Tesina Statistica

Nicola Lancellotti

30/01/2019

Indice

Capitolo 1

Introduzione

Lo scopo della presente tesina è quello di mettere in pratica le nozioni teoriche apprese nel corso di "Statistica e Analisi dei Dati" (Nobile, 2018) tenuto dalla professoressa Nobile all'Università degli Studi di Salerno.

La prima parte, composta dai capitoli da 2 a 7, si occuperà della statistica descrittiva, cioè la parte della statistica che si occupa della rilevazione, analisi, sintesi, interpretazione e rappresentazione dei dati di una popolazione o di una sua parte, detta campione. L'oggetto di interesse su cui sarà applicata la statistica descrittiva è la soddisfazione degli italiani per quanto riguarda la situazione economica, di salute, delle relazioni con i membri della famiglia, delle relazioni con gli amici e infine del proprio tempo libero. I dati sono stati raccolti dall'ISTAT nel 2017 attraverso sondaggi su un campione casuale. Il sondaggio consisteva nell'esprimere una valutazione qualitativa per ognuna delle caratteristiche di interesse. I valori qualitativi possibili erano: "molto", "abbastanza", "poco" e "per niente". Infine è stata calcolata le percentuale di persone che hanno espresso un voto pari a "molto" per ognuna delle caratteristiche e per ognuna delle regioni italiane. Dopo una prima analisi visiva mediante i boxplot verranno effettuate analisi numeriche volte a spiegare la distribuzione del campione. Verranno quindi calcolate le frequenze e verra analizzata la simmetria e la curtosi. In seguito verrano analizzati gli indici di posizione e di dispersione. A seguire verranno analizzate le correlazioni tra le caratteristiche del data set e infine l'analisi dei cluster consentirà di scoprire le regioni simili tra loro per quanto riguarda le caratteristiche analizzate.

La seconda parte, composta dai capitoli da 8 a 12, invece si occuperà dell'inferenza statistica che ha lo scopo di estendere le misura ricavate dal campione alla popolazione da cui il campione è stato estratto. Verrà utilizzato un campione estratto da una popolazione descritta da una distribuzione geometrica e verranno mostrate le tecniche che consentono di stimare il parametro non noto p della distribuzione scelta. Oltre a una stima puntuale di tale valore sarà mostrato anche come ottenere un intervallo di confidenza approssimato per il parametro da stimare. Infine gli ultimi capitoli si concluderanno con i test di verifica delle ipotesi del valore medio e su un particolare test, detto test del chi-quadrato, che consente di verificare se il campione proviene realmente da una distribuzione geometrica.

In entrambe le parti oltre a spiegare e mostrare i risultati ottenuti tramite tabelle e grafici verrà anche mostrato il codice in linguaggio R
(R Development Core Team, 2008) che ha permesso di computare tali risultati.

Capitolo 2

Data Set

Di seguito il codice che consente di creare un data frame in R con i valori in percentuale ottenuti dal sondaggio. Tali percentuali rappresentano il grado di soddisfazione per le caratteristiche di interesse. Inoltre sono state definiti alcuni oggetti che saranno utilizzati successivamente.

Per una migliore visualizzazione i valori sono mostrati anche nella tabella 2.1.

```
df <-data.frame(
  economica = c(4.8, 5.1, 3.9, 4.3, 10.8, 4.7, 6.4, 4.1, 3, 4.2,
                3.4, 2.9, 2.8, 2.2, 1.9, 2.6, 2.2, 1.7, 1.9, 1.9),
  salute = c(17.6, 19.5, 18.2, 18.4, 28.7, 18.7, 17.3, 19.2, 17.1, 16,
             15.3, 14.6, 18.3, 15.3, 13, 11, 12.1, 10.2, 16.2, 12),
  famiglia = c(36.2, 35.3, 39.5, 35.3, 46.4, 38.5, 37.3, 38.7, 35.6, 35.7,
               33.7, 31.1, 33.9, 30.6, 24.3, 22.2, 30.5, 28, 30.5, 30.4),
  amici = c(24.5, 25.5, 26.5, 25.5, 35.2, 25.1, 26.8, 28.4, 25.4, 26,
            23.2, 22.8, 23.8, 19.7, 16.5, 16.2, 21.6, 17.8, 20.1, 21.6),
  tempoLibero = c(15.4, 15, 16.4, 16, 23.8, 15.4, 16.6, 15.5, 15.5, 17.3,
                  13.9, 13, 11.9, 12.1, 9.3, 9.3, 10.6, 9.2, 11.6, 10.9)
)

rownames(df) <- c("Piemonte", "Valle d'Aosta", "Liguria", "Lombardia",
                  "Trentino Alto Adige", "Veneto", "Friuli-Venezia Giulia",
                  "Emilia-Romagna", "Toscana", "Umbria", "Marche", "Lazio",
                  "Abruzzo", "Molise", "Campania", "Puglia", "Basilicata",
                  "Calabria", "Sicilia", "Sardegna")
```

```
column.names <- c("Economica", "Salute", "Famiglia", "Amici", "Tempo libero")

rowCount = dim(df)[1]
colCount = dim(df)[2]
```

2.1 Boxplot

Un boxplot, detto anche diagramma a scatola con baffi, è una rappresentazione grafica che permette di descrivere la distribuzione di un campione.

É costituito da una scatola i cui estremi sono il primo e il terzo quartile, divisa al suo interno dal secondo quartile detto anche mediana. In basso e in alto, o a sinistra e destra a seconda della rappresentazione verticale o orizzontale, sono presenti altri due segmenti detti baffi calcolati nel seguente modo:

Tabella 2.1: Data Set

	Economica	Salute	Famiglia	Amici	Tempo libero
Piemonte	4.8	17.6	36.2	24.5	15.4
Valle d'Aosta	5.1	19.5	35.3	25.5	15.0
Liguria	3.9	18.2	39.5	26.5	16.4
Lombardia	4.3	18.4	35.3	25.5	16.0
Trentino Alto Adige	10.8	28.7	46.4	35.2	23.8
Veneto	4.7	18.7	38.5	25.1	15.4
Friuli-Venezia Giulia	6.4	17.3	37.3	26.8	16.6
Emilia-Romagna	4.1	19.2	38.7	28.4	15.5
Toscana	3.0	17.1	35.6	25.4	15.5
Umbria	4.2	16.0	35.7	26.0	17.3
Marche	3.4	15.3	33.7	23.2	13.9
Lazio	2.9	14.6	31.1	22.8	13.0
Abruzzo	2.8	18.3	33.9	23.8	11.9
Molise	2.2	15.3	30.6	19.7	12.1
Campania	1.9	13.0	24.3	16.5	9.3
Puglia	2.6	11.0	22.2	16.2	9.3
Basilicata	2.2	12.1	30.5	21.6	10.6
Calabria	1.7	10.2	28.0	17.8	9.2
Sicilia	1.9	16.2	30.5	20.1	11.6
Sardegna	1.9	12.0	30.4	21.6	10.9

- il baffo inferiore è il valore più piccolo tra le osservazioni maggiore o uguale a $Q_1 1.5(Q_3 Q_1)$
- il baffo superiore è il valore più grande tra le osservazione minore o uguale a $Q_3 + 1.5(Q_3 Q_1)$

Eventuali valori non appartenenti all'intervallo definito tra il baffo inferiore e superiore sono detti valori anomali.

Il seguente codice permette di mostrare i boxplot per le caratteristiche del dataset. I grafici sono mostrati nella figura 2.1.

```
par(mfrow=c(3, 2))
for (i in 1:colCount) {
  boxplot(df[i], horizontal = TRUE, col = i + 1, main = column.names[i])
}
```

Invece il seguente codice permette di mostrare tutti i boxplot nello stesso grafico. Il grafico è mostrato nella figura 2.2.

```
  boxplot(df, col = 2:6, names = column.names)
```

Infine il seguente codice permette di calcolare i baffi i cui valori sono mostrati nella tabella 2.2.

```
calcolaBaffi <- function(values) {
  q <- quantile(values)
  inferiore <- q[[2]] - 1.5 * (q[[4]] - q[[2]])
  superiore <- q[[4]] + 1.5 * (q[[4]] - q[[2]])
  baffoInf <- min(values[values >= inferiore])
  baffoSup <- max(values[values <= superiore])
  c(baffoInf, baffoSup)
}
baffi <- sapply(df, calcolaBaffi)
row.names(baffi) <- c("Inferiore", "Superiore")
```

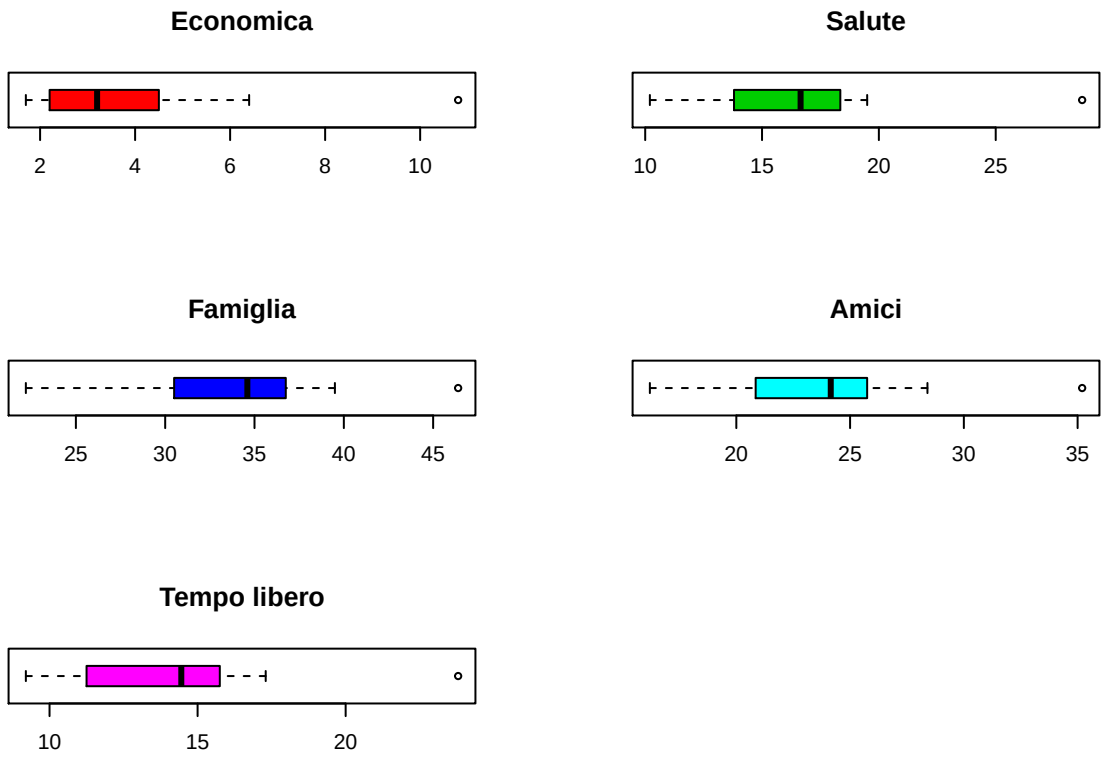

Figura 2.1: Boxplot

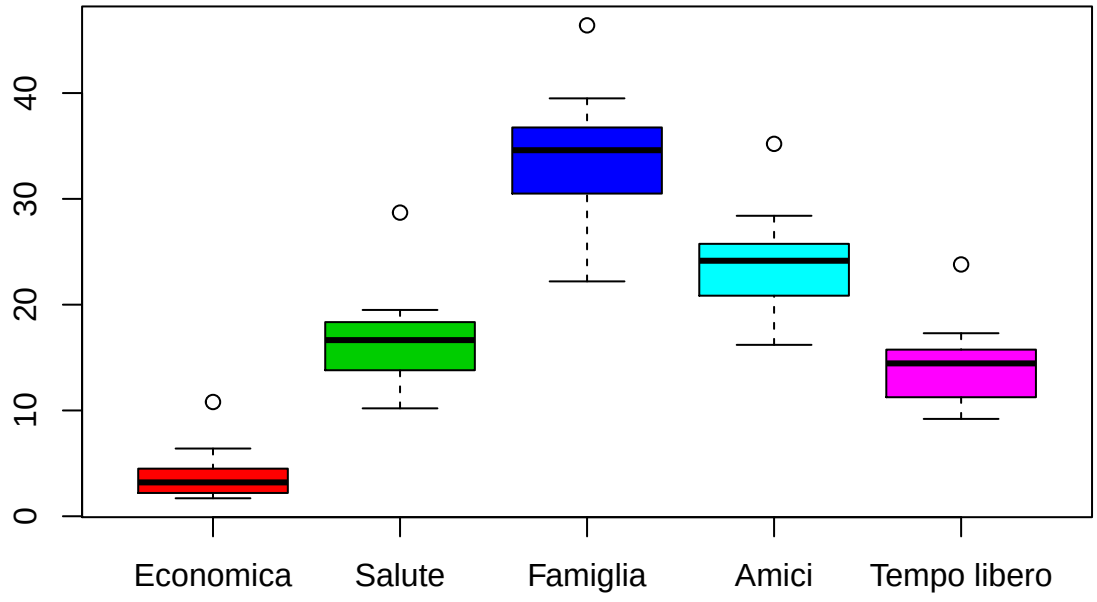

Figura 2.2: Boxplot

Tabella 2.2: Baffi

	Economica	Salute	Famiglia	Amici	Tempo libero
Inferiore	1.7	10.2	22.2	16.2	9.2
Superiore	6.4	19.5	39.5	28.4	17.3

Dall'analisi visiva dei baffi si può trarre la conclusione che in ogni caratteristica è presente un valore anomalo e per tutte le caratteristiche l'individuo nel campione che presenta il valore anomale è il "Trentino Alto Adige". Inoltre si può osservare che gli italiani traggono in media il massimo della soddisfazione nei rapporti familiari seguiti dai rapporti con gli amici, la salute, il tempo libero e infine dalla situazione economica.

Si nota inoltre che la mediana è più vicina al terzo quartile e che il baffo sinistro è più grande del destro per tutte le caratteristiche eccetto per la soddisfazione economica.

Capitolo 3

Distribuzione

In questo capitolo verranno prima calcolate e analizzate le frequenze delle caratteristiche del data set, in seguito le frequenze saranno utilizzate per il calcolo della distribuzione empirica continua di frequenza per ogni caratteristica. Il capitolo si concluderà con un analisi della simmetria e della curtosi delle distribuzioni considerate.

3.1 Frequenze

Se i valori osservati possono assumere k modalità, la frequenza assoluta della modalità i-esima è definita come il numero di occorrenze della modalità i-esima nel campione. Mentre la frequenza relativa della modalità i-esima è definita come il rapporto tra la frequenza assoluta i-esima ed il numero delle osservazioni. Infine la frequenza cumulata assoluta (o relativa) i-esima è definita come la somma di tutte le frequenze assolute (o relative) j-esime con j minore o uguale a i.

Il calcolo delle frequenze diventa problematico nel caso in cui il numero delle modalità è grande (o al limite infinito), rispetto al numero delle osservazioni. Infatti le frequenze tenderanno ad assumere valori pari a zero o uno. In questi casi è utile dividere i valori in classi e considerare tali classi come le modalità osservabili.

Essendo il nostro data set composto da dati reali si è scelto, per il motivo sopra esposto, di dividere i dati in classi in modo tale che ogni numero reale appartenga alla classe del suo intero più vicino.

Il seguente codice permette di calcolare le frequenze relative del dataset. I valori calcolati sono mostrati nelle tabelle 3.1, 3.2, 3.3 3.4 e 3.5, per le classi non presenti nelle tabelle la frequenza relativa associata è nulla. Invece una rappresentazione delle frequenze con grafici a bastoncini è mostrata nelle figure 3.1, 3.2, 3.3, 3.4 e 3.5 rispettivamente per le stesse caratteristiche. Nei grafici la linea rossa indica la mediana mentre la linea verde la media.

```
classi <- c(0, seq(0.5, 100.5, by = 1))

calcolaFrequenzeRelative = function(x) {
  table(cut(x, breaks = classi, right = FALSE)) / rowCount
}

frequenzeRelative <- sapply(df, calcolaFrequenzeRelative)
```

Tabella 3.1: Frequenze relative - Soddisfazione Economica

Classe	Frequenza Relativa
[1.5,2.5)	0.3
[2.5,3.5)	0.25
[3.5,4.5)	0.2
[4.5,5.5)	0.15
[5.5,6.5)	0.05
[10.5,11.5)	0.05

Tabella 3.2: Frequenze relative - Soddisfazione Salute

Classe	Frequenza Relativa
[9.5,10.5)	0.05
[10.5,11.5)	0.05
[11.5,12.5)	0.1
[12.5,13.5)	0.05
[14.5,15.5)	0.15
[15.5,16.5)	0.1
[16.5,17.5)	0.1
[17.5,18.5)	0.2
[18.5,19.5)	0.1
[19.5,20.5)	0.05
[28.5,29.5)	0.05

Tabella 3.3: Frequenze relative - Soddisfazione Famiglia

Classe	Frequenza Relativa
[21.5,22.5)	0.05
[23.5,24.5)	0.05
[27.5,28.5)	0.05
[29.5,30.5)	0.05
[30.5,31.5)	0.2
[33.5,34.5)	0.1
[34.5,35.5)	0.1
[35.5,36.5)	0.15
[36.5,37.5)	0.05
[38.5,39.5)	0.1
[39.5,40.5)	0.05
[45.5,46.5)	0.05

Tabella 3.4: Frequenze relative - Soddisfazione Amici

Classe	Frequenza Relativa
[15.5,16.5)	0.05
[16.5,17.5)	0.05
[17.5,18.5)	0.05
[19.5,20.5)	0.1
[21.5,22.5)	0.1
[22.5,23.5)	0.1
[23.5,24.5)	0.05
[24.5,25.5)	0.15
[25.5,26.5)	0.15
[26.5,27.5)	0.1
[27.5,28.5)	0.05
[34.5,35.5)	0.05

Tabella 3.5: Frequenze relative - Soddisfazione Tempo libero

Classe	Frequenza Relativa
[8.5,9.5)	0.15
[10.5,11.5)	0.1
[11.5,12.5)	0.15
[12.5,13.5)	0.05
[13.5,14.5)	0.05
[14.5,15.5)	0.15
[15.5,16.5)	0.2
[16.5,17.5)	0.1
[23.5,24.5)	0.05

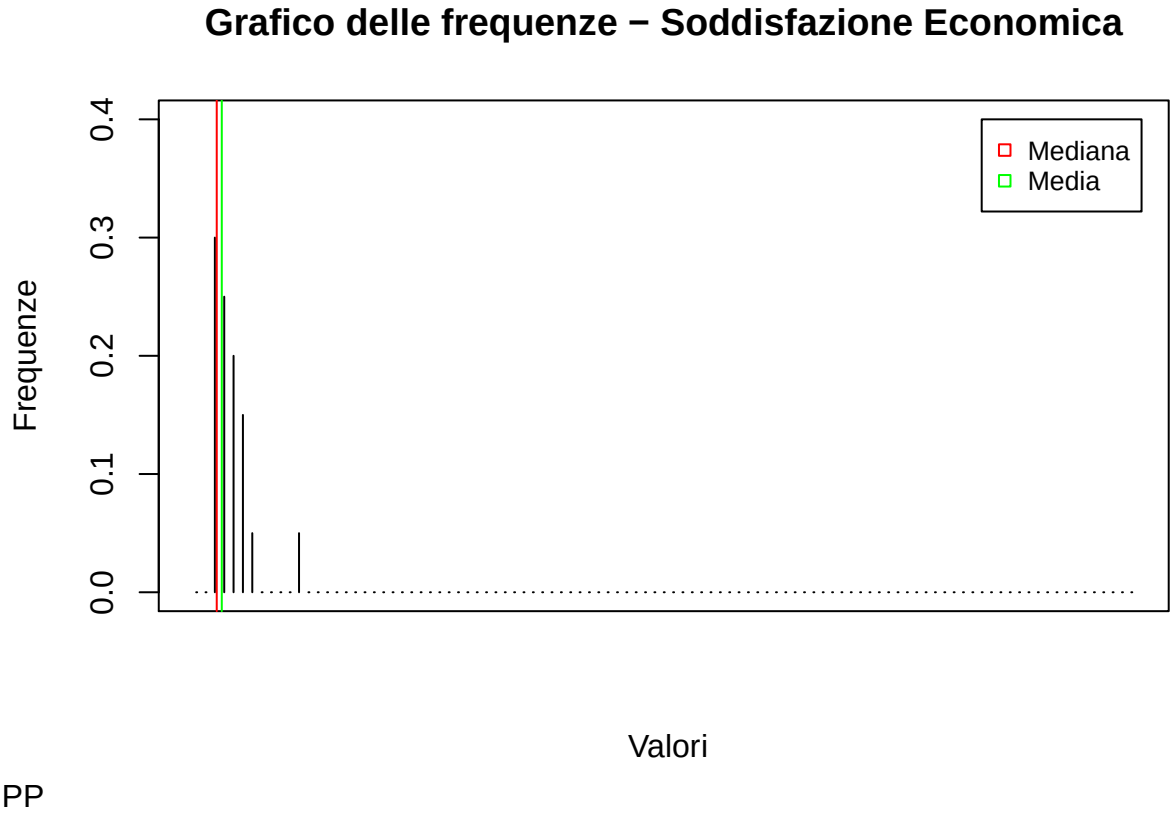

Figura 3.1: Grafico delle frequenze - Soddisfazione Economica

3.2 Funzione di distribuzione empirica continua

La funzione di distribuzione empirica continua è una funzione che associa a ogni valore reale x la frequenza che un valore osservato sia minore o uguale a x. Da notare che nel continuo si preferisce suddividere i valori in k classi $[z_i, z_{i+1})$ con $i = 1, .., k$. Ne consegue che la distribuzione empirica continua è una funzione reale non decrescente, che assume valore zero per ogni reale minore di z_1 e valore uno per ogni reale maggiore o uguale a z_{k+1}, quando $x = z_i$ per $i = 1, .., k$ la funzione assume lo stesso valore della frequenza relativa cumulata F_i mentre per ogni valore compreso tra (z_i, z_{i+1}) la funzione coincide con il segmento passante per i punti (z_i, F_i) e (z_{i+1}, F_{i+1})

La seguente funzione permette di creare i grafici delle distribuzioni empiriche continue delle caratteristiche del dataset, tali grafici sono mostrati nelle figure 3.6, 3.7, 3.8, 3.9, 3.10.

```
distribuzioneEmpiricaContinua = function(index, lim) {
  x <-frequenzeRelative[, index]
  Fi<-cumsum(x)
  Fi<-c(0,Fi)
  main = paste("Distribuzione empirica continua - Soddisfazione", column.names[index])
  plot(classi, Fi, type = "l",axes = FALSE,
       main = main,col="red",
       xlab = " ", ylab = " ",
       xlim = c(0, lim))
  axis(1, classi)
  axis(2, format(Fi, digits = 1), las = 2)
  box()
}
```

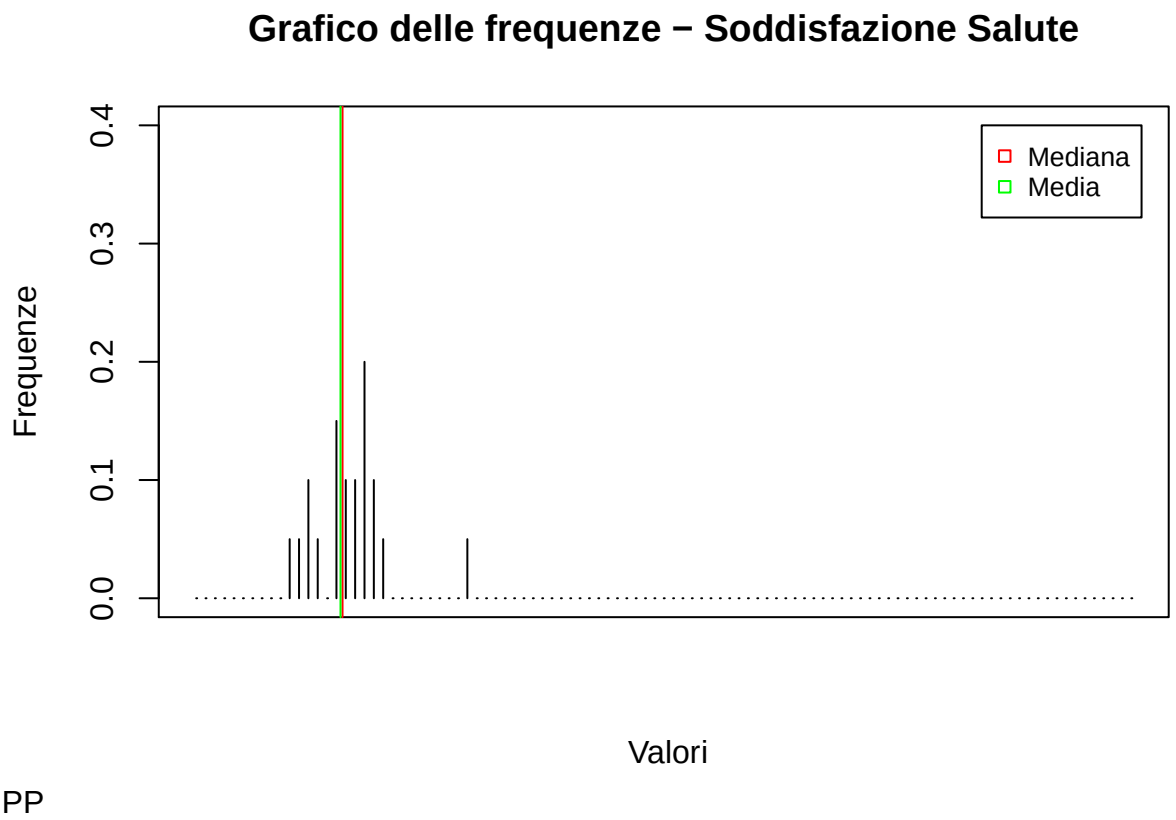

Figura 3.2: Grafico delle frequenze - Soddisfazione Salute

3.3 Istogrammi

Gli istogrammi sono una rappresentazione grafica della distribuzione delle frequenze in classi. Graficamente sono costituiti da rettangoli i cui estremi delle basi coincidono con gli estremi delle classi. Inoltre le aree dei rettangoli sono uguali alla frequenza assoluta o relativa delle rispettive classi.

Il seguente codice permette di disegnare gli istogrammi per le caratteristiche del data set. I grafici sono mostrati in figura 3.11.

```
par(mfrow=c(3, 2))
for (i in 1:colCount) {
  hist(df[[i]], freq = TRUE,
     main = paste("Istogramma - ", column.names[i]),
     ylab = "Frequenza assoluta",
     xlab = paste("Soddifazione - ", column.names[i]))
}
```

3.4 Simmetria

In questo paragrafo analizzeremo la simmetria della distribuzione di frequenza tramite un indice detto Skewness campionario definito con il valore

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

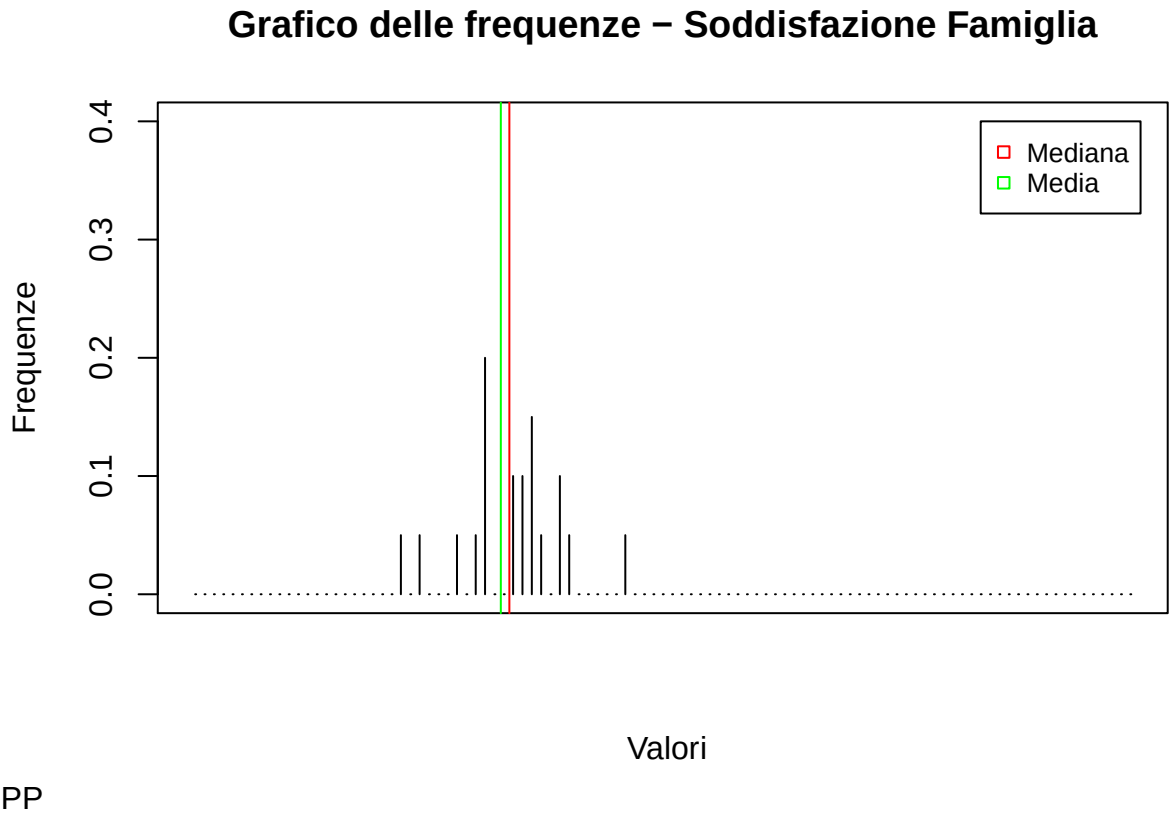

Figura 3.3: Grafico delle frequenze - Soddisfazione Famiglia

dove

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^j$$

è detto momento centrato campionario j-esimo.

La skewness campionaria assume il valore zero se la distribuzione è simmetrica, valori positivi nel caso di asimmetria positiva (sbilanciata a destra) o valori negativi nel caso di asimmetria negativa (sbilanciata a sinistra). Da notare che è Skewness campionaria è un numero puro.

Va inoltre notato che esistono distribuzioni non simmetriche con skewness campionaria pari a zero. Ne consegue che il valore zero è una condizione necessaria ma non sufficiente affinché la distribuzione sia simmetrica.

Il seguente codice permette di calcolare la Skewness campionaria e il tipo di simmetria. I valori ottenuti sono mostrati rispettivamente nelle tabelle 3.6 e 3.7

```
mcc <-function(x, j){
      sum((x - mean(x))^j) / length(x)
}

skw <-function(x) {
    mcc(x, 3) / mcc(x, 2) ^ 1.5
}

tipoSimmetria <- function(x) {
  if (x == 0) {
```

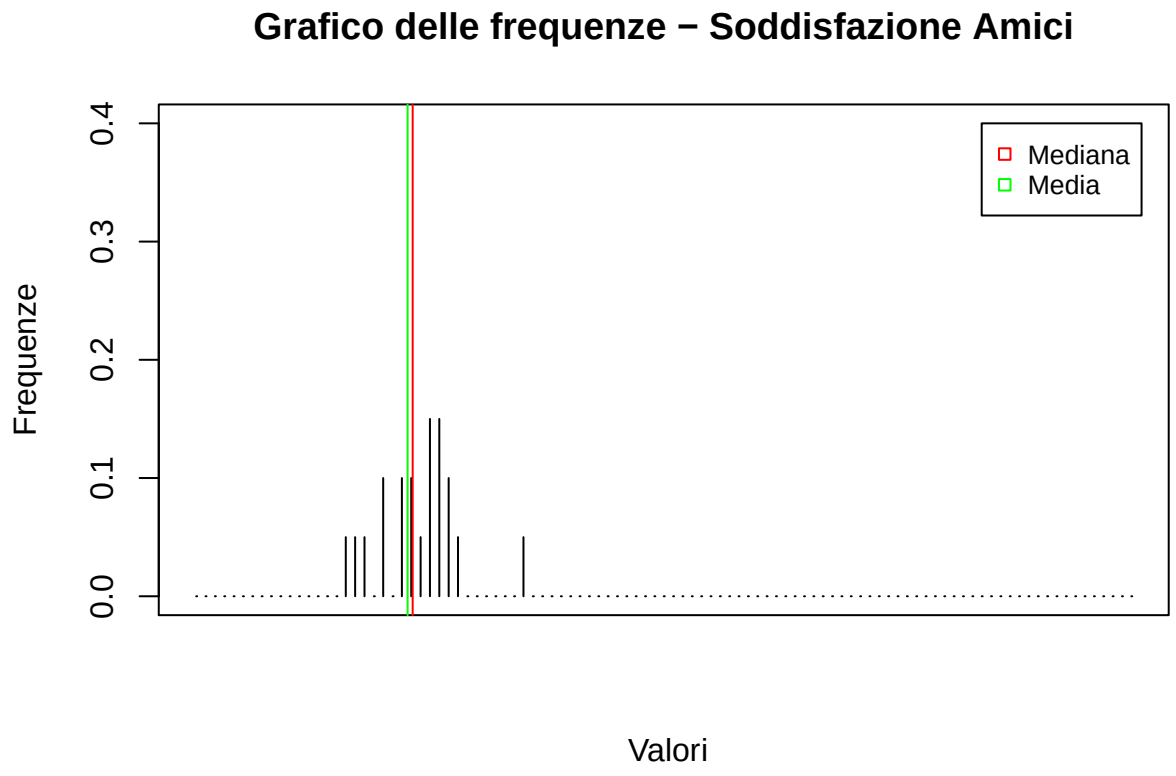

Figura 3.4: Grafico delle frequenze - Soddisfazione Amici

Tabella 3.6: Skewness campionaria

Caratteristica	Skewness campionaria
Economica	1.95847654825419
Salute	1.06874798713143
Famiglia	-0.0308002067356801
Amici	0.424929097467179
Tempo libero	0.811228347832538

```
    "simmetrica"
  } else if (x > 0) {
    "asimmetria positiva"
  } else {
    "asimmetria negativa"
  }
}

skewness <- sapply(df, skw)

tipoSimmetria <- as.matrix(sapply(skewness, tipoSimmetria))
colnames(tipoSimmetria) <- "Simmetria"
```

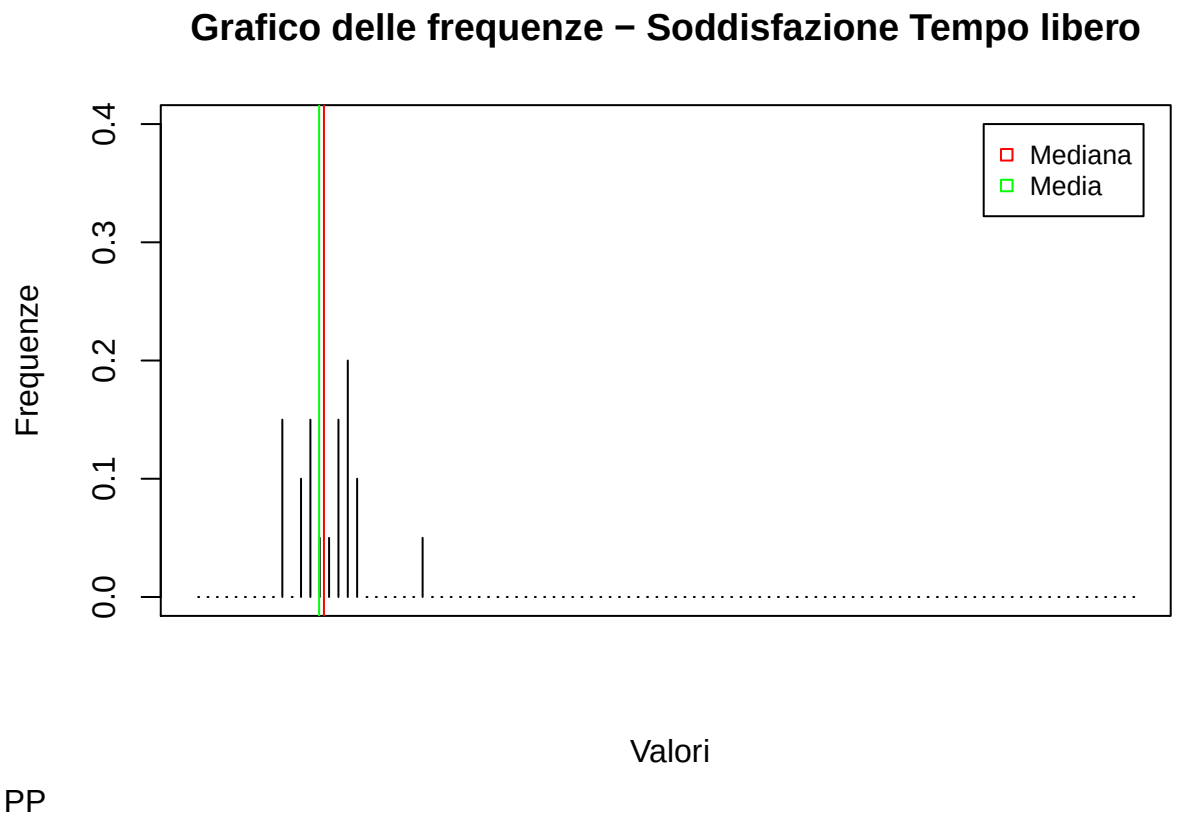

Figura 3.5: Grafico delle frequenze - Soddisfazione Tempo libero

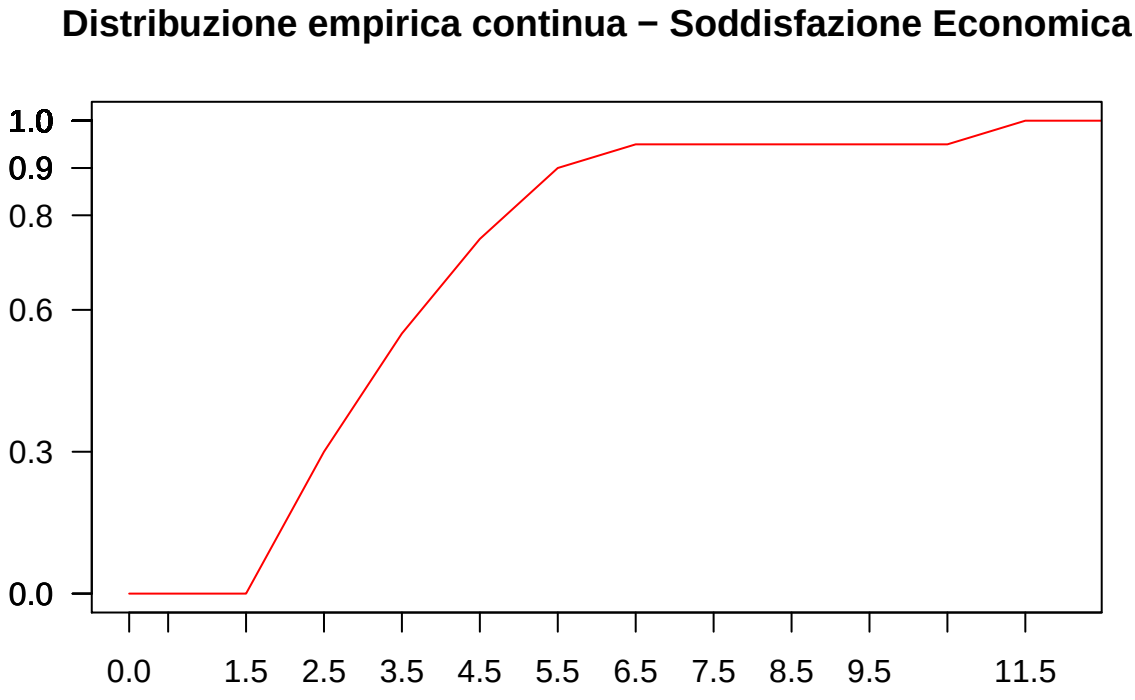

Figura 3.6: Distribuzione empirica continua - Soddisfazione Economica

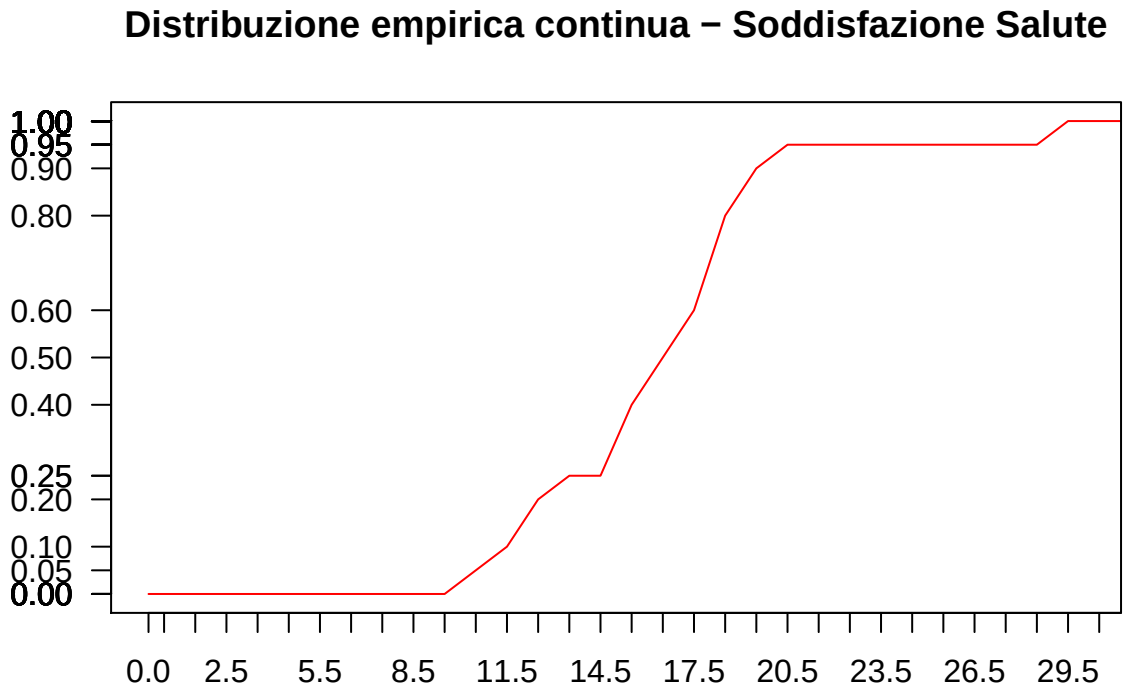

Figura 3.7: Distribuzione empirica continua - Soddisfazione Salute

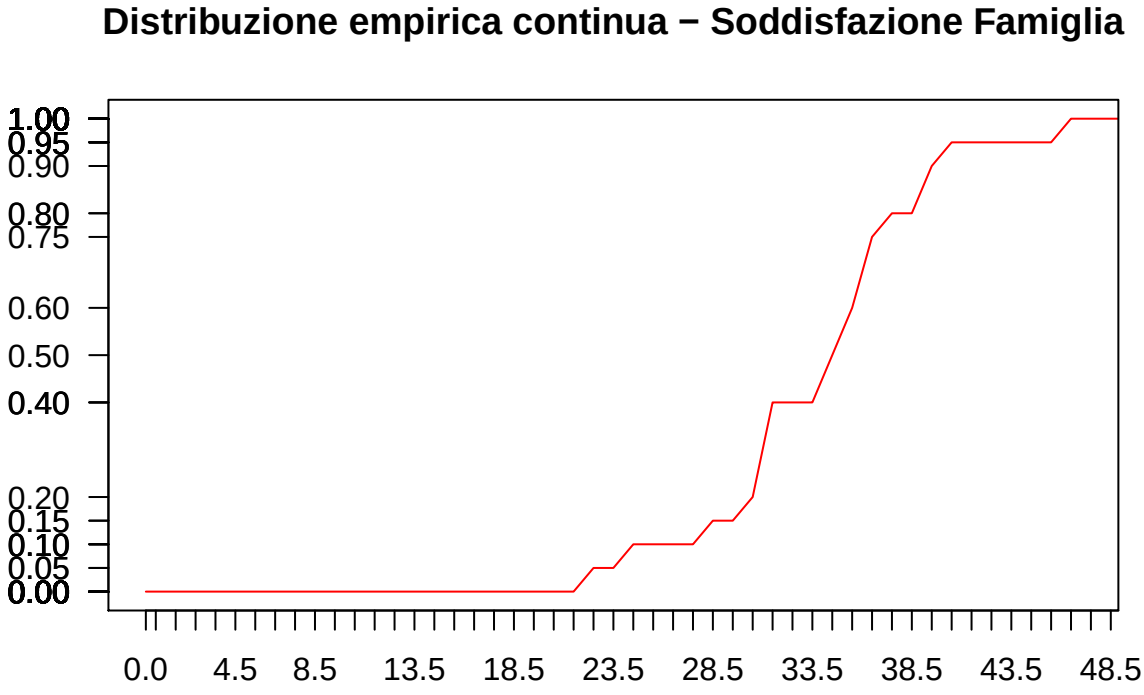

Figura 3.8: Distribuzione empirica continua - Soddisfazione Famiglia

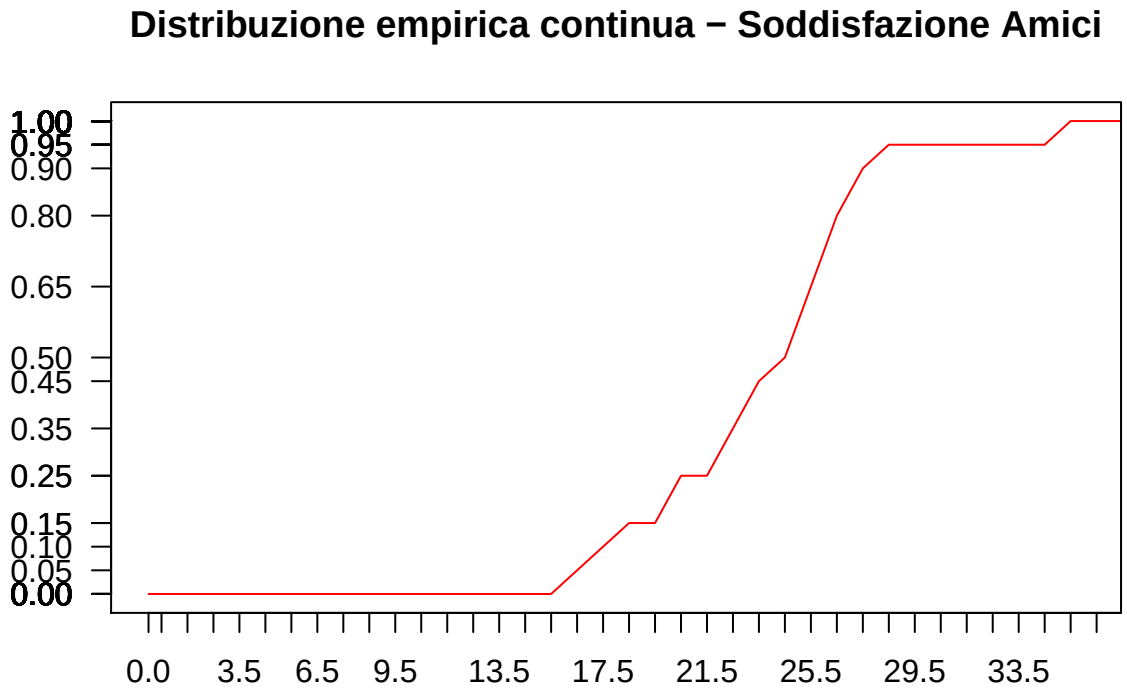

Figura 3.9: Distribuzione empirica continua - Soddisfazione Amici

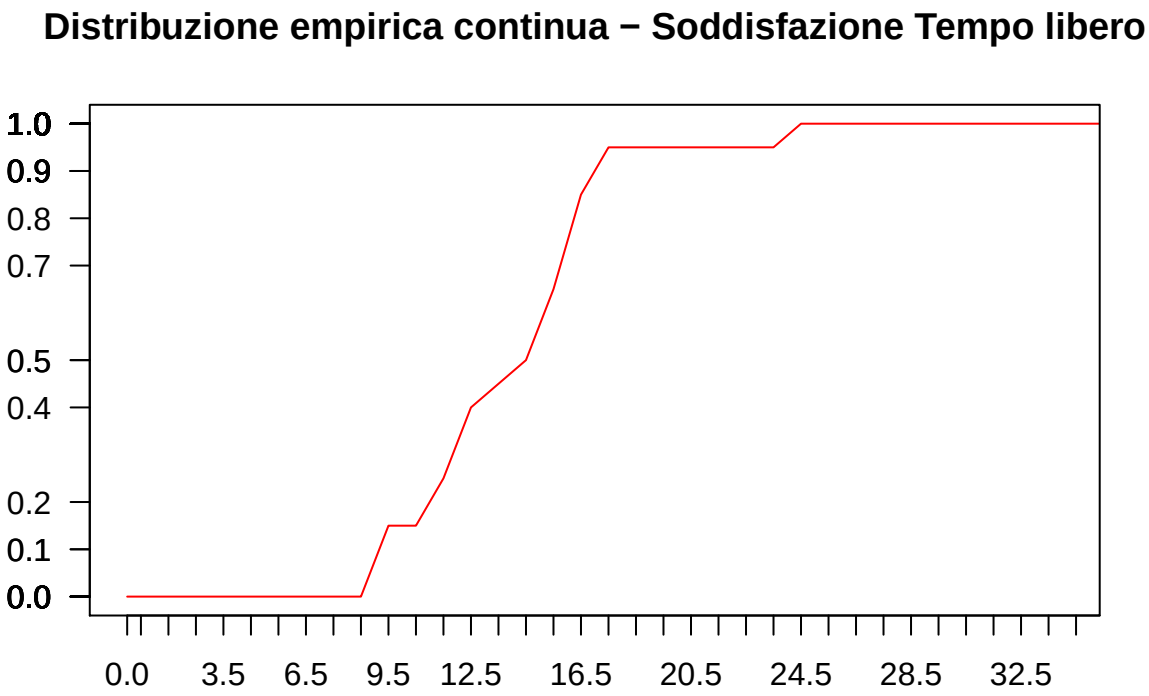

Figura 3.10: Distribuzione empirica continua - Soddisfazione Tempo Libero

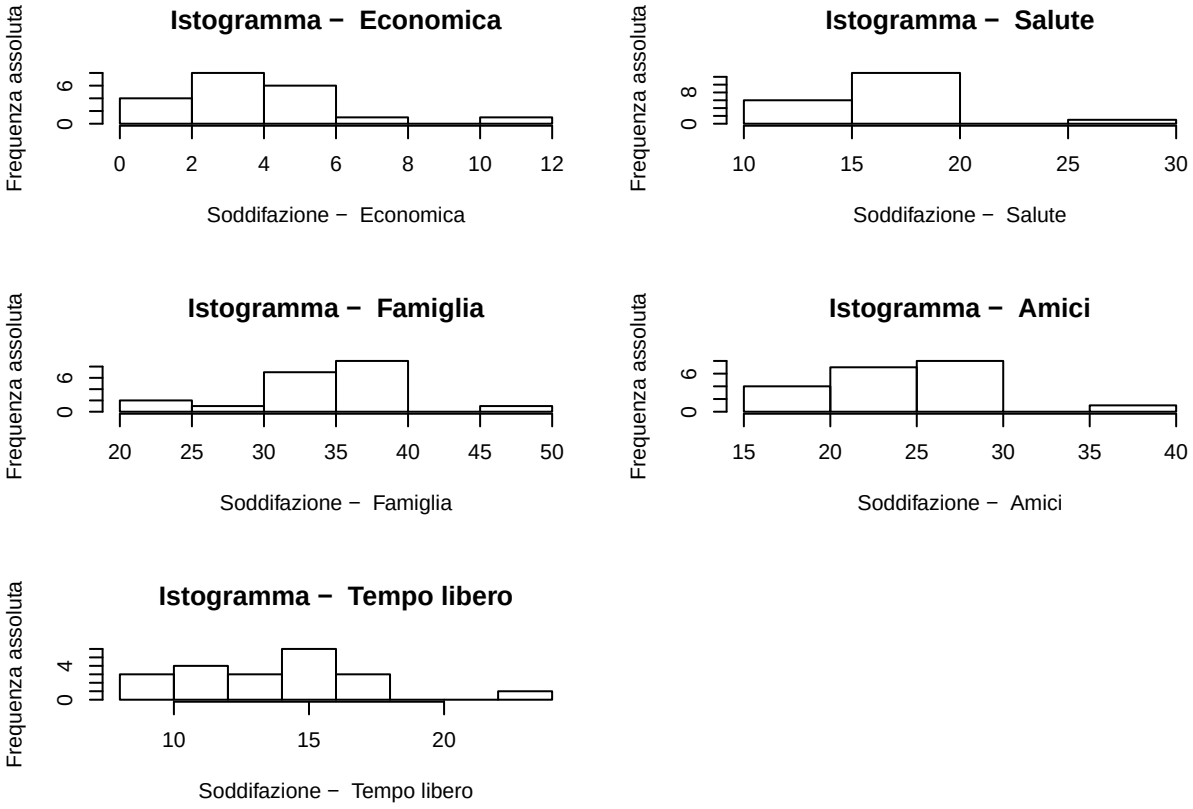

Figura 3.11: Istrogrammi

Tabella 3.7: Simmetria

Caratteristica	Simmetria
Economica	asimmetria positiva
Salute	asimmetria positiva
Famiglia	asimmetria negativa
Amici	asimmetria positiva
Tempo libero	asimmetria positiva

Tabella 3.8: Curtosi campionaria

Caratteristica	Curtosi campionaria
Economica	4.36873112268003
Salute	2.41958559064078
Famiglia	0.339928187129484
Amici	0.777502821228514
Tempo libero	1.10215748669869

Tabella 3.9: Tipo Curtosi campionaria

Caratteristica	Tipo Curtosi campionaria
Economica	leptocurtica
Salute	leptocurtica
Famiglia	leptocurtica
Amici	leptocurtica
Tempo libero	leptocurtica

3.5 Curtosi campionaria

Con curtosi si intende un allontanamento dalla normalità distributiva ed è misurato con il coefficiente di curtosi definito con il valore

$$\gamma_2 = \frac{m_4}{m_2^2} - 3$$

dove m_j è il momento centrato campionario j-esimo. Tale indice consente di dare una misura della piccatezza della distribuzione di frequenza del campione in confronto alla densità normale standard che assume un valore di curtosi campionaria pari a zero.

Se

- $\gamma_2 = 0$ la distribuzione è come una normale (normocurtica)
- $\gamma_2 > 0$ la distribuzione è più piccata di una normale (leptocurtica)
- $\gamma_2 < 0$ la distribuzione è meno piccata di una normale (platicurtica)

Il seguente codice permette di calcolare il valora della curtosi campionaria e il tipo di curtosi per ogni caratterista del dataset. I valori calcolati sono esposti rispettivamente nelle tabelle 3.8, 3.9.

```
calcolaCurtosi <- function(x) {
  mcc(x, 4) / mcc(x, 2) ^ 2 - 3
}

calcolaTipoCurtosi <- function(x) {
  if (x == 0) {
    "normocurtica"
  } else if (x > 0) {
    "leptocurtica"
  } else {
    "platicurtica"
  }
}

curtosi <- sapply(df, calcolaCurtosi)
tipoCurtosi <- sapply(curtosi, calcolaTipoCurtosi)
```

Ne consegue che per tutte le caratteristiche la distribuzione di frequenza è leptocurtica.

Capitolo 4

Indici di posizione

Gli indici di posizione sono indici di sintesi che danno un'idea dell'ordine di grandezza dei valori. Si dividono in indici di posizione centrali e non centrali.

I più importanti indici di posizione centrali sono la media campionaria, la mediana campionaria e la moda campionaria. Nel seguito verranno analizzate le prime due statistiche tralasciando la moda campionaria in quanto poco significativa essendo i dati in forma percentuale. Successivamente verrano analizzati i quantili, che sono indici di posizione non centrali che ripartiscono il campione in parti di uguale numerosità.

4.1 Media campionaria

Dati n valori la media campionaria $\overline{x}$ è definita come la media aritmetica degli n valori.

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

Va notato che il valore della media campionaria è fortemente influenzato dai valori molto grandi o piccoli.

Il seguente codice permette di computare le medie campionarie per le caratteristiche del data set. I risultati sono mostrati nella tabella 4.1.

```
medie <- sapply(df, mean)
```

Come già osservato in forma visiva tramite l'analisi dei boxplot possiamo dire che in media gli italiani traggono il massimo della soddisfazione nei rapporti familiari seguiti dai rapporti con gli amici, la salute, il tempo libero e infine dalla situazione economica.

Tabella 4.1: Medie campionarie

Caratteristica	Media campionaria
Economica	3.74
Salute	16.435
Famiglia	33.685
Amici	23.61
Tempo libero	13.935

Tabella 4.2: Scarti dalle medie campionarie

	Economica	Salute	Famiglia	Amici	Tempo libero
Piemonte	1.06	1.165	2.515	0.89	1.465
Valle d'Aosta	1.36	3.065	1.615	1.89	1.065
Liguria	0.16	1.765	5.815	2.89	2.465
Lombardia	0.56	1.965	1.615	1.89	2.065
Trentino Alto Adige	7.06	12.265	12.715	11.59	9.865
Veneto	0.96	2.265	4.815	1.49	1.465
Friuli-Venezia Giulia	2.66	0.865	3.615	3.19	2.665
Emilia-Romagna	0.36	2.765	5.015	4.79	1.565
Toscana	-0.74	0.665	1.915	1.79	1.565
Umbria	0.46	-0.435	2.015	2.39	3.365
Marche	-0.34	-1.135	0.015	-0.41	-0.035
Lazio	-0.84	-1.835	-2.585	-0.81	-0.935
Abruzzo	-0.94	1.865	0.215	0.19	-2.035
Molise	-1.54	-1.135	-3.085	-3.91	-1.835
Campania	-1.84	-3.435	-9.385	-7.11	-4.635
Puglia	-1.14	-5.435	-11.485	-7.41	-4.635
Basilicata	-1.54	-4.335	-3.185	-2.01	-3.335
Calabria	-2.04	-6.235	-5.685	-5.81	-4.735
Sicilia	-1.84	-0.235	-3.185	-3.51	-2.335
Sardegna	-1.84	-4.435	-3.285	-2.01	-3.035

4.1.1 Scarto dalla media campionaria

Per valutare quanto i valori si discostano dalla media campionaria è utile computare lo scarto dalla media campionaria. Tale quantità è definita per ogni valore di ogni caratteristica come la differenza tra il valore è la media campionaria della caratteristica considerata.

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

Va notato che la somma degli scarti dalla media è sempre nulla, infatti

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

Il seguente codice permette di computare gli scarti dalla media campionaria per ogni valore di ogni caratteristica. I risultati sono mostrati nella tabella 4.2.

```
scarti <- matrix(0, nrow = rowCount, ncol = 0)
for (i in 1:colCount) {
  scarti <- cbind(scarti, df[i] - medie[i])
}
```

Dai risultati ottenuti si può osservare che gli scarti maggiori derivano dai dati del "Trentino Alto Adige" il quale rappresenta un valore anomalo come già riscontrato nell'analisi dei boxplot.

4.2 Mediana campionaria

La mediana campionaria è la statistica che bipartisce il campione in due parti di uguale numerosità.

Tabella 4.3: Mediane campionarie

Caratteristica	Mediana campionaria
Economica	3.2
Salute	16.65
Famiglia	34.6
Amici	24.15
Tempo libero	14.45

La mediana campionaria di un campione di ampiezza n è definita come il valore in posizione $(n+1)/2$ del campione ordinato in modo non decrescente se n è dispari, altrimenti è definito come la media aritmetica dei valori in posizione $n/2$ e $n/2+1$ del medesimo campione ordinato. Di conseguenza, al contrario della media campionaria, la mediana campionaria dipende da al più due valori.

Il seguente codice permette di computare le mediane campionarie per le caratteristiche del data set. I risultati sono mostrati nella tabella 4.3.

```
mediane <- sapply(df, median)
```

Si può infine notare che per tutte le caratteriste eccetto per la soddisfazione economica la mediana campionaria è leggermente maggiore della media campionaria.

4.3 Quantili

I quantili sono indici che dividono il campione in un numero fissato di parti di uguale numerosità. In numero di parti può essere uguale a qualsiasi numero naturale positivo, in genere il campione è diviso in cento quantili detti anche percentili o quattro quantili detti anche quartili.

In R esistono nove algoritmi per il calcolo dei quantili, tuttavia, i valori tendono a coincidere con qualsiasi algoritmo quando il campione ha un'ampiezza elevata.

Il seguente codice consente di calcolare i quartili per la soddisfazione economica utilizzando tutti gli algoritmi. I risultati sono esposti nella tabella 4.4. Come è possibile osservare i quantili ottenuti non presentano significative differenze con i differenti algoritmi.

```
tipiquartili <- function(x) {
  y <- numeric(0)
  for(i in 1:9) {
    y <- rbind(y, c(quantile(x, 0, type = i),
                    quantile(x, 0.25, type = i),
                    quantile(x, 0.5, type = i),
                    quantile(x, 0.75, type = i),
                    quantile(x, 1, type = i)))
  }
  rownames(y) <- paste("type", 1:9)
  return (y)
}

quartili <- tipiquartili(as.matrix(df[1]))
```

La seguente funzione consente di calcolare i quartili per un dato tipo di algoritmo. Tale funzione verrà utilizzata nel seguito per calcolare i quartili di tutte le caratteristiche con alcuni tipi di algoritmi.

Tabella 4.4: Quartili - Soddisfazione Economica

	0%	25%	50%	75%	100%
type 1	1.7	2.2	3.0	4.300000	10.8
type 2	1.7	2.2	3.2	4.500000	10.8
type 3	1.7	2.2	3.0	4.300000	10.8
type 4	1.7	2.2	3.0	4.300000	10.8
type 5	1.7	2.2	3.2	4.500000	10.8
type 6	1.7	2.2	3.2	4.600000	10.8
type 7	1.7	2.2	3.2	4.400000	10.8
type 8	1.7	2.2	3.2	4.533333	10.8
type 9	1.7	2.2	3.2	4.525000	10.8

Tabella 4.5: Quartili tipo 2

	0%	25%	50%	75%	100%
Economica	1.7	2.20	3.20	4.50	10.8
Salute	10.2	13.80	16.65	18.35	28.7
Famiglia	22.2	30.50	34.60	36.75	46.4
Amici	16.2	20.85	24.15	25.75	35.2
Tempo libero	9.2	11.25	14.45	15.75	23.8

```
dataframe.quantile <- function(df, type) {
  y <- numeric(0)
  for (x in 1:colCount) {
    value <- df[[x]]
    row <- c(quantile(value, probs = 0, type = type),
             quantile(value, probs = 0.25, type = type),
             quantile(value, probs = 0.5, type = type),
             quantile(value, probs = 0.75, type = type),
             quantile(value, probs = 1, type = type))
    y <- rbind(y, row)
  }
  # colnames(y) <- colnames(df)
  row.names(y) <- column.names
  return (y)
}
```

4.3.1 Quantili con il tipo 2

Il percentile k-esimo con l'algoritmo di tipo 2 si calcola nel seguente modo:

1. Si ordina il campione v in modo non decrescente.
2. Si calcola l'indice $h = np = \frac{nk}{100}$.
3. Se h è un intero il percentile k-esimo è pari a $(v[h] + v[h+1])/2$ altrimenti è pari a $v[ceiling(h)]$.

Il seguente codice permette di calcolare i quartili con l'algoritmo di tipo 2. I risultati sono mostrati nella tabella 4.5.

```
quartili2 <- dataframe.quantile(df, 2)
```

Tabella 4.6: Quartili tipo 7

	0%	25%	50%	75%	100%
Economica	1.7	2.200	3.20	4.400	10.8
Salute	10.2	14.200	16.65	18.325	28.7
Famiglia	22.2	30.500	34.60	36.475	46.4
Amici	16.2	21.225	24.15	25.625	35.2
Tempo libero	9.2	11.425	14.45	15.625	23.8

Tabella 4.7: Quartili tipo 1

	0%	25%	50%	75%	100%
Economica	1.7	2.2	3.0	4.3	10.8
Salute	10.2	13.0	16.2	18.3	28.7
Famiglia	22.2	30.5	33.9	36.2	46.4
Amici	16.2	20.1	23.8	25.5	35.2
Tempo libero	9.2	10.9	13.9	15.5	23.8

4.3.2 Quantili con il tipo 7

Il percentile k-esimo con l'algoritmo di tipo 7 (default in R) si basa su una tecnica di interpolazione lineare e si calcola nel seguente modo:

1. Si ordina il campione v in modo non decrescente.
2. Si calcola l'indice $h = (n-1)p + 1 = (n-1)\frac{k}{100} + 1$.
3. Calcolare $h^* = floor(h)$.
4. Il percentile k-esimo è pari a $v[h^*] + (h - h^*) * [v(h^* + 1) - v(h^*)]$.

Il seguente codice permette di calcolare i quartili con l'algoritmo di tipo 7. I risultati sono mostrati nella tabella 4.6.

```
quartili7 <- dataframe.quantile(df, 7)
```

4.3.3 Quantili con il tipo 1

Se un campione può assumere k modalità distinte $z_1 < z_2 < ... < z_k$ allora il percentile k-esimo con l'algoritmo di tipo 1 è definito come la modalità i-esima che soddisfa la doppia disuguaglianza $F_{i-1} < k/100$ e $F_i \geq k/100$ dove $F_1, ..., F_k$ sono le frequenze relative cumulate.

Il seguente codice permette di calcolare i quartili con l'algoritmo di tipo 1. I risultati sono mostrati nella tabella 4.7.

```
quartili1 <- dataframe.quantile(df, 1)
```

4.3.4 Rappresentazione grafica dei quartili

Una rappresentazione grafica dei quartili sul diagramma della funzione di distribuzione empirica per ognuna delle caratteristiche è mostrata nelle figure 4.1, 4.2, 4.3, 4.4, 4.5. Nei grafici la linea blu rappresenta il primo quartile, la linea gialla il secondo e la linea verde il terzo.

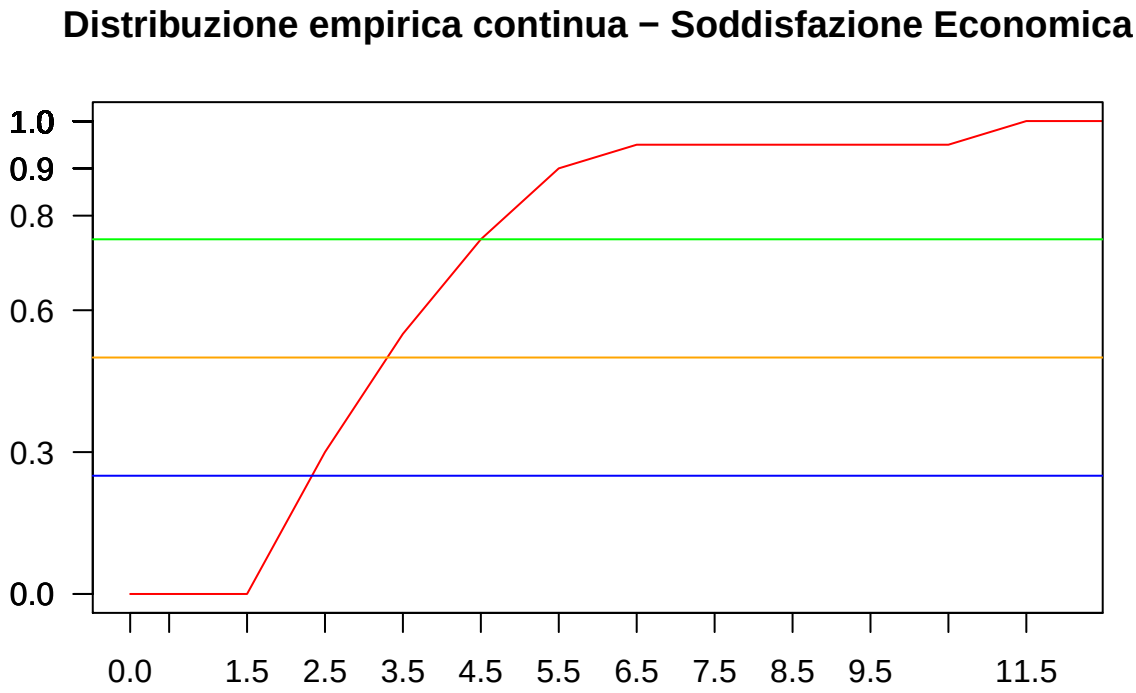

Figura 4.1: Distribuzione empirica continua e quartili - Soddisfazione Economica

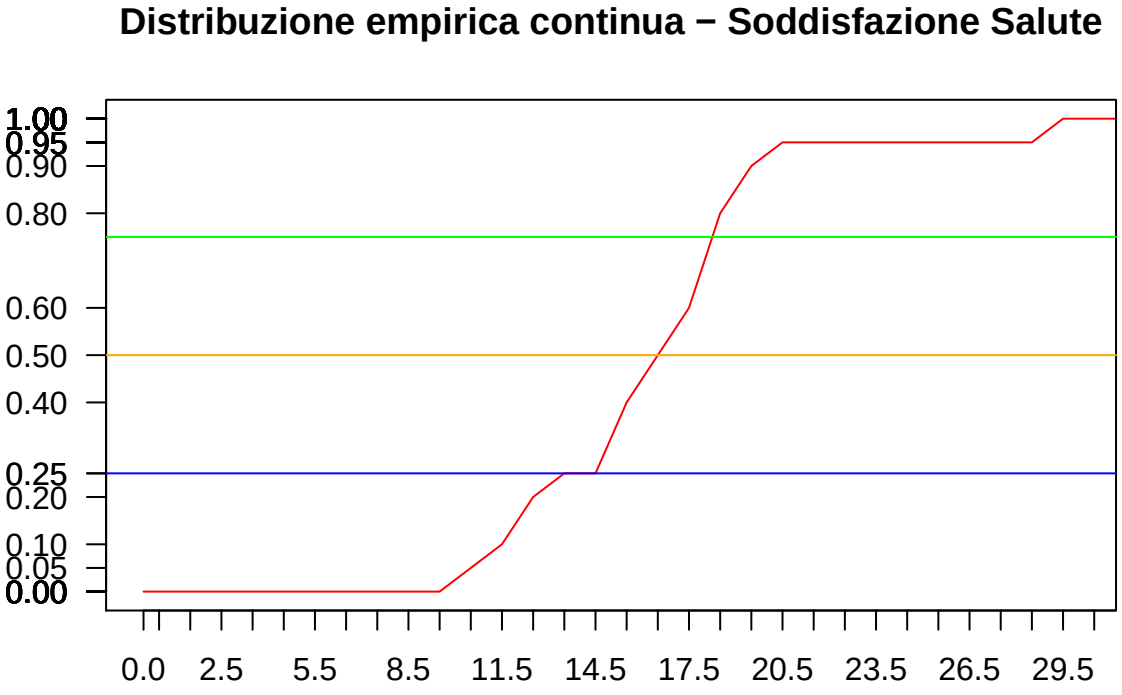

Figura 4.2: Distribuzione empirica continua e quartili - Soddisfazione Salute

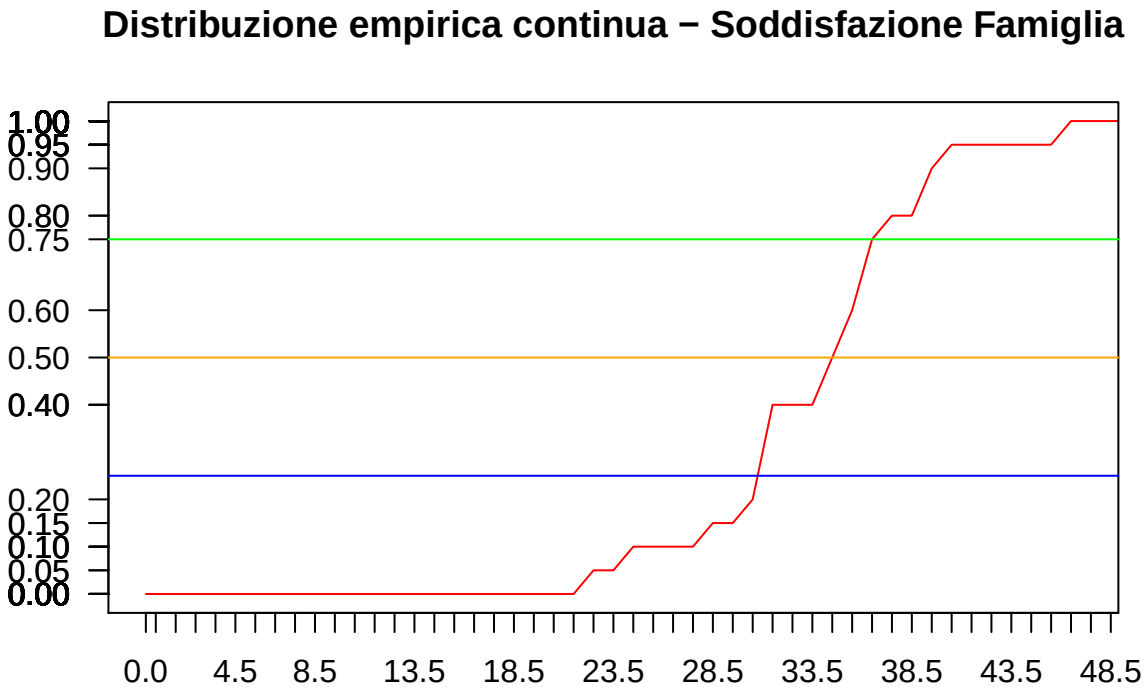

Figura 4.3: Distribuzione empirica continua e quartili - Soddisfazione Famiglia

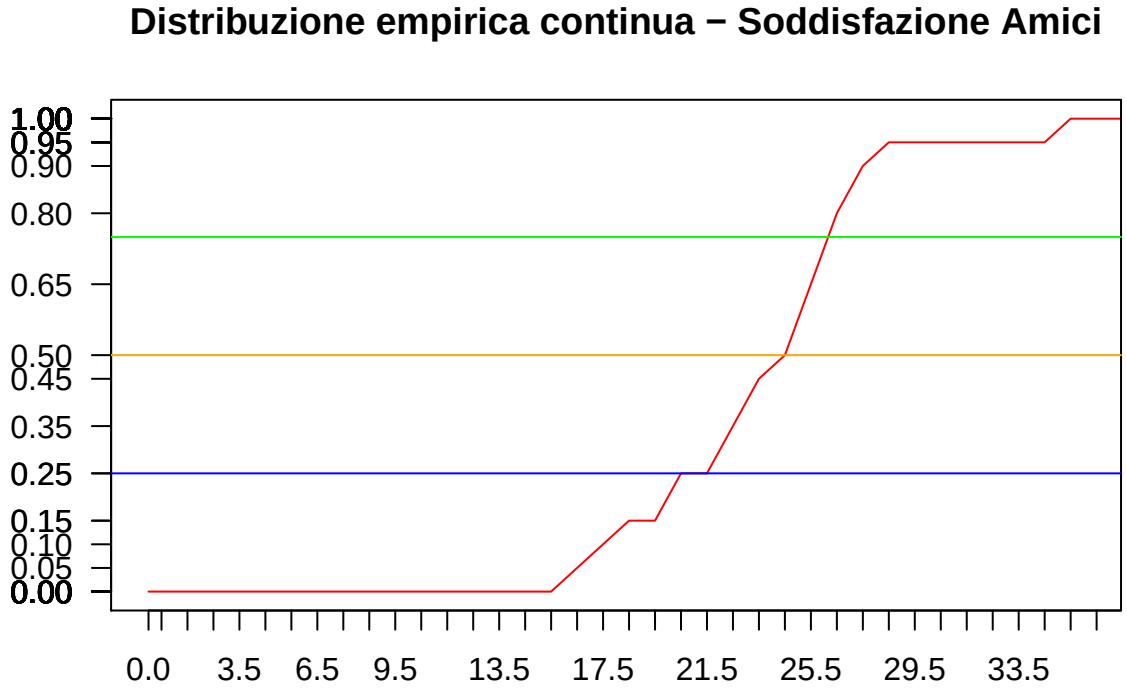

Figura 4.4: Distribuzione empirica continua e quartili - Soddisfazione Amici

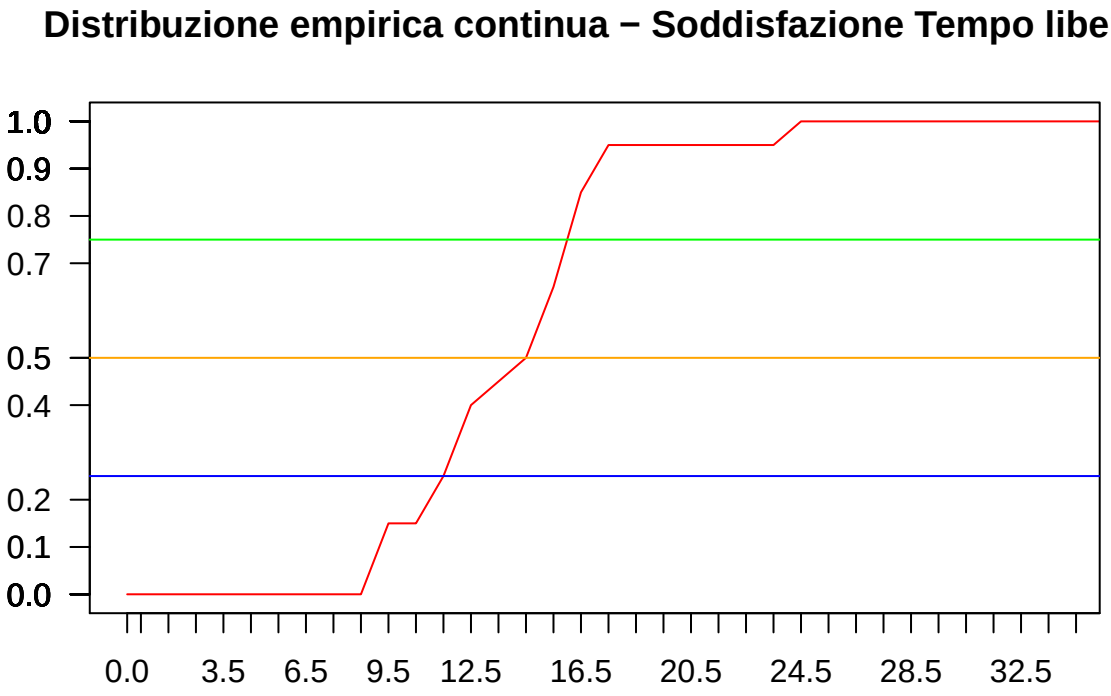

Figura 4.5: Distribuzione empirica continua e quartili - Soddisfazione Tempo libero

Capitolo 5

Indici di dispersione

Gli indici di dispersione sono indici di sintesi che descrivono quanto un valore è distante da un indice centrale. Gli indici di dispersione considerati nel seguito sono la varianza campionaria, la deviazione standard campionaria e il coefficiente di variazione.

5.1 Varianza campionaria

La varianza campionaria fornisce una misura di quanto i valori si discostano quadraticamente dalla media campionaria.

Formalmente dato un campione $x_1, ..., x_n$, la varianza campionaria s^2 è definita con la quantità

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_1 - \overline{x})^2$$

Ne consegue che se i valori del campione sono tutti uguali la varianza campionaria è nulla, altrimenti il valore della varianza campionaria è tanto più grande quando più i dati si discostano dalla media.

Il seguente codice permette di computare la varianza campionaria per ogni caratteristica del data set. I risultati sono mostrati nella tabella 5.1.

```
varianze <- sapply(df, var)
```

Dai valori ottenuti si può notare che la variabilità maggiore si ottiene nella soddisfazione in famiglia seguita da quella con gli amici, della salute, del tempo libero, e infine economica.

Tabella 5.1: Varianze campionarie

Caratteristica	Varianza campionaria
Economica	4.42673684210526
Salute	16.3402894736842
Famiglia	30.3045
Amici	19.5072631578947
Tempo libero	12.4192368421053

Tabella 5.2: Deviazione standard campionarie

Caratteristica	Deviazione standard campionaria
Economica	2.10398118862913
Salute	4.04231239189702
Famiglia	5.50495231586978
Amici	4.41670274728725
Tempo libero	3.52409376182094

Tabella 5.3: Range di variazione

Caratteristica	Range di variazione
Economica	9.1
Salute	18.5
Famiglia	24.2
Amici	19
Tempo libero	14.6

5.2 Deviazione standard campionaria

La deviazione standard campionaria detta anche scarto quadratico medio campionario è definita come la radice quadrata della varianza campionaria.

$$s = \sqrt{s^2}$$

Da notare che la deviazione standard campionaria ha le stesse unità di misura del campione.

Il seguente codice permette di computare la deviazione standard campionaria per ogni caratteristica del data set. I risultati sono mostrati nella tabella 5.2.

```
deviazioniStandard <- sapply(df, sd)
```

Per la deviazione standard campionaria le considerazioni sui risultati sono analoghe a quelle della varianza campionaria.

5.3 Coefficiente di variazione

Il coefficiente di variazione è definito come il rapporto tra la deviazione standard campionaria e il valore assoluto della media campionaria. Essendo un numero puro, cioè adimensionale, può essere utilizzato per confrontare le varianze di campioni con unità di misura diverse o con differenti range di variazione. Dove con range di variazione si intende la differenza tra il massimo e il minimo nel campione. Inoltre va notato che il coefficiente di variazione può essere calcolato solo se la media campionaria non è nulla.

Il seguente codice permette di calcolare il range di variazione per ogni caratteristica. I risultati sono mostrati nella tabella 5.3.

```
range <- sapply(df, max) - sapply(df, min)
```

Possiamo osservare che il range di variazione maggiore è relativo alla soddisfazione in famiglia seguita dalla soddisfazione con gli amici, salute, tempo libero e per ultimo quella economica.

Il seguente codice permette di calcolare il coefficiente di variazione per ogni caratteristica. I valori sono mostrati nella tabella 5.4.

Tabella 5.4: Coefficiente di variazione

Caratteristica	Coefficiente di variazione
Economica	0.56256181514148
Salute	0.245957553507577
Famiglia	0.163424441617034
Amici	0.187069154904161
Tempo libero	0.252895138989662

```
cv <- function(x) {
  sd(x) / abs(mean(x))
}
coeffVariazione <- sapply(df, cv)
```

Il coefficiente di variazione maggiore si ottiene con la soddisfazione economica, seguita da quella per il tempo libero, per la salute, per gli amici e infine per la famiglia.

Capitolo 6

Correlazioni

In questo capitolo l'attenzione verrà posta sulle relazioni tra le caratteristiche del data set. Le relazioni possono essere descritte visivamente tramite diagrammi di dispersione o tramite una misura quantitativa detta covarianza campionaria.

6.1 Diagrammi si dispersione

I diagrammi di dispersione detti anche scatterplot sono rappresentazioni grafiche della correlazione tra variabili.

Dato un campione bivariato (x_i, y_i) per $i = 1, ...n$ rappresentante n osservazioni delle variabili (X, Y) il relativo diagramma di dispersione consiste nel riportare i punti del campione su un piano cartesiano dopo aver scelto una variabile indipendente e una dipendente. Dalla visione del grafico è possibile intuire un'eventuale regolarità tra i valori.

Il seguente codice permette di creare tutti gli scatterplot tra le caratteristiche considerate. I diagrammi ottenuti sono mostrati in figura 6.1.

```
pairs(df)
```

In ogni diagramma la variabile indipendente è descritta dall'etichetta presente nella stessa colonna del diagramma, mentre la variabile dipendente è descritta dall'etichetta presente nella stessa riga del diagramma. Dalla visualizzazione del grafico si può notare una correlazione positiva tra tutte le caratteristiche prese a due a due.

6.2 Covarianza campionaria

La covarianza campionaria consente di dare una misura quantitativa della correlazione tra le variabili. Dato un campione bivariato (x_i, y_i) per $i = 1, ...n$ la covarianza campionaria tra le variabili X e Y è definita con il valore

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

Se la covarianza campionaria assume valore zero ne consegue che le variabili non sono correlate, altrimenti all'aumentare del valore assoluto della covarianza aumenta la correlazione. In particolare la correlazione è positiva se il valore della covarianza è positivo, negativa altrimenti.

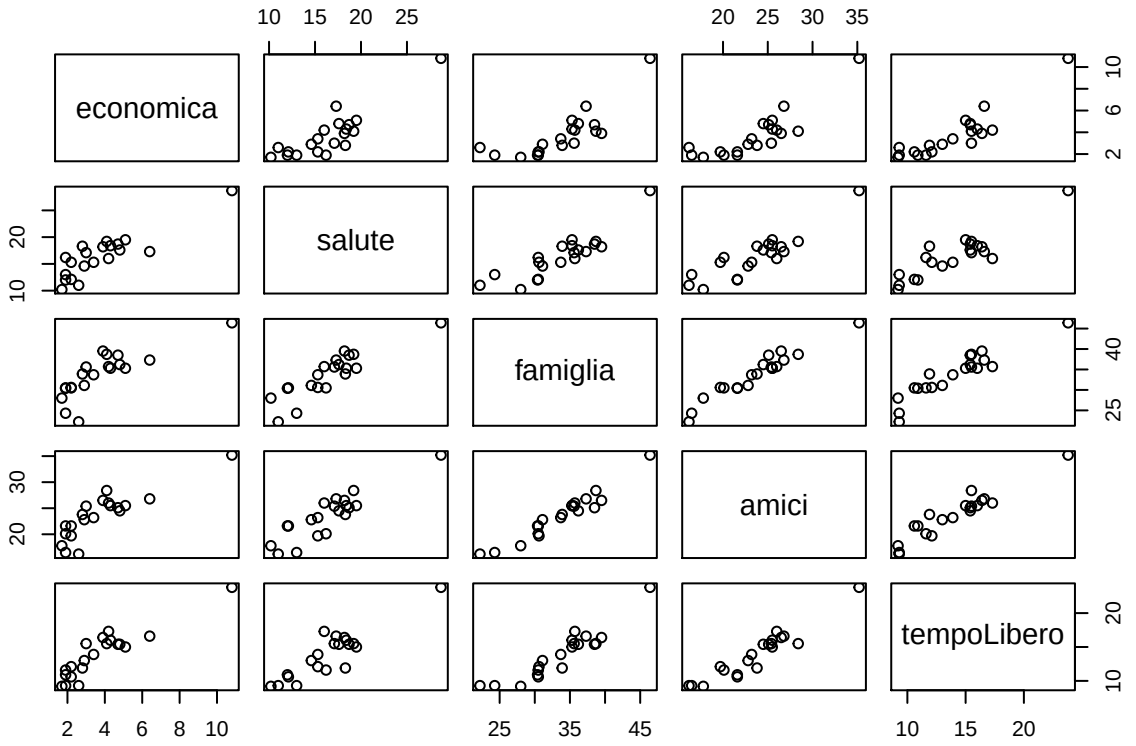

Figura 6.1: Diagrammi si dispersione

Tabella 6.1: Correlazioni

	Economica	Salute	Famiglia	Amici	Tempo libero
Economica	4.426737	7.399053	9.342211	7.977474	6.719579
Salute	7.399053	16.340290	19.653710	15.970684	12.791868
Famiglia	9.342211	19.653710	30.304500	23.511737	18.007395
Amici	7.977474	15.970684	23.511737	19.507263	14.793316
Tempo libero	6.719579	12.791868	18.007395	14.793316	12.419237

Il seguente codice permette di calcolare la covarianza tra le caratteristiche del data set prese a due a due. I risultati ottenuti sono mostrati nella tabella 6.1. Da notare che la correlazione tra le stesse variabili è uguale alla varianza tra le variabili.

```
covarianze <- cov(df)
```

6.3 Coefficiente di correlazione campionario

Il coefficiente di correlazione campionario è un coefficiente per misurare in modo quantitativo la relazione lineare tra le variabili ed è definito come

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

Assume valori compresi tra -1 e 1, in particolare assume valore zero nel caso in cui le variabili non sono correlate, mentre all'aumentare del valore assoluto del coefficiente aumenta la correlazione tra le variabili. La correlazione è positiva se il segno del coefficiente è positivo, negativa altrimenti.

Tabella 6.2: Correlazioni

	Economica	Salute	Famiglia	Amici	Tempo libero
Economica	1.0000000	0.8699702	0.8065926	0.8584705	0.9062599
Salute	0.8699702	1.0000000	0.8832042	0.8945312	0.8979593
Famiglia	0.8065926	0.8832042	1.0000000	0.9670145	0.9282177
Amici	0.8584705	0.8945312	0.9670145	1.0000000	0.9504295
Tempo libero	0.9062599	0.8979593	0.9282177	0.9504295	1.0000000

Tabella 6.3: Coefficiente di correlazione campionario

	Coef. Correlazione
Famiglia - Amici	0.9670145
Amici - Tempo libero	0.9504295
Famiglia - Tempo libero	0.9282177
Economica - Tempo libero	0.9062599
Salute - Tempo libero	0.8979593
Salute - Amici	0.8945312
Salute - Famiglia	0.8832042
Economica - Salute	0.8699702
Economica - Amici	0.8584705
Economica - Famiglia	0.8065926

Inoltre il coefficiente di correlazione campionario non fa distinzione tra la variabile indipendente e la variabile dipendente.

Il seguente codice permette di calcolare il coefficiente di correlazione campionario tra le caratteristiche del data set prese a due a due. I risultati ottenuti sono mostrati nella tabella 6.2.

```
correlazioni <- cor(df)
```

Il seguente codice permette di calcolare i valori della tabella 6.3 dove è mostrato il coefficiente di correlazione campionario per tutte le caratteristiche prese a due a due in ordine non crescente.

```
cc <- numeric()
for (i in 1:colCount) {
  for (j in 1:colCount) {
    if (i < j) {
      value <-array(correlazioni[i, j], c(1, 1))
      name <- paste(row.names(correlazioni)[i], "-",
                    row.names(correlazioni)[j])
      rownames(value) <- c(name)
      cc <- rbind(cc, value)
    }
  }
}
lista.corr <- as.matrix(cc[order(cc[,1], decreasing = TRUE),])
```

Possiamo osservare che la correlazione maggiore si presenta tra la soddisfazione in famiglia e la soddisfazione con gli amici. Mentre la correlazione minore tra la soddisfazione in famiglia e quella economica. Si nota inoltre che tutte le correlazioni sono positive.

Tabella 6.4: Coefficienti di determinazione

	Coef. determinazione
Famiglia - Amici	0.9351170
Amici - Tempo libero	0.9033162
Famiglia - Tempo libero	0.8615882
Economica - Tempo libero	0.8213070
Salute - Tempo libero	0.8063309
Salute - Amici	0.8001860
Salute - Famiglia	0.7800496
Economica - Salute	0.7568482
Economica - Amici	0.7369715
Economica - Famiglia	0.6505916

6.4 Coefficiente di determinazione

Per determinare se il modello di regressione spiega i dati è utile considerare il coefficiente di determinazione, anche chiamato r-square, e definito con il valore

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

Dove con $\hat{y}_i$ e y_i sono rispettivamente è l'i-esimo valore stimato e osservato, mentre $\bar{y}$ è la media campionaria dei valori osservati.

Il coefficiente di determinazione può essere anche visto come il rapporto tra la varianza dei valori stimati e quella dei valori osservati.

Se il coefficiente di determinazione assume valore zero vi è una completa incapacità del modello di spiegare i valori, altrimenti all'aumentare del valore del coefficiente di determinazione aumenta la capacità del modello di spiegare i valori. Nel caso limite in cui il coefficiente assume valore uno il modello di regressione spiega perfettamente i valori.

6.4.0.1 Coefficiente di determinazione nella regressione lineare

Nella regressione lineare il coefficiente di determinazione è equivalente al quadrato del coefficiente di determinazione

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

Il seguente codice permette di calcolare il coefficiente di determinazione per il modello lineare per tutte le caratteristiche prese a due a due. I risultati sono mostrati nella tabella 6.4.

```
coeffDeterminazione <- lista.corr ^ 2
```

Il coefficiente di determinazione minore è associato alle caratteristiche soddisfazione economica e soddisfazione in famiglia con un valore 0.6505916. Ne consegue che la regressione lineare non approssima in modo soddisfacente i valori. Per tutte le altre caratteristiche prese a due a due il modello lineare approssima i valori con buona soddisfazione, infatti il coefficiente di determinazione è maggiore di 0.7369715, ottenendo un massimo pari a 0.935117 tra la soddisfazione per la famiglia e quella per gli amici.

Nel seguito verrà mostrato il modello di regressione lineare e relativa analisi tra la soddisfazione per la famiglia e quella per gli amici. Verrà inoltre anche mostrato come la soddisfazione economica e la soddisfazione per la salute influiscono sulla soddisfazione nel tempo libero tramite una regressione lineare multipla. Infine verrà mostrato un modello di regressione non lineare per la correlazione tra la soddisfazione economica e quella in famiglia.

Tabella 6.5: Dati modello lineare

Intercetta	Coef. Angolare	Coef. Determinazione
-2.524497	0.7758497	0.935117

6.5 Regressione lineare

Il modello di regressione lineare è un metodo di stima che mira a trovare la retta che meglio interpola la nuvola dei punti.

I valori del coefficiente angolare β e dell'intercetta α vengono calcolati tramite il metodo dei minimi quadrati e hanno valori pari a

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

Il seguente codice calcola il modello lineare per la variabile indipendente associata alla soddisfazione per la famiglia e la variabile dipendente associata alla soddisfazione per gli amici. L'intercetta, il coefficiente angolare e il coefficiente di determinazione per tale modello sono mostrati nella tabella 6.5

```
y <- df$amici
x <- df$famiglia
modello <- lm(y~x)
intercetta <- modello$coefficients[[1]]
coeffAngolare <- modello$coefficients[[2]]
coeffDeterminazione <- summary(modello)$r.squared
```

Poiché il coefficiente angolare è positivo ne consegue che la retta di regressione è una retta crescente.

Invece il diagramma del modello di regressione è disegnato con il seguente codice ed è mostrato in figura 6.2

```
plot(x, y, xlab = "Soddisfazione Famiglia",
     ylab = "Soddisfazione Amici",
     main = "Regressione lineare semplice")
abline(modello, col="blue")
```

La stima dei valori tramite il modello lineare può differire dai valori osservati nel caso in cui i punti non sono tutti sulla retta interpolante. Con residuo si definisce la differenza tra il valore osservato e il valore stimato.

Può essere notato che la media campionaria dei residui è nulla, di conseguenza gli scostamenti positivi e negativi si compensano. Inoltre la media campionaria dei valori stimati è uguale alla media campionaria dei valori osservati.

Il seguente codice permette di calcolare i valori stimati, i residui, e i residui standardizzati. I risultati sono mostrati nella tabella 6.6.
Inoltre vengono calcolate alcune statistiche sui residui: la media campionaria, la varianza campionaria e deviazione standard campionaria. Non può essere calcolato il coefficiente di variazione poiché la media dei residui è zero. I risultati sono mostrati nella tabella 6.7.

```
osservati <- y
stime <- fitted(modello)
residui <-resid(modello)
residui.standard <- residui / sd(residui)

mediaResidui <- mean(residui)
varianzaResidui <- var(residui)
deviazioneStandardResidui <- sd(residui)
```

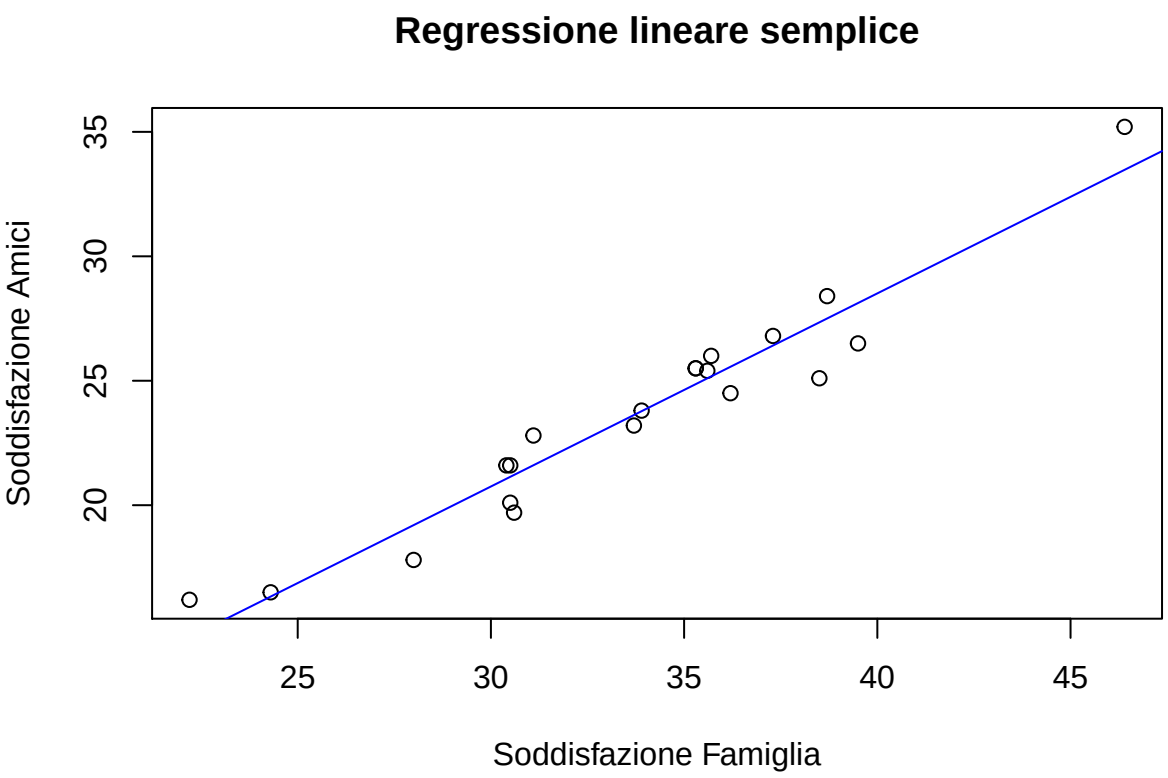

Figura 6.2: Diagramma di dispersione

Tabella 6.6: Residui

Osservati	Stimati	Residui	Residui standardizzati
24.5	25.56126	-1.0612620	-0.9433199
25.5	24.86300	0.6370028	0.5662102
26.5	28.12157	-1.6215659	-1.4413552
25.5	24.86300	0.6370028	0.5662102
35.2	33.47493	1.7250712	1.5333576
25.1	27.34572	-2.2457162	-1.9961414
26.8	26.41470	0.3853034	0.3424832
28.4	27.50089	0.8991138	0.7991919
25.4	25.09575	0.3042478	0.2704356
26.0	25.17334	0.8266629	0.7347927
23.2	23.62164	-0.4216377	-0.3747796
22.8	21.60443	1.1955714	1.0627031
23.8	23.77681	0.0231923	0.0206149
19.7	21.21650	-1.5165037	-1.3479690
16.5	16.32865	0.1713493	0.1523066
16.2	14.69937	1.5006337	1.3338626
21.6	21.13892	0.4610813	0.4098396
17.8	19.19929	-1.3992945	-1.2437857
20.1	21.13892	-1.0389187	-0.9234598
21.6	21.06133	0.5386662	0.4788022

Tabella 6.7: Statistiche Residui

Media	Varianza	Deviazione Standard
0	1.265689	1.125029

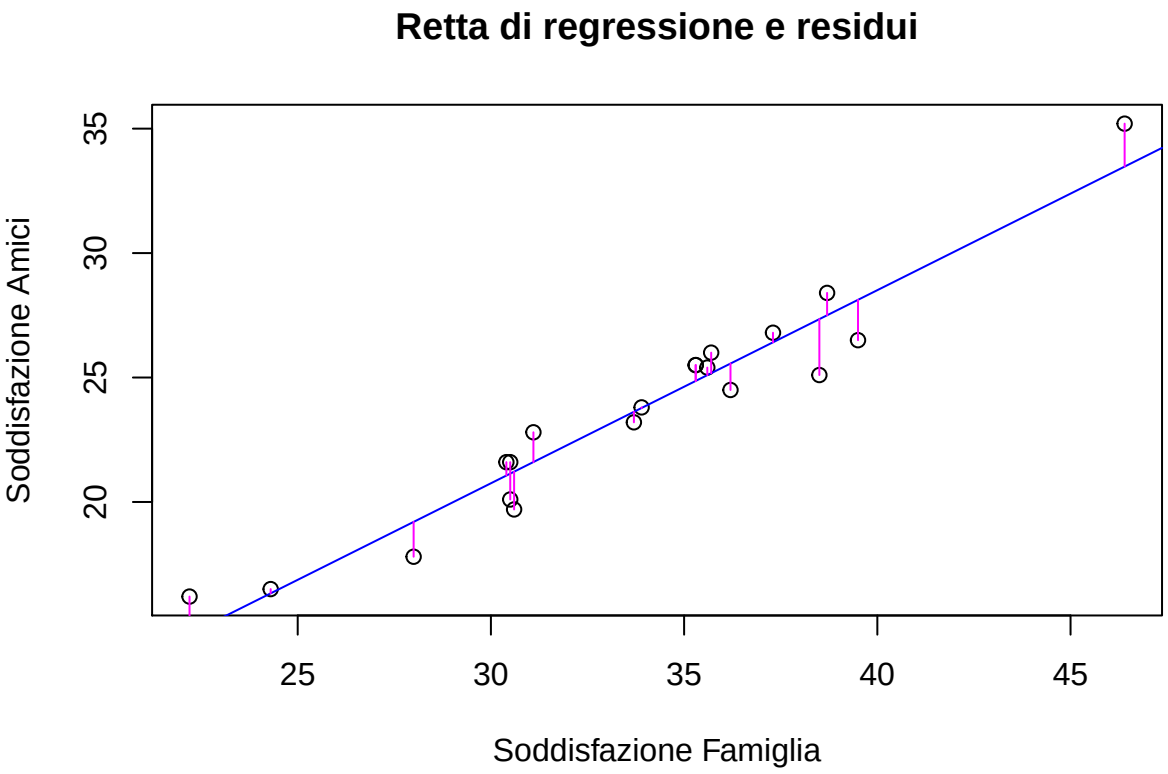

Figura 6.3: Diagramma di dispersione

Una visualizzazione grafica dei residui sulla retta di regressione è mostrata in figura 6.3 mentre un diagramma mostrante i valori dei residui rispetto ai valori stimati è mostrato in figura 6.4. In quest'ultimo grafico la retta orizzontale rappresenta la media campionaria dei residui di valore zero.

Come è possibile notare alcuni residui si discostano in modo maggiore degli altri dal valore medio influenzando l'andamento della retta interpolante

6.6 Regressione lineare multipla

In questo paragrafo si desidera analizzare quanto la soddisfazione economica e la soddisfazione per la salute influiscono sulla soddisfazione nel tempo libero. La correlazione tra più variabili viene analizzata per mezzo del modello di regressione lineare multivariato. Il quale è descritto dalla retta

$$Y = \alpha = \beta_1 X_1 + \cdots + \beta_p X_p$$

dove α è l'intercetta e i β_i per $i = 1, .., p$ sono i regressori. L'intercetta e i regressori sono calcolati con il metodo dei minimi quadrati.

Il seguente codice calcola il modello lineare multivariato per le variabili indipendenti associate alla soddisfazione economica e alla soddisfazione per la salute e la variabile dipendente associata alla soddisfazione nel tempo libero. I parametri e il coefficiente di determinazione per tale modello sono mostrati nella tabella 6.8

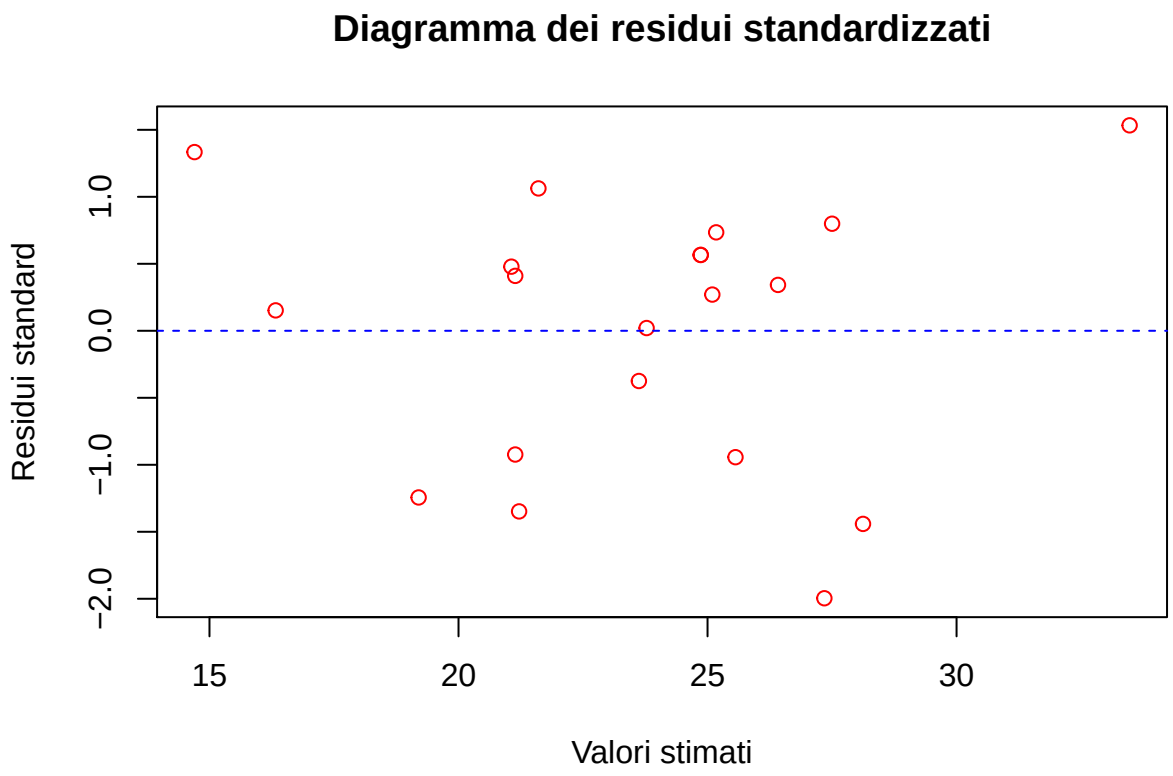

Figura 6.4: Diagramma di dispersione

Tabella 6.8: Dati modello lineare

	1	2	Coef. Determinazione
4.258195	0.8614965	0.3927476	0.870655

```
y <- df$tempoLibero

x1 <- df$economica
x2 <- df$salute
modello <- lm(y~x1+x2)

alfa <- modello$coefficients[[1]]
beta1 <- modello$coefficients[[2]]
beta2 <- modello$coefficients[[3]]
coeffDeterminazione <- summary(modello)$r.squared
```

Poiché entrambi i regressori sono positivi ne consegue che sia la soddisfazione economica sia quella della salute hanno un effetto positivo sulla soddisfazione nel tempo libero.

Inoltre poiché il valore del coefficiente di determinazione assume valore 0.870655 ne consegue che il modello di regressione lineare multipla spiega in modo soddisfacente i valori.

Il seguente codice permette di calcolare i valori stimati, i residui, e i residui standardizzati. I risultati sono mostrati nella tabella 6.9.
Inoltre vengono calcolate alcune statistiche sui residui: la media campionaria, la varianza campionaria e la deviazione standard campionaria. I risultati sono mostrati nella tabella 6.10.

Tabella 6.9: Residui

Osservati	Stimati	Residui	Residui standardizzati
15.4	15.305737	0.0942627	0.0743734
15.0	16.310407	-1.3104068	-1.0339125
16.4	14.766039	1.6339610	1.2891971
16.0	15.189187	0.8108128	0.6397323
23.8	24.834215	-1.0342154	-0.8159972
15.4	15.651610	-0.2516101	-0.1985207
16.6	16.566308	0.0336925	0.0265834
15.5	15.331086	0.1689140	0.1332734
15.5	13.558670	1.9413302	1.5317119
17.3	14.160443	3.1395568	2.4771141
13.9	13.196323	0.7036774	0.5552023
13.0	12.490651	0.5093490	0.4018770
11.9	13.857668	-1.9576676	-1.5446021
12.1	12.162527	-0.0625268	-0.0493337
9.3	11.000758	-1.7007582	-1.3419003
9.3	10.818311	-1.5183105	-1.1979488
10.6	10.905734	-0.3057343	-0.2412247
9.2	9.728765	-0.5287655	-0.4171966
11.6	12.257551	-0.6575507	-0.5188083
10.9	10.608011	0.2919894	0.2303800

Tabella 6.10: Statistiche Residui

Media	Varianza	Deviazione Standard
0	1.606367	1.267425

```
osservati <- y
stime <- fitted(modello)
residui <-resid(modello)
residui.standard <- residui / sd(residui)

mediaResidui <- mean(residui)
varianzaResidui <- var(residui)
deviazioneStandardResidui <- sd(residui)
```

Infine una visualizzazione grafica dei residui rispetto ai valori stimati è mostrata in figura 6.5.

6.7 Regressione polinomiale

Come visto precedentemente il modello lineare non riesce a spiegare la correlazione tra la soddisfazione in famiglia e la soddisfazione economica. Nel seguito verrà utilizzato la regressione polinomiale per modellare la correlazione tra queste due caratteristiche.

Al modello di regressione polinomiale è associata la curva

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

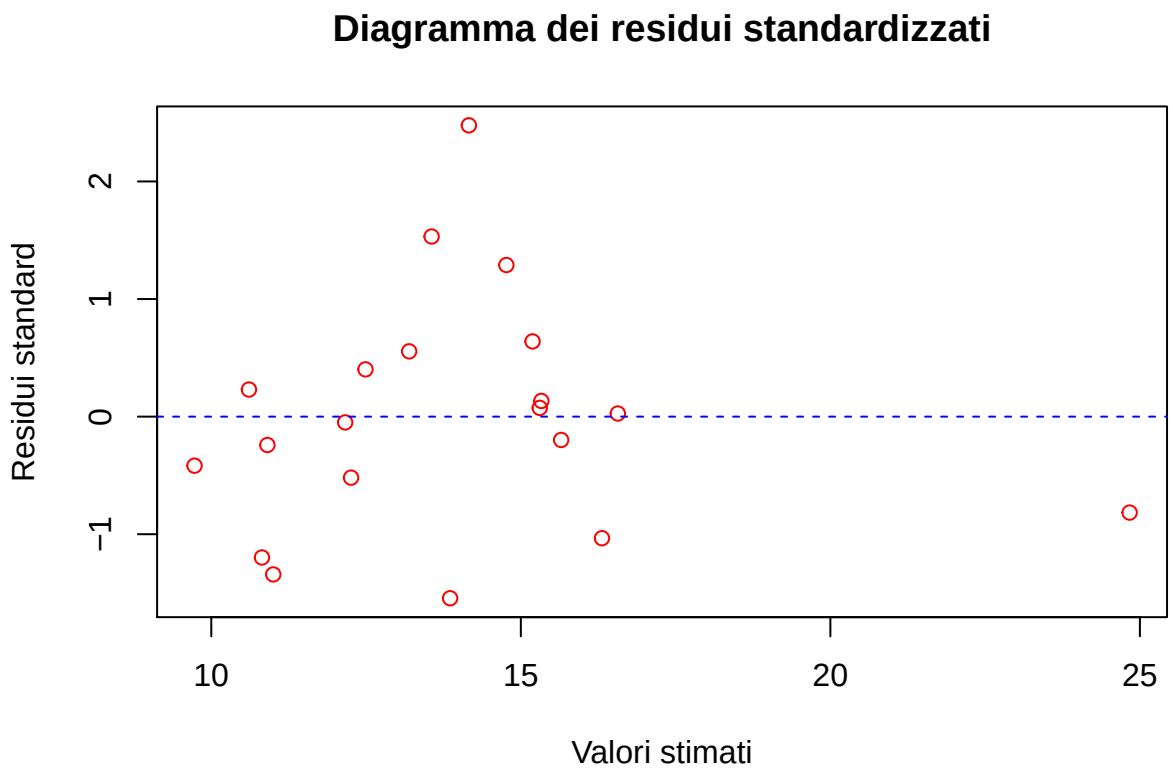

Figura 6.5: Diagramma di dispersione

Tabella 6.11: Dati modello polinomiale

			Coef. Determinazione
16.06647	-1.078837	0.0206401	0.8469753

e per stimare i parametri α, β e γ si utilizza la regressione lineare multipla

$$Y = \alpha + \beta X_1 + \gamma X_2^2$$

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

Il seguente codice calcola il modello polinomiale per la variabile indipendente associata alla soddisfazione in famiglia e la variabile dipendente associata alla soddisfazione economica. I parametri α, β e γ e il coefficiente di determinazione per tale modello sono mostrati nella tabella 6.11

```
y <- df$economica
x <- df$famiglia
modello <- lm(y~x + I(x ^ 2))
alpha <- modello$coefficients[[1]]
beta <- modello$coefficients[[2]]
gamma <- modello$coefficients[[3]]
coeffDeterminazione <- summary(modello)$r.squared
```

Essendo il coefficiente di determinazione pari a 0.8469753 ne consegue che il modello spiega i valori in modo soddisfacente.

Invece il diagramma del modello di regressione è disegnato con il seguente codice ed è mostrato in figura 6.6

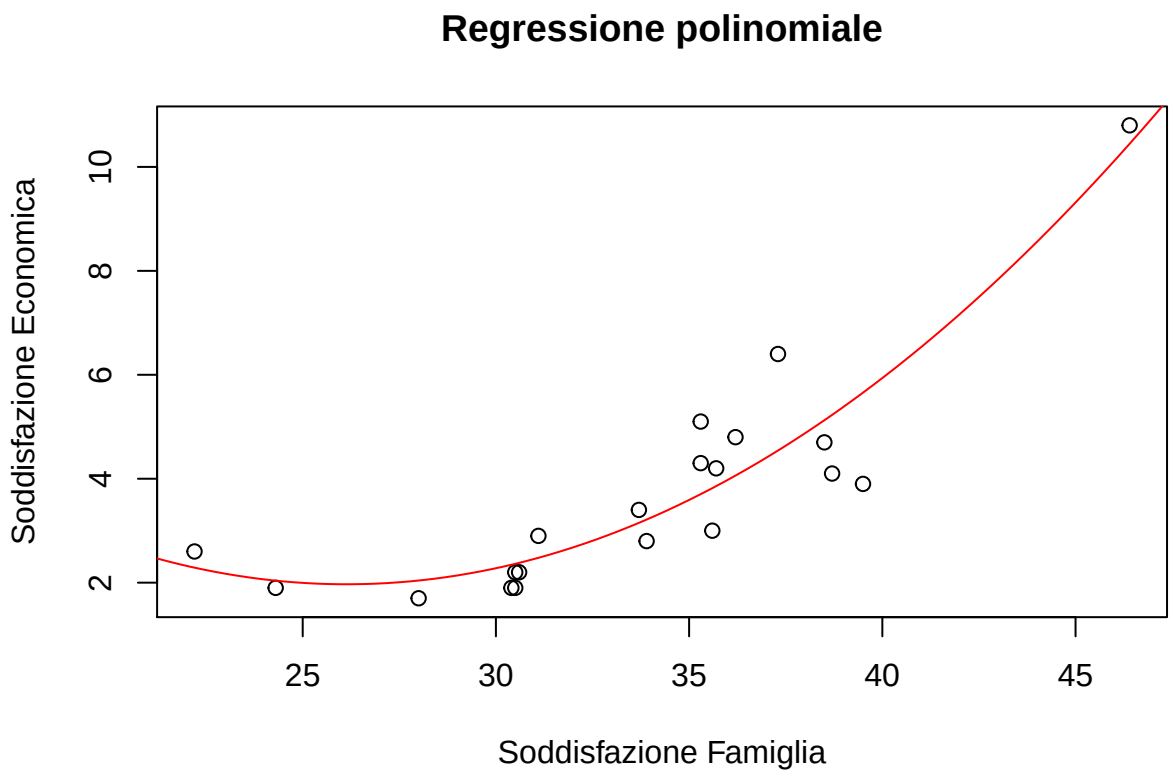

Figura 6.6: Diagramma di dispersione

```
plot(x, y, xlab = "Soddisfazione Famiglia",
     ylab = "Soddisfazione Economica",
     main = "Regressione polinomiale")
curve(alpha+beta*x+gamma*(x^2) ,from=20,to=50, col="red", add = TRUE)
```

Il seguente codice permette di calcolare i valori stimati, i residui, e i residui standardizzati. I risultati sono mostrati nella tabella 6.12.
Inoltre vengono calcolate alcune statistiche sui residui: la media campionaria, la varianza campionaria e deviazione standard campionaria. I risultati sono mostrati nella tabella 6.13.

```
osservati <- y
stime <- fitted(modello)
residui <-resid(modello)
residui.standard <- residui / sd(residui)

mediaResidui <- mean(residui)
varianzaResidui <- var(residui)
deviazioneStandardResidui <- sd(residui)
```

Una visualizzazione grafica dei residui sulla retta di regressione è mostrata in figura 6.7 mentre un diagramma mostrante i valori dei residui rispetto ai valori stimati è mostrato in figura 6.8.

Tabella 6.12: Residui

Osservati	Stimati	Residui	Residui standardizzati
4.8	4.060237	0.7397631	0.8988145
5.1	3.702997	1.3970026	1.6973626
3.9	5.656187	-1.7561871	-2.1337730
4.3	3.702997	0.5970026	0.7253600
10.8	10.445826	0.3541737	0.4303222
4.7	5.125094	-0.4250935	-0.5164900
6.4	4.542271	1.8577288	2.2571464
4.1	5.228010	-1.1280098	-1.3705356
3.0	3.818362	-0.8183620	-0.9943125
4.2	3.857642	0.3423575	0.4159654
3.4	3.150466	0.2495344	0.3031851
2.9	2.477992	0.4220084	0.5127416
2.8	3.213753	-0.4137528	-0.5027110
2.2	2.380662	-0.1806618	-0.2195047
1.9	2.038530	-0.1385305	-0.1683149
2.6	2.288579	0.3114214	0.3783781
2.2	2.362434	-0.1624343	-0.1973582
1.7	2.040907	-0.3409067	-0.4142027
1.9	2.362434	-0.4624343	-0.5618591
1.9	2.344620	-0.4446196	-0.5402142

Tabella 6.13: Statistiche Residui

Media	Varianza	Deviazione Standard
0	0.6773999	0.8230431

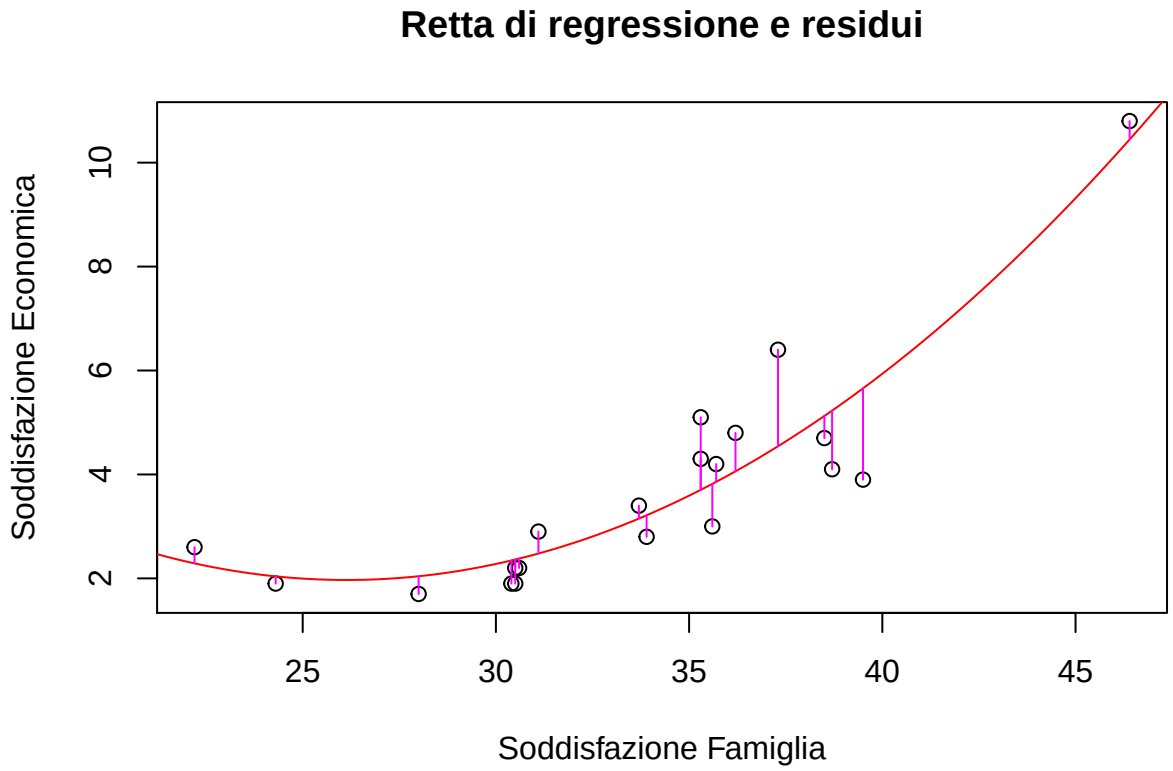

Figura 6.7: Diagramma di dispersione

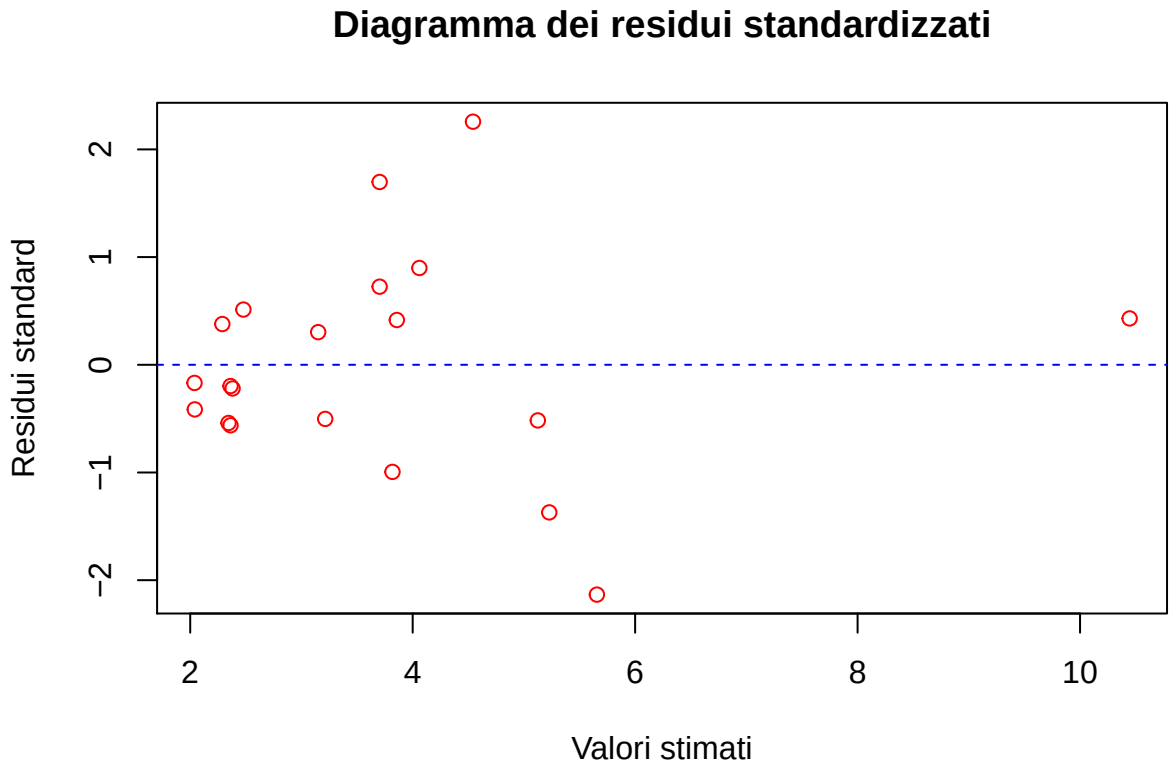

Figura 6.8: Diagramma di dispersione

Capitolo 7

Analisi dei cluster

In questo capitolo verrà effettuata l'analisi dei cluster che ha come obbiettivo il raggruppamento degli individui del campione in gruppi con caratteristiche simili.

Poiché gli individui del nostro campione sono le regioni italiane cercheremo di determinare dei gruppi costituiti da regioni aventi un grado di associazione elevato, e aventi un grado di associazione basso con le regioni presenti negli altri gruppi.

Per determinare quantitativamente la somiglianza tra gli individui è possibile utilizzare misure di distanza o misure di similarità.

Essendo i metodi di enumerazione completa troppo costosi l'analisi dei cluster sarà effettuata prima con metodi gerarchici agglomerativi e poi con un metodo non gerarchico. I metodi gerarchici consentiranno di avere una visione in termini di distanze e consentiranno di scegliere il numero di cluster. Il metodo non gerarchico k-mean cercherà di migliorare i risultati ottenuti precedentemente grazie alla riallocazione degli individui tra i cluster.

Per ogni metodo gerarchico sarà mostrato anche il dendrogramma, un grafico che riporta sull'asse delle ascisse gli individui e sull'asse delle ordinate le distanze di aggregazione.

Con i metodi per cui è possibile verrà effettuata anche un'analisi con lo screeplot. Lo screeplot è un grafico in cui sulle ordinate vengono posti i numeri di gruppi che si possono ottenere con il metodo gerarchico e sull'asse delle ascisse le distanze di aggregazione. Calcolate le differenze tra le distanze di aggregazione per formare k gruppi e $k-1$ gruppi per $k = 2, .., n$ dove n è il numero di individui, la procedura consiglia di scegliere il numero di cluster pari a l se la distanza di aggregazione per formare l gruppi e $l-1$ gruppi è la distanza più grande tra le distanze possibili. Va notato che l'analisi dello screen plot non suggerisce sempre il numero ottimale di cluster e che non è consigliabile usare il metodo del centroide e della mediana in quanto le distanze di agglomerazioni potrebbero essere non crescenti.

7.1 Misure di distanza

Dati due individui I_i e I_j, e definito con X_i il vettore delle caratteristiche dell'individuo I_i appartenente a uno spazio euclideo a p dimensioni E_p, una funzione di distanza $d(X_i, X_j)$ è una funzione a valori reali tale che:

1. $d(X_i, X_j) = 0$ se e sole se $X_i = X_j$ con X_i e X_j in E_p
2. $d(X_i, X_j) \geq 0$ per ogni X_i e X_j in E_p
3. $d(X_i, X_j) = d(X_j, X_i)$ per ogni X_i e X_j in E_p
4. $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ per ogni X_i, X_j e X_k in E_p

Inoltre le funzioni di distanza godono delle seguenti proprietà:

1. Se d e d' sono misure di distanza anche $d + d'$ è una misura di distanza
2. Se d è una misura di distanza e c è un numero reale positivo allora cd è una misura di distanza
3. Se d è una misura di distanza e c è un numero reale positivo allora $\frac{d}{c+d}$ è una misura di distanza

Poiché il prodotto di due misure di distanza (e quindi il quadrato di una misura di distanza) non sempre soddisfa la disuguaglianza triangolare, ne consegue che non può essere considerato una misura di distanza.

Esempi di metriche di distanza sono:

- metrica euclidea
- metrica di Manhattan
- metrica di Checycev
- metrica di Minkowski
- metrica di Camberra
- metrica di Jaccard

Nel seguito utilizzeremo la metrica euclidea definita nel seguente modo

$$d_2(X_i, X_j) = \sqrt{\sum_{k=1}^{p}(x_{i,k} - x_{j,k})^2}$$

dove $x_{i,k}$ è il valore della k-esima caratteristica dell'individuo I_i.

Da notare che la distanza euclidea è influenzata dall'unità di misura delle caratteristiche, il problema non si pone con il nostro data set in quanto i valori sono espressi in forma percentuale. In caso contrario sarebbe stato necessario scalare e standardizzare i valori.

7.2 Misura di non omogeneità

La misura di non omogeneità statistica è una misura che consente di dire quanto è buona la partizione degli individui in cluster.

Matrice di non omogeneità statistica

Dato un insieme di individui distinti $I = \{I_1, \ldots, I_n\}$ ognuno associato a p caratteristiche, possiamo considerare la matrice W_I delle varianze e covarianze tra le caratteristiche dell'insieme in modo tale che $W_{I_{ij}}$ sia la covarianza tra la caratteristica i-esima e j-esima.

Viene definita la matrice statistica di non omogeneità H_I il valore $H_I = (n-1)W_I$ e si definisce misura di non omogeneità statistica il valore

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

Matrice di non omogeneità statistica dell'unione di cluster

La matrice statistica di non omogeneità T dell'unione di m cluster $G_1, \ldots, G_m$ è definita dal valore:

$$T = H_{G_1 \cup \cdots \cup G_m} = S + B$$

dove S è detta matrice di non omogeneità all'interno dei cluster (within) ed è definita dal valore

$$S = \sum_{i=1}^{m} H_i$$

e B è detta matrice di non omogeneità tra i cluster (between) ed è definita dal valore

$$B = \sum_{i<j}^{m} H_{G_i \cap G_j} + \cdots + G_1 \cap \cdots \cap G_m$$

Ne consegue che $trT = trS + trB$ che è equivalente a $1 = \frac{trS}{trT} + \frac{trB}{trT}$

Essendo la traccia della matrice di non omogeneità totale T fissata e poiché le matrici S e B dipendono dalla partizione dei cluster, si desidera, per avere una buona partizione, massimizzare il rapporto $\frac{trB}{trT}$ tra la misura di non omogeneità tra cluster e la misura di non omogeneità totale

Nei seguenti paragrafi verrà utilizzato tale rapporto per analizzare la bontà delle classificazioni generate dai metodi analizzati.

7.3 Funzioni utili per l'analisi dei cluster

In questo paragrafo sono mostrate funzioni e valori che saranno utilizzati nei seguenti paragrafi per l'analisi dei cluster.

Sono definite funzioni e valori per il calcolo della misura di non omogeneità totale e non omogeneità tra cluster, per calcolare le distanze di aggregazione, per calcolare il numero di cluster consigliato dallo screeplot, e per la visualizzazione del dendrogramma. Inoltre è presente anche il codice per il calcolo delle distanze utilizzando la metrica euclidea.

```
nonOmogeneitaTotale <- function(df) {
  n <- nrow(df)
  trHI <- (n - 1) * sum(apply(df, 2, var))
}

nonOmogeneitaCluster <- function(hls, numCluster) {
  individuiInCluster <- cutree(hls, k = as.integer(numCluster), h = NULL)
  indici <- list(individuiInCluster)
  varianze <- aggregate(df, indici, var)
  frequenzeAssolute <- table(individuiInCluster)


  agvar <- varianze[, -1]
  misuraNonOmogCluster <- numeric(0)
  for (i in 1:numCluster) {
    misuraNonOmogCluster[i] <- (frequenzeAssolute[[i]]-1) * sum(agvar[i, ])
    if (is.na(misuraNonOmogCluster[i])) {
      misuraNonOmogCluster[i] = 0
    }
  }
  misuraNonOmogCluster
}

plotCluster <- function(hls, numCluster, tipo) {
   plot(hls,
        hang = -1,
        main = paste("Metodo gerarchico agglomerativo\n", tipo),
        xlab = "",
        sub = "")
  axis(side = 4, at = round(c(0, hls$height),2))
```

```
  rect.hclust(hls, k = numCluster, border = "red")
}

calcoloMisure <- function (df, hls, numCluster) {
  totale <- nonOmogeneitaTotale(df)
  misuraNonOmogCluster <- nonOmogeneitaCluster(hls, numCluster)
  within <- sum(misuraNonOmogCluster)
  between <- totale - within
  rapporto <- between / totale

  list(totale = totale, misuraNonOmogCluster = misuraNonOmogCluster,
       within = within, between = between, rapporto = rapporto)
}

screeplot <- function(hls, tipo) {
  plot(rev(c(0,hls$height)),
     seq(1,rowCount),
     type = "b",
     main = paste("Screeplot - Metodo", tipo),
     xlab = "Distanza di aggregazione",
     ylab = "Numero di cluster",
     col = "red")
}

distancesCluster <-function(hls) {
  c(0, rev(diff(c(0, hls$height))))
}

computeNumCluster <- function(hls) {
  d <- distancesCluster(hls)
  which(d == max(d))
}

d2 <- dist(df, method = "euclidean", diag = TRUE, upper = TRUE)
square.d2 = d2 ^ 2
```

7.4 Metodi gerarchici

I metodi gerarchici agglomerativi partono da n cluster contenenti i singoli individui ed ad ogni passo uniscono i cluster più vicini in modo tale da ottenere in $n-1$ passi un singolo cluster.

L'algoritmo è il seguente:

1. Considerare gli individui come singoli cluster.
2. Calcolare la matrice delle distanza tra i cluster
3. Raggruppare in un unico cluster i due cluster a distanza minima e calcolare la distanza del nuovo cluster da tutti gli altri cluster.
4. Ripartire dal passo 2 finché non rimane un singolo cluster.

I metodi si distinguono per la scelta della misura di distanza e per come si individuano i cluster più vicini.

Nel seguito vengono mostrati i metodi non gerarchici del legame completo, del legame singolo, del legame medio, del centroide e della mediana.

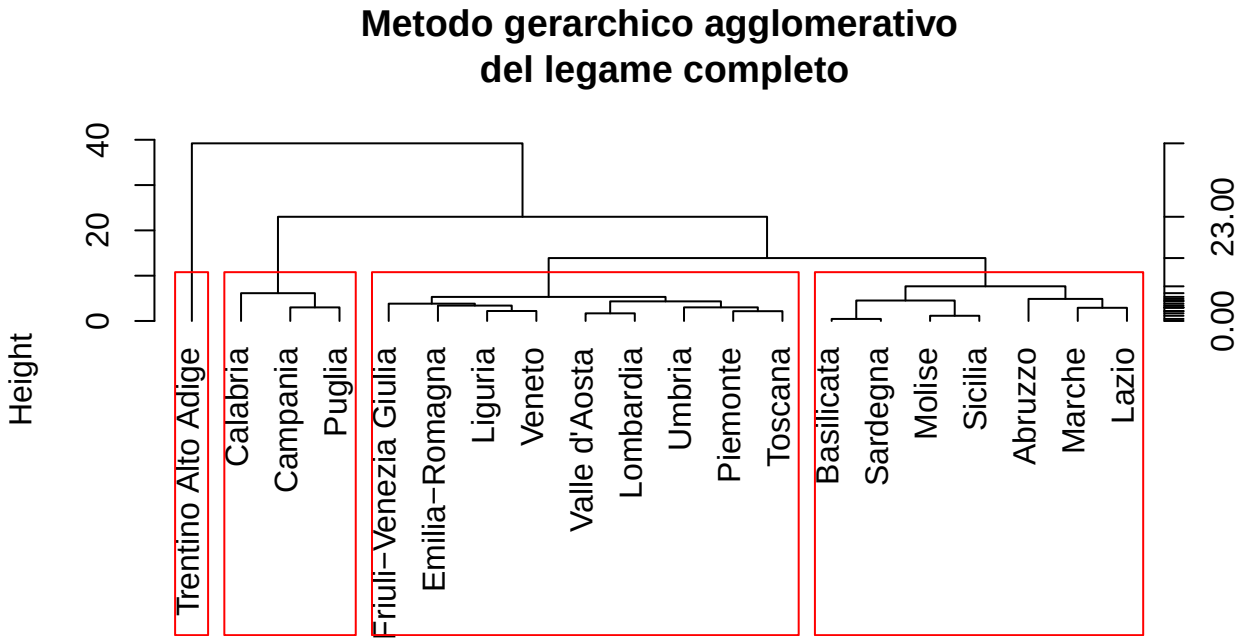

Figura 7.1: Cluster

Tabella 7.1: Misura di non omogeneità

Totale	Within	Between	Between / Totale
1576.9625	145.40380952381	1431.55869047619	0.907795011280351

7.4.1 Metodo del legame completo

Nel metodo del legame completo la distanza tra due gruppi è pari alla massima distanza tra tutte le distanze tra gli individui presi a due a due in cui uno appartiene a un gruppo e uno all'altro gruppo. Tale distanza rappresenta il diametro della sfera contenente tutti gli individui dei due gruppi.

Questo metodo riesce ad identificare soprattutto gruppi di forma ellissoidale, cioè punti addensati intorno a un nucleo. Inoltre privilegia l'omogeneità tra gli individui del gruppo.

Il seguente codice è utilizzato per effettuare l'analisi con il metodo del legame completo.

```
hls <- hclust(d2, method = "complete")
numCluster <- 4
misure <- calcoloMisure(df, hls, numCluster)
```

Nelle figure 7.1 e 7.2, sono mostrati il dendrogramma e lo screeplot. L'analisi dello screeplot suggerisce un numero di cluster pari a 2, tuttavia si è scelto un numero di cluster pari a 4 in quanto la distanza per formare 4 cluster è sensibilmente maggiore rispetto a quelle per formare un numero maggiore di cluster.

Nella tabella 7.1 sono mostrate le misure di non omogeneità per il metodo considerato.

7.4.2 Metodo del legame singolo

Nel metodo del legame singolo la distanza tra due gruppi è pari alla minima distanza tra tutte le distanze tra gli individui presi a due a due in cui uno appartiene a un gruppo e uno all'altro gruppo.

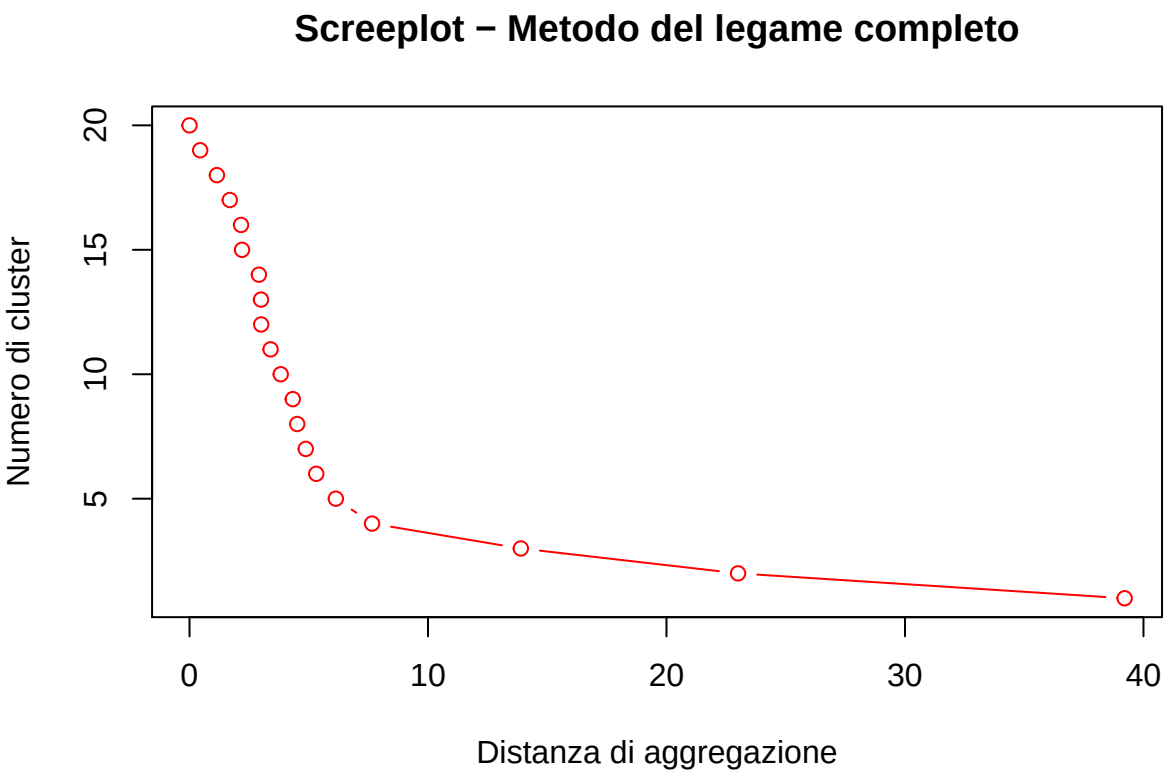

Figura 7.2: Screeplot

Tabella 7.2: Misura di non omogeneità

Totale	Within	Between	Between / Totale
1576.9625	423.31875	1153.64375	0.731560674397774

Questo metodo riesce a individuare cluster di qualsiasi forma e mette in luce valori anomali, di contro può provocare la formazione di catene in quanto basa l'unione tra i cluster su un solo legame. Di conseguenza ci può essere un'incapacità di individuare cluster distinti ma non ben separati.

Il seguente codice è utilizzato per effettuare l'analisi con il metodo del legame singolo.

```
hls <- hclust(d2, method = "single")
numCluster <- 4
misure <- calcoloMisure(df, hls, numCluster)
```

Nelle figure 7.3 e 7.4, sono mostrati il dendrogramma e lo screeplot. L'analisi dello screeplot suggerisce un numero di cluster pari a 2, tuttavia si è scelto un numero di cluster pari a 4 per permettere un confronto con altri metodi.

Nella tabella 7.2 sono mostrate le misure di non omogeneità per il metodo considerato.

7.4.3 Metodo del legame medio

Nel metodo del legame medio la distanza tra due gruppi è pari alla media aritmetica tra tutte le distanze tra gli individui presi a due a due in cui uno appartiene a un gruppo e uno all'altro gruppo.

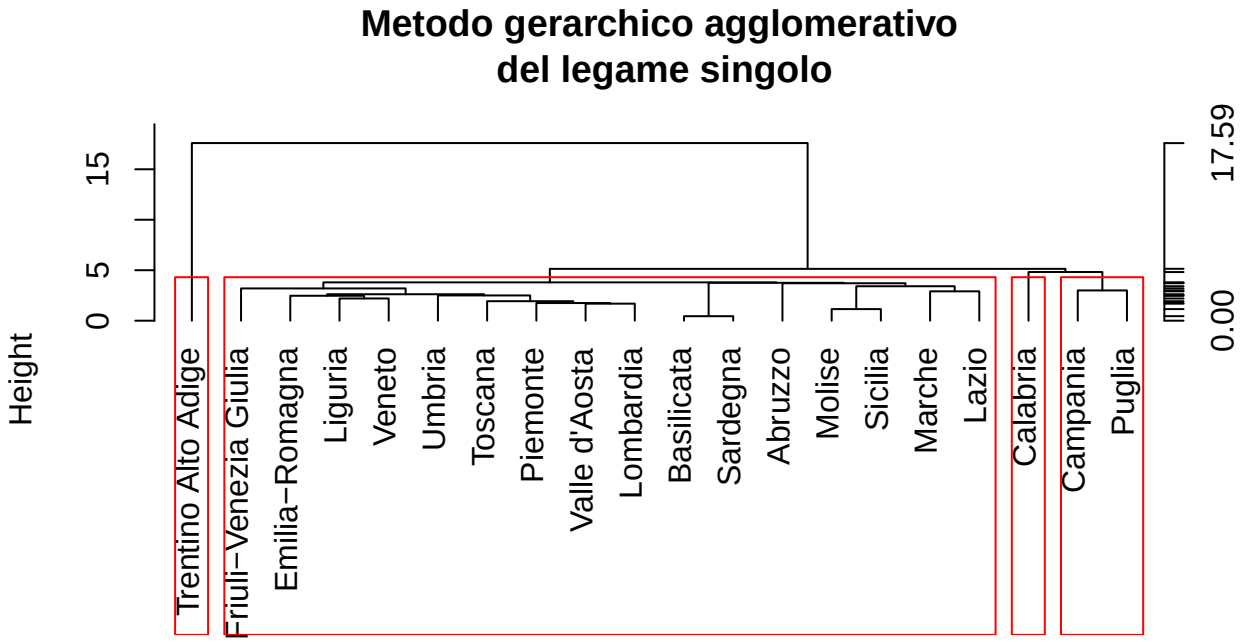

Figura 7.3: Cluster

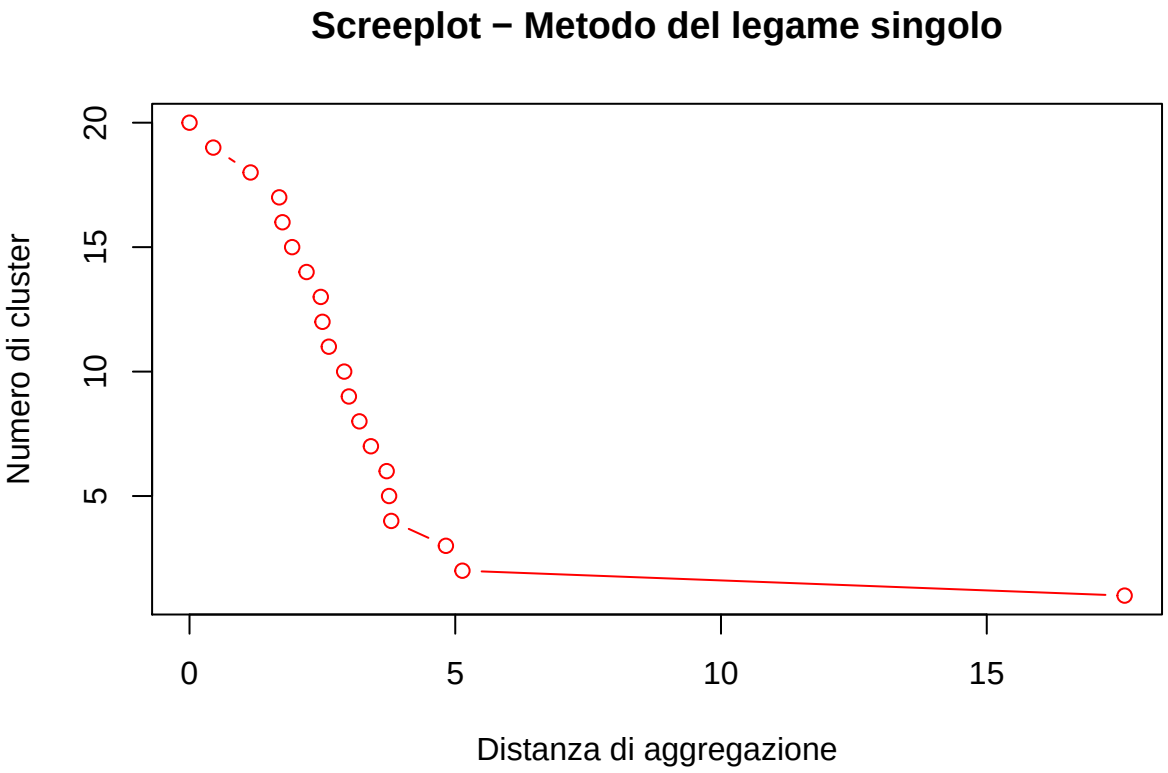

Figura 7.4: Screeplot

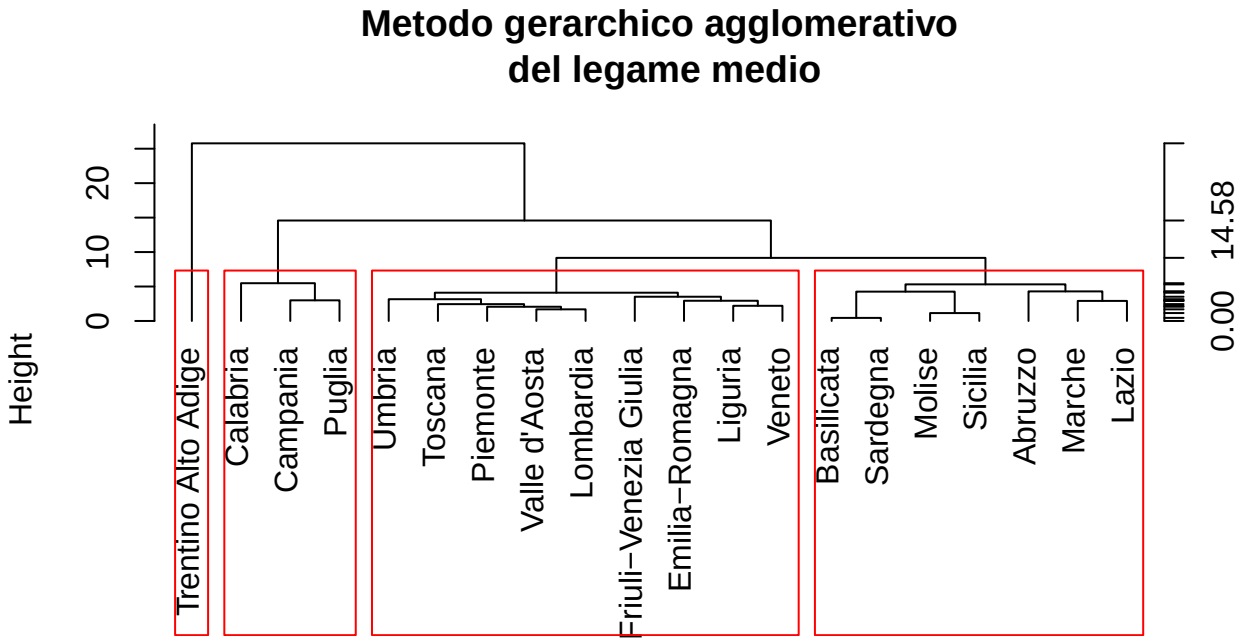

Figura 7.5: Cluster

Tabella 7.3: Misura di non omogeneità

Totale	Within	Between	Between / Totale
1576.9625	145.40380952381	1431.55869047619	0.907795011280351

Va notato che se le dimensioni tra i cluster sono molto differenti la distanza calcolata sarà molto più vicina a quella del cluster più numeroso.

Il seguente codice è utilizzato per effettuare l'analisi con il metodo del legame medio.

```
hls <- hclust(d2, method = "average")
numCluster <- 4
misure <- calcoloMisure(df, hls, numCluster)
```

Nelle figure 7.5 e 7.6, sono mostrati il dendrogramma e lo screeplot. L'analisi dello screeplot suggerisce un numero di cluster pari a 2, tuttavia si è scelto un numero di cluster pari a 4 per permettere un confronto con altri metodi.

Nella tabella 7.3 sono mostrate le misure di non omogeneità per il metodo considerato.

7.4.4 Metodo del centroide

Nel metodo della mediana la distanza tra due gruppi è pari alla distanza tra i centroidi, coincidenti con le medie campionarie tra gli individui dei due gruppi.

Da notare che a differenza dei metodi precedenti il metodo nel centroide utilizza i quadrati delle distanze euclidee. Utilizzando questo metodo gruppi grandi potrebbero attrarre gruppi più piccoli, e se le dimensioni dei cluster sono molto differenti il nuovo centroide sarà più vicino al cluster più numeroso. Infine le distanze di aggregazione potrebbero essere non crescenti.

Il seguente codice è utilizzato per effettuare l'analisi con il metodo del centroide.

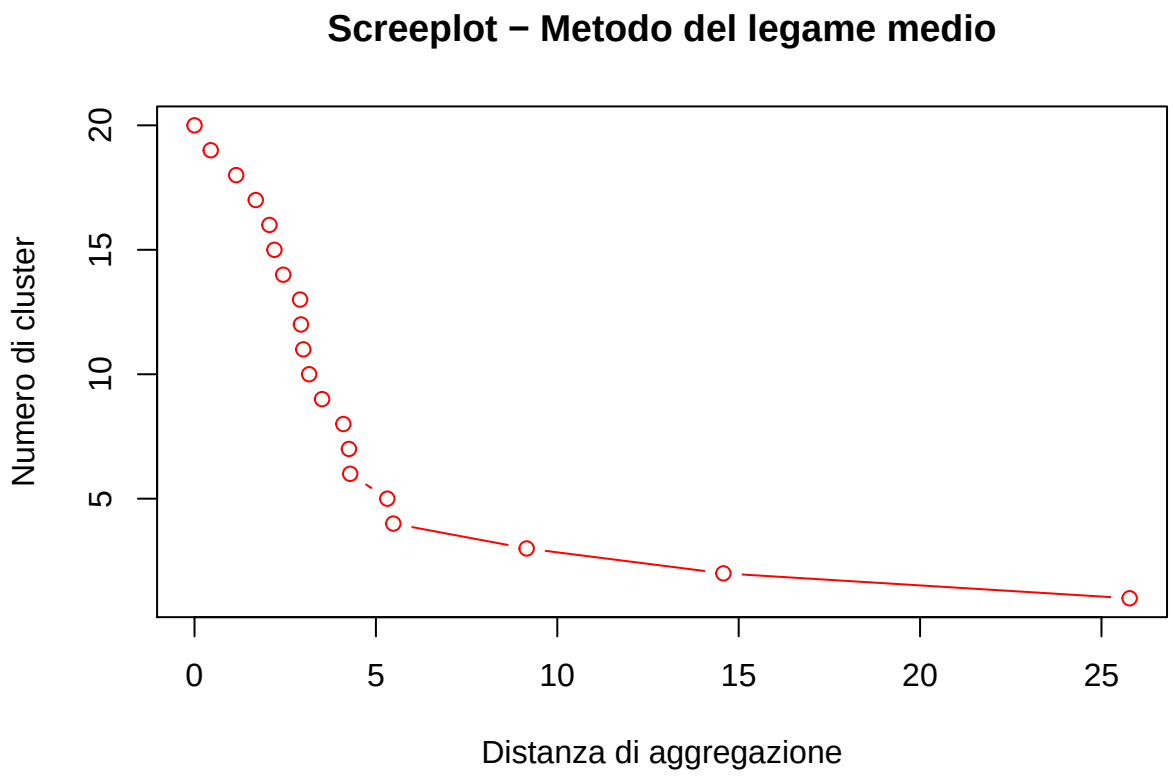

Figura 7.6: Screeplot

Tabella 7.4: Misura di non omogeneità

Totale	Within	Between	Between / Totale
1576.9625	145.40380952381	1431.55869047619	0.907795011280351

```
hls <- hclust(square.d2, method = "centroid")
numCluster <- 4
misure <- calcoloMisure(df, hls, numCluster)
```

Nella figura 7.7 è mostrato il dendrogramma con 4 cluster.

Nella tabella 7.4 sono mostrate le misure di non omogeneità per il metodo considerato.

7.4.5 Metodo della mediana

Nel metodo della mediana la distanza tra due gruppi è pari alla distanza tra i centroidi, coincidenti con la media aritmetica dei centroidi dei due gruppi e di conseguenza è indipendente dalla numerosità dei cluster.

In modo simile al metodo del legame singolo, il metodo nel centroide può provocare una formazione di una catena.

Il seguente codice è utilizzato per effettuare l'analisi con il metodo della mediana.

```
hls <-hclust(square.d2, method = "median")
numCluster <- 4
misure <- calcoloMisure(df, hls, numCluster)
```

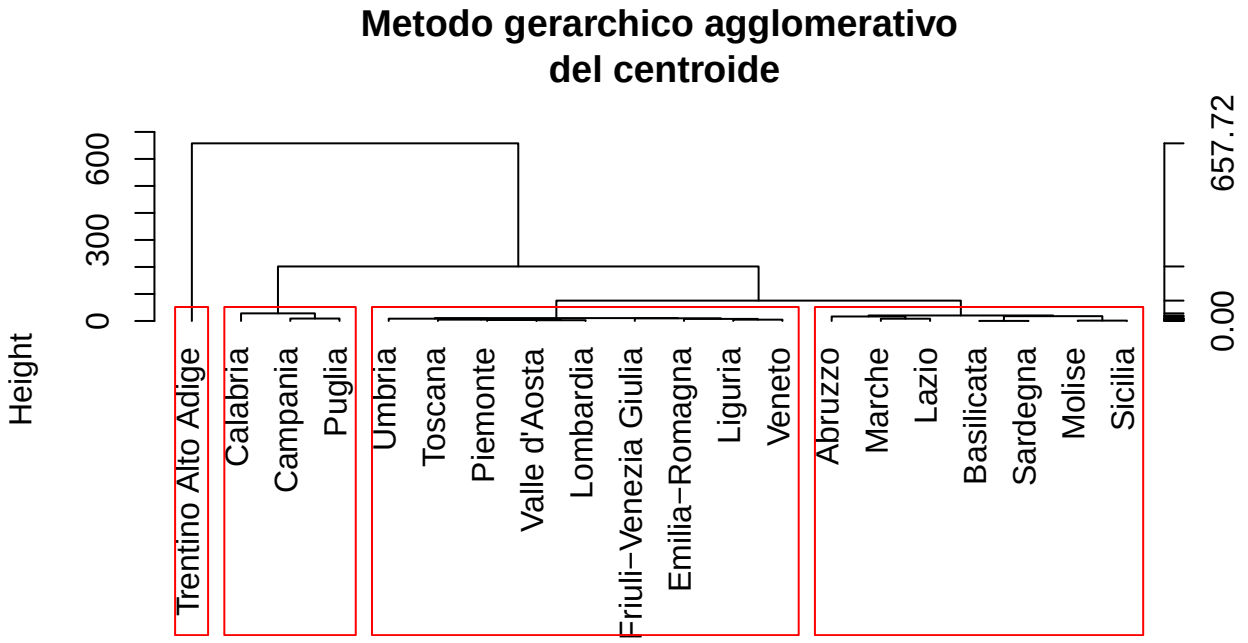

Figura 7.7: Cluster

Tabella 7.5: Misura di non omogeneità

Totale	Within	Between	Between / Totale
1576.9625	145.40380952381	1431.55869047619	0.907795011280351

Nella figura 7.8 è mostrato il dendrogramma con 4 cluster.

Nella tabella 7.5 sono mostrate le misure di non omogeneità per il metodo considerato.

7.4.6 Analisi metodi gerarchici

Nonostante l'analisi degli screen plot abbia suggerito un numero di cluster pari a due, e stato scelto empiricamente un numero di cluster pari a quattro. Infatti una suddivisione in due cluster avrebbe avuto in un cluster l'individuo anomalo "Trentino Alto-Adige" e nel restante gruppo tutte le rimanenti regioni italiane. Con una suddivisione uguale a quattro si ottengono ottimi risultati, infatti per tutti i metodi eccetto il metodo del legame singolo, il rapporto tra la misura di non omogeneità between e quella totale è uguale per tutti i metodi ed ha valore maggiore di 0.9. Con il metodo nel legame singolo si ottiene invece un valore inferiore.

Nel seguito vengono analizzati i cluster con il metodo del legame completo, ottenendo informazioni su alcune misure di sintesi.

Il seguente codice permette di calcolare a quale cluster appartiene ogni individuo, il numero di individui per cluster e le altezze di aggregazione. I risultati ottenuti sono mostrati rispettivamente nelle tabelle: 7.6, 7.7 e 7.8.

```
individuiInCluster <- cutree(hls, k = as.integer(numCluster), h = NULL)
frequenzeAssolute <- table(individuiInCluster)
```

Invece il seguente codice permette di calcolare le medie campionarie, le varianze campionarie e le deviazioni standard campionarie. I risultati sono mostrati rispettivamente nelle tabelle 7.9, 7.10 e 7.11.

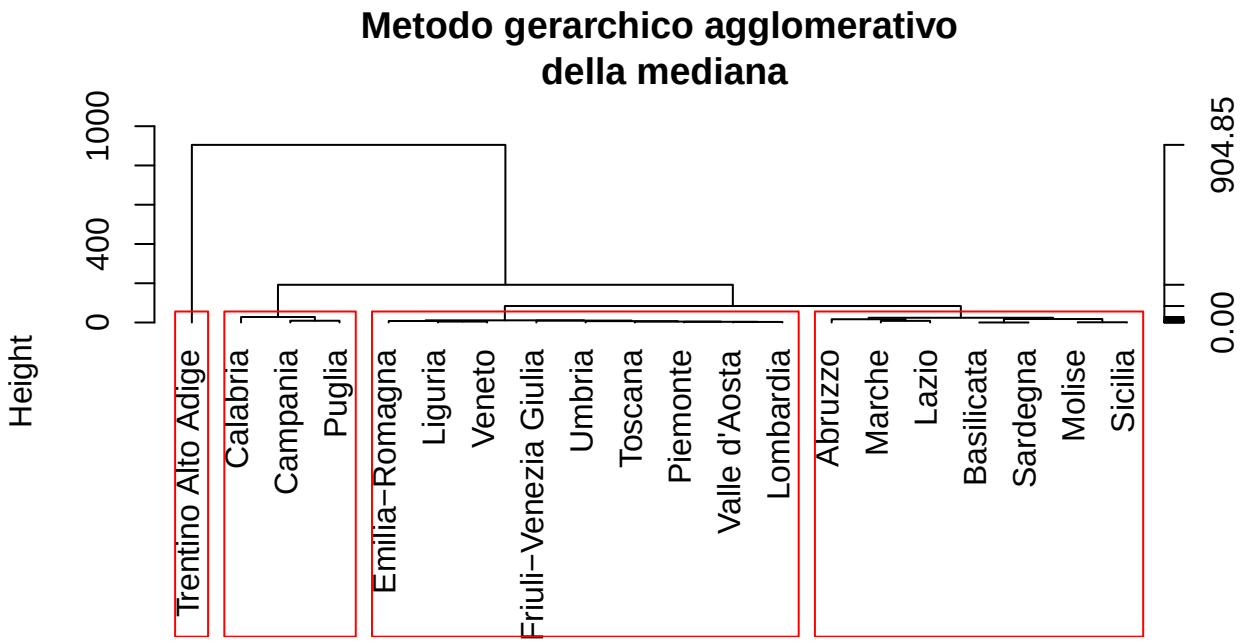

Figura 7.8: Cluster

Tabella 7.6: Individui nei cluster

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

Tabella 7.7: Numero individui nei cluster

Cluster	Numero individui
1	9
2	1
3	7
4	3

Tabella 7.8: Altezze di aggregazione - I numeri negativi indicano gli individui, i numeri positivi i cluster

Primo individuo/cluster	Secondo individuo/cluster	Altezza
-17	-20	0.200000
-14	-19	1.320000
-2	-4	2.850000
-1	3	3.652500
-9	4	4.628125
-3	-6	4.850000
-10	5	6.924531
-8	6	7.612500
-11	-12	8.470000
-15	-16	8.990000
-7	7	9.874883
8	11	11.103252
-13	9	16.637500
1	2	17.800000
13	14	24.216875
-18	10	28.217500
12	15	84.044055
16	17	192.296307
-5	18	904.848374

Tabella 7.9: Medie

Cluster	Economica	Salute	Famiglia	Amici	Tempo libero
1	4.500000	18.00000	36.90000	25.96667	15.900000
2	10.800000	28.70000	46.40000	35.20000	23.800000
3	2.471429	14.82857	31.52857	21.82857	12.000000
4	2.066667	11.40000	24.83333	16.83333	9.266667

Tabella 7.10: Varianze

Cluster	Economica	Salute	Famiglia	Amici	Tempo libero
1	0.8750000	1.230000	2.682500	1.3200000	0.5425000
2	NA	NA	NA	NA	NA
3	0.3257143	4.979048	2.462381	2.3890476	1.3266667
4	0.2233333	2.080000	8.623333	0.7233333	0.0033333

```
hls <- hclust(d2, method = "complete")
individuiInCluster <- cutree(hls, k = as.integer(numCluster), h = NULL)
indici <- list(individuiInCluster)
medie <- aggregate(df, indici, mean)
varianze <- aggregate(df, indici, var)
deviazioniStandard <- aggregate(df, indici, sd)
```

Va notato che la varianza e la variazione standard possono essere calcolate solo se sono presenti almeno due individui. Di conseguenza per il cluster in cui è presente l'unico individuo "Trentino Alto-Adige" tali valori non sono calcolati.

7.5 Metodo non gerarchico

Nei paragrafi precedenti abbiamo utilizzato i metodi non gerarchici per individuare il numero di cluster, in questo paragrafo invece utilizzeremo il metodo non gerarchico k-means per cercare di ottenere un migliore raggruppamento degli individui per il numero di cluster scelto. Questo è possibile in quanto il metodo k-means al contrario dei metodi gerarchici visti precedentemente consente di riallocare gli individui nei cluster.

Il metodo funziona nel seguente modo:

1. Fissare il numero k di cluster, e i punti di riferimento iniziali.
2. Associare ogni individuo al cluster più vicino.
3. Calcolare il baricentro di ogni cluster.
4. Ricalcolare le distanze di ogni individuo da tutti i cluster e riallocare gli individui nel cluster più vicino se necessario.
5. Ricalcolare il baricentro di ogni cluster.

Tabella 7.11: Deviazioni standard

Cluster	Economica	Salute	Famiglia	Amici	Tempo libero
1	0.9354143	1.109054	1.637834	1.1489125	0.736546
2	NA	NA	NA	NA	NA
3	0.5707138	2.231378	1.569198	1.5456544	1.151810
4	0.4725816	1.442220	2.936551	0.8504901	0.057735

Tabella 7.12: Misura di non omogeneità

Totale	Within	Between	Between / Totale
1576.963	145.4038	1431.559	0.907795

6. Ripetere il procedimento dal punto 4 finché i centroidi non subiscono modifiche rispetto all'iterazione precedente.

Va notato che nel metodo k-means si considera la matrice contenente i quadrati delle distanze euclidee, e che la classificazione finale può dipendere dalle scelte iniziali.

I punti di riferimento possono essere scelti in modo casuale o utilizzare i centroidi calcolati con il metodo del centroide, inoltre può essere fissato a priori il numero massimo di iterazioni. Nel seguito il numero massimo di iterazioni sarà posto pari a dieci, un numero maggiore non migliora la classificazione.

Il seguente codice utilizza il metodo k-means scegliendo per dieci volte i punti di riferimento iniziale in modo casuale.

```
numCluster <- 4
km <- kmeans(df, center = numCluster, iter.max = 10, nstart = 10)
```

Invece il seguente codice utilizza come centroidi i centroidi calcolati con il metodo del centroide. Tuttavia con tali punti si ottengono gli stessi risultati.

```
hls <-  hclust(square.d2, method = "centroid")
individuiInCluster <- cutree(hls, k = numCluster, h = NULL)
indici <- list(individuiInCluster)
centroidiIniziali<-aggregate(df, indici, mean)[,-1]

km <- kmeans(df, center = centroidiIniziali, iter.max = 10)
```

Nella tabella 7.12 vengono mostrate le misure di non omogeneità utilizzando il metodo k-means, mentre nella tabella 7.13 viene mostrato il cluster di appartenenza di ogni individuo. Come è possibile osservare si ottengono i medesimi risultati del metodo gerarchico del legame completo.

Infine vengono mostrati nella tabella 7.14 i centroidi dei cluster calcolati.

Tabella 7.13: Individui nei cluster

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

Tabella 7.14: Centroidi

	Economica	Salute	Famiglia	Amici	Tempo libero
Cluster 1	4.500000	18.00000	36.90000	25.96667	15.900000
Cluster 2	10.800000	28.70000	46.40000	35.20000	23.800000
Cluster 3	2.471429	14.82857	31.52857	21.82857	12.000000
Cluster 4	2.066667	11.40000	24.83333	16.83333	9.266667

Capitolo 8

Distribuzione geometrica

La distribuzione geometrica è una distribuzione di probabilità discreta il cui nome deriva dal fatto che segue una progressione geometrica, cioè una successione di numeri tali che è costante il rapporto di un elemento con il precedente.

Considerato un esperimento consistente nella successione di prove indipendenti di Bernoulli di parametro $p \in \{0, 1\}$ e considerato l'evento

$$E_r = \{\text{il primo successo si verifica alla prova r-esima}\}$$

allora la probabilità di tale evento è

$$P(E_r) = p(1-p)^{r-1}$$

La distribuzione di una variabile aleatoria X che descrive il numero di prove per ottenere il primo successo in una successione di prove di Bernoulli è detta distribuzione geometrica di parametro p

$$P(X = r) = P(E_r) \quad \text{per } r = 1, 2...$$

e la sua funzione di probabilità per $0 < p < 1$ è

$$p_X(x) = \begin{cases} p(1-p)^{x-1}, & \text{se } x \geq 1 \\ 0, & \text{altrimenti} \end{cases}$$

8.1 Probabilità teorica e frequenze del campione

Il seguente codice permette di calcolare le probabilità che la variabile aleatoria con distribuzione geometrica di parametro $p = 0.4$ assuma i valori $1, ..., 20$

```
p = 0.4
minX = 0
maxX = 19
x = minX:maxX

probabilita.x = x + 1
probabilita.y = dgeom(x, prob = p)
```

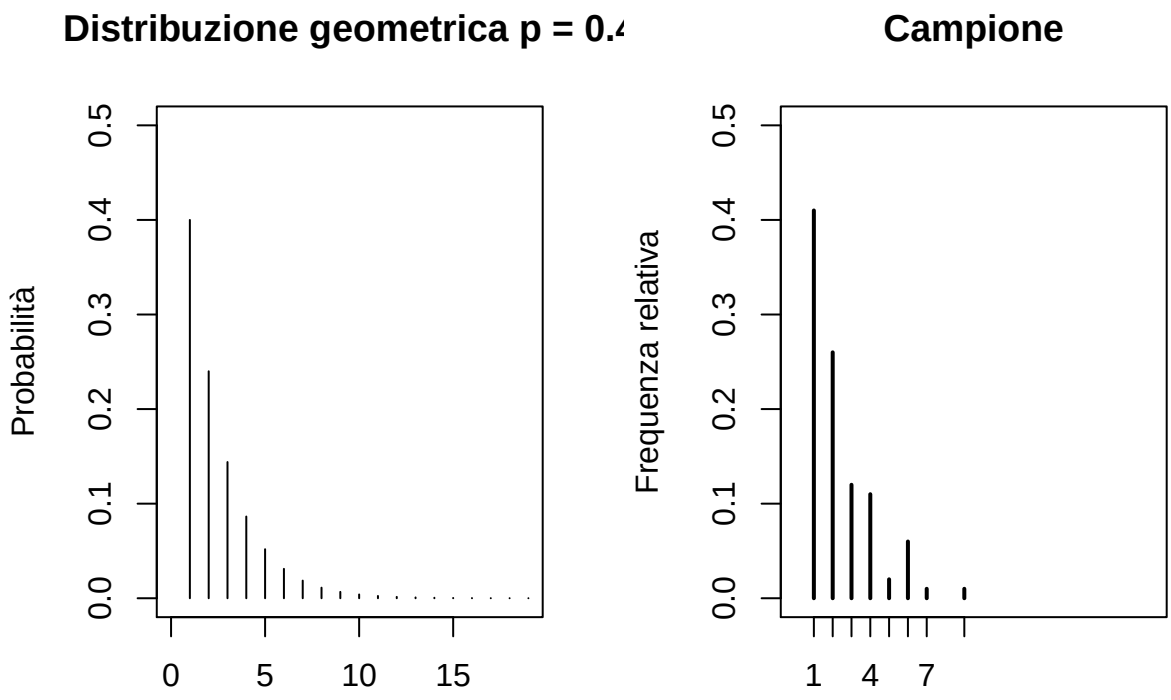

Figura 8.1: Probabilità teoriche e frequenze relative del campione

Invece il seguente codice permette di definire un campione di lunghezza 200 estratto da una popolazione descritta da una variabile geometrica di parametro $p = 0.4$ e di calcolare le frequenze relative dei valori assunti dal campione.

```
p = 0.4
#campione <- rgeom(100, prob = p) + 1
campione <- c(2, 1, 1, 3, 1, 3, 4, 5, 1, 1, 1, 1, 4, 2, 2, 2, 4, 3, 3, 1,
              1, 1, 2, 2, 1, 1, 1, 2, 3, 6, 3, 1, 6, 1, 1, 2, 1, 4, 1, 1,
              4, 3, 2, 1, 4, 6, 7, 2, 1, 1, 3, 1, 1, 4, 1, 1, 4, 1, 4, 1,
              2, 1, 1, 1, 4, 2, 1, 2, 3, 2, 2, 1, 2, 1, 2, 1, 2, 2, 6, 5,
              6, 2, 6, 2, 3, 2, 2, 2, 4, 1, 1, 1, 1, 2, 3, 9, 3, 1, 2, 1)
frequenzeRelative <- table(campione) / length(campione)
```

In figura 8.1 sono mostrate le probabilità della distribuzione geometrica e le frequenze del campione calcolate dal codice precedente. Come è possibile notare, per campioni numerosi, le frequenze relative tendono alla probabilità teoriche.

8.2 Funzione di distribuzione

Calcolata la somma delle prime k probabilità

$$\sum_{r=1}^{k} p_X(r) = \sum_{r=1}^{k} p(1-p)^{r-1} = p\sum_{s=0}^{k-1}(1-p)^s = p\frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k$$

La funzione di distribuzione per una distribuzione geometrica di parametro p è definita con il valore

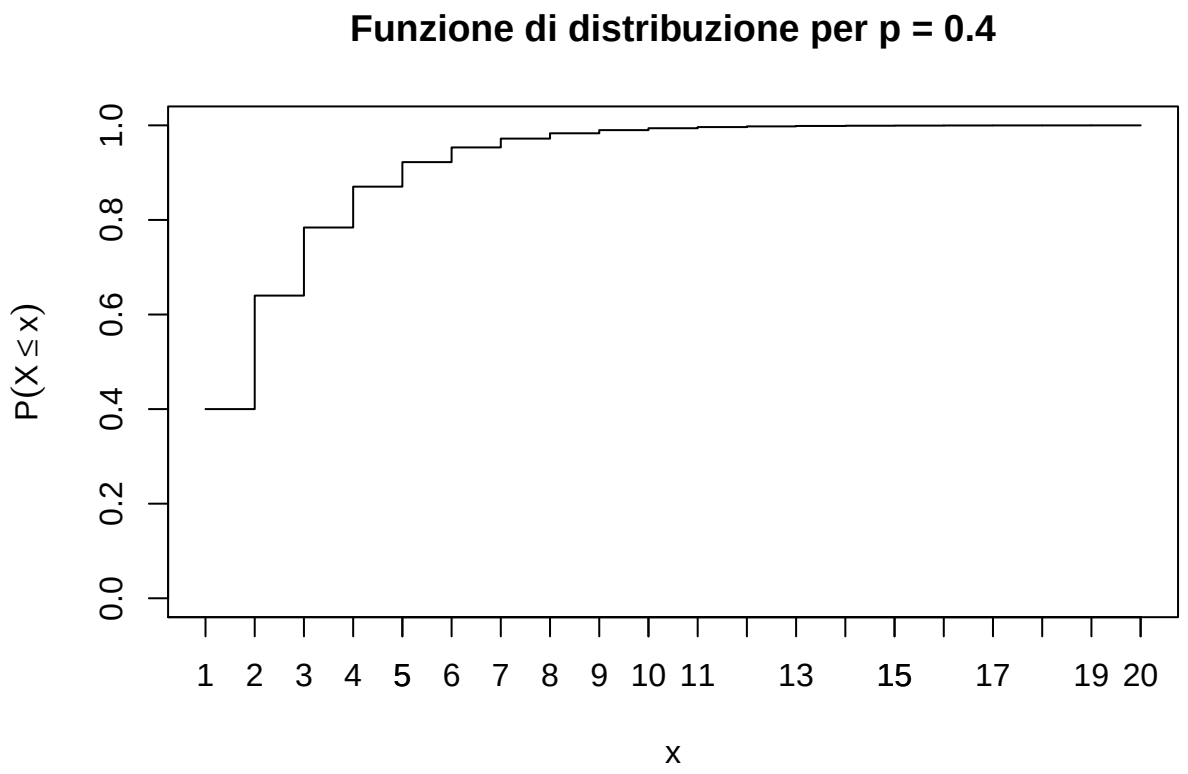

Figura 8.2: Funzione di distribuzione

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{se } x < 1 \\ 1 - (1-p)^k, & \text{se } k \leq x < k+1 \end{cases}$$

Il seguente codice consente di calcolare i valori che assume la funzione di distribuzione di frequenza tra 1 e 20. Il relativo grafico è mostrato in figura 8.2

```
distribuzione <- pgeom(x, prob = p, lower.tail = TRUE)
```

8.3 Quantili

Il percentile $z * 100$-esimo è il più piccolo k tale che

$$P(X \leq x) = 1 - (1-p)^k > z$$

Il valore di k può essere calcolato algebricamente con nel seguente modo

$$1 - (1-p)^k > z \Leftrightarrow (1-p)^k < 1 - z \Leftrightarrow k \geq \frac{log(1-z)}{log(1-p)}$$

Mediana

Ad esempio se vogliamo calcolare la mediana, cioè il percentile $0.5 * 100$-esimo allora bisogna scegliere il più piccolo k tale che $k \geq \frac{log(1-0.5)}{log(1-0.4)}$

Il seguente codice effettua il calcolo

Tabella 8.1: Quartili

Primo quartile	Secondo quartile	Terzo quartile	Quarto quartile
1	2	3	Inf

Tabella 8.2: Statistiche

Media	Media campionaria	Varianza	Varianza campionaria
2.5	2.35	3.75	2.75505

```
mediana <- ceiling(log(1-0.5) / log(1-p))
```

Il risultato ottenuto è 2.

Quartili

Invece il seguente codice calcola i quartili utilizzando la funzione *qgeom*. I risultati ottenuti sono mostrati in tabella 8.1

```
quartili <- qgeom(c(0.25, 0.5, 0.75, 1), prob = p, lower.tail = TRUE) + 1
```

Come è possibile notare il valore della mediana (secondo quartile) assume valore pari a 2 con entrambi i metodi.

8.4 Valore atteso e varianza

Per una variabile aleatoria descritta da una distribuzione geometrica di parametro p il valore atteso è

$$E(X) = \frac{1}{p}$$

mentre la varianza è

$$Var(X) = \frac{1-p}{p^2}$$

Il seguente codice calcola il valore atteso e la varianza della variabile geometrica di parametro p, e la media campionaria e la varianza campionaria del campione simulato. I risultati ottenuti sono mostrati in tabella 8.2

```
valore.atteso <- 1 / p
media.campionaria <- mean(campione)

varianza <- (1 - p) / p^2
varianzaCampionaria <- var(campione)
```

8.5 Assenza di memoria

Una variabile aleatoria X con distribuzione geometrica gode della proprietà di assenza di memoria

$$P(X > r + n \mid X > r) = P(X > n) \quad \text{con r e n interi non negativi}$$

Ciò significa che pur sapendo che nelle prime r prove non c'è stato un successo, la probabilità che che non si verifica un successo fino alla prova $r + n$ dipende soltanto dalle n prove da effettuare e non dalle r prove già effettuate.

Capitolo 9

Stima puntuale

Un problema dell'inferenza statistica è lo studio di una popolazione descritta da una variabile aleatoria osservabile e avente una funzione di distribuzione nota eccetto per uno o più parametri non noti

Uno stimatore è una funzione misurabile e osservabile che associa a un campione un valore per il parametro da stimare. Il valore assunto dallo stimatore è detto stima.

Nel seguito verranno mostrati due metodi di stima dei parametri, il metodo dei momenti e il metodo della massima verosimiglianza.

9.1 Metodo dei momenti

Si definisce momento campionari r-esimo di un campione $(x_1, ..., x_n)$ il valore

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

mentre si definisce momento campionario di una variabile aleatoria il valore

$$M_r(X) = E(X^r)$$

Il metodo dei momenti consiste nell'uguagliare i momenti campionari e i momenti non osservabili della variabile aleatoria caratterizzante la popolazione. Le soluzioni risultanti sono gli stimatori dei parametri non noti.

Formalmente, se si desidera stimare k parametri $\theta_1, ..., \theta_k$ non noti di una distribuzione di probabilità $f_X(x; \theta_1, ..., \theta_k)$ della variabile aleatoria X si calcolano i primi k momenti della variabile aleatoria X

$$\mu_i = E[X^i] = g_i(\theta_1, ..., \theta_k) \quad \text{per } i = 1, ..., k$$

e dato un campione estratto $(x_1, ..., x_n)$ si calcolano i primi k momenti del campione

$$\hat{\mu_i} = M_r(x_1, ..., x_n) \quad \text{per } i = 1, ..., k$$

Lo stimatore del metodo dei momenti per $\theta_1, ..., \theta_k$ denotato da $\hat{\theta}_1, ..., \hat{\theta}_k$ se esiste è la soluzione del sistema di equazioni

$$\hat{\mu_i} = g_i(\hat{\theta}_1, ..., \hat{\theta}_k) \quad \text{per } i = 1, ..., k$$

Essendo gli stimatori dipendenti dal campione ne consegue che al variare dei campioni si potrebbero ottenere stimatori diversi.

9.1.1 Stima per la popolazione geometrica

In questo paragrafo verrà stimato il parametro p di una distribuzione geometrica con il metodo dei momenti.

Il momento campionario primo della variabile geometrica è

$$\mu_1 = E[X^1] = \frac{1}{p}$$

mentre il momento campionario primo del campione è

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

Ne consegue che la media campionaria $\overline{x}$ è uno stimatore per $\frac{1}{p}$

Il seguente codice definisce un campione proveniente da una distribuzione geometrica di parametro p non noto e stima il parametro p con il metodo dei momenti.

```
campione <- c(2, 1, 1, 3, 1, 3, 4, 5, 1, 1, 1, 1, 4, 2, 2, 2, 4, 3, 3, 1,
         1, 1, 2, 2, 1, 1, 1, 2, 3, 6, 3, 1, 6, 1, 1, 2, 1, 4, 1, 1,
         4, 3, 2, 1, 4, 6, 7, 2, 1, 1, 3, 1, 1, 4, 1, 1, 4, 1, 4, 1,
         2, 1, 1, 1, 4, 2, 1, 2, 3, 2, 2, 1, 2, 1, 2, 1, 2, 2, 6, 5,
         6, 2, 6, 2, 3, 2, 2, 2, 4, 1, 1, 1, 1, 2, 3, 9, 3, 1, 2, 1)

stima.p <- 1 / mean(campione)
```

Il valore calcolato pari a 0.4255319 è il valore stimato del parametro p.

9.2 Metodo della massima verosimiglianza

Se $X_1, ..., X_k$ è un campione casuale si definisce funzione di verosimiglianza

$$L(\theta_1, ..., \theta_k) = L(\theta_1, ..., \theta_k; x_1, ..., x_k) = f(x_1; \theta_1, ..., \theta_k) \cdots f(x_n; \theta_1, ..., \theta_k)$$

del campione osservato $(x_1, ..., x_k)$ la funzione di probabilità congiunta se la popolazione è discreta o la funzione di densità di probabilità congiunta se la popolazione è assolutamente continua, del campione casuale $X_1, ..., X_k$.

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri non noti e quindi trovare i parametri della funzione di probabilità o di densità di probabilità per la quale sia più verosimile la provenienza del campione.

I valori $\hat{\theta}_1, ..., \hat{\theta}_k$ che massimizzano la funzione di verosimiglianza sono detti stime di massima verosimiglianza dei parametri non noti $\theta_1, ..., \theta_k$. Essendo le stime dipendenti dal campione ne consegue che al variare dei campioni si possono ottenere stimatori diversi dei parametri non noti.

9.2.1 Stima per la popolazione geometrica

In questo paragrafo verrà stimato il parametro p di una distribuzione geometrica con il metodo della massima verosimiglianza.

Posto $\theta = 1/p$ viene calcolata la funzione di massima verosimiglianza

$$L(\theta) = \left(\frac{1}{\theta}\right)^n \left(\frac{\theta - 1}{\theta}\right)^{\sum_{i=1}^{n} x_i - n} = \left(\frac{1}{\theta}\right)^{\sum_{i=1}^{n} x_i} (\theta - 1)^{\sum_{i=1}^{n} x_i - n}$$

per semplificare i calcoli viene preso il logaritmo della funzione di massima verosimiglianza

$$\log L(\theta) = -\sum_{i=1}^{n} x_i \log(\theta) + \left(\sum_{i=1}^{n} x_i - n\right) \log(\theta - 1)$$

viene poi calcolata la derivata del logaritmo della funzione di massima verosimiglianza

$$\frac{d \log L(\theta)}{d\theta} = -\frac{\sum_{i=1}^{n} x_i}{\theta} + \frac{\sum_{i=1}^{n} x_i - n}{\theta - 1} = \sum_{i=1}^{n} x_i \left(-\frac{1}{\theta} + \frac{1}{\theta - 1}\right) - \frac{n}{\theta - 1} =$$

$$= \frac{1}{\theta(\theta - 1)} \sum_{i=1}^{n} x_i - \frac{n}{\theta - 1} = \frac{\sum_{i=1}^{n} x_i - \theta n}{\theta(\theta - 1)}$$

e infine viene calcolato per quale valore la derivata è uguale a zero

$$\frac{\sum_{i=1}^{n} x_i - \theta n}{\theta(\theta - 1)} = 0 \Leftrightarrow \theta = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ne consegue che $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$ è uno stimatore di $\theta = 1/p$.

Lo stimatore ottenuto con il metodo della massima verosimiglianza è lo stesso di quello ottenuto con il metodo dei momenti, quindi la stima per parametro p utilizzando il nostro campione è 0.4255319.

9.3 Proprietà degli stimatori

Stimatore corretto

Uno stimatore $\hat{\Theta}$ di un parametro non noto θ è detto corretto o non distorto se per ogni $\theta \in \Theta$ il valore medio dello stimatore $\hat{\Theta}$ è uguale al parametro non noto.

$$E[\hat{\Theta}] = \theta \quad \forall\ \theta \in \Theta$$

Da notare che possono esistere diversi stimatori corretti del parametro non noto.

Errore quadratico medio

L'errore quadratico medio fornisce una misura di quanto lo stimatore si discosta dal parametro non noto ed è definito con il valore

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2]$$

Se $\hat{\Theta}$ è uno stimatore corretto del parametro θ allora

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - E[\hat{\Theta}])^2] = Var(\hat{\Theta})$$

Stimatore corretto con varianza uniformemente minima

Uno stimatore è corretto con varianza uniformemente minima se per ogni $\theta \in \Theta$:

- $E[\hat{\Theta}] = \theta$
- $Var(\hat{\Theta}) \leq Var(\hat{\Theta}^*)$ per ogni stimatore $\hat{\Theta}^*$ corretto del parametro θ

Disuguaglianza di Cramer

Se $\hat{\theta}$ è uno stimatore corretto del parametro non noto θ di una popolazione caratterizzata da una funzione di probabilità (nel caso discreto) o densità di probabilità (nel caso assolutamente continuo) $f(x; \theta)$ e se:

- $\frac{\partial}{\partial\theta}\log f(x;\theta)$esiste per ogni x e per ogni $\theta \in \Theta$
- $E\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right]$ esiste finito per ogni $\theta \in \Theta$

Allora

$$Var(\hat{\Theta}) \geq \frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right]}$$

Di conseguenza se la varianza dello stimatore $\hat{\Theta}$ è

$$Var(\hat{\Theta}) = \frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right]}$$

allora lo stimatore $\hat{\Theta}$ è corretto con varianza uniformemente minima.

Stimatore consistente

Uno stimatore $\hat{\theta}$ è detto consistente se converge in probabilità a θ

$$\lim_{n\to+\infty} P(|\hat{\Theta} - \theta| < \epsilon) = 1 \quad \forall\theta \in \Theta$$

Una condizione sufficiente alla consistenza di uno stimatore $\hat{\theta}$ è il verificarsi delle seguenti uguaglianze:

$$\lim_{n\to+\infty} E(\hat{\Theta}) = \theta \quad \forall\theta \in \Theta$$

$$\lim_{n\to+\infty} Var(\hat{\Theta}) = 0 \quad \forall\theta \in \Theta$$

9.3.1 Analisi dello stimatore per la distribuzione geometrica

In questo paragrafo verrà mostrato come la varianza campionaria è uno stimatore corretto con varianza minima per il parametro p di una distribuzione geometrica.

Dato che $E(X) = 1/p$ allora il parametro da stimare è $\theta = 1/p$

Inoltre vale la seguente uguaglianza $Var(X) = \frac{1-p}{p^2} = \theta^2 - \theta$

Stimatore corretto

Per dimostrare che lo stimatore θ è corretto dobbiamo dimostrare che

$$E[\hat{\Theta}] = \theta \quad \forall\ \theta \in \Theta$$

Poiché

$$E[\overline{X}] = E[X]$$

ne consegue che

$$E[\hat{\Theta}] = E[\overline{X}] = E[X] = \theta$$

Di conseguenza Θ è uno stimatore corretto.

Stimatore corretto con varianza uniformemente minima

Per dimostrare che la stimatore è corretto con varianza uniformemente minima utilizziamo la disuguaglianza di Cramer.

Poiché la seguente derivata esiste ed è finita

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{\partial}{\partial\theta}\log(p(1-p)^{x-1}) = \frac{\partial}{\partial\theta}\log\left(\theta^{-1}(1-\theta^{-1})^{x-1}\right) = \frac{x-\theta}{\theta^2-\theta}$$

è possibile calcolare il seguente valore

$$\frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right]} = \frac{1}{n\frac{E[(X-\theta)^2]}{(\theta^2-\theta)^2}} = \frac{(\theta^2-\theta)^2}{nE[(X-\theta)^2]} = \frac{(\theta^2-\theta)^2}{nVar(X)} = \frac{\theta^2-\theta}{n}$$

che risulta essere uguale al valore della varianza dello stimatore

$$Var(\overline{X}) = \frac{Var(X)}{n} = \frac{(\theta^2-\theta)}{n}$$

Ne consegue che $\overline{X}$ è uno stimatore con varianza uniformemente minima.

Stimatore consistente

Inoltre essendo verificati i due limiti:

- $\lim_{n\to+\infty} E[\overline{X_n}] = \theta$
- $\lim_{n\to+\infty} Var(\overline{X_n}) = \lim_{n\to+\infty} \frac{1-p}{np^2} = 0$

Ne consegue che $\overline{X}$ è uno stimatore consistente.

Capitolo 10

Intervalli di fiducia approssimati

Intervalli di fiducia

Sia $X_1, .., X_n$ un campione di ampiezza n estratto da una popolazione con funzione di probabilità (nel caso discreto) o funzione di densità di probabilità (nel caso continuo) $F(x;\theta)$ dove θ è il parametro non noto della popolazione.

Se per $0 < \alpha < 1$ esistono due statistiche $\underline{C_n} = g_1(X_1, .., X_n)$ e $\overline{C_n} = g_2(X_1, .., X_n)$ tali che

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

allora $(\underline{C_n}, \overline{C_n})$ è un intervallo di confidenza di grado $1 - \alpha$ per il parametro non noto θ.

Mentre l'intervallo $(g_1(X_1, .., X_n), g_2(X_1, .., X_n))$ è detto stima dell'intervallo di confidenza di grado $1 - \alpha$ per il parametro non noto θ.

Da notare che possono esistere più intervalli di confidenza dello stesso grado per un parametro non noto di una popolazione.

Metodo pivotale

Il metodo pivotale è un metodo per la ricerca di intervalli di confidenza che consiste nel cercare una variabile pivot $\gamma(X_1, .., X_n; \theta)$ dipendente dal campione e dal parametro non noto θ. La variabile non è osservabile perché dipende dal parametro non noto, ne consegue che la variabile pivot non è una statistica.

Per un fissato α con $0 < \alpha < 1$ se esistono α_1 e α_2 dipendenti soltanto dal campione con $\alpha_1 < \alpha_2$ tali che per ogni $\theta \in \Theta$

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

e se per ogni campione osservato $(x_1, .., x_n)$ e per ogni $\theta \in \Theta$ si può dimostrare che

$$\alpha_1 < \gamma(x_1, .., x_n; \theta) < \alpha_2 \Leftrightarrow g_1(x_1, .., x_n) < \theta < g_2(x_1, .., x_n)$$

dove $g_1(x_1, .., x_n)$ e $g_2(x_1, .., x_n)$ dipendono soltanto dal campione osservato, allora la probabilità

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

è equivalente a

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

Se $\underline{C_n} = g_1(X_1,..,X_n)$ e $\overline{C_n} = g_2(X_1,..,X_n)$

allora $(\underline{C_n}, \overline{C_n})$ è un intervallo di confidenza di grado $1-\alpha$ per il parametro non noto θ.

Teorema centrale di convergenza

Siano $X_1, ...X_n$ n variabili aleatorie indipendenti e identicamente distribuite tali che

$$E(X_i) = \mu \quad e \quad Var(X_i) = \sigma^2 \quad per \quad i = 1,..n$$

e posto $Y_n = \sum_{i=1}^{n} X_i$ tale che

$$E(Y_n) = n\mu \quad e \quad Var(Y_n) = \sigma^2 n$$

$$\lim_{n\to+\infty} P\left(\frac{Y_n - E(Y_n)}{\sqrt{Var(Y_n)}}\right) \leq x = \Phi(x) \quad \forall n \in \mathbf{Z},\ n > 0\ \forall x \in \mathbf{R}$$

dove $\Phi(x)$ è la funzione di distribuzione della normale standardizzata.

Di conseguenza per n grande ($n > 30$) la variabile Y_n standardizzata converge in distribuzione alla normale standard.

Intervalli di fiducia approssimati

Se X è la variabile aleatoria che descrive la popolazione, se $E(X) = \mu$ e $Var(X) = \sigma^2$ e se $X_1, ..., X_n$ è un campione casuale estratto dalla popolazione, allora per il teorema centrale di convergenza

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{\sigma^2 n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

converge in distribuzione alla variabile aleatoria normale standard $Z \sim \mathcal{N}(0,1)$.

Di conseguenza se il campione è numeroso è possibile utilizzare il metodo pivotale in forma approssimata per determinare un intervallo di confidenza approssimato

$$P\left(-z_{\alpha/2} < Y_n < z_{\alpha/2}\right) \backsimeq 1-\alpha$$

dove $-z_{\alpha/2}$ e $z_{\alpha/2}$ sono tali che

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

In figura 10.1 è rappresentata una densità normale standard e i valori $-z_{\alpha/2}$ e $z_{\alpha/2}$

10.1 Intervalli di fiducia approssimati per la popolazione geometrica

Nel caso della popolazione geometrica il valore medio e la varianza della popolazione, e il valore medio e la varianza della media campionaria hanno valori

$$E(X) = 1/p \quad Var(X) = (1-p)/p^2$$

$$E(\overline{X_n}) = 1/p \quad Var(\overline{X}_n) = (1-p)/(np^2)$$

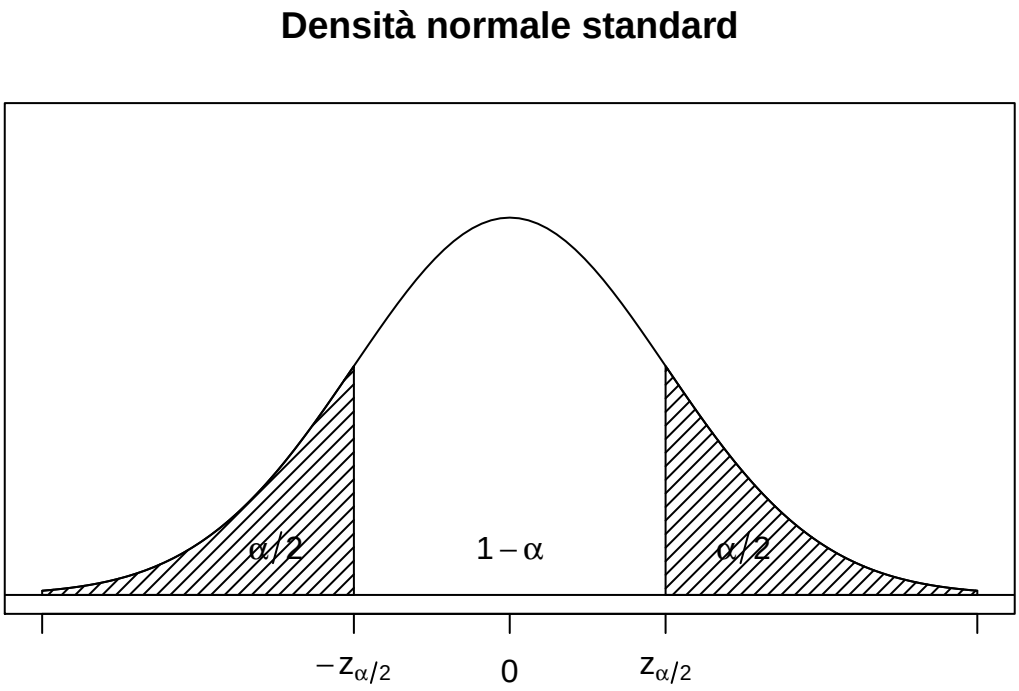

Figura 10.1: Densità normale standard

per il teorema centrale di convergenza la variabile aleatoria

$$Z_n = \frac{\overline{X} - E(\overline{X_n})}{Var(\overline{X}_n)} = \frac{\overline{X} - 1/p}{\sqrt{(1-p)/(np^2)}} = \frac{p\overline{X} - 1}{p} \frac{p\sqrt{n}}{\sqrt{1-p}} = \sqrt{n}\frac{p\overline{X} - 1}{\sqrt{1-p}}$$

converge in distribuzione a una normale standard.

La disuguaglianza

$$-z_{\alpha/2} < \sqrt{n}\frac{p\overline{x} - 1}{\sqrt{1-p}} < z_{\alpha/2}$$

è equivalente alla disequazione di secondo grado

$$\left[\sqrt{n}\frac{p\overline{x} - 1}{\sqrt{1-p}}\right]^2 < z_{\alpha/2}$$

espressa nel seguito in forma canonica

$$n\overline{x}^2p^2 + p(z^2_{\alpha/2} - 2n\overline{x_n}) + n - z^2_{\alpha/2} < 0$$

La soluzioni della disequazione sono interne all'intervallo formato dalle radici del polinomio. Tali radici possono essere calcolate esplicitamente o mediante la funzione *polyroot*.

La seguente funzione consente di calcolare l'intervallo di confidenza per il parametro p di una distribuzione geometrica dati in input il grado $1 - \alpha$, la lunghezza del campione e la media campionaria.

```
intervalloDiConfidenza <- function(grado, lunghezzaCampione, mediaCampionaria) {
  alpha <- 1 - grado
  zalpha <- qnorm(1 - alpha / 2, mean = 0, sd = 1)
  a2 <- lunghezzaCampione * mediaCampionaria^2
  a1 <- zalpha^2 - 2 * lunghezzaCampione * mediaCampionaria
```

```
  a0 <- lunghezzaCampione - zalpha^2
  radici <- polyroot(c(a0,a1,a2))
  radici
}
```

Il seguente codice definisce una variabile contenente i dati del campione e calcola l'intervallo di confidenza di grado $1 - \alpha = 0.95$ per p.

```
campione <- c(2, 1, 1, 3, 1, 3, 4, 5, 1, 1, 1, 1, 4, 2, 2, 2, 4, 3, 3, 1,
         1, 1, 2, 2, 1, 1, 1, 2, 3, 6, 3, 1, 6, 1, 1, 2, 1, 4, 1, 1,
         4, 3, 2, 1, 4, 6, 7, 2, 1, 1, 3, 1, 1, 4, 1, 1, 4, 1, 4, 1,
         2, 1, 1, 1, 4, 2, 1, 2, 3, 2, 2, 1, 2, 1, 2, 1, 2, 2, 6, 5,
         6, 2, 6, 2, 3, 2, 2, 2, 4, 1, 1, 1, 1, 2, 3, 9, 3, 1, 2, 1)

intervallo <- intervalloDiConfidenza(grado = 0.95,
                                     lunghezzaCampione = length(campione),
                                     mediaCampionaria = mean(campione))
```

L'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro p calcolato è

(0.358744314001264-0i, 0.48536350162446+0i)

Da notare che la stima puntuale di valore 0.4255319 risulta interna all'intervallo.

10.2 Differenza tra valori medi

Siano $X_1, .., X_n$ e $Y_1, .., Y_m$ due campioni casuali di ampiezza n_1 e n_2 estratti da due popolazioni geometriche con parametri p_1 e p_2 rispettivamente. Vogliamo determinare un intervallo di confidenza di grado $1 - \alpha$ per la differenza $p_1 - p_2$.

Essendo

$$E(\overline{X_{n_1}} - \overline{Y_{n_2}}) = \frac{1}{p_1} - \frac{1}{p_2}$$

$$Var(\overline{X_{n_1}} - \overline{Y_{n_2}}) = \frac{1 - p_1}{n_1 p_1^2} + \frac{1 - p_2}{n_2 p_2^2}$$

allora per il teorema centrale di convergenza la variabile aleatoria

$$\frac{\overline{X_{n_1}} - \overline{Y_{n_2}} - \left(\frac{1}{p_1} - \frac{1}{p_2}\right)}{\sqrt{\frac{1-p_1}{n_1 p_1^2} + \frac{1-p_2}{n_2 p_2^2}}}$$

converge in distribuzione alla variabile aleatoria normale standard $Z \sim \mathcal{N}(0, 1)$.

Inoltre essendo

$$lim_{n \to +\infty} E[\overline{X_{n_1}}(\overline{X_{n_1}} - 1)] = \frac{1 - p_1}{p_1^2}$$

$$lim_{n \to +\infty} E[\overline{Y_{n_2}}(\overline{Y_{n_2}} - 1)] = \frac{1 - p_2}{p_2^2}$$

Per campioni sufficientemente grandi

$$P\left(-z_{\alpha/2} < \frac{\overline{X_{n_1}} - \overline{Y_{n_2}} - \left(\frac{1}{p_1} - \frac{1}{p_2}\right)}{\sqrt{\frac{\overline{X_{n_1}}(\overline{X_{n_1}}-1)}{n_1} + \frac{\overline{Y_{n_2}}(\overline{Y_{n_2}}-1)}{n_1}}} < z_{\alpha/2}\right) \backsimeq 1-\alpha$$

La disuguaglianza

$$-z_{\alpha/2} < \frac{\overline{x_{n_1}} - \overline{y_{n_2}} - \left(\frac{1}{p_1} - \frac{1}{p_2}\right)}{\sqrt{\frac{\overline{x_{n_1}}(\overline{x_{n_1}}-1)}{n_1} + \frac{\overline{y_{n_2}}(\overline{y_{n_2}}-1)}{n_1}}} < z_{\alpha/2}$$

è equivalente a

$$\overline{x_{n_1}} - \overline{y_{n_2}} - z_{\alpha/2}\sqrt{\frac{\overline{x_{n_1}}(\overline{x_{n_1}}-1)}{n_1} + \frac{\overline{y_{n_2}}(\overline{y_{n_2}}-1)}{n_1}} < \left(\frac{1}{p_1} - \frac{1}{p_2}\right)$$

$$\left(\frac{1}{p_1} - \frac{1}{p_2}\right) < \overline{x_{n_1}} - \overline{y_{n_2}} + z_{\alpha/2}\sqrt{\frac{\overline{x_{n_1}}(\overline{x_{n_1}}-1)}{n_1} + \frac{\overline{y_{n_2}}(\overline{y_{n_2}}-1)}{n_1}}$$

La seguente funzione consente di calcolare l'intervallo di confidenza per la differenza dei valori medi dati in input il grado $1-\alpha$, la lunghezza del primo campione e del secondo campione e la media campionaria del primo e del secondo campione.

```
differenzeValoriMedi <- function(grado, n1, n2, m1, m2) {
  alpha <- 1 - grado
  delta <- m1 * (m1 - 1) / n1 + m2 * (m2 - 1) / n2
  radice <- sqrt(delta)
  a <- m1 - m2
  b <- qnorm(1 - alpha / 2, mean = 0, sd = 1) * radice
  left <-  a - b
  right <- a + b
  return(c(left, right))
}
```

Interpretazione intervallo

- Se gli estremi dell'intervallo sono entrambi positivi allora il valore medio della prima popolazione è maggiore del valore medio della seconda popolazione.
- Se gli estremi dell'intervallo sono entrambi negativi allora il valore medio della prima popolazione è minore del valore medio della seconda popolazione.
- Altrimenti esiste la possibilità che i valori medi siano uguali, quindi non è possibile dire quale sia maggiore.

Il seguente codice definisce due campione e calcola l'intervallo di confidenza di grado $1-\alpha = 0.99$ per le differenze dei valori medi.

```
#campione1 <- rgeom(150, prob = 0.3) + 1
campione1 <- c(2, 4, 7, 5, 2, 2, 1, 2, 6, 5, 3, 2, 6, 1, 4, 1, 5, 1, 1, 1,
               2, 1, 5, 4, 9, 6, 3, 5, 3, 4, 3, 1, 6, 1, 3, 1, 1, 4, 3, 1,
               3, 1, 1, 1, 1, 4, 2, 3, 6, 1, 3, 1, 3, 7, 12, 5, 5, 3, 1, 4,
               2, 3, 1, 7, 4, 3, 5, 1, 3, 11, 9, 1, 3, 4, 1, 2, 4, 2, 1, 4,
               4, 2, 5, 8, 12, 5, 2, 1, 1, 6, 2, 2, 5, 1, 4, 7, 1, 3, 1, 2,
               2, 2, 4, 4, 3, 4, 2, 4, 1, 5, 1, 2, 1, 3, 4, 1, 6, 1, 2, 3,
               12, 2, 3, 4, 13, 1, 1, 2, 3, 3, 3, 4, 2, 9, 1, 5, 1, 2, 1, 5,
               3, 5, 8, 1, 9, 9, 4, 2, 1, 5)
```

```
#campione2 <- rgeom(100, prob = 0.6) + 1
campione2 <- c(3, 1, 1, 1, 2, 2, 2, 1, 1, 1, 3, 1, 1, 1, 4, 2, 3, 2, 1, 4,
               4, 4, 1, 1, 1, 1, 3, 1, 1, 2, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1,
               1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 6,
               1, 1, 2, 3, 2, 4, 2, 1, 1, 2, 1, 1, 4, 6, 1, 2, 1, 1, 1, 1,
               2, 2, 3, 1, 1, 2, 1, 4, 1, 1, 1, 1, 1, 1, 2, 1, 6, 1, 2, 1)

intervallo <- differenzeValoriMedi(grado = 0.99,
                                   n1 = length(campione1), n2 = length(campione2),
                                   m1 = mean(campione1), m2 = mean(campione2))
```

L'intervallo ottenuto è

(1.0797592, 2.4269075)

Essendo gli estremi ottenuti entrambi positivi ne consegue che la media della prima popolazione è maggiore della media della seconda popolazione.

Capitolo 11

Verifica delle ipotesi

Il test di verifica delle ipotesi è un test che mira a verificare la bontà di un'ipotesi rappresentante un'affermazione su un parametro non noto della distribuzione che descrive la popolazione.

L'ipotesi da verificare tramite il test è detta ipotesi nulla H_0, se l'ipotesi nulla non può essere accettata allora viene accettata un'ipotesi alternativa H_1 espressa in contrapposizione all'ipotesi nulla.

Se Θ_0 e Θ_1 sono due sottoinsiemi disgiunti dello spazio dei parametri Θ allora l'ipotesi nulla e alternativa possono essere espresse come

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$$

Se l'ipotesi specifica completamente la funzione di probabilità o densità di probabilità allora l'ipotesi è detta semplice, altrimenti è detta composta.

La verifica del test può incorrere in due tipologie di errore:

- Errore di tipo I - rifiuto dell'ipotesi nulla quando l'ipotesi è vera
- Errore di tipo II - accettazione dell'ipotesi nulla quando l'ipotesi è falsa

Se per effettuare il test vengono utilizzati campioni di ampiezza fissa allora, in genere, al diminuire della probabilità di commettere un errore di un dato tipo aumenta la probabilità di commettere un errore dell'altro tipo.

Di conseguenza, se l'errore di tipo I è più grave dell'errore di tipo II, si fissa la probabilità di commettere errori di tipo I e si minimizza la probabilità di commettere errori di tipo II, altrimenti il contrario.

In genere le probabilità di errore scelte sono di 0.5 (test statisticamente significativo), 0.01 (test statisticamente molto significativo) e 0.001 (test statisticamente estremamente significativo).

Nel seguito verranno mostrati test di verifica delle ipotesi unilaterali e bilaterali con una probabilità α di commettere errori di tipo I sul valore medio di una distribuzione geometrica nel caso in cui la varianza sia nota.

In modo analogo a quanto fatto per il calcolo degli intervalli di fiducia approssimati, utilizzeremo una variabile pivot che converge in distribuzione alla normale standard.

Infatti se $X_1, \ldots, X_n$ è un campione casuale estratto da una popolazione geometrica con valore medio μ e varianza σ^2 allora la variabile aleatoria

$$Y_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

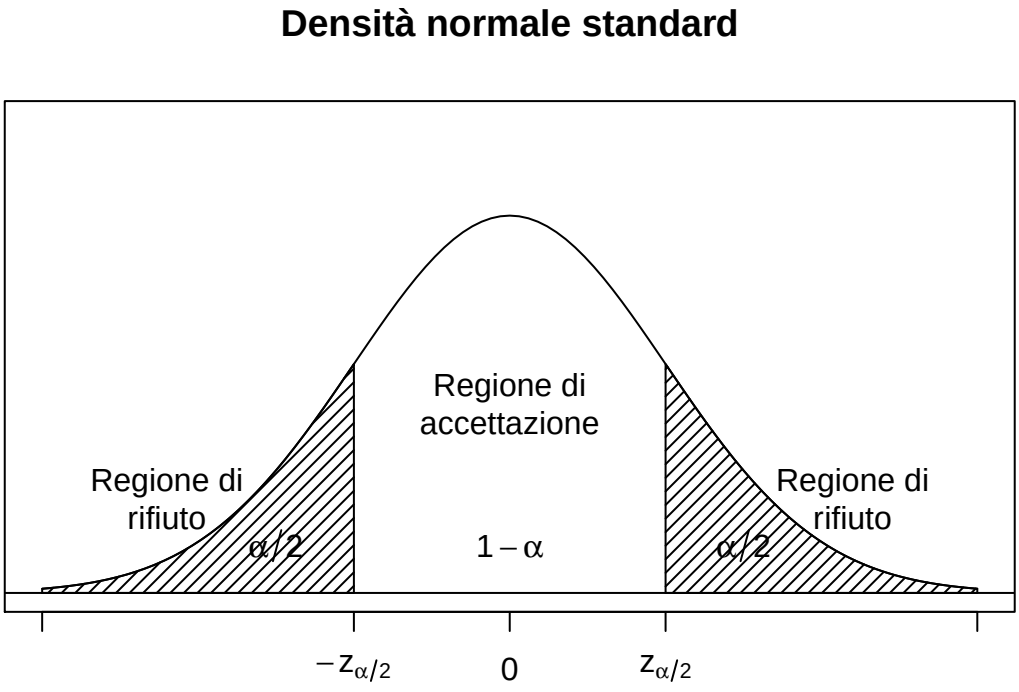

Figura 11.1: Densità normale standard

converge in distribuzione alla variabile normale standard $Z \sim \mathcal{N}(0,1)$.

11.1 Test bilaterale approssimato

Se $(x_1, \ldots, x_n)$ è un campione osservato di ampiezza n estratto da una popolazione geometrica con varianza nota σ^2 e se consideriamo le ipotesi:

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

allora il test bilaterale Φ di misura α per le ipotesi è:

Accettiamo l'ipotesi H_0 se

$$-z_{\alpha/2} < \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

Rifiutiamo l'ipotesi H_0 se

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{oppure} \quad \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

dove $-z_{\alpha/2}$ e $z_{\alpha/2}$ sono tali che

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

In figura 11.1 è rappresentata una densità normale standard e sono mostrate le regioni di rifiuto e accettazione per il test bilaterale.

La seguente funzione effettua il test bilaterale approssimato. La funzione prende in input la media e la deviazione standard della popolazione, la media campionaria e la misura α.

```
testBilaterale <- function(media, deviazioneStandard, mediaCampionaria, n, alpha) {
  zAlphaMezzi <- qnorm(1 - alpha / 2, mean = 0, sd = 1)
  valore <- (mediaCampionaria - media) / (deviazioneStandard / sqrt(n))
  -zAlphaMezzi < valore && valore < zAlphaMezzi
}
```

Il seguente codice effettua il test bilaterale sul valore medio utilizzando un campione che si ritiene estratto da una popolazione geometrica di parametro $p = 0.4$. Vengono quindi calcolate la media e la deviazione standard della popolazione e viene poi invocata la precedente funzione per ottenere il risultato del test.

```
p <- 0.4
media <- 1 / p
varianza <- (1 - p) / p^2
deviazioneStandard <- sqrt(varianza)

#campione <- rgeom(100, prob = 0.4) + 1
campione <- c(2, 1, 1, 3, 1, 3, 4, 5, 1, 1, 1, 1, 4, 2, 2, 2, 4, 3, 3, 1,
              1, 1, 2, 2, 1, 1, 1, 2, 3, 6, 3, 1, 6, 1, 1, 2, 1, 4, 1, 1,
              4, 3, 2, 1, 4, 6, 7, 2, 1, 1, 3, 1, 1, 4, 1, 1, 4, 1, 4, 1,
              2, 1, 1, 1, 4, 2, 1, 2, 3, 2, 2, 1, 2, 1, 2, 1, 2, 2, 6, 5,
              6, 2, 6, 2, 3, 2, 2, 2, 4, 1, 1, 1, 1, 2, 3, 9, 3, 1, 2, 1)

esito <- testBilaterale(media = media,
                        deviazioneStandard = deviazioneStandard,
                        mediaCampionaria = mean(campione),
                        n = length(campione),
                        alpha = 0.05)
```

Il test da esito positivo.

11.2 Test unilaterale sinistro approssimato

Se $(x_1, \ldots, x_n)$ è un campione osservato di ampiezza n estratto da una popolazione geometrica con varianza nota σ^2 e se consideriamo le ipotesi:

$$H_0 : \mu \leq \mu_0 \quad H_1 : \mu > \mu_0$$

allora il test unilaterale sinistro Φ di misura α per le ipotesi è:

Accettiamo l'ipotesi H_0 se

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$$

Rifiutiamo l'ipotesi H_0 se

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

dove z_α è tale che:

$$P(Z > z_\alpha) = \alpha$$

La seguente funzione effettua il test unilaterale sinistro approssimato. La funzione prende in input la media e la deviazione standard della popolazione, la media campionaria e la misura α.

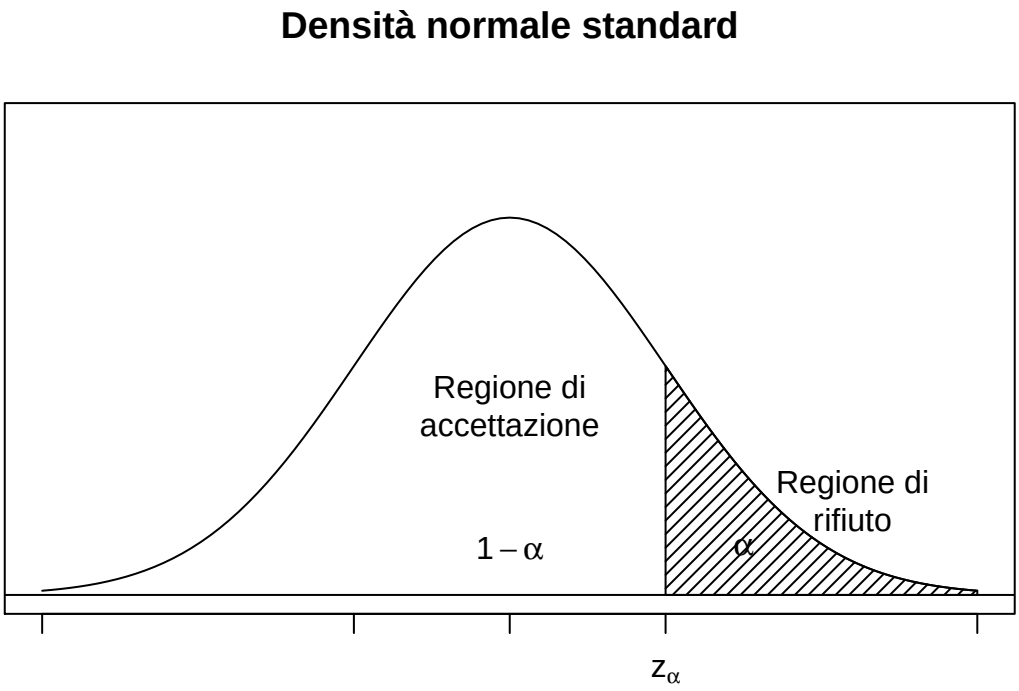

Figura 11.2: Densità normale standard

```
testUnilateraleSinistro <- function(media, deviazioneStandard, mediaCampionaria, n, alpha) {
  zAlpha <- qnorm(1 - alpha, mean = 0, sd = 1)
  valore <- (mediaCampionaria - media) / (deviazioneStandard / sqrt(n))
  valore < zAlpha
}
```

Il seguente codice effettua il test unilaterale sinistro sul valore medio utilizzando un campione che si ritiene estratto da una popolazione geometrica di parametro $p = 0.4$. Vengono quindi calcolati il valore atteso e la deviazione standard della popolazione e viene poi invocata la precedente funzione per ottenere il risultato del test.

```
p <- 0.4
media <- 1 / p
varianza <- (1 - p) / p^2
deviazioneStandard <- sqrt(varianza)

#campione <- rgeom(100, prob = 0.7) + 1
campione <- c(2, 2, 1, 1, 3, 1, 2, 1, 1, 2, 3, 1, 1, 1, 3, 2, 2, 1, 1, 2,
              2, 2, 1, 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
              2, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1,
              1, 1, 1, 1, 3, 1, 1, 3, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1,
              2, 3, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 3, 1, 1, 1, 1, 2)

esito <- testUnilateraleSinistro(media = media,
                                 deviazioneStandard = deviazioneStandard,
                                 mediaCampionaria = mean(campione),
                                 n = length(campione),
                                 alpha = 0.05)
```

Il test da esito positivo.

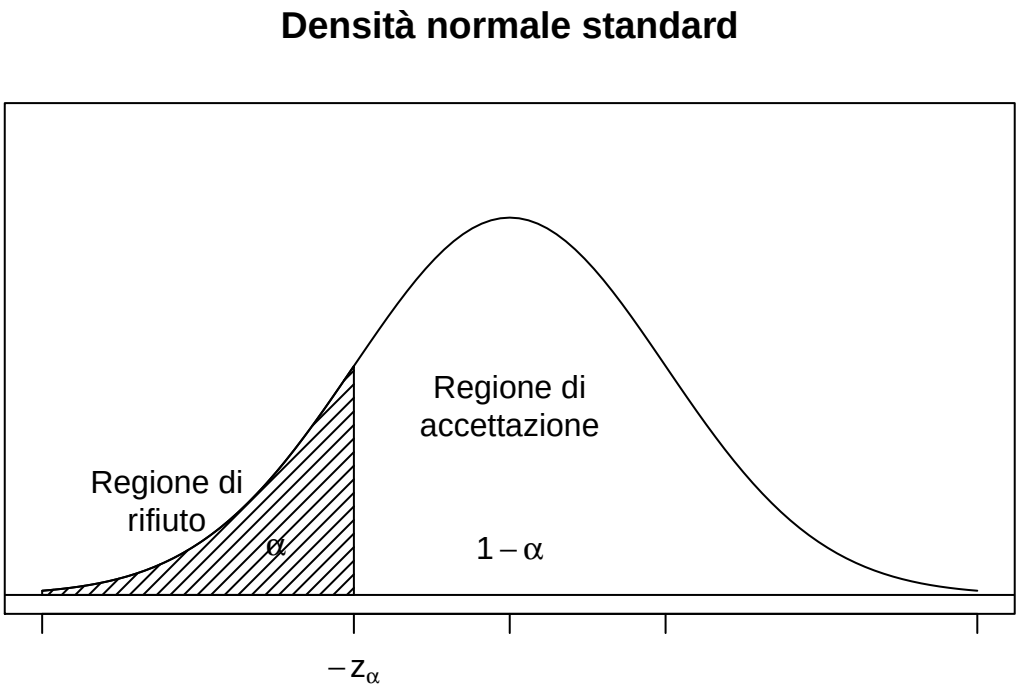

Figura 11.3: Densità normale standard

11.3 Test unilaterale destro approssimato

Se $(x_1, \ldots, x_n)$ è un campione osservato di ampiezza n estratto da una popolazione geometrica con varianza nota σ^2 e se consideriamo le ipotesi:

$$H_0 : \mu \geq \mu_0 \quad H_1 : \mu < \mu_0$$

allora il test unilaterale destro Φ di misura α per le ipotesi è:

Accettiamo l'ipotesi H_0 se

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha$$

Rifiutiamo l'ipotesi H_0 se

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

dove $-z_\alpha$ è tale che:

$$P(Z < -z_\alpha) = \alpha$$

La seguente funzione effettua il test unilaterale sinistro approssimato. La funzione prende in input la media e la deviazione standard della popolazione, la media campionaria e la misura α.

```
testUnilateraleDestro <- function(media, deviazioneStandard, mediaCampionaria, n, alpha) {
  menoAlpha <- qnorm(alpha, mean = 0, sd = 1)
  valore <- (mediaCampionaria - media) / (deviazioneStandard / sqrt(n))
  valore > menoAlpha
}
```

Il seguente codice effettua il test unilaterale destro sul valore medio utilizzando un campione che si ritiene estratto da una popolazione geometrica di parametro $p = 0.4$. Vengono quindi calcolati il valore atteso e la

deviazione standard della popolazione e viene poi invocata la precedente funzione per ottenere il risultato del test.

```
p <- 0.4
media <- 1 / p
varianza <- (1 - p) / p^2
deviazioneStandard <- sqrt(varianza)

#campione <- rgeom(100, prob = 0.7) + 1
campione <- c(2, 2, 1, 1, 3, 1, 2, 1, 1, 2, 3, 1, 1, 1, 3, 2, 2, 1, 1, 2,
              2, 2, 1, 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
              2, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1,
              1, 1, 1, 1, 3, 1, 1, 3, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1,
              2, 3, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 3, 1, 1, 1, 1, 2)

esito <- testUnilateraleDestro(media = media,
                               deviazioneStandard = deviazioneStandard,
                               mediaCampionaria = mean(campione),
                               n = length(campione),
                               alpha = 0.05)
```

Il test da esito negativo.

Capitolo 12

Criterio del chi-quadrato

Il criterio del chi-quadrato è un test di verifica delle ipotesi che consente di verificare se le frequenze dei valori osservati si adattano alle frequenze teoriche (probabilità) di una data distribuzione di probabilità e quindi poter dire, con una certa probabilità di errore, se il campione proviene dalla fissata distribuzione di probabilità.

Consideriamo una popolazione descritta da una variabile aleatoria X e avente una distribuzione di probabilità $F_X(x)$ con k parametri non noti da stimare. I parametri non noti possono essere stimati utilizzando il campione.

Il test di misura α consiste del verificare l'ipotesi nulla:

H_0 *: X ha una funzione di distribuzione* $F_X(x)$

o l'ipotesi alternativa:

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

La misura α è la probabilità di rifiutare l'ipotesi nulla.

Il test consiste nel suddividere l'insieme dei valori che può assumere la variabile aleatoria X in r sottoinsiemi $I_1, ..., I_r$ e calcolare le probabilità che la variabile aleatoria assuma un valore appartenente all'i-esimo sottoinsieme $p_i = P(X \in I_i)$ per $i = 1, ..., r$

Preso un campione $x_1, ...x_n$ si calcolano le frequenze assolute $n_1, ..., n_r$ con cui con i valori osservati si distribuiscono nei sottoinsiemi $I_1, ..., I_r$

Si calcola il valore chi-quadrato

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

pari alla somma dei quadrati degli scarti tra le frequenze teoriche e quelle osservate pesati sulle frequenze teoriche.

E si considera la statistica

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

dove N_i è la variabile aleatoria che descrive il numero di elementi del campione che appartengono all'intervallo i-esimo.

Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti, allora per n sufficientemente grande la variabile aleatoria Q converge in distribuzione alla variabile chi-quadrato con

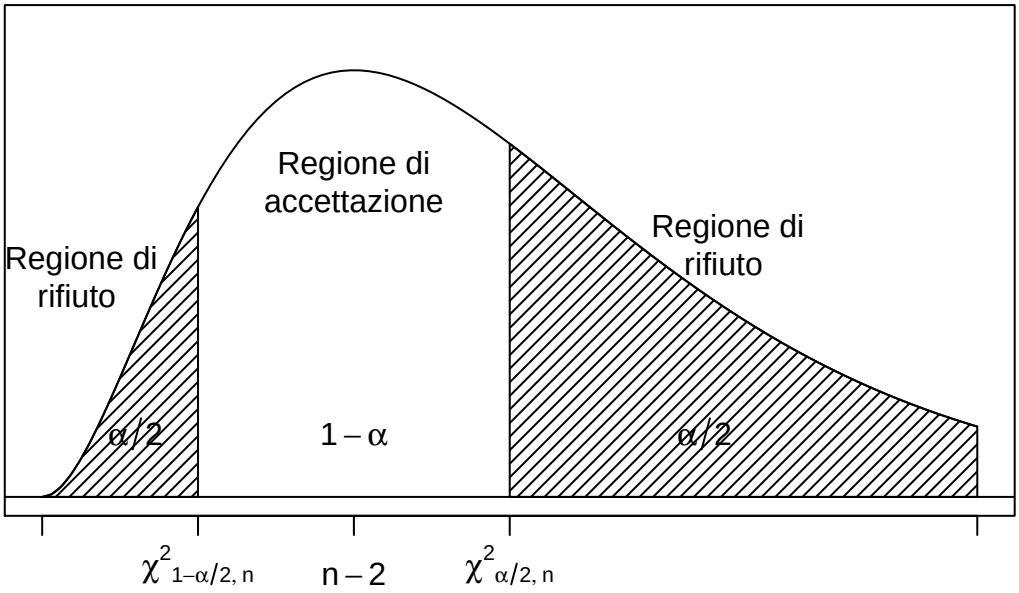

Figura 12.1: Densità chi-quadrato con n gradi di libertà

$r - k - 1$ gradi di libertà.
Il significato dei gradi di libertà è il seguente: se r viene sottratto 1 perché se conosciamo $r - 1$ probabilità possiamo calcolare l'r-esima invece k indica il numero di parametri non noti sostituiti da stime.

Il test del chi-quadrato di misura α per le ipotesi è:

Accettiamo l'ipotesi H_0 se

$$\chi^2_{1-\frac{\alpha}{2},r-k-1} < \chi^2 < \chi^2_{\frac{\alpha}{2},r-k-1}$$

Rifiutiamo l'ipotesi H_0 se

$$\chi^2 > \chi^2_{\frac{\alpha}{2},r-k-1} \quad \text{oppure} \quad \chi^2 < \chi^2_{1-\frac{\alpha}{2},r-k-1}$$

Dove i valori $\chi^2_{\frac{\alpha}{2},r-k-1}$ e $\chi^2_{1-\frac{\alpha}{2},r-k-1}$ sono tali che

$$P\left(Q < \chi^2_{1-\frac{\alpha}{2},r-k-1}\right) \backsimeq \alpha/2 \quad \text{e} \quad P\left(Q > \chi^2_{\frac{\alpha}{2},r-k-1}\right) \backsimeq \alpha/2$$

Il test funziona se ogni classe contiene in media almeno 5 elementi.

In figura 12.1 è rappresentata una densità chi-quadrato e sono mostrate le regioni di rifiuto e di accettazione.

12.1 Test per la distribuzione geometrica

In questo paragrafo verrà effettuato il test del chi-quadrato di misura $\alpha = 0.01$ su un campione per verificare la provenienza da una distribuzione geometrica.

Consideriamo il seguente campione

```
campione <- c(2, 1, 2, 1, 1, 1, 6, 1, 1, 2, 1, 1, 1, 2, 1, 2, 2, 3, 1, 1,
              2, 4, 1, 3, 1, 1, 2, 2, 1, 4, 4, 1, 3, 4, 2, 2, 2, 3, 1, 3,
              1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 2, 1, 1,
              2, 1, 1, 1, 1, 4, 1, 1, 2, 3, 2, 1, 1, 2, 1, 1, 2, 4, 2, 3,
              1, 1, 1, 1, 2, 1, 3, 1, 3, 6, 5, 1, 1, 1, 1, 1, 2, 2, 1, 1)
```

Con il seguente codice viene stimato il parametro non noto della distribuzione geometrica tramite inferenza sul campione.

```
stima.p <- 1 / mean(campione)
```

Nel seguente codice consideriamo 4 sottoinsiemi dei valori che può assumere una variabile aleatoria distribuita in modo geometrico. E per ogni sottoinsieme calcoliamo la probabilità che il valore della variabile aleatoria appartenga all'insieme.

```
r = 4
p <- numeric(r)
p[1] <- dgeom(0, prob = stima.p) # {1}
p[2] <- dgeom(1, prob = stima.p) # {2}
p[3] <- dgeom(2, prob = stima.p) # {3}
p[4] <- pgeom(2, prob = stima.p, lower.tail = FALSE) # [4, +infinity)
```

É possibile effettuare il test del chi-quadrato se ogni classe contiene in media almeno 5 elementi. Il mimino numero di elementi contenuti in media in una classe è calcolato con il seguente codice

```
n <- length(campione)
minimo <- min(n * p[1], n * p[2], n * p[3], n * p[4])
```

ed il risultato è 8.4144125 di conseguenza è possibile effettuare il test.

Il seguente codice calcola le frequenze dei valori osservati per ogni intervallo.

```
nint <- numeric(4)
nint[1] <- length(which(campione == 1))
nint[2] <- length(which(campione == 2))
nint[3] <- length(which(campione == 3))
nint[4] <- length(which(campione > 3))
```

Il seguente codice calcola il valore χ^2

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

Il valore di χ^2 calcolato è 0.4746316

Mentre il seguente codice calcola l'intervallo $\left(\chi^2_{1-\frac{\alpha}{2},r-k-1}, \quad \chi^2_{\frac{\alpha}{2},r-k-1}\right)$

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

L'intervallo calcolato è (0.0100251, 10.5966347).

Infine il seguente codice effettua il test del chi-quadrato

```
esito <- valori[1] < chi2 && chi2 < valori[2]
```

L'esito ottenuto del test è positivo, di conseguenza possiamo affermare che il campione proviene da una popolazione descritta da una distribuzione geometrica.

Bibliografia

Nobile, A. G. (2018). *Statistica e analisi dei dati (prima parte e seconda parte)*. Dipartimento di Informatica, Università degli Studi di Salerno.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.