PROGETTO DI STATISTICA E ANALISI DEI DATI

Anno accademico 2017/2018

Docente

Prof.ssa A. Nobile

Studente

Sara Volpe

MATR. : 0522500468

Prima parte

La prima parte di questo progetto consiste nell'analizzare i dati forniti da una statistica, quindi applicarne tutti gli argomenti trattati a lezione, tramite il linguaggio R.

L'argomento selezionato tramite il sito ISTAT su cui ho incentrato il mio progetto è:

"**Interruzioni volontarie della gravidanza – Tasso di abortività per classi di età (valori per 1000 donne) – livello regionale**"

Il dataset preso in analisi è costituito da una tabella avente come righe le venti regioni italiane e come colonne le diverse fasce di età in cui vi è una situazione di aborto.

Di seguito è riportata una tabella con i relativi valore divisi per regione.

Età e classe di età	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
Territorio di residenza								
Italia	4,49	9,74	10,69	10,29	8,58	3,91	0,35	6,37
Piemonte	5,87	13,5	13,48	13,51	10,54	4,56	0,36	7,93
Valle d'Aosta	6,45	8,95	14,91	9,59	8,51	4,28	1,27	6,94
Liguria	7,15	15,76	15,73	14,36	11,34	4,76	0,47	8,62
Lombardia	4,65	10,38	11,11	10,22	8,49	3,68	0,27	6,33
Trentino Alto Adige	2,86	7,35	8,46	8,65	6,21	2,94	0,42	4,9
Veneto	2,66	7,43	8,23	7,98	6,62	3,28	0,41	4,8
Friuli-Venezia Giulia	4,85	8,74	8,94	9,82	7,76	3,18	0,28	5,55
Emilia-Romagna	4,79	11,97	13,48	12,59	10,2	4,81	0,44	7,58
Toscana	4,86	11,08	12,6	12,21	9,7	4,17	0,35	7,11
Umbria	4,29	10,95	12,66	9,78	9,26	3,97	0,31	6,75
Marche	3,24	7,52	9,37	7,96	6,73	2,95	0,23	5,02
Lazio	5,87	11,94	12,41	10,89	8,98	4,09	0,37	7,05
Abruzzo	4,81	9,09	9,89	9,78	8,9	4,2	0,44	6,34
Molise	4,62	8,47	8,01	8,34	9,39	4,09	0,65	5,95
Campania	3,36	7,16	8,48	8,86	7,41	3,54	0,26	5,39
Puglia	5,55	11,26	12,28	12,66	11,26	5,16	0,48	8
Basilicata	3,93	7,59	9,52	8,34	7,77	4	0,31	5,68
Calabria	3,02	6,31	7,53	8,01	6,95	3,4	0,31	4,96
Sicilia	4,3	8,1	8,82	8,09	7,05	3,3	0,33	5,49
Sardegna	3,8	7,39	8,06	8,65	6,67	3,08	0,36	5,05

1. TABELLE E GRAFICI

Iniziamo col costruire le tabelle e le *distribuzioni di frequenza* tramite R.

1.1. Distribuzioni di frequenza semplici

Considerando le 20 regioni, è stato costruito un vettore contenente le regioni ed i valori associati a “15-19 anni”.

```
> Quindici_Diciannove_anni<-c(rep("Piemonte",5.87),rep("Valle_d_Aosta",6.45),rep("Liguria",7.15),rep("Lombardia",4.65),rep("Trentino_Alto_Adige",2.86),rep("Veneto",2.66),rep("Friuli_Venezia_Giulia",4.85),rep("Emilia_Romagna",4.79),rep("Toscana",4.86),rep("Umbria",4.29),rep("Marche",3.24),rep("Lazio",5.87),rep("Abruzzo",4.81),rep("Molise",4.62),rep("Campania",3.36),rep("Puglia",5.55),rep("Basilicata",3.93),rep("Calabria",3.02),rep("Sicilia",4.3),rep("Sardegna",3.8))
```

In R la costruzione di una distribuzione di frequenza viene effettuata utilizzando la funzione *table()*. Quindi, per calcolare le frequenze assolute del nostro vettore è bastato eseguire il seguente comando

```
> table(Quindici_Diciannove_anni) #calcola le frequenze assolute
Quindici_Diciannove_anni
              Abruzzo             Basilicata               Calabria               Campania         Emilia_Romagna 
                    4                      3                      3                      3                      4 
Friuli_Venezia_Giulia                  Lazio                Liguria              Lombardia                Marche 
                    4                      5                      7                      4                      3 
              Molise               Piemonte                 Puglia               Sardegna               Sicilia 
                    4                      5                      5                      3                      4 
              Toscana    Trentino_Alto_Adige                 Umbria          Valle_d_Aosta                 Veneto 
                    4                      2                      4                      6                      2 
```

Si noti che *table(Quindici_Diciannove_anni)* ordina le regioni ordine alfabetico.

Per ottenere la distribuzione delle frequenze relative è bastato utilizzare il comando *table(vettore)/length(vettore)*

```
> table(Quindici_Diciannove_anni)/length(Quindici_Diciannove_anni) #calcola le frequenze relative
Quindici_Diciannove_anni
              Abruzzo             Basilicata               Calabria               Campania         Emilia_Romagna 
           0.05063291             0.03797468             0.03797468             0.03797468             0.05063291 
Friuli_Venezia_Giulia                  Lazio                Liguria              Lombardia                Marche 
           0.05063291             0.06329114             0.08860759             0.05063291             0.03797468 
              Molise               Piemonte                 Puglia               Sardegna               Sicilia 
           0.05063291             0.06329114             0.06329114             0.03797468             0.05063291 
              Toscana    Trentino_Alto_Adige                 Umbria          Valle_d_Aosta                 Veneto 
           0.05063291             0.02531646             0.05063291             0.07594937             0.02531646 
```

Per calcolare le frequenze assolute cumulate si deve utilizzare la funzione cumsum().

Per ottenere le frequenze relative cumulate si divide per length() del vettore.

Riferendoci ai dati del nostro progetto le seguenti linee di codice

```
> cumsum(table(Quindici_Diciannove_anni))
              Abruzzo             Basilicata               Calabria               Campania         Emilia_Romagna 
                    4                      7                     10                     13                     17 
Friuli_Venezia_Giulia                  Lazio                Liguria              Lombardia                Marche 
                   21                     26                     33                     37                     40 
              Molise               Piemonte                 Puglia               Sardegna               Sicilia 
                   44                     49                     54                     57                     61 
              Toscana    Trentino_Alto_Adige                 Umbria          Valle_d_Aosta                 Veneto 
                   65                     67                     71                     77                     79 
> cumsum(table(Quindici_Diciannove_anni)/length(Quindici_Diciannove_anni))
              Abruzzo             Basilicata               Calabria               Campania         Emilia_Romagna 
           0.05063291             0.08860759             0.12658228             0.16455696             0.21518987 
Friuli_Venezia_Giulia                  Lazio                Liguria              Lombardia                Marche 
           0.26582278             0.32911392             0.41772152             0.46835443             0.50632911 
              Molise               Piemonte                 Puglia               Sardegna               Sicilia 
           0.55696203             0.62025316             0.68354430             0.72151899             0.77215190 
              Toscana    Trentino_Alto_Adige                 Umbria          Valle_d_Aosta                 Veneto 
           0.82278481             0.84810127             0.89873418             0.97468354             1.00000000 
```

hanno permesso di calcolare le frequenze assolute cumulate e le frequenze relative cumulate.

1.2 Le rappresentazioni grafiche

È stata considerata una *variabile* "età" e 8 modalità distinte da essa assunte rappresentate dai valori in Italia.

È stato rappresentato un grafico disponendo sull'asse orizzontale ed in modo equi spaziato le modalità assunte dalla variabile "età" e sull'asse verticale riportiamo le frequenze assolute. Sono stati tracciati dei rettangoli centrati sulle modalità di "età" tutti della stessa base e altezzapari alle frequenze, ottenendo così un grafico a barre.

In R si ottiene un *grafico a barre* utilizzando barplot(table()). Per colorare i rettangoli con differenti colori basta aggiungere in barplot() il parametro col=1:8, essendo 8 le fasce d'età, e quindi le modalità, considerate.

Le seguenti linee di codice

```
> eta<-c(rep("15-19",4),rep("20-24",9),rep("25-29",10),rep("30-34",10),rep("35-39",8),rep("40-44",3)
,rep("45-49",0),rep("15-49",6))
> barplot(table(eta),col=1:8)
```

hanno prodotto il grafico a barre illustrato in Figura 1.1.

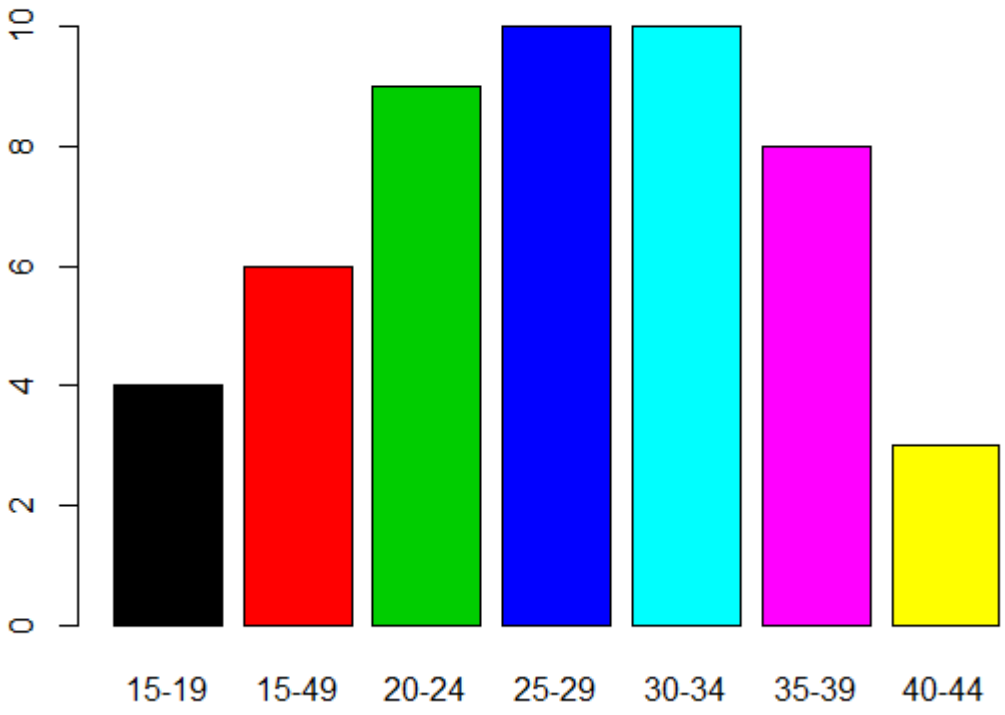


Figura 1.1: Frequenza assoluta della variabile qualitativa "età" tramite un grafico a barre.

Dalla figura 1.1 possiamo dedurre che le fasce d'età in cui si presentano maggiormente situazioni di aborto in Italia sono "25-29" e "30-34", mentre la fascia d'età in cui si presentano di meno risulta essere "40-44". La fascia d'età "45-49" non viene proprio riportata in quanto la sua frequenza assoluta è uguale a 0.

Le modalità della variabile "eta" sono state ordinate nella Figura 1.1 in ordine alfabetico (come possiamo vedere la modalità "15-49"). Affinchè le modalità siano ordinate in modo differente occorre trasformare il vettore "eta" in un fattore "eta1". Le seguenti linee di codice

```
> eta1<-ordered(eta,levels=c("15-19","20-24","25-29","30-34","35-39","40-44","45-49","15-49"))
> barplot(table(eta1), col=1:8)
```

producono il grafico a barre illustrato in Figura 1.2; si nota che le modalità della variabile "eta" sono state ordinate come specificato in *levels*.

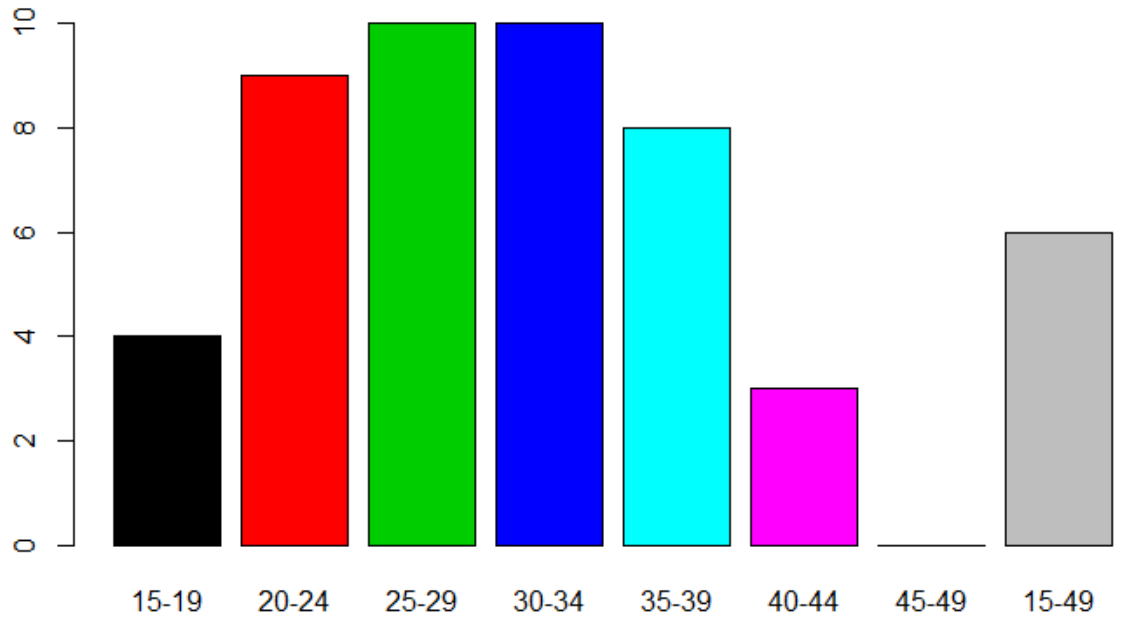


Figura 1.2: Frequenza assoluta della variabile qualitativa "età" tramite un grafico a barre ordinando le modalità

Un altro tipo di rappresentazione si ottiene mediante i *diagrammi a torta* che permettono di attribuire ciascuna modalità della variabile qualitativa in esame ad un settore circolare di un cerchio, la cui ampiezza è proporzionale alle frequenze.
Il sistema R sceglie il tipo di diagramma a torta con i diversi settori colorati differentemente utilizzando il comando pie(table()). Riferendosi sempre all'esempio precedente, il seguente codice

```
> pie(table(eta),col = 1:8)
```

ha prodotto il diagramma a torta mostrato in Figura 1.3.

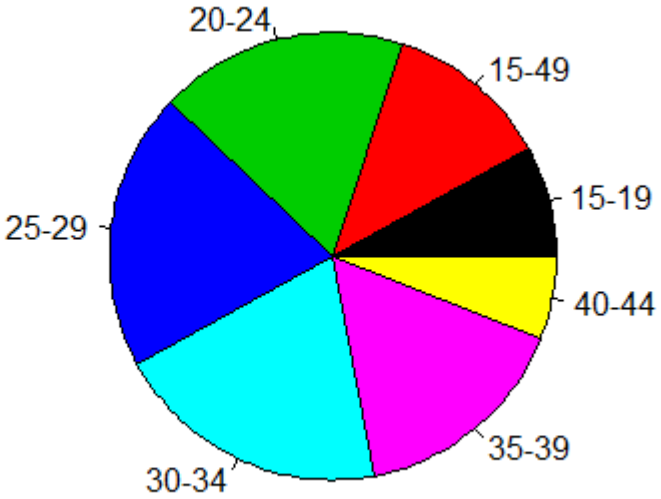


Figura 1.3:Frequenza della variabile qualitativa "età" utilizzando un diagramma a torta.

Anche in questo caso, dalla figura 1.3 possiamo dedurre che le fasce d'età in cui si presentano magg iormente situazioni di aborto in Italia sono "25-29" e "30-34", mentre la fascia d'età in cui si presen tano di meno risulta essere "40-44". In più si può notare che manca la sezione per la modalità "45-4 9" poiché la sua frequenza assoluta è pari a 0.

Si può anche scegliere un *tratteggio particolare* da utilizzare nei diagrammi a torta per tratteggiare differentemente i diversi settori utilizzando i comandi *density= e angle=.* Riferendosi al precedente esempio il seguente codice

```
> pie(table(eta),density = 10,angle = 18+10*(1:8),col = 1:8)
```

ha prodotto il diagramma a torta mostrato in Figura 1.4.

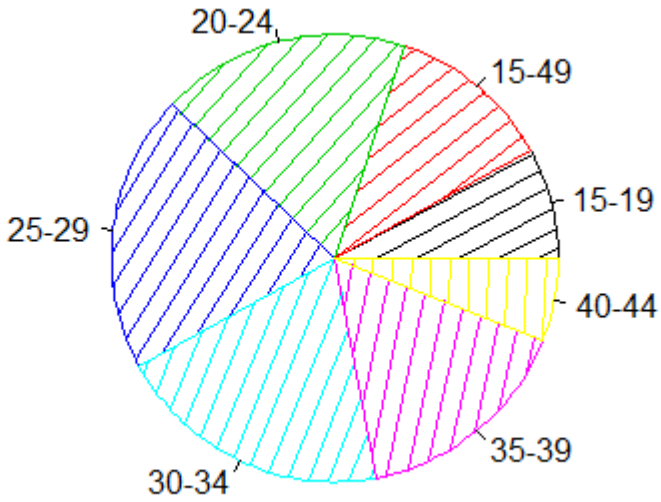


Figura 1.4:Frequenza della variabile qualitativa "età" utilizzando un particolare diagramma a torta.

È stata considerata ora una *variabile quantitativa* "quindici_diciannove_anni" e valori numerici assunti dalle varie regioni.
In questo caso la funzione plot() illustra l'andamento dei valori assunti dalla fascia d'età presa in considerazione rispetto alle relative regioni. Tale comando, quindi, non fornisce informazioni significative circa la distribuzione di frequenza. Ad esempio, il seguente codice

```
> quindici_diciannove_anni<-c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3)
> plot(quindici_diciannove_anni,ylab = "15-19_anni",col="red")
```

ha prodotto il grafico in Figura 1.5, che illustra la percentuale di aborto per ognuna delle 20 regioni individuate dalle loro posizioni nel vettore.

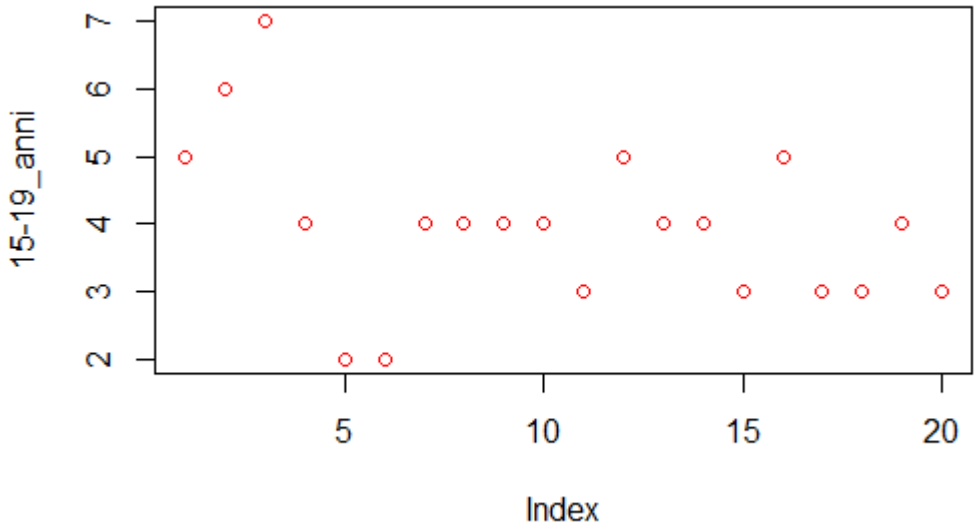


Figura 1.5: Rappresentazione dei valori assunti dal vettore numerico "quindici_diciannove_anni"

Per rappresentare invece correttamente la distribuzione di frequenza della variabile quantitativa "quindici_diciannove_anni" è stato necessario utilizzare il comando plot(table()) che ha prodotto un grafico a bastoncini in cui sull'asse orizzontale sono riportati i valori assunti dalle regioni e sull'asse verticale le frequenze assolute dei valori distinti assunti nel vettore. Riferendosi all'esempio precedente, il comando

```
> plot(table(quindici_diciannove_anni),ylab="15-19_anni",col=1:20)
```

ha prodotto il grafico in Figura 1.6.

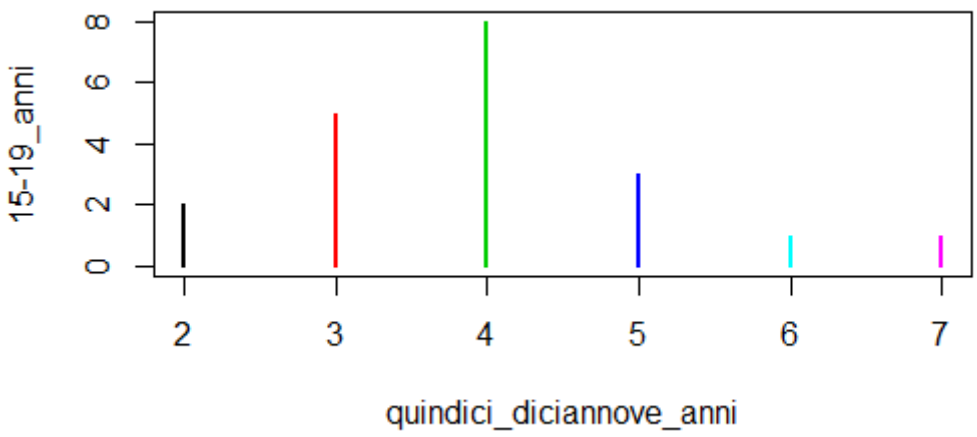


Figura 1.6:Distribuzione di frequenza del vettore numerico "quindici_diciannove_anni"

Dalla figura 1.6 si può notare che il valore numerico 4 assunto dal vettore nelle varie regioni è quello che si presenta con più frequenza.

Sono stati connessi mediante linee i punti della Figura 1.5 creando una *serie storica* della fascia d'età 15-19 anni. A tal fine, è stata utilizzata la funzione plot() con l'opzione *type=l,* che permette di creare delle linee interconnesse. Quindi, il comando

```
> plot(quindici_diciannove_anni,type="l",ylab = "15-19_anni",col="blue")
```

ha prodotto il grafico illustrato in Figura 1.7.

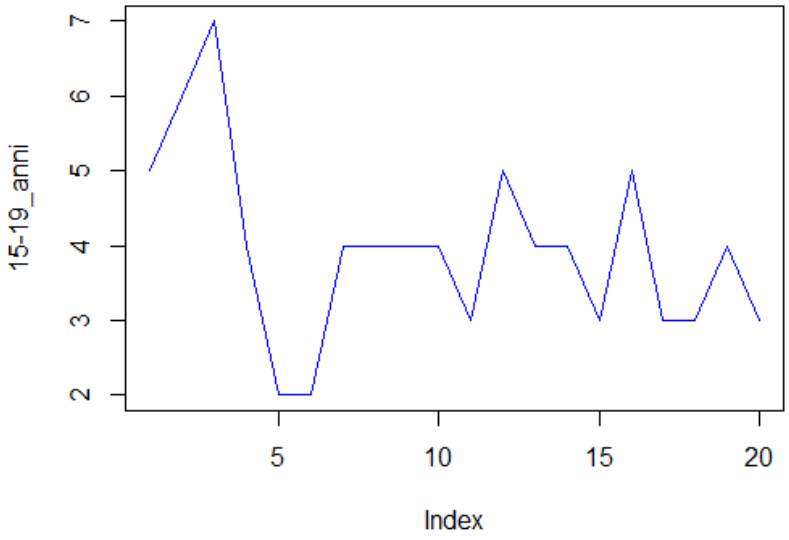


Figura 1.7:Rappresentazione dei valori assunti dal vettore numerico "quindici_diciannove_anni"

Dalla figura 1.7 si nota che la regione che presenta un picco è la Liguria, mentre si ha una discesa per quanto riguarda le regioni Trentino Alto Adige e Veneto.

Si può anche procedere in modo diverso utilizzando il comando di basso livello *lines().* Infatti i comandi

```
> plot(quindici_diciannove_anni,ylab = "15-19_anni",col="red")
> lines(quindici_diciannove_anni,col="blue")
```

hanno prodotto il grafico mostrato in Figura 1.8.

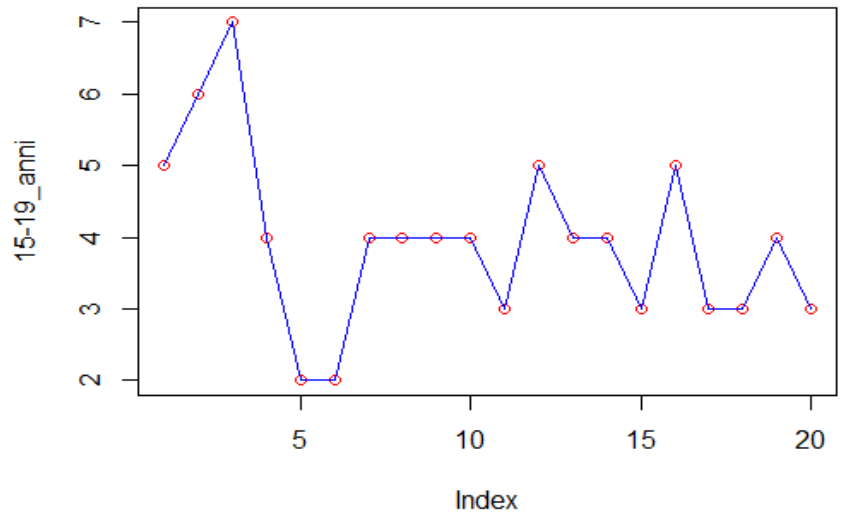


Figura 1.8:Rappresentazione dei valori assunti dal vettore numerico "quindici_diciannove_anni"

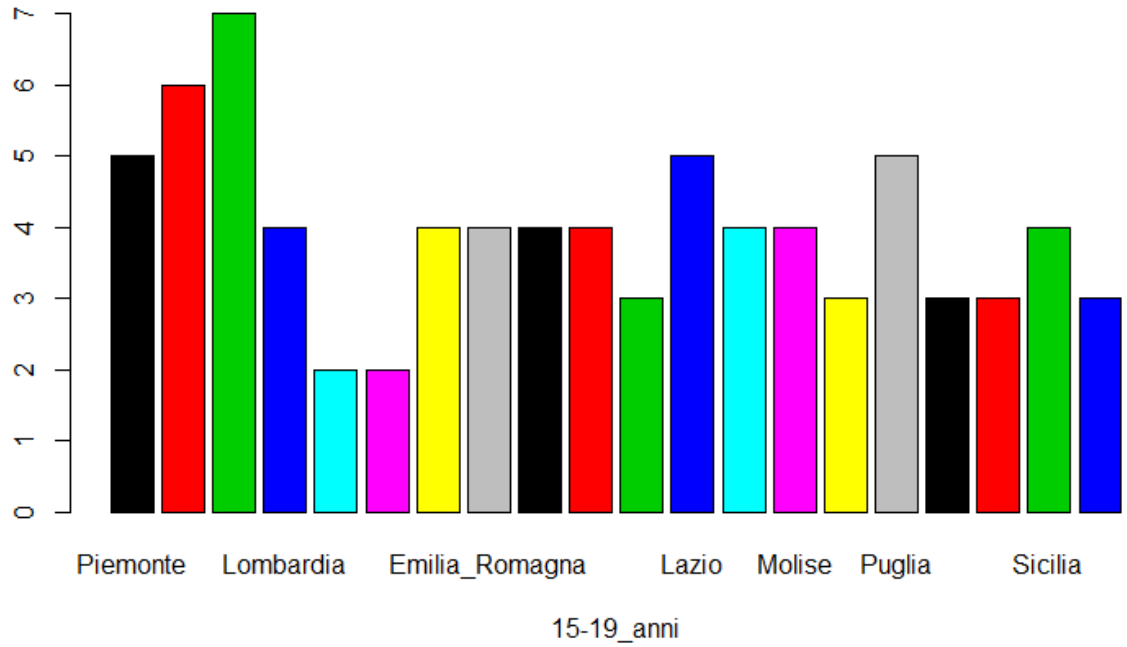
Grafici per matrici di dati

Supponendo ora di disporre di una matrice con i nostri dati. A partire da essa è stato possibile creare dei vettori contenenti gli elementi delle singole colonne. Utilizzando poi la funzione barplot() è stato possibile creare dei grafici a barre.

```
> matrix_eta<-cbind(c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),c(13,8,15,10,7,7,8,11,11,10,7,11,9,
8,7,11,7,6,8,7),c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,12,9,7,8,8),c(13,9,14,10,8,7,9,12,12,9,7,1
0,9,8,8,12,8,8,8,8),c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3
,5,4,3,3,3),c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),c(7,6,8,6,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
> rownames(matrix_eta)<-c("Piemonte","Valle_d_Aosta","Liguria","Lombardia","Trentino_Alto_Adige","
Veneto","Friuli_Venezia_Giulia","Emilia_Romagna","Toscana","Umbria","Marche","Lazio","Abruzzo","Mo
lise","Campania","Puglia","Basilicata","Calabria","Sicilia","Sardegna")
> colnames(matrix_eta)<-c("quindici_diciannove_anni","venti_ventiquattro_anni","venticinque_ventin
ove_anni","trenta_trentaquattro_anni","trentacinque_trentanove_anni","quaranta_quarantaquattro_ann
i","quarantacinque_quarantanove_anni","quindici_quarantanove_anni")
> b1<-matrix_eta[,1]
> b1
             Piemonte         Valle_d_Aosta               Liguria              Lombardia
                    5                     6                     7                     4
  Trentino_Alto_Adige                Veneto Friuli_Venezia_Giulia        Emilia_Romagna
                    2                     2                     4                     4
              Toscana               Umbria               Marche                 Lazio
                    4                     4                     3                     5
             Abruzzo               Molise              Campania                Puglia
                    4                     4                     3                     5
          Basilicata             Calabria               Sicilia              Sardegna
                    3                     3                     4                     3
> b2<-matrix_eta[,2]
> b2
             Piemonte         Valle_d_Aosta               Liguria              Lombardia
                   13                     8                    15                    10
  Trentino_Alto_Adige                Veneto Friuli_Venezia_Giulia        Emilia_Romagna
                    7                     7                     8                    11
              Toscana               Umbria               Marche                 Lazio
                   11                    10                     7                    11
             Abruzzo               Molise              Campania                Puglia
                    9                     8                     7                    11
          Basilicata             Calabria               Sicilia              Sardegna
                    7                     6                     8                     7
```

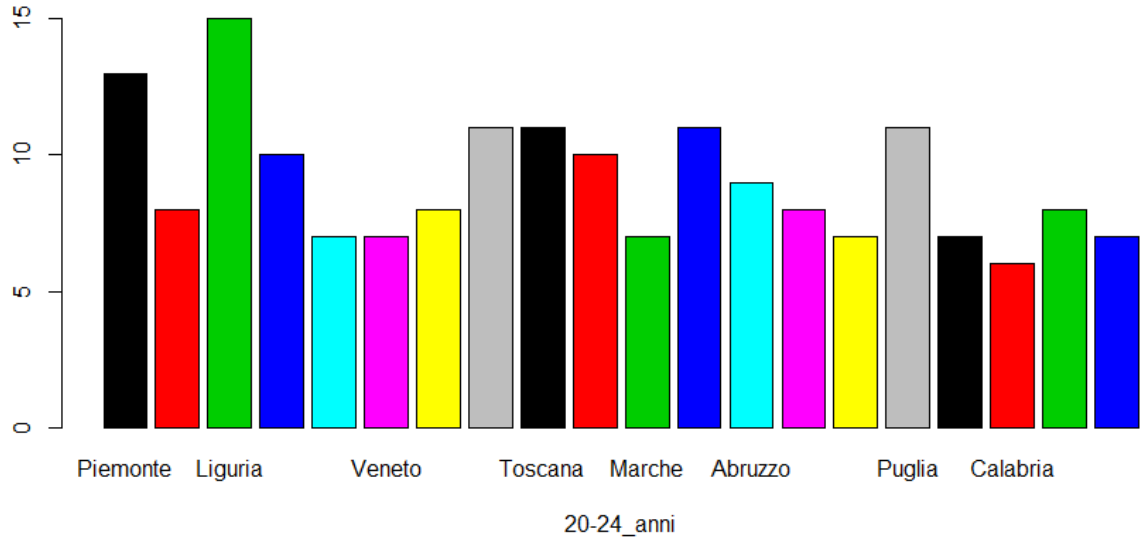
```
> b3<-matrix_eta[,3]
> b3
            Piemonte         Valle_d_Aosta               Liguria           Lombardia
                  13                    14                    15                  11
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   8                     8                     8                  13
             Toscana               Umbria                Marche               Lazio
                  12                    12                     9                  12
             Abruzzo               Molise              Campania              Puglia
                   9                     8                     8                  12
          Basilicata              Calabria               Sicilia            Sardegna
                   9                     7                     8                   8
> b4<-matrix_eta[,4]
> b4
            Piemonte         Valle_d_Aosta               Liguria           Lombardia
                  13                     9                    14                  10
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   8                     7                     9                  12
             Toscana               Umbria                Marche               Lazio
                  12                     9                     7                  10
             Abruzzo               Molise              Campania              Puglia
                   9                     8                     8                  12
          Basilicata              Calabria               Sicilia            Sardegna
                   8                     8                     8                   8
> b5<-matrix_eta[,5]
> b5
            Piemonte         Valle_d_Aosta               Liguria           Lombardia
                  10                     8                    11                   8
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   6                     6                     7                  10
             Toscana               Umbria                Marche               Lazio
                   9                     9                     6                   8
             Abruzzo               Molise              Campania              Puglia
                   8                     9                     7                  11
          Basilicata              Calabria               Sicilia            Sardegna
                   7                     6                     7                   6
> b6<-matrix_eta[,6]
> b6
            Piemonte         Valle_d_Aosta               Liguria           Lombardia
                   4                     4                     4                   3
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   2                     3                     3                   4
             Toscana               Umbria                Marche               Lazio
                   4                     3                     2                   4
             Abruzzo               Molise              Campania              Puglia
                   4                     4                     3                   5
          Basilicata              Calabria               Sicilia            Sardegna
                   4                     3                     3                   3
> b7<-matrix_eta[,7]
> b7
            Piemonte         Valle_d_Aosta               Liguria           Lombardia
                   0                     1                     0                   0
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   0                     0                     0                   0
             Toscana               Umbria                Marche               Lazio
                   0                     0                     0                   0
             Abruzzo               Molise              Campania              Puglia
                   0                     0                     0                   0
          Basilicata              Calabria               Sicilia            Sardegna
                   0                     0                     0                   0
> b8<-matrix_eta[,8]
> b8
            Piemonte         Valle_d_Aosta               Liguria           Lombardia
                   7                     6                     8                   6
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   4                     4                     5                   7
             Toscana               Umbria                Marche               Lazio
                   7                     6                     5                   7
             Abruzzo               Molise              Campania              Puglia
                   6                     5                     5                   8
          Basilicata              Calabria               Sicilia            Sardegna
                   5                     4                     5                   5
```

```
> barplot(b1,xlab = "15-19_anni",col=1:20)
```



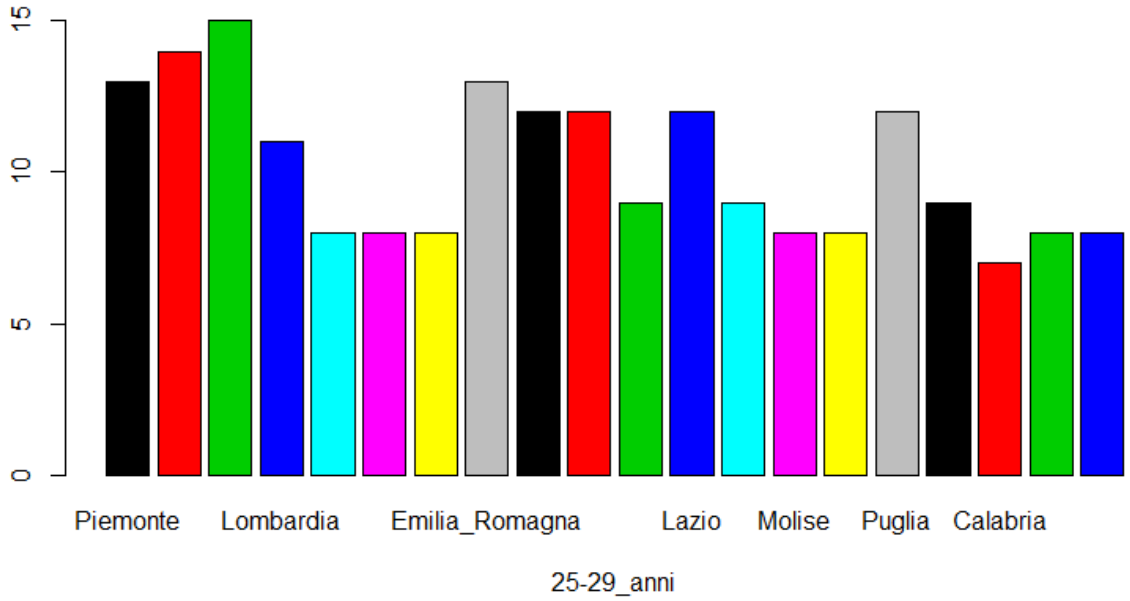
Dal grafico si può notare che la regione in cui si presentano maggiormente situazioni di aborto nella fascia d'età 15-19 anni è la Liguria, mentre le regioni in cui vi sono poche situazioni di aborto sono Trentino Alto Adige e Veneto.

```
> barplot(b2,xlab = "20-24_anni",col=1:20)
```



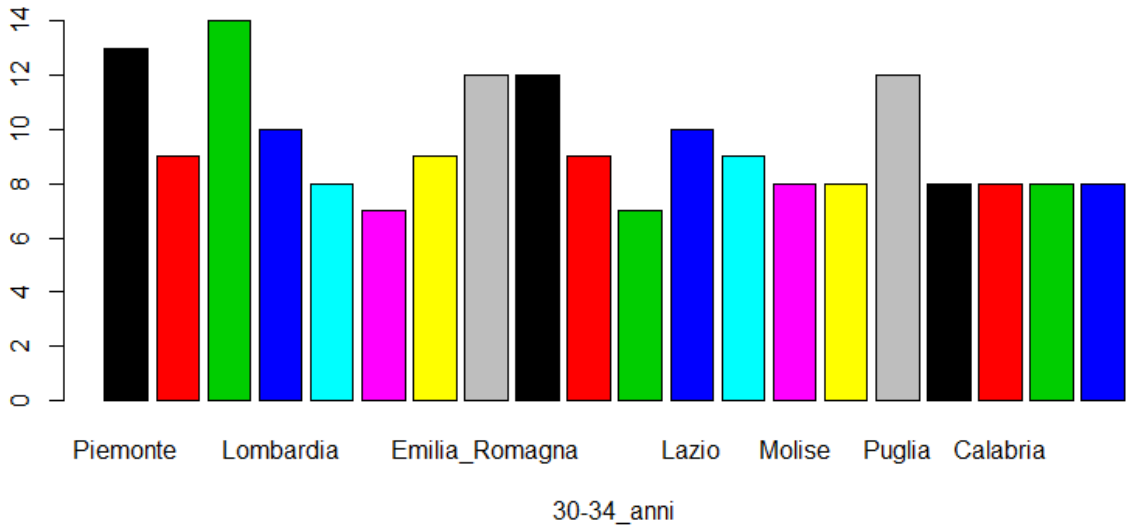
Dal grafico si può notare che la regione in cui si presentano maggiormente situazioni di aborto nella fascia d'età 20-24 anni è la Liguria, mentre la regione in cui vi sono poche situazioni di aborto è la Calabria.

```
> barplot(b3,xlab = "25-29_anni",col=1:20)
```



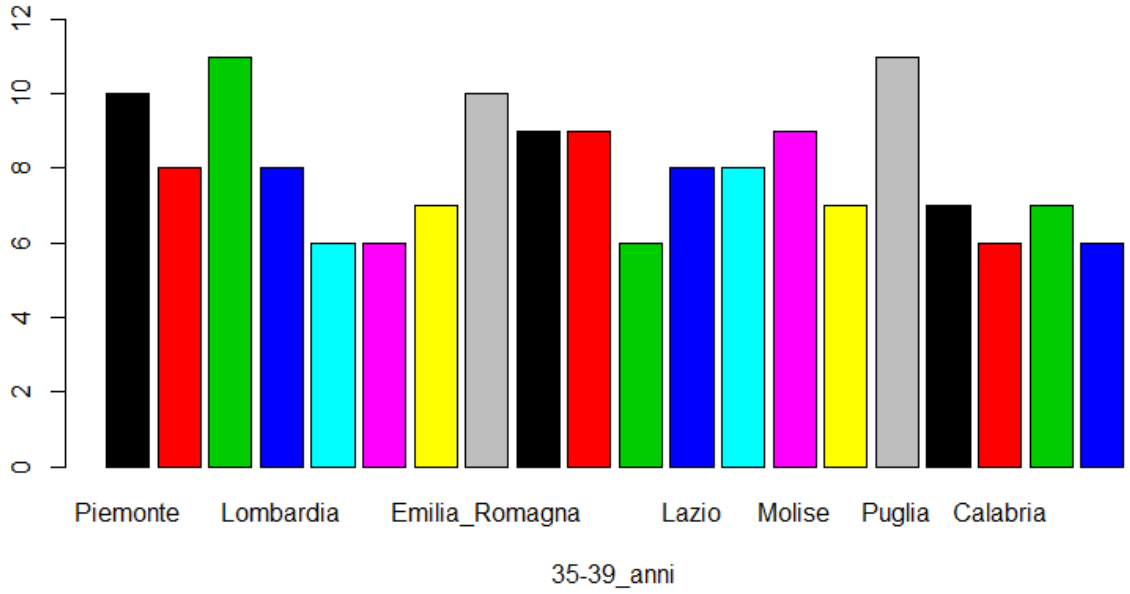
Dal grafico si può notare che la regione in cui si presentano maggiormente situazioni di aborto nella fascia d'età 25-29 anni è la Liguria, mentre la regione in cui vi sono poche situazioni di aborto è la Calabria.

```
> barplot(b4,xlab = "30-34_anni",col=1:20)
```



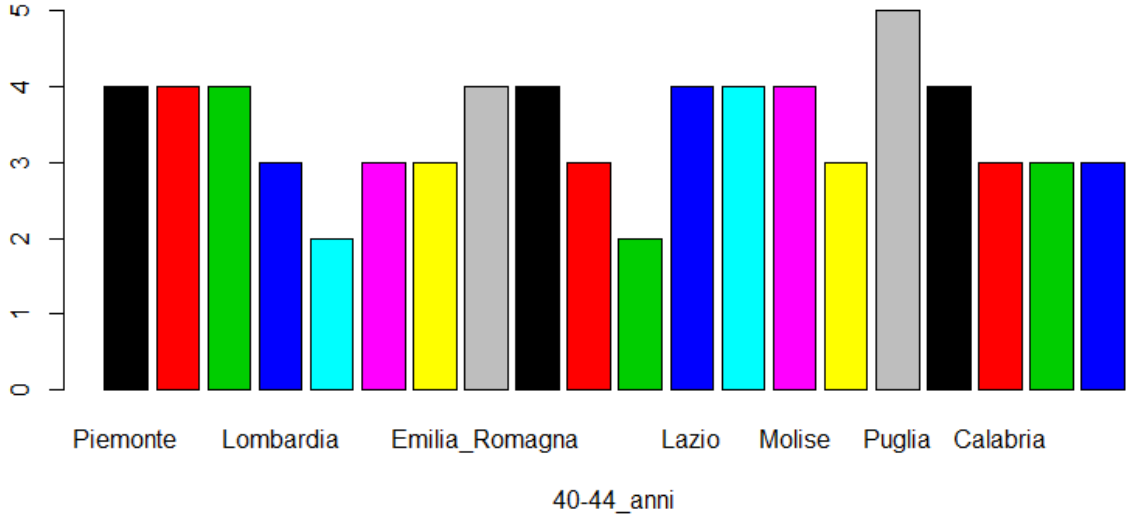
Dal grafico si può notare che la regione in cui si presentano maggiormente situazioni di aborto nella fascia d'età 30-34 anni è la Liguria, mentre le regioni in cui vi sono poche situazioni di aborto sono Veneto e Marche.

```
> barplot(b5,xlab = "35-39_anni",col=1:20)
```



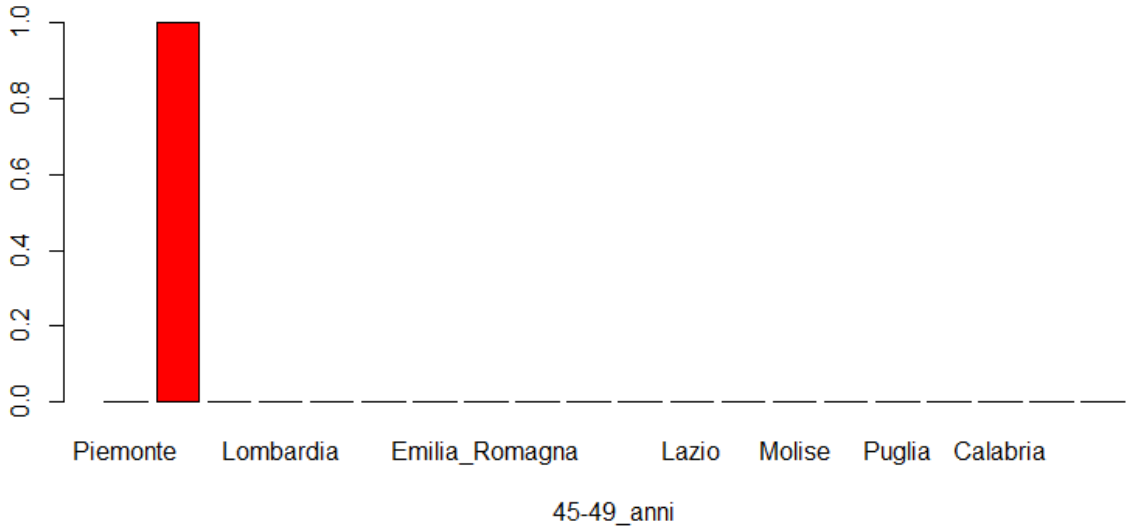
Dal grafico si può notare che le regioni in cui si presentano maggiormente situazioni di aborto nella fascia d'età 35-39 anni sono la Liguria e la Puglia, mentre le regioni in cui vi sono poche situazioni di aborto sono Trentino Alto Adige, Veneto, Marche, Calabria e Sardegna.

```
> barplot(b6,xlab = "40-44_anni",col=1:20)
```



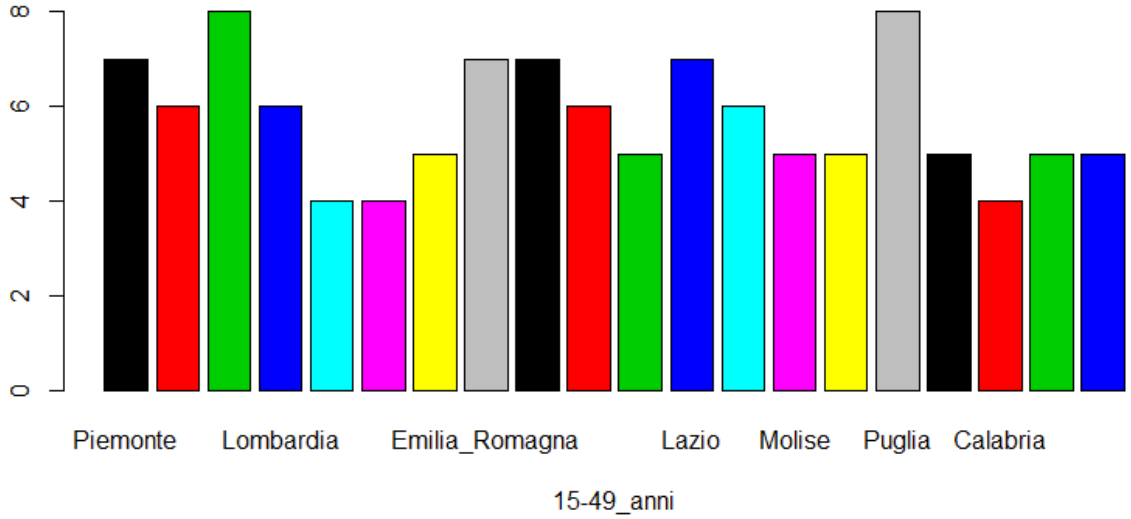
Dal grafico si può notare che la regione in cui si presentano maggiormente situazioni di aborto nella fascia d'età 40-44 anni è la Puglia, mentre le regioni in cui vi sono poche situazioni di aborto sono Trentino Alto Adige e Marche.

```
> barplot(b7,xlab = "45-49_anni",col=1:20)
```



Dal grafico si può notare che la regione in cui si presentano maggiormente situazioni di aborto nella fascia d'età 45-49 anni è la Valle d'Aosta, mentre le regioni in cui non vi sono quasi situazioni di aborto sono le rimanenti.

```
> barplot(b8,xlab = "15-49_anni",col=1:20)
```



Dal grafico si può notare che le regioni in cui si presentano maggiormente situazioni di aborto nella fascia d'età 15-49 anni sono la Puglia e la Liguria, mentre le regioni in cui vi sono poche situazioni di aborto sono il Trentino Alto Adige,Veneto e Calabria.

1.2.1 Diagramma di Pareto

Il *diagramma di Pareto* consiste di un diagramma a barre verticali con le modalità ordinate in ordine decrescente rispetto alle loro frequenza relativa; inoltre le frequenze relative sono visualizzate anche nella loro forma cumulata, mediante una sequenza di segmenti crescenti.

L'analisi basata sul diagramma di Pareto consiste nelle seguenti fasi:

1. decidere come classificare i dati;
2. rilevare i dati ed ordinarli;
3. disegnare il diagramma;
4. costruire la linea cumulativa;
5. aggiungere le informazioni di base.

Prendendo in considerazione il vettore "quindici_diciannove_anni", la costruzione del diagramma di Pareto consiste nelle seguenti fasi:

Fase 1:

Consideriamo la tabella:

	Regioni	**Frequenza**
a	Abruzzo	4
b	Basilicata	3
c	Calabria	3
d	Campania	3
e	Emilia_Romagna	4
f	Friuli_Venezia_Giulia	4
g	Lazio	5
h	Liguria	7
i	Lombardia	4
l	Marche	3
m	Molise	4
n	Piemonte	5
o	Puglia	5
p	Sardegna	3
q	Sicilia	4
r	Toscana	4
s	Trentino_Alto_Adige	2
t	Umbria	4
u	Valle_d_Aosta	6
v	Veneto	2

Fase 2:

Prima di costruire il diagramma di Pareto risulta utile, per una visione immediata, riordinare le voci in una tabella in base alla rilevanza del parametro in esame, essendo stata scelta la frequenza degli aborti, si elenca prima la regione con maggior frequenza, poi la successiva e così via.

	Regioni	**Frequenza**	**Frequenza relativa (Freq/tot)**	**Frequenza cumulata (Freq1+Freq2/tot)**
h	Liguria	7	0.0886	8%
u	Valle_d_Aosta	6	0.0759	16%

g	Lazio	5	0.0633	22%
n	Piemonte	5	0.0633	29%
o	Puglia	5	0.0633	35%
a	Abruzzo	4	0.0506	41%
e	Emilia_Romagna	4	0.0506	46%
f	Friuli_Venezia_Giulia	4	0.0506	51%
i	Lombardia	4	0.0506	56%
m	Molise	4	0.0506	61%
q	Sicilia	4	0.0506	66%
r	Toscana	4	0.0506	71%
t	Umbria	4	0.0506	76%
b	Basilicata	3	0.0380	80%
c	Calabria	3	0.0380	84%
d	Campania	3	0.0380	87%
l	Marche	3	0.0380	91%
p	Sardegna	3	0.0380	94%
s	Trentino_Alto_Adige	2	0.0253	97%
v	Veneto	2	0.0253	100%
	Totale	79		

I dati così ordinati costituiscono la base per la costruzione del diagramma di Pareto.

Fase 3-4:

La linea dei valori percentuali cumulativi è rappresentata da una spezzata.

Nel linguaggio R non è presente alcuna funzione in grado di generare un diagramma di Pareto. Tuttavia è stato possibile creare facilmente questo tipo di grafico con il seguente codice:

```
> Quindici_diciannove_anni<-c(rep("a",4),rep("b",3),rep("c",3),rep("d",3),rep("e",4),rep("f",4),r
ep("g",5),rep("h",7),rep("i",4),rep("l",3),rep("m",4),rep("n",5),rep("o",5),rep("p",3),rep("q",4)
,rep("r",4),rep("s",2),rep("t",4),rep("u",6),rep("v",2))
> tab<-table(Quindici_diciannove_anni)
> ord<-rev(sort(tab))
> propOrd<-prop.table(ord)
> x<-barplot(propOrd,ylim = c(0,1.05),main = "Diagramma_di_Pareto",col = 1:20)
> lines(x,cumsum(propOrd),type = "b",pch=16)
> text(x-0.2,cumsum(propOrd)+0.03,paste(format(cumsum(propOrd)*100,digits=2),"%"))
```

che permette di ottenere il grafico illustrato in Figura 1.10.

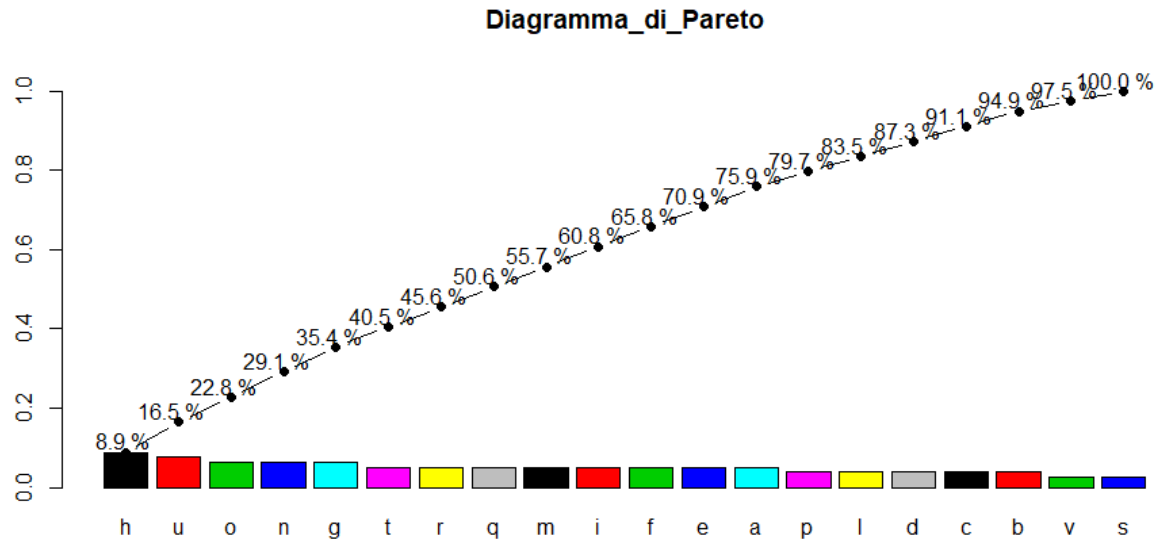


Figura 1.10: Diagramma di Pareto ottenuto osservando la numerosità delle regioni in cui si ha un aborto in un fissato intervallo temporale

Osservando la figura 1.10 è possibile determinare in quali regioni si presenta maggiormente l'aborto. Nel caso specifico, circa il 70% del numero degli aborti è presente nelle prime 12 regioni presenti nella tabella riportata precedentemente.

1.2.2 Istogrammi

Gli istogrammi, che si utilizzano per *variabili quantitative*, sono una particolare rappresentazione grafica di una distribuzione di frequenza in classi. Gli istogrammi sono quindi 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. Sono state utilizzate le frequenze assolute delle classi. La funzione che realizza in R un istogramma è hits().
Ad esempio, riferendoci al vettore *"*quindici_diciannove_anni*"* delle 20 regioni le seguenti linee di codice

```
> table(quindici_diciannove_anni)
quindici_diciannove_anni
2 3 4 5 6 7
2 5 8 3 1 1
> hist(quindici_diciannove_anni,freq = TRUE,main = "Istogramma_dei_15-19_anni",ylab = "Frequen
za_assoluta_delle_classi")
```

hanno prodotto l'istogramma rappresentato in Figura 1.11 in cui il numero di classi è stato scelto automaticamente da R.

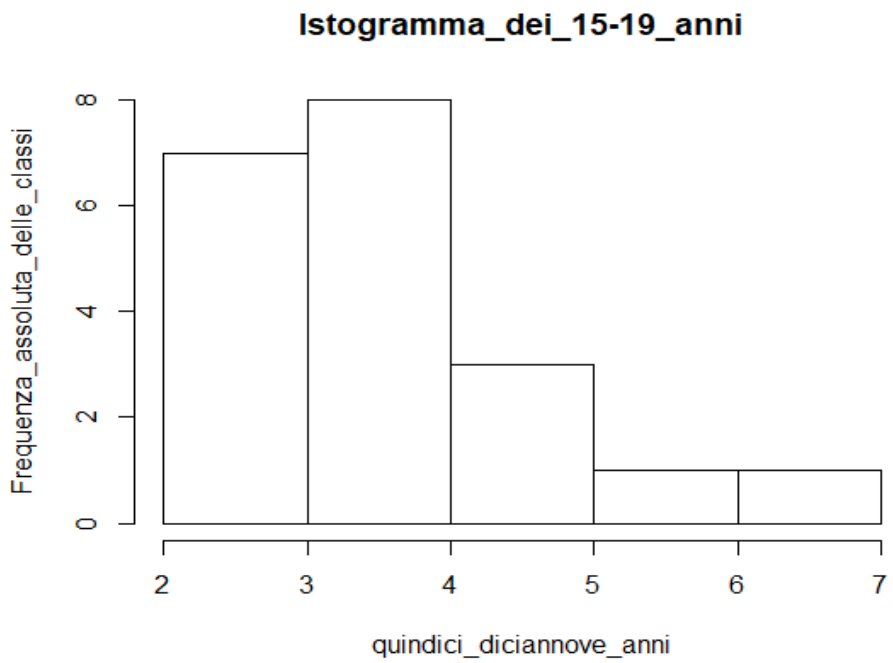


Figura 1.11:Istogramma relativo al vettore numerico "quindici_diciannove_anni" in base alle frequenze assolute delle classi

La suddivisione in classi scelta automaticamente da R è la seguente: [2,3], (3,4],(4,5],(5,6],(6,7]. Infatti dei 20 valori, 7 cadono nella prima classe, 8 nella seconda classe, 3 nella terza classe e 1 nella seconda e prima classe.

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

```
> h<-hist(quindici_diciannove_anni,freq = TRUE,main = "Istogramma_dei_15-19_anni",ylab = "Freq
uenza_assoluta_delle_classi")
> str(h)
List of 6
 $ breaks  : int [1:6] 2 3 4 5 6 7
 $ counts  : int [1:5] 7 8 3 1 1
 $ density : num [1:5] 0.35 0.4 0.15 0.05 0.05
 $ mids    : num [1:5] 2.5 3.5 4.5 5.5 6.5
 $ xname   : chr "quindici_diciannove_anni"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

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

Ad esempio se utilizziamo l'informazione density, attraverso questa possiamo calcolare la frequenza relativa associate alle classi dell'istogramma. Per calcolarla bisogna moltiplicare la density per l'ampiezza di ogni intervallo dell'istogramma (in questo caso è 1):

```
> f<-1*h$density
> f
[1] 0.35 0.40 0.15 0.05 0.05
> sum(f)
[1] 1
```

Effettuando la somma delle frequenze relative il risultato è uguale ad 1, il risultato che ci si aspettava.

1.2.3 Boxplot

Considerando il nostro campione dei valori assunti da una variabile quantitativa "15-19_anni". Si è proceduto ordinando 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, ossia il valore per il quale 50% dei dati sono alla sua sinistra e il restante 50% è alla sua destra è detto *mediana*. Q_0 e Q_4 forniscono il *minimo* e il *massimo* dei valori del campione. In R i quartili si calcolano tramite la funzione quantile() e la funzione summary() permette di determinare i valori precisi del minimo, del massimo, della media, della mediana, del primo e del terzo quantile.

Ad esempio, riferendoci al vettore "quindici_diciannove_anni" delle 20 regioni risulta:

```
> quantile(quindici_diciannove_anni)
  0%  25%  50%  75% 100%
2.00 3.00 4.00 4.25 7.00
```

da cui si deduce che Q_0=2 (minimo), Q_1=3 (primo quartile), Q_2=4 (secondo quartile), Q_3=4.25 (terzo quartile), Q_4=7 (massimo). Inoltre, si ha

```
> summary(quindici_diciannove_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    3.00    4.00    3.95    4.25    7.00
```

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 orizzontali, dette i *baffi*. Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni ed è maggiore o uguale del valore $Q1 - 1.5 \cdot (Q3 - Q1)$, mentre il baffo superiore corrisponde al valore più grande delle osservazioni, ed è minore o uguale del valore $Q3+1.5 \cdot (Q3-Q1)$. I valori dei dati che sono al di sopra del baffo superiore o al di sotto del baffo inferiori costituiscono un'anomalia nei dati, e sono rappresentati da punti.
Il boxplot viene utilizzato per illustrare alcune caratteristiche di una distribuzione di frequenza: la *centralità*, la *forma*, la *dispersione* e la *presenza di eventuali valori anomali*.
In R un boxplot è ottenuto tramite la funzione boxplot().

Ad esempio, è stato costruito il boxplot a partite dal vettore "quindici_diciannove_anni" usando il seguente comando

```
> boxplot(quindici_diciannove_anni,xlab="15-19_anni",main="Boxplot_di_15-19_anni",col = "gre
en")
```

che ha condotto al grafico illustrato di Figura 1.12.

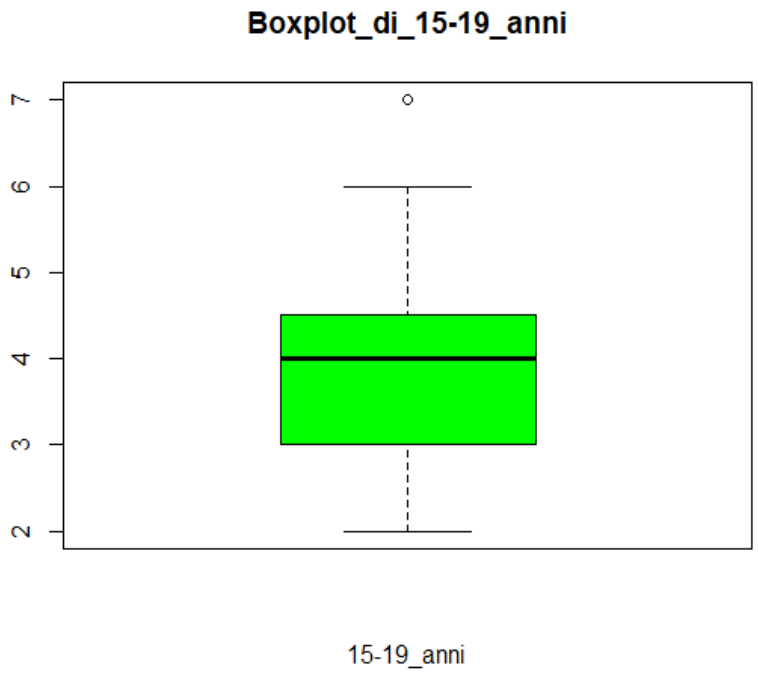


Figura 1.12: Boxplot relativo al vettore numerico "quindici_diciannove_anni"

Dalla figura 1.12 si nota che gli estremi della scatola sono Q1=3 e Q3=4.25; essa è tagliata da una linea orizzontale in corrispondenza di Q2=4. Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale di Q1-1.5(Q3-Q1)= 1.125, ossia 2, mentre il baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a Q3+1.5(Q3-Q1)=6.125, ossia 6. Quindi i baffi sono stati posti in corrispondenza del minimo valore 2 e del massimo valore 6. Il valore al di fuori dell'intervallo [1.125,6.125] è 7, che risulta essere in questo caso un valore anomalo al di sopra del baffo superiore che si può notare nel boxplot , e riguarda il numero di aborti presenti in Liguria.

1.3 Grafici per tabelle e matrici di dati

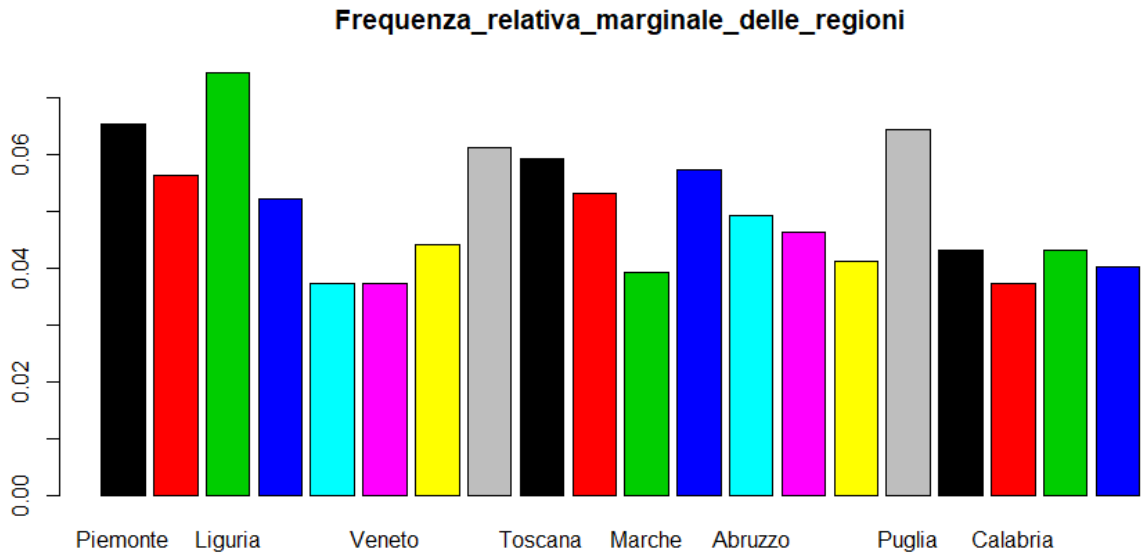
Sono state fornite ora delle rappresentazioni grafiche utili per visualizzare tabelle di contingenza e per matrici di dati numerici.

Grafici per matrici di dati numerici

```
> matrix_eta<-cbind(c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),c(13,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,6
,8,7),c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,12,9,7,8,8),c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8),
c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3,5,4,3,3,3),c(0,1,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0),c(7,6,8,6,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
> rownames(matrix_eta)<-c("Piemonte","Valle_d_Aosta","Liguria","Lombardia","Trentino_Alto_Adige","Veneto","F
riuli_Venezia_Giulia","Emilia_Romagna","Toscana","Umbria","Marche","Lazio","Abruzzo","Molise","Campania","Pu
glia","Basilicata","Calabria","Sicilia","Sardegna")
> colnames(matrix_eta)<-c("15-19_anni","20-24_anni","25-29_anni","30-34_anni","35-39_anni","40-44_anni","45-
49_anni","15-49_anni")
> matrix_eta_r<-prop.table(matrix_eta)
> barplot(margin.table(matrix_eta_r,1),main="Frequenza_relativa_marginale_delle_regioni",col = 1:20)
> barplot(margin.table(matrix_eta_r,2),main="Frequenza_relativa_marginale_delle_fasce_età",col = 1:8)
> barplot(margin.table(matrix_eta,1),main="Frequenza_assoluta_marginale_delle_regioni",col = 1:20)
> barplot(margin.table(matrix_eta,2),main="Frequenza_assoluta_marginale_delle_fasce_età",col = 1:8)
```

```
> barplot(prop.table(matrix_eta,2),beside=TRUE,main="Frequenza_relativa_condizionata_f(regioni|età)",col = 1:20)
> barplot(prop.table(matrix_eta,1),beside=TRUE,main="Frequenza_relativa_condizionata_f(età|regioni)",col = 1:8)
```

A partire dalla matrice dei dati *matrix_eta* attraverso la funzione prop.table(matrix_eta) è stata costruita la matrice *matrix_eta_r* delle frequenze relative e attraverso le funzioni margin.table(matrix_eta_r, 1) e margin.table(matrix_eta_r, 2) è stata calcolata la distribuzione di frequenza relativa marginale ed è stata visualizzata attraverso la funzione barplot() come evidenziato in Figura 1.13, mentre attraverso le funzioni margin.table(matrix_eta, 1) *e* margin.table(matrix_eta, 2) è stata calcolata la distribuzione di frequenza assoluta marginale ed è stata visualizzata anch'essa attraverso la funzione barplot() come evidenziato in Figura 1.14.
Inoltre attraverso la funzione prop.table(matrix_eta, 2) e prop.table(matrix_eta, 1) è stata calcolata la matrice delle frequenza relativa condizionata dalle ordinate e dalle ascisse rispettivamente e sono state visualizzate tali probabilità attraverso la funzione barplot(), come evidenziato nella Figura 1.15.



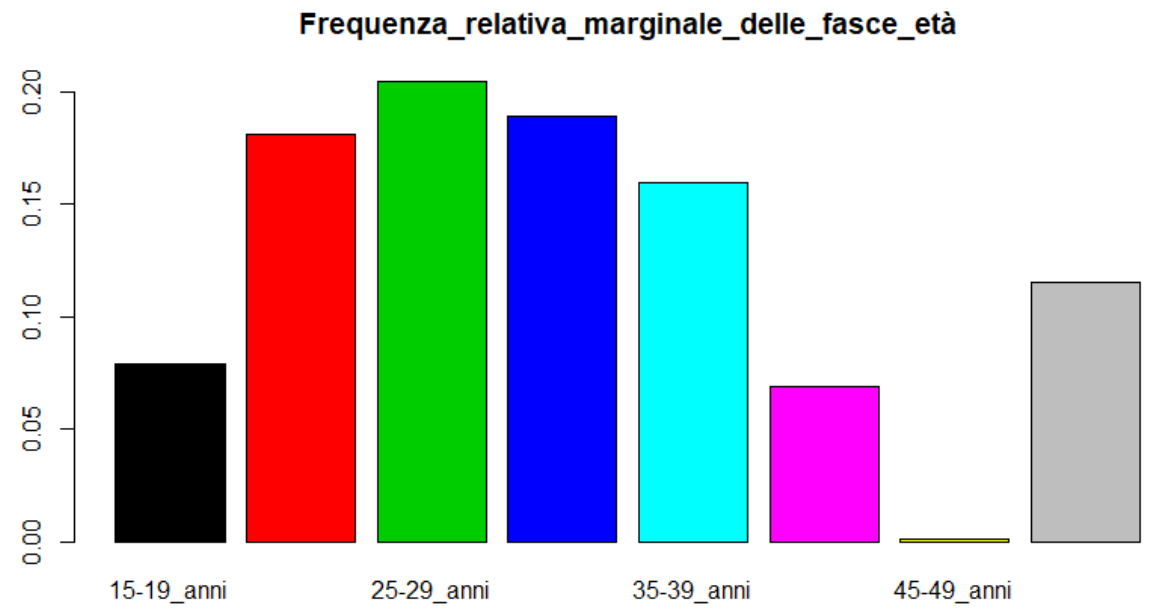
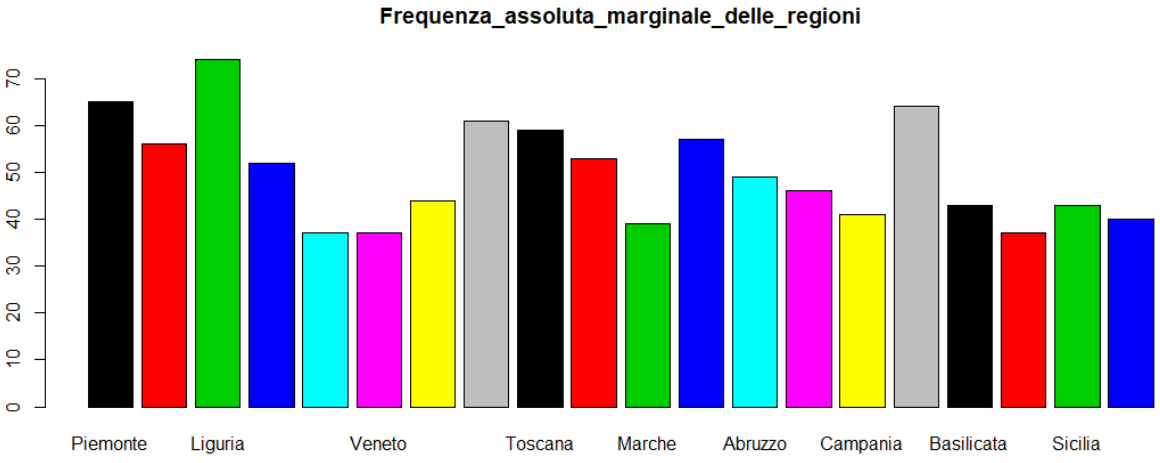


Figura 1.13:Frequenze relative marginali



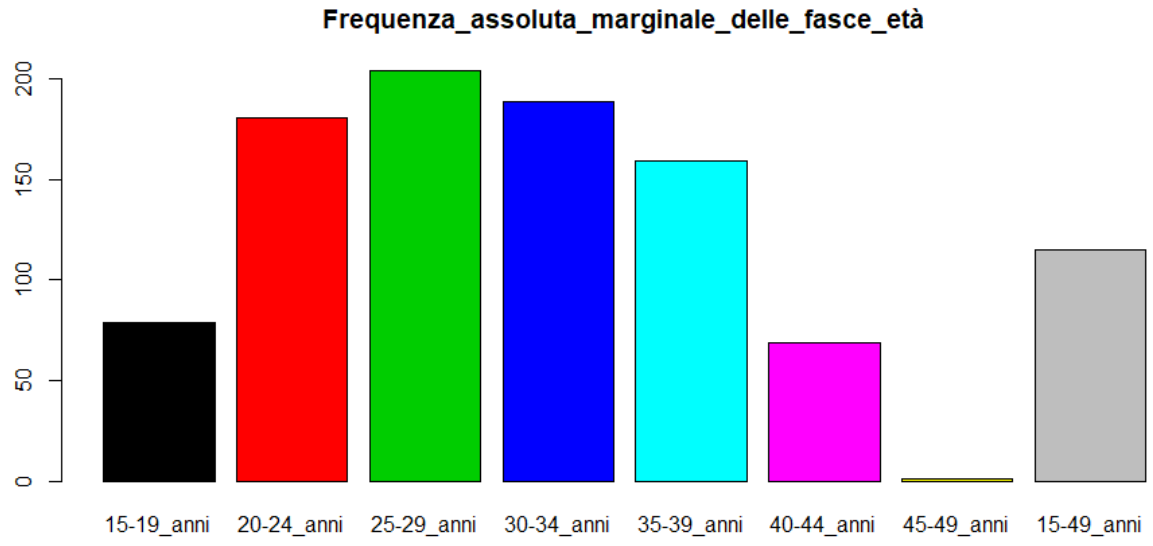
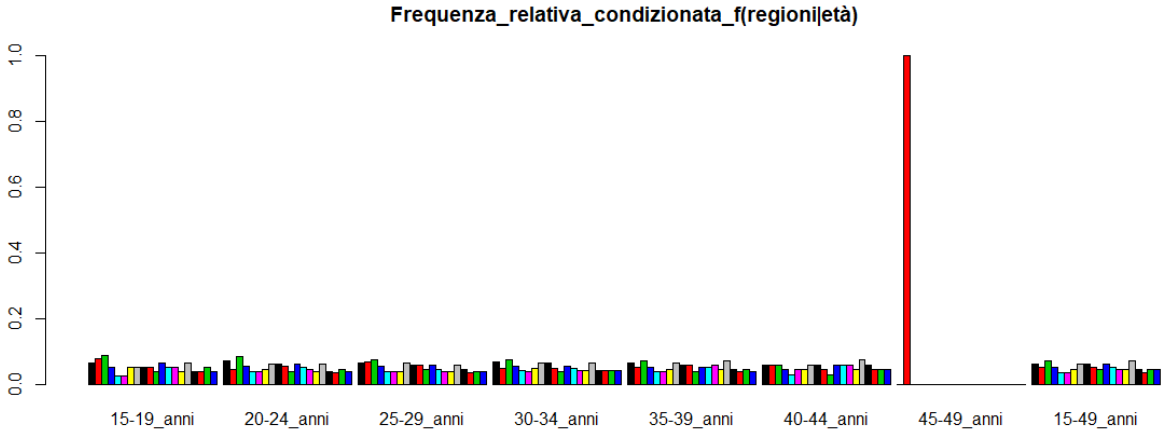


Figura 1.14:Frequenze assolute marginali



-

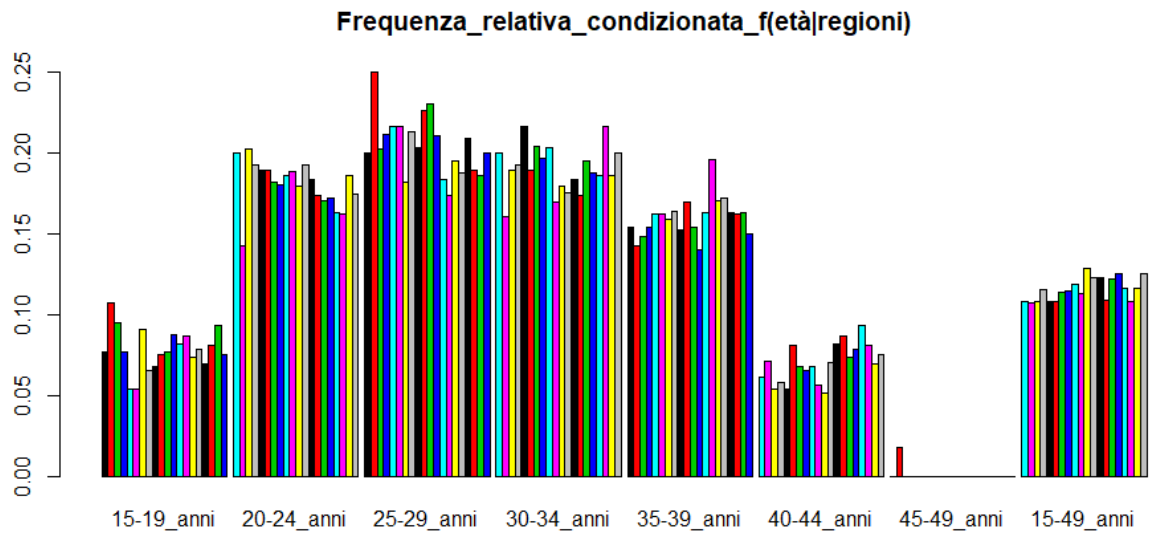


Figura 1.15:Frequenza relativa condizionata

1.2.2 Scatterplot

I diagrammi di dispersione (Scatterplot) sono delle rappresentazioni grafiche delle relazioni tra variabili quantitative. Ciò che si fa è scegliere la variabile da porre sull'asse delle ascisse (variabile indipendente) e quella da porre sull'asse delle ordinate (variabile dipendente), utilizzando la funzione plot, il risultato è una nuvola di punti che può presentare o meno una certa regolarità. Inoltre il grafico di dispersione può evidenziare se è presente una certa relazione tra le variabili e di che tipo di relazione si tratta (lineare, quadratica ,…).

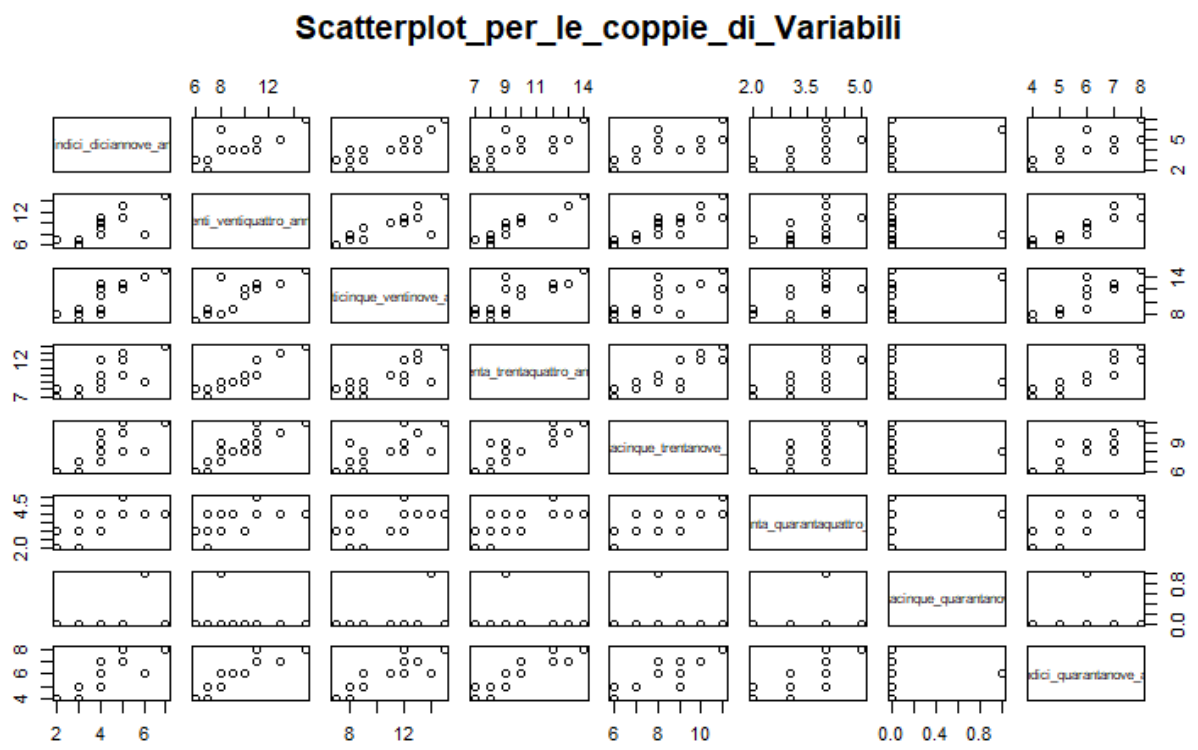
Le seguenti linee di codice

```
> Eta<-data.frame(quindici_diciannove_anni=c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),venti_ventiquattro_anni=c
(13,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,6,8,7),venticinque_ventinove_anni=c(13,14,15,11,8,8,8,13,12,12,9,12,
9,8,8,12,9,7,8,8),trenta_trentaquattro_anni=c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8),trentacinque_tren
tanove_anni=c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),quaranta_quarantaquattro_anni=c(4,4,4,3,2,3,3,4,4,3,
2,4,4,4,3,5,4,3,3,3),quarantacinque_quarantanove_anni=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),quindici_quara
ntanove_anni=c(7,6,8,6,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
```

definiscono un data frame contenente le otto fasce di età delle 20 regioni italiane.
È stato realizzato uno scatterplot tramite la funzione pairs() che è in grado di visualizzare in un'unica finestra una pluralità di grafici per punti ottenuti mettendo in relazione tutte le coppie di variabili definite all'interno del nostro data frame, la seguente linea di codice

```
> pairs(Eta,main="Scatterplot_per_le_coppie_di_Variabili")
```

ha prodotto il grafico visualizzato in Figura 1.16.



Le diverse immagini illustrano le nuvole di punti che si ottengono prendendo in considerazione tutte le differenti coppie di variabili.

2 STATISTICA DESCRITTIVA UNIVARIATA

2.1 Funzione di distribuzione empirica

Funzione di distribuzione empirica discreta

È stato considerato il nostro vettore “15-19_anni” delle 20 regioni italiane, ed è stata costruita la seguente tabella in cui possiamo vedere z_i i valori da essa assunti e assumiamo che essi siano ordinati in ordine crescente ,le frequenze assolute n_i, le frequenze relative f_i, le frequenze relative cumulate F_i. La funzione di distribuzione empirica discreta, è una funzione a gradini in cui ogni gradino indica quale proporzione di dati presenta un valore minore e uguale di quello indicato sull’asse delle ascisse ed è così definita:

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

i	z_i	n_i	f_i	F_i
1	2	2	2/20	2/20
2	3	5	5/20	7/20
3	4	8	8/20	15/20
4	5	3	3/20	18/20
5	6	1	1/20	19/20
6	7	1	1/20	20/20

Le seguenti linee di codice

```
> quindici_diciannove_anni<-c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3)
> round(cumsum(table(quindici_diciannove_anni)/length(quindici_diciannove_anni)),3)
   2    3    4    5    6    7
0.10 0.35 0.75 0.90 0.95 1.00
```

hanno permesso di ottenere le frequenze relative cumulate relative al vettore "quindici_diciannove_anni", arrotondate alla terza cifra decimale. Il linguaggio R dispone della classe stepfun che implementa una serie di metodi per trattare funzioni a gradino. In particolare, la funzione ecdf() permette di disegnare il grafico della funzione di distribuzione empirica per variabili quantitative discrete. La seguente linea di codice

```
> plot(ecdf(quindici_diciannove_anni),main="Funzione_di_distribuzione_empirica_discreta",col="red")
```

ha prodotto il grafico illustrato in Figura 2.1.

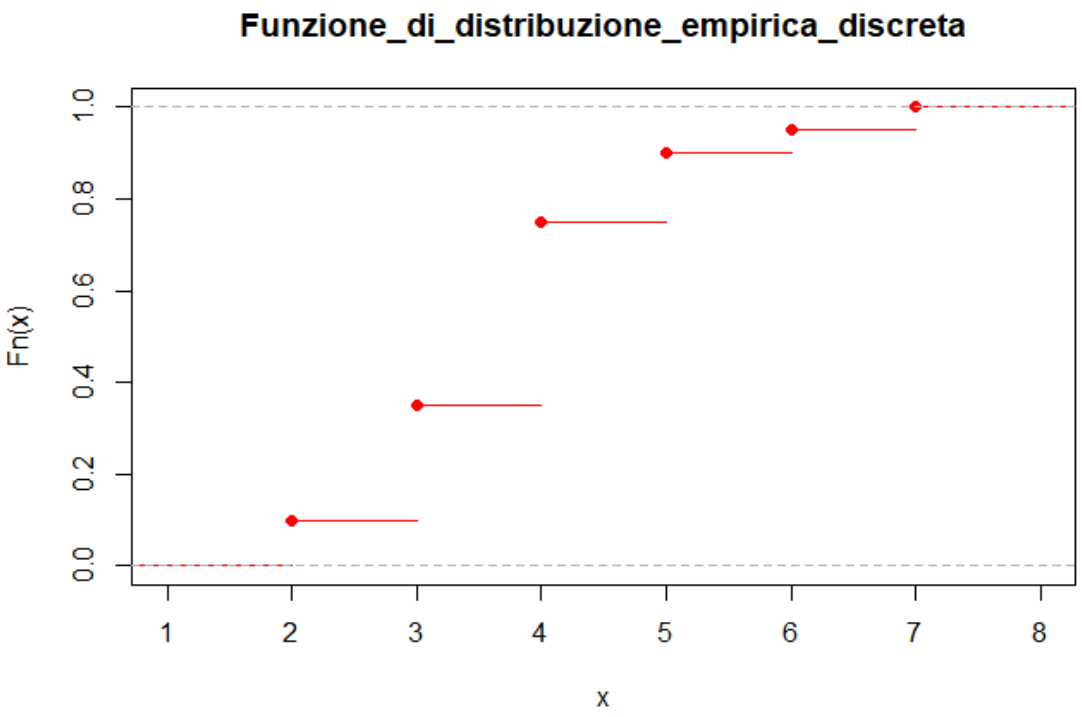


Figura 2.1: Funzione di distribuzione empirica discreta del vettore "quindici_diciannove_anni"

Funzione di distribuzione empirica continua

Per fenomeni quantitativi continui la funzione di distribuzione empirica è una funzione continua. In particolare, se i dati sono raccolti in k distinte classi C1 = [z1, z2),C2 = [z2, z3) . . . ,Ck = [zk, zk+1), con z1 < z2 < . . . zk < zk+1, la funzione di distribuzione empirica è così definita:

$$F(x) = \begin{cases} 0, & x < z_1 \\ \dots & \\ F_i, & x = z_i \\ \dfrac{F_{i+1} - F_i}{z_{i+1} - z_i}\, x + \dfrac{z_{i+1}F_i - z_i F_{i+1}}{z_{i+1} - z_i}, & z_i < x < z_{i+1} \\ F_{i+1}, & x = z_{i+1} \\ \dots & \\ 1, & x \geq z_{k+1} \end{cases}$$

La funzione di distribuzione empirica continua coincide con il segmento che passa per i punti (z_i, F_i) e (z_{i+1}, F_{i+1}), ossia:

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

Riferendoci alla "quindici_diciannove_anni" delle 20 regioni italiane, sono state introdotte le seguenti classi [0,2),[2,4), [4,6), [6,8), [8,10), in cui le classi [0,2) e [8,10) sono delle classi fittizie che ci servono per tracciare la linea y=0 nell'intervallo [0,2) e la linea y=1 nell'intervallo [8,10) nel grafico della funzione di distribuzione empirica continua.
Nella seguente tabella sono indicate le frequenze assolute, le frequenze relative e le frequenze relative cumulate associate al vettore "quindici_diciannove_anni".

i	$[z_i, z_{i+1}]$	n_{i+1}	f_{i+1}	F_{i+1}
0	[0,2)	0	0	0
1	[2,4)	7	7/20	7/20
2	[4,6)	11	11/20	18/20
3	[6,8)	2	2/20	20/20
4	[8,10)	0	0	1

Di seguito è indicata la funzione di distribuzione empirica continua utilizzando le classi [0,2), [2,4), [4,6), [6,8), [8,10).

Le seguenti linee di codice

```
> classi<-c(0,2,4,6,8,10)
> Fi<-cumsum(table(cut(quindici_diciannove_anni, breaks = classi, right=FALSE)))/length(quindici_diciannove_anni)
> Fi
 [0,2)  [2,4)  [4,6)  [6,8) [8,10)
  0.00   0.35   0.90   1.00   1.00
```

hanno permesso di visualizzare le frequenze relative cumulate associate alle classi scelte.
Successivamente sono state utilizzate delle linee di codice in R che consentono di ottenere il grafico della distribuzione empirica continua.

Infatti le seguenti linee di codice

```
> Fi<-c(0,Fi)
> plot(classi,Fi,type="b",axes = FALSE, main="Funzione_di_distribuzione_empirica_continua",col="red")
> axis(1,classi)
> axis(2,format(Fi,digits = 2))
> box()
```

hanno prodotto il grafico illustrato in Figura 2.2.

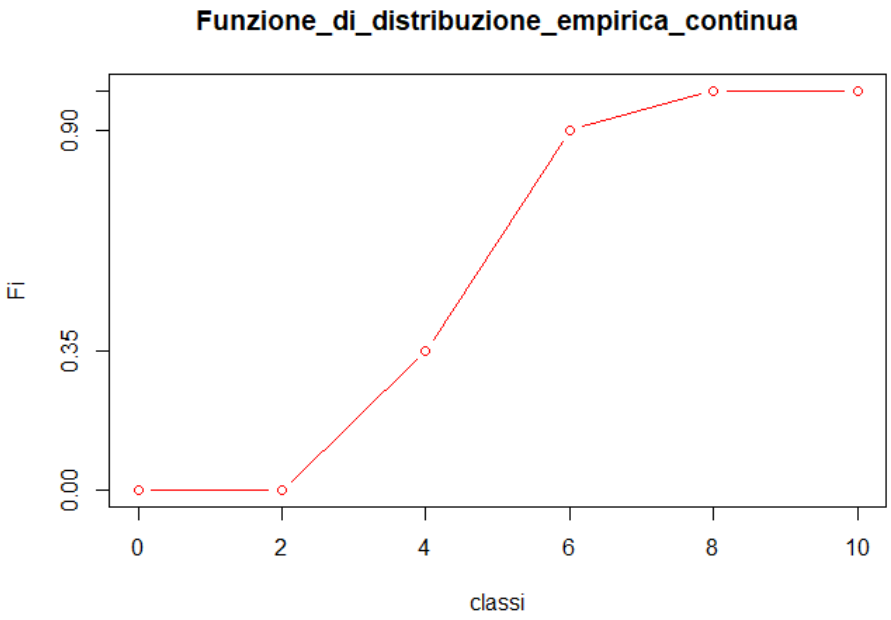


Figura 2.2: Grafico della funzione di distribuzione empirica continua del vettore "quindici_diciannove_anni" utilizzando le [0,2),[2,4), [4,6), [6,8), [8,10).

2.2 Indici di posizione e di dispersione

Alcuni *indici di sintesi*, detti anche *statistiche*, utili a descrivere dei dati numerici, sono media, mediana, moda, varianza, deviazione standard e 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.

Media, mediana e moda

- Dato un insieme di dati x1, x2,…., xn di n valori, detto campione di ampiezza, la media campionaria è la media aritmetica di questi valori.

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

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

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

$$s_i = x_i - \bar{x} \qquad (i=1,2,\ldots,n)$$

Che indica quanto ogni valore sia distante dalla media campionaria. La somma degli scarti dalla media campionaria è nulla.

- 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 dal minore al maggiore. 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 assicura che lo stesso numero di valori cada sia a sinistra che a destra della mediana stessa.

Il sistema R mette a disposizione le funzioni mean() e median() per calcolare rispettivamente la media e la mediana di un insieme di dati.
Considerando sempre i nostri otto vettori che rappresentano le fasce di età, il seguente codice R ci ha permesso di ricavare la media e la mediana di essi.

```
> mean(quindici_diciannove_anni)
[1] 3.95
> median(quindici_diciannove_anni)
[1] 4
> mean(venti_ventiquattro_anni)
[1] 9.210526
> median(venticinque_ventinove_anni)
[1] 9
> median(venti_ventiquattro_anni)
[1] 8
> mean(venticinque_ventinove_anni)
[1] 10.2
> mean(trenta_trentaquattro_anni)
[1] 9.45
> median(trenta_trentaquattro_anni)
[1] 9
> mean(trentacinque_trentanove_anni)
[1] 7.95
> median(trentacinque_trentanove_anni)
[1] 8
> mean(quaranta_quarantaquattro_anni)
[1] 3.45
> median(quaranta_quarantaquattro_anni)
[1] 3.5
> mean(quarantacinque_quarantanove_anni)
[1] 0.05
> median(quarantacinque_quarantanove_anni)
[1] 0
> mean(quindici_quarantanove_anni)
[1] 5.75
> median(quindici_quarantanove_anni)
[1] 5.5
```

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.

- La terza statistica utilizzata per descrivere la centralità di una distribuzione di dati è la *moda campionaria*.

La moda campionaria di un insieme di dati, se esiste, è la modalità a cui è associata la frequenza (assoluta o relativa) più elevata. Se esistono più modalità con frequenza massima, ciascuna di esse è detto valore modale.
La moda, quindi, rappresenta il valore prevalente nell'insieme dei dati, ovvero quello che si presenta con maggiore frequenza. Non esiste invece in R una funzione per estrarre la moda da una distribuzione di dati poiché è facilmente ricavabile osservando il grafico delle frequenze assolute. E' stata definita una funzione che calcola i valori modali di un vettore numerico.

```
> moda<-function(v){
+ y<-table(v)
+ z<-which(y==max(y))
+ return(c(z))
+ }
```

Ed è stata applicata tale formula ai nostri otto vettori.

```
> moda(quindici_diciannove_anni)
4
3
> moda(venti_ventiquattro_anni)
7
1
> moda(venticinque_ventinove_anni)
8
2
> moda(trenta_trentaquattro_anni)
8
2
> moda(trentacinque_trentanove_anni)
6
1
> moda(quaranta_quarantaquattro_anni)
4
3
> moda(quarantacinque_quarantanove_anni)
0
1
> moda(quindici_quarantanove_anni)
5
2
```

È stata costruita una tabella per avere un'idea più intuitiva dei valori ottenuti.

	15-19 anni	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
media	3.95	9.210526	10.2	9.45	7.95	3.45	0.05	5.75
mediana	4	8	9	9	8	3.5	0	5.5
moda	4	7	8	8	6	4	0	5

Sono state considerate le otto fasce d'età di numerosità uguale alle 20 regioni italiane.
Le seguenti linee di codice

```
> quindici_diciannove_anni<-c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3)
> venti_ventiquattro_anni<-c(13,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,8,7)
> venticinque_ventinove_anni<-c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,12,9,7,8,8)
> trenta_trentaquattro_anni<-c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8)
> trentacinque_trentanove_anni<-c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6)
> quaranta_quarantaquattro_anni<-c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3,5,4,3,3,3)
> quarantacinque_quarantanove_anni<-c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
> quindici_quarantanove_anni<-c(7,6,8,6,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5)
> par(mfrow=c(2,4))
> plot(table(quindici_diciannove_anni),col="red")
> plot(table(venti_ventiquattro_anni),col="blue")
> plot(table(venticinque_ventinove_anni),col="brown")
> plot(table(trenta_trentaquattro_anni),col="magenta")
> plot(table(trentacinque_trentanove_anni),col="black")
> plot(table(quaranta_quarantaquattro_anni),col="green")
> plot(table(quarantacinque_quarantanove_anni),col="orange")
> plot(table(quindici_quarantanove_anni),col="grey")
```

hanno prodotto il grafico illustrato in Figura 2.3 che rappresenta le distribuzioni di frequenza per le otto differenti fasce di età.

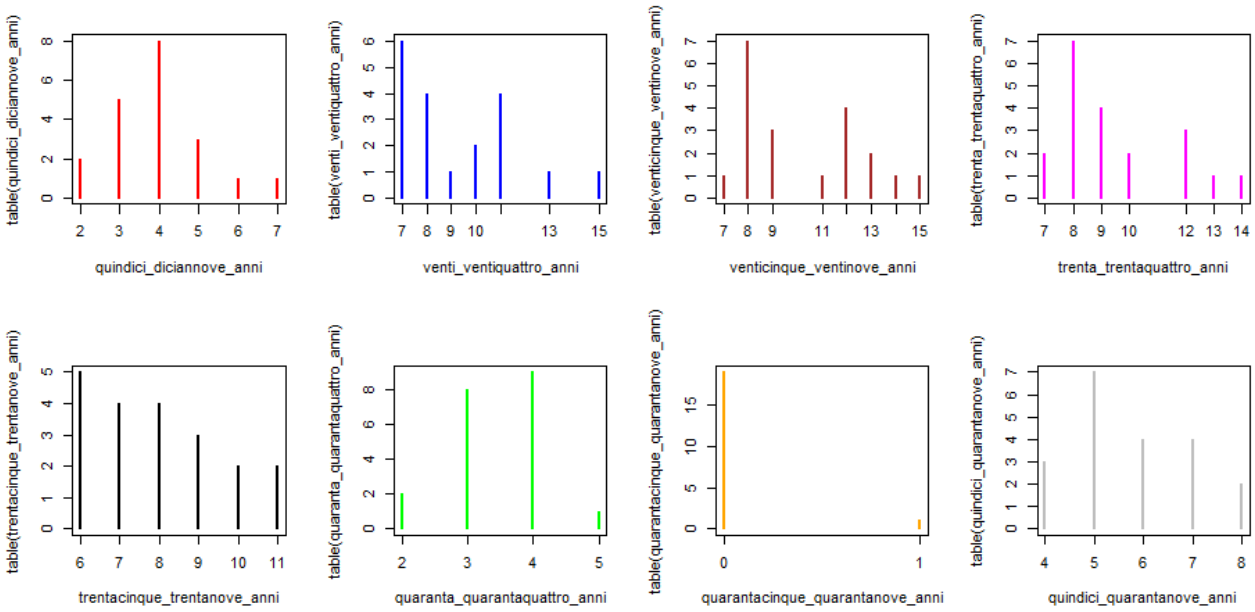


Figura 2.3: Distribuzione delle frequenze assolute delle otto fasce d'età

Per descrivere la forma di una distribuzione si può confrontare la media e la mediana campionaria. Se queste due misure sono uguali allora la distribuzione è simmetrica, se la media è maggiore della mediana, la distribuzione è più sbilanciata verso destra, altrimenti è più sbilanciata verso sinistra, come è possibile vedere nella figura 2.3.

Mediana per una distribuzione di frequenze

Un modo di procedere diverso per definire la mediana consiste nel considerare le frequenze relative cumulate. Sia X una variabile quantitativa e siano z1, z2, . . . , zk le modalità distinte da essa assunte, con z1 < z2 < . . . < zk.
Considerato un campione (x1, x2, . . . , xn), siano $F_i = f_1 + f_2 + \cdots + f_i$ (i = 1, 2, . . . , k) le frequenze relative cumulate.

Definizione La mediana per una distribuzione di frequenze è definita come la modalità i-esima (i=1,2,. . . , k) che soddisfa la doppia disuguaglianza:

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

Come si evince dalla definizione, la mediana di una distribuzione di frequenza è un valore di sintesi che indica un punto centrale intorno al quale si dispone la distribuzione di frequenza.

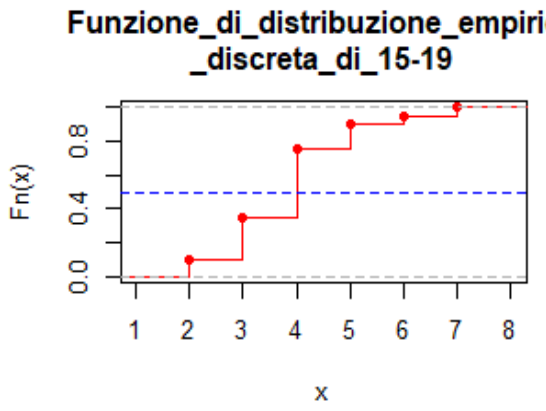
```
> Fdati1<-cumsum(table(quindici_diciannove_anni))/length(quindici_diciannove_anni)
> round(Fdati1,2)
   2    3    4    5    6    7
0.10 0.35 0.75 0.90 0.95 1.00
> Fdati2<-cumsum(table(venti_ventiquattro_anni))/length(venti_ventiquattro_anni)
> round(Fdati2,2)
   7    8    9   10   11   13   15
0.32 0.53 0.58 0.68 0.89 0.95 1.00
```

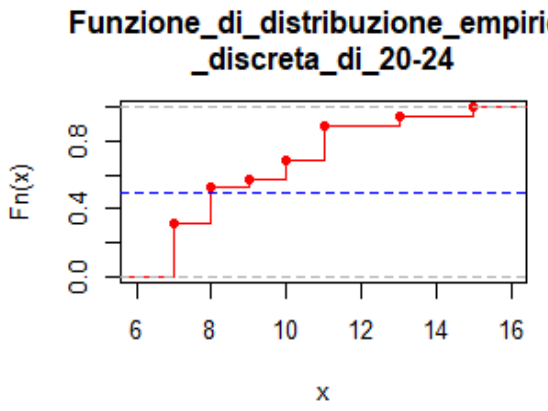
```
> Fdati3<-cumsum(table(venticinque_ventinove_anni))/length(venticinque_ventinove_anni)
> round(Fdati3,2)
   7    8    9   11   12   13   14   15
0.05 0.40 0.55 0.60 0.80 0.90 0.95 1.00
> Fdati4<-cumsum(table(trenta_trentaquattro_anni))/length(trenta_trentaquattro_anni)
> round(Fdati4,2)
   7    8    9   10   12   13   14
0.10 0.45 0.65 0.75 0.90 0.95 1.00
> Fdati5<-cumsum(table(trentacinque_trentanove_anni))/length(trentacinque_trentanove_anni)
> round(Fdati5,2)
   6    7    8    9   10   11
0.25 0.45 0.65 0.80 0.90 1.00
> Fdati6<-cumsum(table(quaranta_quarantaquattro_anni))/length(quaranta_quarantaquattro_anni)
> round(Fdati6,2)
   2    3    4    5
0.10 0.50 0.95 1.00
> Fdati7<-cumsum(table(quarantacinque_quarantanove_anni))/length(quarantacinque_quarantanove_anni)
> round(Fdati7,2)
   0    1
0.95 1.00
> Fdati8<-cumsum(table(quindici_quarantanove_anni))/length(quindici_quarantanove_anni)
> round(Fdati8,2)
   4    5    6    7    8
0.15 0.50 0.70 0.90 1.00
```

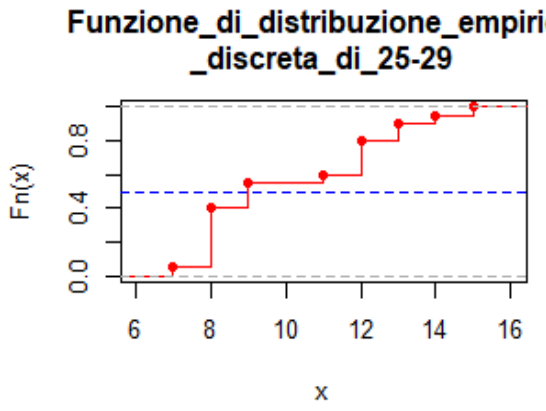
Segue che la mediana per la distribuzione di frequenze per “quindici_diciannove_anni” è 4, per venti_ventiquattro_anni” è 8, per “venticinque_ventinove_anni” è 9, per “trenta_trentaquattro_anni” è 9, per “trentacinque_trentanove_anni” è 8, per “quaranta_quarantaquattro_anni” è 3, per “quarantacinque_quarantanove_anni” è 0 , per “quindici_quarantanove_anni” è 5.

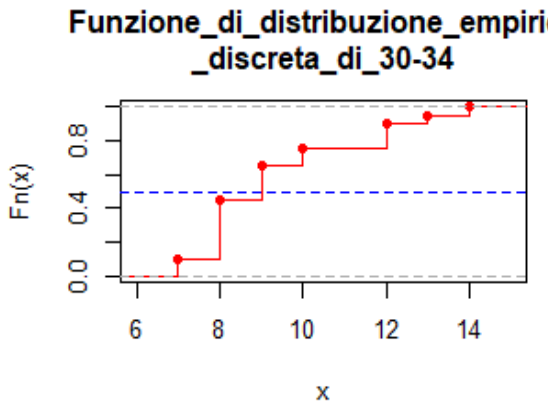
La mediana di una distribuzione di frequenze può essere ricavata graficamente a partire dalla funzione di distribuzione empirica discreta. Si traccia la funzione di distribuzione empirica e sull’asse delle ordinate si individua il punto 0.5 e da questo si traccia una linea orizzontale. Il minimo valore osservato la cui funzione di distribuzione empirica supera 0.5 è proprio la mediana per una distribuzione di frequenze.
Le seguenti linee di codice

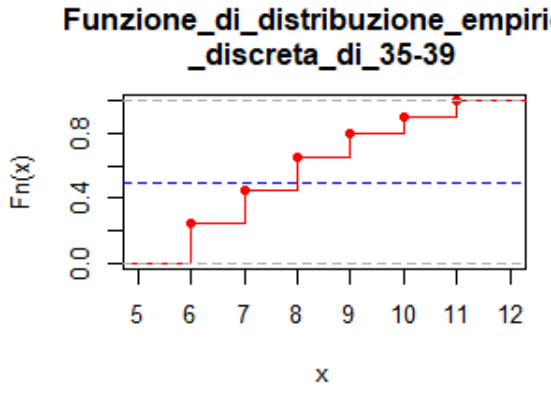
```
> par(mfrow=c(2,2))
> plot(ecdf(quindici_diciannove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_15-19
",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(venti_ventiquattro_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_20-24"
,verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(venticinque_ventinove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_25-
29",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(trenta_trentaquattro_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_30-3
4",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(trentacinque_trentanove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_3
5-39",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(quaranta_quarantaquattro_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_
40-44",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(quarantacinque_quarantanove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_
di_45-49",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
> plot(ecdf(quindici_quarantanove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_15-
49",verticals = TRUE, col="red")
> abline(h=0.5,lty=2,col="blue")
```

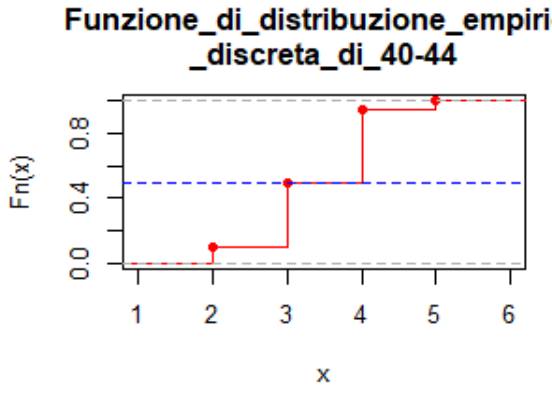

Funzione_di_distribuzione_empirica
_discreta_di_15-19
Fn(x)
0.0
0.4
0.8
1
2
3
4
5
6
7
8
x

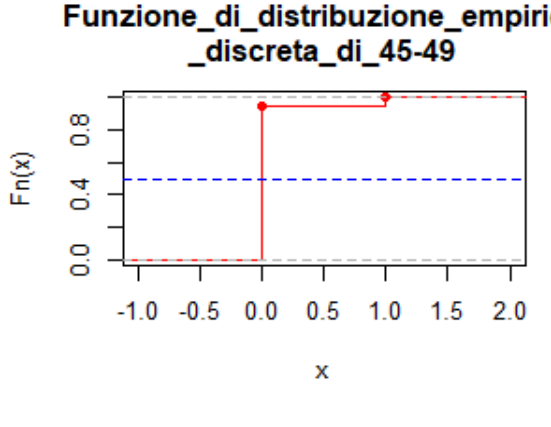

Funzione_di_distribuzione_empirica
_discreta_di_20-24
Fn(x)
0.0
0.4
0.8
6
8
10
12
14
16
x

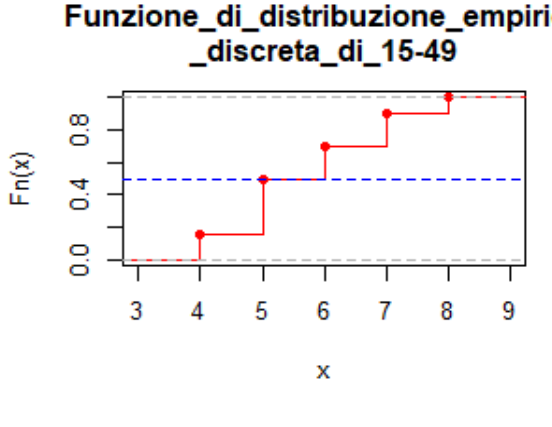

Funzione_di_distribuzione_empirica
_discreta_di_25-29
Fn(x)
0.0
0.4
0.8
6
8
10
12
14
16
x


Funzione_di_distribuzione_empirica
_discreta_di_30-34
Fn(x)
0.0
0.4
0.8
6
8
10
12
14
x


Funzione_di_distribuzione_empirica
_discreta_di_35-39
Fn(x)
0.0
0.4
0.8
5
6
7
8
9
10
11
12
x


Funzione_di_distribuzione_empirica
_discreta_di_40-44
Fn(x)
0.0
0.4
0.8
1
2
3
4
5
6
x


Funzione_di_distribuzione_empirica
_discreta_di_45-49
Fn(x)
0.0
0.4
0.8
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
x


Funzione_di_distribuzione_empirica
_discreta_di_15-49
Fn(x)
0.0
0.4
0.8
3
4
5
6
7
8
9
x

Quantili, percentili, decili e quartili

Oltre alla mediana che divide a metà un insieme di dati ordinati, si possono definire altri indici di posizione, detti *quantili*, che suddividono l'insieme dei dati ordinati in un fissato numero di parti uguali.
In R esistono 9 differenti algoritmi per calcolare i quantili ottenibili utilizzando la funzione quantile(v, probs=,type=j), dove v è un vettore numerico, probs è il vettore delle probabilità e j=1,2,…,9 denota il tipo di algoritmo selezionato.
Scegliendo type=2 non si tiene conto dell'interpolazione dei dati.
Se si omette probs vengono calcolati di default i quartili. Se si omette type per default viene implementato l'algoritmo di tipo 7.
La funzione quantile() restituisce il minimo, il massimo e i tre vettori quartili Q_1, Q_2 e Q_3.

```
> quantile(quindici_diciannove_anni)
  0%  25%  50%  75% 100%
2.00 3.00 4.00 4.25 7.00
> quantile(venti_ventiquattro_anni)
  0%  25%  50%  75% 100%
   7    7    8   11   15
> quantile(venticinque_ventinove_anni)
  0%  25%  50%  75% 100%
   7    8    9   12   15
> quantile(trenta_trentaquattro_anni)
  0%  25%  50%  75% 100%
 7.0  8.0  9.0 10.5 14.0
> quantile(trentacinque_trentanove_anni)
   0%   25%   50%   75%  100%
 6.00  6.75  8.00  9.00 11.00
> quantile(quaranta_quarantaquattro_anni)
  0%  25%  50%  75% 100%
 2.0  3.0  3.5  4.0  5.0
> quantile(quarantacinque_quarantanove_anni)
  0%  25%  50%  75% 100%
   0    0    0    0    1
> quantile(quindici_quarantanove_anni)
  0%  25%  50%  75% 100%
 4.0  5.0  5.5  7.0  8.0
```

La funzione summary() restituisce oltre al minimo, massimo e ai tre quartili Q_1, Q_2 e Q_3 anche la media campionaria.

```
> summary(quindici_diciannove_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    3.00    4.00    3.95    4.25    7.00
> summary(venti_ventiquattro_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.000   7.000   8.000   9.211  11.000  15.000
> summary(venticinque_ventinove_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    7.0     8.0     9.0    10.2    12.0    15.0
> summary(trenta_trentaquattro_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00    8.00    9.00    9.45   10.50   14.00
> summary(trentacinque_trentanove_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00    6.75    8.00    7.95    9.00   11.00
> summary(quaranta_quarantaquattro_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    3.00    3.50    3.45    4.00    5.00
> summary(quarantacinque_quarantanove_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00    0.05    0.00    1.00
> summary(quindici_quarantanove_anni)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00    5.00    5.50    5.75    7.00    8.00
```

I risultati ottenuti sono stati elencati nella seguente tabella.

	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
Q_0 (Minimo)	2	7	7	7	6	2	0	4
Q_1	3	7	8	8	6.75	3	0	5
Q_2 (Mediana)	4	8	9	9	8	3.50	0	5.50
Media	3.95	9.211	10.2	9.45	7.95	3.45	0.05	5.75
Q_3	4.25	11	12	10.50	9	4	0	7
Q_4(Massimo)	7	15	15	14	11	5	1	8

La seguente linea di codice

```
> boxplot(quindici_diciannove_anni,venti_ventiquattro_anni,venticinque_ventinove_anni,trenta_trenta
quattro_anni,trentacinque_trentanove_anni,quaranta_quarantaquattro_anni,quarantacinque_quarantanove
_anni,quindici_quarantanove_anni,names = c("15-19_anni","20-24_anni","25-29_anni","30-34_anni","35-
39_anni","40-44_anni","45-49_anni","15-49_anni"),col = c("red","orange","green","yellow","blue","ma
genta"))
```

ha prodotto il grafico in Figura 2.4 in cui sono confrontati nella stessa finestra grafica i boxplot degli otto vettori. È stato possibile analizzare i boxplot in Figura 2.4 facendo uso dei risultati in tabella.

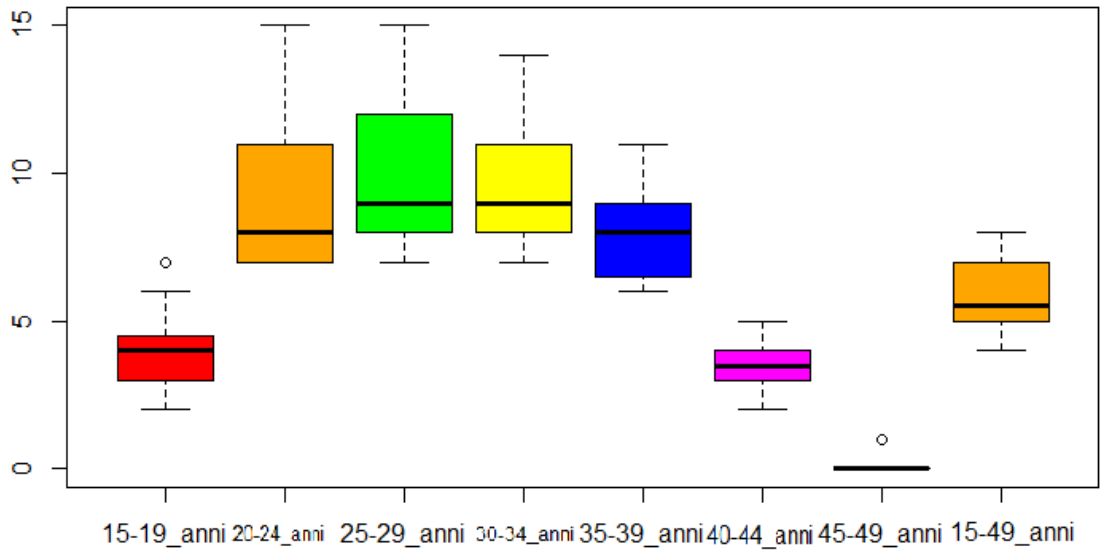


Figura 2.4: Confronto tra i boxplot dei vettori delle otto fasce d'età

Facendo uso dei risultati in tabella, per il primo boxplot si ha Q1-1.5(Q3-Q1)=1.125, ossia 2 e Q3+1.5(Q3-Q1)=6.125, ossia 6 , da cui segue che il baffo inferiore è posto 2 (minimo) e il baffo superiore in 7(massimo); per il secondo boxplot si ha Q1-1.5(Q3-Q1)=1, ossia 6 e Q3+1.5(Q3-Q1)=17, ossia 15 , da cui segue che il baffo inferiore è posto 7 (minimo) e il baffo superiore in 15(massimo); per il terzo boxplot si ha Q1-1.5(Q3-Q1)=2, ossia 7 e Q3+1.5(Q3-Q1)=18, ossia 15 , da cui segue che il baffo inferiore è posto 7 (minimo) e il baffo superiore in 15(massimo); per il quarto boxplot si ha Q1-1.5(Q3-Q1)=4.25, ossia 7 e Q3+1.5(Q3-Q1)=14.25, ossia 13, da cui segue che il baffo inferiore è posto 7 (minimo) e il baffo superiore in 14(massimo); per il quinto boxplot si ha Q1-1.5(Q3-Q1)=3.375, ossia 6 e Q3+1.5(Q3-Q1)=12.375, ossia 11 , da cui segue che il baffo

inferiore è posto 6 (minimo) e il baffo superiore in 11(massimo); per il sesto boxplot si ha Q1-1.5(Q3-Q1)=1.5, ossia 2 e Q3+1.5(Q3-Q1)=5.5, ossia 5, da cui segue che il baffo inferiore è posto 2 (minimo) e il baffo superiore in 5(massimo); per il settimo boxplot si ha Q1-1.5(Q3-Q1)=0, ossia 0 e Q3+1.5(Q3-Q1)=0, ossia 0 , da cui segue che il baffo inferiore è posto 0 (minimo) e il baffo superiore in 1(massimo); per l'ottavo boxplot si ha Q1-1.5(Q3-Q1)=2, ossia 4 e Q3+1.5(Q3-Q1)=10, ossia 8 , da cui segue che il baffo inferiore è posto 4 (minimo) e il baffo superiore in 8(massimo);.

Quantili per una distribuzione di frequenze

Un modo di procedere diverso per definire i quantili consiste nel considerare le frequenze relative cumulate. Sia X una variabile quantitativa e siano z1, z2, . . . , zk le modalità distinte da essa assunte, con z1 < z2 < . . . < zk. Considerato un campione (x1, x2, . . . , xn), siano $F_1, F_2 \ldots, F_k$ le frequenze relative cumulate.

Definizione 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:

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

Se p = 0.5 si ottiene la mediana per la distribuzione di frequenze.
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 definita 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 definita 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 della Definizione è implementato in R scegliendo j = 1 e usando la funzione

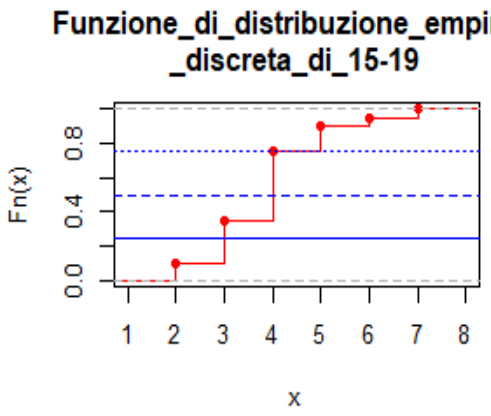
quantile(v, prob =, type = 1)

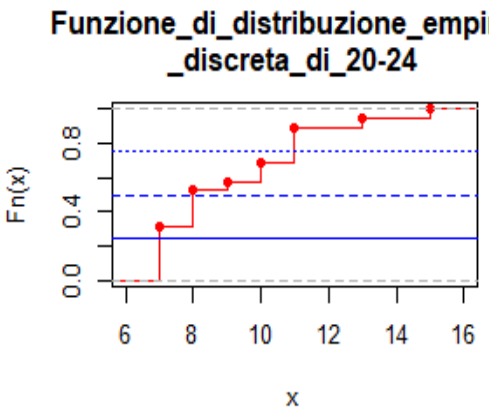
I quartili per una distribuzione di frequenze possono essere ricavati graficamente a partire dalla funzione di distribuzione empirica discreta. Si traccia la funzione di distribuzione empirica e sull'asse delle ordinate si individuano i punti 0.25,0.5, 0.75 e da questi si tracciano delle linee orizzontali:
- 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.

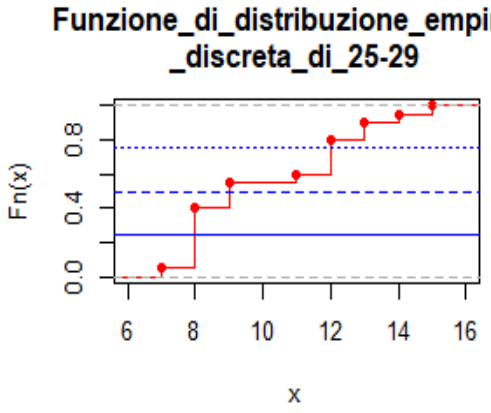
```
> quantile(quindici_diciannove_anni, type=1)
  0%  25%  50%  75% 100%
   2    3    4    4    7
> quantile(venti_ventiquattro_anni, type=1)
  0%  25%  50%  75% 100%
   7    7    8   11   15
> quantile(venticinque_ventinove_anni, type=1)
  0%  25%  50%  75% 100%
   7    8    9   12   15
> quantile(trenta_trentaquattro_anni, type=1)
  0%  25%  50%  75% 100%
   7    8    9   10   14
> quantile(trentacinque_trentanove_anni, type=1)
  0%  25%  50%  75% 100%
   6    6    8    9   11
> quantile(quaranta_quarantaquattro_anni, type=1)
  0%  25%  50%  75% 100%
   2    3    3    4    5
> quantile(quarantacinque_quarantanove_anni, type=1)
  0%  25%  50%  75% 100%
   0    0    0    0    1
> quantile(quindici_quarantanove_anni, type=1)
  0%  25%  50%  75% 100%
   4    5    5    7    8
```

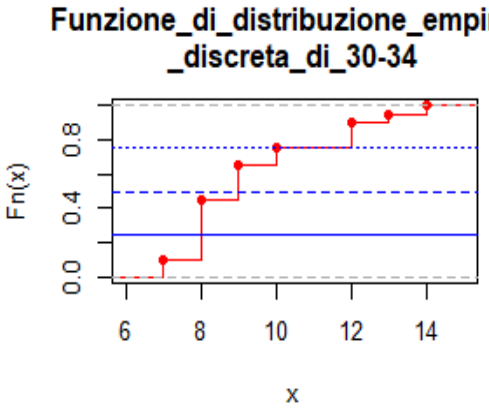
Inoltre, le seguenti linee di codice

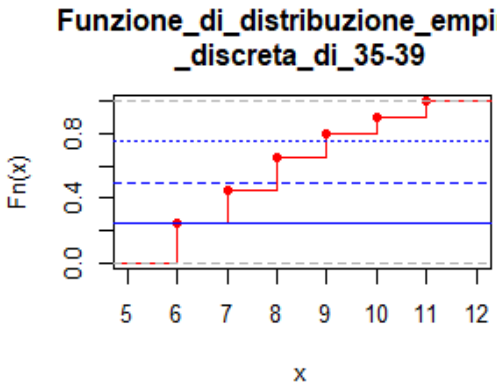
```
> par(mfrow=c(2,2))
> plot(ecdf(quindici_diciannove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_15-19",verticals
= TRUE, col="red")
> abline(h=0.25,lty=2,col="blue")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(venti_ventiquattro_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_20-24",verticals =
 TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(venticinque_ventinove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_25-29",vertical
s = TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(trenta_trentaquattro_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_30-34",verticals
 = TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(trentacinque_trentanove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_35-39",vertic
als = TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(quaranta_quarantaquattro_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_40-44",verti
cals = TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(quarantacinque_quarantanove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_45-49",ve
rticals = TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
> plot(ecdf(quindici_quarantanove_anni),main="Funzione_di_distribuzione_empirica\n_discreta_di_15-49",vertical
s = TRUE, col="red")
> abline(h=0.25,lty=1,col="blue")
> abline(h=0.5,lty=2,col="blue")
> abline(h=0.75,lty=3,col="blue")
```

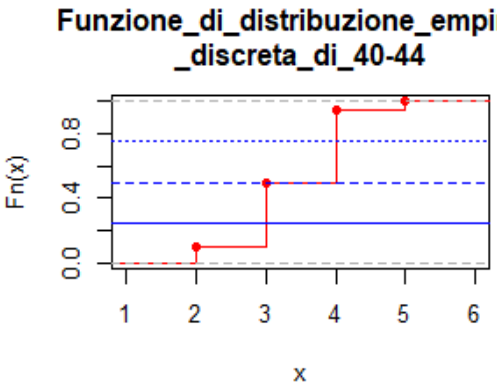

Funzione_di_distribuzione_empirica
_discreta_di_15-19
Fn(x)
0.0
0.4
0.8
1
2
3
4
5
6
7
8
x

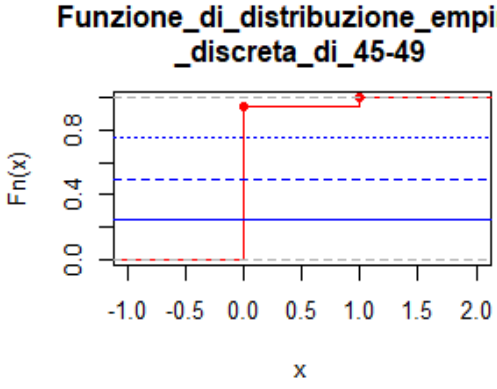

Funzione_di_distribuzione_empirica
_discreta_di_20-24
Fn(x)
0.0
0.4
0.8
6
8
10
12
14
16
x

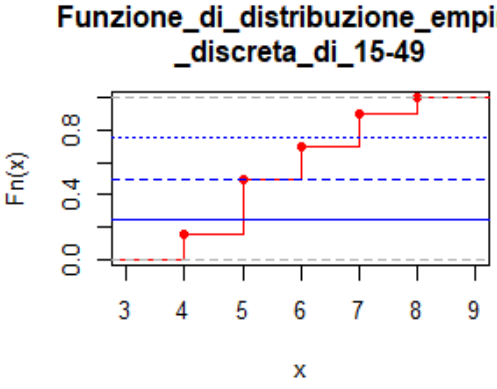

Funzione_di_distribuzione_empirica
_discreta_di_25-29
Fn(x)
0.0
0.4
0.8
6
8
10
12
14
16
x


Funzione_di_distribuzione_empirica
_discreta_di_30-34
Fn(x)
0.0
0.4
0.8
6
8
10
12
14
x


Funzione_di_distribuzione_empirica
_discreta_di_35-39
Fn(x)
0.0
0.4
0.8
5
6
7
8
9
10
11
12
x


Funzione_di_distribuzione_empirica
_discreta_di_40-44
Fn(x)
0.0
0.4
0.8
1
2
3
4
5
6
x


Funzione_di_distribuzione_empirica
_discreta_di_45-49
Fn(x)
0.0
0.4
0.8
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
x


Funzione_di_distribuzione_empirica
_discreta_di_15-49
Fn(x)
0.0
0.4
0.8
3
4
5
6
7
8
9
x

Variazione, deviazione standard e coefficiente di variazione

Gli indici di posizione non tengono conto delle variabilità esistente tra i dati; infatti esistono distribuzioni di frequenza, che pur avendo la stessa media campionaria, sono molto diverse tra di loro. Indici significativi per misurare la variabilità di una distribuzione di frequenza sono la *varianza campionari*a e la *deviazione standard.*

La varianza campionaria e la deviazione standard campionaria, detta anche scarto quadratico medio campionario, sono indici utili per misurare la variabilità di una distribuzione di frequenza.

Definzione varianza campionaria: Assegnato un insieme di dati numerici x1, x2, . . . , xn, si definisce varianza campionaria, e si denota con s^2, la quantità:

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

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

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

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

In R la varianza campionaria di un vettore numeri si calcola utilizzando la funzione var() e la deviazione standard campionaria si calcola utilizzando la funzione sd().

```
> var(quindici_diciannove_anni)
[1] 1.523684
> var(venti_ventiquattro_anni)
[1] 5.397661
> var(venticinque_ventinove_anni)
[1] 6.063158
> var(trenta_trentaquattro_anni)
[1] 4.260526
> var(trentacinque_trentanove_anni)
[1] 2.786842
> var(quaranta_quarantaquattro_anni)
[1] 0.5763158
> var(quarantacinque_quarantanove_anni)
[1] 0.05
> var(quindici_quarantanove_anni)
[1] 1.565789
> sd(quindici_diciannove_anni)
[1] 1.234376
> sd(venti_ventiquattro_anni)
[1] 2.323287
> sd(venticinque_ventinove_anni)
[1] 2.462348
> sd(trenta_trentaquattro_anni)
[1] 2.064104
> sd(trentacinque_trentanove_anni)
[1] 1.669384
> sd(quaranta_quarantaquattro_anni)
[1] 0.7591547
> sd(quarantacinque_quarantanove_anni)
[1] 0.2236068
> sd(quindici_quarantanove_anni)
[1] 1.251315
```

La media campionaria e la varianza campionaria sono i due indici di posizione e di dispersione dei dati maggiormente utilizzati. I valori della varianza campionaria e della deviazione standard campionaria dipendono dall'unità di misura dei dati. In particolare, 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 confrontare le variazioni esistenti tra diversi campioni di dati è utile introdurre *coefficiente di variazione.*

Definizione coefficiente di variazione: Assegnato un insieme di dati numerici x1, x2, . . . , xn, 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, poichè la media campionaria e la deviazione standard campionaria sono espressi in identiche unità di misura. Segue che il coefficiente di variazione è un indice di dispersione che ha senso soltanto per campioni aventi la media campionaria non nulla.

In R non è definita una funzione che calcola il coefficiente di variazione. Tale funzione può essere comunque facilmente implementate in R nel seguente modo:

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

Ed è stata applicata tale formula ai nostri otto vettori.

```
> cv(quindici_diciannove_anni)
[1] 0.3125003
> cv(venti_ventiquattro_anni)
[1] 0.2522425
> cv(venticinque_ventinove_anni)
[1] 0.2414067
> cv(trenta_trentaquattro_anni)
[1] 0.2184237
> cv(trentacinque_trentanove_anni)
[1] 0.2099854
> cv(quaranta_quarantaquattro_anni)
[1] 0.2200448
> cv(quarantacinque_quarantanove_anni)
[1] 4.472136
> cv(quindici_quarantanove_anni)
[1] 0.21762
```

Nella seguente tabella sono stati riportati la media, la varianza, la deviazione standard e il coefficiente di variazione dei vettori delle otto fasce di età.

	15-19 anni	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
media	3.95	9.210526	10.2	9.45	7.95	3.45	0.05	5.75
varianza	1.523684	5.397661	6.063158	4.260526	2.786842	0.5763158	0.05	1.565789
deviazione standard	1.234376	2.323287	2.462348	2.064104	1.669384	0.7591547	0.2236068	1.251315
coefficiente di variazione	0.3125003	0.2522425	0.2414067	0.2184237	0.2099854	0.2200448	4.472136	0.21762

La media campionaria nel nostro caso è differente per tutti i vettori e il coefficiente di variazione più alto si ottiene per il vettore "45-49_anni".

2.2 Forma di una distribuzione di frequenza

La media, la mediana e la moda sono utili a comprendere la forma delle distribuzioni di frequenza, nel senso che differenze sostanziali tra questi indici indicano uno *sbilanciamento eccessivo della distribuzione di frequenza verso destra o verso sinistra.* Inoltre è stato dimostrato che anche se la media e la mediana coincidono, le misure di dispersione dei dati possono essere sostanzialmente differenti implicando una differente forma delle distribuzioni di frequenza.
Esistono degli indici statistici che permettono di misurare quando una distribuzione di frequenza presenta *simmetria o asimmetria* oppure se essa è *più o meno piccata*.
Un indice che permette di misurare la simmetria di una distribuzione è la *skewness campionaria*.

Definizione skewness campionaria: Assegnato un insieme di dati numerici x1, x2, . . . , xn, si definisce skewness campionaria il valore:

$$\gamma_1 = \frac{m_3}{m_2^{3/2}}$$

dove m_3 denota il momento centrato campionario di ordine 3. In generale, il momento centrato campionario di ordine j è così definito:

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^j \qquad \text{(j=1,2,…)}$$

Se γ_1 =0, allora la distribuzione di frequenza è simmetrica, mentre se γ_1>0 c'è un'asimmetria positiva (ossia la distribuzione di frequenza ha la coda di destra più allungata), infine se γ_1 <0 c'è un'asimmetria negativa (ossia la distribuzione di frequenza ha la coda di sinistra più allungata).
Si nota che γ_1 è un indice adimensionale, ossia è indipendente dall'unità di misura dei dati.

Il calcolo della skewness campionaria può essere così implementato in R:

```
> skw<-function(x){
+   n<-length(x)
+   m2<-(n-1)*var(x)/n
+   m3<-(sum((x-mean(x))^3))/n
+   m3/(m2^1.5)
+ }
```

Riferendoci ai dati degli otto vettori che indicano fasce di età, si nota che

```
> skw(quindici_diciannove_anni)
[1] 0.6128276
> skw(venti_ventiquattro_anni)
[1] 0.9124037
> skw(venticinque_ventinove_anni)
[1] 0.4366319
> skw(trenta_trentaquattro_anni)
[1] 0.8467169
> skw(trentacinque_trentanove_anni)
[1] 0.4288167
> skw(quaranta_quarantaquattro_anni)
[1] -0.1980929
> skw(quarantacinque_quarantanove_anni)
[1] 4.129483
> skw(quindici_quarantanove_anni)
[1] 0.3203889
```

che mostra che la skewness presenta un'asimmetria negativa per il vettore "quaranta_quarantaquattro_anni" e presenta un'asimmetria positiva per i rimanenti sette vettori.

Un indice che permette di misurare la densità dei dati intorno alla media è la *curtosi campionaria.*
Definizione curtosi campionaria: Assegnato un insieme di dati numerici x1, x2, . . . , xn, si definisce curtosi campionaria il valore:

$$\beta_2 = \frac{m_4}{m_2^2}$$

dove m4 denota il momento centrato campionario di ordine 4.

Questo indice ci permette di *confrontare la distribuzione di frequenza dei dati con una densità di probabilità normale* che è caratterizzata da curtosi uguale a 3.
Se la distribuzione di frequenza è *più piatta di una normale(platicurtica)* tale indice è inferiore a 3, mentre se è *più piccata di una normale (leptocurica)* tale indice è maggiore di 3. Se invece tale indice è uguale a 3 la distribuzione di frequenza si definisce *normocurtica*, ossia piatta come una normale.
Il calcolo della curtosi campionaria può essere implementato in R nel seguente modo:

```
> curt<-function(x){
+   n<-length(x)
+   m2<-(n-1)*var(x)/n
+   m4<-(sum((x-mean(x))^4))/n
+   m4/(m2^2)
+ }
```

Riferendoci ai dati degli otto vettori che indicano le fasce di età, si nota che

```
> curt(quindici_diciannove_anni)
[1] 3.360797
> curt(venti_ventiquattro_anni)
[1] 3.066111
> curt(venticinque_ventinove_anni)
[1] 1.777922
> curt(trenta_trentaquattro_anni)
[1] 2.495427
> curt(trentacinque_trentanove_anni)
[1] 2.051535
> curt(quaranta_quarantaquattro_anni)
[1] 2.629574
> curt(quarantacinque_quarantanove_anni)
[1] 18.05263
> curt(quindici_quarantanove_anni)
[1] 2.065179
```

che mostra che la curtosi è minore di 3 per il terzo,quarto,quinto,sesto e ottavo vettore (la distribuzione di frequenza è più piatta di una normale) ed è maggiore di 3 per il primo,secondo e settimo vettore (la distribuzione di frequenza è più piccata di una normale).

3 STATISTICA DESCRITTIVA BIVARIATA

Considerando il seguente data frame

```
> Aborto<-data.frame(quindici_diciannove_anni=c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),venti_ventiquattro_anni=c(1
3,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,6,8,7),venticinque_ventinove_anni=c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,1
2,9,7,8,8),trenta_trentaquattro_anni=c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8),trentacinque_trentanove_anni=
c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),quaranta_quarantaquattro_anni=c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3,5,4,3,3
,3),quarantacinque_quarantanove_anni=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),quindici_quarantanove_anni=c(7,6,8,6
,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
```

è stata inoltre sviluppata la statistica descrittiva bivariata, ossia il ramo della statistica che si occupa dei metodi grafici e statistica atti a descrivere le relazioni che intercorrono tra due variabili. Sono state considerate due variabili “quindici_diciannove_anni” e “venti_ventiquattro_anni” di tipo quantitativo. Successivamente calcoliamo gli indici statistici di posizione e di dispersione relativi alle singole variabili.

```
> median(Aborto$quindici_diciannove_anni)
[1] 4
> mean(Aborto$quindici_diciannove_anni)
[1] 3.95
> sd(Aborto$quindici_diciannove_anni)
[1] 1.234376
>
> median(Aborto$venti_ventiquattro_anni)
[1] 8
> mean(Aborto$venti_ventiquattro_anni)
[1] 9.05
> sd(Aborto$venti_ventiquattro_anni)
[1] 2.37254
```

	quindici_diciannove_anni	**venti_ventiquattro_anni**
Mediana campionaria	4	8
Media campionaria	3.95	9.05
Deviazione standard	1.234376	2.37254

Successivamente realizziamo lo scatterplot considerando "quindici_diciannove_anni" come variabile indipendente e "venti_ventiquattro_anni" come variabile dipendente; nello scatterplot sono visualizzate le 20 coppie del data frame "Aborto". Tracciamo anche nello scatterplot delle linee orizzontali e verticali in corrispondenza delle mediane campionarie e delle medie campionarie dei due vettori. Le seguenti linee di codice

```
> plot(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni,main = "20-24_ann
i_in_funzione_di_15-19_anni",xlab = "15-19_anni",ylab = "20-24_anni",col="red")
> abline(v=median(Aborto$quindici_diciannove_anni),lty=1,col="magenta")
> abline(v=mean(Aborto$quindici_diciannove_anni),lty=2,col="blue")
> abline(h=median(Aborto$venti_ventiquattro_anni),lty=1,col="magenta")
> abline(h=mean(Aborto$venti_ventiquattro_anni),lty=2,col="blue")
> legend(18,30,c("Mediana","Media"),pch = 0,col = c("magenta","blue"),cex=0.8)
```

producono il diagramma di dispersione (scatterplot) di Figura 3.1. Si nota che i dati sono posizionati attorno ad una retta ascendente e ciò induce a pensare che esista una correlazione lineare positiva tra le variabili.

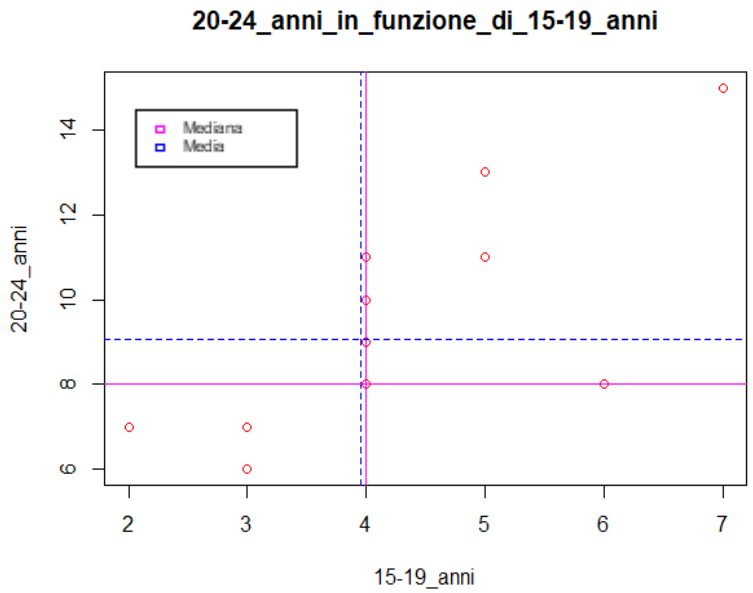


Figura 3.1:Scatterplot dei vettori "quindici_diciannove_anni" e "venti_ventiquattro_anni"

3.1 Covarianza e correlazione campionaria

Per ottenere una misura quantitativa della correlazione tra le variabili si considera la *covarianza campionaria*.

Definizione covarianza campionaria: Assegnato un campione bivariato (x1, y1), (x2, y2), . . . , (xn, yn) di una variabile quantitativa bidimensionale (X, Y), siano $\bar{x}$ e $\bar{y}$ rispettivamente le medie campionarie di x1, x2, . . . , xn e di y1, y2, . . . , yn. La covarianza campionaria tra le due variabili X e Y è così definita:

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

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

Per ottenere una misura quantitativa della correlazione tra le variabili si può anche considerare il *coefficiente di correlazione campionario*.

Definizione coefficiente di correlazione campionario: Assegnato un campione bivariato (x1, y1), (x2, y2), . . . , (xn, yn) di una variabile quantitativa bidimensionale (X, Y), siano $\bar{x}$ e s_x la media campionaria e la deviazione standard campionaria di x1, x2, . . ., xn ed inoltre siano $\bar{y}$ e s_y la media campionaria e la deviazione standard campionaria di y1, y2, . . . , yn. Il coefficiente di correlazione campionario tra le due variabili X e Y è così definito:

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

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

Occorre ricordare che il coefficiente di correlazione campionario r_{xy} misura la *forza del legame di natura lineare esistente tra le due variabili quantitative*. Eventuali relazioni tra le variabili che assumono una forma curvilinea non possono pertanto essere individuati con tale coefficiente. Riassumendo si può dire che il segno di r_{xy} indica la *direzione della retta interpolante* e indica la presenza di una tra le seguenti relazioni:
- $r_{xy} = 1$ (correlazione perfetta positiva) tutti i punti sono allineati su una retta ascendente;
- r_{xy} compreso tra 0 e 1 estremi esclusi (correlazione positiva) i punti sono posizionati in una nuvola attorno ad una *linea retta interpolante ascendente*;
- $r_{xy} = 0$ (nessuna correlazione) i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare;
- r_{xy} compreso tra -1 e 0 estremi esclusi (correlazione negativa) i punti sono posizionati in una nuvola attorno ad una *linea retta interpolante discendente*;
- $r_{xy} = -1$ (correlazione perfetta negativa) tutti i punti sono allineati su una linea retta discendente.

Nel linguaggio R le covarianze campionarie e le correlazioni campionarie fra una coppia X e Y di variabili numeriche possono essere immediatamente ottenute con le rispettive funzioni cov(X,Y) e cor(X,Y).

```
> cov(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni)
[1] 2.265789
> cor(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni)
[1] 0.7736748
```

I dati dei due vettori "quindici_diciannove_anni" e "venti_ventiquattro_anni" sono positivamente correlati essendo la covarianza campionaria uguale a 2.265789, ossia i valori assunti dal primo e dal secondo vettore tendono ad essere grandi e piccoli insieme.

Inoltre il coefficiente di correlazione è uguale a 0.7736748. Ciò indica, come è evidenziato nel diagramma di dispersione (scatterplot) di Figura 3.2, che esiste una forte correlazione lineare tra la "quindici_diciannove_anni" e "venti_ventiquattro_anni".

Lo scatterplot di Figura 3.2, è stato ottenuto con le seguenti linee di codice:

```
> plot(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni,main = "Retta_di_regressione
",xlab = "15-19_anni",ylab = "20-24_anni",col="red")
> abline(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni),col="blue")
```

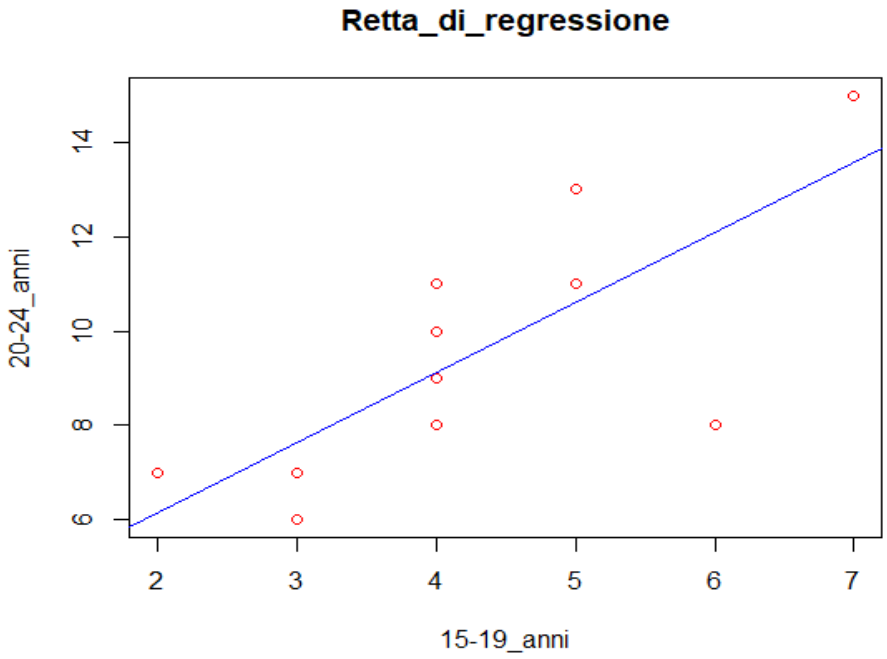


Figura 3.2:Scatterplot retta interpolante stimata (retta di regressione)

La funzione abline(lm(progr2~progr1)) permette di aggiungere allo scatterplot relativo ai valori dei due vettori presi in esame la linea interpolante stimata. Gli scatterplot sono dei potenti mezzi per visualizzare le eventuali relazioni che possono intercorrere tra variabili quantitative. Infatti, alcune volte osservando il grafico si nota che i punti si dispongono attorno a qualche linea orientata in qualche direzione, come ad esempio si verifica in Figura 3.2. Tuttavia per avere un'idea più accurata del fenomeno, è necessario utilizzare altre tecniche statistiche in grado di misurare con maggiore precisione questo legame.

Il modello lineare viene di solito utilizzato per spiegare, descrivere, o anche prevedere un andamento futuro sulla base della relazione che si instaura tra una variabile Y, chiamata variabile dipendente, e una o più altre variabili che assumono il significato di variabili indipendenti X1,X2, . . . ,Xp. Nel caso in cui p = 1, l'analisi prende il nome di regressione semplice, mentre se p = 2, 3 . . . si parla di regressione multipla. Per poter utilizzare un modello di regressione e fondamentale individuare in primo luogo quale è la variabile indipendente e quali sono le variabili dipendenti.

3.2 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 le altre possibili rette.

Consideriamo l'equazione della retta: $Y = \alpha + \beta X$ dove a è l'intercetta e β è il coefficiente angolare.
Il *coefficiente angolare* β esprime quantitativamente la *pendenza (inclinazione) della retta.*
L'*intercetta* α invece corrisponde all'ordinata dei punti 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.*
Le medie campionarie, le deviazioni standard e il coefficiente di correlazione permettono di stimare i parametri a e b della retta di regressione.
I coefficienti di regressione sono i valori α e β per i quali la somma Q dei quadrati degli errori

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

sia minima. Con il metodo dei minimi quadrati si giunge a :

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

Nel nostro esempio, riferendoci ai vettori "quindici_diciannove_anni" e "venti_ventiquattro_anni" le seguenti linee di codice calcolano il coefficiente angolare β e l'intercetta α.

```
> beta<-(sd(Aborto$venti_ventiquattro_anni)/sd(Aborto$quindici_diciannove_anni))*cor(Aborto$quindici_
diciannove_anni,Aborto$venti_ventiquattro_anni)
> alpha<-mean(Aborto$venti_ventiquattro_anni)-beta*mean(Aborto$quindici_diciannove_anni)
> c(alpha,beta)
[1] 3.176166 1.487047
```

Inoltre è stata utilizzata la funzione lm(y~x) per eseguire le analisi di regressione lineari ed è stata utilizzata la funzione abline(lm(y~x)) che permette di aggiungere la retta di regressione al grafico dello scatterplot, con "venti_ventiquattro_anni", variabile dipendente, che dipende da "quindici_diciannove_anni", variabile indipendente.

Ad esempio le seguenti linee di codice

```
> lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni)

Call:
lm(formula = Aborto$venti_ventiquattro_anni ~ Aborto$quindici_diciannove_anni)

Coefficients:
                          (Intercept)
                                3.176
Aborto$quindici_diciannove_anni
                                1.487
```

mostrano che la retta di regressione ha intercettato α = 3.176 e coefficiente angolare β = 1.487.
La retta di regressione ha quindi equazione y = 3.176 + 1.487 x.

Residui

Una volta calcolati i coefficienti α e β, e dopo aver trovato la retta di regressione, possiamo vedere quanto i punti appartenenti alla nuvola di punti sia distante dalla retta di regressione. Per vedere ciò possiamo calcolare degli scostamenti (residui) tra le ordinate dei punti y_i e i corrispondenti valori stimati

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

Ottenuti mediante la retta di regressione e presenti sulla retta di regressione.
I residui $E_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i)$(i=1,2,…,n) mostrano quanto si discostano i valori osservati da quelli stimati con l'aiuto della retta di regressione.
La mediana, la varianza campionaria e la deviazione standard campionaria dei residui sono:

```
> linearmodel<-lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni)
> median(linearmodel$residuals)
[1] 0.1321244
> var(linearmodel$residuals)
[1] 2.259613
> sd(linearmodel$residuals)
[1] 1.503201
```

Non si può invece calcolare il coefficiente di variazione, essendo la media campionaria dei residui nulla, poiché in media gli scostamenti positivi e negativi si compensano.

➡ ***Segmenti che congiungono i valori stimati e i valori osservati***

È stato realizzato il grafico dei residui ottenuto aggiungendo, al grafico contenente lo scatterplot e la retta di regressione, dei segmenti verticali che visualizzano i residui.
Le seguenti linee di codice

```
> plot(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni,main = "Retta_di_regressione
_e_residui",xlab = "15-19_anni",ylab = "20-24_anni",col="red")
> abline(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni),col="blue")
> stime<-fitted(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni))
> segments(Aborto$quindici_diciannove_anni,stime,Aborto$quindici_diciannove_anni,Aborto$venti_vent
iquattro_anni,col="magenta")
```

hanno prodotto il grafico illustrato in Figura 3.3.

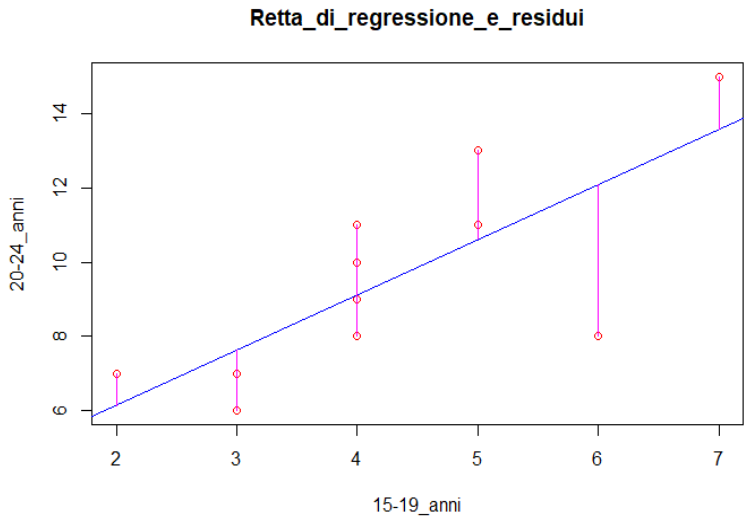


Figura 3.3: Grafico dei residui (tra valori osservati e valori stimati con la retta di regressione).

➡ ***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 di dispongano intorno alla retta interpolante influenzandone la posizione, può essere ottenuto attraverso il *diagramma dei residui*.
Si è desiderato realizzare un diagramma dei residui ponendo i valori dei residui sull'asse delle ordinate e quelli della variabile indipendente sull'asse delle ascisse.
Le seguenti linee di codice

```
> residui<-resid(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni))
> plot(Aborto$quindici_diciannove_anni, residui, main="Diagramma_dei_residui",xlab = "15-19_anni",
ylab = "Residui",pch=9, col="red")
> abline(h=0, col="blue", lty=2)
```

hanno prodotto il grafico in Figura 3.4. I punti indicano la posizione dove si collocano i residui rispetto ai valori del vettore "quindici_diciannove_anni".

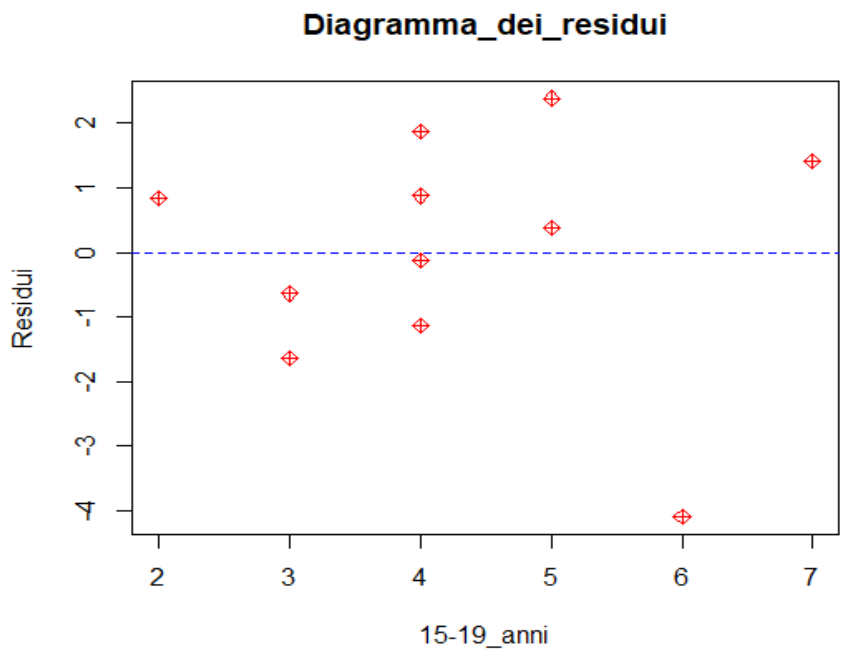


Figura 3.4: Residui in funzione dei valori del vettore "quindici_diciannove_anni"

La retta orizzontale è posizionata nello zero e corrisponde alla media campionaria dei residui. Si nota che i punti che sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia nessun comportamento nella distribuzione dei punti.
Il diagramma dei residui aiuta a comprendere quale è l'andamento della retta di regressione rispetto ai dati.
La presenza di valori anomali può ulteriormente influenzare la direzione e la collocazione della retta di regressione. Con l'analisi dei residui si possono capire quali sono i valori anomali presenti nei dati.

➡ ***Valori dei residui standardizzati rispetto ai valori stimati***

È spesso interessante calcolare i *residui standardizzati* così definiti:

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

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

```
> residuistandard<-residui/sd(residui)
> plot(stime,residuistandard, main = "Residui_standard_rispetto_ai_valori_stimati",xlab = "Valori_
stimati", ylab = "Residui_standard", pch=5, col="red")
> abline(h=0, col="blue", lty=2)
```

ha prodotto il grafico in Figura 3.5. I punti indicano la posizione dove si collocano i residui standardizzati rispetto ai valori stimati con la retta di regressione. La retta orizzontale è posizionata nello zero, che corrisponde alla media campionaria dei residui standardizzati. Anche in questo caso i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenzia nessuna tendenza particolare nella distribuzione dei punti.

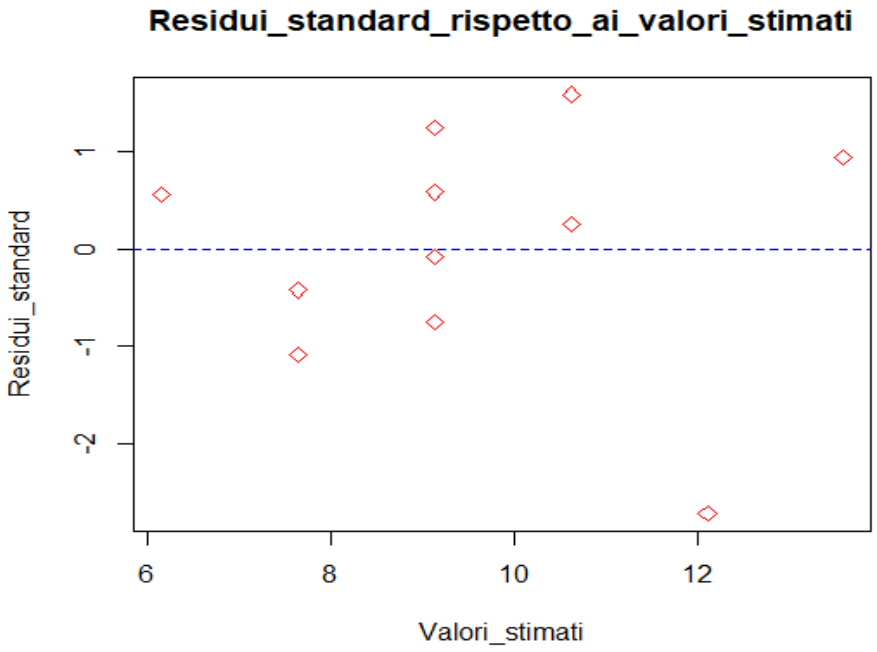


Figura 3.5: Residui in funzione dei valori stimati con la retta di regressione

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. E' chiaro che r_{xy}^2 molto vicino ad 1 indicherà che tutti i punti tenderanno ad allinearsi lungo la retta di regressione, mentre r_{xy}^2 prossimo a 0 esprime una completa incapacità della retta di rappresentare la distribuzione dei dati considerati.

Definizione Se si denota con (y1, y2, . . . , yn) il vettore dei dati della variabile dipendente, con 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, il coefficiente di determinazione (detto anche r-square) è così definito

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

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

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

Per calcolare tale indice in R possiamo utilizzare il quadrato del coefficiente di correlazione oppure summary(lm(y x))$r.square. Riferendoci al nostro caso si ha:

```
> (cor(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni))^2
[1] 0.5985728
> summary(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni))$r.square
[1] 0.5985728
```

E' possibile notare quindi che il coefficiente di determinazione, essendo nel caso di regressione semplice, è proprio uguale al quadrato del coefficiente di correlazione che era 0.7736748.

3.3 Regressione lineare multipla

In molte applicazioni dell'analisi di regressione sono coinvolte situazioni con più di una singola variabile indipendente. La regressione lineare multipla è utile per spiegare la relazione tra una variabile Y detta dipendente, e le variabili X1, X2,…, Xp dette indipendenti. Definendo un data frame con le p+1 variabili, ovvero Y X1,X2,..,Xp avremo che le funzioni R cov(dfm) e cor(dfm) restituiranno matrici di dimensioni (p+1)*(p+1) i cui elementi sono le covarianze e gli indici di correlazioni tra le variabili. In particolare, queste matrici sono simmetriche; la matrice delle covarianze contiene sulla diagonale principale la varianza delle singole colonne del data frame, mentre la matrice degli indici di correlazione contiene il numero 1 sulla diagonale principale; quest'ultima indica le correlazioni che ci sono tra le variabili.
Partendo sempre dal data frame iniziale

```
> Aborto<-data.frame(quindici_diciannove_anni=c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),venti_ventiquattro_anni=c(1
3,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,6,8,7),venticinque_ventinove_anni=c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,1
2,9,7,8,8),trenta_trentaquattro_anni=c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8),trentacinque_trentanove_anni=
c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),quaranta_quarantaquattro_anni=c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3,5,4,3,3
,3),quarantacinque_quarantanove_anni=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),quindici_quarantanove_anni=c(7,6,8,6
,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
```

stato utilizzato il modello di regressione lineare multipla per spiegare la relazione tra la variabile quantitativa "venti_ventiquattro_anni", considerata come variabile dipendente,
e tutte le altre variabili dette variabili indipendenti.
È stato definito un data frame e sono state utilizzate le funzioni cov() e cor() fornendo due matrici con le covarianze e le correlazioni tra coppie di variabili.

È stata visualizzata la matrice delle covarianze campionarie i cui valori sono riportati nella seguente Tabella

```
> cov(Aborto)
```

cov	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
15-19 anni	1.5236842	2.26578947	2.484211	1.86578947	1.576315789	0.60263158	0.107894737	1.25000000
20-24 anni	2.2657895	5.62894737	4.936842	4.60789474	3.476315789	1.02894737	-0.055263158	2.69736842
25-29 anni	2.4842105	4.93684211	6.063158	4.16842105	3.273684211	1.06315789	0.200000000	2.68421053
30-34 anni	1.8657895	4.60789474	4.168421	4.26052632	3.023684211	0.99736842	-0.023684211	2.32894737
35-39 anni	1.5763158	3.47631579	3.273684	3.02368421	2.786842105	0.97105263	0.002631579	1.88157895
40-44 anni	0.6026316	1.02894737	1.063158	0.99736842	0.971052632	0.57631579	0.028947368	0.69736842
45-49 anni	0.1078947	-0.0552631 6	0.200000	-0.0236842 1	0.002631579	0.02894737	0.050000000	0.01315789
15-49 anni	1.2500000	2.69736842	2.684211	2.32894737	1.881578947	0.69736842	0.013157895	1.56578947

Mentre i valori ottenuti calcolando la matrice delle correlazioni sono riportati nella seguente tabella.

```
> cor(Aborto)
```

cor	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
15-19 anni	1.0000000	0.7736748	0.8173188	0.73229072	0.764961469	0.6430935	0.390901894	0.8092744 8
20-24 anni	0.7736748	1.0000000	0.8450574	0.94093003	0.877706764	0.5712804	-0.104168666	0.9085733 4
25-29 anni	0.8173188	0.8450574	1.0000000	0.82014476	0.796399838	0.5687456	0.363241579	0.8711650 7
30-34 anni	0.7322907	0.9409300	0.8201448	1.00000000	0.877503029	0.6364931	-0.051314758	0.9016985 3
35-39 anni	0.7649615	0.8777068	0.7963998	0.87750303	1.000000000	0.7662250	0.007049774	0.9007402 7
40-44 anni	0.6430935	0.5712804	0.5687456	0.63649314	0.766225019	1.0000000	0.170527265	0.7341170 9

45-49 anni	0.3909019	-0.1041687	0.3632416	-0.05131476	0.007049774	0.1705273	1.000000000	0.04702564
15-49 anni	0.8092745	0.9085733	0.8711651	0.90169853	0.900740269	0.7341171	0.047025641	1.00000000

Si nota che esiste una forte correlazione lineare tra tutte le coppie di variabili ed è anche alta la correlazione lineare tra il vettore "venti_ventiquattro_anni" e tutte le altre variabili.

Il modello di regressione lineare multipla con p variabili indipendenti è esprimibile attraverso l'equazione:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta p \, Xp,$$

dove

-α è l'intercetta, ossia il valore di Y quando $X_1 = X_2 = \cdots = X_p = 0$

-β1, β2, . . . βp sono i regressori , dove β1 indica l'inclinazione di Y rispetto a X1, e così via.

L'equazione si ottiene mediante le seguenti linee di codice

```
> lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+Aborto$venticinque_ventinove_anni+A
borto$trenta_trentaquattro_anni+Aborto$trentacinque_trentanove_anni+Aborto$quaranta_quarantaquattro_ann
i+Aborto$quarantacinque_quarantanove_anni+Aborto$quindici_quarantanove_anni)

Call:
lm(formula = Aborto$venti_ventiquattro_anni ~ Aborto$quindici_diciannove_anni +
    Aborto$venticinque_ventinove_anni + Aborto$trenta_trentaquattro_anni +
    Aborto$trentacinque_trentanove_anni + Aborto$quaranta_quarantaquattro_anni +
    Aborto$quarantacinque_quarantanove_anni + Aborto$quindici_quarantanove_anni)

Coefficients:
                              (Intercept)          Aborto$quindici_diciannove_anni
                                 -0.90585                                  0.68248
      Aborto$venticinque_ventinove_anni          Aborto$trenta_trentaquattro_anni
                                  0.54417                                  0.36011
    Aborto$trentacinque_trentanove_anni      Aborto$quaranta_quarantaquattro_anni
                                  0.05522                                 -0.08133
Aborto$quarantacinque_quarantanove_anni         Aborto$quindici_quarantanove_anni
                                 -4.46537                                 -0.28325
```

Il modello di regressione multipla stimato è

y = -0.90585 + 0.68248x1 + 0.54417x2 + 0.36011x3 + 0.05522x4 - 0.08133x5 – 4.46537x6 – 0.28325 x7.

Osserviamo che il segno dei regressori β_5 , β_6, β_7 sono negativi; quindi "quaranta_quarantaquattro_anni" , "quarantacinque_quarantanove_anni" e "quindici_quarantanove_anni" hanno un effetto negativo su "venti_ventiquattro_anni", cioè all'aumentare di questi vettori diminuisce "venti_ventiquattro_anni", cosa contraria accade per i regressori $\beta_1, \beta_2, \beta_3, \beta_4$ che sono positivi.

3.3.1 Residui

Una volta calcolati i valori dei coefficienti è stato possibile osservare gli scostamenti (*residui*) tra le ordinate dei punti Y_1 (valori osservati) e i corrispondenti valori stimati

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

ottenuti mediante la regressione lineare multipla. I residui

$$E_i = y_i - by_i = y_i - (\alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p}) \qquad (i=1,2,\ldots,n)$$

mostrano di quanto si discostano i valori osservati dai valori stimati con la retta di regressione. Anche per la regressione lineare multipla utilizziamo per calcolare i valori stimati e i residui le funzioni fitted e resid. I residui multipli sono:

```
> residuimult<-resid(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+Aborto$venticinque_ventinove_anni+Aborto$trenta_trentaquattro_anni+Aborto$trentacinque_trentanove_anni+Aborto$quaranta_quarantaquattro_anni+Aborto$quarantacinque_quarantanove_anni+Aborto$quindici_quarantanove_anni))
> residuimult
            1             2             3             4             5             6             7
 4.936517e-01  1.249001e-16 -9.172814e-02  9.066429e-02  2.709709e-01  7.124064e-01 -1.447385e-01
            8             9            10            11            12            13            14
-4.637631e-01  1.356294e-01 -1.486202e-01 -3.123231e-01  2.285895e-01  6.204428e-01  1.862548e-01
           15            16            17            18            19            20
-1.021534e-01 -2.927159e-01 -5.649975e-01 -7.860073e-01  2.153696e-01 -4.693232e-02
```

La mediana, la varianza campionaria e la deviazione standard dei residui sono:

```
> multiplelinearmodel<-(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+Aborto$venticinque_ventinove_anni+Aborto$trenta_trentaquattro_anni+Aborto$trentacinque_trentanove_anni+Aborto$quaranta_quarantaquattro_anni+Aborto$quarantacinque_quarantanove_anni+Aborto$quindici_quarantanove_anni))
> median(multiplelinearmodel$residuals)
[1] -0.02346616
> var(multiplelinearmodel$residuals)
[1] 0.1457345
> sd(multiplelinearmodel$residuals)
[1] 0.3817519
```

È poi stato possibile realizzare un grafico in cui i residui standardizzati (ordinate) vengono disegnati in funzione dei valori stimati (ascisse) con il metodo dei minimi quadrati. Il seguente codice

```
> stimemult<-fitted(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+Aborto$venticinque_ventinove_anni+Aborto$trenta_trentaquattro_anni+Aborto$trentacinque_trentanove_anni+Aborto$quaranta_quarantaquattro_anni+Aborto$quarantacinque_quarantanove_anni+Aborto$quindici_quarantanove_anni))
> residuimult<-resid(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+Aborto$venticinque_ventinove_anni+Aborto$trenta_trentaquattro_anni+Aborto$trentacinque_trentanove_anni+Aborto$quaranta_quarantaquattro_anni+Aborto$quarantacinque_quarantanove_anni+Aborto$quindici_quarantanove_anni))
> residuimultstandard<-residuimult/sd(residuimult)
> plot(stimemult, residuimultstandard,main="Residui_standard_rispetto_ai_valori_stimati",xlab="Valori_stimati",ylab="Residui_standard",pch=5,col="red")
> abline(h=0,col="blue",lty=2)
```

ha prodotto il grafico in Figura 3.7. I punti indicano dove si collocano i residui standardizzati rispetto ai valori stimati con la retta di regressione. La retta orizzontale è posizionata nello zero, che corrisponde alla media campionaria dei residui standardizzati. Anche in questo caso i punti sono disposti quasi casualmente attorno alla linea orizzontale e non si evidenza nessuna tendenza particolare nella distribuzione dei punti.

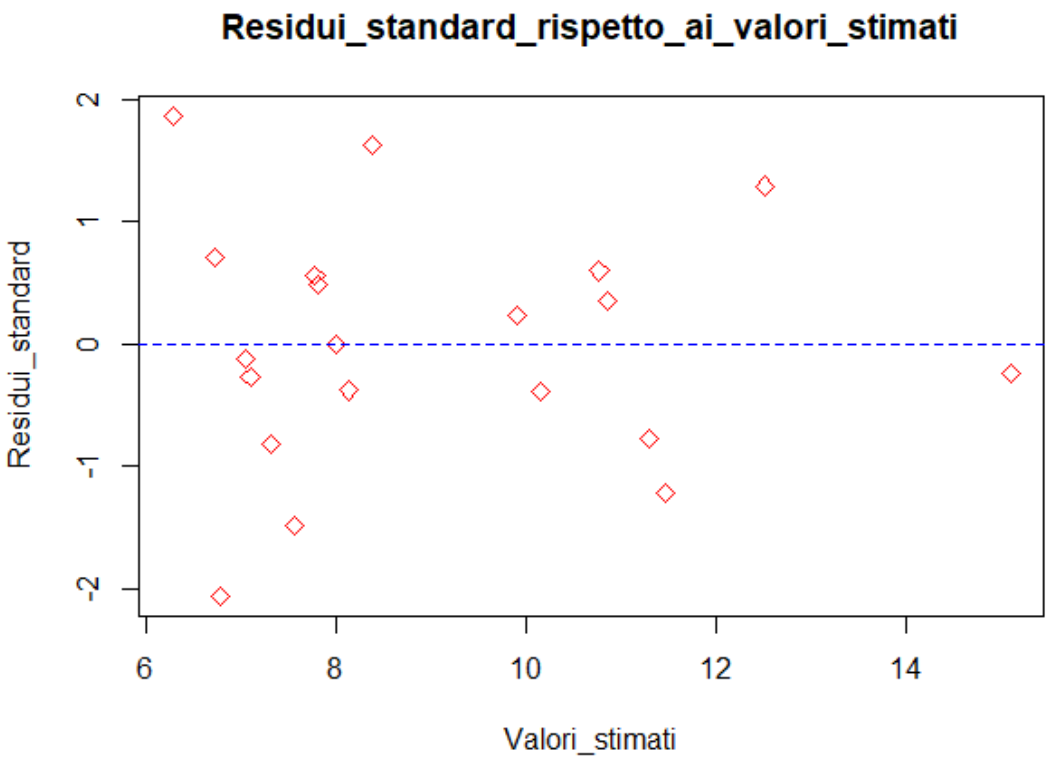


Figura 3.7: Residui in funzione dei valori stimati

Coefficiente di determinazione

Il coefficiente di determinazione di un modello di regressione lineare multipla è il rapporto tra la varianza dei valori stimati tramite la funzione di regressione e la varianza i valori osservati della variabile dipendente. Se si denota con (y1, y2, . . . , yn) 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 funzione di regressione, il coefficiente di determinazione è:

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

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 basta utilizzare
summary(lm(y˜x1 + x2 + . . . xp))$r.square. Riferendoci al nostro esempio calcoliamo D^2 in due diversi modi:

```
> num<-sum((stimemult-mean(Aborto$venti_ventiquattro_anni))^2)
> den<-sum((Aborto$venti_ventiquattro_anni-mean(Aborto$venti_ventiquattro_anni))^2)
> d2<-num/den
> d2
[1] 0.9741098
> summary(lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+Aborto$venticinque_ventinove_anni+Aborto$trenta_trentaquat
tro_anni+Aborto$trentacinque_trentanove_anni+Aborto$quaranta_quarantaquattro_anni+Aborto$quarantacinque_quarantanove_anni+Aborto$quind
ici_quarantanove_anni))$r.square
[1] 0.9741098
```

Il coefficiente di determinazione è quindi 0.9741098, ossia il modello di regressione multipla utilizzato può spiegare significativamente i dati.

3.4 Regressione non lineare

Considerando, sempre, lo stesso data frame

```
> Aborto<-data.frame(quindici_diciannove_anni=c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),venti_ventiquattro_anni=c(1
3,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,6,8,7),venticinque_ventinove_anni=c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,1
2,9,7,8,8),trenta_trentaquattro_anni=c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8),trentacinque_trentanove_anni=
c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),quaranta_quarantaquattro_anni=c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3,5,4,3,3
,3),quarantacinque_quarantanove_anni=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),quindici_quarantanove_anni=c(7,6,8,6
,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
```

sono state considerate “venti_ventiquattro_anni” come variabile dipendente e “quindici_diciannove_anni” come variabile indipendente.

(i) ***Regressione polinomiale***

Si è visto se l’approssimazione polinomiale è adeguata per stimare i nostri dati. Per la stima dei parametri α, β e γ si può ricorrere alla regressione multipla

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

con regressori X_1 = X e $X_2 = X^2$.

Con R è facile stimare i parametri α, β, γ tramite la funzione lm(y˜x + I(x ∧ 2)) dove I() `e un identificatore di variabile e viene inserito quando si debbono effettuare operazioni matematiche (divisione, elevamento a potenza) nelle variabili della regressione.
Vediamo se l’approssimazione polinomiale è adeguata per stimare i nostri dati.
Dalle seguenti linee di codice

```
> pol2<-lm(Aborto$venti_ventiquattro_anni~Aborto$quindici_diciannove_anni+I((Aborto$quindici_diciannove
_anni)^2))
> pol2

Call:
lm(formula = Aborto$venti_ventiquattro_anni ~ Aborto$quindici_diciannove_anni +
    I((Aborto$quindici_diciannove_anni)^2))

Coefficients:
                           (Intercept)         Aborto$quindici_diciannove_anni
                               3.60987                                 1.26756
I((Aborto$quindici_diciannove_anni)^2)
                               0.02541
```

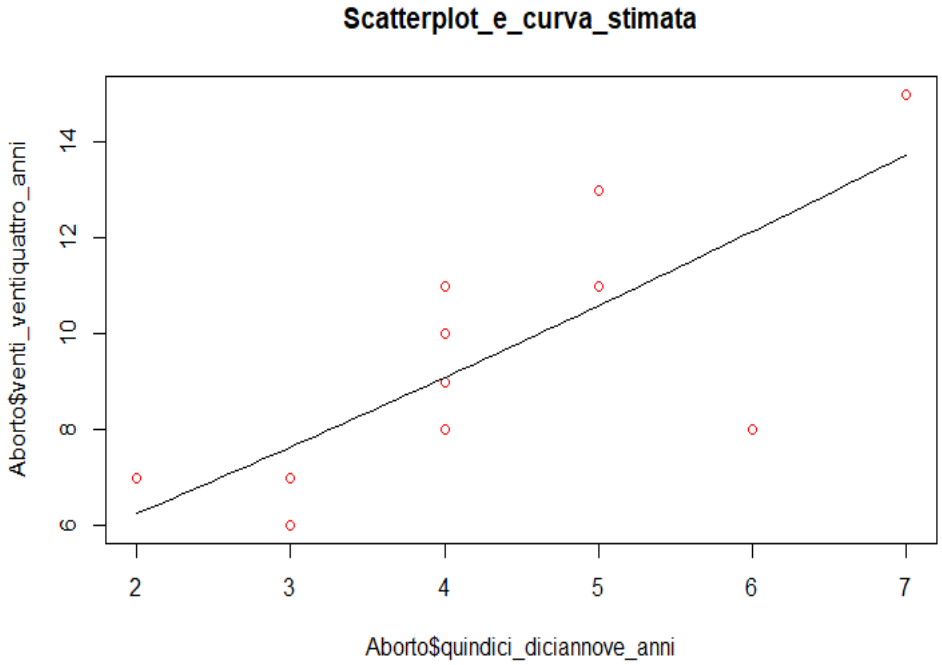
è stato possibile ricavare il modello di regressione che è
$Y = 3.60987 + 1.25756\ X + 0.02541X^2$

Per poter recuperare i parametri si è proceduto nel seguente modo

```
> alpha<-pol2$coefficients[[1]]
> beta<-pol2$coefficients[[2]]
> gamma<-pol2$coefficients[[3]]
```

Si è disegnata poi la curva stimata sullo scatterplot.
Le seguenti linee di codice

```
> plot(Aborto$quindici_diciannove_anni,Aborto$venti_ventiquattro_anni,col="red",main="Sc
atterplot_e_curva_stimata")
> curve(alpha+beta*x+gamma*x^2,add=TRUE)
```



4 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.
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. Gli individui che sono assegnati allo stesso cluster sono detti simili mentre gli individui che sono assegnati a differenti cluster sono detti dissimili.
Uno dei problemi che si presenta con l'analisi dei cluster riguarda la standardizzazione o meno delle variabili poiché attribuire un peso diverso a ciascuna caratteristica potrebbe condurre a risultati differenti circa la classificazione a seconda delle tecniche di clustering utilizzate. In molti metodi di clustering si raccomanda la *standardizzazione di ogni variabile* (caratteristica) usando la media campionaria e la deviazione standard campionaria entrambe derivate dall'insieme completo di individui della popolazione. Poichè in molte tecniche di clustering il tempo di calcolo cresce drammaticamente con il crescere del numero delle variabili, in più casi è desiderabile, prima di utilizzare tale analisi, ridurre il numero di variabili a quelle più direttamente collegate al fenomeno in esame. Un metodo che permette di effettuare tale riduzione delle variabili originarie è l'analisi delle componenti principali. Un problema simile sorge se le caratteristiche osservate sono correlate poiché qualora caratteristiche diverse sono correlate esse tendono a falsare i risultati ottenibili con l'analisi dei

cluster. In alcuni metodi di clustering si richiede che le variabili siano non correlate.
Per risolvere il problema di clustering è chiaramente desiderabile definire i termini *somiglianza* o *differenza* in modo quantitativo. La somiglianza tra due individui tra due individui I_ie I_j (i≠j) la si può definire mediante un coefficiente di similarità s_{ij} = s(X_i,X_j) oppure mediante una misura di distanza d_{ij}= d(X_i,X_j). Mentre i coefficienti di similarità assumono valori compresi tra 0 e 1, le misure di distanza possono assumere qualsiasi valore reale maggiore o uguale a zero. Quindi con questi due parametri possiamo capire come inserire due individui nello stesso cluster, cioè se il coefficiente di similarità è prossimo a 1 allora sono molto simili, oppure se la distanza tra i due individui è abbastanza piccola allora possono essere messi nello stesso cluster; Altrimenti se il coefficiente di similarità è prossimo a 0, oppure la distanza è molto grande, allora i due individui sono dissimili e quindi non possono essere messi nello stesso cluster.

4.1 Distanza e similarità

Le *misure metriche di somiglianza* sono soprattutto basate sulle *funzioni distanza* tra i vettori delle caratteristiche.

Definizione : Una funzione a valori reali d(Xi,Xj) è detta funzione distanza se e soltanto se essa soddisfa le seguenti condizioni:

(i) $d(X_i, X_j) = 0$ se e solo se Xi = Xj , con Xi e Xj in E_p;
(ii) $d(X_i, X_j) \geq 0$ per ogni Xi e Xj in E_p;
(iii) $d(X_i, X_j) = d(X_j, X_i)$ per ogni Xi e Xj in E_p;
(iv) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ per ogni Xi, Xj e Xk in E_p.

le distanze tra tutte le possibili coppie di unità sono inserite in una matrice simmetrica D di cardinalità n × n, dove dij indica la distanza d(Xi,Xj) con (i,j=1,2,…,n). Questa matrice è simmetrica e presenta tutti 0 sulla diagonale, e i valori uguali sono presenti due a due. Dato che per conoscere le distanze di ogni variabile rispetto alle n-1 variabili, basta conoscere n(n-1)/2 distanze della matrice, basta considerare solo la matrice triangolare superiore o inferiore della matrice D delle distanze.

In R la funzione: dist(X,method = ”euclidean”, diag = FALSE, upper = FALSE)
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).

e ritorna la matrice delle distanze D calcolata utilizzando le misure di distanza tra le righe della matrice X dei dati.

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 Minkowsky (minKowski);

(5) Distanza di Canberra (canberra);
(6) Distanza di Jaccard (binary);

Nella funzione dist() potrebbe essere presente anche un ulteriore parametro finale r che indica la potenza della matrice di Minkowski (di default è r=2, che corrisponde alla metrica euclidea).

Metrica Euclidea

La più familiare misura di distanza è la metrica Euclidea, cosi definita:

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

dove x_{ik} è il valore della k–esima caratteristica dell'individuo I_i..

La metrica Euclidea risulta essere molto influenzata dalle unità di misura, infatti se si usano due dataframe contenenti gli stessi dati con gli stessi valori, ma in un primo caso con un tipo di unità di misura e nel secondo un'altra unità di misura, avremmo valori delle distanze diverse fra di loro, anche se si tratta degli stessi dati. Un metodo per ovviare a questo problema consiste nello scalare e standardizzare inizialmente le misure in maniera tale da realizzare la possibilità di un confronto tra le misure, cioè considerare delle nuove variabili così calcolate:

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

Dove $\overline{x_j}$ e s_j sono rispettivamente la media campionaria e la deviazione standard campionaria della j-esima caratteristica.

In R per scalare e standardizzare 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).

e si ottengono dei nuovi dati le cui medie campionarie sono nulle e le varianze campionarie unitarie.

```
> AbortoScaled<-scale(Aborto)
```

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

(1)Metrica euclidea, come definita precedentemente e calcolata tramite il seguente codice:

```
> distEuclidean<-dist(AbortoScaled, method="euclidean", diag=TRUE, upper=TRUE)
```

(2) Metrica di Manhattan(o valore assoluto) così definita:

$$d_1(X_i, X_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

(3) Metrica di Chebycev (o del massimo)

$$d_\infty(X_i, X_j) = \max_{k=1,2\ldots,p} |x_{ik} - x_{jk}|.$$

Entrambe le metriche 2) e 3) sono computazionalmente semplici da calcolare con l'unica differenza che la metrica di Chebycev coinvolge anche una procedura di ordinamento.

(4) Metrica di Minkowski

$$d_r(X_i, X_j) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right]^{1/r},$$

con r≥1. Se r=2 si ottiene la metrica Euclidea, se r=1 si ottiene la metrica di Manhattan ed infine se r=∞ si ottiene la metrica di Chebycev.

(5) Metrica di Canberra

$$d_c(X_i, X_j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{|x_{ik} + x_{jk}|}$$

La metrica di Canberra è definita per variabili non negative ed ha la caratteristica di essere sensibile alle differenze *relative* piuttosto che a quelle *assolute*.

Misure di similarità: Nella maggior parte delle tecniche di clustering occorre inizialmente calcolare la matrice D delle distanze oppure una matrice S delle similarità. Una misura di similarità fornisce un valore numerico compreso tra 0 e 1 e permette di definire in modo quantitativo la somiglianza o differenza tra due individui Ii e Ij, intendendo ovviamente con 0 l'assoluta assenza e con 1 la massima presenza di somiglianza.
Una funzione a valori reali sij = s(Xi,Xj) è detta misura di similarità se e soltanto se essa soddisfa le seguenti condizioni:

- s(Xi,Xi) = 1;
- 0 ≤ s(Xi,Xj) ≤ 1;
- s(Xi,Xj) = s(Xj ,Xi) per ogni Xi e Xj.

La quantità sij è chiamata coefficiente di similarità ed è l'elemento nella riga i-esima e colonna j-esima della matrice di similarità S, che presenta la diagonale con tutti elementi pari ad 1.

Matrice di non omogeneità totale: Alla matrice X delle misure si può associare una matrice WX di cardinalità p × p, detta *matrice delle varianze e covarianze* così definita:

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

Dove l'elemento wrl è: $$w_{r\ell} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)(x_{i\ell} - \overline{x}_\ell) \qquad (r, \ell = 1, 2, \ldots, p)$$. Si nota che se r=l l'elemento wrl è la varianza campionaria relativa alla caratteristica r-esima effettuata su tutti gli n individui, mentre se r ≠ **l** l'elemento wrl è la covarianza campionaria effettuta su tutti gli n indivi-

dui. In R utilizzando le funzione apply(X,2,mean), apply(X, 2, var) e apply(X, 2, sd) è possibile calcolare la media campionaria, la varianza campionaria e la deviazione standard campionaria delle colonne di una matrice X. Inoltre applicando la
funzione cov(X) è possibile ottenere la matrice WX delle covarianze campionarie tra le *varie caratteristiche*.

Matrice statistica di non omogeneità: La matrice statistica di non omogeneità (statistical scatter matrix) per l'insieme I di individui, di cardinalità p × p, è così definita**:**

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

Dove l'elemento generico è :

$$h_{rl} = \sum_{i=1}^{n} (x_{ir} - \overline{x_r})(x_{il} - \overline{x_l}) = (n-1)w_{rl}$$

Quando r = ℓ, si nota che h_{rr}

corrisponde a n−1 volte la varianza campionaria della caratteristica r-esima effettuata su tutti gli n individui.

Traccia della matrice: La traccia della matrice H_I, denotata con trH_I, detta misura di non omogeneità statistica dell'insieme I di individui è

$$\mathrm{tr}H_I = \sum_{k=1}^{p} h_{kk} = (n-1)\sum_{k=1}^{p} s_k^2 = \sum_{i=1}^{n} d_2^2(X_i, \overline{X}).$$

dove d_2 indica la distanza euclidea e dove $\bar{X}$ è un vettore di cardinalità p il cui generico elemento $\bar{x}_j$ rappresenta la media campionaria relativa alla caratteristica j-esima effettuata sugli n individui.

Vediamo in R come ricavarci la matrice di non omogeneità statistica e la traccia della matrice:

```
> n<-nrow(Aborto)
> WI<-cov(Aborto)
> HI<-(n-1)*WI
> HI #visualizza la matrice di non omogeneità statistica
                                quindici_diciannove_anni
quindici_diciannove_anni                          28.95
venti_ventiquattro_anni                           43.05
venticinque_ventinove_anni                        47.20
trenta_trentaquattro_anni                         35.45
trentacinque_trentanove_anni                      29.95
quaranta_quarantaquattro_anni                     11.45
quarantacinque_quarantanove_anni                   2.05
quindici_quarantanove_anni                        23.75
                                venti_ventiquattro_anni
quindici_diciannove_anni                          43.05
venti_ventiquattro_anni                          106.95
venticinque_ventinove_anni                        93.80
trenta_trentaquattro_anni                         87.55
trentacinque_trentanove_anni                      66.05
quaranta_quarantaquattro_anni                     19.55
quarantacinque_quarantanove_anni                  -1.05
quindici_quarantanove_anni                        51.25
                                venticinque_ventinove_anni
quindici_diciannove_anni                              47.2
venti_ventiquattro_anni                               93.8
venticinque_ventinove_anni                           115.2
trenta_trentaquattro_anni                             79.2
trentacinque_trentanove_anni                          62.2
quaranta_quarantaquattro_anni                         20.2
```

```
quarantacinque_quarantanove_anni                      3.8
quindici_quarantanove_anni                           51.0
                          trenta_trentaquattro_anni
quindici_diciannove_anni                          35.45
venti_ventiquattro_anni                           87.55
venticinque_ventinove_anni                        79.20
trenta_trentaquattro_anni                         80.95
trentacinque_trentanove_anni                      57.45
quaranta_quarantaquattro_anni                     18.95
quarantacinque_quarantanove_anni                  -0.45
quindici_quarantanove_anni                        44.25
                          trentacinque_trentanove_anni
quindici_diciannove_anni                          29.95
venti_ventiquattro_anni                           66.05
venticinque_ventinove_anni                        62.20
trenta_trentaquattro_anni                         57.45
trentacinque_trentanove_anni                      52.95
quaranta_quarantaquattro_anni                     18.45
quarantacinque_quarantanove_anni                   0.05
quindici_quarantanove_anni                        35.75
                          quaranta_quarantaquattro_anni
quindici_diciannove_anni                          11.45
venti_ventiquattro_anni                           19.55
venticinque_ventinove_anni                        20.20
trenta_trentaquattro_anni                         18.95
trentacinque_trentanove_anni                      18.45
quaranta_quarantaquattro_anni                     10.95
quarantacinque_quarantanove_anni                   0.55
quindici_quarantanove_anni                        13.25
                          quarantacinque_quarantanove_anni
quindici_diciannove_anni                              2.05
venti_ventiquattro_anni                              -1.05
venticinque_ventinove_anni                            3.80
trenta_trentaquattro_anni                            -0.45
trentacinque_trentanove_anni                          0.05
quaranta_quarantaquattro_anni                         0.55
quarantacinque_quarantanove_anni                      0.95
quindici_quarantanove_anni                            0.25

quindici_diciannove_anni                          23.75
venti_ventiquattro_anni                           51.25
venticinque_ventinove_anni                        51.00
trenta_trentaquattro_anni                         44.25
trentacinque_trentanove_anni                      35.75
quaranta_quarantaquattro_anni                     13.25
quarantacinque_quarantanove_anni                   0.25
quindici_quarantanove_anni                        29.75
> trHI<-sum(diag(HI))
> trHI
[1] 426.65
```

Si è inoltre calcolato la misura di non omogeneità statistica, che corrisponde alla somma della diagonale principale della tabella, il risultato ottenuto è 426.65.

Misure di ottimizzazione

Dopo aver scelto la misura di distanza (o di similarità), bisogna un algoritmo di raggruppamento delle unità osservate che sia idoneo. Esistono tre metodi di raggruppamento:

- Metodo di enumerazione completa
- Metodi gerarchici
- Metodi non gerarchici

Un metodo di enumerazione completa consiste nel valutare la traccia della somma delle matrici di non omogeneità relative ai singoli cluster per ogni possibile partizione dell'insieme totale degli n individui in m cluster. La traccia di una matrice di non omogeneità di un insieme di individui fornisce una misura della dispersione dei dati intorno al valore medio dell'insieme dal quale è stata ricavata. È intuitivo pensare che più un insieme di dati è addensato più piccola è la traccia della matrice di non omogeneità. Il problema principale che si presenta utilizzando le tecniche di ottimizzazione è che esse sono computazionalmente onerose poichè prevedono il calcolo della funzione obiettivo per ogni possibile partizione dell'insieme totale di n individui in m cluster. Per questo si adottano altri metodi, quali i metodi gerarchici di clustering e i metodi non gerarchici. I metodi gerarchici hanno due vantaggi: quello di fornire una visione completa dell'insieme in termini di distanza (o di coefficienti di similarità) seppure condizionata dalla scelta del metodo seguito e quello di non comportare la scelta a priori del numero di cluster oppure la scelta a priori di parametri per la determinazione automatica del loro numero. Invece, uno svantaggio di tali metodi è

che essi non consentono di riallocare gli individui che sono stati già classificati ad un livello precedente. dell'analisi. I metodi non gerarchici di clustering consentono, a differenza delle tecniche di tipo gerarchico, di riallocare gli individui già classificati ad un livello precedente dell'analisi.

4.2 Metodi gerarchici

Molti metodi di analisi gerarchica sono caratterizzati da una struttura comune che si riflette in un algoritmo generale che può essere esplicitato come segue:

Algoritmo

• **Step 1:** A partire dalla matrice X (o quella scalata) calcolare la matrice delle distanze;
• **Step 2:** individuare la coppia di cluster meno distanti e raggruppare in un unico cluster;
• **Step 3:** costruire una nuova matrice delle distanze ridotta di una riga e di una colonna e calcolare le nuove distanze tra i cluster;
• **Step 4:** ripetere la procedura a partire dal passo 2 fino ad ottenere una matrice 2x2;
• **Step 5:** rappresentare graficamente attraverso un dendrogramma che riporta sull'asse verticale il livello di distanza a cui avviene l'agglomerazione e sull'asse orizzontale riporta gli individui.

L'analisi gerarchica di tipo agglomerativo viene effettuata in R attraverso la funzione hclust(d, method="complete"). Dove d rappresenta un oggetto 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).

Verranno utilizzati tutti poiché nessuno è un metodo di ottimizzazione.
Considerando il seguente data frame.

```
> Aborto<-data.frame(quindici_diciannove_anni=c(5,6,7,4,2,2,4,4,4,4,3,5,4,4,3,5,3,3,4,3),venti_ventiquattro_anni=c(1
3,8,15,10,7,7,8,11,11,10,7,11,9,8,7,11,7,6,8,7),venticinque_ventinove_anni=c(13,14,15,11,8,8,8,13,12,12,9,12,9,8,8,1
2,9,7,8,8),trenta_trentaquattro_anni=c(13,9,14,10,8,7,9,12,12,9,7,10,9,8,8,12,8,8,8,8),trentacinque_trentanove_anni=
c(10,8,11,8,6,6,7,10,9,9,6,8,8,9,7,11,7,6,7,6),quaranta_quarantaquattro_anni=c(4,4,4,3,2,3,3,4,4,3,2,4,4,4,3,5,4,3,3
,3),quarantacinque_quarantanove_anni=c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),quindici_quarantanove_anni=c(7,6,8,6
,4,4,5,7,7,6,5,7,6,5,5,8,5,4,5,5))
> rownames(Aborto)<-c("Piemonte","Valle_d_Aosta","Liguria","Lombardia","Trentino_Alto_Adige","Veneto",
"Friuli_Venezia_Giulia","Emilia_Romagna","Toscana","Umbria","Marche","Lazio","Abruzzo","Molise","Campa
nia","Puglia","Basilicata","Calabria","Sicilia","Sardegna")
```

Per tutti i metodi utilizzeremo la distanza ricavata utilizzando la metrica euclidea.

(1) Metodo del legame singolo

In questo metodo la distanza tra i due gruppi è definita come la minima tra tutte le distanze che si possono calcolare tra ogni individuo del primo gruppo e ogni individuo del secondo gruppo. Nel primo passo si vedono gli individui più vicini attraverso la matrice delle distanze, e si raggruppano nello stesso cluster, e si va avanti per tutti gli individui.
Un vantaggio del metodo del legame singolo è di consentire di individuare gruppi di qualsiasi forma e di mettere in luce eventuali valori anomali meglio di altre tecniche gerarchiche.
Uno svantaggio di tale metodo è che invece esso può dare origine alla formazione di una catena tra gli individui.
È stato applicato ora, alla matrice delle distanze, il metodo del legame singolo

```
> hls<-hclust(distEuclidean, method="single")
> str(hls) #visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:19, 1:2] -7 -15 -8 -4 1 -18 -6 -13 -1 -5 ...
 $ height     : num [1:19] 0.484 0.599 0.724 0.871 0.913 ...
 $ order      : int [1:20] 2 3 16 12 1 8 9 4 10 13 ...
 $ labels     : chr [1:20] "Piemonte" "Valle_d_Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "single"
 $ call       : language hclust(d = distEuclidean, method = "single")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

I risultati di $merge sono stati disposti su due colonne: i numeri con il segno negativo indicano i singoli individui, mentre i numeri positivi indicano i cluster che si formano.

```
> hls$merge
      [,1] [,2]
 [1,]   -7  -19
 [2,]  -15  -20
 [3,]   -8   -9
 [4,]   -4  -10
 [5,]    1    2
 [6,]  -18    5
 [7,]   -6    6
 [8,]  -13  -14
 [9,]   -1    3
[10,]   -5  -11
[11,]  -17    7
[12,]  -12    9
[13,]   10   11
[14,]    8   13
[15,]    4   14
[16,]   12   15
[17,]  -16   16
[18,]   -3   17
[19,]   -2   18
```

Inoltre, $height indica la distanza in cui è avvenuta l'aggiornamento tra i due cluster.

```
> hls$height
 [1] 0.4844717 0.5990234 0.7237124 0.8709032 0.
9132125 0.9905751 1.1106757 1.2549028 1.2655549
[10] 1.3017691 1.3784377 1.3978498 1.4035215 1.
5601936 1.6754844 1.8364920 1.8851734 2.2864269
[19] 4.8773223
```

		Agglomerazione	**Distanza**
-7	-19	Al livello 1 si uniscono Friuli Venezia Giulia e Sicilia	0.484
-15	-20	Al livello 2 si uniscono Campania e Sardegna	0.599
-8	-9	Al livello 3 si uniscono Emilia Romagna e Toscana	0.724
-4	-10	Al livello 4 si uniscono Lombardia e Umbria	0.871
1	2	Al livello 5 si uniscono il primo cluster(formato da Friuli Venezia Giulia e Sicilia) con il secondo cluster (formato da Campania e Sardegna)	0.913
-18	5	Al livello 6 si uniscono Calabria e il quinto cluster (formato da Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	0.991
-6	6	Al livello 7 si uniscono Veneto e il sesto cluster(formato da Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.111

-13 -14	Al livello 8 si uniscono Abruzzo e Molise	1.255
-1 3	Al livello 9 si uniscono Piemonte e il terzo cluster(formato da Emilia Romagna e Toscana)	1.266
-5 -11	Al livello 10 si uniscono Trentino Alto Adige e Marche	1.302
-17 7	Al livello 11 si uniscono Basilicata e il settimo cluster (formato da Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.378
-12 9	Al livello 12 si uniscono Lazio e il nono cluster (formato da Piemonte, Emilia Romagna e Toscana)	1.398
10 11	Al livello 13 si uniscono il decimo (formato da Trentino Alto Adige e Marche) e l'undicesimo cluster (formato da Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.404
8 13	Al livello 14 si uniscono l'ottavo (formato da Abruzzo e Molise) e tredicesimo cluster (formato da Trentino Alto Adige, Marche, Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.560
4 14	Al livello 15 si uniscono il quarto cluster (formato da Lombardia e Umbria) e il quattordicesimo cluster (formato da Abruzzo, Molise, Trentino Alto Adige, Marche, Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.675
12 15	Al livello 16 si uniscono il dodicesimo cluster (formato da Lazio, da Piemonte, Emilia Romagna e Toscana) e il quindicesimo cluster (formato da Lombardia, Umbria, Abruzzo, Molise, Trentino Alto Adige, Marche, Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.836
-16 16	Al livello 17 si uniscono Puglia e il sedicesimo cluster (formato da Lazio, Piemonte, Emilia Romagna e Toscana, Lombardia, Umbria, Abruzzo, Molise, Trentino Alto Adige, Marche, Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	1.885
-3 17	Al livello 18 si uniscono Liguria e il diciassettesimo cluster (formato da Puglia, Lazio, da Piemonte, Emilia Romagna e Toscana, Lombardia, Umbria, Abruzzo, Molise, Trentino Alto Adige, Marche, Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	2.286
-2 18	Al livello 19 si uniscono Valle d'Aosta e il diciottesimo cluster (formato da Liguria, Puglia, Lazio, da Piemonte, Emilia Romagna e Toscana, Lombardia, Umbria, Abruzzo, Molise, Trentino	4.877

	Alto Adige, Marche, Basilicata, Veneto, Calabria, Friuli Venezia Giulia, Sicilia, Campania e Sardegna)	

È stato costruito il dendrogramma utilizzando le seguenti linee di codice

```
> plot(hls, hang = -1, xlab = "Metodo_gerarchico_agglomerativo", sub = "del_legame_singolo")
> axis(side = 4, at=round(c(0,hls$height),2))
```

che hanno prodotto il grafico in Figura 4.1.

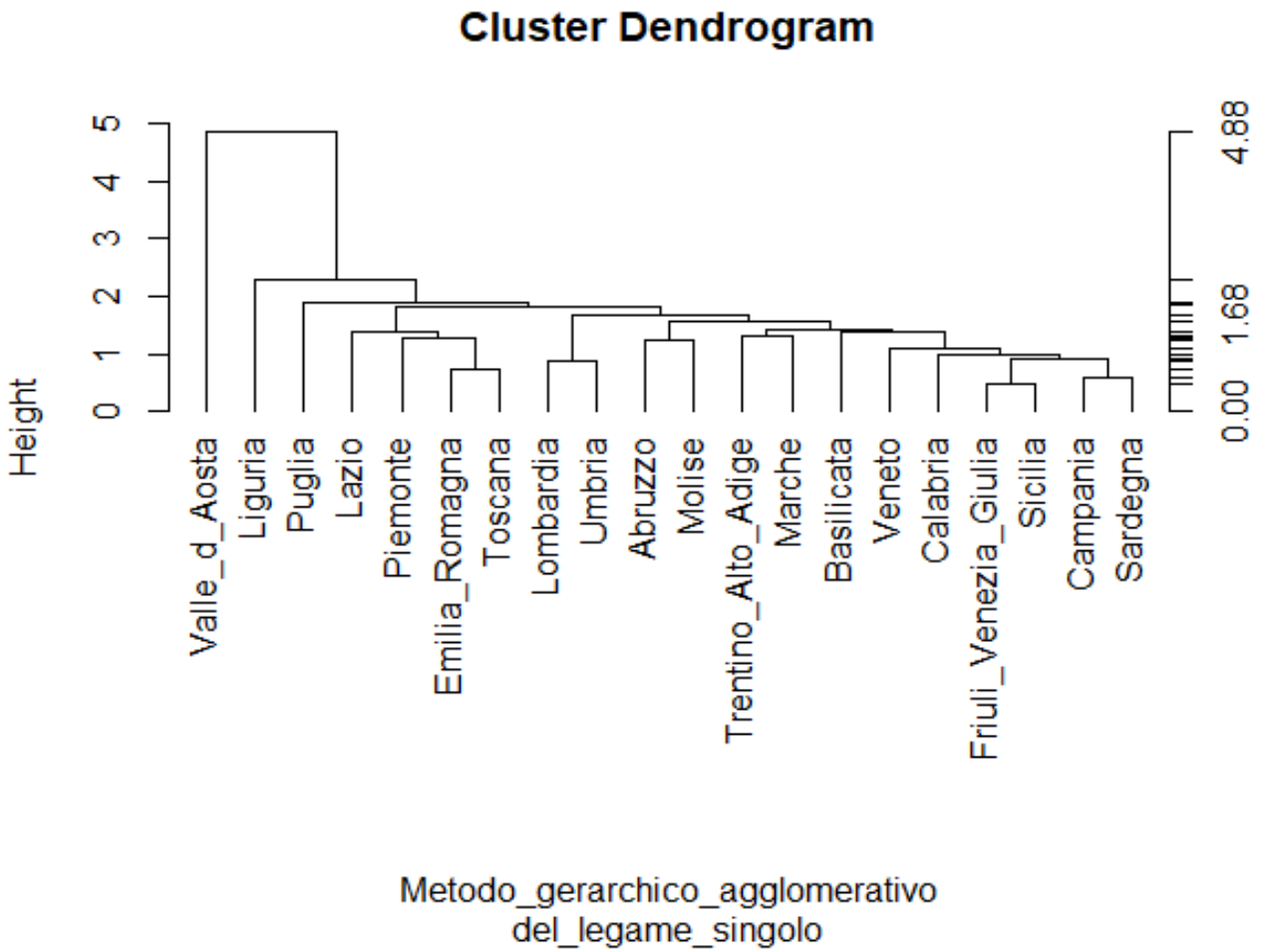


Figura 4.1:Dendrogramma ottenuto con il metodo del legame singolo.

Il dendogramma in figura 4.1 ha i rami più corti poiché i gruppi si formano a livello di distanza minore.

(2) Metodo del legame completo

In questo metodo la distanza tra i due gruppi è definita come la massima tra tutte le distanze che si possono calcolare tra ogni individuo del primo gruppo e ogni individuo del secondo gruppo.

Alla matrice delle distanze, ricavata dal nostro data frame, è stato applicato ora il metodo del legame completo.

```
> hlc<-hclust(distEuclidean, method = "complete")
> str(hlc) #visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:19, 1:2] -7 -15 -8 -4 -6 1 -13 -5 -1 -17 ...
 $ height     : num [1:19] 0.484 0.599 0.724 0.871 1.111 ...
 $ order      : int [1:20] 7 19 15 20 6 18 5 11 17 13 ...
 $ labels     : chr [1:20] "Piemonte" "Valle_d_Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "complete"
 $ call       : language hclust(d = distEuclidean, method = "complete")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> hlc$merge
      [,1] [,2]
 [1,]   -7  -19
 [2,]  -15  -20
 [3,]   -8   -9
 [4,]   -4  -10
 [5,]   -6  -18
 [6,]    1    2
 [7,]  -13  -14
 [8,]   -5  -11
 [9,]   -1    3
[10,]  -17    7
[11,]    5    8
[12,]  -12    4
[13,]  -16    9
[14,]    6   11
[15,]   10   12
[16,]   -3   13
[17,]   14   15
[18,]   -2   16
[19,]   17   18

> hlc$height
 [1] 0.4844717 0.5990234 0.7237124 0.8709032 1.1106757 1.1947799 1.2549028 1.3017691 1.4578713
[10] 1.6121767 1.8563145 1.9496959 2.1131577 2.4021295 3.2869960 3.6395581 5.1430322 6.4170196
[19] 8.3927653
```

È stato costruito il dendrogramma utilizzando le seguenti linee di codice

```
> plot(hlc, hang = -1, xlab = "Metodo_gerarchico_agglomerativo", sub = "del_legame_completo")
> axis(side = 4, at=round(c(0,hlc$height),2))
```

che hanno prodotto il grafico in Figura 4.2.

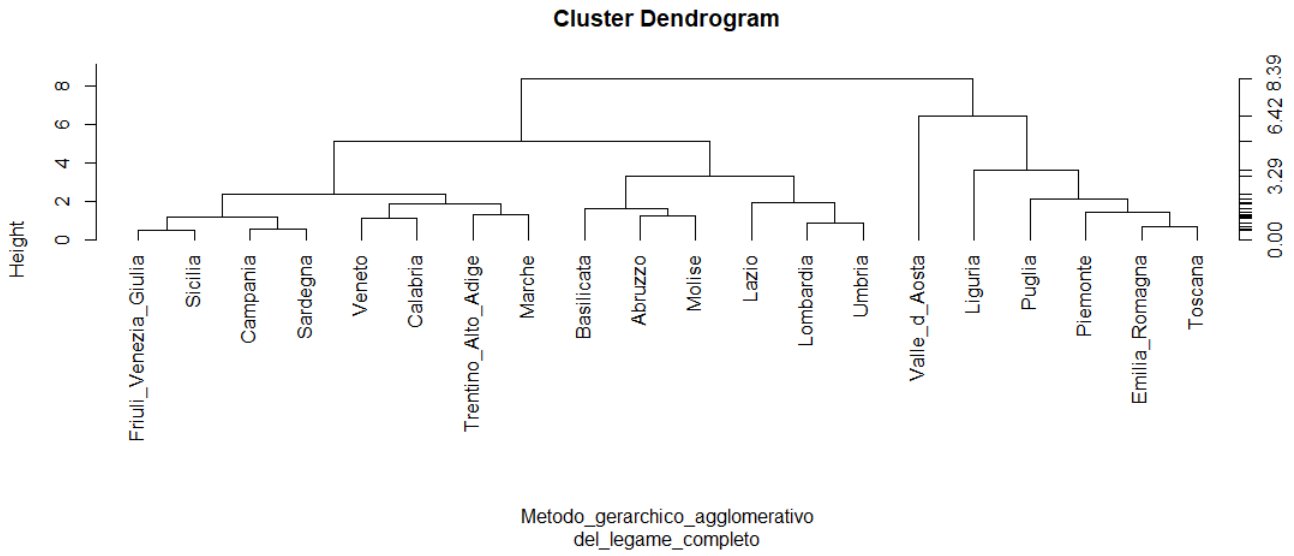


Figura 4.2:Dendrogramma ottenuto con il metodo del legame completo.

Il dendrogramma costruito con questo metodo ha i rami molto più lunghi rispetto al dendrogramma ottenuto con il metodo del legame singolo poichè i gruppi si formano a livelli di distanza maggiori.

(3) Metodo del legame medio

In questo metodo la distanza tra i due gruppi è definita come la media aritmetica delle distanze tra tutte le coppie di unità che compongono i due gruppi e si va avanti fino ad arrivare ad avere un unico cluster con n individui.

Uno svantaggio del metodo del legame medio è che se le misure dei due cluster da unire sono molto differenti la distanza sarà molto vicina a quella del cluster più numeroso.

Alla matrice delle distanze, ricavata dal nostro data frame, è stato applicato il metodo del legame medio

```
> hln<-hclust(distEuclidean, method = "average")
> hlm<-hclust(distEuclidean, method = "average")
> str(hlm) #visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:19, 1:2] -7 -15 -8 -4 1 -6 -13 -5 -1 5 ...
 $ height     : num [1:19] 0.484 0.599 0.724 0.871 1.058 ...
 $ order      : int [1:20] 2 5 11 7 19 15 20 6 18 4 ...
 $ labels     : chr [1:20] "Piemonte" "Valle_d_Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "average"
 $ call       : language hclust(d = distEuclidean, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> hlm$merge
      [,1] [,2]
 [1,]   -7  -19
 [2,]  -15  -20
 [3,]   -8   -9
 [4,]   -4  -10
 [5,]    1    2
 [6,]   -6  -18
 [7,]  -13  -14
 [8,]   -5  -11
 [9,]   -1    3
[10,]    5    6
[11,]  -17    7
[12,]  -12    9
[13,]    8   10
[14,]  -16   12
[15,]    4   11
[16,]   13   15
[17,]   -3   14
[18,]   16   17
[19,]   -2   18

> hlm$height
 [1] 0.4844717 0.5990234 0.7237124 0.8709032 1.0584762 1.1106757
 [7] 1.2549028 1.3017691 1.3617131 1.5247328 1.5861851 1.7628589
[13] 1.8258031 2.1291053 2.3054653 2.6211274 3.2429277 4.6985578
[19] 5.7413187
```

È stato costruito il dendrogramma utilizzando le seguenti linee di codice

```
> plot(hlm, hang = -1, xlab = "Metodo_gerarchico_agglomerativo", sub = "del_legame_medio")
> axis(side = 4, at=round(c(0,hlm$height),2))
```

che hanno prodotto il grafico in Figura 4.3.

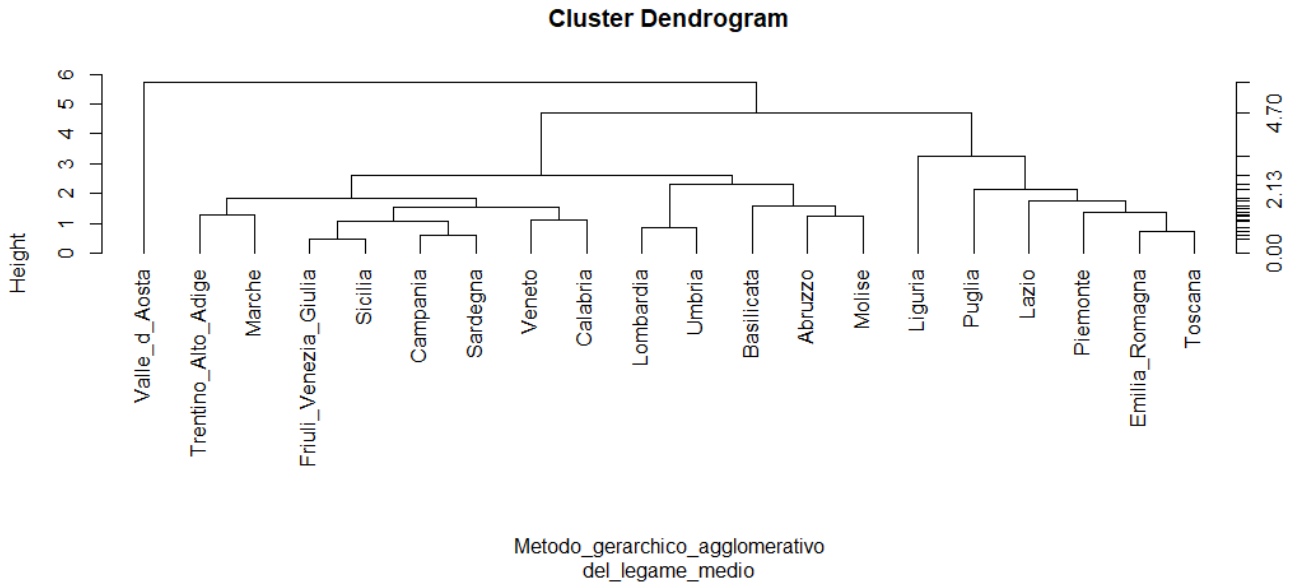


Figura 4.3:Dendrogramma ottenuto con il metodo del legame medio.

Nei metodi agglomerativi del legame singolo, del legame completo e del legame medio si può utilizzare una qualsiasi misura di distanza. Invece, nel metodo del centroide e nel metodo della mediana si considera la distanza euclidea e si lavora con una matrice $D^{(2)}$ che contiene i quadrati delle singole distanze euclidee.

(4) Metodo del centroide

In questo metodo la distanza tra i due gruppi è definita come la distanza tra i centroidi, ossia tra le medie campionarie calcolate sugli individui appartenente ai due gruppi. Si giunge poi ad avere un unico cluster con gli n individui.

Partendo dalla nostra matrice delle distanze, è stata calcolata la matrice contenente i quadrati delle distanze euclidee:

```
> distEuclidean2<-distEuclidean^2
> hc<-hclust(distEuclidean2, method = "centroid")
> str(hc) #visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:19, 1:2] -7 -15 -8 -4 1 -6 -13 -5 -1 5 ...
 $ height     : num [1:19] 0.235 0.359 0.524 0.758 0.982 ...
 $ order      : int [1:20] 2 17 13 14 5 11 7 19 15 20 ...
 $ labels     : chr [1:20] "Piemonte" "Valle_d_Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = distEuclidean2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> hc$merge
      [,1] [,2]
 [1,]   -7  -19
 [2,]  -15  -20
 [3,]   -8   -9
 [4,]   -4  -10
 [5,]    1    2
 [6,]   -6  -18
 [7,]  -13  -14
 [8,]   -5  -11
 [9,]   -1    3
[10,]    5    6
[11,]  -17    7
[12,]    8   10
[13,]  -12    9
[14,]  -16   13
[15,]   11   12
[16,]    4   14
[17,]   -3   16
[18,]   15   17
[19,]   -2   18
```

```
> hc$height
 [1]  0.2347128  0.3588291  0.5237596  0.7584724  0.9823425
 [6]  1.2336004  1.5747810  1.6946028  1.7325691  1.8477125
[11]  2.1229636  2.2712136  2.7185782  3.7371208  4.5294715
[16]  5.8121711 12.3685787 17.2919138 26.5914742
```

È stato costruito il dendrogramma utilizzando le seguenti linee di codice

```
> plot(hc, hang = -1, xlab = "Metodo_gerarchico_agglomerativo", sub = "del_centroide")
> axis(side = 4, at=round(c(0,hc$height),2))
```

che hanno prodotto il grafico in Figura 4.4.

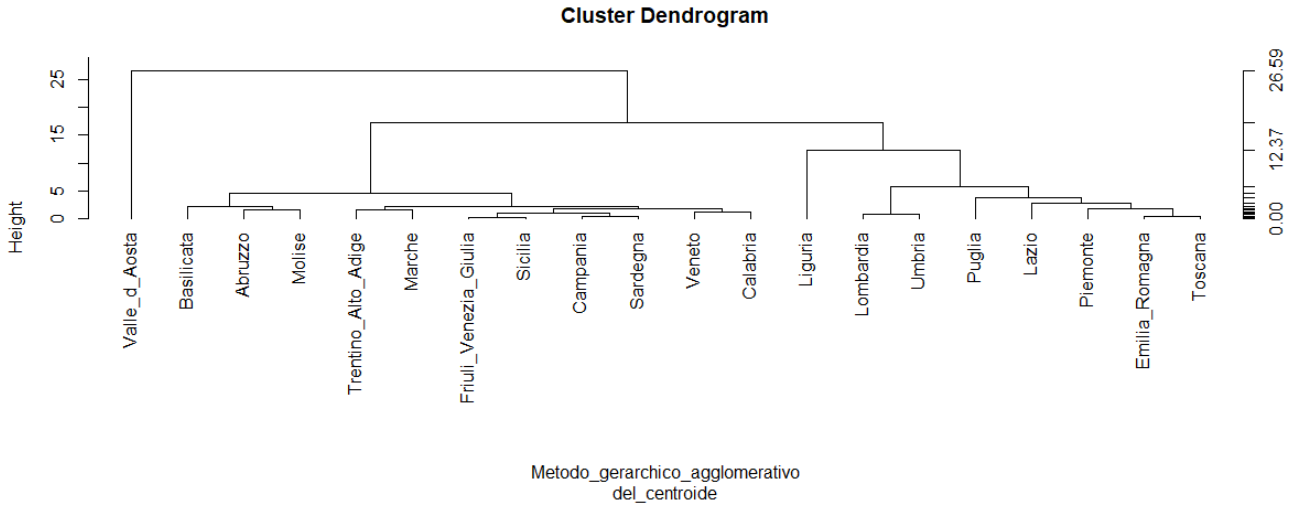


Figura4.4:Dendrogramma ottenuto con il metodo del centroide.

Il metodo del centroide può avere come problema il fatto che gruppi molto grandi possono attrarre al loro interno gruppi piccoli. 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.

(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.

È stato applicato alla matrice contenente i quadrati delle distanze euclidee il metodo della mediana

```
> hmed<-hclust(distEuclidean2, method = "median")
> str(hmed) #visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:19, 1:2] -7 -15 -8 -4 1 -6 -13 -5 -1 5 ...
 $ height     : num [1:19] 0.235 0.359 0.524 0.758 0.982 ...
 $ order      : int [1:20] 2 5 11 7 19 15 20 6 18 17 ...
 $ labels     : chr [1:20] "Piemonte" "Valle_d_Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "median"
 $ call       : language hclust(d = distEuclidean2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> hmed$merge
      [,1] [,2]
 [1,]   -7  -19
 [2,]  -15  -20
 [3,]   -8   -9
 [4,]   -4  -10
 [5,]    1    2
 [6,]   -6  -18
 [7,]  -13  -14
 [8,]   -5  -11
 [9,]   -1    3
[10,]    5    6
[11,]    8   10
[12,]  -17    7
[13,]  -12    9
[14,]    4   13
[15,]   11   12
[16,]  -16   14
[17,]   -3   16
[18,]   15   17
[19,]   -2   18

> hmed$height
 [1]  0.2347128  0.3588291  0.5237596  0.7584724
 [5]  0.9823425  1.2336004  1.5747810  1.6946028
 [9]  1.7325691  1.8477125  2.0754761  2.1229636
[13]  3.0003715  4.0382490  5.7471923  8.1478837
[17] 10.8520569 30.0065827 24.6250229
```

È stato costruito il dendrogramma utilizzando le seguenti linee di codice

```
> plot(hmed, hang = -1, xlab = "Metodo_gerarchico_agglomerativo", sub = "della_mediana")
> axis(side = 4, at=round(c(0,hmed$height),2))
```

che hanno prodotto il grafico in Figura 4.5.

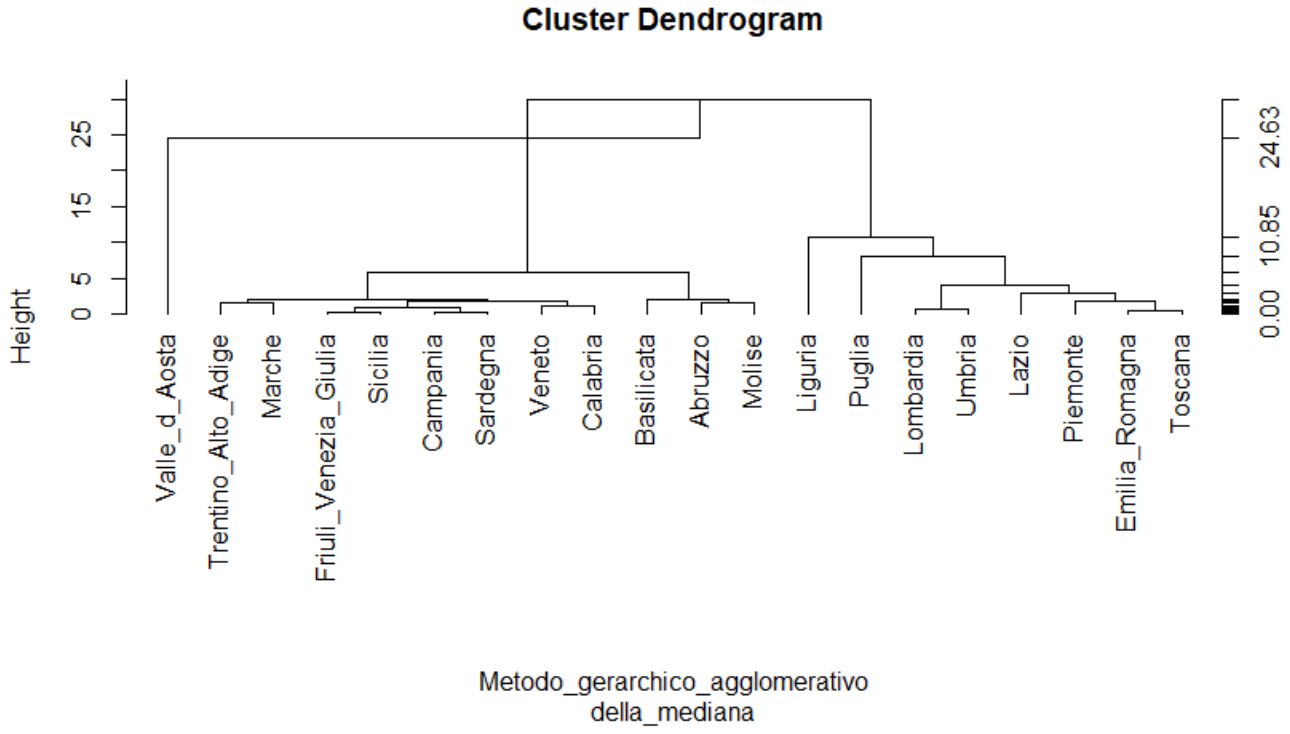


Figura 4.5:Dendrogramma ottenuto con il metodo della mediana.

Si può notare che il dendrogramma creato con il metodo della mediana è uguale a quello del centroide, solo che alcuni livello di distanza relativi alle aggregazioni sono diversi rispetto a quelli del centroide.

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 sia poiché i cluster così formati sono sicuramente ben separati sia poiché questo metodo è in grado di individuare cluster di qualsiasi forma.
Terminato un qualsiasi algoritmo gerarchico si possono selezionare il numero di cluster che il ricercatore ritiene più adeguato al problema oggetto di studio.

4.3 Screeplot

Al fine di scegliere una buona partizione del dendrogramma, si può costruire lo *screeplot* in cui si pongono sull'asse delle ordinate i numeri di gruppi ottenibili con il metodo gerarchico e sull'asse delle ascisse le distanze a cui avvengono le successive aggregazioni tra gruppi. Se si taglia il dendrogramma in k gruppi, e a k-1 si ha un forte incremento della distanza di aggregazione è meglio tagliare il dendrogramma in k gruppi. Per vedere come si incrementano le distanze causate dalle successive agglomerazioni tra individui si può usare la seguente misura:

$$\delta_k = d_{k-1} - d_k \qquad (k = 2, \ldots, n),$$

dk indica la distanza a cui è stata effettuata l'agglomerazione in k gruppi, mentre δk è la misura dell'incremento, che quando è molto elevato significa che i gruppi sono abbastanza dissimili tra loro per cui è possibile tagliare il dendrogramma all'altezza corrispondente alla partizione in k gruppi.
È preferibile costruire lo screeplot a partire dal metodo del legame singolo, del legame completo o del legame medio in cui è utilizzata la funzione distanza. Invece, nel metodo del centroide e della mediana (che utilizzano i quadrati delle distanze) le successive agglomerazioni potrebbero non verificarsi ad un livello di distanza maggiore o uguale rispetto alle precedenti agglomerazioni. Ciò comporta che lo screeplot ottenuto a partire dal metodo del centroide o della mediana potrebbe essere non regolare.

- ✓ Si è costruito lo screeplot relativo al metodo del legame singolo.

```
> hls$height
 [1] 0.4844717 0.5990234 0.7237124 0.8709032 0.
9132125 0.9905751 1.1106757 1.2549028 1.2655549
[10] 1.3017691 1.3784377 1.3978498 1.4035215 1.
5601936 1.6754844 1.8364920 1.8851734 2.2864269
[19] 4.8773223
```

Considerando la seguente linea di codice

```
> plot(rev(c(0,hls$height)), seq(1,20), type = "b", main = "Screeplot", xlab = "Distanza_di_aggrega
zione", ylab = "Numero_di_cluster", col="red")
```

la funzione c(0, hls$height) permette di concatenare 0 con il vettore hls$height delle altezze a cui sono avvenute le successive agglomerazioni.

La precedente linea di codice ha prodotto il grafico illustrato in Figura 4.6.

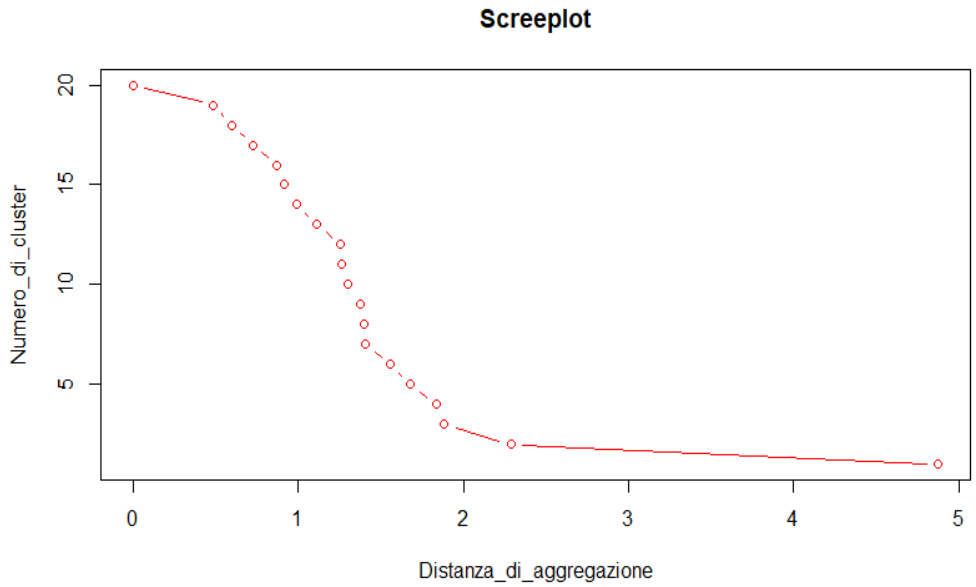


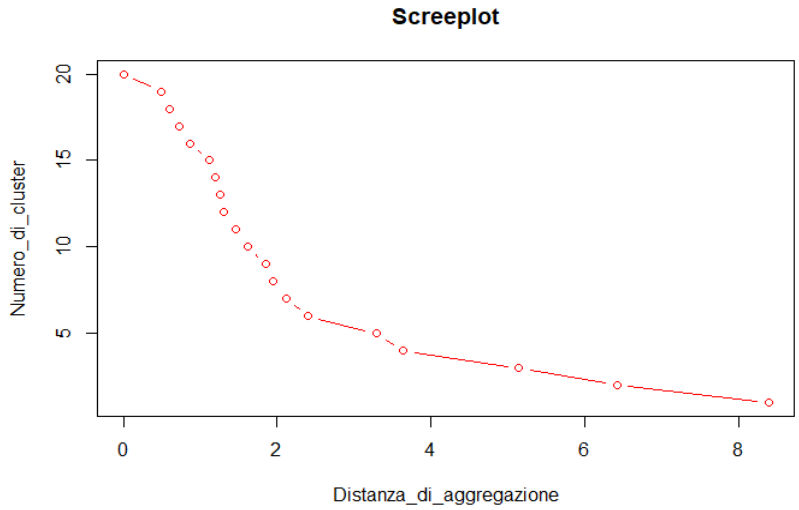
Figura 4.6:Sull'asse delle ordinate è presente il numero di gruppi e sull'asse delle ascisse la distanza in cui è avvenuta

Lo screeplot in Figura 4.6 suggerisce di considerare una suddivisione in due gruppi. Infatti eseguendo i vari calcoli, applicando la formula abbiamo che il valore di k per il quale δk è massima è k=2 dato da $h_1 - h_2 = 4.8773223 - 2.2864269 = 2.5908954$. E' preferibile quindi considerare una suddivisione nei due cluster.

✓ Si è costruito lo screeplot relativo al metodo del legame completo.

```
> hlc$height
 [1] 0.4844717 0.5990234 0.7237124 0.8709032 1.1106757 1.1947799 1.2549028 1.3017691 1.4578713
[10] 1.6121767 1.8563145 1.9496959 2.1131577 2.4021295 3.2869960 3.6395581 5.1430322 6.4170196
[19] 8.3927653

> plot(rev(c(0,hlc$height)), seq(1,20), type = "b", main = "Screeplot", xlab = "Distanza_di_aggregazione", yl
ab = "Numero_di_cluster", col="red")
```



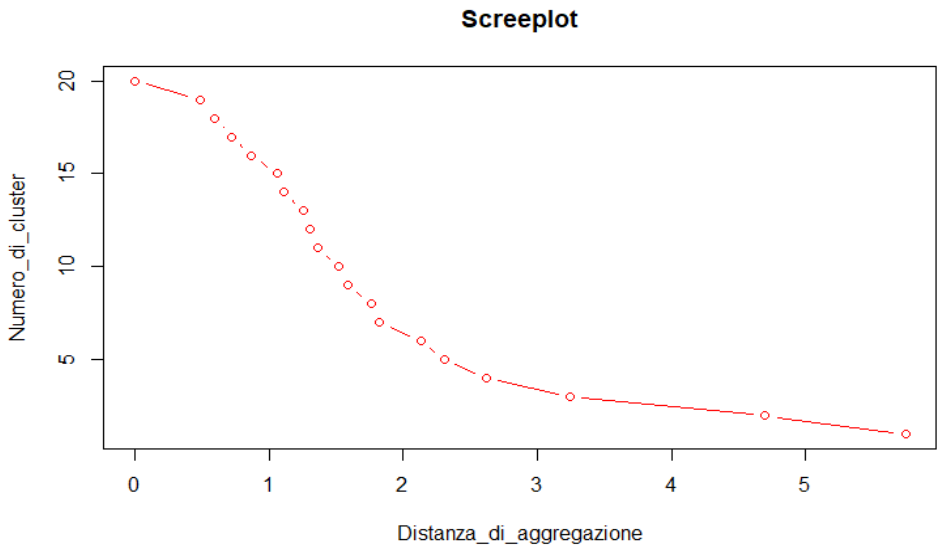
Lo screeplot suggerisce di considerare una suddivisione in due gruppi. Infatti eseguendo i vari calcoli, applicando la formula abbiamo che il valore di k per il quale δk è massima è k=2 dato da

$h_1 - h_2 = 8.3927653 - 6.4170196 = 1.9759457$. E' preferibile quindi considerare una suddivisione nei due cluster.

✓ Si è costruito lo screeplot relativo al metodo del legame medio.

```
> hlm$height
 [1] 0.4844717 0.5990234 0.7237124 0.8709032 1.0584762 1.1106757
 [7] 1.2549028 1.3017691 1.3617131 1.5247328 1.5861851 1.7628589
[13] 1.8258031 2.1291053 2.3054653 2.6211274 3.2429277 4.6985578
[19] 5.7413187

> plot(rev(c(0,hlm$height)), seq(1,20), type = "b", main = "Screeplot", xlab = "Distanza_di_aggregazione", y
lab = "Numero_di_cluster", col="red")
```



Lo screeplot suggerisce di considerare una suddivisione in tre gruppi. Infatti eseguendo i vari calcoli, applicando la formula abbiamo che il valore di k per il quale δk è massima è k=3 dato da $h_2 - h_3 = 4.6985578 - 3.2429277 = 1.4556301$. E' preferibile quindi considerare una suddivisione nei tre cluster.

Occorre infine sottolineare che lo screeplot è soltanto un grafico basato sulle altezze a cui sono avvenute le agglomerazioni e non sempre fornisce il numero ottimale di cluster in cui suddividere gli individui.

4.4 Analisi del dendrogramma

Ci siamo proposti ora di analizzare i dendrogrammi ottenuti con i vari metodi gerarchici e di calcolare, fissato il numero di cluster, le misure di non omogeneità della partizione individuata.

4.4.1 Disegnare rettangoli che evidenziano i cluster

Ad esempio, prendendo in considerazione il dendrogramma ottenuto applicando il metodo del legame singolo si sono potute evidenziare due partizioni mediante rettangoli colorati in rosso. La seguente linea di codice

```
> rect.hclust(hls, k=2, border = "red")
```

h indica l'altezza del taglio, e k il numero di cluster che si vogliono ottenere, ha prodotto il grafico di Figura 4.7.

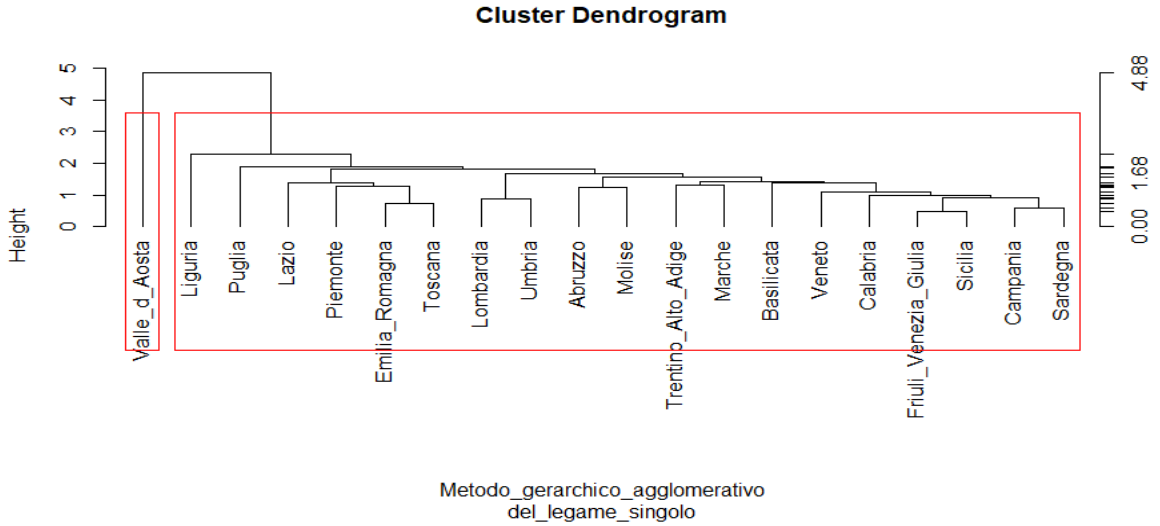


Figura 4.7:Rettangoli che evidenziano due partizioni

Se invece si vogliono evidenziare tre partizioni mediante rettangoli colorati in verde.
La seguente linea di codice

```
> rect.hclust(hls, k=3, border = "green")
```

ha prodotto il grafico di Figura 4.8.

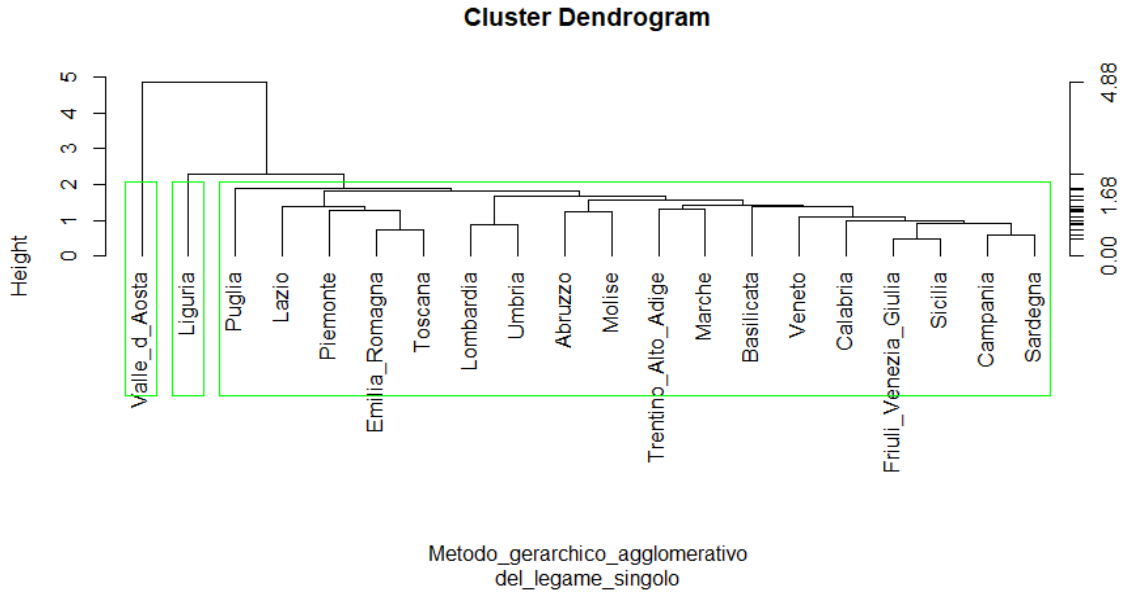


Figura4.8:Rettangoli che evidenziano tre partizioni

Per confrontare differenti partizioni alternative si può utilizzare più volte sullo stesso grafico la funzione rect.hclust(). Ad esempio, le seguenti linee di codice

```
> rect.hclust(hls, k=2, border = "red")
> rect.hclust(hls, k=3, border = "green")
```

hanno prodotto il grafico di Figura 4.9 che permette di confrontare su uno stesso grafico le partizioni di Figura 4.7 e di Figura 4.8.

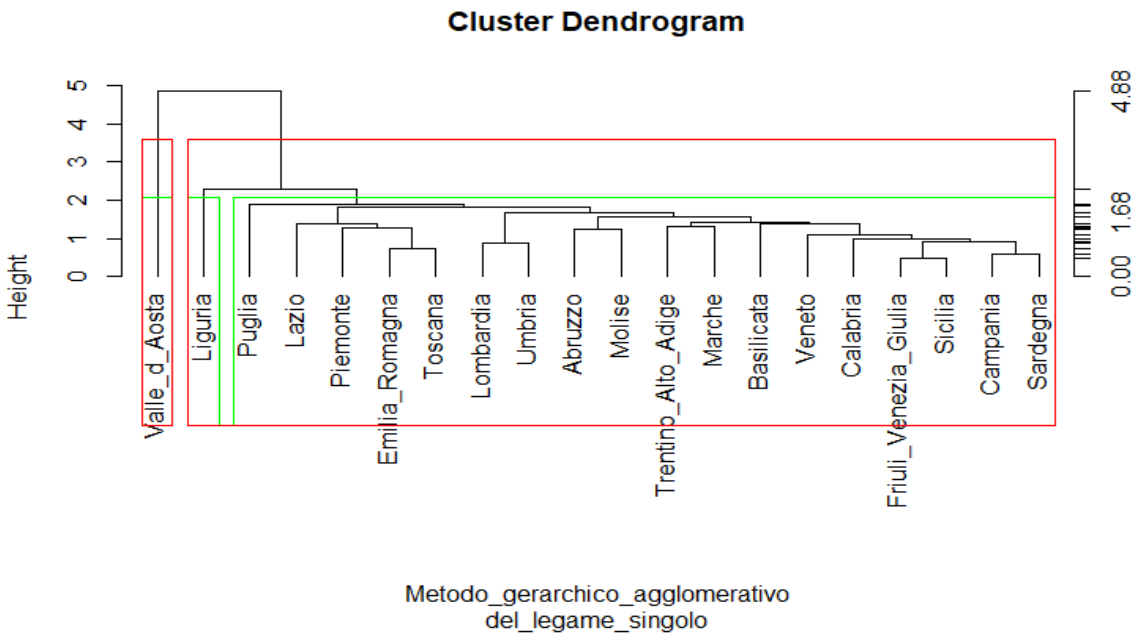


Figura 4.9:Rettangoli che evidenziano due partizioni (in rosso) e tre partizioni (in verde).

Si è supposto invece di voler evidenziare quattro partizioni mediante rettangoli colorati in blu.

Quindi, prendendo sempre in considerazione il dendrogramma ottenuto applicando il metodo del legame singolo si sono evidenziate quattro partizioni mediante rettangoli colorati in blue.
La seguente linea di codice

```
> rect.hclust(hls, k=4, border = "blue")
```

ha prodotto il grafico di Figura 4.10.

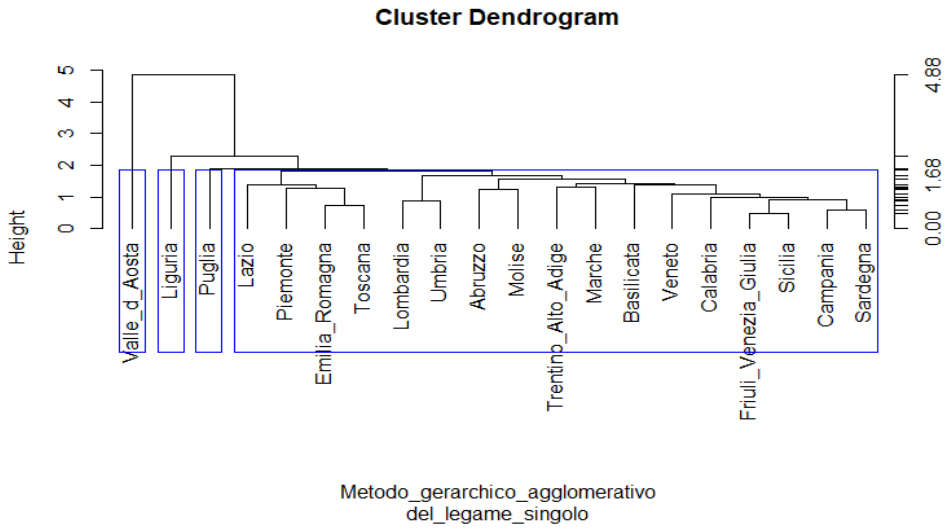


Figura4.10:Rettangoli che evidenziano quattro partizioni

4.4.2 Inserire gli individui nei cluster

Considerando 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.
Inoltre, per vedere come vengono classificati gli individui all'aumentare del numero di cluster si può considerare la funzione cutree(tree, k=1:n), dove *n* indica il numero di individui. L'output di tale funzione cutree() è una matrice in cui la colonna *k*-esima contiene numeri interi positivi associati ai cluster in cui sono stati inseriti i vari individui.

(1) Per il metodo del legame singolo si è applicata la funzione cutree() fissando a 4 il numero di cluster in cui tagliare il dendrogramma e si ha:

```
> cutree(hls, k=4)
            Piemonte         Valle_d_Aosta               Liguria             Lombardia
                   1                     2                     3                     1
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   1                     1                     1                     1
             Toscana               Umbria              Marche               Lazio
                   1                     1                     1                     1
             Abruzzo               Molise            Campania             Puglia
                   1                     1                     1                     4
          Basilicata             Calabria             Sicilia             Sardegna
                   1                     1                     1                     1
```

che individua la partizione in quattro cluster G_1*{Piemonte, Lombardia, Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Emilia Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Basilicata, Calabria, Sicilia, Sardegna}*, G_2*{Valle d'Aosta}*, G_3*{Liguria}*, G_4*{Puglia}*.

Per ottenere il numero di unità (individui) in ciascun cluster si può applicare la funzione table() al risultato della funzione cutree() ed è stato ottenuto:

```
> table(cutree(hls,k=4))

 1  2  3  4
17  1  1  1
```

Mentre, la funzione

```
> cutree(hls, k=1:20)
                     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Piemonte             1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1
Valle_d_Aosta        1 2 2 2 2 2 2 2 2  2  2  2  2  2  2  2  2  2  2  2
Liguria              1 1 3 3 3 3 3 3 3  3  3  3  3  3  3  3  3  3  3  3
Lombardia            1 1 1 1 4 4 4 4 4  4  4  4  4  4  4  4  4  4  4  4
Trentino_Alto_Adige  1 1 1 1 4 5 5 5 5  5  5  5  5  5  5  5  5  5  5  5
Veneto               1 1 1 1 4 5 5 6 6  6  6  6  6  6  6  6  6  6  6  6
Friuli_Venezia_Giulia 1 1 1 1 4 5 5 6 6  6  6  6  6  7  7  7  7  7  7  7
Emilia_Romagna       1 1 1 1 1 1 1 1 1  1  1  7  7  8  8  8  8  8  8  8
Toscana              1 1 1 1 1 1 1 1 1  1  1  7  7  8  8  8  8  9  9  9
Umbria               1 1 1 1 4 4 4 4 4  4  4  4  4  4  4  4  9 10 10 10
Marche               1 1 1 1 4 5 5 5 5  5  7  8  8  9  9  9 10 11 11 11
Lazio                1 1 1 1 1 1 1 1 7  7  8  9  9 10 10 10 11 12 12 12
Abruzzo              1 1 1 1 4 5 6 7 8  8  9 10 10 11 11 11 12 13 13 13
Molise               1 1 1 1 4 5 6 7 8  8  9 10 11 12 12 12 13 14 14 14
Campania             1 1 1 1 4 5 5 6 6  6  6  6  6  7  7 13 14 15 15 15
Puglia               1 1 1 4 5 6 7 8 9  9 10 11 12 13 13 14 15 16 16 16
Basilicata           1 1 1 1 4 5 5 6 6 10 11 12 13 14 14 15 16 17 17 17
Calabria             1 1 1 1 4 5 5 6 6  6  6  6  6  7 15 16 17 18 18 18
Sicilia              1 1 1 1 4 5 5 6 6  6  6  6  6  7  7  7  7  7  7 19
Sardegna             1 1 1 1 4 5 5 6 6  6  6  6  6  7  7 13 14 15 19 20
```

ha permesso di classificare gli individui all'aumentare del numero di cluster. Ad esempio la partizione in tre cluster è *G_1{ Piemonte, Lombardia, Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Emilia Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Basilicata, Calabria, Sicilia, Sardegna }, G_2{Valle d'Aosta}, G_3{Liguria}*.

(2) Per il metodo del legame completo si è applicata la funzione cutree() fissando a 4 il numero di cluster in cui tagliare il dendrogramma si ha:

```
> cutree(hlc, k=4)
             Piemonte         Valle_d_Aosta               Liguria             Lombardia
                    1                     2                     1                     3
  Trentino_Alto_Adige                Veneto Friuli_Venezia_Giulia        Emilia_Romagna
                    4                     4                     4                     1
              Toscana                Umbria               Marche                 Lazio
                    1                     3                     4                     3
             Abruzzo               Molise             Campania                Puglia
                    3                     3                     4                     1
           Basilicata             Calabria              Sicilia              Sardegna
                    3                     4                     4                     4
```

che individua la partizione in quattro cluster *G_1{Piemonte, Liguria, Emilia Romagna, Toscana, Puglia}, G_2{Valle d'Aosta}, G_3{Lombardia, Umbria, Lazio, Abruzzo, Molise, Basilicata}, G_4{Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Marche, Campania, Calabria, Sicilia, Sardegna}*.

(3) Per il metodo del legame medio si è applicata la funzione cutree() fissando a 4 il numero di cluster in cui tagliare il dendrogramma si ha:

```
> cutree(hlm, k=4)
             Piemonte         Valle_d_Aosta               Liguria             Lombardia
                    1                     2                     3                     4
  Trentino_Alto_Adige                Veneto Friuli_Venezia_Giulia        Emilia_Romagna
                    4                     4                     4                     1
              Toscana                Umbria               Marche                 Lazio
                    1                     4                     4                     1
             Abruzzo               Molise             Campania                Puglia
                    4                     4                     4                     1
           Basilicata             Calabria              Sicilia              Sardegna
                    4                     4                     4                     4
```

che individua la partizione in quattro cluster *G_1{Piemonte, Emilia Romagna, Toscana, Lazio, Puglia}, G_2{Valle d'Aosta}, G_3{Liguria}, G_4{Lombardia, Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Umbria, Marche, Abruzzo, Molise, Campania, Basilicata, Calabria, Sicilia, Sardegna}*.

(4) Per il metodo del centroide si è applicata la funzione cutree() fissando a 4 il numero di cluster in cui tagliare il dendrogramma si ha:

```
> cutree(hc, k=4)
            Piemonte         Valle_d_Aosta              Liguria             Lombardia
                   1                     2                    3                     1
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   4                     4                    4                     1
              Toscana               Umbria               Marche                 Lazio
                   1                     1                    4                     1
             Abruzzo               Molise             Campania                Puglia
                   4                     4                    4                     1
          Basilicata             Calabria              Sicilia              Sardegna
                   4                     4                    4                     4
```

che individua la partizione in quattro cluster *G1{Piemonte, Lombardia, Emilia Romagna, Toscana, Umbria, Lazio, Puglia}*, *G2{Valle d'Aosta}*, *G3{Liguria}*, *G4{Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Marche, Abruzzo, Molise, Campania, Basilicata, Calabria, Sicilia, Sardegna}*.

(5) Per il metodo della mediana si è applicata la funzione cutree() fissando a 4 il numero di cluster in cui tagliare il dendrogramma si ha:

```
> cutree(hmed, k=4)
            Piemonte         Valle_d_Aosta              Liguria             Lombardia
                   1                     2                    3                     1
  Trentino_Alto_Adige               Veneto Friuli_Venezia_Giulia       Emilia_Romagna
                   4                     4                    4                     1
              Toscana               Umbria               Marche                 Lazio
                   1                     1                    4                     1
             Abruzzo               Molise             Campania                Puglia
                   4                     4                    4                     1
          Basilicata             Calabria              Sicilia              Sardegna
                   4                     4                    4                     4
```

che individua la partizione in quattro cluster *G1{Piemonte, Lombardia, Emilia Romagna, Toscana, Umbria, Lazio, Puglia}*, *G2{Valle d'Aosta}*, *G3{Liguria}*, *G4{Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Marche, Abruzzo, Molise, Campania, Basilicata, Calabria, Sicilia, Sardegna}*.

Si noti che le quattro partizioni ottenute per il metodo del centroide e per il metodo della mediana sono le stesse.

4.4.3 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, …) sui 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. L'output della funzione aggregate() è una struttura contenente i valori ottenuti applicando la funzione FUN ad ognuna delle caratteristiche associate ai diversi cluster che sono stati aggregati.

Riconsiderando il nostro data frame si è desiderato utilizzare la funzione aggregate() per calcolare le medie campionarie, le varianze campionarie e le deviazioni standard delle caratteristiche dei quattro cluster precedentemente individuati.

(1) Per il metodo del legame singolo

```
> taglio<-cutree(hls, k=4)
> tagliolist<-list(taglio)
```

è stata ricavata:

- Media campionaria

```
> aggregate(Aborto, tagliolist, mean)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	3.588235	8.647059	9.588235	9.058824	7.588235	3.294118	0	5.470588
2	6.000000	8.000000	14.000000	9.000000	8.000000	4.000000	1	6.000000
3	7.000000	15.000000	15.000000	14.000000	11.000000	4.000000	0	8.000000
4	5.000000	11.000000	12.000000	12.000000	11.000000	5.000000	0	8.000000

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

- Varianza campionaria

```
> aggregate(Aborto, tagliolist, var)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	0.7573529	3.992647	4.257353	3.183824	2.007353	0.4705882	0	1.139706
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	NA	NA

Il valore della varianza campionaria per il cluster 2,3 e 4 è NA poiché questo cluster è composto da un solo elemento.

- Deviazione standard

```
> aggregate(Aborto, tagliolist, sd)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	0.8702603	1.998161	2.063335	1.784327	1.416811	0.6859943	0	1.06757
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	NA	NA

Il valore della deviazione standard per il cluster 2,3 e 4 è NA poiché questo cluster è composto da un solo elemento.

(2) Per il metodo del legame completo è stata ricavata:

```
> taglio<-cutree(hlc, k=4)
> tagliolist<-list(taglio)
```

- Media campionaria

```
> aggregate(Aborto, tagliolist, mean)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	5	12.200000	13.00000	12.600	10.200000	4.200000	0	7.400000
2	6	8.000000	14.00000	9.000	8.000000	4.000000	1	6.000000
3	4	9.166667	10.16667	9.000	8.166667	3.666667	0	5.833333
4	3	7.125000	8.00000	7.875	6.375000	2.750000	0	4.625000

- Varianza campionaria

```
> aggregate(Aborto, tagliolist, var)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	1.5000000	3.2000000	1.5000000	0.8000000	0.7000000	0.2000000	0	0.3000000
2	NA	NA	NA	NA	NA	NA	NA	NA
3	0.4000000	2.1666667	2.9666667	0.8000000	0.5666667	0.2666667	0	0.5666667
4	0.5714286	0.4107143	0.2857143	0.4107143	0.2678571	0.2142857	0	0.2678571

Il valore della varianza campionaria per il cluster 2 è NA poiché questo cluster è composto da un solo elemento.

- Deviazione standard

```
> aggregate(Aborto, tagliolist, sd)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	1.2247449	1.7888544	1.2247449	0.8944272	0.8366600	0.4472136	0	0.5477226
2	NA	NA	NA	NA	NA	NA	NA	NA
3	0.6324555	1.4719601	1.7224014	0.8944272	0.7527727	0.5163978	0	0.7527727
4	0.7559289	0.6408699	0.5345225	0.6408699	0.5175492	0.4629100	0	0.5175492

Il valore della deviazione standard per il cluster 2 è NA poiché questo cluster è composto da un solo elemento.

3) Per il metodo del legame medio è stata ricavata:

```
> taglio<-cutree(hlm, k=4)
> tagliolist<-list(taglio)
```

- Media campionaria

```
> aggregate(Aborto, tagliolist, mean)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	4.600000	11.400000	12.400000	11.800000	9.600000	4.200000	0	7.2
2	6.000000	8.000000	14.000000	9.000000	8.000000	4.000000	1	6.0
3	7.000000	15.000000	15.000000	14.000000	11.000000	4.000000	0	8.0
4	3.307692	7.769231	8.692308	8.230769	7.076923	3.076923	0	5.0

- Varianza campionaria

```
> aggregate(Aborto, tagliolist, var)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	0.3000000	0.800000	0.300000	1.2000000	1.30000	0.2000000	0	0.2
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	0.5641026	1.525641	1.897436	0.6923077	1.24359	0.4102564	0	0.5

Il valore della varianza campionaria per il cluster 2 e 3 è NA poiché questo cluster è composto da un solo elemento.

- Deviazione standard

```
> aggregate(Aborto, tagliolist, sd)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	1.2247449	1.7888544	1.2247449	0.8944272	0.8366600	0.4472136	0	0.5477226
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	0.7510676	1.2351684	1.3774745	0.8320503	1.115164	0.6405126	0	0.7071068

Il valore della deviazione standard per il cluster 2 e 3 è NA poiché questo cluster è composto da un solo elemento.

(4) Per il metodo del centroide è stata ricavata:

```
> taglio<-cutree(hc, k=4)
> tagliolist<-list(taglio)
```

- Media campionaria

```
> aggregate(Aborto, tagliolist, mean)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	4.428571	11.000000	12.142857	11.14286	9.285714	3.857143	0	6.857143
2	6.000000	8.000000	14.000000	9.00000	8.000000	4.000000	1	6.000000
3	7.000000	15.000000	15.000000	14.00000	11.000000	4.000000	0	8.000000
4	3.181818	7.363636	8.181818	8.00000	6.818182	3.090909	0	4.818182

- Varianza campionaria

```
> aggregate(Aborto, tagliolist, var)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	0.2857143	1.0000000	0.4761905	2.142857	1.2380952	0.4761905	0	0.4761905
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	0.5636364	0.6545455	0.3636364	0.400000	0.9636364	0.4909091	0	0.3636364

Il valore della varianza campionaria per il cluster 2 e 3 è NA poiché questo cluster è composto da un solo elemento.

- Deviazione standard

```
> aggregate(Aborto, tagliolist, sd)
```

Group 1	**15-19 anni**	**20-24 anni**	**25-29 anni**	**30-34 anni**	**35-39 anni**	**40-44 anni**	**45-49 anni**	**15-49 anni**
1	0.5345225	1.0000000	0.6900656	1.4638501	1.1126973	0.6900656	0	0.6900656
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	0.7507572	0.8090398	0.6030227	0.6324555	0.9816498	0.7006490	0	0.6030227

Il valore della deviazione standard per il cluster 2 e 3 è NA poiché questo cluster è composto da un solo elemento.

(5) Per il metodo della mediana è stata ricavata:

```
> taglio<-cutree(hmed, k=4)
> tagliolist<-list(taglio)
```

- Media campionaria

```
> aggregate(Aborto, tagliolist, mean)
```

Group 1	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
1	4.428571	11.000000	12.142857	11.14286	9.285714	3.857143	0	6.857143
2	6.000000	8.000000	14.000000	9.00000	8.000000	4.000000	1	6.000000
3	7.000000	15.000000	15.000000	14.00000	11.000000	4.000000	0	8.000000
4	3.181818	7.363636	8.181818	8.00000	6.818182	3.090909	0	4.818182

- Varianza campionaria

```
> aggregate(Aborto, tagliolist, var)
```

Group 1	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
1	0.2857143	1.0000000	0.4761905	2.142857	1.2380952	0.4761905	0	0.4761905
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	0.5636364	0.6545455	0.3636364	0.400000	0.9636364	0.4909091	0	0.3636364

Il valore della varianza campionaria per il cluster 2 e 3 è NA poiché questo cluster è composto da un solo elemento.

- Deviazione standard

```
> aggregate(Aborto, tagliolist, sd)
```

Group 1	15-19 anni	20-24 anni	25-29 anni	30-34 anni	35-39 anni	40-44 anni	45-49 anni	15-49 anni
1	0.5345225	1.0000000	0.6900656	1.4638501	1.1126973	0.6900656	0	0.6900656
2	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA
4	0.7507572	0.8090398	0.6030227	0.6324555	0.9816498	0.7006490	0	0.6030227

Il valore della deviazione standard per il cluster 2 e 3 è NA poiché questo cluster è composto da un solo elemento.

Si nota che la deviazione standard, la varianza campionaria e la media campionaria ottenute per il metodo del centroide, per il metodo della mediana sono le stesse.

4.4.4 Misure di non omogeneità statistiche

Dopo aver effettuato il taglio, siamo stati 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 non omogeneità tra i cluster ($tr\ B$):

$$tr\ T\ =\ tr\ S\ +\ tr\ B,$$

o equivalentemente:

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

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)*.
Quindi, se due differenti metodi gerarchici conducono a due diverse partizioni con lo stesso numero di cluster, occorre scegliere quella partizione con misura di non omogeneità statistica all'interno dei cluster ($tr\ S$) più piccola, che corrisponde a maggiore omogeneità interna.

Riconsiderando, quindi, il nostro data frame sono state calcolate le misure di non omogeneità statistica relative all'insieme totale *I* e ai cluster che sono stati individuati con l'analisi della gerarchia.
Per l'insieme totale I si è avuto:

```
> n<-nrow(AbortoScaled)
> trHI<-(n-1)*sum(apply(AbortoScaled,2,var))
> trHI #visualizza la misura di non omogeneità totale
[1] 152
```

trHI indica la misura di non omogeneità totale, adesso vediamo la misura di non omogeneità statistica dei quattro gruppi creati prima.

La misura di non omogeneità totale è quindi $tr\ HI\ =\ 152$.

Si è calcolato ora le omogeneità relative ai gruppi.

(1) Per il metodo del legame singolo si è avuto:

```
> taglio<-cutree(hls, k=4)
> tagliolist<-list(taglio)
> num<-table(taglio)
> agvar<-aggregate(AbortoScaled, tagliolist, var)[,-1]
> trH1<-(num[[1]]-1)*sum(agvar[1,])
> trH2<-0
> trH3<-0
> trH4<-0
> trH1
[1] 78.72853
```

Riassumendo la misura di non omogeneità statistica totale è 152,
la misura di non omogeneità statistica all'interno dei quattro gruppi *(within)* è:
$tr\ HG1\ +\ tr\ HG2\ +\ tr\ HG3\ +\ tr\ HG4$ = 78.72853 + 0 + 0 + 0 = **78.72853**
e la misura di non omogeneità tra i cluster *(between)* è
$tr\ H\ (G1 \cap\ G2 \cap\ G3 \cap\ G4)\ =\ tr\ HI\ -\ tr\ HG1\ +\ tr\ HG2\ +\ tr\ HG3\ +\ tr\ HG4$ = 152 – 78.72853 = **73.27147**.
Da questo si è ricavato facilmente

$$\frac{tr\ S}{tr\ T}=\frac{78.72853}{152}\ =\ 0.517951$$

e che

$$\frac{tr\ B}{tr\ T}\ =\frac{73.27147}{152}\ =\ 0.482049$$

In conclusione è stato possibile osservare che la misura di non omogeneità all'interno dei gruppi e più grande rispetto alla misura di non omogeneità tra i cluster, quindi il numero di cluster scelti non era ottimale.

(2) Per il metodo del legame completo si è avuto:

```
> taglio<-cutree(hlc, k=4)
> tagliolist<-list(taglio)
> num<-table(taglio)
> agvar<-aggregate(AbortoScaled, tagliolist, var)[,-1]
> trH1<-(num[[1]]-1)*sum(agvar[1,])
> trH2<-0
> trH3<-(num[[3]]-1)*sum(agvar[3,])
> trH4<-(num[[4]]-1)*sum(agvar[4,])
> trH1 #visualizza la misura di non omogeneità del primo gruppo
[1] 11.11168
> trH3 #visualizza la misura di non omogeneità del terzo gruppo
[1] 11.76226
> trH4 #visualizza la misura di non omogeneità del quarto gruppo
[1] 8.613652
```

Riassumendo la misura di non omogeneità statistica totale è 152,
la misura di non omogeneità statistica all'interno dei quattro gruppi *(within)* è:
$tr\,HG1 + tr\,HG2 + tr\,HG3 + tr\,HG4$ = 11.11168 + 0 + 11.76226 + 8.613652 = **31.487592**
e la misura di non omogeneità tra i cluster *(between)* è
$tr\,H\,(G1 \cap G2 \cap G3 \cap G4) = tr\,HI - tr\,HG1 + tr\,HG2 + tr\,HG3 + tr\,HG4$ = 152 – 31.487592=**120.512408**.
Da questo si è ricavato facilmente

$$\frac{tr\,S}{tr\,T} = \frac{31.487592}{152} = 0.207155$$

e che

$$\frac{tr\,B}{tr\,T} = \frac{120.512408}{152} = 0.792845$$

In conclusione è stato possibile osservare che la misura di non omogeneità all'interno dei gruppi è più piccola rispetto alla misura di non omogeneità tra i cluster.

(3) Per il metodo del legame medio si è avuto:

```
> taglio<-cutree(hlm, k=4)
> tagliolist<-list(taglio)
> num<-table(taglio)
> agvar<-aggregate(AbortoScaled, tagliolist, var)[,-1]
> trH1<-(num[[1]]-1)*sum(agvar[1,])
> trH2<-0
> trH3<-0
> trH4<-(num[[4]]-1)*sum(agvar[4,])
> trH1 #visualizza la misura di non omogeneità del primo gruppo
[1] 6.445556
> trH4 #visualizza la misura di non omogeneità del quarto gruppo
[1] 31.12945
```

Riassumendo la misura di non omogeneità statistica totale è 152,
la misura di non omogeneità statistica all'interno dei quattro gruppi *(within)* è:
$tr\,HG1 + tr\,HG2 + tr\,HG3 + tr\,HG4$ = 6.445556 + 0 + 0 + 31.12945 = **37.575006**
e la misura di non omogeneità tra i cluster *(between)* è
$tr\,H\,(G1 \cap G2 \cap G3 \cap G4) = tr\,HI - tr\,HG1 + tr\,HG2 + tr\,HG3 + tr\,HG4$ = 152 – 37.575006=**114.424994**.
Da questo si è ricavato facilmente

$$\frac{tr\,S}{tr\,T} = \frac{37.575006}{152} = 0.247204$$

e che

$$\frac{tr\ B}{tr\ T} = \frac{114.424994}{152} = 0.752796$$

In conclusione è stato possibile osservare che la misura di non omogeneità all'interno dei gruppi è più piccola rispetto alla misura di non omogeneità tra i cluster.

(4) Per il metodo del centroide si è avuto:

```
> taglio<-cutree(hc, k=4)
> tagliolist<-list(taglio)
> num<-table(taglio)
> agvar<-aggregate(AbortoScaled, tagliolist, var)[,-1]
> trH1<-(num[[1]]-1)*sum(agvar[1,])
> trH2<-0
> trH3<-0
> trH4<-(num[[4]]-1)*sum(agvar[4,])
> trH1 #visualizza la misura di non omogeneità del primo gruppo
[1] 15.12789
> trH4 #visualizza la misura di non omogeneità del quarto gruppo
[1] 20.69884
```

(5) Per il metodo della mediana si è avuto:

```
> taglio<-cutree(hmed, k=4)
> tagliolist<-list(taglio)
> num<-table(taglio)
> agvar<-aggregate(AbortoScaled, tagliolist, var)[,-1]
> trH1<-(num[[1]]-1)*sum(agvar[1,])
> trH2<-0
> trH3<-0
> trH4<-(num[[4]]-1)*sum(agvar[4,])
> trH1 #visualizza la misura di non omogeneità del primo gruppo
[1] 15.12789
> trH4 #visualizza la misura di non omogeneità del quarto gruppo
[1] 20.69884
```

Riassumendo la misura di non omogeneità statistica totale è 152,
la misura di non omogeneità statistica all'interno dei quattro gruppi *(within)* è:
$tr\ HG1\ +\ tr\ HG2\ +\ tr\ HG3\ +\ tr\ HG4$ = 15.12789 + 0 + 0 + 20.69884 = **35.82673**
e la misura di non omogeneità tra i cluster *(between)* è
$tr\ H\ (G1 \cap\ G2 \cap\ G3 \cap\ G4)\ =\ tr\ HI\ -\ tr\ HG1\ +\ tr\ HG2\ +\ tr\ HG3\ +\ tr\ HG4$ = 152 – 35.82673=**116.17327**.
Da questo si è ricavato facilmente

$$\frac{tr\ S}{tr\ T} = \frac{35.82673}{152} = 0.235702$$

e che

$$\frac{tr\ B}{tr\ T} = \frac{116.17327}{152} = 0.764298$$

In conclusione la misura di non omogeneità all'interno dei gruppi è quindi più piccola rispetto la misura di non omogeneità tra i cluster.

Quindi c'è da dire però che il rapporto $\frac{tr\ B}{tr\ T}$ ha valore un abbastanza alto per 2), 3), 4), 5), rispetto a quello che si sarebbe ottenuto utilizzando un partizionamento con meno cluster. Quindi il partizionamento in quattro cluster è una soluzione accettabile (essendo il valore di almeno 70%) per quei metodi, tranne per il metodo del legame singolo poiché tale rapporto risulta essere inferiore.

4.5 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.
Si può dire che non esiste un unico tipo di metodo non gerarchico, come invece esisteva per metodi gerarchici, in quanto in alcuni metodi si può decidere a priori il numero di cluster in cui suddividere gli n individui, in altri invece è determinato durante l'analisi. 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 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.
Vediamo i passi del metodo k-means:
1. Fissare a priori il numero k di cluster specificando k punti di riferimento iniziali che portano a una prima partizione provvisoria;
2. Considerare tutti gli individui e attribuire ciascuno di essi al cluster individuato dal punto di riferimento da cui ha distanza minore;
3. Calcolare il baricentro (il centroide) di ognuno dei k gruppi così ottenuti. Tali centroidi costituiscono i punti di riferimento per i nuovi cluster;
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;
5. Ricalcolare i centroidi dei k gruppi così ottenuti;
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.

I vantaggi di questo metodo sono la velocità di esecuzione e la libertà che viene lasciata agli individui di raggrupparsi e allontanarsi; uno svantaggio invece riguarda il fatto che la scelta iniziale dei k vettori delle caratteristiche può influenzare la classificazione finale.
Nel metodo *k*-means 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, start=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, 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 riferimenti, nel caso in cui centers è il minimo (di default nstart = 1).
Si nota che nell'algoritmo *k*-means *non occorre calcolare la matrice iniziale delle distanze* così come invece si richiede nei metodi gerarchici.
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.

Riconsiderando il nostro data frame riguardante le otto fasce d'età osservate per le nostre venti regioni.
Si sono considerate tre differenti scelte iniziali: *(i)* scelta casuale dei punti di riferimento, *(ii)* ripetizione della procedura di scelta casuale dei punti di riferimento e *(iii)* scelta dei centroidi come punti di riferimento.

Per la scelta dei numero dei cluster, osservando i dendrogrammi la decisione migliore da prendere è quella di effettuare la suddivisione in quattro cluster, in quanto i dendrogrammi ci fanno notare che questa sia la scelta migliore.

(i) Scelta casuale dei punti di riferimento

È stato applicato ai dati del nostro data frame il metodo non gerarchico *k*-means considerando una suddivisione in quattro cluster ed effettuando un'unica *scelta casuale dei punti di riferimento* con un numero massimo di iterazione pari a 10.

```
> km<-kmeans(AbortoScaled, center = 4, iter.max = 10, nstart = 1 )
> km #visualizza i risultati ottenuti con kmeans
K-means clustering with 4 clusters of sizes 1, 7, 6, 6

Cluster means:
  quindici_diciannove_anni venti_ventiquattro_anni venticinque_ventinove_anni trenta_trentaquattro_anni
1               1.66075809             -0.4425636                1.5432424               -0.2180122
2              -0.07522598             -0.2017127               -0.3713065               -0.3564327
3              -1.03966157             -0.9343010               -0.8934561               -0.8639745
4               0.85063219              1.2433930                1.0694399                1.3161480
  trentacinque_trentanove_anni quaranta_quarantaquattro_anni quarantacinque_quarantanove_anni quindici_quarantanove_anni
1                   0.02995117                    0.72449006                        4.2485292                  0.1997898
2                  -0.05562361                   -0.02822689                       -0.2236068                 -0.2568726
3                  -1.06825847                   -1.03184948                       -0.2236068                 -0.9989490
4                   1.12816082                    0.94403250                       -0.2236068                  1.2653354

Clustering vector:
             Piemonte         Valle_d_Aosta               Liguria             Lombardia   Trentino_Alto_Adige
                    4                     1                     4                     2                     3
               Veneto Friuli_Venezia_Giulia        Emilia_Romagna               Toscana                Umbria
                    3                     2                     4                     4                     2
              Marche                 Lazio               Abruzzo                Molise              Campania
                    3                     4                     2                     2                     3
              Puglia            Basilicata              Calabria               Sicilia              Sardegna
                    4                     2                     3                     2                     3

Within cluster sum of squares by cluster:
[1]  0.000000 11.094667  5.236482 14.374795
 (between_SS / total_SS =  79.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

Il metodo *k*-means individua la seguente partizione in quattro cluster G_1{*Valle d'Aosta*}, G_2{*Lombardia, Friuli Venezia Giulia, Umbria, Abruzzo, Molise, Basilicata, Sicilia*}, G_3{*Trentino Alto Adige, Veneto, Marche, Campania, Calabria, Sardegna*}, G_4{*Piemonte, Liguria, Emilia Romagna, Toscana, Lazio, Puglia*}.

Dopo l'esecuzione dell'algoritmo K-means sono state estratte le misure di non omogeneità statistica dalla variabile km:

```
> km$totss
[1] 152
> km$withinss
[1]  0.000000 11.094667
[3]  5.236482 14.374795
> km$betweenss
[1] 121.2941
```

da cui si definiscono:
trHI=152
trS=trH1+trH2+trH3 + trH4 =0+11.094667+5.236482+14.374795=**30.705944**
trB= 121.294056

$$\frac{trB}{trT} = \frac{121.294056}{152} = 0.798$$

(ii) Ripetizione della procedura di scelta casuale dei punti di riferimento

È stato applicato ai dati del nostro data frame il metodo non gerarchico *k*-means richiedendo che l'algoritmo di aggregazione venga ripetuto otto volte in corrispondenza di otto ripetizioni della procedura di scelta casuale dei punti di riferimento con un numero massimo di iterazioni pari a 10.

```
> km2<-kmeans(AbortoScaled, center = 4, iter.max = 10, nstart = 8 )
> km2 #visualizza i risultati ottenuti con kmeans
K-means clustering with 4 clusters of sizes 6, 6, 7, 1

Cluster means:
  quindici_diciannove_anni venti_ventiquattro_anni venticinque_ventinove_anni trenta_trentaquattro_anni
1              -1.03966157             -0.9343010                -0.8934561               -0.8639745
2               0.85063219              1.2433930                 1.0694399                1.3161480
3              -0.07522598             -0.2017127                -0.3713065               -0.3564327
4               1.66075809             -0.4425636                 1.5432424               -0.2180122
  trentacinque_trentanove_anni quaranta_quarantaquattro_anni quarantacinque_quarantanove_anni quindici_quarantanove_anni
1                   -1.06825847                   -1.03184948                      -0.2236068                -0.9989490
2                    1.12816082                    0.94403250                      -0.2236068                 1.2653354
3                   -0.05562361                   -0.02822689                      -0.2236068                -0.2568726
4                    0.02995117                    0.72449006                       4.2485292                 0.1997898

Clustering vector:
            Piemonte         Valle_d_Aosta               Liguria             Lombardia   Trentino_Alto_Adige
                   2                     4                     2                     3                     1
      Veneto Friuli_Venezia_Giulia        Emilia_Romagna               Toscana                Umbria
                   1                     3                     2                     2                     3
              Marche                 Lazio               Abruzzo               Molise              Campania
                   1                     2                     3                     3                     1
              Puglia            Basilicata              Calabria               Sicilia              Sardegna
                   2                     3                     1                     3                     1

Within cluster sum of squares by cluster:
[1]  5.236482 14.374795 11.094667  0.000000
 (between_SS / total_SS =  79.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

In questo caso il metodo *k*-means individua la stessa partizione in quattro cluster G_1*{Trentino Alto Adige, Veneto, Marche, Campania, Calabria, Sardegna}*, G_2*{Piemonte, Liguria, Emilia Romagna, Toscana, Lazio, Puglia}*, G_3*{Lombardia, Friuli Venezia Giulia, Umbria, Abruzzo, Molise, Basilicata, Sicilia}*, G_4*{Valle d'Aosta}*.

In generale non esiste una regola per determinare in numero ottimale di ripetizioni della procedura di scelta casuale dei punti di riferimento per ottenere un risultato stabile.
Empiricamente, si potrebbe provare con diversi valori crescenti di nstart fino a che il risultato non cambia.

(iii) Scelta dei centroidi come punti di riferimento

Sono stati impiegati i *centroidi dei quattro cluster ottenuti con la tecnica gerarchica del centroide* utilizzando la funzione aggregate().
Effettuiamo la suddivisione in quattro cluster, prima di tutto calcoliamo i centroidi e poi vediamo i risultati di k-means:

```
> distEuclidean<-dist(AbortoScaled, method="euclidean", diag=TRUE, upper=TRUE)
> distEuclidean2<-distEuclidean^2
> hc<-hclust(distEuclidean2, method="centroid")
> taglio<-cutree(hc, k=4)
> tagliolist<-list(taglio)
> centroidiIniziali<-aggregate(AbortoScaled, tagliolist, mean)[,-1]
> centroidiIniziali #visualizza i centroidi iniziali
  quindici_diciannove_anni venti_ventiquattro_anni
1                0.3877031              0.8219039
2                1.6607581             -0.4425636
3                2.4708840              2.5078605
4               -0.6223240             -0.7107840
  venticinque_ventinove_anni trenta_trentaquattro_anni
1                  0.7890262                0.8201413
2                  1.5432424               -0.2180122
3                  1.9493589                2.2043460
4                 -0.8196168               -0.7024839
  trentacinque_trentanove_anni quaranta_quarantaquattro_anni
1                   0.80012417                     0.5363108
2                   0.02995117                     0.7244901
3                   1.82702150                     0.7244901
4                  -0.67798562                    -0.4730142
  quarantacinque_quarantanove_anni quindici_quarantanove_anni
1                       -0.2236068                  0.8847834
2                        4.2485292                  0.1997898
3                       -0.2236068                  1.7981082
4                       -0.2236068                 -0.7446711
```

Utilizzando tali centroidi si è potuto applicare il metodo *k*-means:

```
> km3<-kmeans(AbortoScaled, center = centroidiIniziali, iter.max = 10)
> km3 #visualizza i risultati ottenuti con kmeans
K-means clustering with 4 clusters of sizes 6, 1, 3, 10

Cluster means:
  quindici_diciannove_anni venti_ventiquattro_anni venticinque_ventinove_anni trenta_trentaquattro_anni
1                0.1755273              0.5409111                 0.5279514                0.4279500
2                1.6607581             -0.4425636                 1.5432424               -0.2180122
3                1.3907161              1.6648822                 1.2724982                1.7198744
4               -0.6886070             -0.7797549                -0.8528445               -0.7509311
  trentacinque_trentanove_anni quaranta_quarantaquattro_anni quarantacinque_quarantanove_anni quindici_quarantanove_anni
1                   0.42930013                     0.2854052                       -0.2236068                  0.5993694
2                   0.02995117                     0.7244901                        4.2485292                  0.1997898
3                   1.62734702                     1.1635749                       -0.2236068                  1.5317218
4                  -0.74877930                    -0.5927646                       -0.2236068                 -0.8391172

Clustering vector:
            Piemonte         Valle_d_Aosta               Liguria             Lombardia   Trentino_Alto_Adige
                   3                     2                     3                     1                     4
              Veneto Friuli_Venezia_Giulia        Emilia_Romagna               Toscana               Umbria
                   4                     4                     1                     1                     1
             Marche                 Lazio               Abruzzo                Molise              Campania
                   4                     1                     1                     4                     4
              Puglia            Basilicata              Calabria               Sicilia              Sardegna
                   3                     4                     4                     4                     4

Within cluster sum of squares by cluster:
[1]  9.364217  0.000000  6.232233 16.202738
 (between_SS / total_SS =  79.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

Il metodo k-means individua di nuovo la partizione in quattro cluster G_1*{Lombardia, Emilia Romagna, Toscana, Umbria, Lazio, Abruzzo}*, G_2*{Valle d'Aosta}*, G_3*{Piemonte, Liguria, Puglia}*, G_4*{Trentino Alto Adige, Veneto, Friuli Venezia Giulia, Marche, Molise, Campania, Basilicata, Calabria, Sicilia, Sardegna}*.

Come si può osservare dai risultati, l'algoritmo con dei fissati centroidi iniziali restituisce una partizione degli algoritmi diversa da quella precedentemente eseguita, però con valori pressochè simili.

Seconda parte

Di particolare importanza in statistica è *l'interferenza statistica*. Essa ha lo scopo di *estendere le misure ricavate dall'esame di un campione alla popolazione da cui è stato estratto.*
Uno dei problemi centrali dell'interferenza statistica è il seguente: *si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro non noto (o più parametri non noti).*
Il termine *osservabile* significa che si possono osservare i valori assunti dalla variabile aleatoria X e quindi il parametro non noto è presente soltanto nella legge di probabilità.
Affinché le conclusioni dell'interferenza statistica siano valide il campione deve essere scelto in modo tale da essere *rappresentativo della popolazione*.
L'interferenza statistica si basa su due metodi fondamentali di indagine: la *stima dei parametri* e la *verifica delle ipotesi*.

La *stima dei parametri* ha lo scopo di determinare i valori non noti dei parametri di una popolazione (come il valore medio, la varianza, …) per mezzo dei corrispondenti parametri derivati dal campione estratto dalla popolazione (come la media campionaria, la varianza campionaria, …). Si possono usare *stime puntuali* o *stime per intervallo*.
Si parla di *stima puntuale* quando si stima un parametro non noto di una popolazione usando un singolo valore reale.
Alla stima puntuale di un parametro non noto di una popolazione spesso si preferisce sostituire un intervallo di valori, detto *intervallo di confidenza*, ossia si cerca di determinare in base al campione osservato due limiti entro i quali sia compreso il parametro non noto con un certo *grado di confidenza*, detto anche *grado di fiducia*.

La *verifica delle ipotesi* è un procedimento che consiste nel fare una *congettura* o ipotesi *sul parametro non noto o sulla distribuzione di probabilità* e nel decidere, sulla base del campione estratto se essa è accettabile.
Per affrontare i problemi di stima (puntuale o per intervallo) dei parametri e della verifica delle ipotesi statistiche possono essere prese in considerazione variabili aleatorie discrete o continue con l'ausilio di R.

Mi occupo della stima puntuale e per intervallo e affronto alcuni problemi di verifica di ipotesi statistiche utilizzando una variabile aleatoria continua con una funzione di distribuzione normale.

1 DISTRIBUZIONE NORMALE

La funzione di distribuzione normale, detta anche di Gauss o gaussiana, riveste estrema importanza nel calcolo della probabilità e nella statistica.

Definizione: *Una variabile aleatoria X di densità di probabilità*

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad x \in \mathbb{R} \quad (\mu \in \mathbb{R}, \sigma > 0),$$

si dice avere distribuzione normale di parametri μ *e* σ.
La suddetta distribuzione è simmetrica rispetto all'asse $x = \mu$, infatti, per ogni $x \in \mathbb{R}$ risulta

$f_x(\mu - x) = f_x(\mu + x)$.
La densità $f_x(x)$ esibisce una caratteristica forma a campana, simmetrica rispetto a $x = \mu$.
La notazione $X \sim \mathcal{N}(\mu, \sigma)$ è utilizzata per indicare che X ha distribuzione normale dei parametri μ e σ, o più semplicemente che è una *variabile normale*.

In R la densità normale si calcola attraverso la funzione
dnorm(x, mean = mu, sd = sigma)
dove
- x è il valore assunto (o i valori assunti) dalla variabile aleatoria normale;
- mean e sd sono il valore medio e la deviazione standard della densità normale.

Ad esempio il seguente codice permette di visualizzare la densità di $X \sim \mathcal{N}(\mu, 1)$ con μ = -3, -2, -1, 0, 1, 2, 3.

```
> curve(dnorm(x, mean=-3, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", main="mu=-3,-2,-1,0,1,2,3; sigma=1")
> curve(dnorm(x, mean=-2, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", add=TRUE)
> curve(dnorm(x, mean=-1, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", add=TRUE)
> curve(dnorm(x, mean=0, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", add=TRUE, lty=2)
> curve(dnorm(x, mean=1, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", add=TRUE)
> curve(dnorm(x, mean=2, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", add=TRUE)
> curve(dnorm(x, mean=3, sd=1), from=-6, to=6, xlab="x", ylab="f(x)", add=TRUE)
```

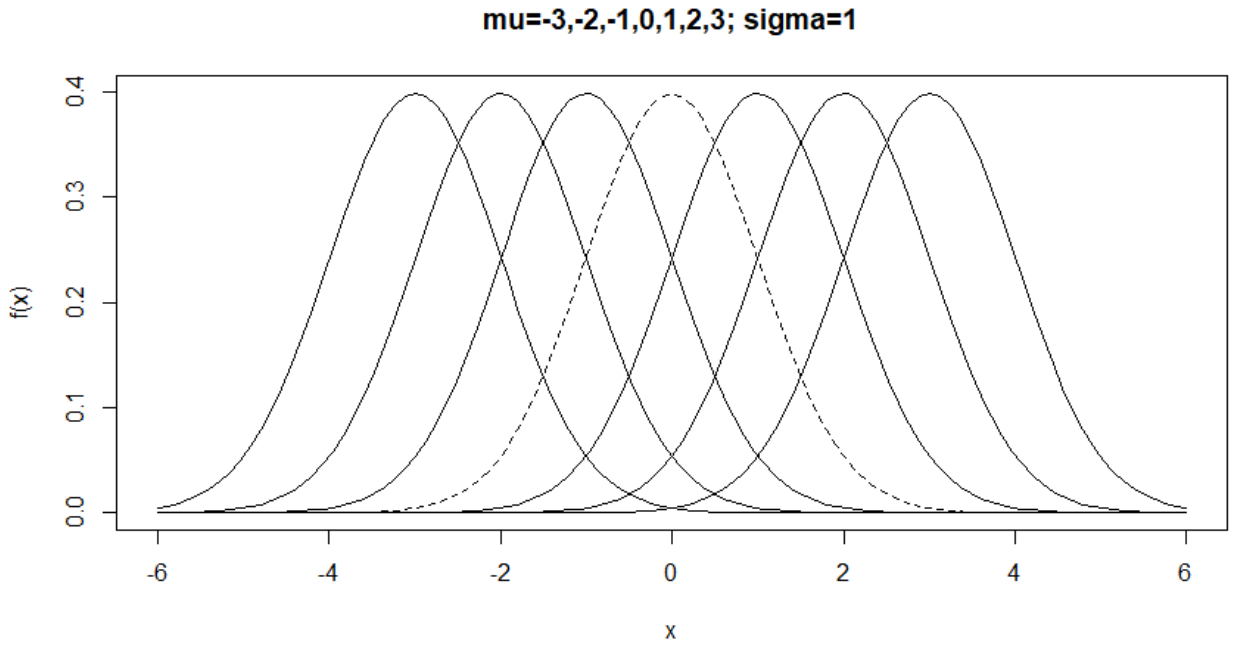


Figura1.1:Densità normale al variare di μ = -3,-2,-1,0,1,2,3 (da sinistra verso destra).

Come illustrato nella Figura 1.1 le variazioni del parametro μ comportano traslazioni della curva lungo l'asse delle ascisse; infatti al crescere del parametro μ la curva si sposta lungo l'asse delle ascisse senza cambiare forma.

Il parametro σ, pari alla semiampiezza tra i due punti di flesso, caratterizza la larghezza della funzione. Poiché l'ordinata massima è inversamente proporzionale a σ questa decresce, mentre l'area sottesa dalla densità deve rimanere unitaria.

Il seguente codice permette di visualizzare la densità di $X \sim \mathcal{N}(0, \sigma)$ con σ = 0.5, 1, 1.5.

```
> curve(dnorm(x, mean=0, sd=0.5), from=-4, to=4, xlab="x", ylab="f(x)", main="mu=0; sigma=0.5,1,1.5")
> curve(dnorm(x, mean=0, sd=1), from=-4, to=4, xlab="x", ylab="f(x)", add=TRUE, lty=2)
> curve(dnorm(x, mean=0, sd=1.5), from=-4, to=4, xlab="x", ylab="f(x)", add=TRUE)
```

il cui grafico è riportato in Figura 1.2. Si nota che al crescere di σ la curva diventa sempre più piatta, mentre al decrescere di σ essa si allunga verso l'alto restringendosi contemporaneamente ai lati.

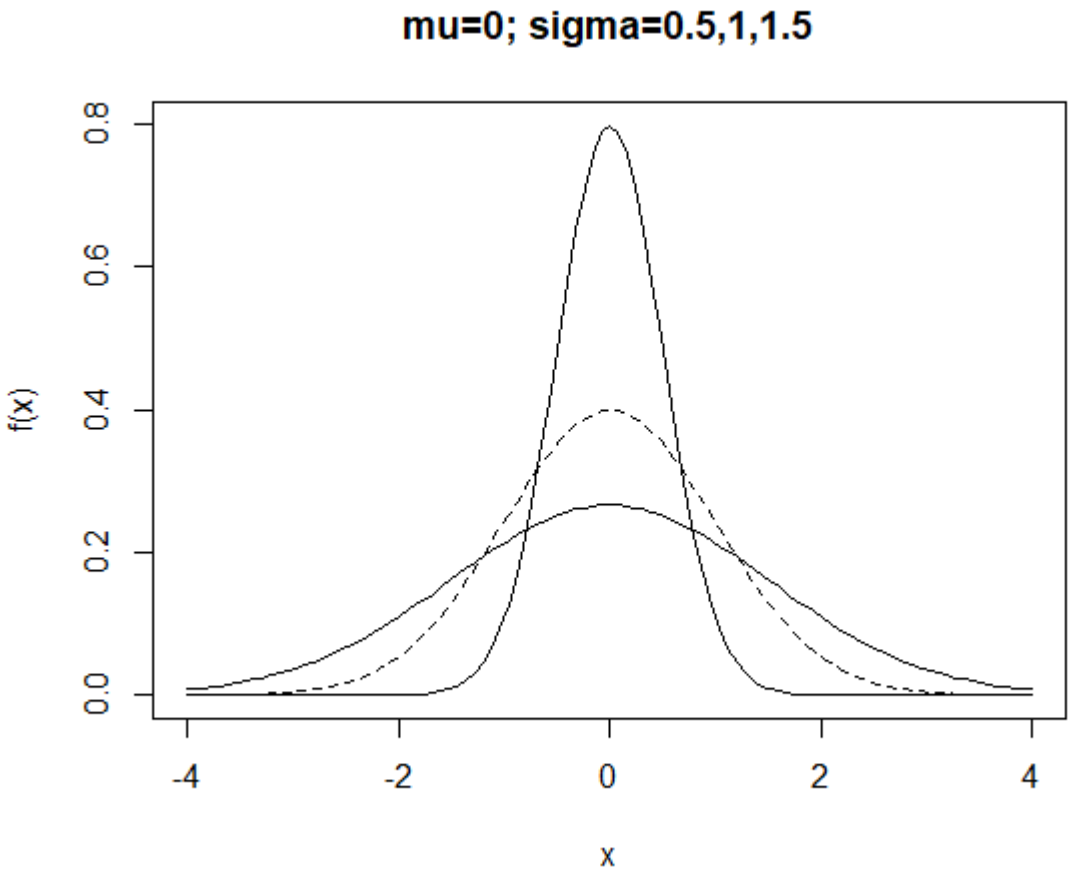


Figura 1.2:Densità normale al variare di σ = 0.5, 1, 1.5 (dall'alto verso il basso in prossimità dell'origine).

La funzione di distribuzione di una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ è:

$$F_x(x) = \int_{-\infty}^{x} f_x(y)\, dy \;=\; \phi\left(\frac{x-\mu}{\sigma}\right) \qquad\qquad x \in \mathbb{R}$$

è la funzione di distribuzione di una variabile aleatoria $Z \sim \mathcal{N}(0, 1)$, detta *normale standard*.

Pertanto, se $X \sim \mathcal{N}(\mu, \sigma)$ si ha:

$$P\,(a < X < b) = F_x(b) \;- F_x(a) \;=\; \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right).$$

In R la funzione di distribuzione di una variabile $X \sim \mathcal{N}(\mu, \sigma)$ si calcola tramite la funzione: pnorm(x, mean = mu, sd = sigma).

Il seguente codice permette di visualizzare la funzione di distribuzione di $X \sim \mathcal{N}(0, \sigma)$ con $\sigma = 0.5$, 1, 1.5:

```
> curve(pnorm(x,mean=0,sd=0.5), from=-4, to=4, xlab = "x", ylab = expression(P(X<=x)),main="mu
=0; sigma=0.5,1,1.5",lty=2)
> text(-0.4,0.8,"sigma=0.5")
> curve(pnorm(x,mean=0,sd=1),add=TRUE)
> arrows(-1,0.1,0.5,0.2, code = 1, length = 0.10)
> text(0.8,0.2,"sigma=1")
> curve(pnorm(x,mean=0,sd=1.5),add=TRUE, lty=3)
> text(-2.2,0.2,"sigma=1.5")
```

il cui grafico è riportato in Figura 1.3. La funzione arrows() ha come argomenti le due coordinate della linea della freccia, il parametro code può assumere valori 1,2,3 a seconda se la freccia deve essere unidirezionale verso sinistra, unidirezionale verso destra oppure bidirezionale; il parametro length fornisce invece la grandezza della freccia.

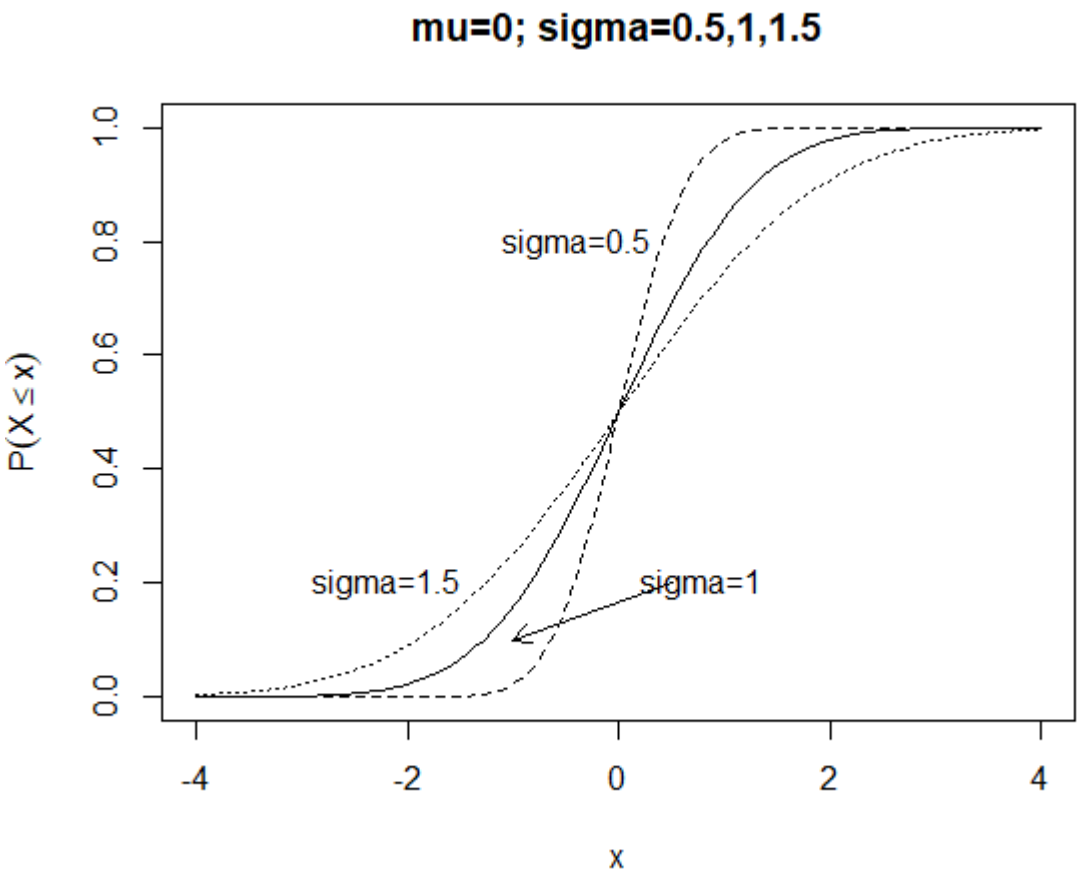


Figura 1.3:Funzione di distribuzione normale al variare di σ = 0.5, 1. 1.5.

Se si considera una variabile aleatoria normale $Z \sim \mathcal{N}(0, 1)$ si nota che

```
> pnorm(3,mean=0, sd=1)-pnorm(-3,mean=0,sd=1)
[1] 0.9973002
```

che mostra che per una qualsiasi variabile aleatoria normale $X \sim \mathcal{N}(\mu, \sigma)$ risulta

$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < \frac{x-\mu}{\sigma} < 3) = P(-3 < Z < 3) = 0.9973002.$

Quindi la probabilità che una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ assuma valori in un intervallo avente come centro μ e semiampiezza 3 σ è prossima all'unità.
Questa proprietà delle variabili aleatorie normali è nota come *regola del 3 σ*.
La regola del 3 σ permette di individuare l'intervallo (μ-3 σ, μ+3 σ) in cui rappresentare la funzione densità di una variabile normale di valore medio μ e varianza σ^2 in maniera tale che l'area sottesa dalla curva sia circa unitaria e l'area delle code destra e sinistra sia trascurabile.

In R si possono calcolare anche i quantili (percentili) della distribuzione normale attraverso la funzione
qnorm(z, mean = mu, sd = sigma).

Ad esempio, se si considera una variabile normale standard $Z \sim \mathcal{N}(0, 1)$, le seguenti linee di codice forniscono i quartili Q_1, Q_2, Q_3, Q_4

```
> z<-c(0,0.25,0.5,0.75,1)
> qnorm(z,mean=0,sd=1)
[1]       -Inf -0.6744898  0.0000000  0.6744898        Inf
```

che mostra che il primo quartile (25-esimo percentile) è Q_1 = - 0.6744898, il secondo quartile o mediana (50-esimo percentile) è Q_2= 0 e il terzo quartile (75-esimo percentile) è Q_3 = 0.6744898 (per la simmetria intorno all'origine della densità normale). Il minimo è $Q_0 = -\infty$ e il massimo è $Q_4 = \infty$.

È possibile simulare in R la variabile aleatoria normale generando una sequenza di numeri pseudocasuali mediante la funzione
rnorm(N, mean = mu, sd = sigma)

dove

- N è la lunghezza della sequenza da generare.
- mean e sd sono il valore medio e la deviazione standard della densità normale

Il seguente codice

```
> par(mfrow=c(2,2))
> curve(dnorm(x,mean=2,sd=1),from=-2,to=6,xlab="x",ylab="f(x)",ylim=c(0,0.5),main="Densità_normale, mu=2,sigma=1")
> sim1<-rnorm(500, mean=2, sd=1)
> hist(sim1,freq=F,xlim=c(-2,6),ylim=c(0,0.5),breaks = 100,xlab = "x",ylab = "Istogramma",main = "Densità_simulata,N=500")
> sim2<-rnorm(5000, mean=2, sd=1)
> hist(sim2,freq=F,xlim=c(-2,6),ylim=c(0,0.5),breaks = 100,xlab = "x",ylab = "Istogramma",main = "Densità_simulata,N=5000")
> sim3<-rnorm(50000, mean=2, sd=1)
> hist(sim3,freq=F,xlim=c(-2,6),ylim=c(0,0.5),breaks = 100,xlab = "x",ylab = "Istogramma",main = "Densità_simulata,N=50000")
```

permette di confrontare in Figura 1.4 la densità normale teorica con μ = 2, σ = 1 con la densità simulata scegliendo N=500, 5000, 50000. All'aumentare del numero di similitudini l'istogramma delle frequenze relative si avvicina sempre di più alla densità esponenziale teorica avente una forma a campana.

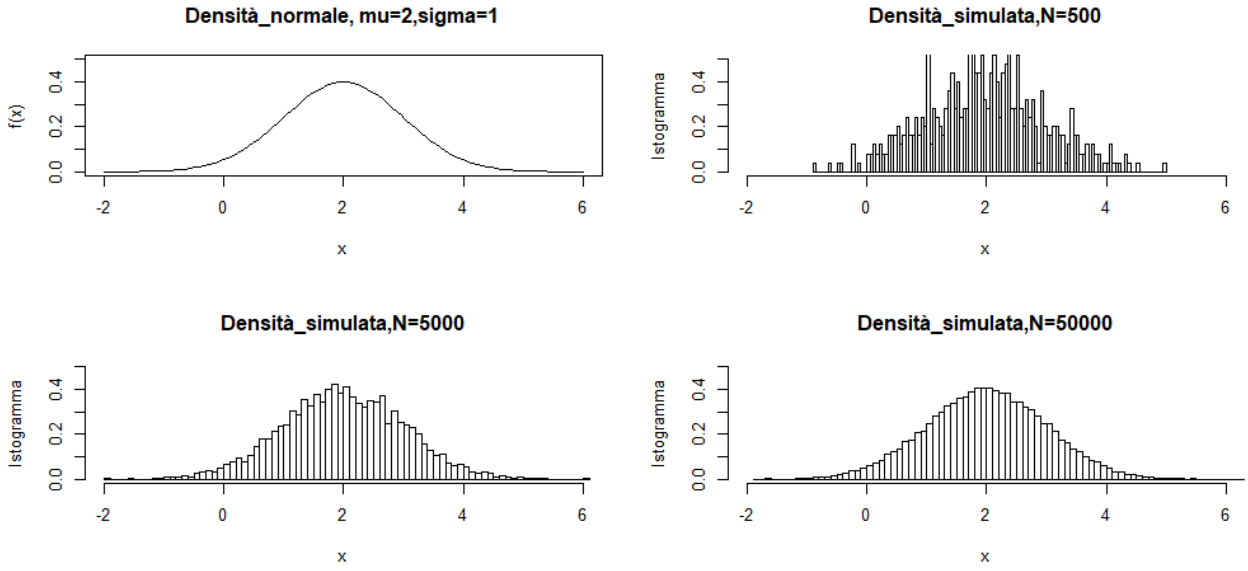


Figura 1.4:Confronto della densità normale con μ = 2 e σ = 1 con la densità simulata.

1.1 Approssimazione della distribuzione binomiale con la distribuzione Normale

Il calcolo delle probabilità binomiali diviene rapidamente oneroso al crescere di *n*. È quindi utile ricercare delle formule approssimate in grado di rendere agevole tale calcolo e, al contempo, accettabile l'errore derivante dall'approssimazione.
Si è preso in primo luogo in considerazione il *teorema di De Moivre-Laplace* e successivamente il *teorema centrale di convergenza*.

(Teorema di De Moivre-Laplace) *Sia $X_1, X_2, \ldots$ una successione di variabili aleatorie indipendenti distribuite alla Bernoulli con parametro p (0 < p < 1), e sia $Y_n = X_1 + X_2 + \ldots + X_n$. Allora per ogni x* € $\mathbb{R}$ *risulta:*

$$\lim_{n\to+\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\, dy$$

ossia $\frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z$

converge in distribuzione alla variabile aleatoria Z normale standard.

Ricordiamo che se $X_1, X_2, \ldots$ sono variabili aleatorie indipendenti di Bernoulli di parametro *p*, allora $Y_n = X_1 + X_2 + \ldots + X_n$ è una variabile aleatoria binomiale di valore medio *np* e varianza *np(1-p)*. Quindi sottraendo a Y_n la sua media e dividendo la differenza per la deviazione standard $\sqrt{\mathrm{np}(1-\mathrm{p})}$ si ottiene una variabile aleatoria standardizzata ($\phi(\frac{x-\mu}{\sigma})$) la cui funzione di distribuzione è per *n* grande approssimativamente normale standard. La bontà dell'approssimazione dipende da *n* e da *p* e migliora al tendere di *p* a 1/2. In generale si vuole assumere che l'approssimazione sia soddisfacente per $n > 10$ e per $5/n < p < 1 - 5/n$.

Si è esaminato l'approssimazione della binomiale alla normale

$$Yn \simeq np + \sqrt{np(1-p)}Z,$$

al variare di *n* con *p* fisso.

Il codice seguente confronta la densità normale di valore medio *np* e varianza *np(1-p)* e la funzione di probabilità binomiale per *n* = 25, 50, 75, 100 e *p* = 0.2

```
> par(mfrow=c(2,2))
> p<-0.2
> q<-1-p
> x<-0:25
> n<-25
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",m
ain="Binomiale,n=25,p=0.2")
> lines(x,dbinom(x,n,0.2), type = "h")
> x<-0:50
> n<-50
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",m
ain="Binomiale,n=50,p=0.2")
> lines(x,dbinom(x,n,0.2), type = "h")
> x<-0:75
> n<-75
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",m
ain="Binomiale,n=75,p=0.2")
> lines(x,dbinom(x,n,0.2), type = "h")
> x<-0:100
> n<-100
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",m
ain="Binomiale,n=100,p=0.2")
> lines(x,dbinom(x,n,0.2), type = "h")
```

il cui grafico è riportato in Figura 1.5. Si nota che l'approssimazione migliora al crescere di *n*.

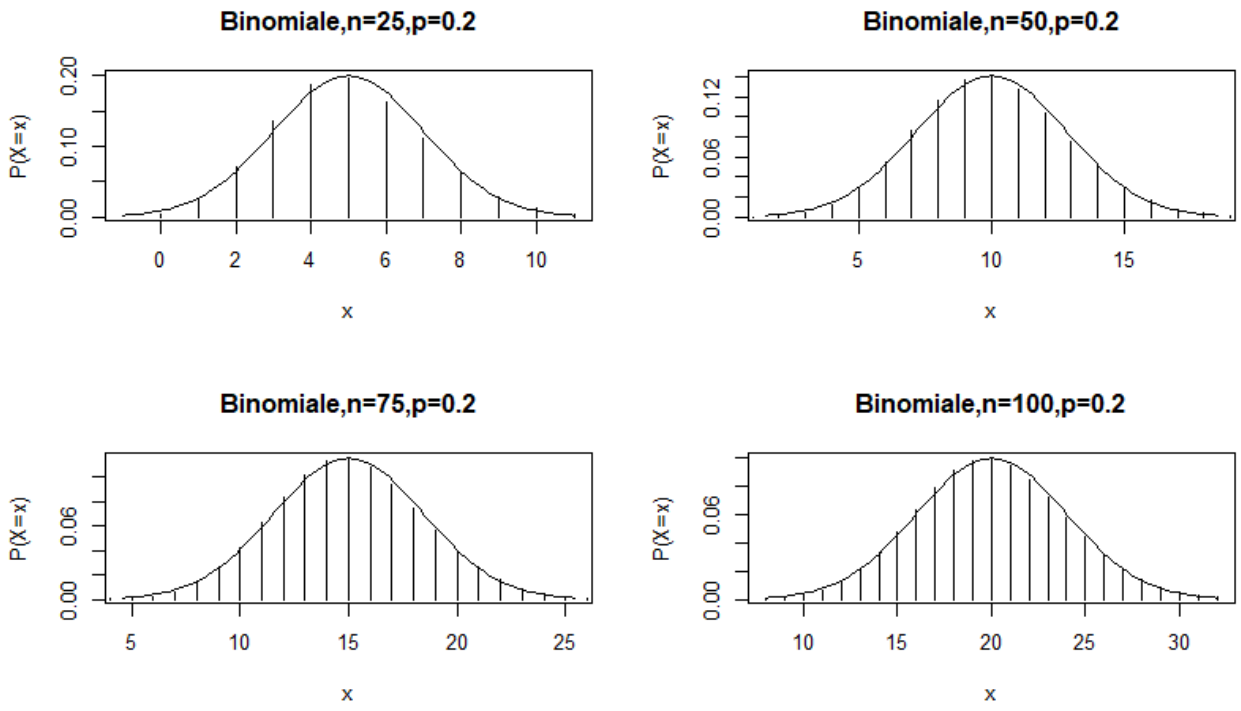


Figura 1.5 Confronto della probabilità della variabile Yn ~ $\mathcal{B}(n, 0.2)$ con la densità normale di valore medio μ= np e deviazione standard $\sigma = \sqrt{np(1-p)}$ per varie scelte di n.

Si è esaminata poi l'approssimazione della binomiale alla normale al variare di *p* con *n* fissato. Il seguente codice confronta la densità normale di valore medio *np* e varianza *np(1-p)* e la funzione di

probabilità binomiale per $n = 20$ e $p = 0.125, 0.25, 0.375, 0.5$

```
> par(mfrow=c(2,2))
> p<-0.125
> x<-0:20
> n<-20
> q<-1-p
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",
main="Binomiale,n=20,p=0.125")
> lines(x,dbinom(x,n,0.125), type = "h")
> p<-0.25
> q<-1-p
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",
main="Binomiale,n=20,p=0.25")
> lines(x,dbinom(x,n,0.25), type = "h")
> p<-0.375
> q<-1-p
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",
main="Binomiale,n=20,p=0.375")
> lines(x,dbinom(x,n,0.375), type = "h")
> p<-0.5
> q<-1-p
> curve(dnorm(x,n*p,sqrt(n*p*q)),from=n*p-3*sqrt(n*p*q),to=n*p+3*sqrt(n*p*q),xlab = "x", ylab = "P(X=x)",
main="Binomiale,n=20,p=0.5")
> lines(x,dbinom(x,n,0.5), type = "h")
```

il cui grafico è riportato in Figura 1.6. Si nota che l'approssimazione non è buona per piccoli valori di p e migliora al tendere di p a 1/2, diventando poi eccellente quando p = 1/2.

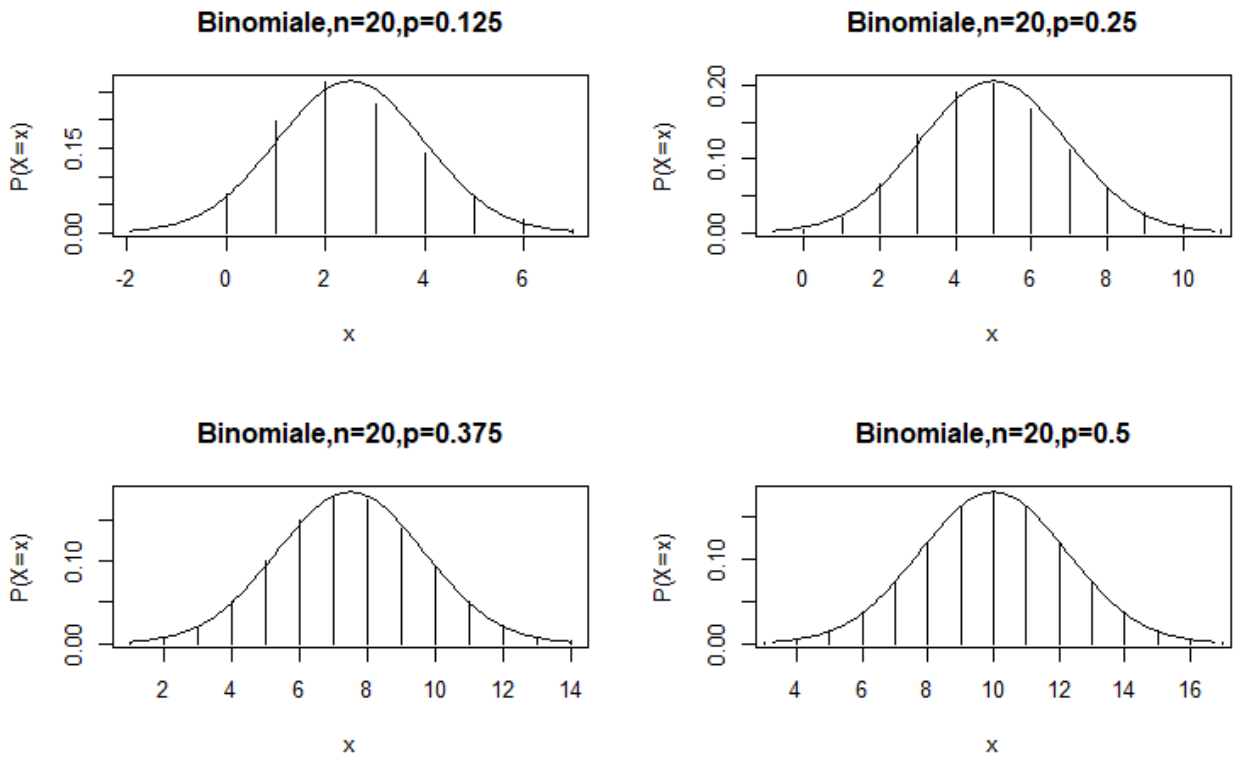


Figura1.0:6:Confronto della probabilità della variabile $YX \sim \mathcal{B}(20, p)$ con la densità normale di valore medio μ= np e deviazione standard $\sigma = np(1 - p)$ per varie scelte di p.

1.2 Teorema centrale di convergenza

Si è poi introdotto uno dei più importanti risultati della teoria della probabilità, noto quale *teorema centrale di convergenza* o *teorema centrale di limite*, che fornisce una semplice ed utile approssimazione alla distribuzione della somma di variabili aleatorie indipendenti.

(Teorema centrale di convergenza) *Sia* $X_1, X_2, \ldots$ *una successione di variabili aleatorie, definite nello stesso spazio di probabilità, indipendenti e identicamente distribuite con valore medio* μ *finito e varianza* σ^2 *finita e positiva. Posto per ogni intero n positivo* $Y_n = X_1 + X_2 + \ldots + X_n$, *per ogni* $x \in \mathbb{R}$ *risulta:*

$$\lim_{n\to+\infty} P\left(\frac{Y_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\, dy = \theta(x)$$

ossia la successione delle variabili aleatorie standardizzate $\frac{y_n - E(y_n)}{\sqrt{var(Y_n)}} = \frac{Y_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z$, *converge in distribuzione alla variabile aleatoria normale standard.*

(Caso particolare) Se, ad esempio, supponiamo che $X_1, X_2, \ldots$ è una *successione di variabili aleatorie indipendenti di Poisson* di parametro λ allora $Y_n = X_1 + X_2 + \ldots + X_n$ è ancora una variabile aleatoria di Poisson di parametro $n\,\lambda$. Quindi, il teorema centrale di convergenza afferma che per n grande la distribuzione di $Y_n = X_1 + X_2 + \ldots + X_n$ è approssimativamente normale con valore medio $n\,\lambda$ e varianza $n\,\lambda$, ossia

$$Y_n \sim n\lambda + \sqrt{n\lambda}\, Z$$

dove $n\lambda + \sqrt{n\lambda}\, Z$ è una variabile aleatoria con densità normale di valore medio e varianza $n\,\lambda$.

Si è esaminata l'approssimazione della distribuzione di Poisson di parametro $n\,\lambda$ alla normale di valore medio e varianza $n\,\lambda$ al variare del parametro $n\,\lambda$. Il seguente codice

```
> par(mfrow=c(2,2))
> x<-0:100
> curve(dnorm(x,5,sqrt(5)),from=5-3*sqrt(5),to=5+3*sqrt(5),xlab = "x", ylab = "P(X=x)",main="Poisson,n_la
mbda=5")
> lines(x,dpois(x,5), type = "h")
> curve(dnorm(x,10,sqrt(10)),from=10-3*sqrt(10),to=10+3*sqrt(10),xlab = "x", ylab = "P(X=x)",main="Poisso
n,n_lambda=10")
> lines(x,dpois(x,10), type = "h")
> curve(dnorm(x,25,sqrt(25)),from=25-3*sqrt(25),to=25+3*sqrt(25),xlab = "x", ylab = "P(X=x)",main="Poisso
n,n_lambda=25")
> lines(x,dpois(x,25), type = "h")
> curve(dnorm(x,50,sqrt(50)),from=50-3*sqrt(50),to=50+3*sqrt(50),xlab = "x", ylab = "P(X=x)",main="Poisso
n,n_lambda=50")
> lines(x,dpois(x,50), type = "h")
```

permette di visualizzare la Figura 1.7 in cui si confronta la probabilità di Poisson di parametro $n\,\lambda$ con la densità normale di valore medio e varianza $n\,\lambda = 5, 10, 25, 50$. Si nota che al crescere di $n\,\lambda$ aumenta l'accuratezza dell'approssimazione.

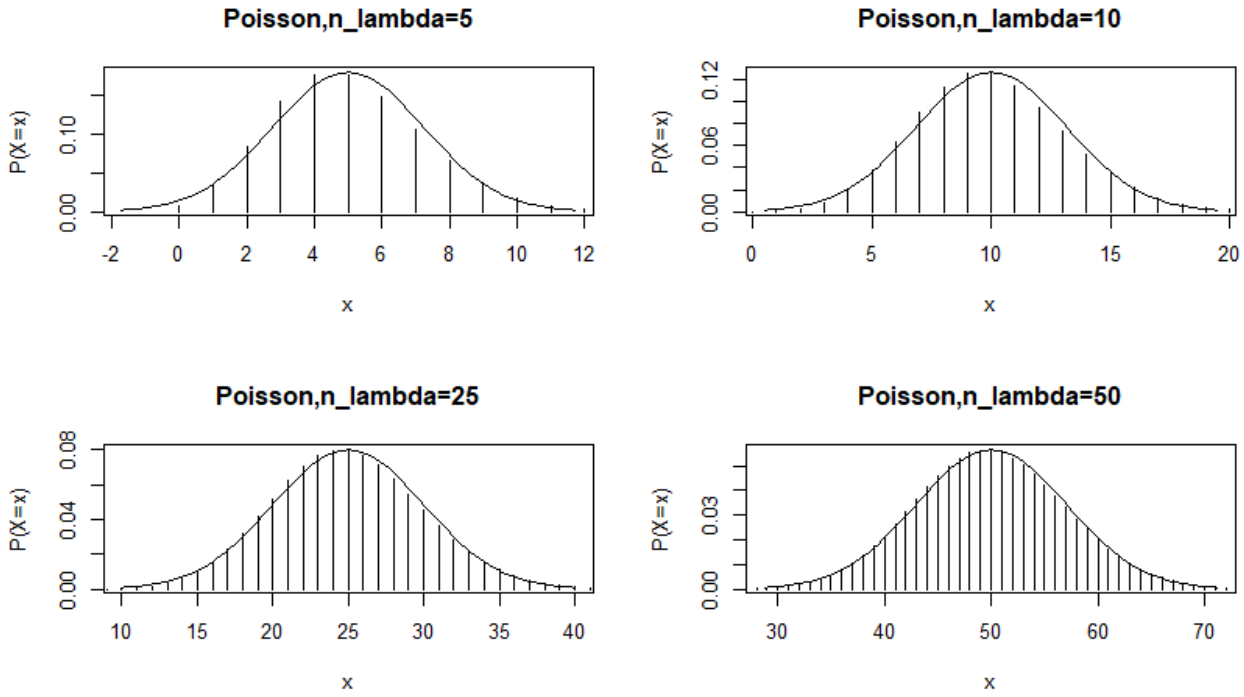


Figura 1.7 Confronto della probabilità di Poisson della variabile $X \sim \mathcal{P}(n\lambda)$ con la densità normale di valore medio μ=n λ e deviazione standard $\sigma = n\,\lambda$ per varie scelte di n λ.

2 ANALISI DI UN CAMPIONE NORMALE

Al fine di effettuare varie analisi e stime relative ad una variabile aleatoria normale, è stato generato un campione facente parte di una popolazione normale, servendosi delle seguenti istruzioni:

```
> camp<-rnorm(100,mean=2,sd=2)
```

Tale campione ha N=100, valore medio $\mu = 2$, varianza $\sigma^2 = 4$ e deviazione standard $\sigma = 2$. Tuttavia, essendo stato generato in maniera pseudocasuale, i valori di media campionaria, varianza campionaria e deviazione standard campionaria del campione sono i seguenti:

Media campionaria	2.146789
Varianza campionaria	4.881814
Deviazione standard	2.209483

Calcolati utilizzando le seguenti linee di codice:

```
> mean(camp)
[1] 2.146789
> var(camp)
[1] 4.881814
> sd(camp)
[1] 2.209483
```

Altre misure di centralità e dispersione sono le seguenti:

```
> median(camp)
[1] 2.171051
> quantile(camp)
       0%       25%       50%       75%      100%
-3.520922  0.940640  2.171051  3.293973  9.131958
```

che si possono vedere anche nel boxplot in Figura 2.1 generato mediante il seguente codice:

```
> boxplot(camp,horizontal = T,main="Boxplot_relativo_al_Campione_normale",col="green")
```

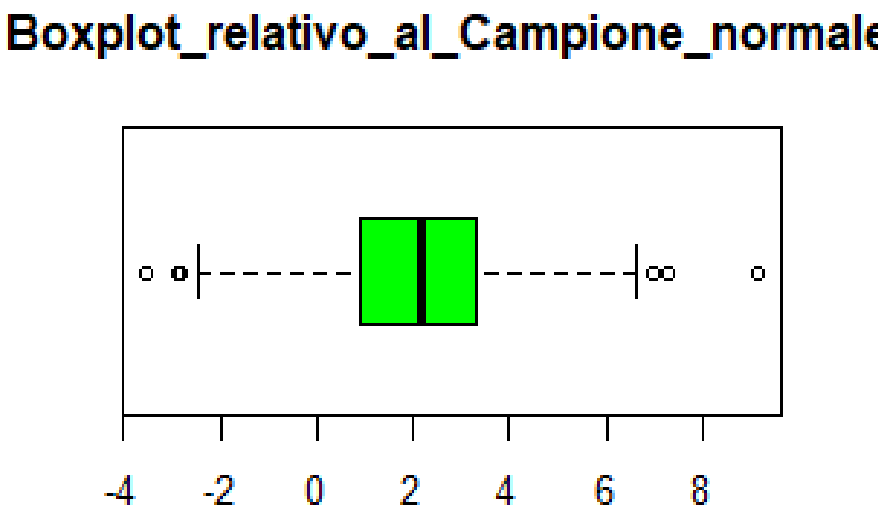


Figura 2.1: Boxplot relativo al campione normale

Che mostra proprio che il primo quartile (25-esimo percentile) è $Q_1 = 0.940640$, il secondo quartile o mediana (50-esimo percentile) è $Q_2 = 2.171051$ e il terzo quartile (75-esimo percentile) è $Q_3 = 3.293973$ (per la simmetria intorno all'origine della densità normale). Il minimo è $Q_0 = -3.520922$ e il massimo è $Q_4 = 9.131958$.

Inoltre, da una prima osservazione al boxplot, il campione risulta centrato con mediana posta quasi in corrispondenza del valore 2. La distribuzione è anche abbastanza simmetrica da entrambi i lati. Questa ipotesi può essere verificata mediante il seguente grafico in Figura 2.2 prodotto dalle seguenti linee di codice:

```
> hist(camp,freq = F,xlim = c(-4,8),ylim = c(0,0.5),breaks = 100,xlab = "x",ylab = "Istogramma",
main = "Densità_simulata,N=100")
```

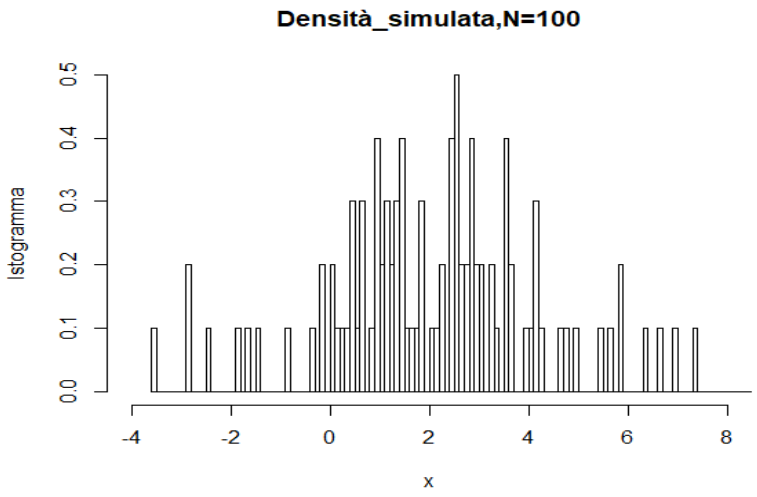


Figura 2.2: Grafico della densità simulata.

3 STIMA PUNTUALE

3.1 Campioni casuali e stimatori

Uno dei problemi centrali dell'inferenza statistica è il seguente: *si desidera studiare una popolazione descritta da una variabile aleatoria X la cui funzione di distribuzione ha una forma nota ma contiene un parametro $\vartheta \epsilon \theta$ non noto (o più parametri non noti).*

Quello che verrà fatto sarà generare dei valori casuali così da poter calcolare una stima puntuale dei valori non noti.
Poi si estrarrà un campione rappresentativo della popolazione così da poter fare delle inferenze su queste ultime, così da validare i risultati ottenuti.

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

Definizione stimatore: *Uno stimatore $\theta = t(X1, X2, \ldots, Xn)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, \ldots, X_n$ 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 ϑ.*
Statistiche tipiche sono la *media campionaria* e la *varianza campionaria*.

Definizione statistica: *Sia $X_1, X_2, \ldots, X_n$ un campione casuale. La statistica*

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

è detta media campionaria, mentre la statistica

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

è detta varianza campionaria.

Proposizione *Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione descritta da una variabile osservabile X caratterizzata da valore medio $E(X) = \mu$ finito e varianza $Var(X) = \sigma^2$. Risulta:*

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

Per la proprietà di linearità del valore medio e l'identica distribuzione delle variabili aleatorie che costituiscono il campione, dalla proposizione si ha:

$$E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\, E(\mathrm{X}) = \mu$$

Inoltre, poiché le variabili aleatorie che costituiscono sono indipendenti ed identicamente distribuite, dalla proposizione si ottiene:

$$Var(\bar{X}) = Var[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\, n\, Var(\mathrm{X}) = \frac{\sigma^2}{n}$$

La proposizione mostra che al crescere dell'ampiezza del campione la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione. Inoltre dal teorema centrale di convergenza della probabilità scaturisce che per n sufficientemente grande (campioni di grande ampiezza) la funzione di distribuzione della media campionaria $\bar{X}$ è approssimativamente normale con valore medio μ e varianza $\frac{\sigma^2}{n}$.

3.2 Metodi per la ricerca di stimatori

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione con funzione di probabilità oppure densità di probabilità $f(x; \vartheta 1, \vartheta 2, \ldots, \vartheta k)$ dove $\vartheta 1, \vartheta 2, \ldots, \vartheta n$ denotano i parametri non noti della popolazione. Lo scopo del decisore è quello di stimare i parametri non noti della popolazione. I principali metodi di stima puntuale dei parametri sono il *metodo dei momenti* e *il metodo della massima verosimiglianza*.

3.2.1 Metodo dei momenti

Il metodo dei momenti è uno de più antichi metodi di stima dei parametri. Per illustrarlo occorre in primo luogo definire i *momenti campionari*.

Definizione momento campionario: *Si definisce momento campionario r-esimo relativo ai valori osservati ($x_1, x_2, \ldots, x_n$) del campione casuale il valore*

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

Si nota che il momento campionario *r*-esimo è la media aritmetica delle potenze *r*-esime delle *n* osservazioni effettuate sulla popolazione. In particolare, se $r = 1$ il momento campionario $M_1 = (x_1, x_2, \ldots, x_n)$ coincide con il valore osservato dalla media campionaria $\bar{X}$, ossia $M_1 = (x_1 + x_2 + \ldots + x_n)/n$. 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

$$\mathrm{E}(X^r) = M_r(x_1, x_2, \ldots, x_n) \qquad (r = 1,2,\ldots,\mathrm{k})$$

Le incognite del sistema sono i parametri $\vartheta 1, \vartheta 2, \ldots, \vartheta k$
Affinché il metodo dei momenti sia utilizzabile occorre che il sistema ammetta un'unica soluzione. Le stime calcolate dipenderanno quindi dal campione osservato, quindi al variare del campione, cambieranno anche gli stimatori di questi ultimi.
Questi stimatori vengono chiamati *stimatori del metodo dei momenti*.

- **(Popolazione normale)** Si è interessati a determinare con il metodo dei momenti gli stimatori dei parametri μ e σ^2 di una popolazione normale.

Occorre quindi stimare due parametri μ e σ^2. Poiché $E(X) = \mu$ e $E(X^2) = \sigma^2 + \mu^2$ si ha un sistema di due equazioni, poiché sono due i parametri da stimare:

$$\begin{cases} E(X) = \bar{x} \\ E(X^2) = M_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \widehat{\sigma^2} = M_2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \end{cases}$$

Il metodo dei momenti fornisce quindi come stimatore del valore medio μ la media campionaria $\overline{X}$ e come stimatore della varianza σ^2 la variabile aleatoria $(n-1)\,S^2/n$. Si osserva quindi che per stimare la varianza il campione deve essere sufficientemente grande.

Considerando il nostro campione camp

```
> stimamu<-mean(camp)
> stimamu
[1] 2.146789
> stimasigma2<-(length(camp)-1)*var(camp)/length(camp)
> stimasigma2
[1] 4.832995
```

la stima del parametro μ con il metodo dei momenti è $\hat{\mu}$=2.146789 e la stima del parametro σ^2 con il metodo dei momenti è $\widehat{\sigma^2}$=4.832995.

3.2.2 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 illustrare il metodo della massima verosimiglianza occorre introdurre in primo luogo la *funzione di verisimiglianza.*

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

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

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$. 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 continua) è *più verosimile* che provenga il campione osservato *$(x_1, x_2, \ldots, x_n)$*.

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

Le stime dipendono dal campione osservato al variare dei possibili campioni osservati, si ottengono gli stimatori di massima verosimiglianza $\widehat{\Theta_1}$,$\widehat{\Theta_2}$, …, $\widehat{\Theta_k}$ dei parametri non noti detti *stimatori di massima verosimiglianza.*

- **(Popolazione normale)** Si desidera determinare lo stimatore di massima verosimiglianza dei parametri μ e σ^2 di una popolazione normale.

Le stime di massima verosimiglianza dei parametri μ e σ^2 sono rispettivamente

$$\hat{\mu} = \overline{x}, \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Lo stimatore di massima verosimiglianza di μ è la media campionaria $\overline{X}$. Invece lo stimatore di σ^2 è $(n-1)S^2/n$. Entrambi gli stimatori coincidono con quelli ottenuti con il metodo dei momenti.

Quindi considerando il nostro campione la stima del parametro μ con il metodo della massima verosimiglianza è $\hat{\mu}$=2.146789 e la stima del parametro σ^2 con il metodo della massima verosimiglianza è $\widehat{\sigma^2}$=4.832995.

3.2 Proprietà degli stimatori

In generale esistono molti stimatori che possono essere utilizzati per stimare il parametro non noto di una popolazione. Occorre quindi definire alcune proprietà di cui può o meno godere uno stimatore. Uno stimatore può essere: *corretto* (o equivalentemente *non distorto*), *più efficiente di un altro, corretto e con varianza uniformemente minima, asintoticamente corretto* e *consistente*.

Definizione stimatore corretto: *Uno stimatore* $\widehat{\Theta} = t(X_1, X_2, \dots, X_n)$ *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.*

Nel caso degli stimatori trovati precedentemente, sia nel metodo della massima verosimiglianza sia del metodo dei momenti, abbiamo riscontrato che il valore medio dello stimatore è proprio uguale alla media campionaria $\bar{X}$, quindi si può affermare che lo stimatore risulta essere corretto per il parametro non noto della popolazione.

Occorre sottolineare che possono esistere differenti stimatori corretti di un parametro non noto di una popolazione, dovendo scegliere quello più efficiente, abbiamo bisogno di un metodo per confrontare questi stimatori.

Un modo è quello di utilizzare l'*errore quadratico medio.*

Definizione errore quadratico medio: *Sia* $\widehat{\Theta} = t(X_1, X_2, \dots, X_n)$ *uno stimatore del parametro non noto* ϑ *della popolazione. Si chiama errore quadratico medio la quantità*

$$MSE(\widehat{\Theta}) = E[(\widehat{\Theta} - \vartheta)^2].$$

Si può scrivere anche in termini di varianza nel seguente modo

$$MSE(\widehat{\Theta}) = Var(\widehat{\Theta}) + [E(\widehat{\Theta}) - \vartheta]^2.$$

Il problema principale del decisore consiste nello specificare lo stimatore migliore del parametro ϑ, ossia lo stimatore che ha il più piccolo errore quadratico medio per ogni valore ammissibile di $\vartheta \in \Theta$.

La ricerca dello stimatore con errore quadratico uniformemente minimo deve essere effettuata in opportune classi, ad esempio nella classe degli stimatori corretti.

4 INTERVALLI DI CONFIDENZA

4.1 Intervalli di confidenza

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*, ossia si cerca di determinare in base ai dati del campione, due limiti (uno inferiore e uno superiore) entro i quali sia compreso il parametro non noto con un certo *coefficiente di confidenza* (detto anche *grado di fiducia*).

Sia $X_1, X_2, \ldots, X_n$ un campione casuale di ampiezza *n* estratto da una popolazione con funzione di probabilità oppure densità di probabilità $f(x;\vartheta)$, dove ϑ denota il parametro non noto della popolazione.

Denotiamo con $\underline{C_n}$=g1(X1,X2,…,Xn) e con $\overline{C_n}$=g2(X1,X2,…,Xn) due statistiche, cioè due funzioni osservabili del campione casuale, che soddisfino la condizione $\underline{C_n} < \overline{C}_n$,dove per ogni campione fissato x=(x1,x2,…,xn) valga g1(x)<g2(x).

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

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

allora si dice che $\left(\underline{C_n}, \overline{C}_n\right)$ *è un intervallo di confidenza 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.*

In generale ci sono numerosi intervalli di confidenza dello stesso grado 1- α per un parametro non noto δ della popolazione. La scelta dell'intervallo deve essere fatta in base ad alcune proprietà statistiche, quali ad esempio, fissato un intervallo 1- α, si vuole che la lunghezza dell'intervallo di confidenza:

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

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

Metodo pivotale

Un metodo per la costruzione degli intervalli di confidenza è il *metodo pivotale*. Tale metodo consiste essenzialmente nel determinare una variabile aleatoria di pivot $\gamma(X_1, X_2, \ldots, X_n; \vartheta)$ che dipende dal campione casuale $X_1, X_2, \ldots, X_n$ e dal parametro non noto ϑ e la cui *funzione di distribuzione non contiene il parametro da stimare*. Tale variabile aleatoria non è una statistica poiché dipende dal parametro non noto ϑ e quindi non è osservabile.

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

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

Se per ogni possibile campione osservato (x_1, x_2, …, x_n) e per ogni $\vartheta \in \Theta$, si riesce a dimostrare che

$$\alpha_1 < \gamma(\mathrm{x}; \vartheta) < \alpha_2 \Leftrightarrow \qquad\qquad g_1(\mathrm{x}) < \vartheta < g_2(\mathrm{x})$$

con $g_1(\mathrm{x})$ e $g_2(\mathrm{x})$ dipendenti soltanto dal campione osservato, che è equivalente dire

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

Denotando con $\underline{C}_n = g_1(X_1, X_2, \dots, X_n)$ e $\overline{C}_n = g_2(X_1, X_2, \dots, X_n)$, dalla definizione segue che $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado 1- α per il parametro non noto ϑ della popolazione.

4.2 Popolazione normale

Sia X_1, X_2, …, X_n un campione casuale di ampiezza *n* estratto da una popolazione normale con valore medio μ e varianza σ^2. Si possono analizzare i seguenti problemi:

- *(i)* determinare un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota;
- *(ii)* determinare un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
- *(iii)* determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;
- *(iv)* determinare un intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

➢ (Intervallo di confidenza per μ con σ^2 nota)

Proposizione *Sia (x_1, x_2, …, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2. Una stima dell'intervallo di confidenza di grado* $1 - \alpha$ *per il valore medio μ è*

$$\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

dove $\overline{x}_n$ denota la media campionaria delle n osservazioni.

Considerato il campione precedentemente generato formato da n = 100 osservazioni si è trovata che $\overline{x}_{100} = 2.146789$. Supponendo che sia nota la varianza $\sigma^2 = 4$, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il valore medio μ.

In questo caso $\alpha = 0.05$ e $\alpha/2 = 0.025$. Il valore $z_{\alpha/2} = z_{0.025}$ può essere determinato tramite R.

```
> alpha<-1-0.95
> qnorm(1-alpha/2, mean = 0, sd=1)
[1] 1.959964
> n<-length(camp)
> mean(camp)-qnorm(1-alpha/2, mean = 0, sd=1)*2/sqrt(n)
[1] 1.754796
> mean(camp)+qnorm(1-alpha/2, mean = 0, sd=1)*2/sqrt(n)
[1] 2.538782
```

Si nota che $z_{0.025} = 1.959964$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il valore medio μ è quindi (1.754796, 2.538782). Si nota che la media campionaria $\overline{x}_{100}$ è compresa nell'intervallo.

- **(Intervallo di confidenza per μ con varianza non nota)**

Proposizione *Sia (x_1, x_2, …, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con varianza non nota. Una stima dell'intervallo di confidenza di grado* $1 - \alpha$ *per il valore medio* μ *è*

$$\overline{x}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} < \mu < \overline{x}_n + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}},$$

dove $\overline{x}_n$ *e* s_n *denotano rispettivamente la media campionaria e la deviazione standard campionaria delle n osservazioni.*

Considerato il campione precedentemente generato formato da n = 100 osservazioni si sono trovati che $\overline{x}_{100} = 2.146789$ e $s_{100} = 2.209483$. Si è determinata una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il valore medio μ.

In questo caso $\alpha = 0.05$ e $\alpha/2 = 0.025$. Il valore $t_{\alpha/2,n-1} = t_{0.025,99}$ può essere determinato tramite R.

```
> alpha<-1-0.95
> n<-length(camp)
> qt(1-alpha/2, df=n-1)
[1] 1.984217
> mean(camp)+qt(1-alpha/2, df=n-1)*sd(camp)/sqrt(n)
[1] 2.585198
> mean(camp)-qt(1-alpha/2, df=n-1)*sd(camp)/sqrt(n)
[1] 1.70838
```

Si nota che $t_{0.025,99} = 1.984271$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il valore medio μ è quindi (1.70838, 2.585198). Si nota che la media campionaria $\overline{x}_{100}$ è compresa nell'intervallo.

Inoltre, si è determinata una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per il valore medio μ.

In questo caso $\alpha = 0.01$ e $\alpha/2 = 0.005$. Il valore $t_{\alpha/2,n-1} = t_{0.005,99}$ può essere determinato tramite R.

```
> alpha<-1-0.99
> n<-length(camp)
> qt(1-alpha/2, df=n-1)
[1] 2.626405
> mean(camp)-qt(1-alpha/2, df=n-1)*sd(camp)/sqrt(n)
[1] 1.566489
> mean(camp)+qt(1-alpha/2, df=n-1)*sd(camp)/sqrt(n)
[1] 2.727089
```

Si nota che $t_{0.025,99} = 2.626405$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per il valore medio μ è quindi (1.566489, 2.727089). Si nota che la media campionaria $\overline{x}_{100}$ è compresa nell'intervallo.

Si nota che aumentando il grado di fiducia aumenta la lunghezza dell'intervallo di confidenza.

- **(Intervallo di confidenza per σ^2 con μ noto)**

Proposizione *Sia (x_1, x_2, …, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio noto μ. Una stima dell'intervallo di confidenza di grado* $1 - \alpha$ *per la varianza* σ^2 *è*

$$\frac{(n-1)s_n^2 + n(\overline{x_n} - \mu)^2}{x^2_{\frac{\alpha}{2},n}} < \sigma^2 < \frac{(n-1)s_n^2 + n(\overline{x_n} - \mu)^2}{x^2_{1-\frac{\alpha}{2},n}}$$

dove $\overline{x}_n$ *e* s_n^2 *denotano rispettivamente la media campionaria e la varianza campionaria delle n osservazioni.*

Considerato il campione precedentemente generato formato da n = 100 osservazioni si sono trovati che $\overline{x}_{100} = 2.146789$ e $s^2_{100} = 4.881814$.. Supponendo che sia noto il valore medio $\mu = 2$ e varianza non nota σ^2, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza σ^2.

In questo caso $\alpha = 0.05$ e $\alpha/2 = 0.025$ e 1-$\alpha/2 = 0.975$. I valori $x^2_{1-\frac{\alpha}{2},n} = x^2_{0.975,100}$ possono essere ottenuti tramite R.

```
> n<-length(camp)
> mu<-2
> alpha<-1-0.95
> qchisq(alpha/2, df=n)
[1] 74.22193
> qchisq(1-alpha/2, df=n)
[1] 129.5612
> ((n-1)*var(camp)+n*(mean(camp)-mu)**2)/qchisq(1-alpha/2,df=n)
[1] 3.746911
> ((n-1)*var(camp)+n*(mean(camp)-mu)**2)/qchisq(alpha/2,df=n)
[1] 6.540577
```

Si nota che $x^2_{1-\frac{\alpha}{2},n} = x^2_{0.975,100}$= 129.5612 e $x^2_{\frac{\alpha}{2},n} = x^2_{0.025,100} = 74.22193$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza è quindi (3.746911, 6.540577). Si nota che la varianza campionaria σ^2 è compresa nell'intervallo.

- **(Intervallo di confidenza per σ^2 con valore medio non noto)**

Proposizione *Sia (x_1, x_2, …, x_n) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio non noto. Una stima dell'intervallo di confidenza di grado* $1 - \alpha$ *per la varianza* σ^2 *è*

$$\frac{(n-1)s_n^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s_n^2}{\chi^2_{1-\alpha/2,n-1}},$$

dove s_n^2 *denota la varianza campionaria delle n osservazioni.*

Considerato il campione precedentemente generato formato da n = 100 osservazioni si è trovata che $s^2_{100} = 4.881814$. Si è determinata una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza σ^2.

In questo caso $\alpha = 0.05$ e $\alpha/2 = 0.025$ e 1-$\alpha/2 = 0.975$. I valori $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,99}$ e possono essere ottenuti tramite R.

```
> n<-length(camp)
> alpha<-1-0.95
> qchisq(alpha/2, df=n-1)
[1] 73.36108
> qchisq(1-alpha/2, df=n-1)
[1] 128.422
> (n-1)*var(camp)/qchisq(1-alpha/2,df=n-1)
[1] 3.763371
> (n-1)*var(camp)/qchisq(alpha/2,df=n-1)
[1] 6.587956
```

Si nota che $\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,99} = 128.422$ e $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,99} = 73.36108$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza σ^2 è quindi (3.763371, 6.587956).

Per una popolazione normale le stime per intervallo del valore medio e della varianza della popolazione possono essere effettuate qualsiasi sia la dimensione del campione casuale osservato. Occorre infine sottolineare che per una popolazione normale i metodi di stima maggiormente utilizzati sono il *(ii)* e il *(iv)*, ossia la determinazione di un intervallo di confidenza di grado $1 - \alpha$ per il valore medio μ nel caso in cui la varianza della popolazione normale è non nota e la determinazione di un intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

5 INTERVALLI DI FIDUCIA APPROSSIMATI

Desideriamo analizzare alcuni problemi in cui è richiesto il confronto tra i valori medi di due differenti popolazioni.

5.1 Differenza tra i valori medi

Si sono costruiti ora degli intervalli di confidenza per la differenza tra i valori medi di due popolazioni normali.

Popolazioni normali

Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$. Vogliamo analizzare i seguenti problemi:

(i) determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 sono note;

(ii) determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando le varianze σ_1^2 e σ_2^2 sono non note per campioni numerosi estratti dalle due popolazioni.

Al fine di costruire gli intervalli di confidenza per la differenza tra i valori medi di due popolazioni normali, è stato generato un altro campione di taglia 150 facente parte di una popolazione normale servendosi delle seguenti istruzioni:

```
> camp2<-rnorm(150, mean = 3, sd=3)
```

Tale campione ha valore medio $\mu = 3$, varianza $\sigma^2 = 9$ e deviazione standard $\sigma = 3$. Tuttavia, essendo stato generato in maniera pseudocasuale, i valori di media campionaria, varianza campionaria e deviazione standard campionaria del campione sono i seguenti:

Media campionaria	3.19262
Varianza campionaria	12.46269
Deviazione standard campionaria	3.530253

Calcolati utilizzando le seguenti linee di codice:

```
> mean(camp2)
[1] 3.19262
> var(camp2)
[1] 12.46269
> sd(camp2)
[1] 3.530253
```

Altre misure di centralità e dispersione sono le seguenti:

```
> median(camp2)
[1] 3.341747
> quantile(camp2)
        0%        25%        50%        75%       100%
-8.0095468  0.8492563  3.3417469  5.9063511 11.4303793
```

Il campione generato ha il seguente grafico in Figura 5.1 prodotto dalle seguenti linee di codice:

```
> hist(camp2, freq = F, xlim=c(-5,13), ylim = c(0,0.3),breaks = 100, xlab = "x", ylab = "Istogramma",
 main = "Densità_simulata,N=150")
```

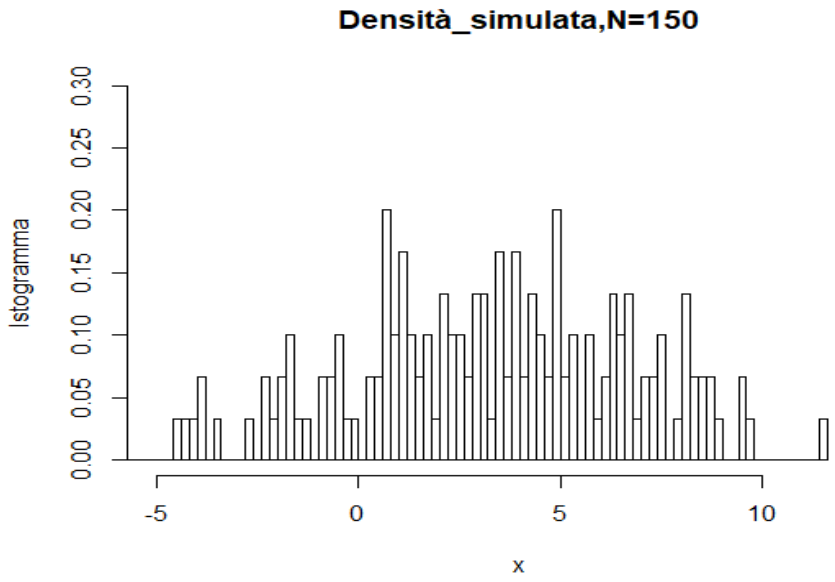


Figura 5.1: Grafico della densità simulata.

Siano quindi $X \sim \mathcal{N}(2,2)$ e $Y \sim \mathcal{N}(3,3)$ le due popolazioni normali dalle quali sono stati estratti due campioni di ampiezza rispettivamente 100 e 150.

- **(Intervallo di confidenza per $\mu_1 - \mu_2$ con σ_1^2 e σ_2^2 note)**

Proposizione *Siano $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$ le cui varianze sono note. Una stima dell'intervallo di confidenza di grado $1 - \alpha$ per la differenza tra le due medie $\mu_1 - \mu_2$ è*

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

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

Considerando i due campioni generati $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$ con rispettive deviazioni standard $\sigma_1 = 2$ e $\sigma_2 = 3$, si è determinata una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la differenza tra le due medie $\mu_1 - \mu_2$.

In questo caso $\overline{x}_{100} = 2.146789$, $\overline{y}_{150} = 3.19262$, $\sigma_1^2 = 4$, $\sigma_2^2 = 9$; inoltre essendo $\alpha = 0.05$ e $\alpha/2 = 0.025$, il valore $z_{\alpha/2} = z_{0.025}$ può essere determinato in R. Infatti, si ha:

```
> alpha<-1-0.95
```

```
> qnorm(1-alpha/2, mean = 0, sd=1)
[1] 1.959964
> n1<-length(camp)
> n2<-length(camp2)
> mean1<-mean(camp)
> mean2<-mean(camp2)
> sigma1<-2
> sigma2<-3
> mean1-mean2-qnorm(1-alpha/2, mean=0, sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] -1.665626
> mean1-mean2+qnorm(1-alpha/2, mean=0, sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] -0.4260357
```

Si nota che $z_{\alpha/2} = z_{0.025} = 1.959964$. La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la differenza tra le due medie $\mu_1 - \mu_2$ è (-1.665626, -0.4260357). Si nota che la differenza dei valori medi è uguale a -1.045831 e quindi è compresa nell'intervallo.

- **(Intervallo di confidenza per $\mu_1 - \mu_2$ con varianze non note)**

Proposizione *Siano $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ due campioni osservati indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$ le cui varianze sono non note. Una stima dell'intervallo di confidenza di grado $1 - \alpha$ per la differenza tra le due medie $\mu_1 - \mu_2$ è*

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{\tilde{s}_{n_2}^2}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{s_{n_1}^2}{n_1} + \frac{\tilde{s}_{n_2}^2}{n_2}},$$

dove $\overline{x}_n$ e $\overline{y}_n$ denotano rispettivamente le medie campionarie delle due osservazioni e dove $s_{n_1}^2$ e $\tilde{s}_{n_2}^2$ denotano rispettivamente le varianze campionarie delle due osservazioni.

Considerando i due campioni generati $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$ con varianze non note, si è determinata una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la differenza tra le due medie $\mu_1 - \mu_2$.

In questo caso $\overline{x}_{100} = 2.146789$, $\overline{y}_{150} = 3.19262$, $s_{n_1}^2 = 4.881814$, $s_{n_2}^2 = 12.46269$; inoltre $\alpha = 0.05$ e quindi $\alpha/2 = 0.025$. Utilizzando R si ha:

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
> n1<-length(camp)
> n2<-length(camp2)
> mean1<-mean(camp)
> mean2<-mean(camp2)
> sigma1<-sd(camp)
> sigma2<-sd(camp2)
> mean1-mean2-qnorm(1-alpha/2, mean=0, sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] -1.757659
> mean1-mean2+qnorm(1-alpha/2, mean=0, sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] -0.3340029
```

Si nota che $z_{\alpha/2} = z_{0.025} = 1.959964$. Una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la differenza tra le due medie $\mu_1 - \mu_2$ è (-1.757659, - 0.3340029). Si nota che la differenza dei valori medi è uguale a -1.045831 e quindi è compresa nell'intervallo.

6 VERIFICA DELLE IPOTESI

Le aree più importanti dell'inferenza statistica sono la stima dei parametri e la verifica delle ipotesi. In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità f(x; ϑ), un'ipotesi su di un parametro non noto della popolazione ed un campione casuale X1,X2, . . . ,Xn estratto dalla popolazione. Occorre in primo luogo precisare il significato di ipotesi statistica.

Definizione ipotesi statistica: *Un'ipotesi statistica è un'affermazione o una congettura sul parametro non noto ϑ. Se l'ipotesi statistica specifica completamente f(x; ϑ) è detta ipotesi semplice, altrimenti è chiamata ipotesi composta. Per denotare un'ipotesi statistica useremo il carattere H seguito dai due punti e successivamente dall'affermazione che specifica l'ipotesi.*

L'ipotesi soggetta a verifica viene in genere denotata con **H0** 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 H0. 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 **H1**.
Il problema della verifica delle ipotesi consiste nel determinare un test che permetta di suddividere, mediante opportuni criteri, l'insieme dei possibili campioni, ossia l'insieme delle n–ple (x1, x2, . . . , xn) assumibili dal vettore aleatorio X1,X2, . . . ,Xn, in due sottoinsiemi: una regione di accettazione A dell'ipotesi nulla ed una regione di rifiuto R dell'ipotesi nulla. Il test ϑ può allora essere così formulato: accettare come valida l'ipotesi nulla se il campione osservato (x1, x2, . . . , xn) $\in$ A e rifiutare l'ipotesi nulla se (x1, x2, . . . , xn) $\in$ R.
Nel caso si verifichi che l'ipotesi nulla sia falsa, l'ipotesi alternativa sarà vera e viceversa. Spesso si usa dire che l'ipotesi **H0** va verificata in alternativa all'ipotesi **H1**. Nel seguire questo tipo di ragionamento si può incorrere in due tipi di errori:

	Rifiutare H_0	Accettare H_0
H_0 vera	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
H_0 falsa	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

-rifiutare l'ipotesi nulla H0 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
-accettare l'ipotesi nulla H0 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

I test statistici sono di due tipi: test unilaterali (detti anche unidirezionali) e test bilaterali (detti anche bidirezionali).

POPOLAZIONE NORMALE

Utilizzando test bilaterali e unilaterali, desideriamo affrontare i seguenti problemi:
(i) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota;

(ii) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
(iii) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;
(iv) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

➢ Test su μ con varianza σ^2 nota

Sia X1,X2, . . . ,Xn un campione casuale estratto da una popolazione normale con varianza nota σ^2.

Test bilaterale: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2. nota. Si considerino le ipotesi:

$$\mathbf{H0} : \mu = \mu 0, \mathbf{H1} : \mu \neq \mu 0$$

Essendo la varianza nota, l'ipotesi **H0** è semplice, mentre l'ipotesi **H1** è composta.
Quando **H0** è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria

$$Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}},$$

che è distribuita normalmente con valore medio nullo e varianza unitaria. Il test bilaterale ϑ di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $-z_{\alpha/2} < Z_n < z_{\alpha/2}$
- si rifiuti **H0** se $Z_n < -z_{\alpha/2}$ oppure $Z_n > z_{\alpha/2}$

Tramite la densità normale standard , sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale.
Il valore $z_{\alpha/2}$ è calcolato tramite qnorm(1 − α/2,mean = 0, sd = 1).
Occorre applicare un test di verifica di ipotesi bilaterale.
Utilizzando R, risulta:

```
> alpha<-0.05
> mu0<-2
> sigma<-2
> qnorm(1-alpha/2,mean = 0, sd=1)
[1] 1.959964
> n<-100
> meancamp<-2.146789
> (meancamp-mu0)/(sigma/sqrt(n))
[1] 0.733945
```

Si nota che $z_{\alpha/2}$ = 1.959964 e z = 0.733945 cade all'interno della regione di accettazione; occorre quindi accettare l'ipotesi nulla.

Test unilaterale sinistro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2. nota. Si considerino le ipotesi:

$$\mathbf{H0} : \mu \leq \mu 0, \mathbf{H1} : \mu > \mu 0$$

Le ipotesi **H0** e **H1** sono entrambe composte. Il test unilaterale sinistro di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $Z_n < z_\alpha$
- si rifiuti **H0** se $Z_n > z_\alpha$

Tramite la densità normale, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale sinistro.
Il valore z_α è calcolato tramite qnorm(1 − α,mean = 0, sd = 1).

Occorre applicare un test di verifica di ipotesi unilaterale sinistro.
Utilizzando R, risulta:

```
> alpha<-0.05
> mu0<-2
> sigma<-2
> qnorm(1-alpha, mean=0, sd=1)
[1] 1.644854
> n<-100
> meancamp<-2.146789
> (meancamp-mu0)/(sigma/sqrt(n))
[1] 0.733945
```

Essendo l'estremo sinistro z_α = 1,64, per poter essere accettato il test laterale sinistro, il valore calcolato sarebbe dovuto essere minore di z_α, in questo caso la proprietà viene soddisfatta, quindi il testo unilaterale sinistro è accettato.

Test unilaterale destro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2 nota. Si
considerano le ipotesi:

$$\mathbf{H0} : \mu \geq \mu 0, \mathbf{H1} : \mu < \mu 0$$

Il test unilaterale destro di misura α per le ipotesi considerate è il seguente

- si accetti **H0** se $Z_n > -z_\alpha$
- si rifiuti **H0** se $Z_n < -z_\alpha$

Tramite la densità normale standard, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale destro.
Il valore $-z_\alpha$ è calcolato tramite qnorm(α,mean = 0, sd = 1).

Occorre applicare un test di verifica di ipotesi unilaterale destro.
Utilizzando R, risulta:

```
> alpha<-0.05
> mu0<-2
> sigma<-2
> qnorm(alpha, mean=0, sd=1)
[1] -1.644854
> (meancamp-mu0)/(sigma/sqrt(n))
[1] 0.733945
```

Ovviamente in questo caso il test Unilaterale destro viene soddisfatto.

- **Test su μ con varianza σ^2 non nota**

Sia X1,X2, . . . ,Xn un campione casuale estratto da una popolazione normale con varianza non nota σ^2.

Test bilaterale: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2non nota. Si considerino le ipotesi:

$$\mathbf{H0} : \mu = \mu 0, \mathbf{H1} : \mu \neq \mu 0$$

Essendo la varianza non nota, entrambe le ipotesi sono composte.
Quando **H0** è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria

$$T_n = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

che è distribuita con legge di Student con n-1 gradi di libertà. Il test bilaterale ϑ di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $-t_{\alpha/2,n-1} < T_n < t_{\alpha/2,n-1}$
- si rifiuti **H0** se $T_n < -t_{\alpha/2,n-1}$ oppure $T_n > t_{\alpha/2,n-1}$

Tramite la densità di Student con n-1 gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale.
Il valore $t_{\alpha/2,n-1}$ è calcolato tramite qt(1 − α/2,df=n-1).
Occorre applicare un test di verifica di ipotesi bilaterale.
Utilizzando R, risulta:

```
> alpha<-0.01
> mu0<-2
> n<-100
> qt(1-alpha/2,df=n-1)
[1] 2.626405
> meancamp<-2.146789
> devcamp<-2.209483
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] 0.664359
```

Si nota che $t_{\alpha/2,n-1}$ = 2.626405 e t = 0.664359 cade all'interno della regione di accettazione; occorre quindi accettare l'ipotesi nulla.

Test unilaterale sinistro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2 non nota. Si considerino le ipotesi:

$$\mathbf{H0} : \mu \leq \mu 0, \mathbf{H1} : \mu > \mu 0$$

Le ipotesi **H0** e **H1** sono entrambe composte. Il test unilaterale sinistro di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $T_n < t_{\alpha,n-1}$
- si rifiuti **H0** se $T_n > t_{\alpha,n-1}$

Tramite la densità di Student con n-1 gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale sinistro.
Il valore $t_{\alpha,n-1}$ è calcolato tramite qt(1 − α,df=n-1).
Occorre applicare un test di verifica di ipotesi unilaterale sinistro.
Utilizzando R, risulta:

```
> alpha<-0.01
> mu0<-2
> meancamp<-2.146789
> devcamp<-2.209483
> qt(1-alpha,df=n-1)
[1] 2.364606
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] 0.664359
```

Essendo l'estremo sinistro $t_{\alpha,n-1}$ = 2.364606, per poter essere accettato il test laterale sinistro, il valore calcolato sarebbe dovuto essere minore di $t_{\alpha,n-1}$, in questo caso la proprietà viene soddisfatta, quindi il testo unilaterale sinistro è accettato.

Test unilaterale destro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con varianza σ^2 non nota. Si
considerano le ipotesi:

$$\mathbf{H0} : \mu \geq \mu 0, \mathbf{H1} : \mu < \mu 0$$

Il test unilateraledestro ϑ di misura α per le ipotesi considerate è il seguente

- si accetti **H0** se $T_n > -t_{\alpha,n-1}$
- si rifiuti **H0** se $T_n < -t_{\alpha,n-1}$

Tramite la densità di Student con n-1 gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale destro.
Il valore $-t_{\alpha,n-1}$ è calcolato tramite qt(α,df=n-1).

Occorre applicare un test di verifica di ipotesi unilaterale destro.
Utilizzando R, risulta:

```
> alpha<-0.01
> mu0<-2
> n<-100
> qt(alpha,df=n-1)
[1] -2.364606
> (meancamp-mu0)/(devcamp/sqrt(n))
[1] 0.664359
```

Ovviamente in questo caso il test Unilaterale destro viene soddisfatto.

- ## Test su σ^2 con valore medio noto

Sia X1,X2, . . . ,Xn un campione casuale estratto da una popolazione normale con valore medio noto.

Test bilaterale: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio noto. Si considerino le ipotesi:

H0 : $\sigma^2 = \sigma^2_0$, **H1** : $\sigma^2 \neq \sigma^2_0$

Essendo il valore medio noto, l'ipotesi **H0** è semplice, mentre l'ipotesi **H1** è composta.
Quando **H0** è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria

$$V_n = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma_0}\right)^2 = \frac{(n-1)\,S_n^2}{\sigma_0^2} + \left(\frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}}\right)^2$$

che è distribuita con legge chi-quadrato con n gradi di libertà. Il test bilaterale ϑ di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $X^2_{1-\alpha/2,n} < V_n < X^2_{\alpha/2,n}$
- si rifiuti **H0** se $V_n < X^2_{1-\alpha/2,n}$ oppure $V_n > X^2_{\alpha/2,n}$

Tramite la densità chi-quadrato con n gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale.
Il valore $X^2_{1-\alpha/2,n}$ è calcolato tramite qchisq(α/2,df=n) mentre il valore $X^2_{\alpha/2,n}$ si calcola con qchisq(1-α/2,df=n).
Occorre applicare un test di verifica di ipotesi bilaterale.
Utilizzando R, risulta:

```
> alpha<-0.05
> mu<-2
> sigma02<-4
> n<-100
> medcamp<-2.146789
> varcamp<-4.881814
> qchisq(alpha/2,df=n)
[1] 74.22193
> qchisq(1-alpha/2,df=n)
[1] 129.5612
> (n-1)*varcamp/sigma02+n*(medcamp-mu)**2/sigma02
[1] 121.3636
```

Si nota che $X^2_{1-\alpha/2,n}$ = 74.22193 e $X^2_{\alpha/2,n}$= 129.5612 e v = 121.3636 cade all'interno della regione di accettazione; occorre quindi accettare l'ipotesi nulla.

Test unilaterale sinistro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio noto. Si considerino le ipotesi:

H0 : $\sigma^2 \leq \sigma^2_0$ **H1** : $\sigma^2 > \sigma^2_0$

Le ipotesi **H0** e **H1** sono entrambe composte. Il test unilaterale sinistro di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $V_n < X^2_{\alpha,n}$
- si rifiuti **H0** se $V_n > X^2_{\alpha,n}$

Tramite la densità chi-quadrato con n gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale sinistro.
Il valore $X^2_{\alpha,n}$ è calcolato tramite qchisq(1 − α,df=n).
Occorre applicare un test di verifica di ipotesi unilaterale sinistro.
Utilizzando R, risulta:

```
> qchisq(1-alpha,df=n)
[1] 124.3421
> (n-1)*varcamp/sigma02+n*(medcamp-mu)**2/sigma02
[1] 121.3636
```

Essendo l'estremo sinistro $X^2_{\alpha,n}$ = 124.3421, per poter essere accettato il test laterale sinistro, il valore calcolato sarebbe dovuto essere minore di $X^2_{\alpha,n}$, in questo caso la proprietà viene soddisfatta, quindi il testo unilaterale sinistro è accettato.

Test unilaterale destro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con valore medio noto. Si
considerano le ipotesi:

H0 : $\sigma^2 \geq \sigma^2$ $_0$

H1 : $\sigma^2 < \sigma^2_0$

Entrambe le ipotesi sono composte. Il test unilateraledestro ϑ di misura α per le ipotesi considerate è il seguente

- si accetti **H0** se $V_n > X^2_{1-\alpha,n}$
- si rifiuti **H0** se $V_n < X^2_{1-\alpha,n}$

Tramite la densità chi-quadrato con n gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale destro.
Il valore $X^2_{1-\alpha,n}$ è calcolato tramite qchisq(α,df=n).

Occorre applicare un test di verifica di ipotesi unilaterale destro.
Utilizzando R, risulta:

```
> qchisq(alpha,df=n)
[1] 77.92947
> (n-1)*varcamp/sigma02+n*(medcamp-mu)**2/sigma02
[1] 121.3636
```

Essendo l'estremo destro $X^2_{1-\alpha,n}$ = 77.92947, in questo caso il test Unilaterale destro viene soddisfatto.

- **Test su σ^2 con valore medio non noto**

Sia X1,X2, . . . ,Xn un campione casuale estratto da una popolazione normale con valore medio non noto.

Test bilaterale: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con entrambi i parametri non noti. Si considerino le ipotesi:

H0 : $\sigma^2 = \sigma^2$ $_0$ **H1** : $\sigma^2 \neq \sigma^2{}_0$

Entrambe le ipotesi sono composte. Quando **H0** è vera, in analogia a quanto visto per gli intervalli di confidenza, gioca un ruolo fondamentale la variabile aleatoria

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

che è distribuita con legge chi-quadrato con n-1 gradi di libertà. Il test bilaterale ϑ di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $X^2_{1-\alpha/2,n-1} < Q_n < X^2_{\alpha/2,n-1}$
- si rifiuti **H0** se $Q_n < X^2_{1-\alpha/2,n-1}$ oppure $Q_n > X^2_{\alpha/2,n-1}$

Tramite la densità chi-quadrato con n-1 gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale.
Il valore $X^2_{1-\alpha/2,n-1}$ è calcolato tramite qchisq(α/2,df=n-1) mentre il valore $X^2_{\alpha/2,n-1}$ si calcola con qchisq(1-α/2,df=n-1).
Occorre applicare un test di verifica di ipotesi bilaterale.
Utilizzando R, risulta:

```
> alpha<-0.05
> sigma02<-4
> n<-100
> qchisq(alpha/2,df=n-1)
[1] 73.36108
> qchisq(1-alpha/2,df=n-1)
[1] 128.422
> varcamp<-4.881814
> (n-1)*varcamp/sigma02
[1] 120.8249
```

Si nota che $X^2_{1-\alpha/2,n-1}$ = 73.36108 e $X^2_{\alpha/2,n-1}$= 128.422 e q = 120.8249 cade all'interno della regione di accettazione; occorre quindi accettare l'ipotesi nulla.

Test unilaterale sinistro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con entrambi i parametri non noti. Si considerino le ipotesi:

H0 : $\sigma^2 \leq \sigma^2{}_0$ **H1** : $\sigma^2 > \sigma^2{}_0$

Il test unilaterale sinistro di misura α per le ipotesi considerate è il seguente:

- si accetti **H0** se $Q_n < X^2_{\alpha,n-1}$
- si rifiuti **H0** se $Q_n > X^2_{\alpha,n-1}$

Tramite la densità chi-quadrato con n-1 gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale sinistro.
Il valore $X^2_{\alpha,n-1}$ è calcolato tramite qchisq(1 − α,df=n-1).
Occorre applicare un test di verifica di ipotesi unilaterale sinistro.
Utilizzando R, risulta:

```
> qchisq(1-alpha,df=n-1)
[1] 123.2252
> (n-1)*varcamp/sigma02
[1] 120.8249
```

Essendo l'estremo sinistro $X^2_{\alpha,n-1}$ = 123.2252, per poter essere accettato il test laterale sinistro, il valore calcolato sarebbe dovuto essere minore di $X^2_{\alpha,n-1}$, in questo caso la proprietà viene soddisfatta, quindi il testo unilaterale sinistro è accettato.

Test unilaterale destro: Sia (x1, x2, . . . , xn) un campione osservato di ampiezza n estratto da una popolazione normale con entrambi i parametri non noti. Si
considerano le ipotesi:

H0 : $\sigma^2 \geq \sigma^2$ $_0$, **H1** : $\sigma^2 < \sigma^2$ $_0$

Il test unilateraledestro ϑ di misura α per le ipotesi considerate è il seguente

- si accetti **H0** se $Q_n > X^2_{1-\alpha,n-1}$
- si rifiuti **H0** se $Q_n < X^2_{1-\alpha,n-1}$

Tramite la densità chi-quadrato con n-1 gradi di libertà, sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla del test unilaterale destro.
Il valore $X^2_{1-\alpha,n}$ è calcolato tramite qchisq(α,df=n-1).

Occorre applicare un test di verifica di ipotesi unilaterale destro.
Utilizzando R, risulta:

```
> qchisq(alpha,df=n-1)
[1] 77.04633
> (n-1)*varcamp/sigma02
[1] 120.8249
```

Essendo l'estremo destro $X^2_{1-\alpha,n-1}$ = 77.04633, in questo caso il test Unilaterale destro viene soddisfatto.

7 CRITERIO DEL CHI-QUADRATO

In questo capitolo si è dedicata l'attenzione al criterio di verifica delle ipotesi del chi-quadrato.

Spesso si desidera verificare se il campione osservato può essere stato estratto da una popolazione caratterizzata da una funzione di distribuzione $F_X(x)$. A questo scopo utilizzeremo il *criterio di verifica delle ipotesi del chi-quadrato*, detto anche *test del chi-quadrato* o *test del buon adattamento*.

7.1 Criterio del chi-quadrato

Con il criterio del chi-quadrato si desidera verificare l'ipotesi che una certa popolazione, descritta da una variabile aleatoria *X*, sia caratterizzata da una funzione di distribuzione $F_X(x)$ con k parametri non noti da stimare.

Denotando con $\mathbf{H}_0$ l'ipotesi soggetta a verifica, detta *ipotesi nulla*, e con $\mathbf{H}_1$ l'*ipotesi alternativa*, il test chi-quadrato di misura α mira a verificare l'ipotesi nulla

$\mathbf{H}_0$ *: X ha una funzione di distribuzione* $F_X(x)$ *(avendo stimato i k parametri non noti in base al campione)*

In alternativa all'ipotesi

$\mathbf{H}_1$ *: X non ha una funzione di distribuzione* $F_X(x)$,

dove α è la *probabilità massima di rifiutare l'ipotesi nulla quando essa è vera*.

Di solito il decisore sceglie α uguale a 0.05, 0.01, 0.001 e il test viene rispettivamente detto *significativo*, *abbastanza significativo* e *molto significativo*.

Infatti, quando più piccolo è il valore di α tanto maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla.

Occorre determinare un test di misura α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. A tal fine, suddividiamo l'insieme dei valori che la variabile aleatoria *X* può assumere in *r* sottoinsiemi $I_1, I_2, \dots, I_r$ (classi o categorie) in modo che risulti essere uguale a p_i la probabilità che la variabile aleatoria assuma un valore appartenente a I_i, ossia

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

Si estrae poi un campione $x_1, x_2, \dots, x_n$ di ampiezza *n* e si osservano le frequenze assolute $n_1, n_2, \dots, n_r$ con cui gli *n* elementi si distribuiscono nei rispettivi insiemi $I_1, I_2, \dots, I_r$. Quindi n_i rappresenta il numero degli elementi del campione che cadono nell'intervallo I_i $(i = 1,2, \dots, r)$. È chiaro che

$$p_i \geq 0 \qquad \sum_{i=1}^{r} p_i = 1\,;$$

$$n_i \geq 0 \qquad \sum_{i=1}^{r} n_i = n.$$

Il numero medio di elementi che cadono nell'intervallo I_i è np_i.

Si calcola poi la quantità

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

Il criterio chi-quadrato si basa sulla statistica

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

dove N_i è la variabile aleatoria che descrive il numero degli elementi del campione casuale X_1, X_2, …, X_n che cadono nell'intervallo I_i $(i = 1,2, \dots, r)$.

Si può dimostrare che per n sufficientemente grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con $r - k - 1$ gradi di libertà.

Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta

$$\min\{np_1, np_2, \dots, np_r\} \geq 5.$$

Il *test chi-quadrato* può essere formulato nel seguente modo:

- *si rifiuta l'ipotesi H_0 se* $\chi^2 < \chi^2_{1-\alpha/2,r-k-1}$ *oppure* $\chi^2 > \chi^2_{\alpha/2,r-k-1}$
- *si accetti l'ipotesi H_0 se* $\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$

Considerato il campione precedentemente generato formato da n = 100

```
> n<-length(camp)
> n
[1] 100
> m<-mean(camp)
> m
[1] 2.146789
> d<-sd(camp)
> d
[1] 2.209483
```

Si nota che la media campionaria $\overline{x} = 2.146789$ e la deviazione standard campionaria è $s = 2.209483$.

Applicando il test chi-quadrato di misura $\alpha = 0.05$, si è verificato se la popolazione da cui proviene il campione può essere descritta da una variabile aleatoria X normale di densità

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad x \in \mathbb{R} \quad (\mu \in \mathbb{R}, \sigma > 0),$$

Si è supposto di suddividere l'insieme dei valori che tale variabile aleatoria normale X può assumere in $r = 5$ sottoinsiemi $I_1, I_2, \dots, I_5$ in modo che risulti essere uguale a $p_i = 0.2$ la probabilità che X assuma valore appartenente a I_i $(i = 1,2, \dots, r)$. La condizione che la classe contenga almeno 5

elementi è verificata essendo $np_i = 100 \cdot 0.2 = 20 \geq 5$. Ricordando che uno stimatore di μ è la media campionaria e uno stimatore di σ^2 è la varianza campionaria, utilizzando i quantili della distribuzione normale si sono potuti determinare i sottoinsiemi $I_1, I_2, \dots, I_5$

```
> a<-numeric(4)
> for(i in 1:4)
+ a[i]<-qnorm(0.2*i, mean=m, sd=d)
> a
[1] 0.2872414 1.5870229 2.7065549 4.0063364
```

Gli intervalli $I_1, I_2, \dots, I_5$ sono:

$$I_1 = (-\infty, 0.2872414), \quad I_2 = [0.2872414, 1.5870229), \quad I_3 = [1.5870229, 2.7065549),$$
$$I_4 = [2.7065549, 4.0063364), \quad I_5 = [4.0063364, +\infty).$$

Si è determinato ora il numero di elementi del campione che cadono negli intervalli $I_1, I_2, \dots, I_5$:

```
> r<-5
> nint<-numeric(r)
> nint[1]<-length(which(camp<a[1]))
> nint[2]<-length(which((camp>=a[1]) & (camp<a[2])))
> nint[3]<-length(which((camp>=a[2]) & (camp<a[3])))
> nint[4]<-length(which((camp>=a[3]) & (camp<a[4])))
> nint[5]<-length(which(camp>=a[4]))
> nint
[1] 15 28 20 20 17
> sum(nint)
[1] 100
```

Segue che $n_1 = 15$, $n_2 = 28$, $n_3 = 20$, $n_4 = 20$ e $n_5 = 17$. Si è calcolato ora χ^2 definita precedentemente

```
> chi2<-sum(((nint-n*0.2)/sqrt(n*0.2))^2)
> chi2
[1] 4.9
```

ossia $\chi^2 = 4.9$. La distribuzione normale ha due parametri non noti (μ e σ^2) e quindi $k = 2$ (numero paramentri non noti) e quindi la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con $r - k - 1 = 2$ gradi di libertà. Quindi si sono calcolati ora $\chi^2_{\alpha/2,2}$ e $\chi^2_{1-\alpha/2,2}$ con $\alpha = 0.05$.

```
> k<-2
> alpha<-0.05
> qchisq(alpha/2, df=r-k-1)
[1] 0.05063562
> qchisq(1-alpha/2, df=r-k-1)
[1] 7.377759
```

da cui segue che $\chi^2_{1-\alpha/2,r-k-1} = 7.377759$ e $\chi^2_{\alpha/2,2} = 0.05063562$. Essendo $0.05063562 < \chi^2 < 7.377759$, l'ipotesi H_0 di popolazione normale può essere accettata.