	43.860347	30.329359	16.422241	37.463582	42.857205	24.650152	50.716861
ING	7.763376	5.170106	20.795432	10.889444	0.000000	13.814123	10.258167	4.835287	24.373141	39.038827	21.528353	6.790434	33.449963	34.140592	21.898173	50.872586
ARC	21.293661	11.290704	7.800000	22.815127	13.814123	0.000000	10.014989	14.417004	16.534207	31.296326	8.987769	7.275988	27.114019	22.537746	19.836834	48.757461
AGR	15.847397	9.118114	17.608521	15.334601	10.258167	10.014989	0.000000	7.601973	15.292482	30.526382	15.707323	8.367795	24.755403	28.131299	13.787676	43.084916
ES	9.119759	5.840377	21.915976	8.515868	4.835287	14.417004	7.601973	0.000000	22.582515	37.399599	21.852917	8.440972	31.420376	34.563854	19.327442	47.895720
PS	29.253034	23.650370	20.062403	29.724905	24.373141	16.534207	15.292482	22.582515	0.000000	15.574980	12.303658	20.820663	10.754069	17.525125	9.240130	33.143476
GIU	42.855105	38.980764	33.025748	43.860347	39.038827	31.296326	30.526382	37.399599	15.574980	0.000000	24.493673	35.925478	6.797794	19.478450	19.543797	24.721044
LET	28.261989	19.934392	8.826098	30.329359	21.528353	8.987769	15.707323	21.852917	12.303658	24.493673	0.000000	15.718142	21.797936	13.615065	19.009997	45.074938
LING	14.510686	4.666905	14.071958	16.422241	6.790434	7.275988	8.367795	8.440972	20.820663	35.925478	15.718142	0.000000	30.914236	28.928360	20.996190	50.622228
INS	36.914089	33.531329	30.215062	37.463582	33.449963	27.114019	24.755403	31.420376	10.754069	6.797794	21.797936	30.914236	0.000000	20.871512	12.956466	23.722563
PSI	40.025617	33.289187	20.101990	42.857205	34.140592	22.537746	28.131299	34.563854	17.525125	19.478450	13.615065	28.928360	20.871512	0.000000	25.537228	43.710639
EF	24.538134	22.596017	25.774794	24.650152	21.898173	19.836834	13.787676	19.327442	9.240130	19.543797	19.009997	20.996190	12.956466	25.537228	0.000000	29.627521
DS	51.469311	51.923309	53.195395	50.716861	50.872586	48.757461	43.084916	47.895720	33.143476	24.721044	45.074938	50.622228	23.722563	43.710639	29.627521	0.000000

Notiamo come sulla diagonale principale i termini sono posti tutti a zero per la proprietà (i). Mentre i termini simmetrici sono uguali a due a due, per questo motivo è sufficiente considerare solo la triangolare superiore o inferiore. Per semplificare infatti, è stata considerata solo la matrice triangolare al di sotto della diagonale principale.



	SC	CF	GB	MD	ING	ARC	AGR	ES	PS	GIU	LET	LING	INS	PSI	EF	DS
SC	0.000000															
CF	12.349089	0.000000														
GB	28.387673	17.863930	0.000000													
MD	8.362416	12.621807	30.110297	0.000000												
ING	7.763376	5.170106	20.795432	10.889444	0.000000											
ARC	21.293661	11.290704	7.800000	22.815127	13.814123	0.000000										
AGR	15.847397	9.118114	17.608521	15.334601	10.258167	10.014589	0.000000									
ES	9.119759	5.840377	21.915976	8.515868	4.835287	14.417004	7.601973	0.000000								
PS	29.253034	23.650370	20.062403	29.724905	24.373141	16.534207	15.292482	22.582515	0.000000							
GIU	42.855105	38.980764	33.025748	43.860347	39.038827	31.296326	30.526382	37.399599	15.574980	0.000000						
LET	28.261989	19.934392	8.826098	30.329359	21.528353	8.987769	15.707323	21.852917	12.303658	24.493673	0.000000					
LING	14.510686	4.666905	14.071858	16.422241	6.790434	7.275988	8.367795	8.440972	20.820663	35.925478	15.718142	0.000000				
INS	36.914089	33.531329	30.215062	37.463582	33.449963	27.114019	24.755403	31.420376	10.754069	6.797784	21.797936	30.914236	0.000000			
PSI	40.025617	33.289187	20.101990	42.857205	34.140592	22.537746	28.131299	34.563854	17.525125	19.478450	13.615065	26.928360	20.871512	0.000000		
EF	24.538134	22.596017	25.774794	24.650152	21.898173	19.836834	13.787676	19.327442	9.240130	19.543797	19.008997	20.996190	12.956466	25.537228	0.000000	
DS	51.469311	51.923309	53.195395	50.716861	50.872586	48.757461	43.084916	47.895720	33.143476	24.721044	45.074938	50.622228	23.722563	43.710639	29.627521	0.000000

La matrice delle distanze è stata calcolata usando la metrica euclidea, questo perché la metrica euclidea può essere utilizzata con tutti i tipi di metodi.

5.5.1 METRICA EUCLIDEA
È così definita:

$$d_2\left(X_i,X_J\right)=\left[\sum_{k=1}^{p}\left(x_{ik}-x_{Jk}\right)^2\right]^{1/2}$$

Dove x_{ik} è il valore della k–esima caratteristica dell'individuo I_i.
Se si considerano due caratteristiche, ossia p = 2, l'espressione corrisponde alla radice quadrata della somma dei quadrati costruiti sui cateti di un triangolo rettangolo e per il teorema di Pitagora tale radice fornisce la misura dell'ipotenusa del triangolo stesso.
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.
Il metodo più utilizzato per ovviare a questo inconveniente è quello di scalare e standardizzare inizialmente le misure in maniera tale rendere possibile un confronto tra le misure.
Vengono quindi considerate delle nuove variabili:

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{s_j} \quad (i=1,2,\ldots,n;\ j=1,2,\ldots,p)$$

dove $\bar{x}_j$ e s_j sono rispettivamente la media campionaria e la deviazione standard campionaria della j–esima caratteristica.
Mediante lo scalamento e la standardizzazione si ottengono dei nuovi dati le cui medie campionarie sono nulle e le varianze campionarie unitarie.

In R per scalare e standardizzare le variabili si utilizza la funzione

```
> scale(X, center = TRUE, scale = TRUE)
```

dove

- X rappresenta una matrice numerica o un data frame;
- center è posta uguale a TRUE se dagli elementi di ogni colonna della matrice X si sottrae il valore medio della corrispondente colonna (di default è TRUE);
- scale è posta uguale a TRUE se si dividono gli elementi centrati di ogni colonna della matrice X per la deviazione standard della corrispondente colonna (di default è TRUE).

Nel nostro caso non vi è stata necessità di applicare la funzione *scale* per scalare e standardizzare i dati in quanto tutte le colonne presentano i valori in termini di percentuali.

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.
Solitamente gli algoritmi di raggruppamento si differenziamo tra:

- **metodi gerarchici** che conducono ad un insieme di gruppi ordinabili secondo livelli crescenti, con un numero di gruppi da n ad 1;
- **metodi non gerarchici** che forniscono un'unica partizione delle n unità in g gruppi, e g deve essere specificato a priori.

5.2 METODI NON GERARCHICI

L'obiettivo dei metodi non gerarchici è quello di ottenere un'unica partizione degli n individui di partenza in cluster.
A differenza dei metodi gerarchici, in tali tecniche è consentito riallocare gli individui già classificati ad un livello precedente dell'analisi.
In letteratura esistono moltissime tecniche non gerarchiche ed è quindi impossibile ricondurre tali metodi ad un unico tipo, come invece avviene per molti metodi gerarchici di tipo agglomerativo.
In molti metodi non gerarchici di clustering si assume che il numero di cluster in cui suddividere l'insieme totale degli n individui sia fissato a priori dal ricercatore, mentre in altri tale numero è determinato nel corso dell'analisi.
Inoltre, molte di queste tecniche richiedono inizialmente la determinazione di un insieme iniziale di punti di riferimento (ad esempio un insieme iniziale di punti attorno ai quali si addensano i cluster) oppure l'individuazione di una partizione iniziale dei n individui in cluster.

Gli algoritmi di tipo non gerarchico procedono, data una prima partizione, a riallocare gli individui nel gruppo con centroide più vicino, fino a che per nessun individuo si verifica che sia minima la distanza rispetto al centroide di un gruppo diverso da quello a cui esso appartiene.
Il metodo più utilizzato prende il nome di **k-means**.
Tale metodo richiede che il numero di cluster sia specificato a priori 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 i punti di riferimento iniziali (scegliendo in maniera opportuna alcuni individui, o unità, o prendendo la configurazione determinata con una tecnica gerarchica) 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 (il centroide) di ognuno dei k gruppi così ottenuti. Tali centroidi costituiscono i punti di riferimento per i nuovi cluster;
- **Step 4**: Valutare la distanza di ogni unità 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 così ottenuti.
- **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, per garantire la convergenza della procedura iterativa, come misura di distanza tra i vettori delle caratteristiche e i centroidi viene utilizzata la distanza euclidea e, come per il metodo del centroide, si considera la matrice contenente i quadrati delle distanze euclidee.
L'analisi con il metodo k-means si effettua in R mediante la funzione

```
> kmeans(X, centers, iter.max = N, nstart = M)
```

dove

- X è la matrice dei dati;
- centers è il numero dei cluster che si vogliono identificare o un vettore di lunghezza pari al numero di cluster contenente un insieme di centroidi iniziali dei cluster. Nel primo caso, ossia se è numero intero, l'algoritmo sceglie casualmente i punti di riferimento e tale insieme è utilizzato per individuare la partizione iniziale. Nel secondo caso, i centroidi iniziali possono essere derivati effettuando preliminarmente un'analisi di tipo gerarchico con il metodo del centroide.
- iter.max è il massimo numero di iterazioni permesse. Di default iter.max = 10.
- nstart fornisce il numero di volte in cui ripetere la procedura di scelta casuale dei punti di riferimento, nel caso in cui centers è il numero. Di default nstart = 1. Se nstart > 1, l'algoritmo fornisce sempre come risultato la partizione con una misura di non omogeneità statistica totale all'interno dei cluster minima.

Si nota che nell'algoritmo k-means non occorre calcolare la matrice iniziale delle distanze (o dei quadrati delle distanze) così come invece si richiede nei metodi gerarchici.

La funzione kmeans() produce come output una lista, i cui elementi sono:
- un vettore di interi che indica il cluster di allocazione di ogni individuo ($cluster)
- una matrice che contiene i centroidi dei cluster ($center)
- un vettore contenente le misure di non omogeneità statistica calcolate all'interno di ognuno dei cluster; tali valori dipendono dall'omogeneità interna e dalla numerosità del gruppo ($withinss)
- dimensione dei gruppi ($size)

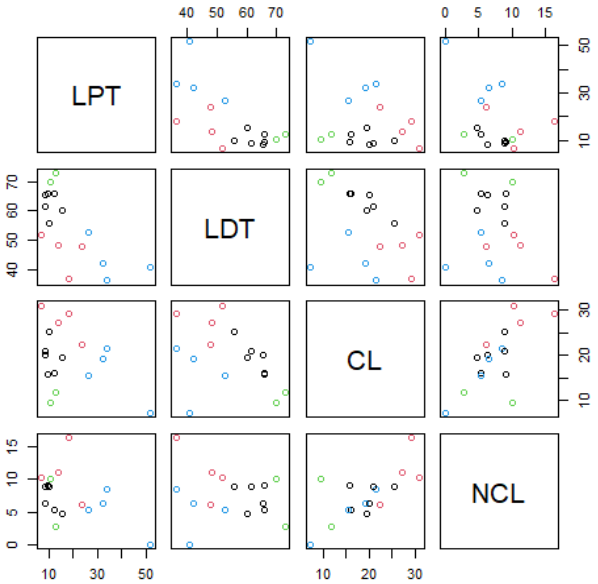
La misura di non omogeneità statistica complessiva all'interno dei vari cluster (within) è quindi la somma delle misure di non omogeneità statistica di ognuno dei cluster.

Gli output di kmeans() vengono mostrati di seguito:

```
> km<-kmeans(occupazione, center=4, iter.max=20, nstart=1)
> km
K-means clustering with 4 clusters of sizes 2, 6, 4, 4

Cluster means:
     LPT      LDT     CL       NCL
1 11.700 71.35000 10.600  6.400000
2 10.700 62.41667 19.700  7.183333
3 15.525 46.02500 27.525 10.950000
4 36.175 42.92500 15.875  5.025000

Clustering vector:
  SC   CF   GB   MD  ING  ARC  AGR   ES   PS  GIU  LET LING  INS  PSI   EF
   1    2    3    1    2    2    2    2    3    4    3    2    4    3    4
  DS
   4

Within cluster sum of squares by cluster:
[1]  34.9650 203.1567 382.0925 661.9300
 (between_SS / total_SS =  76.0 %)


> km<-kmeans(occupazione, center=4, iter.max=20, nstart=10)
> km
K-means clustering with 4 clusters of sizes 4, 7, 4, 1

Cluster means:
       LPT      LDT       CL       NCL
1 29.17500 44.62500 19.67500  6.550000
2 11.08571 65.91429 16.28571  6.728571
3 12.07500 48.07500 28.25000 11.625000
4 51.80000 40.80000  7.30000  0.000000

Clustering vector:
  SC   CF   GB   MD  ING  ARC  AGR   ES   PS  GIU  LET LING  INS  PSI   EF
   2    2    3    2    2    3    2    2    1    1    3    2    1    3    1
  DS
   4

Within cluster sum of squares by cluster:
[1] 259.9125 317.4800 329.2325   0.0000
 (between_SS / total_SS =  83.0 %)
```

Sulla matrice dei dati usando 4 cluster, con un massimo di iterazioni permesse pari a 20 e n_start=10, l'algoritmo mostra una misura di non omogeneità tra i cluster diviso la misura di non omogeneità totale pari a 83.0%.
Ponendo n_start=10 il numero di partizioni rispecchia i dendrogrammi che verranno mostrati successivamente e, inoltre, la misura di non omogeneità tra i cluster diviso la misura di non omogeneità totale risulta essere maggiore rispetto a porre n_start=1 (76.0%).

Per rappresentare graficamente i cluster generati mediante il metodo kmeans possiamo utilizzare le funzioni plot() e points() che producono il grafico seguente:



5.3 METODI GERARCHICI

I metodi di clustering gerarchico possono essere di due tipi: agglomerativi e divisivi.
I metodi gerarchici di tipo **agglomerativo** partono da una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo per giungere, attraverso successive unioni dei cluster meno distanti tra loro, ad una situazione in cui si ha un solo cluster che contiene tutti gli n individui.
Invece, i metodi gerarchici di tipo **divisivo** partono da una situazione in cui si ha un solo cluster che contiene tutti gli n individui per giungere, attraverso successive divisioni dei cluster più distanti tra loro, ad una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo.
Quindi, i metodi gerarchici di tipo agglomerativo procedono con una sequenza di successive unioni degli n individui iniziali in gruppi, mentre i metodi gerarchici di tipo divisivo procedono con una sequenza di successive divisioni dell'insieme degli n individui in partizioni sempre più fini.

L'obiettivo finale dei metodi gerarchici non è quello di ottenere una singola partizione degli n individui di partenza, ma di ottenere una sequenza di partizioni che possono essere rappresentate graficamente mediante una struttura ad albero, detta **dendrogramma**, nella quale sull'asse delle ordinate sono riportati i livelli di distanza mentre sull'asse delle ascisse sono riportati i singoli individui.
Ad ogni livello di distanza corrisponde una partizione, mentre ad ogni partizione corrispondono infiniti livelli di distanza compresi tra quelli che individuano due successive unioni o divisioni.
Il dendrogramma fornisce un quadro completo della struttura dell'insieme in termini delle misure di distanza tra gli individui.
Fissando un opportuno livello della funzione distanza e analizzando tale struttura, il ricercatore può indirettamente stabilire a quale stadio dell'analisi gerarchica occorre fermarsi ottenendo la partizione dell'insieme totale di individui in cluster.

Occorre osservare che è possibile effettuare a priori l'opportuna scelta del livello della funzione distanza se l'analisi gerarchica è effettuata per finalità comparative o se si hanno precedenti conoscenze del fenomeno in osservazione. In generale, comunque, tale scelta può anche essere effettuata a posteriori con il grosso vantaggio di avere una visione globale della struttura dell'insieme totale di individui in termini di distanza.

L'analisi gerarchica di tipo agglomerativo viene effettuata in R attraverso la funzione

```
> hclust(d,method = "complete")
```

dove

- d rappresenta un oggetto (che individua una struttura di similarità o distanza) creato tramite la funzione dist();
- method seleziona il metodo gerarchico agglomerativo (di default è complete).

Alcune delle opzioni disponibili per method sono:

(1) metodo del legame singolo (single);
(2) metodo del legame completo (complete);

(3) metodo del legame medio (average);
(4) metodo del centroide (centroid);
(5) metodo della mediana (median).

La funzione hclust() produce come output una lista, i cui elementi sono:
- la sequenza del processo di agglomerazione ($merge);
- un vettore, la cui lunghezza corrisponde al numero di iterazioni, che indica il livello di distanza alla quale è avvenuta l'unione tra due cluster ($height);
- un'opportuna permutazione delle unità (individui) finalizzata alla costruzione del dendrogramma ($order);
- un vettore delle etichette che contrassegnano le varie unit`a ($labels).

Per ottenere il dendrogramma si impiega la funzione:

> plot(z, labels = NULL, hang = −1,main = "Dendrogramma", sub = NULL, xlab = NULL)

dove
- z è l'oggetto creato (output) dalla funzione hclust();
- labels è un vettore di etichette per i rami del dendrogramma (di default impiega i nomi delle righe del data frame);
- hang determina l'altezza alla quale le etichette vengono visualizzate al di sotto del dendrogramma (un valore negativo pone le etichette al di sotto dell'ordinata nulla);
- main, sub, xlab sono comandi per la finestra grafica.

5.3.1 ANALISI DEL DENDROGRAMMA

Ci proponiamo di analizzare il dendrogramma ottenuto con un particolare metodo gerarchico e di calcolare, fissato il numero di cluster, le misure di non omogeneità della partizione individuata.

5.3.1.1 DISEGNARE RETTANGOLI CHE EVIDENZIANO I CLUSTER

Consideriamo un particolare dendrogramma ottenuto a partire dalla funzione hclust.
La funzione rect.hclust() permette di disegnare dei rettangoli intorno ai cluster, individuati in base all'altezza h alla quale si opera il taglio del dendrogramma oppure in base al numero k di cluster che si vogliono ottenere attraverso la funzione

> rect.hclust(z, h = NULL, k = NULL, border = "color")

dove
- z è l'oggetto creato (output) dalla funzione hclust;
- h è l'altezza alla quale si inserisce il taglio;
- k è il numero di cluster che si vogliono ottenere;
- border è il colore dei contorni dei rettangoli.

5.3.1.2 INSERIRE GLI INDIVIDUI NEI CLUSTER

Considerato un particolare dendrogramma, per ottenere una suddivisione degli individui in cluster in corrispondenza di un determinato livello di distanza oppure in corrispondenza di un prefissato numero di cluster, R utilizza anche la funzione cutree() nel seguente modo:

> cutree(tree, k = NULL, h = NULL)

dove:
- tree rappresenta un oggetto (che individua un dendrogramma) creato tramite la funzione hclust();
- k è il numero prefissato di cluster;
- h è l'altezza alla quale il dendrogramma viene tagliato.

L'output della funzione cutree() è un vettore contenente numeri interi positivi associati ai cluster in cui sono stati inseriti i vari individui.

5.3.1.3 MISURE DI SINTESI ASSOCIATE AI CLUSTER

In R è inoltre possibile ricavare misure di sintesi (ad esempio, la media campionaria, la varianza campionaria, la deviazione standard, …) sulle colonne dei singoli cluster, ottenuti tagliando il dendrogramma tramite la funzione cutree(), utilizzando la funzione aggregate() nel seguente modo:

> aggregate(X, by, FUN)

dove:
- X rappresenta una matrice numerica o un data frame;
- by è una lista di indici sulla base dei quali le colonne di X vanno aggregate;

- FUN è la funzione da applicare alle colonne di X, separatamente per i vari gruppi individuati in base a by.

L'output della funzione aggregate() è una struttura contenente i valori ottenuti applicando la funzione FUN (ad esempio, la media campionaria, la varianza campionaria, la deviazione standard,...) ad ognuna delle caratteristiche associate ai diversi cluster che sono stati aggregati.

Occorre ricordare che per il calcolo della varianza campionaria e della deviazione standard campionaria occorre che nel cluster siano presenti almeno due unità (individui).

5.3.2.3 MISURE DI NON OMOGENEITA' STATISTICHE

Dopo aver effettuato il taglio, siamo interessati a calcolare le misure di non omogeneità statistica relative all'insieme totale di individui (tr T), ai singoli cluster ottenuti effettuando il taglio e alla somma delle loro misure di non omogeneità (tr S) e alla misura di omogeneità tra i cluster (tr B):

$$\text{tr } T = \text{tr } S + \text{tr } B$$

o equivalentemente:

$$1 = \frac{\text{tr } S}{\text{tr } T} + \frac{\text{tr } B}{\text{tr } T}$$

Poiché per ogni fissata matrice X dei dati si ha che la tr T è fissata, i cluster dovrebbero essere individuati in modo da minimizzare la misura di non omogeneità statistica all'interno dei cluster (within) e massimizzare la misura di non omogeneità statistica tra i gruppi (between).

Se, fissato il numero di cluster, due differenti metodi gerarchici conducono a due diverse partizioni, occorre scegliere quella partizione con misura di non omogeneità statistica all'interno dei cluster (tr S) più piccola, che corrisponde a maggiore omogeneità interna.

Per calcolare la misura di non omogeneità statistica totale si usano le seguenti linee di codice in R:

```
> n<-nrow(X)
> trHI<-(n-1)*sum(apply(X, 2, var))
```

```
> n<-nrow(occupazione)
> trHI<-(n-1)*sum(apply(occupazione,2,var))
> trHI
[1] 5335.472
```

Si ha che il valore di trHI = 5335.472.

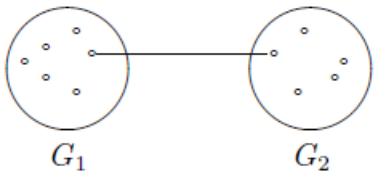
Il prossimo passo, quindi, consiste nel calcolare le misure di non omogeneità statistiche nel caso di quattro gruppi. Tramite R, utilizzando le linee di codice che seguono

```
> agvar<-aggregate(X,tagliolist,var)[,-1]
> trHG1<-(num[[1]]-1)*sum(agvar [1,])
> trHG2<-(num[[2]]-1)*sum(agvar [2,])
> trHG3<-(num[[3]]-1)*sum(agvar [3,])
> trHG4<-(num[[4]]-1)*sum(agvar [4,])
```

5.3.2 APPLICAZIONE METODI

Metodo del legame singolo

In questo metodo la distanza tra i gruppi G_1 (contenente n_1 individui) e G_2 (contenente n_2 individui) è definita come la minima tra tutte le n_1n_2 distanze che si possono calcolare tra ogni individuo di G_1 e ogni individuo di G_2, come mostrato in figura.



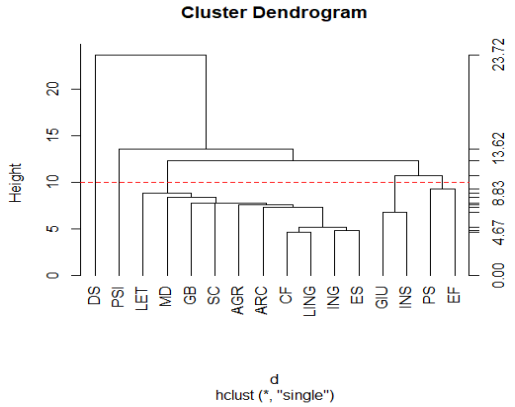
La distanza dell'individuo I_k dal cluster G_{ij} si ottiene scegliendo la più piccola tra le due distanze d_{ik} e d_{jk}. Quindi, al livello 1 si costruisce una nuova matrice D_1 di cardinalità (n-1)*(n-1) costituita da G_{ij}, considerato come un unico elemento, e dagli n-2 individui esterni a G_{ij}. Ad ogni passo successivo, dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

$$d_{(uv),z} = \min (d_{uz}, d_{vz})$$

La quantità $d_{(uv),z}$ rappresenta la misura di distanza tra gli elementi meno distanti dei cluster G_{uv} e G_z.
La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui.
Se volessimo utilizzare R per ottenere il dendrogramma relativo all'uso del metodo del legame singolo, bisogna usare la funzione *hclust* con la variabile method posta a '*single*':

hls <-hclust (d,method =" single").

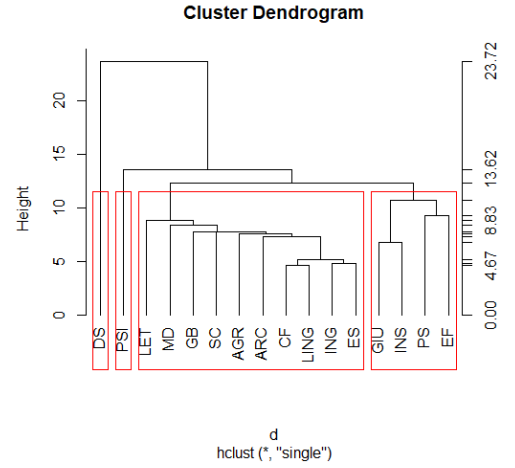
Se in base all'output di tale funzione costruiamo in R il dendogramma, otteniamo:



Tra i metodi considerati questo risulta essere il peggiore poiché spesso porta alla creazione di effetti catena. Tramite tale grafico possiamo infatti notare come si sia verificato anche nel nostro caso.

Evidenziando quattro partizioni mediante rettangoli colorati in rosso, per fare ciò useremo la seguente linea di codice:

> rect.hclust (single, k = 4, border = "red")



Se si considera una suddivisione in k=4 cluster, allora: il primo gruppo fa riferimento a DS; il secondo gruppofa riferimento PSI; il terzo gruppo fa riferimento a LET, MD, GB, SC, AGR, ARC, CF, LING, ING, ES; il quarto gruppo fa riferimento a GIU, INS PS, EF.
DS e PSI sono gruppi contenenti un unico elemento, così come riscontrato nei boxplot tramite anomalie.

Relativamente alle misure di sintesi otteniamo:

```
> taglioHLS<-cutree(hls, k=4, h=NULL)
> tagliolistHLS<-list(taglioHLS)
> aggregate(occupazione, tagliolistHLS, mean)
  Group.1    LPT    LDT     CL   NCL
1       1 10.790 61.720 19.770  7.73
2       2 29.175 44.625 19.675  6.55
3       3 18.000 36.500 29.300 16.30
4       4 51.800 40.800  7.300  0.00
> aggregate(occupazione, tagliolistHLS, var)
  Group.1       LPT      LDT       CL      NCL
1       1  7.585444 61.90178 46.246778 7.582333
2       2 23.242500 51.74917  9.909167 1.736667
3       3        NA       NA        NA       NA
4       4        NA       NA        NA       NA
> aggregate(occupazione, tagliolistHLS, sd)
  Group.1      LPT      LDT       CL      NCL
1       1 2.754169 7.867768 6.800498 2.753604
2       2 4.821048 7.193689 3.147883 1.317826
3       3       NA       NA       NA       NA
4       4       NA       NA       NA       NA

> num<-table(taglioHLS)
> tagliolistHLS<-list(taglioHLS)
> agvar<-aggregate(occupazione, tagliolistHLS, var)[, -1]
> trH1hls<-(num[[1]]-1)*sum(agvar[1,])
> trH1hls
[1] 1109.847
> trH2hls<-(num[[2]]-1)*sum(agvar[2,])
> trH2hls
[1] 259.9125
> trH3hls<-(num[[3]]-1)*sum(agvar[3,])
> trH3hls
[1] NA
> trH4hls<-(num[[4]]-1)*sum(agvar[4,])
> trH4hls
[1] NA
```

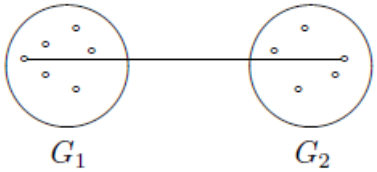
Nel nostro caso, poiché vi sono due gruppi contenente una singola unità (DS e PSI), tutti i valori per questi elementi sono posti ad NA (come mostrato in figura).

Per k=4, è risultato che il valore del primo gruppo trH1hls=1109.847, del secondo gruppo trH2hls=259.9125, del terzo gruppo trH3hls=0, del quarto gruppo trH4hls=0 (nelle operazioni successive per semplicità trH3hls e trH4hls non sono stati riportati in quanto ininfluenti)
Il **valore di within** è pari alla somma tra trH1hls, trH2hls, ovvero trH1hls+trH2hls = 1369.759;
Il **valore di between** è pari alla sottrazione tra trHI e la somma di trH1hls, trH2hls ovvero: trHI − (trHG1+trHG2) = 5335.472 - 1369.759 = 3965.713.

Concludendo, dunque:

between/trHI ≃ 0.74

2. Metodo del legame completo
In questo metodo la distanza tra i gruppi G_1 (contenente n_1 individui) e G_2 (contenente n_2 individui) è definita come la massima tra tutte le n_1n_2 distanze che si possono calcolare tra ogni individuo di G_1 e ogni individuo di G_2, come mostrato in figura.



La distanza dell'individuo I_k dal cluster G_{ij} si ottiene scegliendo la più piccola tra le due distanze d_{ik} e d_{jk}. Quindi, al livello 1 si costruisce una nuova matrice D_1 di cardinalità (n-1)*(n-1) costituita da G_{ij}, considerato come un unico elemento, e dagli n-2 individui esterni a G_{ij}.
Ad ogni passo successivo, dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

$$d_{(uv),z} = \max (d_{uz}, d_{vz})$$

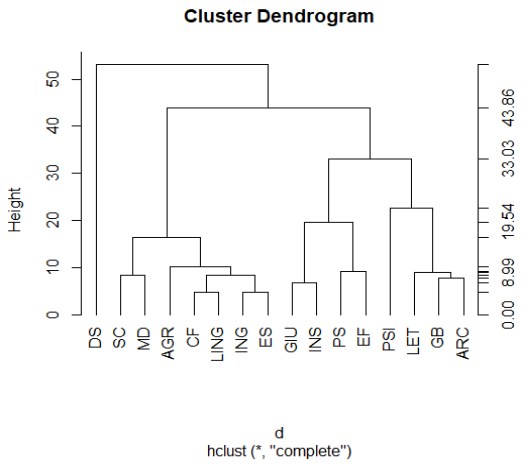
La quantità $d_{(uv),z}$ rappresenta la misura di distanza tra gli elementi meno distanti dei cluster G_{uv} e G_z.
La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui.
Il metodo del legame completo identifica soprattutto gruppi di forma ellissoidale, ossia una serie 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.
Se volessimo utilizzare R per ottenere il dendrogramma relativo all'uso del metodo del legame completo, bisogna usare la funzione *hclust* con la variabile method posta a *'complete'*:

hlc <-hclust (d,method = "complete").

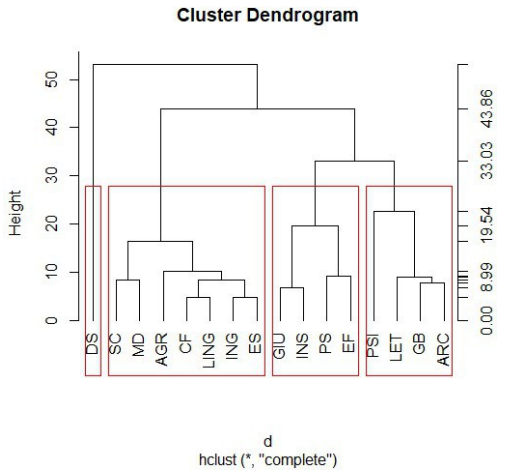
Se in base all'output di tale funzione costruiamo in R il dendogramma, otteniamo:



Possiamo effettuare, anche visivamente, una suddivisione in cluster e per tale motivo il metodo del legame completo risulta essere un buon candidato. Gli altri metodi ritornano, invece, risultati simili tra loro, con cluster più strutturati.

Evidenziando quattro partizioni mediante rettangoli colorati in rosso, per fare ciò useremo la seguente linea di codice:

> rect.hclust (complete, k = 4, border = "red")



Se si considera una suddivisione in k=4 cluster, allora: il primo gruppo fa riferimento a DS; il secondo gruppo fa riferimento a SC, MD, AGR, CF, LING, ING, ES; il terzo gruppo fa riferimento a GIU, INS, PS, EF; il quarto gruppo fa riferimento a PSI, LET, GB, ARC. Il primo gruppo contenente un unico elemento.

Relativamente alle misure di sintesi otteniamo:

```
> taglioHLC<-cutree(hlc, k=4, h=NULL)
> tagliolistHLC<-list(taglioHLC)
> aggregate(occupazione, tagliolistHLC, mean)
  Group.1      LPT      LDT       CL       NCL
1       1 11.08571 65.91429 16.28571  6.728571
2       2 12.07500 48.07500 28.25000 11.625000
3       3 29.17500 44.62500 19.67500  6.550000
4       4 51.80000 40.80000  7.30000  0.000000
> aggregate(occupazione, tagliolistHLC, var)
  Group.1       LPT      LDT        CL       NCL
1       1  7.068095 19.49810 19.328095  7.019048
2       2 23.542500 69.42917  6.150000 10.622500
3       3 23.242500 51.74917  9.909167  1.736667
4       4        NA       NA        NA        NA
> aggregate(occupazione, tagliolistHLC, sd)
  Group.1      LPT      LDT       CL      NCL
1       1 2.658589 4.415665 4.396373 2.649349
2       2 4.852061 8.332417 2.479919 3.259218
3       3 4.821048 7.193689 3.147883 1.317826
4       4       NA       NA       NA       NA

> n
[1] 16
> agvar<-aggregate(occupazione, tagliolistHLC, var)[, -1]
> trH1hlc<-(num[[1]]-1)*sum(agvar[1,])
> trH1hlc
[1] 317.48
> trH2hlc<-(num[[2]]-1)*sum(agvar[2,])
> trH2hlc
[1] 329.2325
> trH3hlc<-(num[[3]]-1)*sum(agvar[3,])
> trH3hlc
[1] 259.9125
> trH4hlc<-(num[[4]]-1)*sum(agvar[4,])
> trH4hlc
[1] NA
```

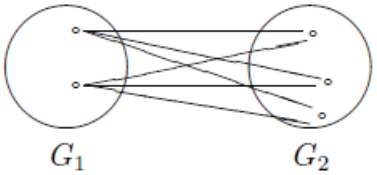
Nel nostro caso, poiché vi è un gruppo contenente una singola unità (DS), tutti i valori per questo elemento sono posti ad NA (come mostrato in figura).

Per k=4, è risultato che il valore del primo gruppo trH1hlc=317.48, del secondo gruppo trH2hlc=329.2325, del terzo gruppo trH3hlc=259.9125, del quarto gruppo trH4hls=0 (nelle operazioni successive per semplicità trH4hls non è stato riportato in quanto ininfluente).
Il **valore di within** è pari alla somma tra trH1hlc, trH2hlc, trH3hlc ovvero 906.625;
Il **valore di between** è pari alla sottrazione tra trHI e la somma di trH1hlc, trH2hlc, trH3hlc = 4428.847.

Concludendo, dunque:

between/trHI ≃0.83

3. Metodo del legame medio
In questo metodo la distanza tra i gruppi G_1 e G_2 è definita come la media aritmetica delle distanze tra tutte le coppie di unità che compongono i due gruppi, così come mostrato in figura.



Nella procedura si considera inizialmente, ossia al livello 0, un insieme di n cluster $\{I_1\}$, $\{I_2\}$, …, $\{I_n\}$. Al passo successivo si cerca nella matrice *D* delle distanze il coefficiente di distanza minima e si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente.
Nel caso i coefficienti di distanza minima siano più di uno, si attua una scelta arbitraria tra di essi. Al livello 1 quindi si modifica la matrice delle distanze valutando le distanze di G_{ij} da ogni altro individuo I_k non appartenente a G_{ij} mediante la seguente relazione

$$d_{(i,j),k}=\frac{1}{2}\left(d_{i,k}+d_{j,k}\right)\ (k=1,2,\ldots,n;k\neq i,j)$$

Quindi, al livello 1 si costruisce una nuova matrice di cardinalità (n-1)*(n-1) costituita da G_{ij}, considerato come un unico elemento, e dagli n − 2 individui esterni a G_{ij}. Ad ogni passo successivo, dopo che i

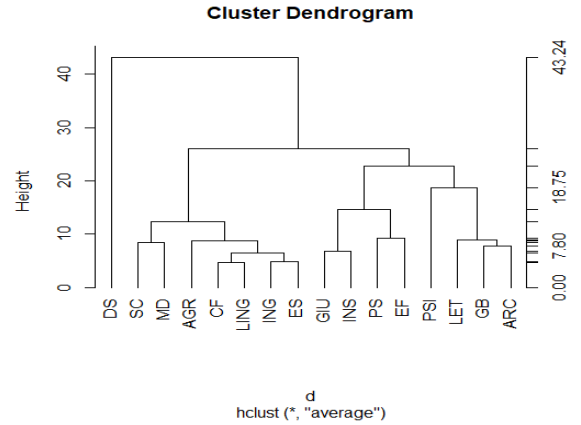
cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

$$d_{(uv),z} = \frac{N_u}{N_u+N_v} d_{uz} + \frac{N_v}{N_u+N_v} d_{vz}$$

dove N_u, N_v e N_z sono rispettivamente il numero di individui del cluster Gu, del cluster G_v e del cluster G_z. La quantità $d_{(uv),z}$ rappresenta la misura di distanza media tra gli elementi dei cluster G_{uv} e G_z.
La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui.
Se volessimo utilizzare R per ottenere il dendrogramma relativo all'uso del metodo del legame medio, bisogna usare la funzione *hclust* con la variabile method posta a '*average*':

hlm <-hclust (d,method = "average").

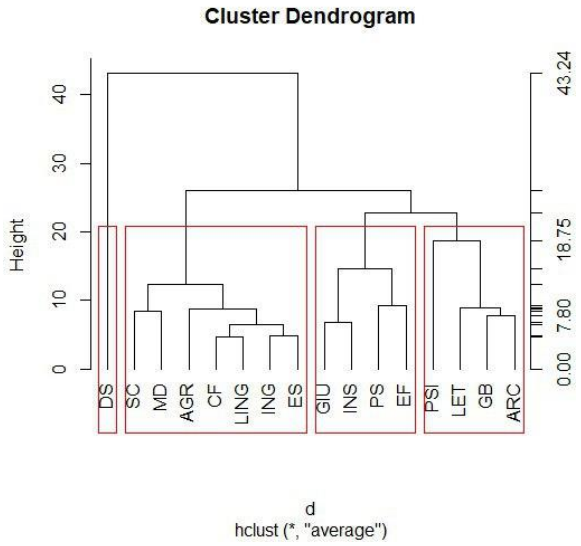
Se in base all'output di tale funzione costruiamo in R il dendogramma, otteniamo:



Osservando questo grafico possiamo banalmente notare che non si verifica un effetto catena.

Evidenziando quattro partizioni mediante rettangoli colorati in rosso, per fare ciò useremo la seguente linea di codice:

> rect.hclust (average, k = 4, border = "red")



Se si considera una suddivisione in k=4 cluster, allora: il primo gruppo fa riferimento a DS; il secondo gruppo sarebbe riferimento SC, MD, AGR, CF, LING, ING, ES; il terzo gruppo fa riferimento a GIU, INS, PS, EF; il quarto gruppo fa riferimento a PSI, LET, GB, ARC. Il primo gruppo contiene un singolo elemento.

Relativamente alle misure di sintesi otteniamo:

```
> taglioHLM<-cutree(hlm, k=4, h=NULL)
> tagliolistHLM<-list(taglioHLM)
> aggregate(occupazione, tagliolistHLM, mean)
  Group.1      LPT      LDT       CL       NCL
1       1 11.08571 65.91429 16.28571  6.728571
2       2 12.07500 48.07500 28.25000 11.625000
3       3 29.17500 44.62500 19.67500  6.550000
4       4 51.80000 40.80000  7.30000  0.000000
> aggregate(occupazione, tagliolistHLM, var)
  Group.1       LPT      LDT        CL       NCL
1       1  7.068095 19.49810 19.328095  7.019048
2       2 23.542500 69.42917  6.150000 10.622500
3       3 23.242500 51.74917  9.909167  1.736667
4       4        NA       NA        NA        NA
> aggregate(occupazione, tagliolistHLM, sd)
  Group.1       LPT      LDT       CL      NCL
1       1 2.658589 4.415665 4.396373 2.649349
2       2 4.852061 8.332417 2.479919 3.259218
3       3 4.821048 7.193689 3.147883 1.317826
4       4       NA       NA       NA       NA
```

```
> agvar<-aggregate(occupazione, tagliolistHLM, var)[, -1]
> trH1hlm<-(num[[1]]-1)*sum(agvar[1,])
> trH1hlm
[1] 317.48
> trH2hlm<-(num[[2]]-1)*sum(agvar[2,])
> trH2hlm
[1] 329.2325
> trH3hlm<-(num[[3]]-1)*sum(agvar[2,])
> trH3hlm<-(num[[3]]-1)*sum(agvar[3,])
> trH3hlm
[1] 259.9125
```

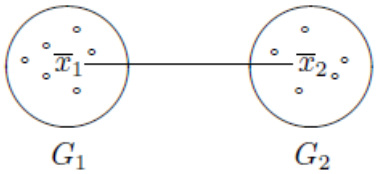
Nel nostro caso, poiché vi è un gruppo contenente una singola unità (DS), tutti i valori per questo elemento sono posti ad NA (come mostrato in figura).

Per k=4, è risultato che il valore del primo gruppo trH1hlm=317.48, del secondo gruppo trH2hlm=329.2325, del terzo gruppo trH3hlm=259.9125, del quarto gruppo trH4hlm=0 (nelle operazioni successive per semplicità trH4hls non è stato riportato in quanto ininfluente).
Il **valore di within** è pari alla somma tra trH1hlm, trH2hlm, trH3hlm ovvero trH1hlm+trH2hlm+trH3hlm = 906.625;
Il **valore di between** è pari alla sottrazione tra trHI e la somma di trH1hls, trH2hls, trH3hls ovvero: trHI − (trH1hlm+trH2hlm+trH3hlm) = 4428.847.

Concludendo, dunque:

between/trHI ≃ 0.83

4. Metodo del legame centroide
In questo metodo la distanza tra il gruppo G_1 e il gruppo G_2 è definita come la distanza tra i centroidi, ossia tra le medie campionarie calcolate sugli individui appartenenti ai due gruppi, così come mostrato in figura.



Nella procedura si considera inizialmente, ossia al livello 0, un insieme di n cluster $\{I_1\}$, $\{I_2\}$, …, $\{I_n\}$. Al passo successivo si cerca nella matrice $D^{(2)}$, contenente i quadrati delle singole distanze euclidee, il coefficiente di distanza minima e si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente. Nel caso i coefficienti di distanza minima siano più di uno, si attua una scelta arbitraria tra di essi. Al livello 1 quindi si modifica la matrice deli quadrati delle distanze valutando i quadrati delle distanze di G_{ij} da ogni altro individuo I_k non appartenente a G_{ij} mediante la relazione:

$$d^2_{(ij),k} = \sum_{r=1}^{p} \left(\overline{x}_{(i,j),r} - \overline{x}_{k,r}\right)^2 = \frac{1}{2}\left(d^2_{ik} + d^2_{jk}\right) - \frac{1}{4} d^2_{ij}, \ (k \neq i,j)$$

dove

$$\bar{x}_{(i,J),r}=\frac{1}{2}\left(x_{i,r}+x_{j,r}\right)\ \bar{x}_{k,r}=x_{k,r}\ (r=1,2,\ldots,p)$$

Quindi, al livello 1 si modifica la matrice delle misure nel seguente modo:

$$X_1=\begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_{i,j} \\ \vdots \\ I_n \end{array}\begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \ldots & C_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_{(i,j),1} & \overline{x}_{(i,j),2} & \ldots & \overline{x}_{(i,j),p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \end{array}$$

ottenendo una matrice di cardinalità (n-1)*p. Ad ogni passo successivo, dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice dei quadrati delle distanze euclidee i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

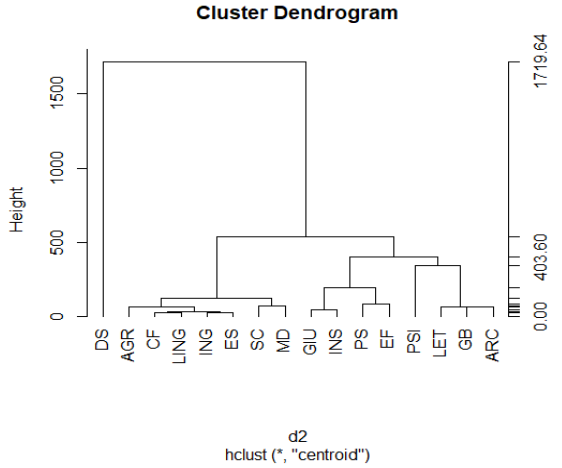
$$d^2_{(u,v),z}=\sum_{u=1}\left(\bar{x}_{(u,v),k}-\bar{x}_{(z)k}\right)^2=\frac{N_u}{N_u+N_v}d^2_{uz}+\frac{N_v}{N_u+N_v}d^2_{vz}-\frac{N_uN_v}{N_u+N_v}d^2_{u,v}$$

dove

$$\bar{x}_{(z),r}=\frac{1}{N_z}\sum_{k:I_k\in G_z}x_{kz}$$

e dove N_u, N_v e N_z denotano rispettivamente il numero di individui del cluster G_u, G_v e G_z.
La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui.
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.

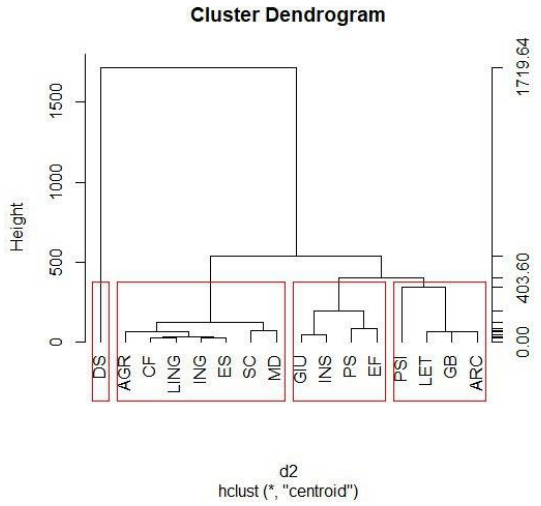
Costruiamo in R il dendogramma, otteniamo:



Osservando questo grafico possiamo banalmente notare come non si verifichi un effetto catena.

Evidenziando quattro partizioni mediante rettangoli colorati in rosso, per fare ciò useremo la seguente linea di codice:

```
> rect.hclust (centroid, k = 4, border = "red")
```



Se si considera una suddivisione in k=4 cluster, allora: il primo gruppo fa riferimento a DS; il secondo gruppo sarebbe riferimento AGR, CF, LING, ING, ES, SC, MD; il terzo gruppo fa riferimento a GIU, INS, PS, EF; il quarto gruppo fa riferimento a PSI, LET, GB, ARC. Il primo gruppo contiene un singolo elemento.
Relativamente alle misure di sintesi otteniamo:

```
> taglioHC<-cutree(hc, k=4, h=NULL)
> tagliolistHC<-list(taglioHC)
> aggregate(occupazione, tagliolistHC, mean)
  Group.1      LPT      LDT       CL       NCL
1       1 11.08571 65.91429 16.28571  6.728571
2       2 12.07500 48.07500 28.25000 11.625000
3       3 29.17500 44.62500 19.67500  6.550000
4       4 51.80000 40.80000  7.30000  0.000000
> aggregate(occupazione, tagliolistHC, var)
  Group.1       LPT      LDT        CL       NCL
1       1  7.068095 19.49810 19.328095  7.019048
2       2 23.542500 69.42917  6.150000 10.622500
3       3 23.242500 51.74917  9.909167  1.736667
4       4        NA       NA        NA        NA
> aggregate(occupazione, tagliolistHC, sd)
  Group.1      LPT      LDT       CL      NCL
1       1 2.658589 4.415665 4.396373 2.649349
2       2 4.852061 8.332417 2.479919 3.259218
3       3 4.821048 7.193689 3.147883 1.317826
4       4       NA       NA       NA       NA

> agvar<-aggregate(occupazione, tagliolistHC, var)[, -1]
> trH1hc<-(num[[1]]-1)*sum(agvar[1,])
> trH1hc
[1] 317.48
> trH2hc<-(num[[2]]-1)*sum(agvar[2,])
> trH2hc
[1] 329.2325
> trH3hc<-(num[[3]]-1)*sum(agvar[3,])
> trH3hc
[1] 259.9125
> trH4hc<-(num[[4]]-1)*sum(agvar[4,])
> trH4hc
[1] NA
```

Nel nostro caso, poiché vi è un gruppo contenente una singola unità (DS), tutti i valori per questo elemento sono posti ad NA (come mostrato in figura).

Per k=4, è risultato che il valore del primo gruppo trH1hlc=317.48, del secondo gruppo trH2hlc=329.2325, del terzo gruppo trH3hlc=259.9125, del quarto gruppo trH4hlc=0 (nelle operazioni successive per semplicità trH4hls non è stato riportato in quanto ininfluente).
Il **valore di within** è pari alla somma tra trH1hlc, trH2hlc, trH3hlc ovvero trH1hlc+trH2hlc+trH3hlc = 906.625;
Il **valore di between** è pari alla sottrazione tra trHI e la somma di trH1hlc, trH2hlc, trH3hlc ovvero: trHI − (trH1hlc+trH2hlc+trH3hlc) = 4428.847.

Concludendo, dunque:

between/trHI ≃ 0.83

5. Metodo della mediana

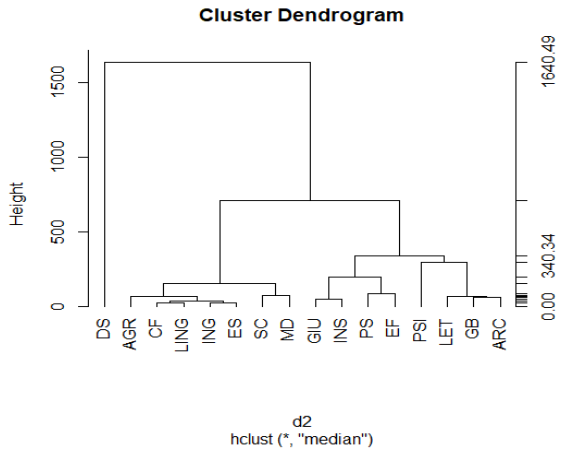
Il metodo della mediana è simile a quello del centroide, con la differenza che la procedura è indipendente dalla numerosità dei cluster. Infatti, quando due gruppi si aggregano, il nuovo centroide è calcolato come la semisomma dei due centroidi precedenti. Il livello 1 della procedura è lo stesso del metodo del centroide. Ad ogni passo successivo, dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice dei quadrati delle distanze euclidee i due cluster più vicini, la distanza tra il nuovo cluster, denotato con Guv, e un altro cluster Gz è così definita:

$$d^2_{(uv),z} = \sum_k^p \left(\bar{x}_{(u,v),k} - \bar{x}_{(z),k}\right)^2 = \frac{1}{2} d^2_{u,z} + \frac{1}{2} d^2_{v,z} - \frac{1}{4} d^2_{u,v}$$

dove

$$\bar{x}_{(u,v),r} = \frac{1}{2} \bar{x}_{(u),r} - \bar{x}_{(v),r} \quad (r=1,2,\ldots,p)$$

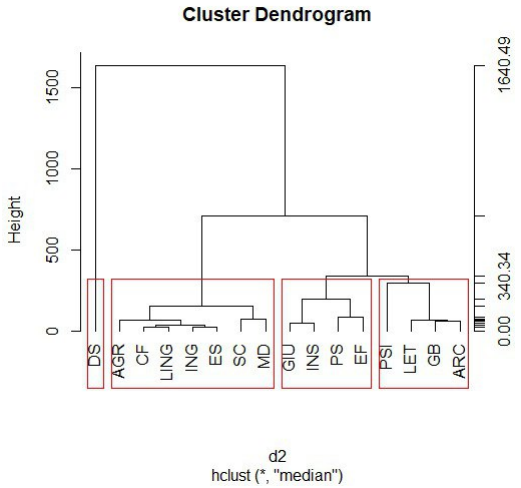
La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui.
Costruiamo in R il dendogramma, otteniamo:



Osservando questo grafico possiamo banalmente notare come non si verifichi un effetto catena.

Evidenziando quattro partizioni mediante rettangoli colorati in rosso, per fare ciò useremo la seguente linea di codice:

> rect.hclust (median, k = 4, border = "red")



Se si considera una suddivisione in k=4 cluster, allora: il primo gruppo fa riferimento a DS; il secondo gruppo sarebbe riferimento AGR, CF, LING, ING, ES, SC, MD; il terzo gruppo fa riferimento a GIU, INS, PS, EF; il quarto gruppo fa riferimento a PSI, LET, GB, ARC.
Il primo gruppo contiene un singolo elemento.

Relativamente alle misure di sintesi otteniamo:

```
> taglioHMED<-cutree(hmed, k=4, h=NULL)
> tagliolistHMED<-list(taglioHMED)
> aggregate(occupazione, tagliolistHMED, mean)
  Group.1      LPT      LDT       CL       NCL
1       1 11.08571 65.91429 16.28571  6.728571
2       2 12.07500 48.07500 28.25000 11.625000
3       3 29.17500 44.62500 19.67500  6.550000
4       4 51.80000 40.80000  7.30000  0.000000
> aggregate(occupazione, tagliolistHMED, var)
  Group.1       LPT      LDT       CL       NCL
1       1  7.068095 19.49810 19.328095  7.019048
2       2 23.542500 69.42917  6.150000 10.622500
3       3 23.242500 51.74917  9.909167  1.736667
4       4        NA       NA        NA        NA
> aggregate(occupazione, tagliolistHMED, sd)
  Group.1      LPT      LDT       CL      NCL
1       1 2.658589 4.415665 4.396373 2.649349
2       2 4.852061 8.332417 2.479919 3.259218
3       3 4.821048 7.193689 3.147883 1.317826
4       4       NA       NA       NA       NA
> agvar<-aggregate(occupazione, tagliolistHMED, var)[, -1]
> trH1hmed<-(num[[1]]-1)*sum(agvar[1,])
> trH1hmed
[1] 317.48
> trH2hmed<-(num[[2]]-1)*sum(agvar[2,])
> trH2hmed
[1] 329.2325
> trH3hmed<-(num[[3]]-1)*sum(agvar[3,])
> trH3hmed
[1] 259.9125
> trH4hmed<-(num[[4]]-1)*sum(agvar[4,])
> trH4hmed
[1] NA
> sum(trH1hmed, trH2hmed, trH3hmed)
[1] 906.625
```

Nel nostro caso, poiché vi è un gruppo contenente una singola unità (DS), tutti i valori per questo elemento sono posti ad NA (come mostrato in figura).

Per k=4, è risultato che il valore del primo gruppo trH1hmed=317.48, del secondo gruppo trH2hmed=329.2325, del terzo gruppo trH3hmed=259.9125, del quarto gruppo trH4hmed=0 (nelle operazioni successive per semplicità trH4hmed non è stato riportato in quanto ininfluente).
Il **valore di within** è pari alla somma tra trH1hmed, trH2hmed, trH3hmed ovvero trH1hmed+trH2hmed+trH3hmed = 906.625;
Il **valore di between** è pari alla sottrazione tra trHI e la somma di trH1hmed, trH2hmed, trH3hmed ovvero: trHI − (trH1hmed+trH2hmed+trH3hmed) = 4428.847.

Concludendo, dunque:

between/trHI ≃ 0.83

5.3.3 CONCLUSIONI

La scelta del metodo gerarchico agglomerativo dipende dagli scopi che il ricercatore si propone poiché ogni metodo definisce un diverso concetto di omogeneità all'interno dei cluster.
Non esiste un metodo migliore, ma ogni metodo ha i suoi vantaggi e i suoi svantaggi.
Se non si ha nessuna informazione sulla struttura dell'insieme da investigare e soprattutto se non si conosce la forma dei cluster da individuare, è sempre interessante applicare il metodo del legame singolo e il metodo del legame completo.
Occorre sottolineare che il metodo del legame singolo è in grado di individuare cluster di qualsiasi forma ma può dare origine alla formazione di una catena.
Con il metodo del legame completo i cluster sono sicuramente ben separati ma l'algoritmo privilegia l'omogeneità tra gli elementi interni ai vari gruppi.

Le tecniche di tipo gerarchico sono sicuramente appropriate per dati numerici di tipo biologico o zoologico per i quali si può ragionevolmente assumere che esista una struttura gerarchica.
Tali tecniche comunque trovano applicazione anche in numerosi altri campi scientifici ed hanno il notevole vantaggio rispetto alle tecniche di enumerazione completa di richiedere minore tempo

computazionale, permettendo così il loro utilizzo anche in presenza di un numero considerevole di dati numerici.
Occorre infine sottolineare che i metodi gerarchici hanno due vantaggi:

- fornire una visione completa dell'insieme in termini di distanze, seppure condizionata dalla scelta del metodo scelto;
- non comportare la scelta a priori del numero di cluster oppure la scelta a priori dei parametri per la determinazione automatica del loro numero.

Terminato un qualsiasi algoritmo gerarchico si possono selezionare il numero di cluster che il ricercatore ritiene più adeguato al problema oggetto di studio.

Relativamente al nostro caso di studio, successivamente viene mostrata la tabella contenente il rapporto between/total per ogni metodo analizzato.

metodo	between/total
metodo del legame singolo	0.74
metodo del legame medio	0.83
metodo del legame completo	0.83
metodo del legame centroide	0.83
metodo della mediana	0.83

Come è possibile notare, il metodo che ha prodotto un risultatato peggiore è il metodo del legame singolo; invece, a parità di cluster (k=4) tutti gli altri metodi hanno prodotto lo stesso risultato.

SECONDA PARTE:
Statistica e Analisi dei Dati

CAPITOLO 6: VARIABILI ALEATORIE

Una variabile aleatoria è e un numero che viene assegnato, mediante una determinata regola, a ciascun punto dello spazio campione, ovvero a ciascuno degli esiti possibili di un esperimento aleatorio.
Il termine "aleatorio" allude al fatto che ci occupiamo degli esiti possibili di un esperimento aleatorio, ovvero, di un esperimento il cui esito è incerto prima che dell'esecuzione dell'esperimento stesso.
Le variabili aleatorie possono essere:

- *discrete:* si dice discreta se può assumere un numero finito, o al più infinito numerabile, di valori;
- *continue*: si dice continua se può assumere tutti gli infiniti valori dell'asse reale R, oppure di un suo intervallo [a,b].

Le variabili aleatorie sono DISCRETE se producono risposte numeriche che derivano da un processo di conteggio. Ad es. "Il numero dei componenti la famiglia", "il numero delle stanze di un'abitazione", ecc.
Le variabili aleatorie sono CONTINUE se generano risposte che derivano da un processo di misurazione. Ad es. "l'altezza", "il reddito", "il fatturato", ecc.
Il sistema R mette a disposizione per ciascuna delle principali variabili aleatorie discrete:

- la funzione di probabilità;
- la funzione di distribuzione;
- la funzione per calcolare i quantili;
- la funzione che simula la variabile aleatoria mediante la generazione di sequenze di numeri pseudocasuali.

Le distribuzioni discrete sono varie, ma noi focalizzeremo l'attenzione sulla distribuzione binomiale.

6.1 DISTRIBUZIONE BINOMIALE

In questa seconda parte, andremo a modellare uno dei problemi analizzabili tramite la distribuzione binomiale.
In particolare, consideriamo il seguente problema: un'azienda produttrice di strumenti elettronici invia una spedizione ad un privato di 50 computer. Durante il trasporto, il veicolo che trasportava la spedizione ha subito un incidente e molte delle confezioni contenenti i computer sono state danneggiate. L'esperimento consiste nel valutare la variabile X (pari al numero di computer che hanno subito dei danni). Ciascuno dei computer può risultare danneggiato con una probabilità del 35%.
L'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche, ed assumiamo che in ogni prova i risultati di interesse siano sintetizzabili nel verificarsi dei seguenti due eventi necessari ed incompatibili: A (interpretabile come successo) e $\bar{A}$ (interpretabile come insuccesso), con P(A) = p (0 < p < 1). Un siffatto esperimento si dice costituito da n prove ripetute indipendenti di Bernoulli.
Sia X la variabile aleatoria che rappresenta il numero di volte in cui si verifica l'evento A nelle n prove.
Una variabile aleatoria X di funzione di probabilità

$$p_X(x)=P(X=x)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x}, & x=0,1,\ldots,n\\ 0, \text{ altrimenti}\end{cases}$$

con 0 < p < 1 e n intero positivo, è detta avere distribuzione binomiale di parametri n e p.
Il termine binomiale deriva dalla contrazione di bi (che significa due) e nom (che significa un numero), riflettendo così il concetto di risultati binari. Con la notazione X ~ B(n, p) intenderemo che X è una variabile aleatoria con distribuzione binomiale di parametri n e p, che chiameremo anche variabile binomiale. Nel caso particolare n = 1, la formula precedente si riduce alla funzione di probabilità di Bernoulli di parametro p. Dalla formula precedente si ricava:

$$\frac{px(r)}{px(r-1)}=\frac{p}{1-p}\frac{n-r+1}{r} \quad r=1,2\ldots n$$

da cui segue che le probabilità binomiali della prima formula sono calcolabili ricorsivamente al seguente modo:

$$px(0)=(1-p)^n\ ,\ px(r)=\frac{p}{1-p}\frac{n-r+1}{r}px(r-1),\ r=1,2,\ldots,n$$

La funzione di distribuzione di X è poi immediatamente ottenibile:

$$F_X(x)=\begin{cases}0, & x<0\\ \sum_{i=0}^{k}\binom{n}{i}P^i(1-p)^{n-i}, & k\le x<k+1 \quad (k=0,1\ldots n-1)\\ 1, & x\ge n\end{cases}$$

Per una variabile aleatoria binomiale si ha $X= X_1 + X_2 + \ldots + X_n$ dove $X_1, X_2 \ldots, X_n$ sono variabili aleatorie di Bernoulli indipendenti ed identicamente distribuite. Pertanto, per una variabile aleatoria binomiale si ottiene:

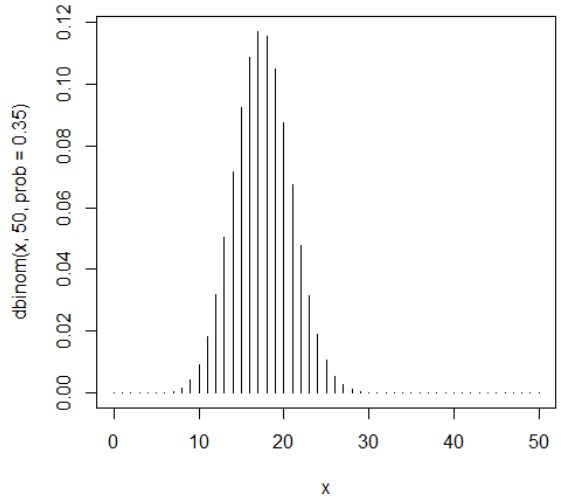
$$E(X)= np,\ Var(X)=np(1-p)$$

Se p = 1/2 la funzione di probabilità binomiale è simmetrica rispetto al suo valore medio E(X) = n/2.

Per il calcolo in R delle probabilità binomiali si utilizza la funzione dbinom (x, size, prob) che calcola P(X = x) dove:

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale;
- size è il numero complessivo delle prove;
- prob è la probabilità di successo in ciascuna prova.

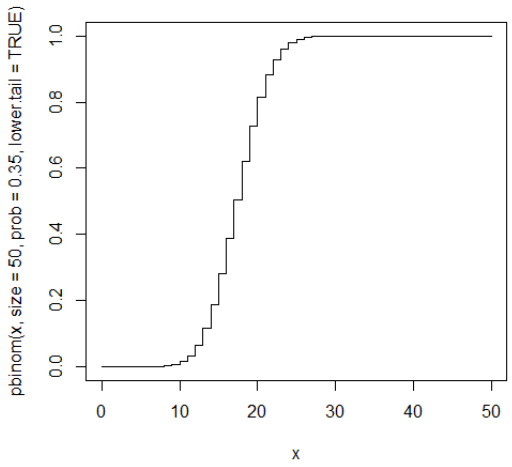
```
> round((dbinom(x, 50, prob=0.35)),2)
 [1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.02 0.03 0.05
[15] 0.07 0.09 0.11 0.12 0.12 0.10 0.09 0.07 0.05 0.03 0.02 0.01 0.01 0.00
[29] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
[43] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```



Per il calcolo della funzione di distribuzione binomiale in R si utilizza la funzione pbinom (x, size , prob, lower.tail = TRUE) dove:

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale;
- size è il numero complessivo delle prove;
- prob è la probabilità di successo in ciascuna prova;
- lower.tail se tale parametro è TRUE (caso di default) calcola $P(X \le x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

```
> round((pbinom(x, size=50, prob=0.35, lower.tail=TRUE)), 2)
 [1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.02 0.03 0.07 0.12
[15] 0.19 0.28 0.39 0.51 0.62 0.73 0.81 0.88 0.93 0.96 0.98 0.99 1.00 1.00
[29] 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00
[43] 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00
```



In R è possibile valutare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione binomiale:

```
> c(Ml, V, sqrt(V), sqrt(V)/Ml)
[1] 17.5000000 11.3750000  3.3726844  0.1927248
```

In R si possono calcolare anche i quantili (percentili) della distribuzione binomiale attraverso la funzione qbinom (z, size, prob) dove:

- z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z·100–esimo;
- size è il numero complessivo delle prove;
- prob è la probabilità di successo in ciascuna prova.

```
> z<-c(0, 0.25, 0.5, 0.75,1)
> qbinom(z, size=50, prob=0.35)
[1]  0 15 17 20 50
```

Mostra che il primo quartile Q1 = 15, la mediana è pari a 17, Q3=20. Il minimo è Q0=0 e il massimo è Q4=50.
Il risultato della funzione è il percentile z ·100–esimo, ossia il più piccolo numero intero k assunto dalla variabile aleatoria binomiale X tale che:

$$F_X(x) = P(X \leq k) \geq z \ (k=0,1,\ldots,n)$$

Per calcolare tale probabilità in R basta utilizzare la funzione pbinom(ne, n, q).

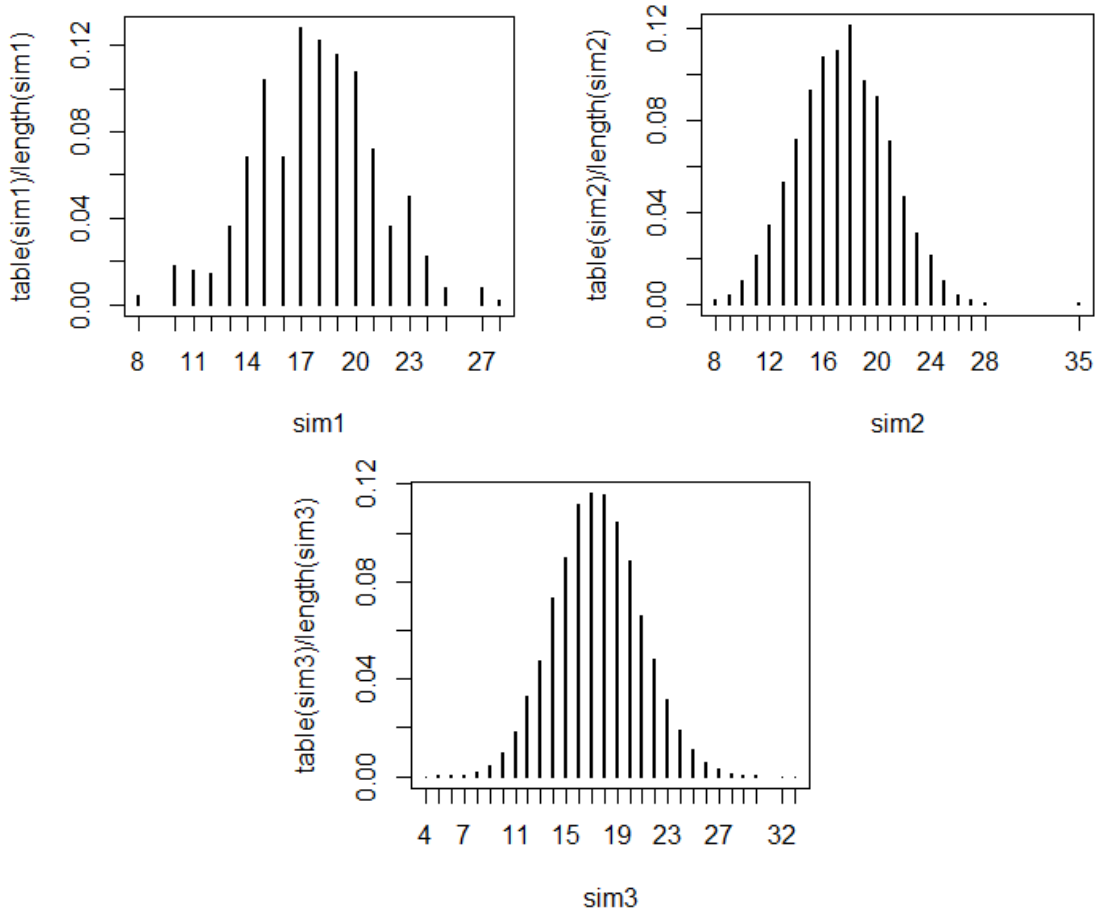
È possibile simulare in R la variabile aleatoria binomiale generando una sequenza di numeri pseudocasuali mediante la funzione rbinom (N, size, prob) dove:

- N è lunghezza della sequenza da generare;
- size è il numero complessivo delle prove;
- prob è la probabilità di successo in ciascuna prova.

Sono state riportate anche le frequenze relative, ottenute dividendo le frequenze generate dalla funzione table() per il numero di volte in cui si è effettuata la simulazione, ovvero 50.

```
> simulazione<-rbinom(50, size=50, prob=0.35)
> simulazione
 [1] 16 15 20 21 15 18 22 16 16 17 22 19 20 14 16 12 18 22 25 13 19 21 19 20
[25] 20 16 16 15 18 13 19 16 23  9 19 14 14 14 21 15 18 14 22 21 16 22 24 16
[49] 13 19
> table(simulazione)
simulazione
 9 12 13 14 15 16 17 18 19 20 21 22 23 24 25
 1  1  3  5  4  9  1  4  6  4  4  5  1  1  1
> table(simulazione)/length(simulazione)
simulazione
   9   12   13   14   15   16   17   18   19   20   21   22   23   24   25
0.02 0.02 0.06 0.10 0.08 0.18 0.02 0.08 0.12 0.08 0.08 0.10 0.02 0.02 0.02
```

Confrontiamo ora la funzione di probabilità binomiale teorica di una variabile binomiale X ~ B(50, 0.35) con quella simulata all'aumentare della lunghezza N= 500, 5000, 50000 della sequenza generata.
Notiamo che all'aumentare della lunghezza N della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità binomiale.



CAPITOLO 7: STIMA PUNTUALE

In statistica uno stimatore (puntuale) è una funzione che associa ad ogni possibile campione un valore del parametro da stimare. Il valore assunto dallo stimatore in corrispondenza a un particolare campione è detto stima. Un problema centrale di inferenza statistica è quello di trovare i valori dei parametri non noti della variabile aleatoria che descrive una certa popolazione. Per fare questo è necessario introdurre la funzione di distribuzione del campione casuale.
Si consideri una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da funzione di distribuzione FX(x). Il vettore aleatorio $X_1, X_2, \dots, Xn$ è detto campione casuale di ampiezza n se le variabili aleatorie del vettore sono osservabili, indipendenti e identicamente distribuite (iid) con la stessa legge di probabilità della popolazione (ossia costituiscono delle osservazioni di X).
La funzione di distribuzione del campione casuale è:

$$F_{x_1,x_{2,\dots}x_n}(x_1,x_{2\dots}x_n)=P\left(X_1\leq x_1,X_2\leq x_2,\ \dots,\ X_n\leq x_n\right)$$
$$=P(X_1\leq x_1)P\left(,X_2\leq x_2\right)\dots P(X_n\leq x_n)=\Pi_{i=1}^{n}F_x(x_i)$$

Il campione casuale può essere estratto da una popolazione illimitata oppure da una popolazione finita; si suppone che l'estrazione avvenga con rimpiazzamento (per garantire l'indipendenza delle variabili aleatorie che costituiscono il campione).
Ogni variabile aleatoria ha uno o più parametri che permettono di determinarne la legge di probabilità. Solo quando tutti i parametri della variabile aleatoria sono già noti la legge di probabilità può dirsi completamente specificata.
Uno stimatore $\widehat{\theta} = t(X_1, X_2, \dots, X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, \dots, X_n$ i cui valori possono essere usati per stimare un parametro non noto ϑ della popolazione. I valori ϑ assunti da tale stimatore sono detti stime del parametro non noto $\bar{\vartheta}$.
Statistiche tipiche sono la media campionaria e la varianza campionaria. Sia $X_1, X_2, \dots, X_n$ un campione casuale.
La statistica

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

è detta **media campionaria**, mentre la statistica

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

è detta **varianza campionaria**.

7.1 METODO DEI MOMENTI

Il metodo dei momenti è uno dei più antichi metodi di stima dei parametri. Occorre preventivamente definire i momenti campionari: si definisce momento campionario r–esimo relativo ai valori osservati $(x_1, x_2, \dots, x_n)$ del campione casuale il valore

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

Il momento campionario r–esimo è la media aritmetica delle potenze r–esime delle n osservazioni effettuate sulla popolazione.
Inoltre, si può affermare che:

- se r = 1 il momento campionario $M_1(x_1, x_2, \dots, x_n)$ coincide con il valore osservato della media campionaria X, ossia $M_1(x_1, x_2, \dots, x_n)/n$.
- 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

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

In particolare, i termini a sinistra di questo sistema di equazioni dipendono dalla legge di probabilità e contengono i parametri non noti della popolazione. I termini a destra invece, possono essere calcolati a partire dai dati osservati del campione estratto dalla popolazione. Le incognite di questo sistema sono ϑ_1, $\vartheta_2, \dots, \vartheta_k$ e sono presenti alla sinistra di questo sistema. Affinché questo metodo sia applicabile occorre che il sistema (a) ammetta un'unica soluzione.

Nel nostro caso p risulta essere l'unico parametro da stimare.

Popolazione Binomiale – Ci proponiamo di determinare con il metodo dei momenti lo stimatore del parametro p di una popolazione binomiale descritta da una variabile aleatoria $X \sim B(k, p)$ con funzione di probabilità

$$p_X(x) = \binom{k}{x} p^x (1-p)^{k-x} \qquad (x=0,1,\dots,k) \qquad (0<p<1)$$

Occorre quindi stimare il parametro p. Poiché $X=Y_1 + Y_2 + \dots + Y_k$, dove $Y_1, Y_2, \dots, Y_k$ sono variabili aleatorie indipendenti di Bernoulli, risulta $E(X)=kp$ e dalla dalla (a) si ha:

$$k\hat{p} = \frac{x_1+x_2+\dots+x_n}{n} \text{ ossia } \hat{p} = \frac{1}{k}\frac{x_1+x_2+\dots+x_n}{n} = \frac{\bar{x}}{k}.$$

Il metodo dei momenti fornisce quindi come stimatore del parametro kp la media campionaria $\bar{X}$.

Sistema R – Utilizzando il sistema R, tali stime vengono effettuate mediante la funzione:

> stima <- mean(valori simulati)/numero di lanci

L'output di tale metodologia risulta essere uguale a 0.3532.

```
> #stima
> stimap<-mean(simulazione)/50
> stimap
[1] 0.3532
```

7.2 METODO DELLA MASSIMA VEROSIMIGLIANZA

Il metodo della massima verosimiglianza è il più importante metodo per la stima dei parametri non noti di una popolazione e solitamente è preferito al metodo dei momenti. Per illustrarlo occorre introdurre in primo luogo la funzione di verosimiglianza.

Sia $X_1, X_2, \dots, X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\vartheta_1, \vartheta_2, \dots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \dots, \vartheta_k; x_1, x_2, \dots, x_n)$ del campione osservato $(x_1, x_2, \dots, x_n)$ è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale $X_1, X_2, \dots, X_n$, ossia

$$L(\vartheta_1, \vartheta_2, \dots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \dots, \vartheta_k; x_1, x_2, \dots, x_n)$$
$$= f(x_1; \vartheta_1, \vartheta_2, \dots, \vartheta_k)\, f(x_2; \vartheta_1, \vartheta_2, \dots, \vartheta_k) \dots f(x_n; \vartheta_1, \vartheta_2, \dots, \vartheta_k).$$

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, \dots, \vartheta_k$.

Tale metodo cerca quindi di determinare da quale funzione di probabilità congiunta (nel caso di popolazione discreta) oppure di densità di probabilità congiunta (nel caso di popolazione assolutamente continua) è più verosimile (è più plausibile) che provenga il campione osservato $(x_1, x_2, \dots, x_n)$.

Pertanto, si cercano di determinare i valori $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ che rendono massima la funzione di verosimiglianza e che quindi offrano, in un certo senso, la migliore spiegazione del campione osservato $(x_1, x_2, \dots, x_n)$.

I valori di $\vartheta_1, \vartheta_2, \dots, \vartheta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $\hat{\vartheta}_1, \hat{\vartheta}_2, \dots, \hat{\vartheta}_k$; essi costituiscono le stime di massima verosimiglianza dei parametri non noti $\vartheta 1, \vartheta 2, \dots, \vartheta k$ della popolazione.

Tali stime dipendono dal campione osservato $(x_1, x_2, \ldots, x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori di massima verosimiglianza $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ della popolazione, detti stimatori di massima verosimiglianza.
Nel nostro caso, come con il metodo dei momenti, il nostro unico parametro da stimare corrisponde a p.
Anche in questo caso, per una popolazione binomiale lo stimatore di massima verosimiglianza per p è la media campionaria $\overline{X}$.
Per tale motivo non occorre effettuare nessun nuovo calcolo in R, in quanto la media campionaria è nota essere 0.3532.

Proprietà degli stimatori – Considerato un certo parametro Θ, esistono diverse funzioni dei dati campionari che possono essere considerate come possibili stimatori del parametro, ma la scelta di una determinata funzione rispetto ad un'altra, porta a stime che sono generalmente diverse fra di loro.
Nei casi reali il parametro in questione è ovviamente ignoto, per cui non c'è nessuna possibilità di quantificare l'errore commesso utilizzando una particolare stima.
Per determinare la bontà di una funzione dei dati campionari rispetto ad altre funzioni diverse ci si basa sulle proprietà degli stimatori.
Si sceglierà quindi lo stimatore con le proprietà migliori, anche se in una particolare occasione di campionamento non si potrà mai sapere se la stima fornita dallo stimatore scelto è effettivamente prossima al valore vero del parametro ignoto.
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:

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

1- Uno stimatore $\widehat{\theta} = t(X_1, X_2, \ldots, X_n)$ del parametro non noto ϑ della popolazione è detto corretto (non distorto) se e solo se per ogni $\vartheta \in \theta$ si ha $E(\widehat{\theta}) = \vartheta$, ossia se il valore medio dello stimatore $\widehat{\theta}$ è uguale al corrispondente parametro non noto della popolazione.
2- Sia $\widehat{\theta} = t(X_1, X_2, \ldots, X_n)$ uno stimatore del parametro non noto ϑ della popolazione. Si chiama errore quadratico medio la quantità $MSE(\widehat{\theta}) = E[(\widehat{\theta} - \vartheta)^2]$. Il principale problema del decisore consiste nello scegliere lo stimatore migliore del parametro ϑ, ossia lo stimatore che ha il più piccolo errore quadratico medio per ogni valore ammissibile di $\vartheta \in \theta$.
Situazioni in cui esiste uno stimatore migliore di tutti gli altri si verificano raramente e spesso sono poco interessanti. La ricerca dello stimatore con errore quadratico uniformemente minimo deve essere quindi effettuata in opportune classi come, ad esempio, nella classe degli stimatori corretti.
3- Uno stimatore $\widehat{\theta}$ si dice corretto con varianza uniformemente minima per il parametro non noto ϑ se e solo se per ogni $\vartheta \in \theta$ risulta:
(i) $E(\widehat{\theta}) = \vartheta$,
(ii) $Var(\widehat{\theta}) \leq Var(\widehat{\theta}\star)$ per ogni altro stimatore $\widehat{\theta}\star$ corretto del parametro ϑ.
La varianza fornisce quindi una misura della dispersione dei valori assunti dallo stimatore intorno al suo valore medio.
4- uno stimatore $\widehat{\theta}$ è detto asintoticamente corretto se, al crescere dell'ampiezza del campione, la precisione del suo valore medio non diminuisce;
5- una condizione sufficiente (ma non necessaria) affinché lo stimatore sia consistente è che sia asintoticamente corretto e la sua varianza tende a zero al crescere del campione.

Considerato lo stimatore analizzato, definito da una variabile aleatoria $X \sim B(k, p)$, le proprietà di cui gode sono: corretto (non distorto) perché è pari proprio al valore medio della popolazione (meglio di così non si può fare) e con varianza minima e consistente per kp .

CAPITOLO 8: INTERVALLI DI CONFIDENZA E FIDUCIA APPROSSIMATI

Alla stima puntuale di un parametro non noto di una popolazione (costituita da un singolo valore reale) 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 ed uno superiore) entro i quali sia compreso il parametro non noto con un certo coefficiente di confidenza (detto anche grado di fiducia).
Sia $X_1, X_2, \dots, X_n$ un campione casuale di ampiezza n estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x; \vartheta)$, dove ϑ denota il parametro non noto della popolazione.
Denotiamo con $\underline{C_n} = g1(X_1, X_2, \dots, X_n)$ e con $\overline{C_n} = g2(X_1, X_2, \dots, X_n)$ due statistiche (funzioni osservabili del campione casuale) che soddisfino la condizione $\underline{C_n} < \overline{C_n}$, cioè che godono della proprietà che per ogni possibile fissato campione osservato x = (x1, x2, . . . , xn) risulti g1(x) < g2(x).

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

$$P(\underline{C_n} < \vartheta < \overline{C_n}) = 1 - \alpha,$$

allora si dice che $(\underline{C_n}, \overline{C_n})$ è un intervallo di confidenza (intervallo di fiducia) di grado $1 - \alpha$ per ϑ.
Inoltre, le statistiche $\underline{C_n}$ e $\overline{C_n}$ sono dette limite inferiore e superiore dell'intervallo di confidenza.
Se g1(x) e g2(x) sono i valori assunti dalle statistiche $\underline{C_n}$ e $\overline{C_n}$ per il campione osservato $x = (x_1, x_2, \dots, x_n)$, allora l'intervallo (g1(x), g2(x)) è detto stima dell'intervallo di confidenza di grado $1 - \alpha$ per ϑ ed i punti finali g1(x) e g2(x) di tale intervallo sono detti rispettivamente stima del limite inferiore e stima del limite superiore dell'intervallo di confidenza.

In generale, esistono numerosi intervalli di confidenza dello stesso grado $1-\alpha$ per un parametro non noto ϑ della popolazione. La scelta dell'intervallo di confidenza deve essere effettuata dal decisore in base ad alcune proprietà statistiche.

Ad esempio, fissato un coefficiente di confidenza $1 - \alpha$, alcune proprietà desiderabili sono che la lunghezza dell'intervallo di confidenza

$$L(X_1, X_2, \dots, X_n; 1 - \alpha) = \overline{C_n} - \underline{C_n}$$

sia la più piccola possibile oppure che la lunghezza media di tale intervallo sia la più piccola possibile.

8.1 METODO PIVOTALE

Uno dei metodi usati molto spesso per effettuare una stima intervallare è il metodo pivotale. Questo consiste nell'ausilio di una variabile aleatoria di Pivot $\gamma(X_1, X_2, \dots, X_n; \vartheta)$ che:

- dipende dal campione casuale $(X_1, X_2, \dots, X_n)$;
- dipende dal parametro non noto ϑ;
- la funzione di distribuzione non contiene il parametro non noto.

Il metodo pivotale è usato su popolazioni normali.
Poiché la dimensione del nostro campione è elevata ossia ≥ 30 (nel nostro caso, infatti, n è pari a 50), 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.
Questo afferma che la seguente variabile aleatoria converge in distribuzione ad una variabile normale standard.

Data una variabile aleatoria X che descrive la popolazione con $E(X) = \mu$ e $Var(X) = \sigma^2$; con il campione casuale $(X_1, X_2, \dots, X_n)$, il teorema della convergenza afferma che la variabile aleatoria

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to Z$$

converge in distribuzione ad una variabile aleatoria normale standard.
Se il valore medio $E(X) = \mu$ e $Var(X) = \sigma^2$ della popolazione dipendono da un parametro non noto ϑ della popolazione, si nota la variabile aleatoria Z_n può essere interpretata come una variabile aleatoria di pivot poiché valgono le seguenti:
• Zn dipende dal campione casuale $X_1, X_2, \dots, X_n$;
• Zn dipende dal parametro non noto ϑ della popolazione attraverso il valore medio $E(X)=\mu$ e la varianza $Var(X) = \sigma^2$;
• per grandi campioni la funzione di distribuzione di Z_n è approssimativamente normale standard e quindi non contiene il parametro non noto ϑ da stimare.

Di fatto quindi la nostra variabile aleatoria binomiale è riconducibile a una variabile aleatoria normale su cui svolgeremo il metodo pivotale. Questo procedimento è chiamato metodo pivotale approssimato. Quindi si determinerà un intervallo di confidenza dato uno specifico grado di fiducia. A questo punto è possibile definire una variabile aleatoria Z_n come se fosse una variabile aleatoria di Pivot che può essere usata per il calcolo dell'intervallo di confidenza. Pertanto, viene applicato *il metodo pivotale in forma approssimata* secondo questa relazione:

$$P\left(-\frac{z_a}{2} < \frac{\overline{X}_m - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{z_a}{2} - \right) \simeq 1-\alpha$$

Per determinare l'intervallo di confidenza per un generico parametro, si cerca una espressione per cui deve comparire solo il parametro da stimare e non altri parametri incogniti e la distribuzione deve essere perfettamente nota.
Una volta individuata questa espressione si può, isolando il parametro, costruire l'intervallo di confidenza. Come detto anche in precedenza, la variabile aleatoria binomiale è tale che:

$$E(x) = kp \text{ e } var(X) = kp(1-p).$$

Entrambi dipendono dal parametro non noto p.

Applicando il teorema centrale di convergenza risulta che

$$\frac{(\overline{X}_n - E(X))}{\sqrt{\frac{var(X)}{n}}} = \sqrt{n}\ \frac{(\overline{X}_n - E(X))}{\sqrt{var(X)}} = \sqrt{n}\ \frac{(\overline{X}_n - kp)}{\sqrt{kp(1-p)}}$$

Converge in distribuzione ad una variabile aleatoria normale standard. Per il campione da noi considerato avente ampiezza 50, è possibile determinare l'intervallo di confidenza richiedendo che:

$$P(-z_{a/2} < \sqrt{n}\ \frac{(\overline{X}_n - kp)}{\sqrt{kp(1-p)}} < z_{a/2}) \cong 1-\alpha$$

Dopo alcune trasformazioni matematiche si ottiene la seguente disequazione di secondo grado in p:

$$k(nk + z_{\frac{\alpha}{2}}^2)p^2 - k(2n\,\overline{X}_n + z_{\frac{\alpha}{2}}^2)p + n\bar{x}^2{}_n < 0$$

Le radici di questa disequazione risultano infine:

$$a_2 = k(nk + {z_{a/2}}^2)$$

$$a_1 = -k(2n\ \text{mean}(X_n) + z_{a/2}{}^2)$$
$$a_0 = n\ \text{mean}(X_n^2)$$

Calcolando i suddetti coefficienti sarà possibile ottenere l'intervallo di confidenza.
L'intervallo di confidenza può essere calcolato in R mediante le linee di codice sottostanti.
Tali calcoli riportano un valore dell'intervallo nel modo che segue (0.3289950, 0.3781821). Il valore di stimap calcolato in precedenza risulta essere pari a circa 0.3532, che è contenuta nell'intervallo calcolato.

Nello specifico:

- α rappresenta il livello di significatività che è pari a 0.01 (1-0.99);
- qnorm(1-α/2,mean=0, sd=1) rappresenta il valore zα (variabile aleatoria di pivot) che risulta essere pari a 2,575829;
- la funzione polyroot() permette di calcolare con le radici α0, α1, α2, l'intervallo di confidenza.

```
> alpha<-1-0.99
> qnorm(1-alpha/2, mean=0, sd=1)
[1] 2.575829
> zalpha<-qnorm(1-alpha/2, mean=0, sd=1)
> a2<-50*(50*50+zalpha^2)
> a1<- -50*(2*50*mean(simulazione)+zalpha^2)
> a0<-50*(mean(simulazione))^2
> polyroot(c(a0,a1,a2))
[1] 0.3289950-0i 0.3781821+0i
```

Notiamo ora come, utilizzando un livello di significatività inferiore, e quindi, scegliendo alpha pari a 0.05 (1-0.95), l'intervallo di confidenza risulta essere più ristretto rispetto al valore calcolato precedentemente.

```
> alpha<-1-0.95
> alpha
[1] 0.05
> qnorm(1-alpha/2, mean=0, sd=1)
[1] 1.959964
> zalpha<-qnorm(1-alpha/2, mean=0, sd=1)
> a2<-50*(50*50+zalpha^2)
> a1<- -50*(2*50*mean(simulazione)+zalpha^2)
> a0<-50*(mean(simulazione))^2
> polyroot(c(a0,a1,a2))
[1] 0.3347024-0i 0.3721481+0i
```

CAPITOLO 9: 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.

L'ipotesi statistica è una affermazione o una congettura riguardante un parametro che caratterizza il modello descrittivo della popolazione.
L'ipotesi soggetta a verifica viene in genere denotata con $\mathbf{H_0}$ e viene 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 $\mathbf{H_1}$.

Si tratta di dividere lo spazio campionario in due sottoinsiemi: zona di accettazione e zona di rifiuto. Se la statistica test cade nella regione di accettazione, l'ipotesi nulla non può essere rifiutata; mentre, se questa cade nella regione di rifiuto, l'ipotesi nulla deve essere rifiutata.
Per prendere una decisione sull'ipotesi nulla dobbiamo definire le regioni di rifiuto e di accettazione e questo viene fatto determinando il **valore critico** della statistica del test. La determinazione di questo valore dipende dall'ampiezza della regione di rifiuto.

Nel seguire questo tipo di ragionamento si può incorrere in **due tipi di errori**:
- I tipo: rifiutare l'ipotesi nulla nel caso in cui tale ipotesi sia vera; si dice allora che si commette un errore di tipo I e si denota la probabilità di commettere tale errore con α.
- II tipo: accettare l'ipotesi nulla nel caso in cui tale ipotesi sia falsa; si dice allora che si commette un errore di tipo II e si denota la probabilità di commettere tale errore con β.

Di norma, il rischio di commettere un errore di I tipo α è sotto il controllo di chi compie l'analisi; mentre, la probabilità di commettere un errore di II tipo dipende dalla differenza tra il valore ipotizzato e il vero valore del parametro della popolazione. Un modo per controllare e ridurre l'errore di seconda specie consiste nell'aumentare la dimensione del campione perché un'elevata dimensione del campione consente di individuare anche piccole differenze tra la statistica campionaria e il parametro della popolazione.

I test statistici si dividono **in due gruppi**:
- test unilaterale, quando la regione di rifiuto è costituita da un intervallo;
- test bilaterale, quando la regione di rifiuto è costituita da due intervalli, ossia da due code della distribuzione.

Per il test bilaterale:

$$H_0: p = p_0$$
$$H_1: p \neq p_0,$$

Mentre, il test unilaterale sinistro e test unilaterale destro sono rispettivamente i seguenti:

$$H_0: p \leq p_0 \qquad H_0: p \geq p_0$$
$$H_1: p > p_0 \qquad H_1: p < p_0,$$

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.

La stima osservata della statistica del test è denotata con z_{os} ed è calcolata nel modo che segue:

$$z_{os} = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} \quad (a)$$

Nel nostro caso, quella che stiamo considerando è la popolazione binomiale, formata da un numero finito di prove di Bernoulli.

Nel caso di una popolazione di Bernoulli, il valore di $\mu_0 = p_0$ e il valore di $\sigma_0^2 = p_0 (1- p_0)$, nei test unilaterali e bilaterali occorre considerare la formula (a) con le opportune sostituzioni:

$$z_{os} = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x}_n - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

La stima osservata della statistica (z_{os}) del test cambia, in quanto dobbiamo considerare un numero pari a k di prove di Bernoulli. Pertanto, il valore di $\mu_0 = p_0 = kp_0$ e il valore di $\sigma_0^2 = p_0 (1 - p_0) = kp_0(1 - p_0)$, nei test unilaterali e bilaterali occorre considerare la formula (a) con le opportune sostituzioni:

$$z_{os} = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x}_n - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{\bar{x}_n - kp_0}{\sqrt{\frac{kp_0(1 - p_0)}{n}}}$$

Il p-value è definito come la probabilità (quindi un numero compreso fra 0 e 1), supposta vera l'ipotesi H_0, che la statistica del test assuma un valore uguale o più estremo di quello effettivamente osservato.
In altri termini, il p-value aiuta a capire se la differenza tra il risultato osservato e quello ipotizzato è dovuta alla casualità introdotta dal campionamento, oppure se tale differenza è statisticamente significativa, cioè difficilmente spiegabile mediante la casualità dovuta al campionamento.

Calcolando il p-value sono possibili **due osservazioni**:
- se $p > \alpha$, l'ipotesi H_0 non può essere rifiutata;
- se $p \leq \alpha$, l'ipotesi H_0 deve essere rifiutata.

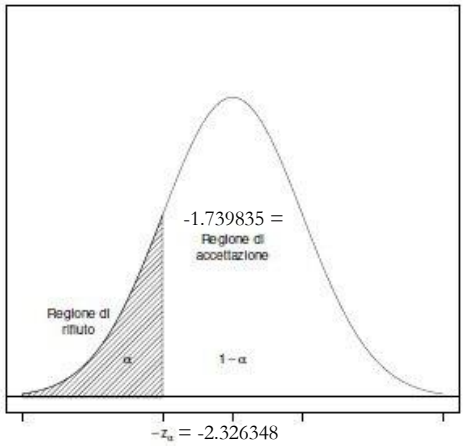
Nel condurre un test statistico è importante fissare **il livello di significatività α** prima di calcolare il p-value. Se si calcola prima il p-value, il decisore potrebbe scegliere il livello di significatività α in funzione del risultato desiderato in modo da accettare o rigettare l'ipotesi nulla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

Precedentemente, abbiamo mostrato che una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ è (0.3289950, 0.3781821).

Vogliamo verificare H_0 con $p_0 \geq 0.37$, in alternativa H_1 con $p_0 < 0.37$, con un livello di significatività $\alpha = 0.01$. Quindi, utilizzando R, si ottiene quello che viene mostrato successivamente.
Nello specifico abbiamo che:
- a p_0 è stato assegnato il valore 0.37;
- ad α è stato assegnato il valore 0.01;
- la funzione qnorm() calcola z_α;
- il valore meanCamp contiene il calcolo della media campionaria;
- il valore (meanCamp-kp_0)/sqrt(kp_0*(1-p_0)/50) rappresenta z_{os}, dove k è posto a 50, in quanto vengono effettuate 50 prove di Bernoulli;

```
> p0<-0.37
> alpha
[1] 0.01
> qnorm(alpha, mean=0, sd=1)
[1] -2.326348
> meanCamp
[1] 17.66
> (meanCamp-(50*p0))/sqrt(((50*p0)*(1-p0))/50)
[1] -1.739835
> qnorm(1-alpha, mean=0, sd=1)
[1] 2.326348
```

Abbiamo che, nel nostro caso:
- $-z_\alpha = -2.326348 \rightarrow z_\alpha = +2.326348$
- $z_{os} = -1.739835$



Il valore di z_{os} ricade all'interno dell'intervallo della **regione di accettazione**, pertanto, occorre **accettare l'ipotesi nulla** con un livello di significatività $\alpha = 0.01$, ovvero del 1%.

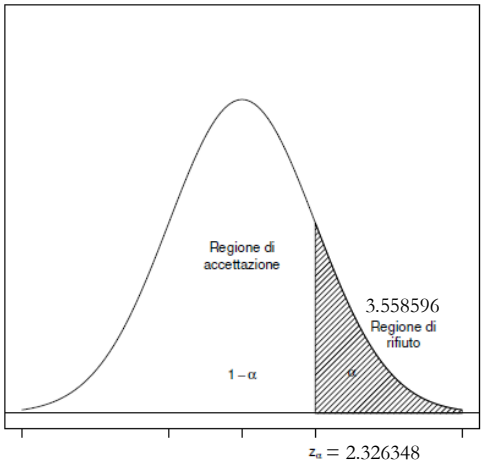
Proviamo ora a cambiare il valore di p_0, ponendolo a 0.32. **Ci proponiamo quindi di verificare H_0 con $p_0 \geq 0.32$, in alternativa H_1 con $p_0 < 0.32$, con un livello di significatività $\alpha = 0.01$.**
Quindi, utilizzando R, si ottiene quello che viene mostrato successivamente.
Nello specifico abbiamo che:

- a p_0 è stato assegnato il valore 0.32;
- ad α è stato assegnato il valore 0.01;
- la funzione qnorm() calcola z_α;
- il valore meanCamp contiene il calcolo della media campionaria;
- il valore (meanCamp-kp_0)/sqrt(kp_0*(1-p_0)/50) rappresenta z_{os}, dove k è posto a 50, in quanto vengono effettuate 50 prove di Bernoulli;

```
> p0<-0.32
> alpha
[1] 0.01
> qnorm(alpha, mean=0, sd=1)
[1] -2.326348
> (meanCamp-(50*p0))/sqrt(((50*p0)*(1-p0))/50)
[1] 3.558596
> p0
[1] 0.32
> qnorm(1-alpha, mean=0, sd=1)
[1] 2.326348
```

Abbiamo che, nel nostro caso:

- $-z_\alpha = -2.326348 \rightarrow z_\alpha = +2.326348$
- $z_{os} = 3.558596$



Il valore di z_{os} ricade all'interno dell'intervallo della **regione di rifiuto**, pertanto, occorre **rifiutare l'ipotesi nulla** con un livello di significatività $\alpha = 0.01$, ovvero del 1%.

CAPITOLO 10: 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 $F_X(x)$, con k parametri non noti da stimare.
A questo scopo, utilizzeremo il criterio di verifica delle ipotesi del chi–quadrato, detto anche test del chi–quadrato o test del buon adattamento.

In altre parole, per criterio del chi-quadrato, si intende uno dei **test di verifica d'ipotesi** usati in statistica che utilizzano la distribuzione chi quadrato per decidere se rifiutare o meno l'ipotesi nulla.

Ricordiamo che, con H_0 intendiamo l'ipotesi nulla; mentre, con H_1 ci riferiamo all'ipotesi alternativa, contrapposta a H_0.
Il test chi–quadrato con livello di significatività α mira a verificare l'ipotesi nulla. α rappresenta la probabilità massima di rifiutare l'ipotesi nulla quando questa è vera.

Occorre determinare **un test ψ 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**, anche detto a due code.
Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi I_1, I_1, ..., I_R (classi o categorie) in modo che risulti essere uguale a $\mathbf{p_i}$ **la probabilità che**, secondo la distribuzione ipotizzata, **la variabile aleatoria assuma un valore appartenente a $\mathbf{I_i}$**.
Si estrae poi un campione di ampiezza n e si osservano le frequenze assolute con cui gli n elementi si distribuiscono nei rispettivi insiemi I_1, I_1, ..., I_R. Quindi n_i rappresenta **il numero degli elementi del campione che cadono nell'intervallo** Ii (i = 1, 2, ..., r).
Il criterio del chi-quadro si basa sulla seguente statistica:

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

dove N_i è la variabile aleatoria che descrive il numero degli elementi del campione casuale che cadono nell'intervallo I_i.
Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta:

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

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

- si **accetti** l'ipotesi nulla H_0 se $\mathbf{X^2_{1-\frac{\alpha}{2},\ r-k-1} < X^2 < X^2_{\frac{\alpha}{2},\ r-k-1}}$
- si **rifiuti** l'ipotesi nulla H_0 se $\mathbf{X^2 < X^2_{1-\frac{\alpha}{2},\ r-k-1}}$ oppure $\mathbf{X^2 > X^2_{\frac{\alpha}{2},\ r-k-1}}$

dove $X^2_{1-\frac{\alpha}{2},\ r-k-1}$ e $X^2_{\frac{\alpha}{2},\ r-k-1}$ sono le soluzioni delle equazioni:

$$P(Q < X^2_{1-\frac{\alpha}{2},r-k-1}) = \frac{\alpha}{2}, \qquad P\left(Q < X^2_{\frac{\alpha}{2},\ r-k-1}\right) = 1-\frac{\alpha}{2}$$

Considerando la statistica (a) riportata, ciò che è stato fatto è stato suddividere l'insieme dei valori che la variabile aleatoria X può assumere in r=4 intervalli ottenuti tramite l'utilizzo dei quartili:

- I1 = (0, 15];
- I2 = (15, 18];
- I3 = (18, 20];
- I4 = (20, 50].

```
> simulazione
 [1] 16 15 20 21 15 18 22 16 16 17 22 19 20 14 16 12 18 22 25 13 19 21 19 20
[25] 20 16 16 15 18 13 19 16 23  9 19 14 14 14 21 15 18 14 22 21 16 22 24 16
[49] 13 19
> qbinom(0.25, 50, stimap)
[1] 15
> qbinom(0.5, 50, stimap)
[1] 18
> qbinom(0.75, 50, stimap)
[1] 20
> p<-numeric(4)
> p[1]<-pbinom(15, 50, stimap)
> p[2]<-dbinom(16, 50, stimap)+dbinom(17,50,stimap)+dbinom(18,50,stimap)
> p[3]<-dbinom(19,50,stimap)+dbinom(20,50,stimap)
> p[4]<- 1-p[1]-p[2]-p[3]
> sum(p)
[1] 1
> min(50*p[1], 50*p[2], 50*p[3], 50*p[4])
[1] 9.874632
> p
[1] 0.2644392 0.3388316 0.1974926 0.1992366
```

Una volta ottenuti gli intervalli, determiniamo il numero di elementi del campione che ricadono negli intervalli I_1, I_2, I_3, I_4; ne segue che:

```
> r<-4
> nint<-numeric(4)
> nint[1]<-length(which(simulazione<=15))
> nint[2]<-length(which((simulazione>15) & (simulazione<=18)))
> nint[3]<-length(which((simulazione>18) & (simulazione<=20)))
> nint[4]<-length(which(simulazione>=21))
> nint
[1] 14 14 10 12
> sum(nint)
[1] 50
```

- $n_1 = 14$;
- $n_2 = 14$;
- $n_3 = 10$;
- $n_4 = 12$.

Dopo di che, viene calcolato il valore di chi-quadro (X^2) che risulta essere pari a 6.65359.

```
> chi2<-sum(((nint-50*0.3532)/sqrt(50*0.3532))^2)
> chi2
[1] 6.65359
```

La distribuzione binomiale ha un solo parametro non noto, ovvero la p; quindi il valore di k è posto pari a 1. Pertanto, la funzione di distribuzione della statistica Q (a) è approssimabile con la funzione di distribuzione chi–quadrato con r - k - 1 pari a 1 grado di libertà. Dunque:

- $k = 1$;
- $r - k - 1 = 1$.

Da queste assunzioni, riassumendo i valori ottenuti in (a), abbiamo che, con $\alpha = 0.01$:

```
> r<-4
> 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
```

- $X^2_{1-\frac{\alpha}{2},\ r-k-1} \simeq 10.59663$
- $X^2 \simeq 6.65359$
- $X^2_{\frac{\alpha}{2},\ r-k-1} \simeq 0.01002508$

Essendo false le seguenti disequazioni, già mostrate nel corso di questo capitolo:

$$X^2_{1-\frac{\alpha}{2},\ r-k-1} > X^2 > X^2_{\frac{\alpha}{2},\ r-k-1} \rightarrow \quad 10.59663 > 6.65359 > 0.01002508$$

si ha che l'ipotesi H_0 può essere accettata, quindi il campione può essere descritto tramite una popolazione binomiale.