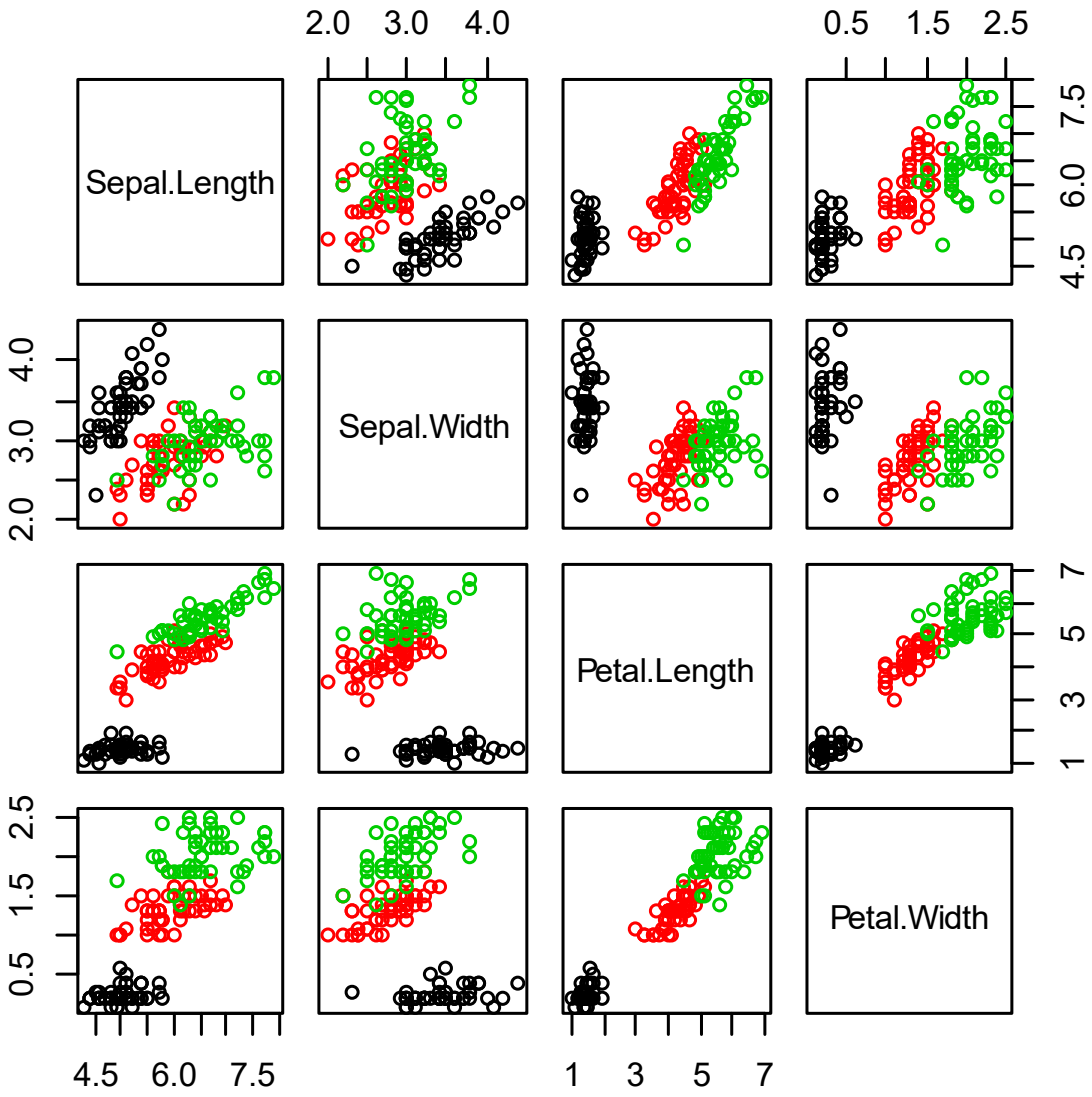
Elementi di Inferenza per Data Science

Lucio Barabesi



Pagina intenzionalmente vuota

Capitolo 1

L'analisi preliminare dei dati

1.1. La matrice dei dati

Se si considerano d variabili su n unità statistiche, in generale le osservazioni raccolte possono essere organizzate nella seguente matrice dei dati

$$\mathbf{D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}.$$

La i-esima riga della matrice **D** rappresenta le osservazioni rilevate sull'i-esima unità, mentre la j-esima colonna della matrice **D** fornisce le osservazioni relative a tutte le unità per la j-esima variabile. Le variabili analizzate possono essere di tipo qualitativo o quantitativo. A loro volta, le variabili quantitative possono essere di tipo continuo o discreto. Le variabili qualitative sono dette anche fattori.

L'analisi delle variabili sulla base della matrice dei dati viene usualmente effettuata sia in modo marginale (ovvero rispetto ad ogni singola variabile) che in modo congiunto (ovvero rispetto a gruppi di variabili o alla totalità delle variabili). Inoltre, può essere conveniente adottare una differente notazione quando si vuole distinguere le variabili esplicative da quelle di risposta. Se vi sono p variabili esplicative e $(d-p)$ variabili di risposta, la matrice **D** può essere opportunamente suddivisa nelle due matrici **X** (relativa alle osservazioni delle variabili esplicative) e **Y** (relativa alle osservazioni delle variabili di risposta), ovvero

$$\mathbf{D} = (\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_{11} & y_{12} & \cdots & y_{1(d-p)} \\ x_{21} & x_{22} & \cdots & x_{2p} & y_{21} & y_{22} & \cdots & y_{2(d-p)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_{n1} & y_{n2} & \cdots & y_{n(d-p)} \end{pmatrix}.$$

Inoltre, quando $d = 1$ si assume la notazione

$$\mathbf{D} = \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

mentre, quando $d = 2$ e $p = 1$, si adotta la notazione

$$\mathbf{D} = (\mathbf{x}, \mathbf{y}) = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}.$$

• **Esempio 1.1.1.** La seguente matrice dei dati è relativa ad un esperimento finalizzato ad analizzare i tempi per completare un semplice gioco enigmistico quando si odora un profumo e quando non lo si odora (Fonte: Hirsch, A.R. e Johnston, L.H., 1996, Odors and learning, *Journal of Neurological & Orthopedic Medicine & Surgery* **17**, 119-124). I dati, contenuti nel file `scent.txt`, vengono letti e resi disponibili mediante i seguenti comandi:

```
> d <- read.table("c:\\Rwork\\examples\\scent.txt", header = T)
> attach(d)
```

Per quanto riguarda l'analisi statistica, vi sono $n = 21$ soggetti su cui vengono misurate $d = 11$ variabili. I nomi delle variabili vengono ottenuti mediante il seguente comando:

```
> names(d)
[1] "Sex"    "Smoker"  "Opinion"  "Age"  "Order"  "U.T1"  "U.T2"
[8] "U.T3"   "S.T1"    "S.T2"     "S.T3"
```

Le prime $p = 5$ variabili (sesso, fumo, effetto percepito del profumo, età, ordine con cui si effettua l'esperimento) sono esplicative, mentre le ultime $d - p = 6$ variabili (tempi di reazione in tre esperimenti sequenziali indipendenti in cui non si è odorato o si è odorato il profumo) sono di risposta. Dunque, le prime tre variabili e la quinta sono di tipo qualitativo, mentre le restanti sono di tipo quantitativo. In particolare, la quarta variabile è quantitativa discreta, mentre le ultime sei sono quantitative continue. La matrice dei dati viene ottenuta esplicitamente mediante il seguente comando:

```
> d
   Sex Smoker Opinion Age Order U.T1 U.T2 U.T3 S.T1  S.T2 S.T3
1    M      N     Pos  23     1 38.4 27.7 25.7 53.1  30.6 30.2
2    F      Y     Neg  43     2 46.2 57.2 41.9 54.7  43.3 56.7
3    M      N     Pos  43     1 72.5 57.9 51.9 74.2  53.4 42.4
4    M      N     Neg  32     2 38.0 38.0 32.2 49.6  37.4 34.4
5    M      N     Neg  15     1 82.8 57.9 64.7 53.6  48.6 44.8
6    F      Y     Pos  37     2 33.9 32.0 31.4 51.3  35.5 42.9
7    F      N     Pos  26     1 50.4 40.6 40.1 44.1  46.9 42.7
8    F      N     Pos  35     2 35.0 33.1 43.2 34.0  26.4 24.8
9    M      N     Pos  26     1 32.8 26.8 33.9 34.5  25.1 25.1
10   F      N     Ind  31     2 60.1 53.2 40.4 59.1  87.1 59.2
11   F      Y     Pos  35     1 75.1 63.1 58.0 67.3  43.8 42.2
12   F      Y     Ind  55     2 57.6 57.7 61.5 75.5 126.6 48.4
13   F      Y     Pos  25     1 55.5 63.3 44.6 41.1  41.8 32.0
14   M      Y     Ind  39     2 49.5 45.8 35.3 52.2  53.8 48.1
15   M      N     Ind  25     1 40.9 35.7 37.2 28.3  26.0 33.7
16   M      N     Pos  26     2 44.3 46.8 39.4 74.9  45.3 42.6
17   M      Y     Neg  33     1 93.8 91.9 77.4 77.5  55.8 54.9
18   M      N     Neg  62     2 47.9 59.9 52.8 50.9  58.6 64.5
19   F      Y     Pos  54     1 75.2 54.1 63.6 70.1  44.0 43.1
20   F      N     Neg  38     2 46.2 39.3 56.6 60.3  47.8 52.8
21   M      N     Neg  65     1 56.3 45.8 58.9 59.9  36.8 44.3
```

In questo esempio, la visualizzazione della matrice dei dati è possibile in quanto le corrispondenti dimensioni sono ridotte. In caso di dati massivi (i cosiddetti "big data" nella terminologia anglosassone), questa operazione non risulta utile e può essere perfino proibitiva. □

Inizialmente, secondo la prassi usuale, l'analisi esplorativa della matrice dei dati viene condotta marginalmente su ogni singola variabile. Nel caso in cui la variabile analizzata sia quantitativa, si considerano i quantili ed il connesso diagramma a scatola e baffi e, in seguito, si implementano

l'istogramma e gli ulteriori indici di sintesi. Se la variabile è qualitativa, si adotta invece il diagramma a nastri.

Successivamente si considera l'analisi esplorativa per coppie di variabili. Di nuovo, questa indagine viene condotta mediante sintesi numeriche e grafiche. Se la coppia di variabili è quantitativa, si rappresenta i dati mediante il diagramma di dispersione e si analizza la relazione fra variabili mediante indici di dipendenza. Se una delle variabili è quantitativa e l'altra è qualitativa, si adottano invece i diagrammi a scatola e baffi condizionati. Se entrambe le variabili sono qualitative i dati vengono sintetizzati in una tabella a doppia entrata e analizzate mediante i diagrammi a nastro condizionati e con opportuni indici di dipendenza.

Quando si vuole analizzare gruppi di variabili o la globalità delle variabili, l'analisi diventa ovviamente più complessa. Tuttavia è possibile introdurre alcuni strumenti che permettono di facilitare l'indagine esplorativa, quali la matrice dei grafici di dispersione e la matrice di correlazione. Inoltre, in questo ambito sono utili le rappresentazioni grafiche per gruppi di variabili, quali il diagramma a stelle e il diagramma a coordinate parallele. Infine, si possono considerare le rappresentazioni grafiche per dati rilevati nel tempo (quali le serie temporali) e/o nello spazio (quali le mappe coropletiche). Le tecniche esplorative descritte verrano analizzate in dettaglio nelle prossime sezioni.

1.2. L'analisi esplorativa marginale delle variabili

L'analisi esplorativa dei dati è usualmente iniziata mediante l'indagine marginale delle d variabili, e viene condotta mediante sintesi numeriche e grafiche. Supponiamo di considerare una singola variabile quantitativa e si desideri effettuare una prima analisi esplorativa. Le osservazioni relative alla variabile, ovvero $x_1, \ldots, x_n$, possono essere convenientemente ordinate e indicate con i simboli $x_{(1)} < \cdots < x_{(n)}$. Queste quantità possono venire rappresentate mediante segmenti su un asse ordinato al fine di graficizzare la relativa distribuzione, ovvero l'insieme di valori assunti dalla variabile.

Il quantile di ordine α, con $\alpha \in [0, 1]$, è un valore $\widetilde{x}_\alpha$ che separa in due gruppi le osservazioni ordinate, ovvero la frazione α di osservazioni più piccole e la frazione $(1 - \alpha)$ di quelle più elevate. Non esiste un valore unico di $\widetilde{x}_\alpha$ eccetto che in alcuni casi particolari. Ad esempio, quando $\alpha = 0.5$ e n è dispari, si ottiene immediatamente il valore unico $\widetilde{x}_{0.5} = x_{(n/2+1/2)}$, mentre se $\alpha = 0.5$ e n è pari, $\widetilde{x}_{0.5}$ può essere scelto come un qualsiasi valore interno all'intervallo $[x_{(n/2)}, x_{(n/2+1)}]$. Esistono varie proposte per la selezione di un valore unico per un generico quantile, che tendono comunque a coincidere per n elevato. Inoltre, $\widetilde{x}_{0.5}$ è detta mediana, $\widetilde{x}_{0.25}$ è detto primo quartile, mentre $\widetilde{x}_{0.75}$ è detto terzo quartile. Infine, per definizione si ha $\widetilde{x}_0 = x_{(1)}$ e $\widetilde{x}_1 = x_{(n)}$, ovvero per $\alpha = 0$ e $\alpha = 1$ si ottengono il minimo e il massimo delle osservazioni. I precedenti cinque quantili sono detti di base, in quanto caratterizzano sommariamente la distribuzione delle osservazioni.

Per quanto riguarda l'interpretazione dei quantili di base, la mediana individua il valore "centrale" della distribuzione. Il primo e terzo quartile individuano un intervallo $[\widetilde{x}_{0.25}, \widetilde{x}_{0.75}]$ che contiene la metà delle osservazioni (ovvero quelle più "interne" della distribuzione) e che fornisce un'informazione sulla dispersione della variabile. Infine, il minimo e il massimo individuano il dominio delle osservazioni, ovvero l'intervallo $[x_{(1)}, x_{(n)}]$ che contiene tutte le osservazioni.

Mediante i cinque quantili di base si può produrre un grafico importante per una prima analisi esplorativa di una variabile quantitativa, ovvero il cosiddetto diagramma a scatola e baffi. Questo diagramma è basato su una rettangolo (la cosiddetta scatola) di larghezza arbitraria, la cui lunghezza è data dalla differenza fra il terzo e il primo quartile, ovvero $(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$. Due segmenti (i cosiddetti baffi) si estendono oltre il rettangolo. Il primo baffo si estende fra il primo quartile e il valore adiacente inferiore, ovvero la più piccola osservazione maggiore di $\widetilde{x}_{0.25} - 1.5(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$. Il secondo baffo si estende fra il terzo quartile e il valore adiacente superiore, ovvero la più grande osservazione minore di $\widetilde{x}_{0.75} + 1.5(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$. La costante 1.5 è arbitraria e dettata da una scelta di compromesso. Un valore anomalo è una osservazione più piccola del valore adiacente inferiore o più

grande del valore adiacente superiore. Parallelamente al diagramma a scatola e baffi vengono usualmente riportati anche i segmenti relativi alle osservazioni ordinate. Un esempio di diagramma a scatola e baffi è dato nella Figura 1.2.1.

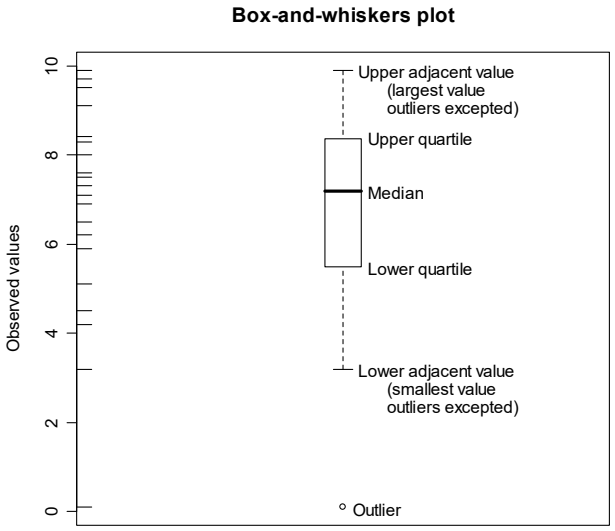


Figura 1.2.1.

• **Esempio 1.2.1.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, ovvero quelli relativi all'esperimento con i profumi, e si analizzano i tempi di risposta alla prima prova quando i soggetti non odorano profumo (ovvero la variabile `U.T1`). I cinque quantili fondamentali vengono calcolati mediante il seguente comando:

```
> summary(U.T1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.80   40.90   49.50   53.92   60.10   93.80
```

Inoltre, il diagramma a scatola e baffi viene ottenuto mediante i seguenti comandi:

```
> boxplot(U.T1, boxwex = 0.3,
+    ylab = "Unscented first trial time (seconds)",
+    main = "Box-and-whiskers plot")
> rug(U.T1, side = 2)
```

I precedenti comandi producono il grafico riportato in Figura 1.2.2. Si osservi che le distribuzioni marginali di più variabili omogenee (quali ad esempio le variabili `U.T1`, `U.T2`, `U.T3`, `S.T1`, `S.T2`, `S.T3`) possono essere confrontate riportando in un unico grafico i vari diagrammi a scatola e baffi corrispondenti ad ogni variabile. Il comando per effettuare questa analisi è il seguente:

```
> boxplot(d[, 6:11], boxwex = 0.3, ylab = "Time (seconds)",
+    main = "Box-and-whiskers plot")
```

Il precedente comando fornisce il grafico riportato in Figura 1.2.3. La Figura 1.2.3 permette dunque di analizzare contemporaneamente i tempi di risposta al variare dell'apprendimento nei due differenti protocolli. □

Figura 1.2.2.

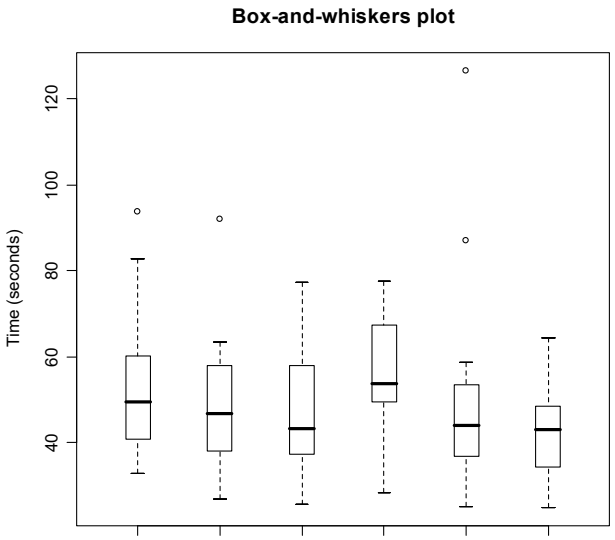


Figura 1.2.3.

Quando si considera una variabile quantitativa discreta o se vi sono arrotondamenti nelle misurazioni di una variabile quantitativa continua, molte determinazioni della variabile possono coincidere. Si supponga che vi siano $r < n$ determinazioni distinte della variabile e che vengano indicate con le quantità $c_1, \ldots, c_r$, che di solito vengono assunte ordinate, ovvero $c_1 < \cdots < c_r$. In questo caso, è conveniente considerare la frequenza delle osservazioni, ovvero il numero di ripetizioni di ogni determinazione distinta della variabile. Le frequenze vengono indicate con i simboli $n_1, \ldots, n_r$. L'insieme delle r coppie $(c_1, n_1), \ldots, (c_r, n_r)$ è detto distribuzione di frequenza e può essere organizzato in una tavola di 2 righe per n colonne.

• **Esempio 1.2.2.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare la variabile `Age`. Il comando per ottenere la distribuzione di frequenza è il seguente:

```
> table(Age)
Age
15 23 25 26 31 32 33 35 37 38 39 43 54 55 62 65
 1  1  2  3  1  1  1  2  1  1  1  2  1  1  1  1
```

Evidentemente, nella precedente distribuzione di frequenza si ha $r = 16$. □

Quando si analizza una variabile quantitativa è comunque conveniente considerare dei raggruppamenti di osservazioni al fine di mitigare gli effetti delle imprecisioni nelle misurazioni e degli arrotondamenti. In questo caso, le osservazioni vengono suddivise in un insieme di r classi contigue (ovvero intervalli di valori) mutuamente esclusive ed esaustive, che sono selezionate opportunamente. Evidentemente, sussiste una certa arbitrarietà nella scelta degli estremi delle classi. Inoltre, le r classi vengono indicate con $]c_0, c_1], \ldots,]c_{r-1}, c_r]$, dove $c_0 < c_1 < \cdots < c_r$. In questo caso, si dispone delle frequenze di classe (ovvero il numero di osservazioni per ogni classe) e le relative densità (ovvero il rapporto fra le frequenze di classe e la lunghezza della relativa classe). Le frequenze di classe vengono indicate con i simboli $n_1, \ldots, n_r$ e le relative densità sono date da $n_1/(c_1 - c_0), \ldots, n_r/(c_r - c_{r-1})$. L'insieme delle classi e delle corrispondenti frequenze è detta distribuzione di frequenza per classi.

Un grafico che permette un'analisi esplorativa di una variabile quantitativa raggruppata in classi è l'istogramma. L'istogramma si ottiene riportando su ogni classe un rettangolo la cui base coincide con la classe stessa, mentre l'altezza è proporzionale alla densità. Dunque, l'area del rettangolo è proporzionale alla frequenza di classe. Le altezze vengono generalmente riproporzionate in modo tale che l'area totale dei rettangoli sia pari ad uno.

• **Esempio 1.2.3.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1 e si analizzano i tempi di risposta alla prima prova quando i soggetti non odorano profumo (variabile `U.T1`). In questo caso, le classi adottate sono $]25, 34],]34, 44],]44, 56],]56, 65],]65, 85],]85, 95]$. La distribuzione di frequenza per classi si ottiene eseguendo il comando:

```
> table(cut(U.T1, breaks = c(25, 34, 44, 56, 65, 85, 95)))
(25,34] (34,44] (44,56] (56,65] (65,85] (85,95]
      2       4       7       3       4       1
```

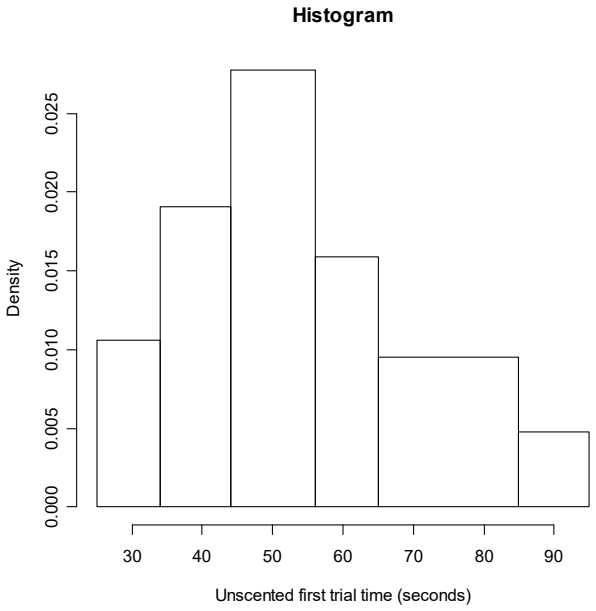


Figura 1.2.4.

Il comando per ottenere l'istogramma è il seguente:

```
> hist(U.T1, breaks = c(25, 34, 44, 56, 65, 85, 95),
+     xlab = "Unscented first trial time (seconds)",
+     ylab = "Density", main = "Histogram")
```

Il precedente comando fornisce il grafico di Figura 1.2.4. Si osservi che una scelta differente delle classi conduce ad un diverso istogramma. □

L'analisi esplorativa marginale di una variabile quantitativa viene usualmente rifinita mediante quattro ulteriori indici di sintesi. Per quanto riguarda la tendenza centrale della distribuzione, il primo indice è dato dalla media aritmentica

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

La mediana $\widetilde{x}_{0.5}$ viene frequentemente preferita alla media come indice di tendenza centrale, in quanto meno sensibile ai valori anomali. Per quanto riguarda la variabilità della distribuzione, il secondo indice è la varianza

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \ .$$

Al fine di ottenere un indice lineare nell'unità di misura, si considera usualmente la radice della varianza, ovvero lo scarto quadratico medio s_x. A loro volta s_x^2 e s_x sono sensibili ai valori anomali e si preferisce adottare come indice di variabilità il rango interquartile, dato da

$$\text{IQR}_x = \widetilde{x}_{0.75} - \widetilde{x}_{0.25} \ ,$$

piuttosto che lo scarto quadratico medio s_x. Inoltre, se si devono confrontare le variabilità di distribuzioni marginali per variabili omogenee, è conveniente adottare indici di variabilità che non dipendono dall'unità di misura, quali il coefficiente di variazione $s_x/|\overline{x}|$ o il rango interquartile standardizzato $\text{IQR}_x/|\widetilde{x}_{0.5}|$.

Per quanto riguarda l'analisi dell'asimmetria della distribuzione, il terzo indice è il coefficiente di asimmetria

$$a_3 = \frac{1}{n s_x^3} \sum_{i=1}^{n} (x_i - \overline{x})^3 \ .$$

Questo indice non dipende dall'unità di misura. Il coefficiente di asimmetria assume valori intorno a zero per distribuzioni approssimativamente simmetriche (ovvero distribuzioni con code simili), valori negativi per distribuzioni con asimmetria negativa (ovvero con code che si allungano verso la parte sinistra della distribuzione) e valori positivi per distribuzioni con asimmetria positiva (ovvero con code che si allungano verso la parte destra della distribuzione). Per quanto riguarda l'analisi della forma della distribuzione, il quarto indice è il coefficiente di curtosi

$$a_4 = \frac{1}{n s_x^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 \ .$$

Il valore di riferimento per questo indice è 3. Il coefficiente di curtosi assume valori elevati per distribuzioni leptocurtiche (ovvero distribuzioni con code molto allungate), mentre assume valori bassi per distribuzioni platicurtiche (ovvero distribuzioni con code molto brevi).

• **Esempio 1.2.4.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare la variabile `U.T1`. Non esiste un comando specifico per calcolare gli indici di sintesi eccetto che per la media, anche se è immediato programmare le seguenti funzioni per il calcolo della varianza e dei coefficienti di asimmetria e curtosi:

```
> variance <- function(x){
+    m2 <- sum((x - mean(x))^2)/length(x)
+    m2}
> skewness <- function(x){
+    s3 <- sum((x - mean(x))^3)/length(x)/sqrt(variance(x))^3
+    s3}
> kurtosis <- function(x){
+    s4 <- sum((x - mean(x))^4)/length(x)/variance(x)^2
+    s4}
```

Gli indici di sintesi per la variabile considerata vengono dunque ottenuti mediante i seguenti comandi:

```
> mean(U.T1)
[1] 53.92381
> variance(U.T1)^(1/2)
[1] 16.7326
> skewness(U.T1)
[1] 0.782112
> kurtosis(U.T1)
[1] 2.677314
```

La distribuzione considerata è dunque moderatamente asimmetrica (con asimmetria positiva) e leggermente platicurtica. La variabilità relativa delle distribuzioni marginali per le variabili `U.T1`, `U.T2`, `U.T3`, `S.T1`, `S.T2`, `S.T3` possono essere confrontate mediante i seguenti comandi che calcolano i coefficienti di variazione:

```
> variance(U.T1)^(1/2)/abs(mean(U.T1))
[1] 0.3103008
> variance(U.T2)^(1/2)/abs(mean(U.T2))
[1] 0.3051359
> variance(U.T3)^(1/2)/abs(mean(U.T3))
[1] 0.2794869
> variance(S.T1)^(1/2)/abs(mean(S.T1))
[1] 0.2513865
> variance(S.T2)^(1/2)/abs(mean(S.T2))
[1] 0.456505
> variance(S.T3)^(1/2)/abs(mean(S.T3))
[1] 0.2418836
```

Le distribuzioni considerate hanno quindi indici di dispersione relativa che sono abbastanza simili. □

Se la variabile analizzata è qualitativa, l'analisi esplorativa si riduce semplicemente nel determinare la distribuzione di frequenza. In questo caso, vi sono $r < n$ determinazioni distinte della variabile qualitativa (le cosiddette modalità), che vengano opportunamente indicate con i simboli $c_1, \ldots, c_r$. Le frequenze delle r modalità vengono indicate con i simboli $n_1, \ldots, n_r$. L'insieme delle r coppie $(c_1, n_1), \ldots, (c_r, n_r)$ è di nuovo detto distribuzione di frequenza e può essere organizzato in una tavola di 2 righe per n colonne. Da un punto di vista grafico la distribuzione di frequenza viene rappresentata mediante il diagramma a nastri, che è un grafico basato su nastri di lunghezza pari alle frequenze di ogni determinazione della variabile e di identica larghezza (che viene scelta in modo soggettivo).

• **Esempio 1.2.5.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare la variabile `Opinion`. Il comando per ottenere la distribuzione di frequenza è il seguente:

```
> table(Opinion)
Opinion
Ind Neg Pos
  4   7  10
```

Inoltre, richiamando la libreria `lattice` che permette di implementare metodi grafici avanzati, il diagramma a nastri si ottiene mediante i seguenti comandi:

```
> library(lattice)
> barchart(table(Opinion), xlab = "Frequency", ylab = "Opinion",
+    main = "Barplot")
```

I precedenti comandi forniscono il grafico contenuto nella Figura 1.2.5.

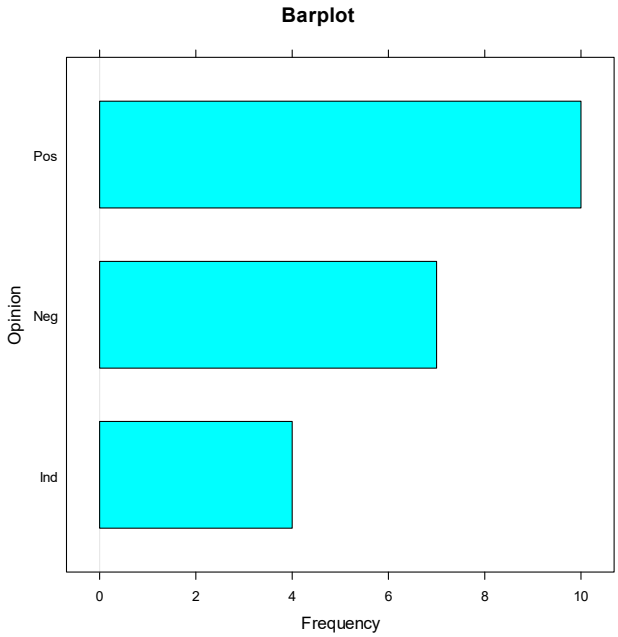


Figura 1.2.5.

Si osservi che, quando il numero delle determinazioni distinte è elevato, è conveniente considerare una larghezza dei nastri più piccola. □

1.3. L'analisi esplorativa di coppie di variabili

L'analisi marginale per coppie di variabili viene condotta di nuovo mediante sintesi numeriche e grafiche. Se entrambe le variabili analizzate sono quantitative, le osservazioni sono costituite da n coppie cartesiane $(x_1, y_1), \ldots, (x_n, y_n)$ che possono venire rappresentate mediante un grafico detto diagramma di dispersione. Il diagramma di dispersione permette di avere una prima impressione sull'esistenza di dipendenza fra le variabili.

• **Esempio 1.3.1.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare le variabili `U.T1` e `S.T1`. Il comando per ottenere il diagramma di dispersione è il seguente

```
> plot(U.T1, S.T1, xlab = "Unscented first trial time (seconds)",
+    ylab = "Scented first trial time (seconds)",
+    main = "Scatter plot")
```

Il precedente comando fornisce il grafico della Figura 1.3.1.

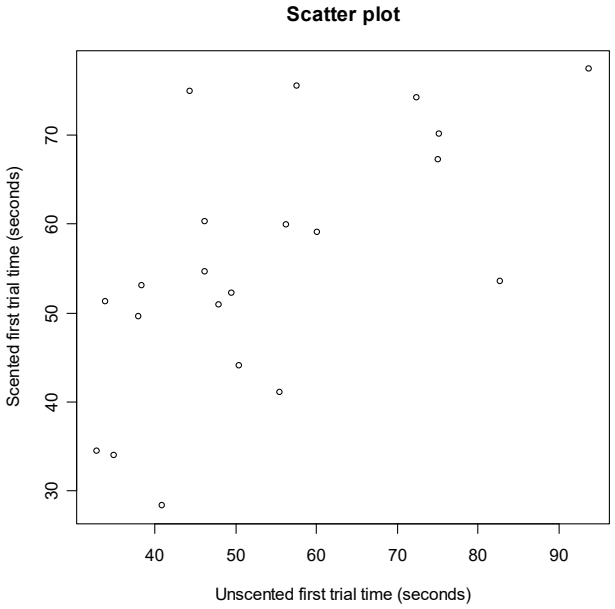


Figura 1.3.1.

Il diagramma di dispersione evidenzia una modesta dipendenza positiva fra le due variabili. □

Una volta che si è verificata graficamente l'esistenza di una relazione fra le due variabili, è conveniente considerare indici per quantificare il grado di dipendenza esistente. Se si sospetta una dipendenza di tipo lineare è opportuno calcolare il coefficiente di correlazione lineare dato da

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

dove s_x e s_y rappresentano rispettivamente gli scarti quadratici della prima e seconda variabile, mentre la quantità

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

è detta covarianza. Risulta $r_{xy} \in [-1, 1]$ e i valori estremi dell'indice sono raggiunti quando vi è dipendenza lineare perfetta inversa ($r_{xy} = -1$) e dipendenza lineare perfetta diretta ($r_{xy} = 1$). Un valore di r_{xy} intorno allo zero denota mancanza di dipendenza lineare. Si noti tuttavia che si può avere $r_{xy} = 0$ nel caso di un legame perfetto di tipo non lineare.

• **Esempio 1.3.2.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare le variabili `U.T1` e `S.T1`. Il comando per ottenere il coefficiente di correlazione è il seguente:

```
> cor(U.T1, S.T1)
[1] 0.6316886
```

Dunque, in questo caso si evidenzia una modesta dipendenza lineare diretta. □

Quando si considerano coppie di variabili quantitative discrete o se vi sono forti arrotondamenti nelle misurazioni di coppie di variabili quantitative continue, molte determinazioni possono coincidere. Analogamente, la stessa situazione si presenta quando una variabile della coppia è discreta

(o arrotondata) e l'altra è qualitativa o quando le osservazioni vengono poste in classi. Si supponga che vi siano r determinazioni distinte della prima variabile (indicate con $c_1, \ldots, c_r$) e s determinazioni distinte della seconda variabile (indicate con $d_1, \ldots, d_s$). In questo caso, è conveniente considerare la frequenza congiunta di (c_j, d_l), ovvero il numero di ripetizioni di ogni coppia di determinazioni distinte. Le frequenza congiunta di (c_j, d_l) viene indicata con il simbolo n_{jl}. La matrice di frequenze (di ordine $r \times s$) che si ottiene in questo modo è detta tabella a doppia entrata. L'insieme delle terne (c_j, d_l, n_{jl}), con $j = 1, \ldots, r, l = 1, \ldots, s$, è detta distribuzione di frequenza congiunta.

La distribuzione di frequenza marginale della prima variabile è data dalle coppie (c_j, n_{j+}) dove

$$n_{j+} = \sum_{l=1}^{s} n_{jl} \, ,$$

mentre la distribuzione di frequenza marginale della seconda variabile è data dalle coppie (d_l, n_{+l}) dove

$$n_{+l} = \sum_{j=1}^{r} n_{jl} \, .$$

Evidentemente, le distribuzioni di frequenza marginali sono quelle che si ottengono considerando una variabile come se l'altra non fosse presente. La tabella a doppia entrata viene usualmente rappresentata come nella Tavola 1.3.1. Si noti che si può sempre ricostruire la matrice dei dati originale a partire dalla tabella a doppia entrata e viceversa.

Tavola 1.3.1.

	d_1	$\cdots$	d_s	
c_1	n_{11}	$\cdots$	n_{1s}	n_{1+}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_r	n_{r1}	$\cdots$	n_{rs}	n_{r+}
	n_{+1}	$\cdots$	n_{+s}	n

• **Esempio 1.3.3.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare le variabili `Sex` e `Age`. Il comando per ottenere la tabella a doppia entrata è il seguente:

```
> table(Sex, Age)
   Age
Sex 15 23 25 26 31 32 33 35 37 38 39 43 54 55 62 65
  F  0  0  1  1  1  0  0  2  1  1  0  1  1  1  0  0
  M  1  1  1  2  0  1  1  0  0  0  1  1  0  0  1  1
```

La prima distribuzione marginale viene ottenuta mediante il seguente comando:

```
> margin.table(table(Sex, Age), 1)
Sex
 F  M
10 11
```

La seconda distribuzione marginale viene ottenuta mediante il seguente comando:

```
> margin.table(table(Sex, Age), 2)
Age
15 23 25 26 31 32 33 35 37 38 39 43 54 55 62 65
 1  1  2  3  1  1  1  2  1  1  1  2  1  1  1  1
```

Si consideri le variabili `U.T1` e `U.T2` e le classi $]25, 34],]34, 44],]44, 56],]56, 65],]65, 85],]85, 95]$. La tabella a doppia entrata con le frequenze di classe si ottiene eseguendo il comando:

```
> table(cut(U.T1, breaks = c(25, 34, 44, 56, 65, 85, 95)),
+     cut(U.T2, breaks = c(25, 34, 44, 56, 65, 85, 95)))
          (25,34] (34,44] (44,56] (56,65] (65,85] (85,95]
  (25,34]       2       0       0       0       0       0
  (34,44]       2       2       0       0       0       0
  (44,56]       0       2       2       3       0       0
  (56,65]       0       0       2       1       0       0
  (65,85]       0       0       1       3       0       0
  (85,95]       0       0       0       0       0       1
```

La prima distribuzione marginale viene ottenuta mediante il seguente comando:

```
> margin.table(table(
+     cut(U.T1, breaks = c(25, 34, 44, 56, 65, 85, 95)),
+     cut(U.T2, breaks = c(25, 34, 44, 56, 65, 85, 95))), 1)
(25,34] (34,44] (44,56] (56,65] (65,85] (85,95]
      2       4       7       3       4       1
```

La seconda distribuzione marginale viene ottenuta mediante il seguente comando:

```
> margin.table(table(
+     cut(U.T1, breaks = c(25, 34, 44, 56, 65, 85, 95)),
+     cut(U.T2, breaks = c(25, 34, 44, 56, 65, 85, 95))), 2)
(25,34] (34,44] (44,56] (56,65] (65,85] (85,95]
      4       4       5       7       0       1
```

□

Quando si analizzano coppie di variabili di cui una è quantitativa e l'altra qualitatitiva, non è possibile dare una rappresentazione cartesiana delle coppie di osservazioni. In questo caso, risulta conveniente implementare un diagramma a scatola e baffi condizionato. Questo tipo di grafico si ottiene riportando un diagramma a scatola e baffi per le osservazioni della variabile quantitativa in corrispondenza di ogni determinazione della variabile qualitativa. I diagrammi a scatola e baffi condizionati differiscono in maniera sostanziale dalla serie di diagrammi a scatola e baffi che si adottano quando si confrontano più variabili omogenee. Infatti, nel primo caso ogni diagramma è riferito alla solita variabile e calcolato solamente sulla parte di osservazioni che manifesta la medesima determinazione della variabile qualitativa, mentre nel secondo caso ogni diagramma è riferito a variabili differenti (anche se omogenee) e calcolato sulla totalità delle n osservazioni.

• **Esempio 1.3.4.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1, e in particolare le variabili `U.T1` e `Sex`. Il comando per ottenere i diagrammi a scatola e baffi condizionati è il seguente:

```
> boxplot(U.T1 ~ Sex, boxwex = 0.3,
+     ylab = "Unscented first trial time (seconds)",
+     main = "Box-and-whiskers plot")
```

Il precedente comando fornisce il grafico della Figura 1.3.2. □

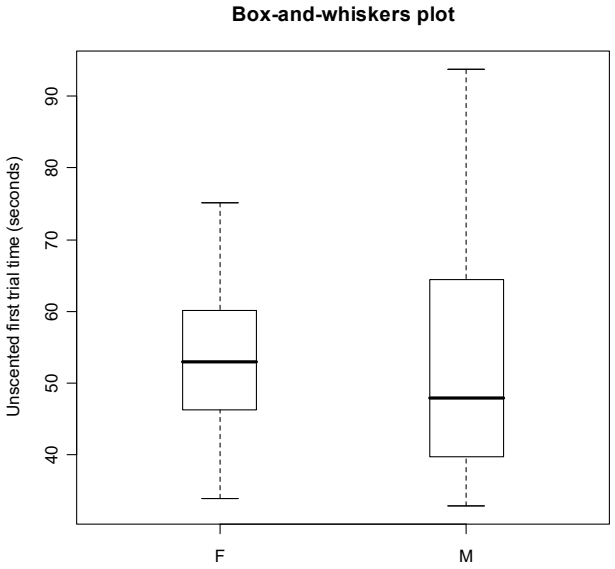


Figura 1.3.2.

Se in una coppia di variabili entrambe le variabili sono qualitative, l'analisi esplorativa si riduce semplicemente nel determinare la distribuzione di frequenza congiunta. Da un punto di vista grafico la distribuzione di frequenza bivariata viene rappresentata mediante i diagrammi a nastri condizionati, che sono basati su nastri (di lunghezza pari alle frequenze di ogni determinazione della prima variabile) che vengono ripartiti rispetto alla composizione della seconda variabile.

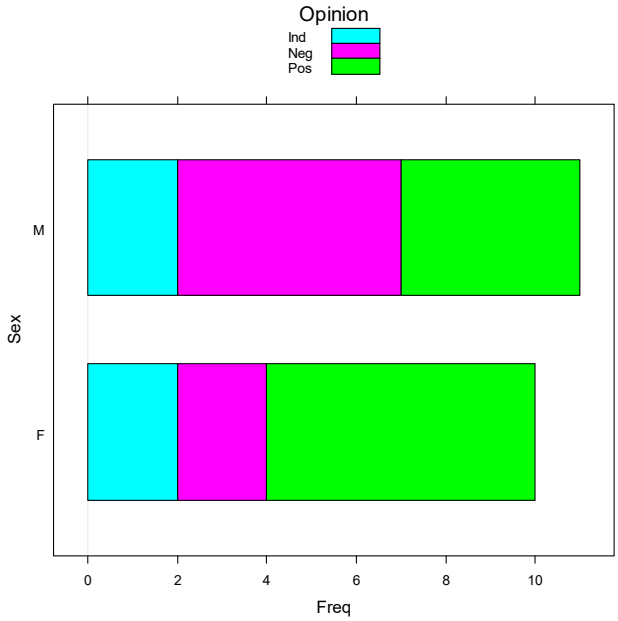


Figura 1.3.3.

● **Esempio 1.3.5.** Si considerano di nuovo i dati relativi all'esperimento con profumi dell'Esempio 1.1.1, e in particolare le variabili `Sex` e `Opinion`. Il comando per ottenere la tabella a doppia entrata è il seguente:

```
> table(Sex, Opinion)
   Opinion
Sex Ind Neg Pos
  F   2   2   6
  M   2   5   4
```

Inoltre, richiamando la libreria `lattice` che permette di implementare metodi grafici avanzati, i diagrammi a nastri condizionati si ottengono mediante i seguenti comandi:

```
> library(lattice)
> barchart(table(Sex, Opinion), ylab = "Sex",
+    auto.key = list(title = "Opinion", cex = 0.8))
```

I precedenti comandi forniscono il grafico nella Figura 1.3.3. □

1.4. L'analisi esplorativa di gruppi di variabili

Quando si analizza un gruppo di variabili quantitative si può indagare inizialmente la dipendenza fra coppie di variabili organizzando la cosiddetta matrice dei diagrammi di dispersione. Questo strumento è costituito da una matrice di grafici che rappresentano i diagrammi di dispersione per tutte le coppie di variabili del gruppo. La matrice dei diagrammi di dispersione consente di evidenziare una parte della struttura di dipendenza fra le variabili, ovvero la dipendenza per coppie di variabili. Tuttavia, la matrice dei grafici di dispersione può non rilevare caratteristiche salienti della dipendenza congiunta globale. Ad esempio, può esistere una relazione lineare perfetta fra un gruppo di variabili e non esistere nessuna dipendenza marginale fra tutte le coppie di variabili.

• **Esempio 1.4.1.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1 e in particolare le variabili `U.T1`, `U.T2`, `U.T3`, `S.T1`, `S.T2`, `S.T3`. Il comando per ottenere la matrice dei diagrammi di dispersione è il seguente:

```
> pairs(d[, 6:11], main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 1.4.1.

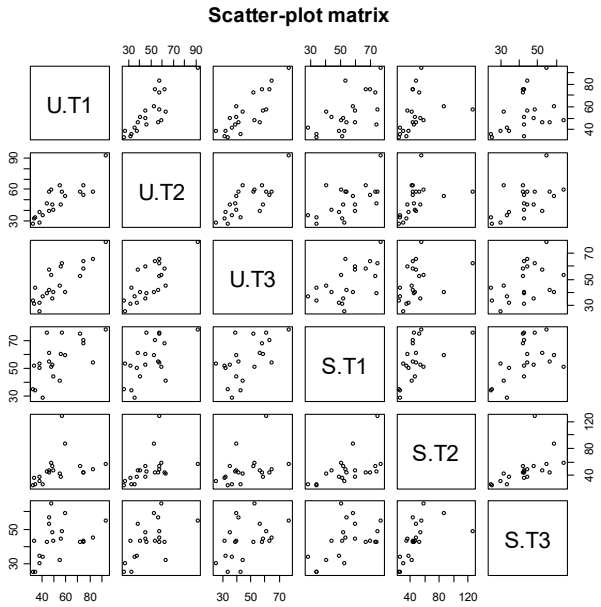


Figura 1.4.1.

Il comportamento di una ulteriore variabile qualitativa può essere analizzato introducendo nella matrice dei diagrammi di dispersione differenti colori (o simboli) dei punti per ogni livello del fattore.

Ad esempio, la variabile `Sex` può essere analizzata nella matrice dei diagrammi di dispersione mediante il seguente comando:

```
> pairs(d[, 6:11], pch = 21, bg = c("red", "blue")[as.integer(Sex)],
+     main = "Scatter-plot matrix (Red=F, Blue=M)")
```

Il precedente comando fornisce il grafico della Figura 1.4.2. □

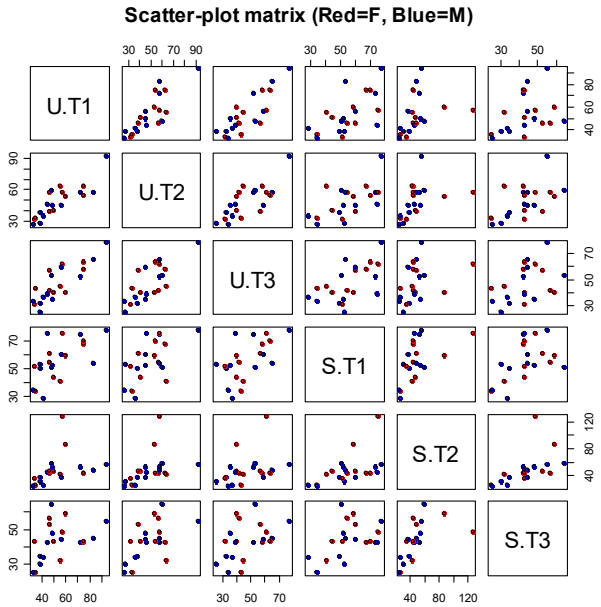


Figura 1.4.2.

Parallelamente alla matrice dei diagrammi di dispersione si considera anche la matrice di correlazione, ovvero la matrice che contiene tutti i coefficienti di correlazione fra coppie di variabili. Allo stesso modo della matrice dei diagrammi di dispersione, la matrice di correlazione non permette di analizzare in modo globale la dipendenza fra le variabili, ma offre solamente una interpretazione della dipendenza lineare per coppie di variabili.

• **Esempio 1.4.2.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1 e in particolare le variabili `U.T1`, `U.T2`, `U.T3`, `S.T1`, `S.T2`, `S.T3`. Il comando per ottenere la matrice di correlazione è il seguente:

```
> cor(d[, 6:11])
          U.T1      U.T2      U.T3      S.T1      S.T2      S.T3
U.T1 1.0000000 0.8409657 0.8357371 0.6316886 0.3348490 0.3961762
U.T2 0.8409657 1.0000000 0.7678098 0.5986291 0.4371346 0.5727865
U.T3 0.8357371 0.7678098 1.0000000 0.5879344 0.3745938 0.4432778
S.T1 0.6316886 0.5986291 0.5879344 1.0000000 0.5430833 0.5167140
S.T2 0.3348490 0.4371346 0.3745938 0.5430833 1.0000000 0.5600428
S.T3 0.3961762 0.5727865 0.4432778 0.5167140 0.5600428 1.0000000
```

Le maggiori dipendenze lineari sembrano dunque manifestarsi all'interno dei due gruppi di variabili `U.T1`, `U.T2`, `U.T3` e `S.T1`, `S.T2`, `S.T3`, come si può anche evincere dalla Figura 1.4.2. □

È possibile analizzare la dipendenza di una coppia di variabili quantitative al variare di una terza (o eventualmente di una quarta) mediante i diagrammi di dispersione condizionati. Questi grafici si

ottengono riportando una serie di diagrammi di dispersione condizionati ai vari livelli delle ulteriori variabili.

• **Esempio 1.4.3.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1 e in particolare le variabili `Smoker`, `U.T1`, `U.T2`. I comandi per ottenere i diagrammi di dispersione di `U.T1` e `U.T2` condizionati a `Smoker` sono i seguenti:

```
> library(lattice)
> xyplot(U.T2 ~ U.T1 | Smoker,
+     xlab = "Unscented first trial time (seconds)",
+     ylab = "Unscented second trial time (seconds)",
+     main = "Scatter plot conditioned to smoke")
```

I precedenti comandi forniscono il grafico della Figura 1.4.3.

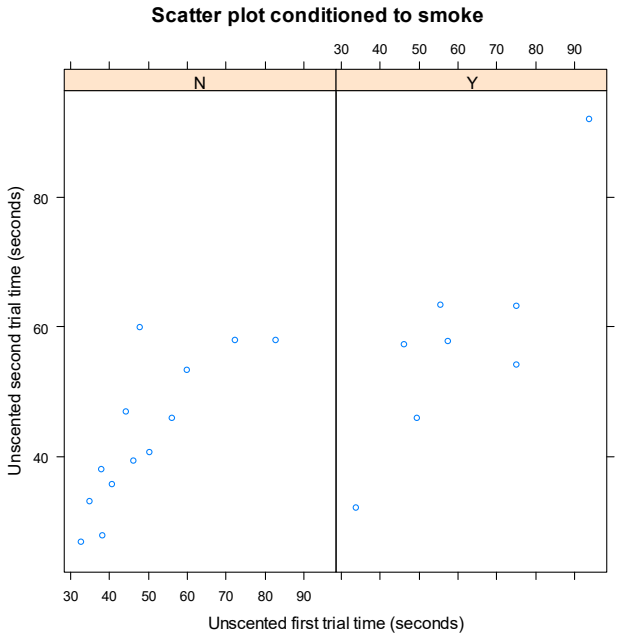


Figura 1.4.3.

Quando la variabile a cui ci si condiziona è quantitativa, allora i diagrammi di dispersione condizionati possono essere implementati suddividendo questa variabile in opportuni intervalli. Ad esempio, le osservazioni corrispondenti alla variabile `Age` possono essere posti nelle classi $[14.5, 34.5[$ e $[34.5, 65.5[$, che rappresentano rispettivamente due grossolane classi per individui giovani e più anziani. L'analisi può essere ulteriormente approfondita condizionandosi anche rispetto ad una seconda variabile ovvero la variabile `Order`. I comandi per ottenere i diagrammi di dispersione di `U.T1` e `U.T2` condizionati alle variabili `Age` e `Order` sono i seguenti:

```
> library(lattice)
> AgeClass = equal.count(Age, number = 2, overlap = 0.0)
> xyplot(U.T2 ~ U.T1 | AgeClass * Order,
+     strip = strip.custom(strip.names = T, strip.levels = T),
+     xlab = "Unscented first trial time (seconds)",
+     ylab = "Unscented second trial time (seconds)",
+     main = "Scatter plot conditioned to age and order")
```

I precedenti comandi forniscono il grafico di Figura 1.4.4. □

Figura 1.4.4.

Il concetto di tabella a doppia entrata può essere generalizzato quando si hanno tre o più variabili. In questo caso si ottengono tabelle a tre o più entrate. Le definizioni di frequenza congiunta e marginale possono essere adattate facilmente a questa struttura (anche se la notazione diviene più complessa). Per la rappresentazione di questi dati è conveniente costruire matrici di diagrammi a nastro condizionati.

• **Esempio 1.4.4.** Si considerano di nuovo i dati relativi all'Esempio 1.1.1 e in particolare le variabili `Sex`, `Opinion`, `Order`. Il comando per ottenere la tabella a tre entrate è il seguente:

```
> table(Sex, Opinion, Order)
, , Order = 1

   Opinion
Sex Ind Neg Pos
  F   0   0   4
  M   1   3   3

, , Order = 2

   Opinion
Sex Ind Neg Pos
  F   2   2   2
  M   1   2   1
```

I comandi per ottenere i diagrammi a nastri condizionati sono i seguenti:

```
> library(lattice)
> barchart(table(Sex, Opinion, Order), ylab = "Sex",
+    auto.key = list(title = "Order", cex = 0.8))
```

I precedenti comandi forniscono il grafico di Figura 1.4.5.

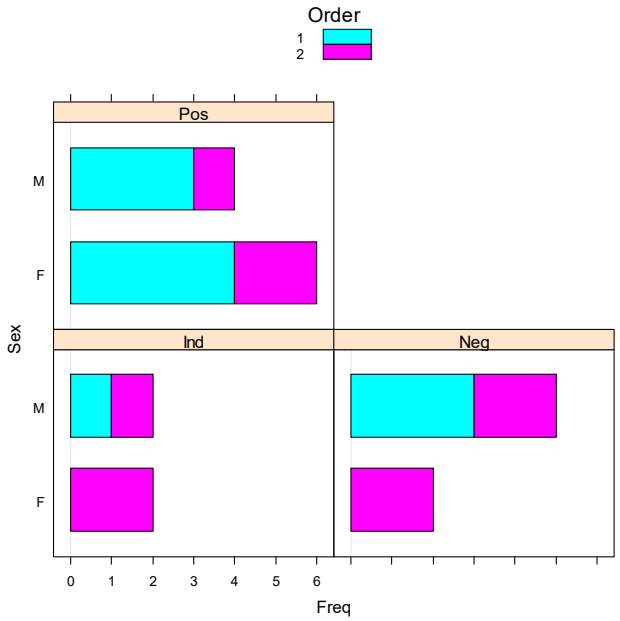


Figura 1.4.5. □

Una prima rappresentazione grafica per gruppi di variabili quantitative consiste nel cosiddetto diagramma a stelle. Nel diagramma a stelle si considera per ogni unità un grafico costituito da un insieme di raggi che partono da un medesimo punto. Il numero di raggi è pari al numero delle variabili da rappresentare. La lunghezza di ogni raggio è proporzionale al valore della corrispondente variabile sull'unità considerata.

• **Esempio 1.4.5.** Si dispone della matrice dei dati relativa alle specie e alle dimensioni dei sepali e dei petali in centimetri per alcuni fiori di iris (Fonte: Fisher, R.A., 1936, The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179-184). Questa classica matrice dei dati è compresa nel database di R e viene resa disponibile mediante i comandi:

```
> data(iris)
> attach(iris)
```

Vi sono $n = 150$ fiori di iris su cui vengono misurate $d = 5$ variabili, ovvero:

```
> names(iris)
[1]  "Sepal.Length"  "Sepal.Width"    "Petal.Length"  "Petal.Width"
"Species"
```

Le prime quattro variabili sono di tipo quantitativo e sono relative alla larghezza e alla lunghezza dei sepali e dei petali del fiore, mentre la restante è di tipo qualitativo ed è relativa alla specie di iris. Il comando per ottenere il diagramma a stelle relativo alle variabili `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` è il seguente:

```
> stars(iris[, 1:4], key.loc = c(14,-3), mar = c(4, 0, 1, 0),
+    main = "Star plot")
```

Il precedente comando fornisce il grafico della Figura 1.4.6.

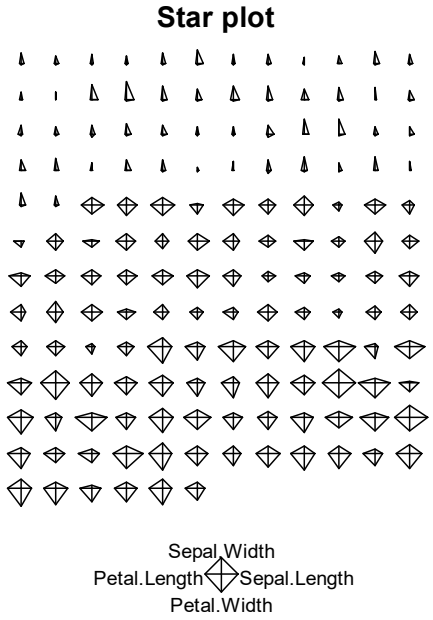


Figura 1.4.6.

Il comportamento di una ulteriore variabile qualitativa può essere analizzato introducendo nel diagramma differenti colori delle stelle per ogni livello del fattore. Ad esempio, la variabile `Species` può essere analizzata nel diagramma a stelle mediante i seguenti comandi:

```
> stars(iris[, 1:4], key.loc = c(6, -3), mar = c(7, 0, 1, 0),
+    main = "Star plot", col.stars = Species)
> legend(x = 15, y = 3, col = c("black", "red", "green"),
+    cex = 0.8, lwd = 1, bty = "n",
+    legend = c("Setosa", "Versicolor", "Virginica"))
```

I precedenti comandi forniscono il grafico della Figura 1.4.7. □

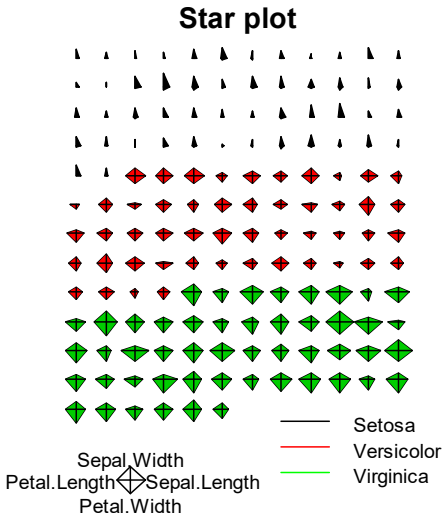


Figura 1.4.7.

Una seconda rappresentazione grafica per gruppi di variabili quantitative consiste nel cosiddetto diagramma a coordinate parallele. Nel diagramma a coordinate parallele si disegna un numero di linee parallele verticali pari al numero delle variabili da rappresentare. Le linee parallele vengono poste ad uguale distanza l'una dall'altra. Per ogni unità i valori assunti dalle variabili considerate vengono

rappresentate come una linea spezzata con i vertici sugli assi paralleli. Le posizioni dei vertici sugli assi corrispondono ai valori relativi alle singole variabili.

• **Esempio 1.4.6.** Si considerano di nuovo i dati relativi ai fiori di iris dell'Esempio 1.4.5. I comandi per ottenere il diagramma a coordinate parallele relativo alle variabili `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` sono i seguenti:

```
> library(MASS)
> parcoord(iris[,-5], main = "Parallel coordinates plot")
```

I precedenti comandi forniscono il grafico della Figura 1.4.8.

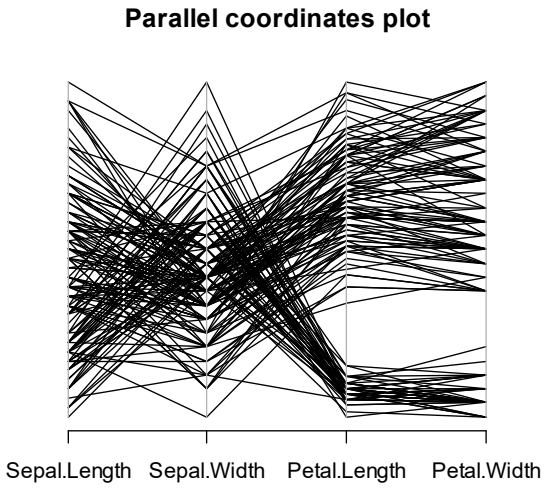


Figura 1.4.8.

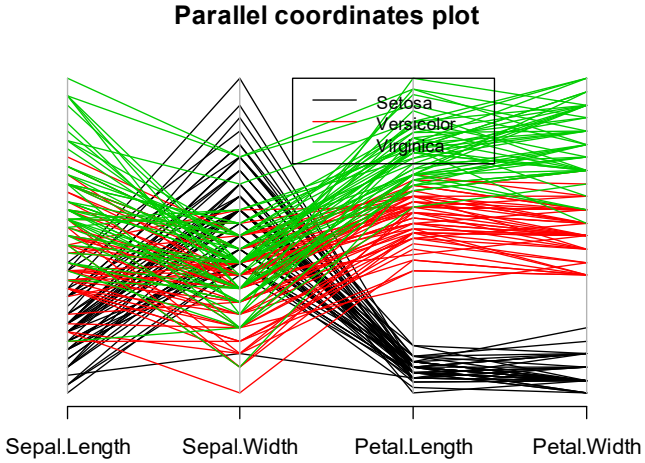


Figura 1.4.9.

Anche in questo caso, il comportamento di una ulteriore variabile qualitativa può essere analizzato introducendo nel diagramma differenti colori delle linee spezzate per ogni livello del fattore. Ad esempio, la variabile `Species` può essere analizzata nel diagramma a coordinate parallele mediante i seguenti comandi:

```
> parcoord(iris[, -5], main = "Parallel coordinates plot",
+    col = as.numeric(iris[, 5]))
> legend(x = 2.3, y = 1, bty = "o", lty = 1, col = 1:3, cex = 0.8,
+    legend = c("Setosa", "Versicolor", "Virginica"))
```

I precedenti comandi forniscono il grafico della Figura 1.4.9. □

1.5. Le rappresentazioni grafiche per serie temporali e spaziali

Se si rileva una variabile a diversi istanti temporali su una unità statistica, si ottiene la cosiddetta serie temporale. Si noti che in questo caso $n = 1$, mentre d è pari al numero di istanti temporali in cui vengono effettuate le rilevazioni. Se la variabile viene rilevata su più unità statistiche si ottengono i cosiddetti dati longitudinali o dati panel. La serie temporale viene usualmente rappresentata mediante una linea spezzata in un diagramma cartesiano. In modo analogo, i dati longitudinali vengono rappresentati come un insieme di linee spezzate in un diagramma cartesiano.

Se si rilevano invece più variabili a diversi istanti temporali su una unità statistica, si ottiene la cosiddetta serie temporale multivariata. Infine, se le variabili vengono rilevate su più unità si hanno i cosiddetti dati longitudinali multivariati. Quando il numero delle variabili è limitato, la serie temporale multivariata viene usualmente rappresentata come un insieme di linee spezzate di differenti colori, dove i colori sono relativi ad ogni variabile. I valori delle osservazioni vengono eventualmente centrati e scalati in modo opportuno al fine di aumentare la leggibilità del grafico. Alternativamente, si considera un insieme di grafici sovrapposti, ognuno dei quali è relativo alla serie temporale di una singola variabile. Simili rappresentazioni possono essere ottenute per i dati longitudinali, anche se con crescente complessità.

• **Esempio 1.5.1.** Si dispone della matrice dei dati relativa alle temperature medie annuali globali in gradi centigradi dal 1880 al 2017 fornite dalla National Oceanic and Atmospheric Administration (NOAA) (Fonte: Shumway, R.H. e Stoffer, D.S., 2017, *Time Series Analysis and its Applications*, quarta edizione, Springer, Switzerland). La matrice dei dati è compresa nel database di R e viene resa disponibile mediante la libreria:

```
> library(astsa)
```

Questa libreria contiene la serie temporale `globtemp`. La serie temporale è relativa agli scarti dalla media del periodo 1951-1980. La serie temporale `globtemp` può essere rappresentata graficamente mediante i comandi:

```
> par(mar = c(2, 2, 0, 0.5) + .5, mgp = c(1.6, 0.6, 0))
> plot(globtemp, ylab = 'Global temperature deviations (°C)',
+    type = 'n')
> grid(lty = 1, col = gray(0.9))
> lines(globtemp, type = 'o')
```

I precedenti comandi forniscono il grafico di Figura 1.5.1.

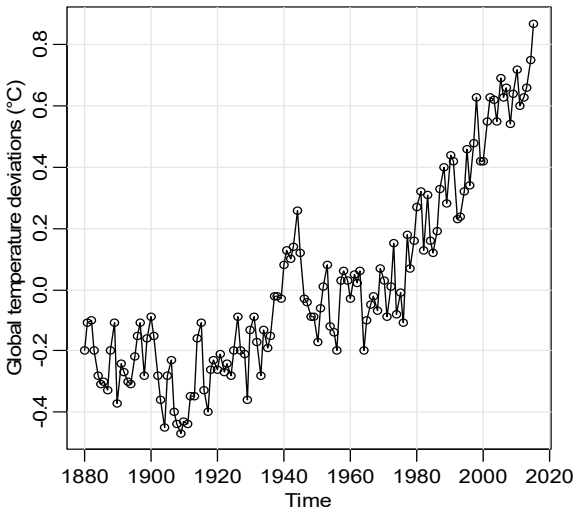


Figura 1.5.1.

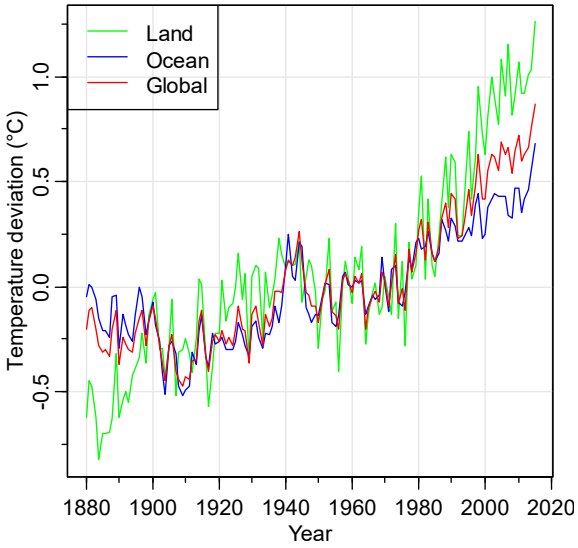


Figura 1.5.2.

Nel medesimo database sono disponibili anche le temperature medie annuali al suolo e le temperature medie annuali sull'oceano in gradi centigradi dal 1880 al 2017 fornite dalla NOAA. Le serie temporali `gtemp_land` e `gtemp_ocean` sono relative ai rispettivi scarti dalle medie del periodo 1951-1980. La serie temporale multivariata composta da `globtemp`, `gtemp_land` e `gtemp_ocean` può essere rappresentata graficamente mediante i comandi:

```
> gtemp.df = data.frame(Year = c(time(globtemp)),
+    gtemp = c(gtemp_ocean)[1:136],
+    gtempl = c(gtemp_land)[1:136], gtemp2 = c(globtemp))
> tsplot(gtemp.df[[1]], gtemp.df[[3]],
+    ylab = "Temperature deviation (°C)",
+    xlab = "Year", lwd = 1, col = "green")
> lines(gtemp.df[[1]],gtemp.df[[2]], lwd = 1, col = "blue")
> lines(gtemp.df[[1]],gtemp.df[[4]], lwd = 1, col = "red")
> legend('topleft', col = c("green", "blue", "red"),
+    lwd = 1, legend = c("Land", "Ocean", "Global"))
```

I precedenti comandi forniscono il grafico di Figura 1.5.2. □

● **Esempio 1.5.2.** Si dispone della matrice dei dati per la serie temporale multivariata relativa alle medie settimanali della mortalità globale, della mortalità per cause respiratorie, della mortalità per cause cardiocircolatorie, della temperatura, dell'umidità, del monossido di carbonio, del diossido di zolfo, del diossido di nitrogeno, degli idrocarburi, dell'ozono e del particolato rilevate nella contea di Los Angeles nel periodo 1970-1979 (Fonte: Shumway, R.H. e Stoffer, D.S., 2017, *Time Series Analysis and its Applications*, quarta edizione, Springer, Switzerland). La matrice dei dati è compresa nel database di R e viene resa disponibile mediante la libreria:

```
> library(astsa)
```

Questa libreria contiene la serie temporale multivariata `lap`. La serie temporale multivariata `lap` può essere rappresentata graficamente mediante i comandi:

```
> library(lattice)
> d <- data.frame(lap)
> tmort <- ts(d[[1]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> rmort <- ts(d[[2]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> cmort <- ts(d[[3]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> tempr <- ts(d[[4]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> rh <- ts(d[[5]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> co <- ts(d[[6]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> so2 <- ts(d[[7]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> no2 <- ts(d[[8]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> hycarb <- ts(d[[9]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> o3 <- ts(d[[10]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> part <- ts(d[[11]], start = c(1970, 1),
+    end = c(1979, 40), frequency = 52)
> xyplot(cbind(Total_mortality = tmort,
+    Respiratory_mortality = rmort,
+    Cardiovascular_mortality = cmort, Temperature = tempr,
+    Relative_humidity = rh, Carbon_monoxide = co,
+    Sulfur_dioxide = so2,
+    Nitrogen_dioxide = no2, Hydrocarbons = hycarb, Ozone = o3,
+    Particulates = part))
```

I precedenti comandi forniscono il grafico di Figura 1.5.3.

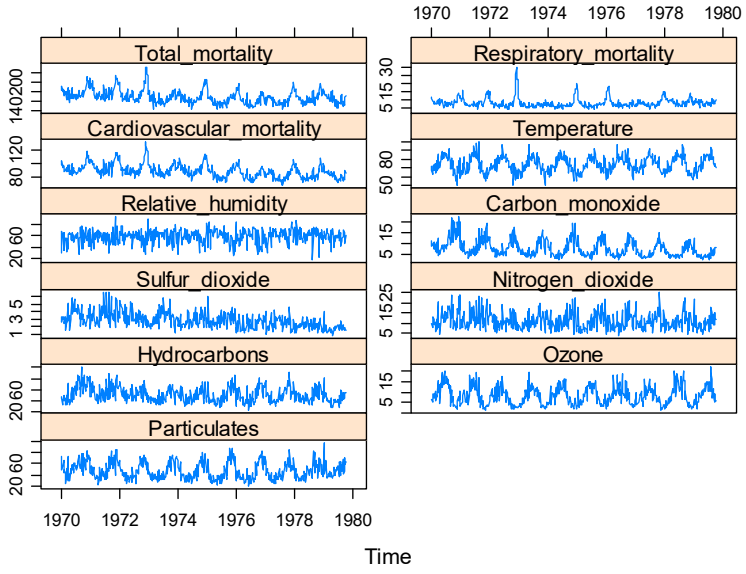


Figura 1.5.3.

Due particolari sottogruppi della serie temporale multivariata `lap` possono essere rappresentati graficamente mediante il comando:

```
> xyplot(cbind(Respiratory_mortality = rmort, Temperature = tempr,
+     Relative_humidity = rh, Particulates = part))
```

e mediante il comando:

```
> xyplot(cbind(Cardiovascular_mortality = cmort,
+     Temperature = tempr,
+     Relative_humidity = rh, Particulates = part))
```

Il primo comando fornisce il grafico della Figura 1.5.4, mentre il secondo comando fornisce il seguente grafico della Figura 1.5.5. □

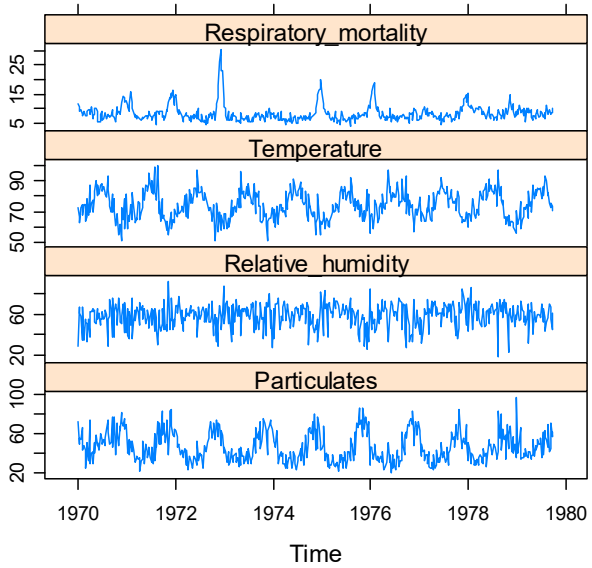


Figura 1.5.4.

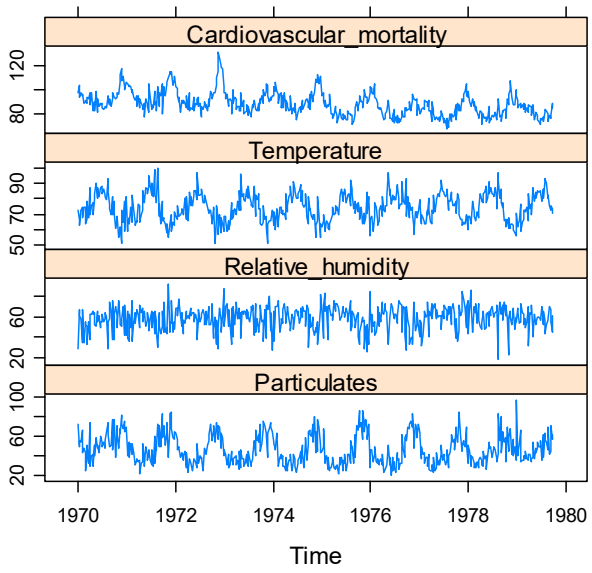


Figura 1.5.5.

In modo analogo a quanto visto per una variabile osservata nel tempo, si può rilevare una variabile in determinate posizioni spaziali in modo da ottenere la cosiddetta serie spaziale. Anche in questo caso risulta $n = 1$, mentre d è pari al numero di posizioni spaziali in cui vengono effettuate le rilevazioni. La serie spaziale può essere rappresentata mediante simboli di differente gradazione di colore o grandezza (proporzionale al valore della variabile) su un riferimento di coordinate spaziali (ad esempio latitudine e longitudine). Se invece si rilevano più variabili in determinate posizioni spaziali, si ottiene la cosiddetta serie spaziale multivariata.

• **Esempio 1.5.3.** Si dispone della matrice dei dati relativa alla serie spaziale delle temperature massime in gradi Fahrenheit rilevate giornalmente in 138 stazioni metereologiche nella parte centrale degli Stati Uniti (fra 32°N-46°N e 80°O-100°O) durante il triennio 1990-1993 (Fonte: Wikle, C.K., Zammit-Mangion, A. e Cressie, N., 2019, *Spatio-temporal Statistics with R*, Chapman&Hall/CRC, Boca Raton). La matrice dei dati è compresa nel database di R e viene resa disponibile mediante i comandi:

```
> library("STRbook")
> data("NOAA_df_1990", package = "STRbook")
```

La rappresentazione della serie spaziale relativa alle temperature massime del 30 maggio 1993 può essere ottenuta mediante i seguenti comandi:

```
> library("dplyr")
> library("ggplot2")
> Tmax <- filter(NOAA_df_1990,
+    proc == "Tmax" & month %in% 5 & year == 1993)
> Tmax$t <- Tmax$julian - 728049
> Tmax_1 <- subset(Tmax, t %in% c(30))
> ggplot(Tmax_1) +
+    geom_point(aes(x = lon, y = lat, colour = z), size = 2) +
+    col_scale(name = "°F") + xlab("Longitude") +
+    ylab("Latitude") + geom_path(data = map_data("state"),
+    aes(x = long, y = lat, group = group)) + theme_bw()
```

I precedenti comandi forniscono il grafico di Figura 1.5.6. □

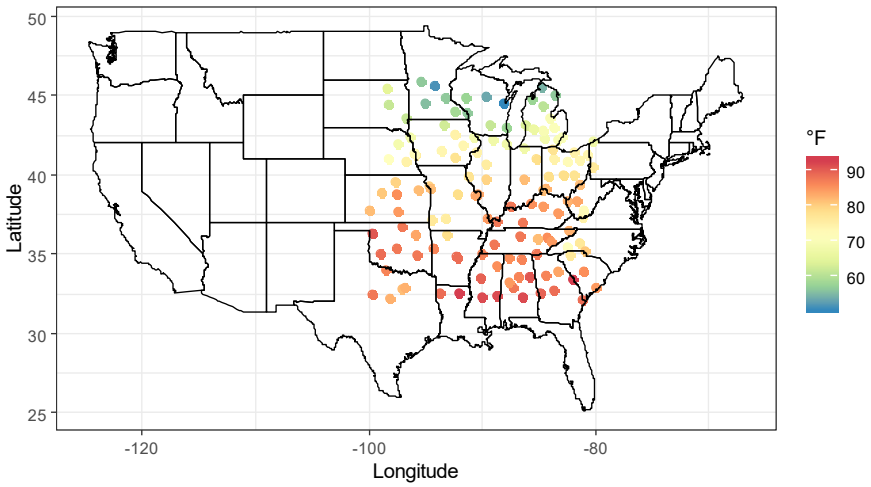


Figura 1.5.6.

Nel caso in cui la variabile assume valori su regioni geografiche, è conveniente rappresentare i dati mediante mappe coropletiche. Questo tipo di grafico è un diagramma in cui le regioni sono colorate o rappresentate con diversi schemi in modo da evidenziare i relativi valori delle variabili. Le mappe coropletiche vengono utilizzate ad esempio per rappresentare graficamente variabili quali la densità di popolazione o la distribuzione del reddito pro capite.

• **Esempio 1.5.4.** Si dispone della matrice dei dati relativa al numero di arresti ogni centomila residenti per omicidio, aggressione e stupro in ogni stato degli USA nel 1973 e alla percentuale di popolazione residente nelle aree urbane (Fonte: McNeil, D.R., 1977, *Interactive Data Analysis*, Wiley, New York). Questa classica matrice dei dati è compresa nel database di R e viene resa disponibile mediante i comandi:

```
> data(USArrests)
> attach(USArrests)
```

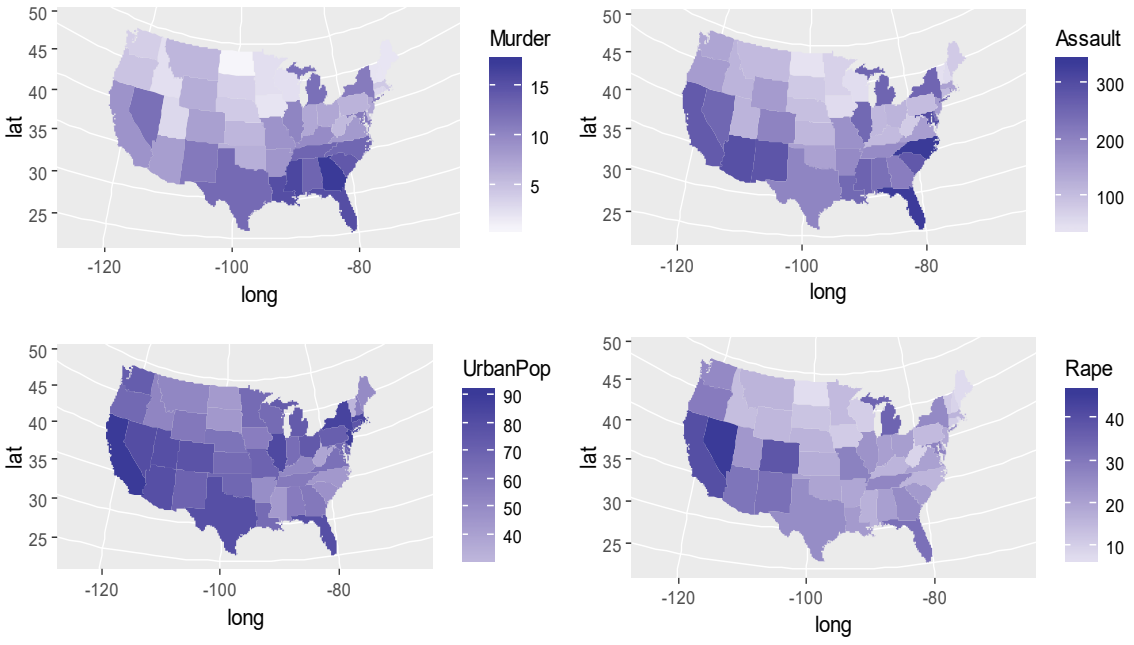


Figura 1.5.7.

Vi sono $n = 50$ stati degli USA in cui vengono misurate $d = 4$ variabili, ovvero:

```
> names(USArrests)
[1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

I comandi per ottenere le mappe coropletiche relative alle variabili `Murder`, `Assault`, `UrbanPop`, `Rape` sono i seguenti:

```
> library(maps)
> library(ggplot2)
> library(mapproj)
> library(plyr)
> states_map <- map_data("state")
> crimes <- data.frame(state = tolower(rownames(USArrests)),
+    USArrests)
> crime_map <- merge(states_map, crimes,
+    by.x = "region", by.y = "state")
> crime_map <- arrange(crime_map, group, order)
> head(crime_map)
> ggplot(crime_map, aes(x = long, y = lat, group = group,
+    fill = Murder)) + geom_polygon(colour = F) +
+    coord_map("polyconic") + scale_fill_gradient2()
> ggplot(crime_map, aes(x = long, y = lat, group = group,
+    fill = Assault)) + geom_polygon(colour = F) +
+    coord_map("polyconic") + scale_fill_gradient2()
> ggplot(crime_map, aes(x = long, y = lat, group = group,
+    fill = UrbanPop)) + geom_polygon(colour = F) +
+    coord_map("polyconic") + scale_fill_gradient2()
> ggplot(crime_map, aes(x = long, y = lat, group = group,
+    fill = Rape)) + geom_polygon(colour = F) +
+    coord_map("polyconic") + scale_fill_gradient2()
```

I precedenti comandi forniscono i grafici della Figura 1.5.7. □

Nel caso in cui una variabile viene osservata in determinate posizioni spaziali e a determinati istanti temporali, si ottiene la cosiddetta serie spazio-temporale. La serie spazio-temporale può essere rappresentata mediante una sequenza di diagrammi per serie spaziali al variare del tempo. Quando il numero di istanti temporali è elevato, può essere utile considerare opportune animazioni.

• **Esempio 1.5.5.** Si considera di nuovo la matrice dei dati relativa alla serie spaziale delle temperature massime in gradi Fahrenheit rilevate giornalmente in 138 stazioni metereologiche nella parte centrale degli Stati Uniti dell'Esempio 1.8.3. La rappresentazione della serie spazio-temporale relativa alle temperature massime dal 16 al 30 maggio 1993 può essere ottenuta mediante i seguenti comandi:

```
> Tmax_2 <- subset(Tmax, t %in% c(16:30))
> ggplot(Tmax_2) +
+    geom_point(aes(x = lon, y = lat, colour = z), size = 1) +
+    col_scale(name = "°F") + xlab("Longitude") +
+    ylab("Latitude") + geom_path(data = map_data("state"),
+    aes(x = long, y = lat, group = group)) +
+    facet_wrap(~ date, ncol = 5) + theme_bw()
```

I precedenti comandi forniscono il grafico della Figura 1.5.8. □

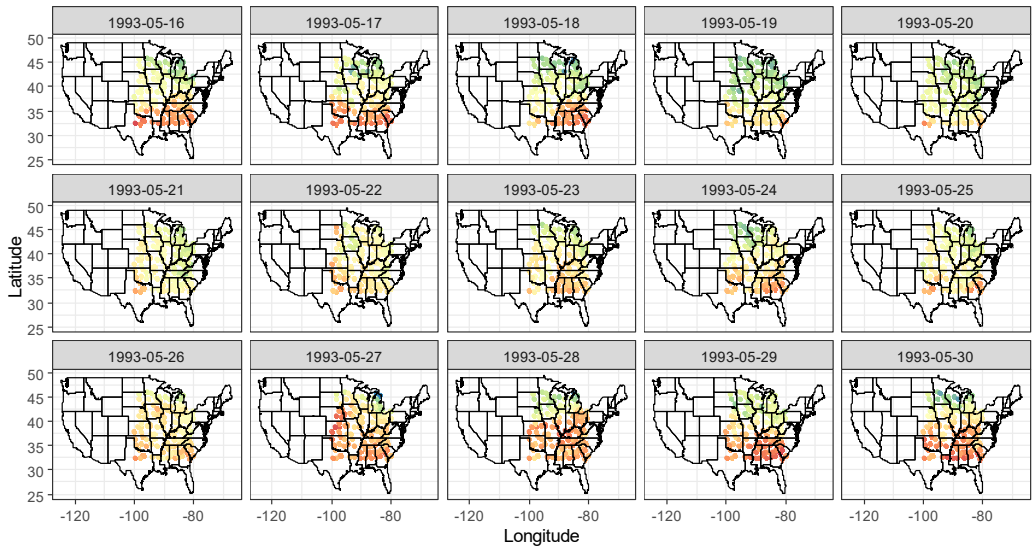


Figura 1.5.8.

● **Esempio 1.5.6.** Si dispone della matrice dei dati relativa al reddito pro capite in dollari per i residenti nelle contee dello stato del Missouri per gli anni 1970, 1980 e 1990 (Fonte: Wikle, C.K., Zammit-Mangion, A. e Cressie, N., 2019, *Spatio-temporal Statistics with R*, Chapman&Hall/CRC, Boca Raton). I redditi sono stati attualizzati rispetto all'inflazione. La matrice dei dati è compresa nel database di R e viene resa disponibile mediante i comandi:

```
> library("STRbook")
> data("BEA", package = "STRbook")
> data("MOcounties", package = "STRbook")
```

I comandi per ottenere la sequenza delle mappe coropletiche negli anni 1970, 1980 e 1990 sono i seguenti:

```
> County1 <- filter(MOcounties, NAME10 == "Clark, MO")
> MOcounties <- left_join(MOcounties, BEA, by = "NAME10")
> g1 <- ggplot(MOcounties) +
+    geom_polygon(aes(x = long, y = lat, group = NAME10,
+    fill = log(X1970))) +
+    geom_path(aes(x = long, y = lat, group = NAME10)) +
+    fill_scale(limits = c(7.5, 10.2), name = "log($)") +
+    coord_fixed() + ggtitle("1970") + theme_bw() +
+    theme(axis.title.x = element_blank(),
+    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
+    axis.title.y = element_blank(),
+    axis.text.y = element_blank(), axis.ticks.y = element_blank())
> g2 <- ggplot(MOcounties) +
+    geom_polygon(aes(x = long, y = lat, group = id,
+    fill = log(X1980))) +
+    geom_path(aes(x = long, y = lat, group = id)) +
+    scale_fill_distiller(palette = "Spectral",
+    limits = c(7.5, 10.2), name = "log($)") +
+    coord_fixed() + ggtitle("1980") + theme_bw() +
+    theme(axis.title.x = element_blank(),
+    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
+    axis.title.y = element_blank(),
+    axis.text.y = element_blank(), axis.ticks.y = element_blank())
```

```
> g3 <- ggplot(MOcounties) +
+    geom_polygon(aes(x = long, y = lat, group = id,
+    fill = log(X1990))) +
+    geom_path(aes(x = long, y = lat, group = id)) +
+    scale_fill_distiller(palette = "Spectral",
+    limits = c(7.5, 10.2),
+    name = "log($)") + coord_fixed() + ggtitle("1990") + theme_bw()
+    theme(axis.title.x = element_blank(),
+    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
+    axis.title.y = element_blank(),
+    axis.text.y = element_blank(), axis.ticks.y = element_blank())
> grid.arrange(g1, g2, g3, nrow = 1)
```

I precedenti comandi forniscono i grafici della Figura 1.5.9. □

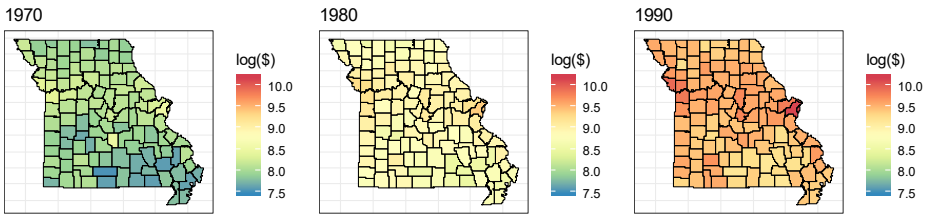


Figura 1.5.9.

1.6. Riferimenti bibliografici

Brunsdon, C. e Comber, L. (2015) *An Introduction to R for Spatial Analysis and Mapping*, Sage, Los Angeles.

Carr, D.B. e Pickle, L.W. (2010) *Visualizing Data Patterns with Micromaps*, Chapman & Hall/CRC Press, Boca Raton.

Chambers, J.M., Cleveland, W.S., Kleiner B. e Tukey, P.A. (1983) *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, Pacific Grove.

Chang, W. (2013) *R Graphics Cookbook*, O'Reilly, Sebastopol.

Cleveland, W.S. (1985) *The Elements of Graphing Data*, Wadsworth, Monterey.

Cleveland, W.S. (1993) *Visualizing Data*, Hobart Press, Summit.

Everitt, B.S. e Hothorn, T. (2010) *A Handbook of Statistical Analyses using R*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.

Healy, K. (2019) *Data Visualization*, Princeton University Press, Princeton.

Horton, N.J. e Kleinman, K. (2015) *Using R and RStudio for Data Management, Statistical Analysis, and Graphics*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.

Lamigueiro, O.P. (2014) *Displaying Time Series, Spatial, and Space-Time Data with R*, Taylor & Francis, Boca Raton.

Mittal H.V. (2011) *R Graphs Cookbook*, Packt Publishing, Birmingham.

Murrell, P. (2006) *R Graphics*, Chapman & Hall/CRC Press, Boca Raton.

Sarkar (2008) *Lattice*, Springer, New York.

Tufte, E.R. (2001) *The Visual Display of Quantitative Information*, seconda edizione, Graphics Press, Cheshire.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading.

Wickham, H. (2016) *ggplot2*, seconda edizione, Springer, New York.

Wickham, H., Çetinkaya-Rundel, M. e Grolemund, G. (2023) *R for Data Science*, seconda edizione, O'Reilly Media, Sebastopol.

Pagina intenzionalmente vuota

Capitolo 2

Le distribuzioni di probabilità

2.1. Alcuni richiami sulle variabili casuali

Dato uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, una variabile casuale X è caratterizzata dalla funzione di ripartizione F_X tale che

$$F_X(x) = P(X \leq x) \, .$$

Si noti che F_X è una funzione monotona non decrescente che assume valori in $[0, 1]$. Una variabile casuale X è detta (assolutamente) continua se F_X è una funzione (assolutamente) continua. Una variabile casuale X è detta discreta se F_X è costante a tratti con un insieme numerabile di salti. Quando il riferimento alla variabile casuale X è evidente, la funzione di ripartizione viene indicata semplicemente con F.

Una variabile casuale continua ammette una funzione di densità (che è unica a meno di insiemi con misura di Lebesgue nulla) data da

$$f_X(x) = F'_X(x) \, .$$

Risulta apparente che f_X è una funzione non negativa. Inoltre, si ha $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Infine, quando il riferimento alla variabile casuale X è evidente, la funzione di densità viene indicata con f.

Una variabile casuale discreta è caratterizzata dalla funzione di probabilità p_X tale che

$$p_X(x) = P(X = x) \, .$$

La funzione di probabilità è non nulla solo nell'insieme numerabile S di valori in cui la funzione di ripartizione effettua un salto. La funzione di probabilità p_X assume valori strettamente positivi solo se $x \in S$. Inoltre, si ha $\sum_{x \in S} p_X(x) = 1$. Infine, quando il riferimento alla variabile casuale X è evidente la funzione di probabilità viene indicata con p.

Assumendo che $\alpha \in [0, 1]$, il quantile di ordine α di una variabile casuale X è dato da

$$x_\alpha = \inf_x \{x : F_X(x) \geq \alpha\} \, .$$

Nel caso di una variabile casuale continua il quantile di ordine α risulta semplicemente $x_\alpha = F_X^{-1}(\alpha)$.

Se $r \in \mathbb{N}$, il momento di ordine r di una variabile casuale continua è dato da

$$\mu_r = \int_{-\infty}^{\infty} x^r f_X(x)dx \, ,$$

mentre il momento di ordine r di una variabile casuale discreta è dato da

$$\mu_r = \sum_{x \in S} x^r p_X(x) \, .$$

Se $r = 1$, il momento è detto media e si adotta la notazione

$$\mu = \mathrm{E}[X] \, .$$

La varianza è invece data da

$$\sigma^2 = \mathrm{Var}[X] = \mu_2 - \mu^2$$

e σ è detto scarto quadratico medio. Il coefficiente di asimmetria risulta

$$\alpha_3 = \frac{1}{\sigma^3}\,\mathrm{E}[(X-\mu)^3]\,,$$

mentre il coefficiente di curtosi è dato da

$$\alpha_4 = \frac{1}{\sigma^4}\,\mathrm{E}[(X-\mu)^4]\,.$$

A partire da una variabile casuale continua standard Z con funzione di ripartizione F_Z e funzione di densità f_Z, la famiglia di distribuzioni di posizione e scala viene generata attraverso la trasformazione lineare

$$X = \lambda + \delta Z\,.$$

Il parametro λ è detto di posizione mentre il parametro δ è detto di scala. La variabile casuale non standard X possiede funzione di ripartizione data da

$$F_X(x) = F_Z\left(\frac{x-\lambda}{\delta}\right)$$

e funzione di densità data da

$$f_X(x) = \frac{1}{\delta}\,f_Z\left(\frac{x-\lambda}{\delta}\right).$$

Se $\mathrm{E}[Z^2] < \infty$, risulta

$$\mathrm{E}[X] = \lambda + \delta\mathrm{E}[Z]$$

e

$$\mathrm{Var}[X] = \delta^2\mathrm{Var}[Z]\,.$$

In particolare, se $\mathrm{E}[Z] = 0$ e $\mathrm{Var}[Z] = 1$, i parametri di posizione e di scala coincidono rispettivamente con la media e lo scarto quadratico medio. I parametri rimanenti di una distribuzione sono detti parametri di forma e vengono eventualmente indicati nel seguito con p e q.

2.2. Alcune variabili casuali continue

La variabile casuale Normale standard Z ammette funzione di densità $f_Z = \phi$, dove

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}z^2}\,.$$

La funzione di ripartizione di Z non è esprimibile in forma analitica e viene indicata come

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}u^2}du\,.$$

Risulta $\mathrm{E}[Z] = 0$ e $\mathrm{Var}[Z] = 1$, mentre si ha $\alpha_3 = 0$ e $\alpha_4 = 3$. Per la variabile casuale Normale non standard X i parametri di posizione e di scala coincidono dunque con la media μ e con lo scarto quadratico medio σ. Per indicare che Z ha distribuzione Normale standard si adotta la notazione

$Z \sim N(0,1)$, mentre se X ha distribuzione Normale non standard si scrive $X \sim N(\mu, \sigma^2)$. Infine, il quantile di ordine α della variabile casuale Normale standard viene indicato con $z_\alpha = \Phi^{-1}(\alpha)$. I grafici della funzione di densità e di ripartizione di Z sono riportati nella Figura 2.2.1.

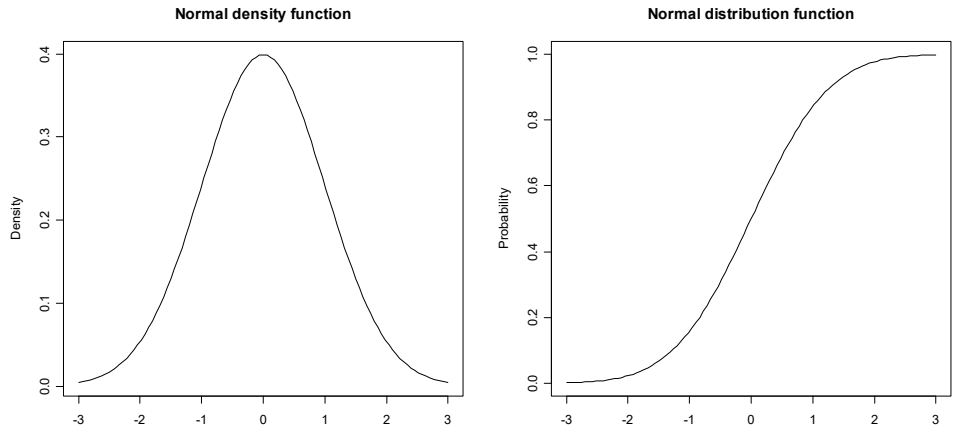


Figura 2.2.1.

La variabile casuale Uniforme standard Z ammette funzione di densità

$$f_Z(z) = \mathbf{1}_{[0,1]}(z)\,,$$

dove $\mathbf{1}_S(x)$ rappresenta la funzione indicatrice dell'insieme S, ovvero $\mathbf{1}_S(x) = 1$ se $x \in S$ e $\mathbf{1}_S(x) = 0$ altrimenti. Risulta $\mathrm{E}[Z] = 1/2$ e $\mathrm{Var}[Z] = 1/12$, mentre $\alpha_3 = 0$ e $\alpha_4 = 9/5$. Per indicare che Z ha distribuzione Uniforme standard si adotta la notazione $Z \sim U(0,1)$, mentre se X ha distribuzione Uniforme non standard si scrive $X \sim U(\lambda, \lambda + \delta)$. La parametrizzazione in termini di λ e $\lambda + \delta$ si usa per evidenziare che la variabile casuale non standard X assume valori in $[\lambda, \lambda + \delta]$ con probabilità 1. I grafici della funzione di densità e di ripartizione di Z sono riportati nella Figura 2.2.2.

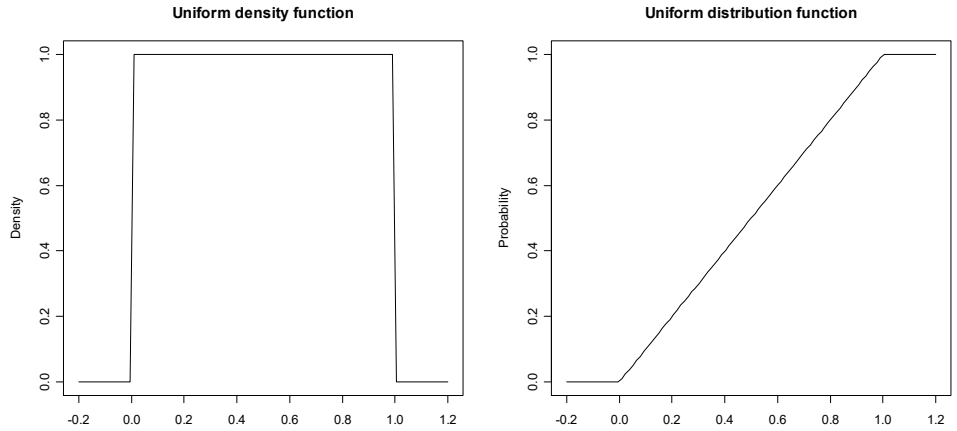


Figura 2.2.2.

La variabile casuale Gamma standard Z ammette funzione di densità

$$f_Z(z) = \frac{1}{\Gamma(p)}\, z^{p-1} e^{-z}\, \mathbf{1}_{[0,\infty[}(z)\,,$$

dove p è un parametro di forma, mentre

$$\Gamma(p) = \int_0^\infty u^{p-1} e^{-u} du$$

è la funzione Gamma di Eulero. Quando $p = 1$ la variabile casuale Z è detta Esponenziale standard. Si ha inoltre $\mathrm{E}[Z] = p$ e $\mathrm{Var}[Z] = p$, mentre $\alpha_3 = 2/\sqrt{p}$ e $\alpha_4 = 3 + 6/p$. Per indicare che Z ha distribuzione Gamma standard con parametro di forma p si adopera la notazione $Z \sim G(0, 1; p)$, mentre se X ha distribuzione Gamma non standard si scrive $X \sim G(\lambda, \delta; p)$. Per indicare che Z ha distribuzione Esponenziale standard si adotta la notazione $Z \sim E(0, 1)$, mentre se X ha distribuzione Esponenziale non standard si scrive $X \sim E(\lambda, \sigma)$, dal momento che il parametro di scala coincide con lo scarto quadratico medio. I grafici della funzione di densità e di ripartizione di Z per $p = 1, 2, 3$ sono riportati nella Figura 2.2.3.

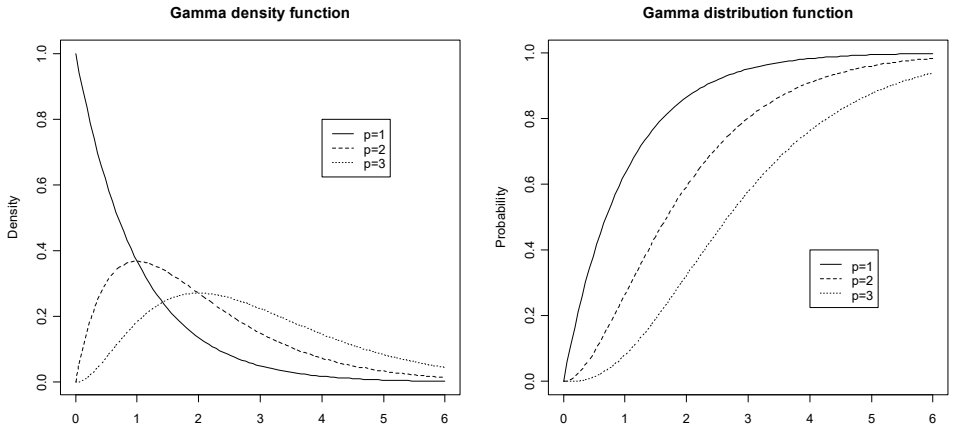


Figura 2.2.3.

La variabile casuale Beta standard Z ammette funzione di densità

$$f_Z(z) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} z^{p-1}(1-z)^{q-1} \mathbf{1}_{[0,1]}(z) ,$$

dove p e q sono parametri di forma. Risulta

$$\mathrm{E}[Z] = \frac{p}{p+q}$$

e

$$\mathrm{Var}[Z] = \frac{pq}{(p+q)^2(p+q+1)} ,$$

mentre

$$\alpha_3 = \frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$$

e

$$\alpha_4 = \frac{3(p+q+1)(2(p+q)^2 + pq(p+q-6))}{pq(p+q+2)(p+q+3)} .$$

Per indicare che Z ha distribuzione Beta standard con parametri di forma p e q si adotta la notazione $Z \sim Be(0,1;p,q)$, mentre se X ha distribuzione Beta non standard si scrive $X \sim Be(\lambda, \lambda+\delta; p, q)$. La parametrizzazione in termini di λ e $\lambda+\delta$ viene impiegata per evidenziare che la variabile casuale non standard X assume valori in $[\lambda, \lambda+\delta]$ con probabilità 1. I grafici della funzione di densità e di ripartizione di Z per $(p,q) = (1.3, 0.7), (0.3, 0.3), (0.7, 1.3)$ sono riportati nella Figura 2.2.4, mentre i grafici della funzione di densità e di ripartizione di Z per $(p,q) = (4,2), (2,2), (2,4)$ sono riportati nella Figura 2.2.5.

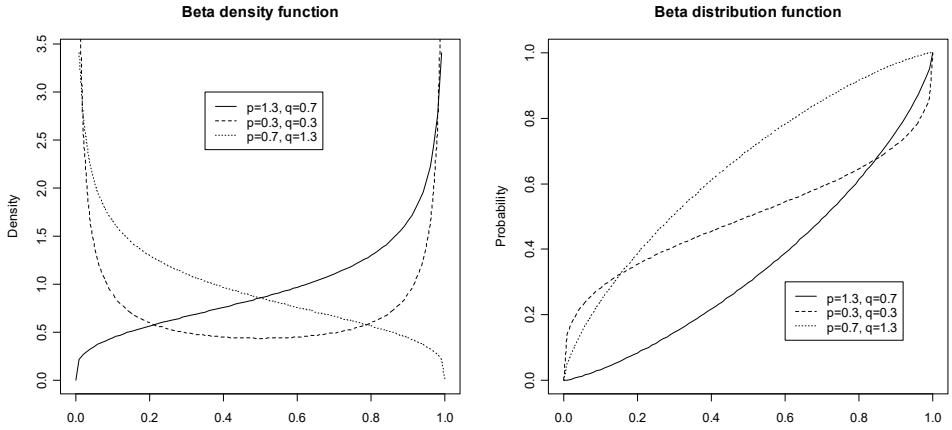


Figura 2.2.4.

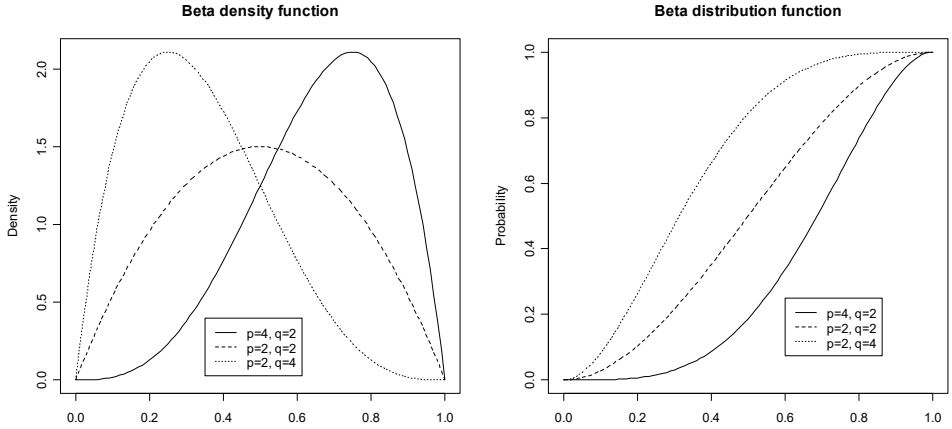


Figura 2.2.5.

La variabile casuale di Cauchy standard Z ammette funzione di densità

$$f_Z(z) = \frac{1}{\pi(1+z^2)} \, .$$

La variabile casuale di Cauchy standard non possiede momenti di alcun ordine. Per indicare che Z ha distribuzione di Cauchy standard si adotta la notazione $Z \sim C(0,1)$, mentre se X ha distribuzione di Cauchy non standard si scrive $X \sim C(\lambda, \delta)$. I grafici della funzione di densità e di ripartizione di Z sono riportati nella Figura 2.2.6.

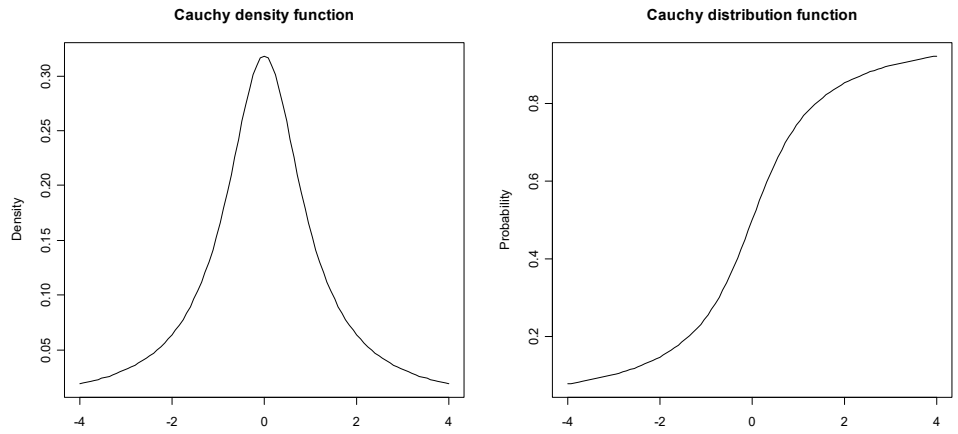


Figura 2.2.6.

2.3. Alcune variabili casuali discrete

La variabile casuale Binomiale Z è caratterizzata dalla funzione di probabilità

$$p_Z(z) = \binom{n}{z} p^z (1-p)^{n-z} \, \mathbf{1}_{\{0,1,\ldots,n\}}(z) \,,$$

dove $p \in]0,1[$ e $n \in \mathbb{N}$. Per la variabile casuale Binomiale risulta $\mathrm{E}[Z] = np$ e $\mathrm{Var}[Z] = np(1-p)$, mentre

$$\alpha_3 = \frac{1-2p}{\sqrt{np(1-p)}}$$

e

$$\alpha_4 = 3 + \frac{1-6p(1-p)}{np(1-p)} \,.$$

Per indicare che Z ha distribuzione Binomiale si adotta la notazione $Z \sim Bi(n,p)$. I grafici della funzione di probabilità di Z per $(n,p) = (10,0.3), (10,0.5)$ sono riportati nella Figura 2.3.1.

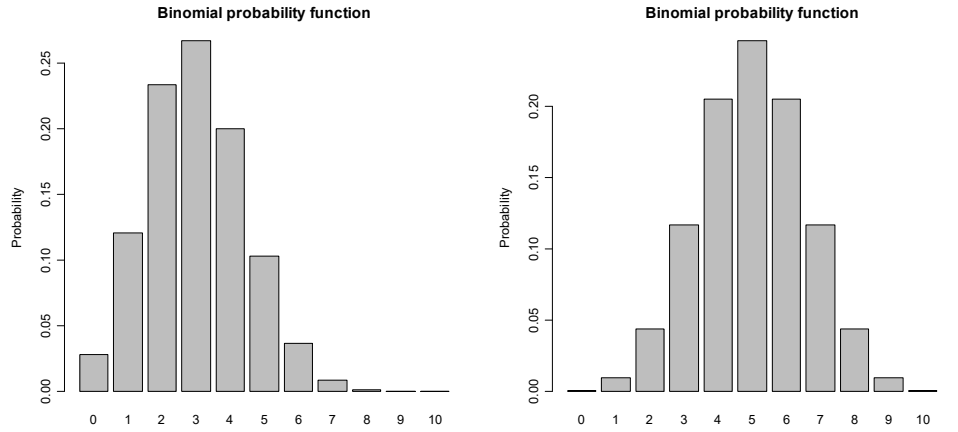


Figura 2.3.1.

La variabile casuale di Poisson Z è caratterizzata dalla funzione di probabilità

$$p_Z(z) = e^{-\mu}\,\frac{\mu^z}{z!}\,\mathbf{1}_{\{0,1,\ldots\}}(z)\,,$$

dove μ è un parametro positivo. Per la variabile casuale di Poisson si ha $\mathrm{E}[Z] = \mu$ e $\mathrm{Var}[Z] = \mu$, mentre $\alpha_3 = 1/\sqrt{\mu}$ e $\alpha_4 = 3 + 1/\mu$. Per indicare che Z ha distribuzione di Poisson si adotta la notazione $Z \sim Po(\mu)$. I grafici della funzione di probabilità di Z per $\mu = 2, 4$ sono riportati nella Figura 2.3.2.

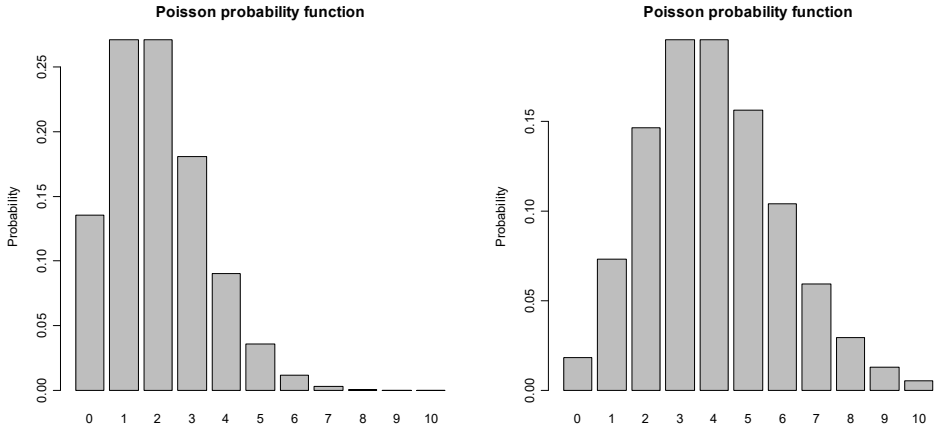


Figura 2.3.2.

La variabile casuale Ipergeometrica Z è caratterizzata dalla funzione di probabilità

$$p_Z(z) = \frac{\binom{D}{z}\binom{N-D}{n-z}}{\binom{N}{z}}\,\mathbf{1}_S(z)\,,$$

dove $S = \{\max(0, n - N + D), \ldots, \min(n, D)\}$ e $n, D, N \in \mathbb{N}$ sono tali che $n, D \leq N$. Posto

$$p = \frac{D}{N}$$

e

$$c = \frac{n(N-n)}{N-1}\,,$$

per la variabile casuale Ipergeometrica si ha $\mathrm{E}[Z] = np$ e $\mathrm{Var}[Z] = cp(1-p)$, mentre

$$\alpha_3 = \frac{N-2n}{N-2}\,\frac{1-2p}{\sqrt{cp(1-p)}}$$

e

$$\alpha_4 = 3 + \frac{N(N+1) - 6N^2p(1-p) - 6(N-1)c}{(N-2)(N-3)cp(1-p)} + \frac{6(5N-6)}{(N-2)(N-3)}\,.$$

Per indicare che Z ha distribuzione Ipergeometrica si adotta la notazione $Z \sim I(n, D, N)$. I grafici della funzione di probabilità di Z per $(n, D, N) = (10, 30, 100), (10, 50, 100)$ sono riportati nella Figura 2.3.3.

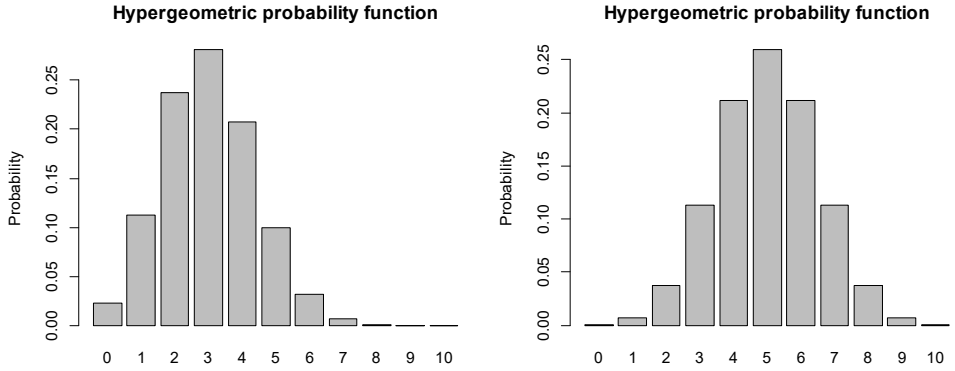


Figura 2.3.3.

2.4. Alcune trasformate notevoli di variabili casuali

Se $Z_1, \ldots, Z_n$ sono variabili casuali indipendenti tali che $Z_i \sim N(0,1)$, la trasformata

$$U = \sum_{i=1}^{n} Z_i^2$$

è detta variabile casuale Chi-quadrato con n gradi di libertà. La variabile casuale Chi-quadrato U ammette funzione di densità data da

$$f_U(u) = \frac{1}{2^{\frac{1}{2}n}\Gamma(\frac{n}{2})}\, u^{\frac{n}{2}-1} e^{-\frac{1}{2}u}\, \mathbf{1}_{[0,\infty[}(z)\,.$$

Risulta $U \sim G(0, 2; n/2)$ e quindi $\mathrm{E}[U] = n$ e $\mathrm{Var}[U] = 2n$, mentre $\alpha_3 = \sqrt{8/n}$ e $\alpha_4 = 3 + 12/n$. Per indicare che U ha distribuzione Chi-quadrato con n gradi di libertà si adotta la notazione $U \sim \chi^2_n$. Inoltre, il quantile di ordine α della Chi-quadrato con n gradi di libertà viene indicato con $\chi^2_{n,\alpha}$. I grafici della funzione di densità di U per i valori di $n = 2, 3, 4$ sono riportati nella Figura 2.4.1.

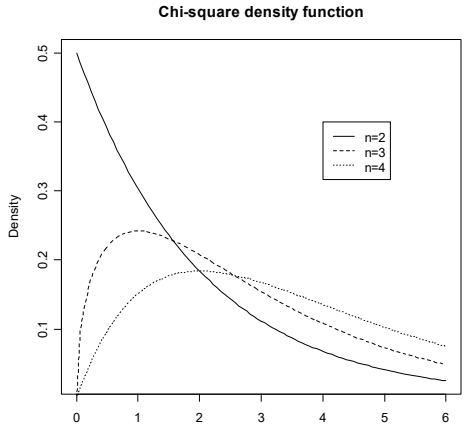


Figura 2.4.1.

Se $Z \sim N(0,1)$ e $U \sim \chi^2_n$ sono variabili casuali indipendenti, la trasformata

$$T = \frac{Z}{\sqrt{U/n}}$$

è detta variabile casuale t di Student con n gradi di libertà. La variabile casuale t di Student T ammette funzione di densità

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}(n+1)} .$$

Inoltre, risulta $\mathrm{E}[T] = 0$ per $n > 1$ e $\mathrm{Var}[T] = n/(n-2)$ per $n > 2$, mentre $\alpha_3 = 0$ per $n > 3$ e $\alpha_4 = 3 + 6/(n-4)$ per $n > 4$. Per indicare che T ha distribuzione t di Student con n gradi di libertà si adotta la notazione $T \sim t_n$. Il quantile di ordine α della t di Student con n gradi di libertà viene indicato con $t_{n,\alpha}$. I grafici della funzione di densità di T per i valori $n = 1, 3, 10$ sono riportati nella Figura 2.4.2.

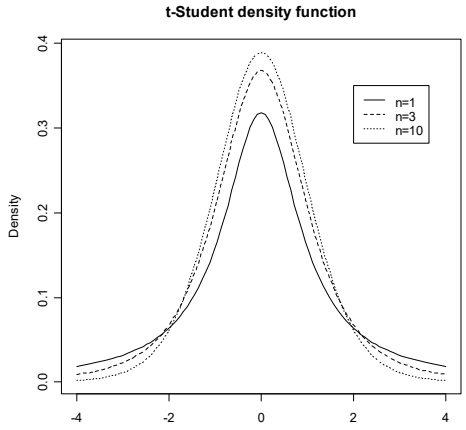


Figura 2.4.2.

Se $U \sim \chi^2_{n_1}$ e $V \sim \chi^2_{n_2}$ sono variabili casuali indipendenti, la trasformata

$$F = \frac{U/n_1}{V/n_2}$$

è detta variabile casuale F di Snedecor con n_1 e n_2 gradi di libertà. La variabile casuale F di Snedecor ammette funzione di densità data da

$$f_F(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{1}{2}n_1} x^{\frac{1}{2}n_1 - 1} \left(1 + \frac{n_1}{n_2}\, x\right)^{-\frac{1}{2}(n_1+n_2)} \mathbf{1}_{]0,\infty[}(x) .$$

Si ha

$$\mathrm{E}[F] = \frac{n_2}{n_2 - 2} \, , \, n_2 > 2 \, ,$$

e

$$\mathrm{Var}[F] = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2-2)^2(n_2-4)} \, , \, n_2 > 4 \, ,$$

mentre

$$\alpha_3 = \frac{(2n_1 + n_2 - 2)\sqrt{8(n_2-4)}}{(n_2-6)\sqrt{n_1(n_1+n_2-2)}} \, , \, n_2 > 6 \, ,$$

e

$$\alpha_4 = 3 + \frac{12(5n_2 - 22)}{(n_2 - 6)(n_2 - 8)} + \frac{12(n_2 - 2)^2(n_2 - 4)}{n_1(n_2 - 6)(n_2 - 8)(n_1 + n_2 - 2)} \text{ , } n_2 > 8 \text{ .}$$

Per indicare che la variabile casuale F ha distribuzione F di Snedecor con n_1 e n_2 gradi di libertà si adotta la notazione $F \sim F_{n_1,n_2}$. Infine, il quantile di ordine α della F di Snedecor con n_1 e n_2 gradi di libertà viene indicato con $F_{n_1,n_2,\alpha}$. I grafici della funzione di densità di F per $(n_1, n_2) = (4, 4)$, $(12, 12)$ sono riportati nella Figura 2.4.3.

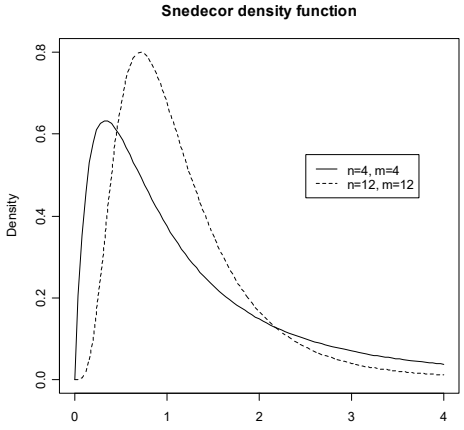


Figura 2.4.3.

2.5. Alcuni richiami sui vettori di variabili casuali

Il concetto di variabile casuale può essere esteso al caso multivariato. Un vettore di variabili casuali (assolutamente) continue $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ ammette una funzione di densità congiunta $f_{\boldsymbol{X}}$ (definita a meno di insiemi di misura di Lebesgue nulla su $\mathbb{R}^d$). Un vettore di variabili casuali discrete $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ è invece caratterizzato da una funzione di probabilità congiunta $p_{\boldsymbol{X}}$.

Il vettore medio è dato da $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ dove $\mu_j = \mathrm{E}[X_j]$. Inoltre, la matrice di varianza-covarianza è data da $\boldsymbol{\Sigma} = (\sigma_{jl})$, dove il j-esimo elemento diagonale è la varianza di X_j, ovvero $\sigma_j^2 = \mathrm{Var}[X_j]$, mentre il generico elemento di posto (j, l) è la covarianza di X_j e X_l, ovvero

$$\sigma_{jl} = \mathrm{Cov}[X_j, X_l] = \mathrm{E}[(X_j - \mu_j)(X_l - \mu_l)] \text{ .}$$

Il vettore di variabili casuali $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ è detto Normale multivariato se ammette funzione di densità congiunta data da

$$f_{\boldsymbol{X}}(\mathbf{x}) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \text{ ,}$$

dove $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$, $\boldsymbol{\mu}$ è il vettore medio e $\boldsymbol{\Sigma}$ è la matrice di varianza-covarianza. Per indicare che il vettore di variabili casuali $\boldsymbol{X}$ ha distribuzione Normale multivariata si adotta la notazione $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Le Figure 2.5.1, 2.5.2, 2.5.3 riportano il grafico della funzione di densità (con relativo grafico di contorno) di un vettore Normale multivariato per $d = 2$, rispettivamente con $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ e $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.

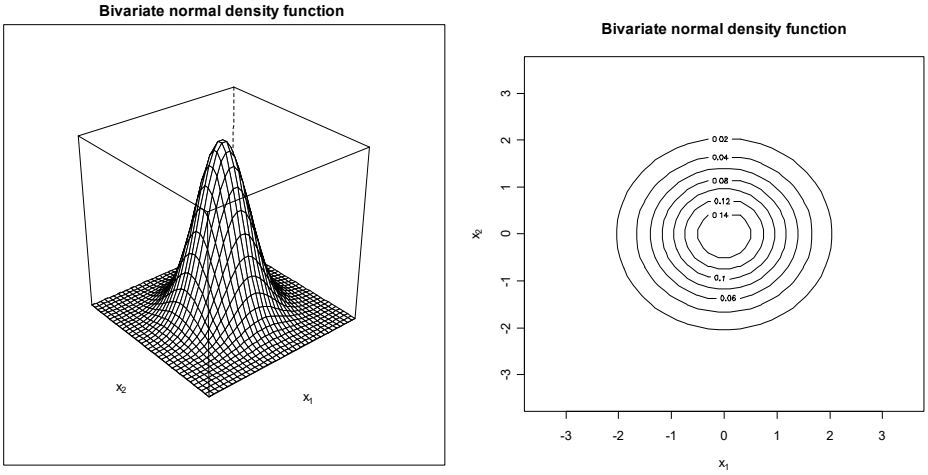


Figura 2.5.1.

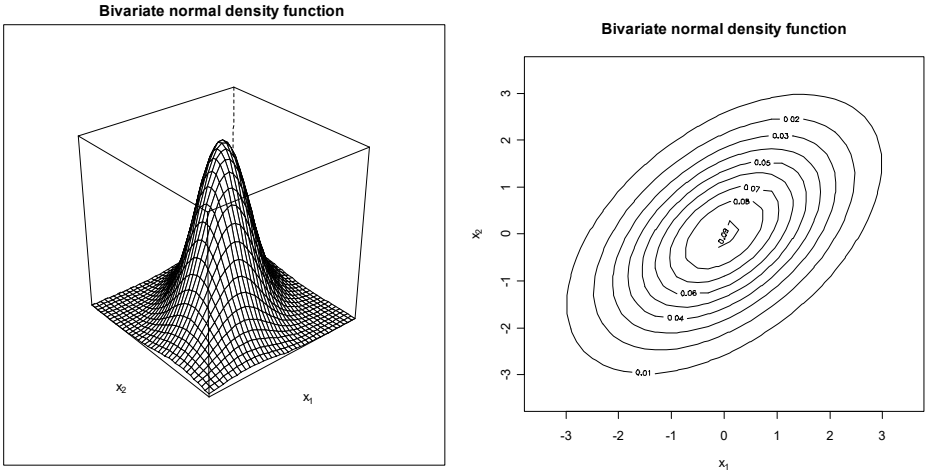


Figura 2.5.2.

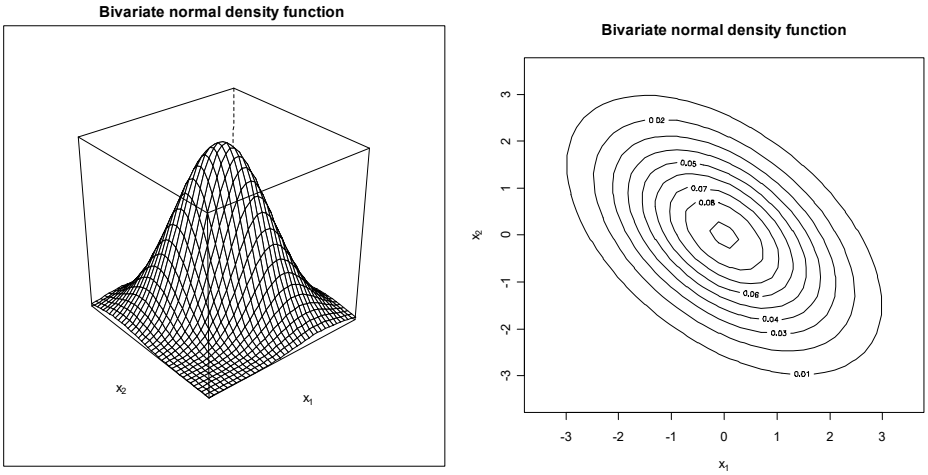


Figura 2.5.3.

Il vettore di variabili casuali $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ è detto Multinomiale se è caratterizzato dalla funzione di probabilità congiunta

$$p_{\boldsymbol{X}}(x_1, \ldots, x_d) = \binom{n}{x_1 \ldots x_d} \prod_{j=1}^{d} \pi_j^{x_j} \, \mathbf{1}_S(x_1, \ldots, x_d) \,,$$

con $\pi_1, \ldots, \pi_d > 0$ e $\sum_{j=1}^{d} \pi_j = 1$, mentre

$$\binom{n}{x_1 \ldots x_d} = \frac{n!}{\prod_{j=1}^{d} x_j!}$$

è il coefficiente multinomiale e

$$S = \{(x_1, \ldots, x_d) : x_j \in \{0, 1, \ldots, n\}, \sum_{j=1}^{d} x_j = n\} .$$

Se si pone $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)^{\mathrm{T}}$, $\boldsymbol{\mu} = n\boldsymbol{\pi}$ è il vettore medio e $\boldsymbol{\Sigma} = n(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}})$ è la matrice di varianza-covarianza. Per indicare che il vettore di variabili casuali $\boldsymbol{X}$ ha distribuzione Multinomiale si adotta la notazione $\boldsymbol{X} \sim M_d(n, \boldsymbol{\pi})$.

2.6. Riferimenti bibliografici

Balakrishnan, N. e Nevzorov, V.B. (2003) *A Primer of Statistical Distributions*, Wiley, New York.

Billingsley, P. (1995) *Probability and Measure*, terza edizione, Wiley, New York.

Fang, K.T., Kotz, S. and Ng, K.W. (1990) *Symmetric Multivariate and Related Distributions*, Springer, New York.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, volume I, terza edizione, Wiley, New York.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, volume II, seconda edizione, Wiley, New York.

Forbes, C., Evans, M., Hastings, N. e Peacock, B. (2011) *Statistical Distributions*, quarta edizione, Wiley, New York.

Gut, A. (2013) *Probability: a Graduate Course*, Springer, New York.

Johnson, N., Kemp, A. e Kotz, S. (2005) *Univariate Discrete Distributions*, terza edizione, Wiley, New York.

Johnson, N. e Kotz, S. (1977) *Urn Models and their Applications*, New York, Wiley.

Johnson, N., Kotz, S. e Balakrishnan, N. (1994) *Continuous Univariate Distributions*, volume 1, seconda edizione, Wiley, New York.

Johnson, N., Kotz, S. e Balakrishnan, N. (1995) *Continuous Univariate Distributions*, volume 2, seconda edizione, Wiley, New York.

Johnson, N., Kotz, S. e Balakrishnan, N. (1997) *Discrete Multivariate Distributions*, New York, Wiley.

Kotz, S., Balakrishnan, N. e Johnson, N. (2000) *Continuous Multivariate Distributions*, volume 1, seconda edizione, Wiley, New York.

Venkatesh, S.S. (2013) *The Theory of Probability*, Cambridge University Press, Cambridge.

Capitolo 3

Il campionamento

3.1. Il modello statistico

La matrice dei dati (o una sua parte) può essere pensata come la realizzazione di un esperimento casuale. In questo caso le colonne di **D** (o alcune sue colonne) sono delle variabili casuali a priori della rilevazione. L'insieme di queste variabili casuali è detto campione, mentre n è detta numerosità campionaria. Se le osservazioni su ogni unità vengono ottenute nelle medesime condizioni sperimentali e se il campionamento è effettuato in modo da assicurare l'indipendenza delle osservazioni fra le unità, il campione è detto casuale.

L'insieme delle distribuzioni di probabilità congiunte ammissibili per il campione delimita una classe $\mathcal{M}$ detta modello statistico. Il modello statistico è detto classico se la morfologia funzionale della distribuzione congiunta è completamente specificata a meno di un insieme di parametri non noti. Il modello statistico è invece detto "distribution-free" se la morfologia funzionale della distribuzione congiunta non è specificata. In modo improprio, spesso un modello statistico classico è detto parametrico, mentre un modello statistico "distribution-free" è detto non parametrico. Questa terminologia è fuorviante, dal momento che in entrambi i casi nella specificazione del modello sono presenti comunque dei parametri.

L'insieme Θ dei valori plausibili per i parametri del modello è detto spazio parametrico e in generale non è uno spazio cartesiano, ma eventualmente anche uno spazio funzionale. In ogni caso, l'obiettivo dell'inferenza statistica si riduce a fare affermazioni sui "veri valori" dei parametri presenti nella specificazione del modello.

• **Esempio 3.1.1.** Il modello statistico più semplice assume una sola variabile, ovvero $d = 1$, e un campione casuale. In questo caso, $x_1, \ldots, x_n$ sono le realizzazioni di n copie indipendenti $X_1, \ldots, X_n$ di una variabile casuale X. In questa situazione statistica, il tipico modello classico assume che $X \sim N(\mu, \sigma^2)$ e quindi la distribuzione congiunta del campione è la fattorizzazione di distribuzioni marginali della stessa forma specificate a meno dei parametri μ e σ. Se $F_n = F_{X_1,\ldots,X_n}$ rappresenta la funzione di ripartizione congiunta del campione, il corrispondente modello statistico è dato da

$$\mathcal{M}_{\mu,\sigma} = \{F_n : F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} \Phi\Big(\frac{x_i - \mu}{\sigma}\Big), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\} .$$

In questo caso, il modello è indicizzato dai parametri μ e σ e lo spazio parametrico è dato da $\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$. Nella medesima situazione, un modello "distribution-free" assume semplicemente che X sia una variabile casuale continua con funzione di ripartizione $F_X(x) = G(x - \lambda)$ tale che $G \in \mathcal{R}_0$. In questo caso, $\mathcal{R}_0$ rappresenta la classe delle funzioni di ripartizione di una variabile casuale continua con mediana pari a 0, ovvero $G(0) = 1/2$. Dunque, la mediana di X è pari a λ. Il relativo modello statistico è dato da

$$\mathcal{M}_{\lambda,G} = \{F_n : F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} G(x_i - \lambda), \lambda \in \mathbb{R}, G \in \mathcal{R}_0\} .$$

Dunque, λ e G sono i "parametri" del modello e con un lieve abuso in notazione lo spazio parametrico è dato da $\Theta = \{(\mu, G) : \lambda \in \mathbb{R}, G \in \mathcal{R}_0\}$. □

• **Esempio 3.1.2.** Nella sua struttura più semplice il modello statistico di regressione assume che vi siano due variabili, ovvero $d = 2$, di cui una sotto controllo dello sperimentatore (detta regressore) e l'altra di risposta. Se $x_1, \ldots, x_n$ rappresentano i valori del regressore le unità statistiche, queste quantità vengono considerate fissate dallo sperimentatore. Per quanto riguarda invece le osservazioni relative alla variabile di risposta, ovvero $y_1, \ldots, y_n$, queste vengono considerate come realizzazioni delle variabili casuali $Y_1, \ldots, Y_n$ e tali che

$$Y_i = m(x_i) + \mathcal{E}_i \,,$$

dove m è la cosiddetta funzione di regressione, mentre $\mathcal{E}_1, \ldots, \mathcal{E}_n$ sono variabili casuali indipendenti dette errori tali che $\mathrm{E}[\mathcal{E}_i] = 0$ e $\mathrm{Var}[\mathcal{E}_i] = \sigma^2$. La formulazione alternativa del modello di regressione è quindi data dalle relazioni $\mathrm{E}[Y_i] = m(x_i)$ e $\mathrm{Var}[Y_i] = \sigma^2$. Evidentemente, in questo caso il campione non è casuale. Il modello di regressione lineare assume che

$$m(x_i) = \beta_0 + \beta_1 x_i \,,$$

ovvero la parte strutturale del modello viene specificata a meno di due parametri. In un approccio classico, il modello lineare viene completato con l'assunzione distribuzionale che richiede $\mathcal{E}_i \sim N(0, \sigma^2)$, ovvero che $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Se $F_n = F_{Y_1,\ldots,Y_n}$ rappresenta la funzione di ripartizione congiunta del campione, il modello statistico è dunque dato da

$$\mathcal{M}_{\beta_0,\beta_1,\sigma} = \{F_n : F_n(y_1, \ldots, y_n) = \prod_{i=1}^{n} \Phi\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right), \beta_0, \beta_1 \in \mathbb{R}, \sigma \in \mathbb{R}^+\} \,.$$

Questo modello è caratterizzato dunque dai parametri β_0, β_1 e σ. In un approccio "distribution-free" non viene specificata nè la funzione di regressione m nè la funzione di ripartizione G che è comune ad ogni $\mathcal{E}_i$. Il modello statistico è dato da

$$\mathcal{M}_{m,G,\sigma} = \{F_n : F_n(y_1, \ldots, y_n) = \prod_{i=1}^{n} G\left(\frac{y_i - m(x_i)}{\sigma}\right), m \in \mathcal{C}, G \in \mathcal{R}, \sigma \in \mathbb{R}^+\} \,,$$

dove $\mathcal{C}$ rappresenta la classe delle funzioni continue e $\mathcal{R}$ rappresenta la classe delle funzioni di ripartizione. In questo caso m, G e σ sono i "parametri" del modello. □

3.2. Le statistiche campionarie

Una statistica campionaria (o semplicemente statistica) è una trasformata del campione. Essendo una trasformata di variabili casuali, anche la statistica campionaria è dunque una variabile casuale. Una statistica è detta "distribution-free" se la sua distribuzione rimane invariata sull'intera classe di distribuzioni definite da un modello "distribution-free". Esistono alcune statistiche fondamentali che vengono descritte di seguito.

Considerato un modello statistico relativo ad un campionamento casuale da una variabile casuale X tale che $\mu = \mathrm{E}[X]$ e $\sigma^2 = \mathrm{Var}[X] < \infty$, la media campionaria è data dalla variabile casuale

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

la cui realizzazione è indicata con $\overline{x}$. Si ha

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

e

$$\text{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

Anche se questi due risultati sono validi per qualsiasi modello con $\sigma^2 < \infty$, la media campionaria non è "distribution-free" in quanto la sua distribuzione dipende dalla variabile casuale X da cui si effettua il campionamento. Inoltre, per il Teorema Fondamentale del Limite, per $n \to \infty$ la variabile casuale standardizzata

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

converge in distribuzione a una variabile casuale con distribuzione $N(0, 1)$ se $\sigma^2 < \infty$. Dunque, assumendo noti μ e σ, la statistica Z risulta "distribution-free" per grandi campioni dal momento che la sua distribuzione asintotica rimane invariata per qualsiasi variabile casuale X.

• **Esempio 3.2.1.** Dato un campione casuale dalla variabile casuale Esponenziale $X \sim E(0, \sigma)$, si dimostra che $\bar{X} \sim G(0, \sigma/n; n)$. Per la proprietà della variabile casuale Gamma si ha $\text{E}[\bar{X}] = \sigma$ e $\text{Var}[\bar{X}] = \sigma^2/n$. Dunque, risultano verificati i risultati generali visti in precedenza, in quanto per la variabile casuale Esponenziale $X \sim E(0, \sigma)$ si ha $\text{E}[X] = \sigma$ e $\text{Var}[X] = \sigma^2$. Assumendo $\sigma = 1$, la Figura 3.2.1 riporta le funzioni di densità di $\bar{X}$ per $n = 5, 10, 20$. Risulta evidente che la distribuzione di $\bar{X}$ si avvicina rapidamente a quella della Normale per $n \to \infty$ anche quando si campiona da una distribuzione asimmetrica come quella Esponenziale. □

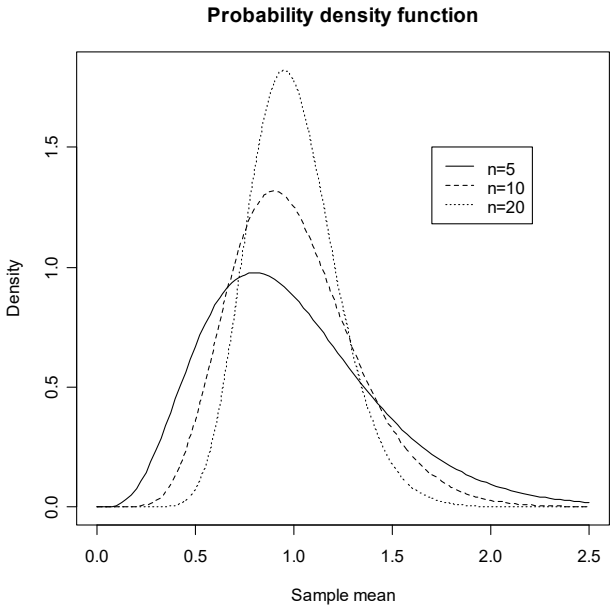


Figura 3.2.1.

Considerato un modello statistico relativo ad un campionamento casuale da una variabile casuale X tale che $\alpha_4 < \infty$, la varianza campionaria è data dalla variabile casuale

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

la cui realizzazione è indicata con s_x^2. Si ha

$$\mathrm{E}[S_x^2] = \frac{n-1}{n}\,\sigma^2 \,.$$

La varianza campionaria corretta è data dalla variabile

$$S_{c,x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \,,$$

la cui realizzazione è indicata con $s_{c,x}^2$, ed è tale che

$$\mathrm{E}[S_{c,x}^2] = \sigma^2$$

e

$$\mathrm{Var}[S_{c,x}^2] = \frac{\sigma^4}{n}\left(\alpha_4 - \frac{n-3}{n-1}\right) .$$

Anche se queste proprietà sono valide per qualsiasi modello con $\alpha_4 < \infty$, la varianza campionaria non è "distribution-free". Per il Teorema Fondamentale del Limite e il Teorema di Slutsky, la variabile casuale standardizzata

$$Z = \frac{\sqrt{n}(S_{c,x}^2 - \sigma^2)}{\sigma^2\sqrt{\alpha_4 - 1}}$$

converge in distribuzione a una variabile casuale $N(0,1)$ per $n \to \infty$ se $\alpha_4 < \infty$. Assumendo note le quantità α_4 e σ, la statistica Z risulta "distribution-free" per grandi campioni dal momento che la sua distribuzione asintotica rimane invariata per qualsiasi variabile casuale X. Inoltre, per il Teorema Fondamentale del Limite e il Teorema di Slutsky, per $n \to \infty$ la media campionaria standardizzata con lo scarto quadratico campionario, ovvero

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{S_{c,x}} \,,$$

converge in distribuzione a una variabile casuale $N(0,1)$ se $\sigma^2 < \infty$. Dunque, anche questa statistica risulta "distribution-free" per grandi campioni.

• **Esempio 3.2.2.** Dato un campione casuale dalla variabile casuale Normale $X \sim N(\mu, \sigma^2)$ è possibile verificare che $\overline{X}$ e $S_{c,x}^2$ sono indipendenti. Si può inoltre dimostrare che questo risultato è valido solo per questo particolare modello statistico. Si ha

$$(n-1)\,\frac{S_{c,x}^2}{\sigma^2} \sim \chi_{n-1}^2 \,.$$

Dal momento che $\mathrm{E}[\chi_{n-1}^2] = n-1$ e $\mathrm{Var}[\chi_{n-1}^2] = 2(n-1)$, risulta

$$\mathrm{E}[S_{c,x}^2] = \sigma^2$$

e

$$\mathrm{Var}[S_{c,x}^2] = \frac{2\sigma^4}{n-1} \,.$$

Essendo $\alpha_4 = 3$ per la variabile casuale Normale $X \sim N(\mu, \sigma^2)$ viene dunque convalidato il risultato generale. Inoltre, risulta $\overline{X} \sim N(\mu, \sigma^2/n)$, ovvero per questo modello la media campionaria ha distribuzione Normale anche per n finito. Assumendo $\sigma = 1$, la Figura 3.2.2 riporta le funzioni di

densità di $S^2_{c,x}$ per $n = 5, 10, 20$. Risulta apparente che la distribuzione di $S^2_{c,x}$ si avvicina rapidamente a quella Normale per $n \to \infty$. □

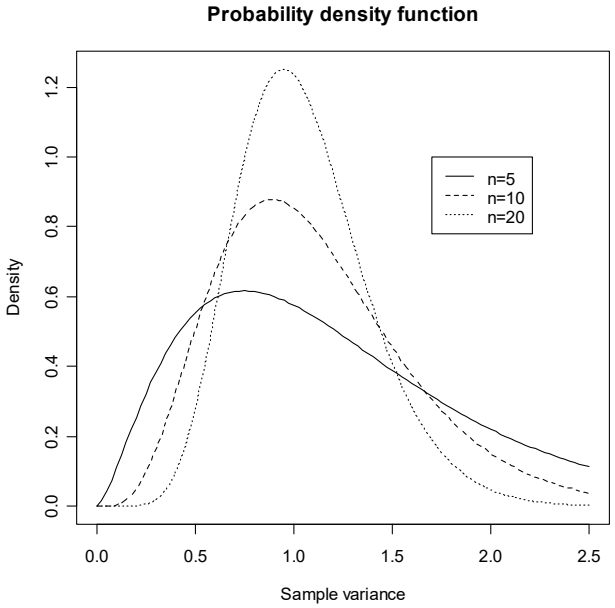


Figura 3.2.2.

Considerato un modello statistico relativo ad un campionamento casuale da una variabile casuale X con funzione di ripartizione F, la funzione di ripartizione empirica è data da

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(X_i) \, .$$

Ovviamente, la funzione di ripartizione empirica fornisce la percentuale di osservazioni campionarie minori od uguali ad un dato valore x. Si ha

$$\mathrm{E}[\widehat{F}(x)] = F(x)$$

e

$$\mathrm{Var}[\widehat{F}(x)] = \frac{F(x)(1 - F(x))}{n} \, .$$

Dunque, la distribuzione della variabile casuale $n\widehat{F}(x)$ è Binomiale con parametri n e $F(x)$. Si noti che per un dato x la funzione di ripartizione empirica è in effetti una media campionaria e quindi ha le proprietà di questa statistica. Inoltre, per $n \to \infty$ la funzione di ripartizione empirica $\widehat{F}$ converge uniformemente ad F, nel senso che $\sup_x |\widehat{F}(x) - F(x)|$ converge quasi certamente a zero per il Teorema di Glivenko-Cantelli.

• **Esempio 3.2.3.** Si dispone delle osservazioni delle precipitazioni medie (in pollici) per 70 città degli Stati Uniti (Fonte: McNeil, D.R., 1977, *Interactive Data Analysis*, Wiley, New York). I dati sono contenuti nel file `rainfall.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\rainfall.txt", header = T)
> attach(d)
```

Il grafico della funzione di ripartizione empirica viene ottenuto mediante il seguente comando:

```
> plot(ecdf(Rainfall), xlab = "Rainfall (inches)",
+    ylab = "Probability",
+    main = "Empirical distribution function")
> rug(Rainfall)
```

Il precedente comando fornisce il grafico della Figura 3.2.3.

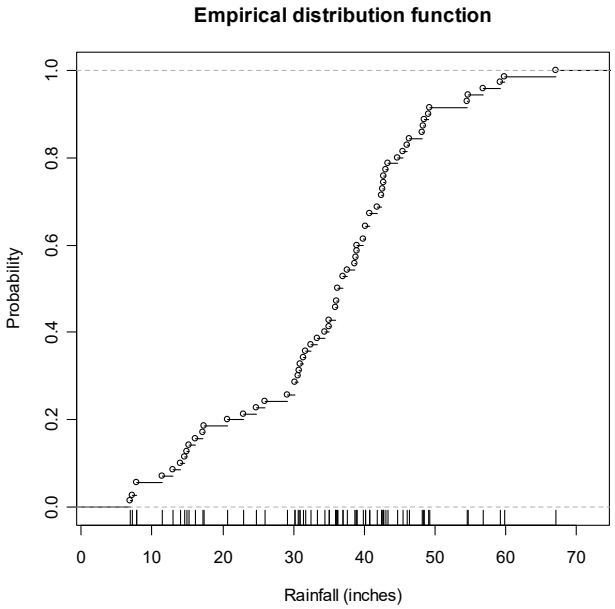


Figura 3.2.3.

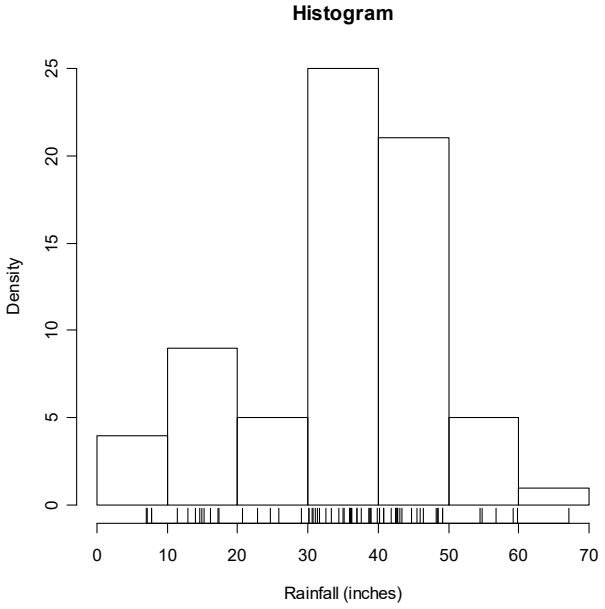


Figura 3.2.4.

Al fine di analizzare la funzione di ripartizione empirica è conveniente comparare il suo grafico con l'istogramma, che si ottiene con il seguente comando:

```
> hist(Rainfall, xlab = "Rainfall (inches)", ylab = "Density",
+    main = "Histogram")
> rug(Rainfall)
```

Il precedente comando fornisce il grafico della Figura 3.2.4. Inoltre, può essere conveniente riportare il grafico della funzione di ripartizione empirica con segmenti uniti per una migliore interpretazione grafica:

```
> plot(ecdf(Rainfall), do.points = F, verticals = T,
+     xlab = "Rainfall (inches)", ylab = "Probability",
+     main = "Empirical distribution function")
> rug(Rainfall)
```

Il precedente comando fornisce il grafico della Figura 3.2.5. □

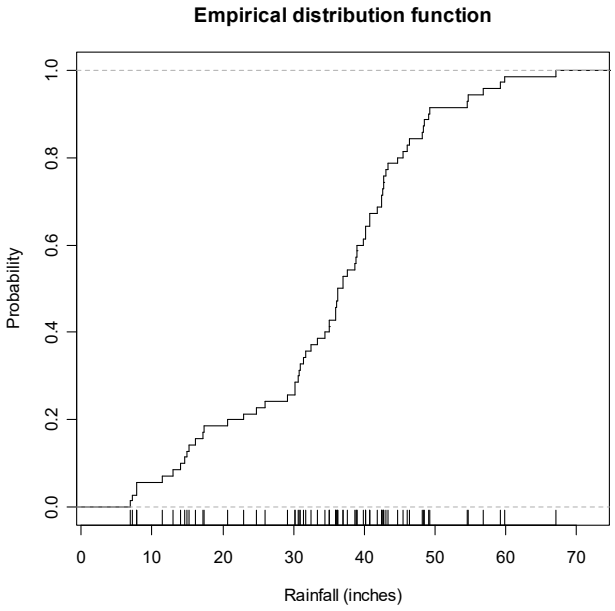


Figura 3.2.5.

Dato un modello statistico relativo ad un campionamento casuale da una variabile casuale X, le osservazioni ordinate $x_{(1)}, \ldots, x_{(n)}$ sono la realizzazione campionaria del vettore di statistiche $(X_{(1)}, \ldots, X_{(n)})$, detto statistica ordinata. La statistica $X_{(i)}$ è detta i-esima statistica ordinata. Se la variabile casuale X da cui si sta campionando è continua con funzione di ripartizione F e funzione di densità f, allora la distribuzione della statistica ordinata può essere ottenuta in forma semplice. In questo caso, la funzione di densità congiunta di $(X_{(1)}, \ldots, X_{(n)})$ risulta

$$f_{(X_{(1)},\ldots,X_{(n)})}(x_{(1)}, \ldots, x_{(n)}) = n! \prod_{i=1}^{n} f(x_{(i)})\, \mathbf{1}_D(x_{(1)}, \ldots, x_{(n)}) \,,$$

dove si assume che $D = \{(x_{(1)}, \ldots, x_{(n)}) : -\infty < x_{(1)} < \cdots < x_{(n)} < \infty\}$. La funzione di densità marginale di $X_{(i)}$ è data da

$$f_{X_{(i)}}(x_{(i)}) = n \binom{n-1}{i-1} F(x_{(i)})^{i-1} (1 - F(x_{(i)}))^{n-i} f(x_{(i)}) \,.$$

Si osservi che la statistica ordinata non è "distribution-free".

I quantili campionari sono funzioni della statistica ordinata. In effetti, anche se esistono diverse proposte in letteratura per la definizione di quantile campionario, tenendo presente che la funzione di ripartizione empirica dipende dalla statistica ordinata dal momento che

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(X_{(i)}) \,,$$

una delle definizioni comunemente adottate per il quantile campionario di ordine $\alpha \in [0,1]$ è data da

$$\widetilde{X}_\alpha = \inf_x \{x : \widehat{F}(x) \geq \alpha\} \,.$$

Ad esempio, mediante questa definizione, la mediana campionaria è definita come $\widetilde{X}_{0.5} = X_{(n/2+1/2)}$ se n è dispari, mentre $\widetilde{X}_{0.5} = X_{(n/2)}$ se n è pari. Infine, se il campionamento è da una variabile casuale continua che ammette funzione di densità f, la variabile casuale standardizzata

$$Z = \frac{\sqrt{n} f(x_\alpha)(\widetilde{X}_\alpha - x_\alpha)}{\sqrt{\alpha(1-\alpha)}}$$

converge in distribuzione a una variabile casuale con distribuzione $N(0,1)$ per $n \to \infty$.

• **Esempio 3.2.4.** Dato un campione casuale da $X \sim U(0,\delta)$, si consideri la massima statistica ordinata, ovvero $X_{(n)}$. La funzione di densità di $X_{(n)}$ risulta

$$f_{X_{(n)}}(x_{(n)}) = \frac{n}{\delta}\Big(\frac{x_{(n)}}{\delta}\Big)^{n-1} \mathbf{1}_{[0,1]}\Big(\frac{x_{(n)}}{\delta}\Big) \,,$$

ovvero $X_{(n)} \sim Be(0,\delta;n,1)$. Analogamente, si può verificare che $X_{(1)} \sim Be(0,\delta;1,n)$. Supponendo la numerosità campionaria dispari con $n = 2l+1$, la mediana campionaria è data da $\widetilde{X}_{0.5} = X_{(l+1)}$. La funzione di densità di $\widetilde{X}$ risulta dunque

$$f_{\widetilde{X}_{0.5}}(\widetilde{x}) = \frac{2l+1}{\delta}\binom{2l}{l}\Big(\frac{\widetilde{x}}{\delta}\Big)^{l}\Big(1-\frac{\widetilde{x}}{\delta}\Big)^{l} \mathbf{1}_{[0,1]}\Big(\frac{\widetilde{x}}{\delta}\Big) \,,$$

ovvero $\widetilde{X}_{0.5} \sim Be(0,\delta;l+1,l+1)$. Inoltre, dal momento che $x_{0.5} = \theta/2$, allora

$$Z = \frac{\sqrt{n}}{\theta}\,(2\widetilde{X}_{0.5} - \theta)$$

converge in distribuzione a una variabile casuale con distribuzione $N(0,1)$ per $n \to \infty$. □

3.3. Alcune statistiche campionarie "distribution-free"

Dato un modello statistico relativo ad un campionamento casuale da una variabile casuale continua X con mediana pari a λ, le statistiche segno sono date dalle n variabili casuali

$$Z_i = \mathbf{1}_{[0,\infty[}(X_i - \lambda) \,,$$

con $i = 1,\ldots,n$. Evidentemente, la variabile casuale Z_i è binaria ed assume valore 1 se X_i è maggiore o uguale alla mediana e valore 0 altrimenti. In particolare, ogni Z_i è distribuita come una variabile casuale Binomiale di parametri 1 e $1/2$, ovvero $Z_i \sim B(1,1/2)$. Le statistiche segno sono indipendenti in quanto trasformate di variabili casuali indipendenti. Le statistiche segno sono anche "distribution-free" nella classe considerata, in quanto la loro distribuzione non dipende dalla distribuzione di X. Infine, anche trasformate di queste statistiche sono "distribution-free".

• **Esempio 3.3.1.** Si consideri la statistica

$$B = \sum_{i=1}^{n} Z_i \, ,$$

che rappresenta il numero di osservazioni maggiori della mediana λ nel campione. Tenendo presente la distribuzione del vettore casuale $(Z_1, \ldots, Z_n)$, si verifica che $B \sim B(n, 1/2)$. Di conseguenza, dal momento che la distribuzione di B rimane invariata per ogni funzione di ripartizione congiunta di un campione casuale da una variabile casuale continua con mediana pari a λ, allora B è una statistica "distribution-free" su questa classe. Questo risultato può essere ottenuto immediatamente tenendo presente che trasformate di statistiche segno sono "distribution-free". □

Dato un modello statistico relativo ad un campionamento casuale da una variabile casuale continua X, le statistiche rango sono date dalle n trasformate

$$R_i = \sum_{j=1}^{n} \mathbf{1}_{[0,\infty[}(X_i - X_j) \, ,$$

con $i = 1, \ldots, n$. Dunque, l'i-esimo rango R_i fornisce il numero di osservazioni minori o uguali a X_i, ovvero R_i rappresenta la posizione di X_i all'interno del campione ordinato e si ha la relazione

$$X_i = X_{(R_i)} \, .$$

Le statistiche rango non sono indipendenti e il vettore casuale $(R_1, \ldots, R_n)$ assume valori sull'insieme $\mathcal{S}_n$ delle permutazioni dei primi n interi. La funzione di probabilità congiunta del vettore casuale $(R_1, \ldots, R_n)$ è uniforme su $\mathcal{S}_n$, ovvero

$$p_{(R_1,\ldots,R_n)}(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{n!} \, \mathbf{1}_{\mathcal{S}_n}(r_1, \ldots, r_n) \, .$$

Di conseguenza, le statistiche rango sono "distribution-free" e quindi trasformate di $(R_1, \ldots, R_n)$ sono "distribution-free". Inoltre, la funzione di probabilità dell'i-esimo rango R_i è uniforme sui primi n interi, ovvero

$$p_{R_i}(r_i) = \frac{1}{n} \, \mathbf{1}_{\{1,\ldots,n\}}(r_i) \, .$$

In particolare, si ha

$$\mathrm{E}[R_i] = \frac{n+1}{2}$$

e

$$\mathrm{Var}[R_i] = \frac{n^2 - 1}{12} \, .$$

In generale, la funzione di probabilità congiunta di una scelta $(i_1, \ldots, i_k)$ di $k < n$ statistiche rango $(R_{i_1}, \ldots, R_{i_k})$, risulta

$$p_{(R_{i_1},\ldots,R_{i_k})}(r_1, \ldots, r_k) = \frac{(n-k)!}{n!} \, \mathbf{1}_{\mathcal{S}_{k,n}}(r_1, \ldots, r_k) \, ,$$

dove $\mathcal{S}_{k,n} = \{(r_1, \ldots, r_k) : (r_1, \ldots, r_k) \in \{1, \ldots, n\}\}$, ovvero $(R_{i_1}, \ldots, R_{i_k})$ ha una distribuzione uniforme su $\mathcal{S}_{k,n}$. Dunque, da questo risultato si ha anche

$$\mathrm{Cov}[R_i, R_j] = -\frac{n+1}{12}$$

per $i \neq j = 1, \ldots, n$. Infine, se $(R_{(i_1)}, \ldots, R_{(i_k)})$ è la statistica ordinata di $(R_{i_1}, \ldots, R_{i_k})$, la relativa funzione di probabilità congiunta risulta

$$p_{(R_{(i_1)},\ldots,R_{(i_k)})}(r_{(1)}, \ldots, r_{(k)}) = \binom{n}{k}^{-1} \mathbf{1}_{\mathcal{S}_{(k),n}}(r_{(1)}, \ldots, r_{(k)}) \,,$$

dove $\mathcal{S}_{(k),n} = \{(r_1, \ldots, r_k) : (r_1, \ldots, r_k) \in \{1, \ldots, n\}, r_{(1)} < \cdots < r_{(k)}\}$, ovvero $(R_{(i_1)}, \ldots, R_{(i_k)})$ ha a sua volta una distribuzione uniforme su $\mathcal{S}_{(k),n}$.

• **Esempio 3.3.2.** Si consideri la suddivisione del campione casuale in due sottocampioni $(X_1, \ldots, X_{n_1})$ e $(X_{n_1+1}, \ldots, X_n)$, rispettivamente di numerosità n_1 e n_2 con $n = n_1 + n_2$. Di conseguenza, siano $(R_1, \ldots, R_{n_1})$ i ranghi assegnati al primo sottocampione e $(R_{n_1+1}, \ldots, R_n)$ i ranghi assegnati al secondo sottocampione nel campione $(X_1, \ldots, X_n)$. Si consideri la statistica

$$W = \sum_{i=1}^{n_1} R_i \,,$$

che fornisce la somma dei ranghi assegnati al primo sottocampione. Quando i ranghi assegnati al primo sottocampione sono i più bassi (ovvero $1, \ldots, n_1$) si ottiene il valore minimo di W, dato da $\sum_{i=1}^{n_1} i = n_1(n_1 + 1)/2$. Quando i ranghi assegnati a $(X_1, \ldots, X_{n_1})$ sono i più elevati (ovvero $n_2 + 1$, $\ldots, n$) si ottiene il valore massimo di W, dato da $\sum_{i=1}^{n_1}(n_2 + i) = n_1(n + n_2 + 1)/2$. Quindi, il supporto di W è $\{n_1(n_1 + 1)/2, n_1(n_1 + 1)/2 + 1, \ldots, n_1(n + n_2 + 1)/2\}$. Se $c_{n_1,n_2}(w)$ rappresenta il numero di sottoinsiemi di n_1 interi di $\{1, \ldots, n\}$ la cui somma è w, dal momento che W può essere eauivalentemente scritto come

$$W = \sum_{i=1}^{n_1} R_{(i)} \,,$$

allora la funzione di probabilità di W è data da

$$p_W(w) = \binom{n}{n_1}^{-1} c_{n_1,n_2}(w) \mathbf{1}_{\{n_1(n_1+1)/2, n_1(n_1+1)/2+1, \ldots, n_1(n+n_2+1)/2\}}(w) \,.$$

Dal momento che la distribuzione di W rimane invariata per ogni funzione di ripartizione congiunta di un campione casuale da una variabile casuale continua, allora W è una statistica "distribution-free" su questa classe. Questo risultato può essere ottenuto immediatamente dal momento che trasformate di statistiche rango sono "distribution-free". □

3.4. Riferimenti bibliografici

Almudevar, A. (2022) *Theory of Statistical Inference*, CRC Press, Boca Raton.

Boos, D.D. e Stefanski, L.A. (2013) *Essential Statistical Inference*, Springer, New York.

Deshmukh, S. e Kulkarni, M. (2022) *Asymptotic Statistical Inference*, Springer, Singapore.

Ferguson, T.S. (1996) *A Course in Large Sample Theory*, Chapman and Hall, London.

Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.

Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.

Huber, P.J. e Ronchetti, E.M. (2009) *Robust Statistics*, seconda edizione, Wiley, New York.

Lauritzen, S. (2023) *Fundamentals of Mathematical Statistics*, Chapman & Hall/CRC Press, Boca Raton.

Lehmann, E.L. (1999) *Elements of Large Sample Theory*, Springer, New York.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*, Wiley, New York.
Shao, J. (2003) *Mathematical Statistics*, seconda edizione, Springer, New York.
van der Vaart, A.W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Pagina intenzionalmente vuota

Capitolo 4
I metodi di stima

4.1. Lo stimatore

Una volta scelto un modello statistico, un primo obiettivo dell'inferenza è quello di selezionare sulla base del campione dei valori "plausibili" per i parametri che caratterizzano il modello. Il procedimento di stima fa corrispondere ad ogni campione un valore per i parametri, ovvero considera una trasformata del campione detta stimatore. Uno stimatore è dunque per definizione una statistica o un insieme di statistiche. La realizzazione campionaria dello stimatore è detta stima. Questo tipo di procedimento inferenziale è detto stima per punti perchè ad ogni campione fa corrispondere una stima (ovvero un singolo "punto" dello spazio parametrico). Anche se uno stimatore gode di proprietà ottimali, la stima può essere molto differente dal "vero valore" del parametro a causa della variabilità campionaria. Dunque, in un procedimento di stima per punti, la stima deve sempre essere accompagnata da un indice di "precisione" dello stimatore nello stimare il vero parametro.

Per semplicità vengono considerate di seguito le proprietà di uno stimatore di un singolo parametro θ, che possono essere comunque estese al caso di più parametri. Dato un modello statistico relativo al campione $X_1, \ldots, X_n$, lo stimatore del parametro θ è dunque una trasformata che può essere indicata come $\widetilde{\Theta} = \widetilde{\Theta}(X_1, \ldots, X_n)$.

La proprietà della correttezza richiede che, *a priori* dal campionamento, la determinazione campionaria dello stimatore sia tendenzialmente prossima al valore vero. Formalmente, lo stimatore $\widetilde{\Theta}$ è detto corretto per il parametro θ se

$$\mathrm{E}[\widetilde{\Theta}] = \theta \, .$$

Uno stimatore non corretto è detto distorto e la distorsione è definita come

$$\mathrm{Bias}[\widetilde{\Theta}] = \mathrm{E}[\widetilde{\Theta}] - \theta \, .$$

Inoltre, uno stimatore è detto asintoticamente corretto per θ se

$$\lim_{n \to \infty} \mathrm{E}[\widetilde{\Theta}] = \theta \, .$$

• **Esempio 4.1.1.** La media campionaria è uno stimatore corretto, essendo $\mathrm{E}[\bar{X}] = \mu$. La proprietà di correttezza della media campionaria è "distribution-free", nel senso che è valida per tutti i modelli con campionamento casuale da una variabile casuale X tale che $\mu < \infty$. Al contrario, la varianza campionaria S_x^2 è uno stimatore distorto per σ^2. La distorsione è pari a

$$\mathrm{Bias}[S_x^2] = \mathrm{E}[S_x^2] - \sigma^2 = -\frac{\sigma^2}{n} \, .$$

Tuttavia, lo stimatore S_x^2 è asintoticamente corretto per σ^2, essendo

$$\lim_{n \to \infty} \mathrm{E}[S_x^2] = \sigma^2 \, .$$

Evidentemente, lo stimatore $S_{c,x}^2$ è corretto per σ^2, dal momento che

$$\mathrm{E}[S_{c,x}^2] = \sigma^2 \, .$$

Anche la proprietà di correttezza di $S_{c,x}$ è "distribution-free", in modo simile alla media campionaria, se si assume che $\sigma^2 < \infty$. □

Si desidera usualmente che la distribuzione dello stimatore si concentri sempre di più intorno a θ all'aumentare della numerosità campionaria, ovvero si richiede la cosiddetta proprietà della coerenza. Formalmente, uno stimatore $\widetilde{\Theta}$ è detto coerente per θ se converge in probabilità a θ per $n \to \infty$. Condizione sufficiente affinchè lo stimatore sia coerente per θ è che sia asintoticamente corretto e che

$$\lim_{n\to\infty} \mathrm{Var}[\widetilde{\Theta}] = 0 \, .$$

• **Esempio 4.1.2.** Per la Legge dei Grandi Numeri la media campionaria $\overline{X}$ converge in probabilità a μ per $n \to \infty$ e quindi è uno stimatore coerente. In effetti, $\overline{X}$ è uno stimatore corretto e

$$\lim_{n\to\infty} \mathrm{Var}[\overline{X}] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \, .$$

La proprietà di coerenza della media campionaria è "distribution-free", se si assume un modello con campionamento casuale da una variabile casuale X tale che $\sigma^2 < \infty$. Anche la varianza campionaria S_x^2 è uno stimatore coerente, essendo asintoticamente corretto e

$$\lim_{n\to\infty} \mathrm{Var}[S_x^2] = \lim_{n\to\infty} \mathrm{Var}[S_{c,x}^2] = \lim_{n\to\infty} \frac{\sigma^4}{n}\left(\alpha_4 - \frac{n-3}{n-1}\right) = 0 \, .$$

Di nuovo, la proprietà di coerenza di S_x^2 è "distribution-free", se si assume che $\alpha_4 < \infty$. □

Quando si deve valutare la precisione di uno stimatore si adotta solitamente il criterio dell'errore quadratico medio, che tiene conto sia della distorsione che della varianza dello stimatore. L'errore quadratico medio è definito come

$$\mathrm{MSE}[\widetilde{\Theta}] = \mathrm{Bias}[\widetilde{\Theta}]^2 + \mathrm{Var}[\widetilde{\Theta}] \, .$$

Basandosi su questo criterio, uno stimatore leggermente distorto e con bassa varianza può essere preferibile ad uno stimatore corretto ma con varianza più elevata.

• **Esempio 4.1.3.** Dato un campione casuale da $X \sim N(0, \upsilon)$, si consideri lo stimatore del parametro υ dato da

$$\widetilde{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \, .$$

Tenendo presente le proprietà della variabile casuale Chi-quadrato, si ottiene

$$\mathrm{E}[\widetilde{\Theta}_1] = \upsilon$$

e

$$\mathrm{Var}[\widetilde{\Theta}_1] = \frac{2\upsilon^2}{n} \, ,$$

per cui lo stimatore è corretto e coerente, con errore quadratico medio

$$\text{MSE}[\widetilde{\Theta}_1] = \frac{2\upsilon^2}{n} .$$

Alternativamente, se si considera lo stimatore

$$\widetilde{\Theta}_2 = \frac{n}{n+2}\,\widetilde{\Theta}_1 ,$$

si ha

$$\text{E}[\widetilde{\Theta}_2] = \frac{n\upsilon}{n+2}$$

e

$$\text{Var}[\widetilde{\Theta}_2] = \frac{2n\upsilon^2}{(n+2)^2} ,$$

ovvero lo stimatore è distorto, anche se asintoticamente corretto e coerente, e quindi l'errore quadratico medio è dato da

$$\text{MSE}[\widetilde{\Theta}_2] = \left(\frac{n\upsilon}{n+2} - \upsilon\right)^2 + \frac{2n\upsilon^2}{(n+2)^2} = \frac{2\upsilon^2}{n+2} .$$

Dal momento che $\text{MSE}[\widetilde{\Theta}_2] < \text{MSE}[\widetilde{\Theta}_1]$, sulla base dell'errore quadratico medio lo stimatore distorto è preferibile a quello corretto. □

Quando si adotta un modello classico, vi sono due ulteriori proprietà desiderabili in uno stimatore. La prima proprietà è quella dell'efficienza, che richiede che uno stimatore corretto abbia varianza minima nell'ambito della classe degli stimatori corretti. Sotto alcune condizioni è possibile dimostrare che per un determinato modello lo stimatore efficiente esiste ed è possibile ottenere una espressione della varianza minima. La seconda proprietà è quella della sufficienza, che assicura che lo stimatore conserva tutta l'informazione fornita dal campione senza alcuna perdita. Queste due proprietà verrano discusse in dettaglio nella prossima Sezione introducendo il concetto fondamentale di veroasimiglianza.

4.2. La verosimiglianza

Si supponga un modello classico indicizzato per semplicità di esposizione mediante il solo parametro θ. La funzione di densità congiunta (o eventualmente la funzione di probabilità congiunta) del campione $X_1, \ldots, X_n$ viene usualmente indicata con

$$f_n(x_1, \ldots, x_n; \theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) ,$$

in modo da enfatizzare l'indicizzazione del modello rispetto al parametro. Quando il campione $X_1, \ldots, X_n$ è stato osservato, $f_n(x_1, \ldots, x_n; \theta)$ è funzione solamente del parametro θ. Dunque, la funzione $f_n(x_1, \ldots, x_n; \theta)$ rappresenta la probabilità di osservare *a priori* esattamente il campione $x_1, \ldots, x_n$ che è stato estratto e contiene tutta l'informazione relativa al campione stesso. In questo caso, si dice funzione di verosimiglianza (o semplicemente verosimiglianza) la funzione data da

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = c f_n(x_1, \ldots, x_n; \theta) ,$$

dove c è una costante che non dipende da θ. Viene usualmente considerata anche la funzione di log-verosimiglianza, definita come

$$l(\theta) = l(\theta; x_1, \ldots, x_n) = \log L(\theta) \,,$$

con la convenzione che $l(\theta) = -\infty$ se $L(\theta) = 0$.

• **Esempio 4.2.1.** Dato un campione casuale da $X \sim N(\mu, 1)$, tenendo presente l'indipendenza delle osservazioni campionarie, la funzione di densità congiunta del campione risulta

$$f_n(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}(x_i - \mu)^2} \,.$$

La verosimiglianza è data da

$$L(\mu) = c \prod_{i=1}^{n} e^{-\frac{1}{2}(x_i - \mu)^2} = c \, e^{-\frac{n}{2}(s_x^2 + (\overline{x} - \mu)^2)} \,,$$

mentre la log-verosimiglianza è data da

$$l(\mu) = \log c - \frac{n}{2} \, (s_x^2 + (\overline{x} - \mu)^2) \,.$$

Il grafico della verosimiglianza per $c = 1$, $n = 5$, $\overline{x} = 1$ e $s_x^2 = 2$ è riportato nella Figura 4.2.1. □

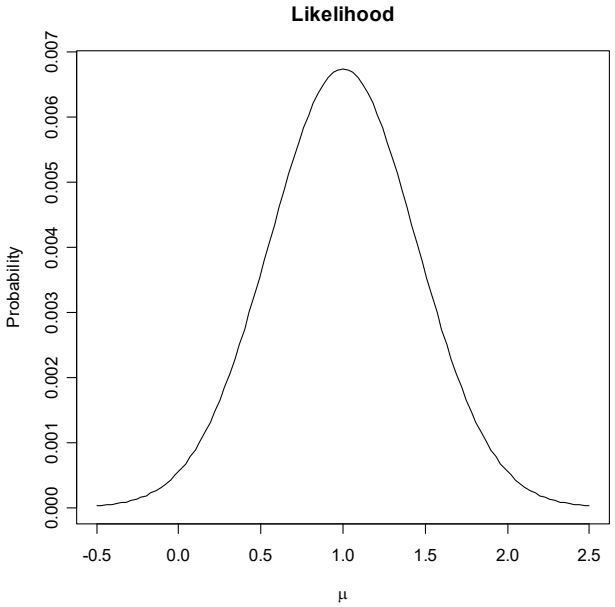


Figura 4.2.1.

• **Esempio 4.2.2.** Dato un campione casuale da $X \sim N(\mu, \upsilon)$, dove si è parametrizzato assumendo che $\upsilon = \sigma^2$, la funzione di densità congiunta del campione risulta

$$f_n(x_1, \ldots, x_n; \mu, \upsilon) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\upsilon}} \, e^{-\frac{1}{2\upsilon}(x_i - \mu)^2} \,.$$

La verosimiglianza è data da

$$L(\mu, \upsilon) = c \prod_{i=1}^{n} \upsilon^{-\frac{1}{2}} e^{-\frac{1}{2\upsilon}(x_i - \mu)^2} = c\upsilon^{-\frac{1}{2}n} \, e^{-\frac{n}{2\upsilon}(s_x^2 + (\overline{x} - \mu)^2)} \,,$$

dal momento che risulta $\sum_{i=1}^{n}(x_i-\mu)^2 = n(s_x^2+(\overline{x}-\mu)^2)$. Si osservi che la verosimiglianza è stata opportunamente espressa in funzione della realizzazione della statistica $(\overline{X}, S_x^2)$. Inoltre, la log-verosimiglianza è data da

$$l(\mu,\upsilon) = \log c - \frac{n}{2}\log\upsilon - \frac{n}{2\upsilon}\left(s_x^2+(\overline{x}-\mu)^2\right).$$

Il grafico della verosimiglianza (e il relativo grafico per linee di livello) per $c=1$, $n=5$, $\overline{x}=1$ e $s_x^2=2$ è riportato nella Figura 4.2.2. □

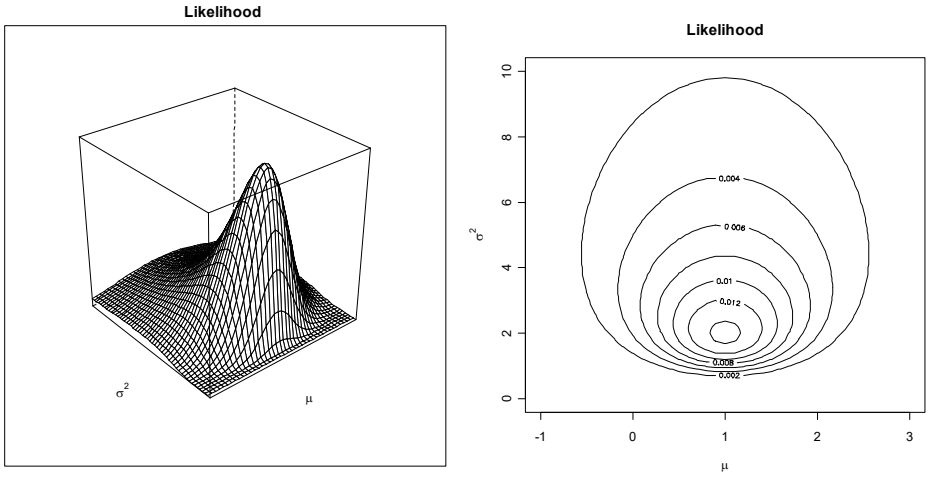


Figura 4.2.2.

• **Esempio 4.2.3.** Dato un campione casuale da $X \sim U(0,\delta)$, la funzione di densità congiunta del campione risulta

$$f_n(x_1,\ldots,x_n;\delta) = \prod_{i=1}^{n}\frac{1}{\delta}\,\mathbf{1}_{[0,1]}\left(\frac{x_i}{\delta}\right).$$

Dunque, la verosimiglianza è data da

$$L(\delta) = c\prod_{i=1}^{n}\frac{1}{\delta}\,\mathbf{1}_{[0,1]}\left(\frac{x_i}{\delta}\right) = c\delta^{-n}\,\mathbf{1}_{[0,1]}\left(\frac{x_{(n)}}{\delta}\right) = c\delta^{-n}\,\mathbf{1}_{[x_{(n)},\infty[}(\delta),$$

dal momento che $\prod_{i=1}^{n}\mathbf{1}_{[0,1]}(x_i/\delta)=1$ quando $\delta > x_i$ per ogni i, ovvero quando si ha $\delta > x_{(n)}$. In questo caso, la verosimiglianza è stata espressa in funzione della realizzazione della statistica $X_{(n)}$, ovvero la massima statistica ordinata. Inoltre, la log-verosimiglianza risulta

$$l(\delta) = \log c - n\log\delta + \log\mathbf{1}_{[x_{(n)},\infty[}(\delta).$$ □

Assumendo alcune condizioni di regolarità che sono soddisfatte dalla maggior parte dei modelli statistici, la cosiddetta funzione punteggio è data da

$$s_n(\theta) = s_n(\theta;x_1,\ldots,x_n) = \frac{d}{d\theta}\,l(\theta;x_1,\ldots,x_n).$$

Si può verificare che si ha

$$\mathrm{E}[s_n(\theta;X_1,\ldots,X_n)] = 0$$

e

$$I_n(\theta) = \text{Var}[s_n(\theta; X_1, \ldots, X_n)] = -\,\text{E}\left[\frac{d}{d\theta}\, s_n(\theta; X_1, \ldots, X_n)\right] = -\,\text{E}\left[\frac{d^2}{d\theta^2}\, l(\theta; X_1, \ldots, X_n)\right].$$

La quantità $I_n(\theta)$ è detta informazione di Fisher ed è fondamentale per determinare uno stimatore efficiente. In effetti, la disugualianza di Rao-Cramér fornisce un limite inferiore per la varianza di uno stimatore corretto $\widetilde{\Theta}$ di θ che dipende da $I_n(\theta)$, ovvero

$$\text{Var}[\widetilde{\Theta}] \geq \frac{1}{I_n(\theta)}\,.$$

Uno stimatore corretto $\widetilde{\Theta}$ per cui si ha $\text{Var}[\widetilde{\Theta}] = 1/I_n(\theta)$ è dunque efficiente.

• **Esempio 4.2.4.** Dato un campione casuale da $X \sim N(\mu, 1)$, si ha la funzione punteggio

$$s_n(\mu; x_1, \ldots, x_n) = \frac{d}{d\mu}\,(\log c - \frac{n}{2}\,(s_x^2 + (\overline{x} - \mu)^2)) = n(\overline{x} - \mu)\,.$$

L'informazione di Fisher è data da

$$I_n(\mu) = -\,\text{E}\left[\frac{d}{d\mu}\, n(\overline{X} - \mu)\right] = n$$

e dunque lo stimatore $\overline{X}$ è efficiente in quanto è corretto e $\text{Var}[\overline{X}] = 1/n$. □

Uno stimatore $\widetilde{\Theta}$ è detto sufficiente per θ se per due realizzazioni campionarie $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_n)$ si ha

$$\widetilde{\Theta}(x_1, \ldots, x_n) = \widetilde{\Theta}(y_1, \ldots, y_n) \Rightarrow L(\theta; x_1, \ldots, x_n) \propto L(\theta; y_1, \ldots, y_n)$$

per ogni $\theta \in \Theta$. La precedente condizione implica effettivamente che lo stimatore mantiene l'informazione contenuta nella verosimiglianza e quindi tutta l'informazione relativa al campione. Il criterio di fattorizzazione di Neyman stabilisce una condizione operativa per verificare la sufficienza di uno stimatore, ovvero lo stimatore $\widetilde{\Theta}$ è sufficiente per θ se

$$f_n(x_1, \ldots, x_n; \theta) = h_n(x_1, \ldots, x_n) g(\widetilde{\theta}; \theta)\,,$$

dove h_n e g sono opportune funzioni.

• **Esempio 4.2.5.** Dato un campione casuale da $X \sim N(\mu, \upsilon)$, tenendo presente l'Esempio 4.2.2, si ha

$$f_n(x_1, \ldots, x_n; \mu, \upsilon) = (2\pi)^{-\frac{1}{2}n} \upsilon^{-\frac{1}{2}n} e^{-\frac{n}{2\upsilon}(s_x^2 + (\overline{x} - \mu)^2)}\,.$$

In questo caso, risulta

$$h_n(x_1, \ldots, x_n) = (2\pi)^{-\frac{1}{2}n}$$

e

$$g(\overline{x}, s_x^2; \mu, \upsilon) = \upsilon^{-\frac{1}{2}n} e^{-\frac{n}{2\upsilon}(s_x^2 + (\overline{x} - \mu)^2)}\,.$$

Dunque, lo stimatore $(\overline{X}, S_x^2)$ è sufficiente per (μ, υ). □

• **Esempio 4.2.6.** Dato un campione casuale da $X \sim U(0, \delta)$, tenendo presente l'Esempio 4.2.3, si ha

$$f_n(x_1, \ldots, x_n; \delta) = \delta^{-n}\, \mathbf{1}_{[0,1]}\left(\frac{x_{(n)}}{\delta}\right).$$

Risulta

$$h_n(x_1, \ldots, x_n) = 1$$

e

$$g(x_{(n)}; \delta) = \delta^{-n}\, \mathbf{1}_{[0,1]}\Big(\frac{x_{(n)}}{\delta}\Big) .$$

Dunque, lo stimatore $X_{(n)}$ è sufficiente per δ. □

4.3. Il metodo della massima verosimiglianza

Anche se esistono diversi metodi di stima, quando si assume un modello statistico classico il cosiddetto metodo della verosimiglianza ha un'importanza fondamentale. Il metodo della massima verosimiglianza consiste nello scegliere il valore del parametro che massimizza la probabilità di ottenere proprio il campione che è stato estratto. Evidentemente, questo metodo di stima può essere adoperato solo quando si considerano modelli classici. Formalmente, si dice stima di massima verosimiglianza di θ quel valore $\widehat{\theta}$ tale che

$$\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\, L(\theta) .$$

Dal momento che la funzione logaritmo è monotona crescente, la precedente condizione è equivalente a

$$\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\, l(\theta) .$$

La stima di massima verosimiglianza $\widehat{\theta}$ è la realizzazione campionaria di $\widehat{\Theta}$, detto appunto stimatore di massima verosimiglianza. Il metodo della massima verosimiglianza fornisce stimatori che hanno proprietà ottimali, sia per campioni finiti che per grandi campioni, assumendo alcune condizioni di regolarità che sono soddisfatte dalla maggior parte dei modelli statistici classici.

La proprietà di equivarianza assicura la congruenza della stima di massima verosimiglianza quando si riparametrizza il modello originale mediante funzioni biunivoche dei parametri. Se g è una funzione biunivoca tale che $\gamma = g(\theta)$ e $L_\gamma(\gamma)$ è la verosimiglianza relativa al parametro γ, si ha

$$L_\gamma(\gamma) = L(g^{-1}(\gamma)) = L(\theta) .$$

Dal momento che il massimo di $L(\theta)$ si ottiene per $\widehat{\theta}$, il massimo di $L_\gamma(\gamma)$ si ottiene per $\widehat{\gamma} = g(\widehat{\theta})$, che è dunque la stima di massima verosimiglianza di γ.

Gli stimatori di massima verosimiglianza sono trasformate di stimatori sufficienti quando questi esistono. In effetti, sulla base del criterio di fattorizzazione di Neyman risulta

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = c\, g(\widetilde{\theta}; \theta)$$

e la stima di verosimiglianza può essere ottenuta semplicemente massimizzando $g(\widetilde{\theta}; \theta)$. Dunque, lo stimatore di massima verosimiglianza $\widehat{\Theta}$ è funzione dello stimatore sufficiente $\widetilde{\Theta}$.

Sotto opportune ipotesi di regolarità lo stimatore di massima verosimiglianza $\widehat{\Theta}$ è anche coerente e la variabile casuale standardizzata $\sqrt{I_n(\theta)}(\widehat{\Theta} - \theta)$ converge in distribuzione a una variabile casuale $N(0, 1)$ per $n \to \infty$. Quindi, lo stimatore di massima verosimiglianza è efficiente e con distribuzione Normale per grandi campioni. Inoltre, uno stimatore coerente per $\text{Var}[\widehat{\Theta}]$ è dato da

$$\widehat{\mathrm{Var}}[\widehat{\Theta}] = \frac{1}{I_n(\widehat{\Theta})} .$$

• **Esempio 4.3.1.** Dato un campione casuale da $X \sim N(\mu, \upsilon)$, si ha

$$l(\mu, \upsilon) = \log c - \frac{n}{2}\log \upsilon - \frac{n}{2\upsilon}\left(s_x^2 + (\overline{x} - \mu)^2\right) \leq \log c - \frac{n}{2}\log \upsilon - \frac{n s_x^2}{2\upsilon} = l(\overline{x}, \upsilon) = \max_{\mu \in \mathbb{R}} l(\mu, \upsilon) ,$$

essendo $(\overline{x} - \mu)^2 \geq 0$. In generale, è valida la relazione $\log x \leq x - 1$ per $x > 0$, essendo la funzione logaritmo concava, e si ha

$$-\log \upsilon - \frac{d}{\upsilon} \leq -\log d - 1$$

per d costante positiva. Dunque, posto $d = s_x^2$ si ottiene

$$l(\overline{x}, \upsilon) = \log c - \frac{n}{2}\log \upsilon - \frac{n s_x^2}{2\upsilon} \leq \log c - \frac{n}{2}\log s_x^2 - \frac{n}{2} = l(\overline{x}, s_x^2) = \max_{\mu \in \mathbb{R}, \upsilon \in \mathbb{R}^+} l(\mu, \upsilon) .$$

Dunque, $(\widehat{\mu}, \widehat{\upsilon}) = (\overline{x}, s_x^2)$ è la stima di massima verosimiglianza di (μ, υ). Di conseguenza, lo stimatore di massima verosimiglianza risulta $(\overline{X}, S_x^2)$. Per la proprietà di equivarianza la stima di σ è data da $\widehat{\sigma} = s_x$. Inoltre, si ha $\overline{X} \sim N(\mu, \upsilon/n)$ e $nS_x^2/\upsilon \sim \chi^2_{n-1}$ e si può verificare che questi due stimatori sono indipendenti. Inoltre, uno stimatore coerente di $\mathrm{Var}[\overline{X}]$ è dato da $\widehat{\mathrm{Var}}[\overline{X}] = S_x^2/n$. □

• **Esempio 4.3.2.** Dato un campione casuale da $X \sim U(0, \delta)$, si ha

$$l(\delta) = \log c - n\log \delta + \log \mathbf{1}_{[x_{(n)}, \infty[}(\delta) \leq \log c - n \log x_{(n)} = l(x_{(n)}) = \max_{\delta \in \mathbb{R}^+} l(\delta) ,$$

dove si è tenuto presente che la funzione logaritmo è crescente e che la log-verosimiglianza può essere definita per $\delta \geq x_{(n)}$. Dunque, $\widehat{\delta} = x_{(n)}$ è la stima di massima verosimiglianza di δ. Lo stimatore di massima verosimiglianza risulta $X_{(n)}$ e si ha $X_{(n)} \sim Be(0, \delta; n, 1)$. Inoltre, si ha

$$\mathrm{E}[X_{(n)}] = \frac{n\delta}{n+1}$$

e

$$\mathrm{Var}[X_{(n)}] = \frac{n\delta^2}{(n+1)^2(n+2)} .$$

Lo stimatore di massima verosimiglianza è asintoticamente corretto, coerente e sufficiente per δ. Le condizioni di regolarità non sono verificate per questo modello e in effetti $n(\delta - X_{(n)})$ converge in distribuzione ad una variabile Esponenziale $E(0, \delta)$ per $n \to \infty$, ovvero lo stimatore di massima verosimiglianza non possiede distribuzione Normale per grandi campioni. Si noti che che lo stimatore $X_{(n)}$ è super efficiente, nel senso che $\mathrm{Var}[X_{(n)}]$ è di ordine n^{-2}. Inoltre, uno stimatore coerente di questa quantità è dato da $\widehat{\mathrm{Var}}[X_{(n)}] = X_{(n)}^2/n^2$. □

• **Esempio 4.3.3.** Quando si considera un campione da una variabile casuale continua si deve avere cautela nella definizione della verosimiglianza. Supponendo per semplicità di disporre di un campione di una sola osservazione x da $X \sim N(\mu, 1)$, essendo $f(x) = \phi(x - \mu)$ la verosimiglianza risulta

$$L(\mu) = \phi(x - \mu) .$$

Dal momento che $\phi(0)$ è il massimo della funzione ϕ, la stima di massima verosimiglianza è data da $\widehat{\mu} = x$. La funzione di densità è in generale definita a meno di insiemi con misura di Lebesgue nulla e dunque una versione legittima di f è anche data da

$$f(x) = \phi(x-\mu)\mathbf{1}_{\mathbb{R}\setminus\{1\}}(x-\mu) + 10\,\mathbf{1}_{\{1\}}(x-\mu)\,.$$

In questo caso, la verosimiglianza risulta

$$L(\mu) = \phi(x-\mu)\mathbf{1}_{\mathbb{R}\setminus\{x-1\}}(\mu) + 10\,\mathbf{1}_{\{x-1\}}(\mu)$$

e la stima di massima verosimiglianza è data da $\widehat{\mu} = x - 1$. Per evitare questo genere di incongruenze esistono definizioni formali della verosimiglianza in ambito probabilistico. In pratica, è sufficiente adottare la versione più "regolare" della funzione di densità. □

• **Esempio 4.3.4.** Si dispone di un campione casuale di diametri di sfere misurate in micron (Fonte: Romano, A., 1977, *Applied Statistics for Science and Industry*, Allyn and Bacon, Boston). I dati sono contenuti nel file `ball.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\ball.txt", header = T)
> attach(d)
```

Assumendo un campionamento casuale da $X \sim N(\mu, \upsilon)$, le stime di massima verosimiglianza di μ e σ^2 risultano:

```
> mean(Diameter)
[1] 1.194
> variance(Diameter)
[1] 0.075524
```

Può essere opportuno verificare la validità del modello controllando i valori dei coefficienti campionari di asimmetria e curtosi (si noti che per una distribuzione Normale i coefficienti di asimmetria e curtosi devono risultare rispettivamente pari a 0 e 3):

```
> skewness(Diameter)
[1] -0.1763099
> kurtosis(Diameter)
[1] 2.170784
```

La validità del modello può essere anche controllata graficamente mediante il diagramma quantile-quantile, che fornisce il diagramma delle osservazioni (standardizzate mediante la media e la varianza campionarie) rispetto ai quantili della distribuzione Normale standardizzata. Questo grafico dovrebbe avere una disposizione dei punti lungo la bisettrice se l'ipotesi di normalità per le osservazioni è valida. I comandi per ottenere il diagramma quantile-quantile sono i seguenti:

```
> qqnorm(Diameter)
> qqline(Diameter)
```

I precedenti comandi forniscono il grafico di Figura 4.3.1. □

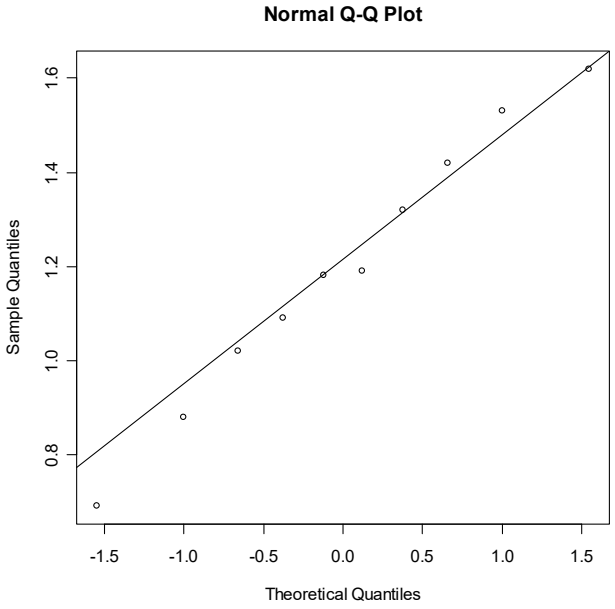


Figura 4.3.1.

Si noti infine che frequentemente il metodo della massima verosimiglianza fornisce stimatori che coincidono con quelli ottenuti mediante il principio di corrispondenza, che è la tecnica di stima più elementare e più intuitiva. Nel principio di corrispondenza si suppone che il parametro venga rappresentato come la media di una opportuna trasformata di X, ovvero $\theta = \mathrm{E}[t(X)]$. In questo caso, lo stimatore di θ è fornito dalla controparte campionaria

$$\widetilde{\Theta} = \frac{1}{n}\sum_{i=1}^{n} t(X_i) \, .$$

Dunque, stimatori come la media e la varianza campionaria, o la funzione di ripartizione empirica, sono giustificati dal principio di corrispondenza. Il metodo della massima verosimiglianza tende quindi a produrre anche stimatori che sono facilmente interpretabili nel loro significato statistico.

4.4. Il metodo dei minimi quadrati

Il metodo dei minimi quadrati viene solitamente applicato quando si considera la stima dei parametri con un modello di regressione. Nel caso semplice di un modello di regressione lineare con un unico regressore, se le osservazioni sono costituite dalle n coppie cartesiane $(x_1, y_1), \ldots, (x_n, y_n)$, il metodo dei minimi quadrati consiste nel minimizzare la somma degli scarti al quadrato dei valori osservati dai valori teorici della variabile di risposta, ovvero nel minimizzare la funzione obiettivo

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \, .$$

In questo caso, la minimizzazione fornisce le stime

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

e

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} \, ,$$

per cui la retta di regressione stimata risulta $\widehat{m}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$. Si noti che il metodo dei minimi quadrati non postula in effetti assunzioni funzionali sulla variabile di risposta. Il metodo può essere anche adoperato in modo generale con modelli complessi, come sarà evidenziato nei Capitoli 9 e 10.

• **Esempio 4.4.1.** Si dispone delle osservazioni del livello del lago Vittoria (in metri) e del numero di macchie solari per gli anni 1902-1921 (Fonte: Shaw, N., 1942, *Manual of Metereology*, Cambridge University Press, London, p.284). La variabile di risposta è il livello del lago (in metri) rispetto ad un valore di riferimento, mentre il regressore è il numero di macchie solari. I dati sono contenuti nel file `lake.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\lake.txt", header = T)
> attach(d)
```

Le stime dei parametri del modello di regressione lineare vengono ottenute mediante il seguente comando:

```
> lm(Level ~ Sunspot)

Call:
lm(formula = Level ~ Sunspot)

Coefficients:
(Intercept)      Sunspot
    -8.0418       0.4128
```

Il diagramma di dispersione con retta di regressione stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Sunspot, Level, xlab = "Number of sunspot",
+    ylab = "Level (meters)", main = "Scatter plot")
> abline(lm(Level ~ Sunspot))
```

I precedenti comandi forniscono il grafico della Figura 4.4.1. □

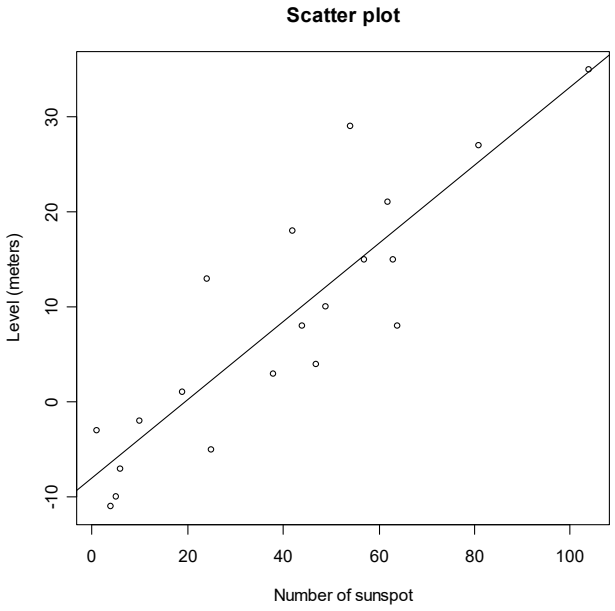


Figura 4.4.1.

Più generalmente, supponendo per semplicità un campione casuale da una singola variabile, i metodi di stima per un parametro θ possono essere basati sulla minimizzazione di una generica funzione obiettivo del tipo

$$\phi(\theta) = \sum_{i=1}^{n} \rho(x_i - \theta) \, ,$$

dove ρ è una opportuna funzione di distanza. Sotto alcune condizioni, la stima basata sulla minimizzazione della funzione obiettivo è equivalente alla (pseudo) soluzione dell'equazione

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

dove $\psi = \rho'$. Gli stimatori basati su questa procedura sono detti stimatori di tipo M.

• **Esempio 4.4.2.** Se θ è un parametro di posizione, allora si può scegliere la funzione di distanza $\rho(x) = x^2$. In questo caso, lo stimatore di θ risulta $\widetilde{\Theta} = \bar{X}$, ovvero la media campionaria. Se invece la funzione di distanza risulta $\rho(x) = |x|$, lo stimatore di θ è dato da $\widetilde{\Theta} = \widetilde{X}_{0.5}$, ovvero la mediana campionaria. Se θ è di nuovo un parametro di posizione, supponendo un approccio classico con un campione casuale, sia $f(x - \theta)$ la funzione di densità di X. In questo caso, lo stimatore di massima verosimiglianza di θ è uno stimatore di tipo M, dove $\rho(x) = \log f(x)$. Evidentemente, anche il metodo dei minimi quadrati si basa su una funzione di distanza del tipo $\rho(x) = x^2$. □

4.5. Metodi di stima Bayesiani

L'approccio all'inferenza statistica considerato sinora è basato sul cosiddetto paradigma frequentista. In questo approccio si assume che il campione rappresenti l'unica fonte di informazione. Il paradigma Bayesiano all'inferenza statistica assume invece che vi sia una distribuzione a priori che descrive la "fiducia" di osservare un valore θ prima di estrarre il campione. Una volta osservato il campione, si ottiene la cosiddetta distribuzione a posteriori che rappresenta la distribuzione sui possibili valori di θ quando viene considerata l'informazione fornita dal campione.

Assumendo che Θ sia una variabile casuale continua, sia f_Θ la funzione di densità che descrive la distribuzione a priori sullo spazio parametrico. In questo caso, la verosimiglianza viene espressa come distribuzione condizionata rispetto a Θ, ovvero come $f_{X_1,\ldots,X_n|\Theta=\theta}$. Per il Teorema di Bayes la distribuzione a posteriori è dunque fornita dalla funzione di densità condizionata

$$f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta) = c f_{X_1,\ldots,X_n|\Theta=\theta}(x_1, \ldots, x_n) f_\Theta(\theta) \, ,$$

dove c è una opportuna costante di proporzionalità (dipendente da $x_1, \ldots, x_n$), che assicura che la precedente funzione di densità sia in effetti tale.

Quando si dispone della distribuzione a posteriori, la stima di θ viene ottenuta come un indice di tendenza centrale. Ad esempio, se si assume la media della distribuzione a posteriori come metodo di stima si ha

$$\widetilde{\theta} = \mathrm{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \int_{-\infty}^{\infty} \theta f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta) d\theta \, .$$

La media della della distribuzione a posteriori è il valore che minimizza la funzione di perdita quadratica, ovvero

$$\widetilde{\theta} = \underset{\theta}{\text{argmin}} \int_{-\infty}^{\infty} (\theta - \vartheta)^2 f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\vartheta) d\vartheta \; .$$

Una stima alternativa di θ è data dalla mediana della della distribuzione a posteriori, ovvero il valore che minimizza la funzione di perdita assoluta data da

$$\widetilde{\theta} = \underset{\theta}{\text{argmin}} \int_{-\infty}^{\infty} |\theta - \vartheta| f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\vartheta) d\vartheta \; .$$

• **Esempio 4.5.1.** Dato un campione casuale da $X \sim B(1, \theta)$, si ha la verosimiglianza

$$L(\theta) = c \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \mathbf{1}_{\{0,1\}}(x_i) = c\, \theta^{n\overline{x}} (1-\theta)^{n-n\overline{x}} \mathbf{1}_{]0,1[}(\theta) \; .$$

In un approccio frequentista la stima di massima verosimiglianza di θ è data da $\widehat{\theta} = \overline{x}$. Dal momento che $\mathrm{E}[X] = \theta$ e $\mathrm{Var}[X] = \theta(1-\theta)$, lo stimatore di massima verosimiglianza $\widehat{\Theta} = \overline{X}$ è corretto con varianza

$$\mathrm{Var}[\widehat{\Theta}] = \frac{\theta(1-\theta)}{n} \; ,$$

che può essere stimata come

$$\widehat{\mathrm{Var}}[\widehat{\Theta}] = \frac{\overline{x}(1-\overline{x})}{n} \; .$$

In un approccio Bayesiano, si consideri come distribuzione a priori una distribuzione Beta di tipo $Be(0, 1; \alpha, \beta)$ con funzione di densità

$$f_{\Theta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \mathbf{1}_{]0,1[}(x) \; .$$

Questa scelta è giustificata dal fatto che la distribuzione Beta è molto flessibile ed è definita sullo spazio parametrico $\Theta = \{\theta : \theta \in]0, 1[\}$. La verosimiglianza viene espressa come

$$f_{X_1,\ldots,X_n|\Theta=\theta}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \mathbf{1}_{]0,1[}(\theta) = \theta^{n\overline{x}} (1-\theta)^{n-n\overline{x}} \mathbf{1}_{]0,1[}(\theta) \; ,$$

mentre la distribuzione a posteriori è data da

$$f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta) = c\, \theta^{n\overline{x}+\alpha-1} (1-\theta)^{n-n\overline{x}+\beta-1} \mathbf{1}_{]0,1[}(\theta) \; .$$

La precedente espressione è effettivamente una funzione di densità, se la costante di proporzionalità è pari a

$$c = \frac{1}{\int_0^1 \theta^{n\overline{x}+\alpha-1} (1-\theta)^{n-n\overline{x}+\beta-1} d\theta} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(n\overline{x}+\alpha)\Gamma(n-n\overline{x}+\beta)} \; ,$$

ovvero la distribuzione a posteriori è una distribuzione Beta di tipo $Be(0, 1; n\overline{x}+\alpha, n-n\overline{x}+\beta)$. Sia la distribuzione a priori, che quella a posteriori, sono della stessa famiglia e per questo motivo le distribuzioni sono dette coniugate con la distribuzione Binomiale. La scelta $\alpha = \beta = 1$, ovvero quella relativa ad un distribuzione a priori di tipo Uniforme, è detta non informativa in quanto tutti i valori di θ vengono ritenuti ugualmente probabili prima del campionamento. Per le proprietà della distribuzione Beta, la media della distribuzione a posteriori è data da

$$\widetilde{\theta} = \mathrm{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{n\overline{x} + \alpha}{n + \alpha + \beta} ,$$

mentre

$$\mathrm{Var}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{(n\overline{x} + \alpha)(n - n\overline{x} + \beta)}{(n + \alpha + \beta)^2 (n + \alpha + \beta - 1)} .$$

Per n elevato la media della distribuzione a posteriori tende a coincidere con la stima di massima verosimiglianza $\widehat{\theta}$, mentre la varianza della distribuzione a posteriori tende a coincidere con la varianza stimata $\widehat{\mathrm{Var}}[\widehat{\Theta}]$ dello stimatore di massima verosimiglianza. Ovviamente, l'interpretazione di queste quantità risulta molto differente nei due paradigmi. □

• **Esempio 4.5.2.** Si dispone dei risultati di un indagine statistica compresa nel "2016 General Social Survey (GSS)" eseguita dal National Opinion Research Center (NORC) presso l'University of Chicago e condotta su $n = 1810$ rispondenti alla domanda sulla possibilità di ridurre i vincoli sull'interruzione legale della gravidanza (Fonte: Agresti, A., 2017, *An Introducion to Categorical Data Analysis*, terza edizione, Wiley, New York, p.7). Fra i 1810 rispondenti 837 erano favorevoli e la stima di massima verosimiglianza è data da $\widehat{\theta} = 837/1810 = 0.4624$ con $\widehat{\mathrm{Var}}[\widehat{\Theta}]^{1/2} = 0.0117$ come risulta dai seguenti comandi:

```
> n = 1810
> t = 837
> test <- t/n
> vtest <- test * (1 - test)/n
> test
[1] 0.4624309
> vtest^(1/2)
[1] 0.01171929
```

Assumendo una distribuzione a priori di tipo $Be(0,1; 1, 1)$, ovvero una distribuzione non informativa, la media e lo scarto della distribuzione a posteriori sono quasi identiche alla stima di massima verosimiglianza e alla relativa stima dello scarto quadratico medio, come risulta dai seguenti comandi:

```
> alpha = 1
> beta = 1
> test <- (t + alpha)/(n + alpha + beta)
> vtest <- (t + alpha)*(n - t + beta)/
+    (n + alpha + beta)^2/(n + alpha + beta -1)
> test
[1] 0.4624724
> vtest^(1/2)
[1] 0.01171613
```

I grafici delle funzione di densità relative alla distribuzione a priori e a posteriori vengono ottenute mediante i seguenti comandi:

```
> plot(function(x) dbeta(x, alpha, beta), -0.1, 1.1, lty = 1,
+    xlab = "", ylab = "Density", main = "Prior distribution")
> plot(function(x) dbeta(x, t + alpha, n - t + beta),
+    0.4, 0.55, lty = 1,
+    xlab = "", ylab = "Density", main = "Posterior distribution")
```

I precedenti comandi forniscono il grafico della Figura 4.5.1.

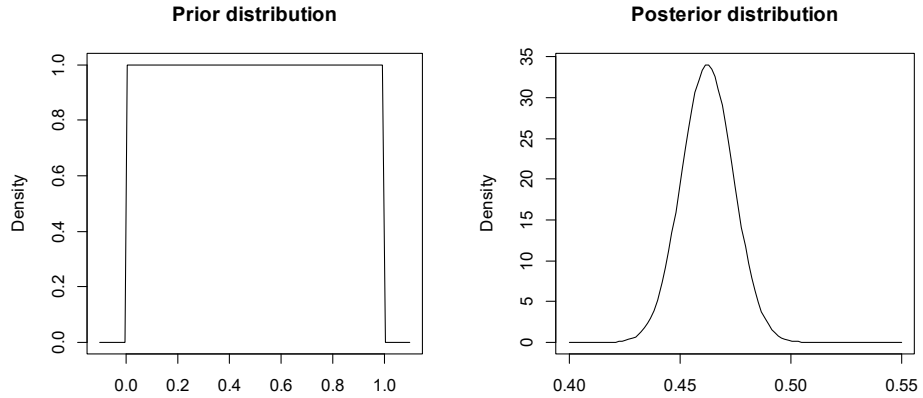


Figura 4.5.1.

Assumendo una distribuzione a priori di tipo $Be(0,1;100,1)$, ovvero una distribuzione per cui si ha notevole fiducia che i rispondenti siano estremamente favorevoli, la media e lo scarto della distribuzione a posteriori sono rispettivamente date da 0.4903 e 0.0114, come risulta dai seguenti comandi:

```
> alpha = 100
> beta = 1
> test <- (t + alpha)/(n + alpha + beta)
> vtest <- (t + alpha)*(n - t + beta)/
+    (n + alpha + beta)^2/(n + alpha + beta -1)
> test
[1] 0.4903192
> vtest^(1/2)
[1] 0.01143857
```

I grafici delle funzione di densità relative alla distribuzione a priori e a posteriori vengono ottenute di nuovo mediante i seguenti comandi:

```
> plot(function(x) dbeta(x, alpha, beta), -0.1, 1.1, lty = 1,
+    xlab = "", ylab = "Density", main = "Prior distribution")
> plot(function(x) dbeta(x, t + alpha, n - t + beta),
+    0.4, 0.55, lty = 1, xlab = "",
+    ylab = "Density", main = "Posterior distribution")
```

I precedenti comandi forniscono il grafico della Figura 4.5.2. □

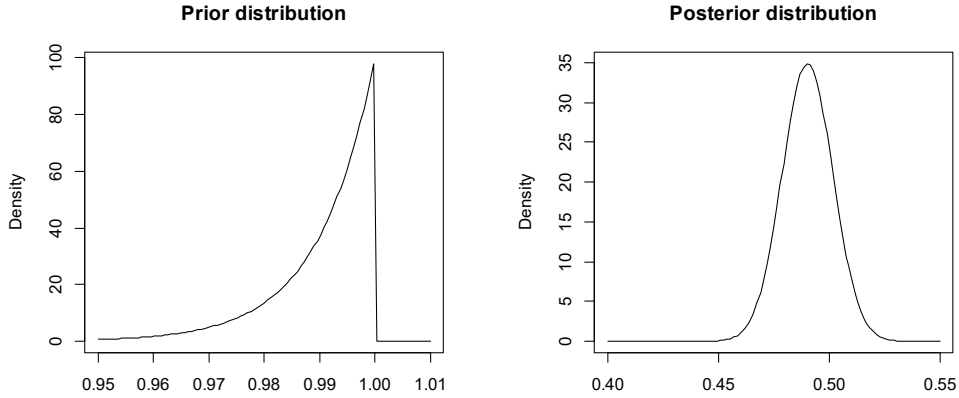


Figura 4.5.2.

4.6. Riferimenti bibliografici

Albert, J. (2009) *Bayesian Computation with R*, seconda edizione, Springer, New York.
Bernardo, J.M. e Smith, A.F.M. (2000) *Bayesian Theory*, Wiley, New York.
Boos, D.D. e Stefanski, L.A. (2013) *Essential Statistical Inference*, Springer, New York.
Cox, D.R. e Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.
Demidenko, E. (2020) *Advanced Statistics with Applications in R*, Wiley, New York.
Ferguson, T.S. (1996) *A Course in Large Sample Theory*, Chapman and Hall, London.
Gelman, A., Carlin, J.B., Stern, H.S. e Dunson, D.B. (2014) *Bayesian Data Analysis*, terza edizione, Chapman & Hall/CRC Press, Boca Raton.
Huber, P.J. e Ronchetti, E.M. (2009) *Robust Statistics*, seconda edizione, Wiley, New York.
Lauritzen, S. (2023) *Fundamentals of Mathematical Statistics*, Chapman & Hall/CRC Press, Boca Raton.
Lehmann, E.L. (1999) *Elements of Large Sample Theory*, Springer, New York.
Lehmann, E.L. e Casella, G. (1998) *The Theory of Point Estimation*, seconda edizione, Springer, New York.
Pruim, R. (2018) *Foundations and Applications of Statistics*, American Mathematical Society, seconda edizione, Providence.
Robert, C.P. (2007) *The Bayesian Choice*, seconda edizione, Springer, New York.
Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, seconda edizione, Wiley, New York.
Shao, J. (2003) *Mathematical Statistics*, seconda edizione, Springer, New York.
Schervish, M.J. (1995) *Theory of Statistics*, Springer, New York.
Schweder, T. e Hjort, N.L. (2016) *Confidence, Likelihood, Probability*, Cambridge University Press, Cambridge.
Tattar, P.N., Ramaiah, S. e Manjunath, B.G. (2016) *A course in Statistics with R*, Wiley, New York.
Wilks, S.S. (1962) *Mathematical Statistics*, Wiley, New York.

Capitolo 5

I metodi di smorzamento

5.1. Lo stimatore di nucleo

Quando si analizza una variabile casuale continua è conveniente effettuare una indagine esplorativa della rispettiva funzione di densità, eventualmente finalizzata alla selezione di un modello. Tuttavia, in questo ambito, l'istogramma fornisce informazioni sulla funzione di densità in modo piuttosto grossolano. Una tecnica più raffinata per stimare la funzione di densità si basa sullo stimatore di nucleo.

Sia $X_1, \ldots, X_n$ un campione casuale da una variabile casuale (assolutamente) continua X con funzione di densità f. Lo stimatore di nucleo per f nel punto x è dato da

$$\widehat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \,,$$

dove

$$K_h(x) = \frac{1}{h} K(h^{-1}x) \,,$$

mentre $h > 0$ è detto parametro di smorzamento. La funzione K è detta nucleo ed è tale che

$$\int_{\mathbb{R}} K(x)dx = 1 \,.$$

Una giustificazione della genesi di questo stimatore può essere data attraverso la seguente rappresentazione ingenua di $f(x)$

$$f(x) = \int_{\mathbb{R}} \mathbf{1}_{\{0\}}(x - y) f(y) dy = \mathrm{E}[\mathbf{1}_{\{0\}}(x - X)] \,.$$

Per $h \to 0$ si ha $K_h(x) \to \mathbf{1}_{\{0\}}(x)$ e dunque dalla precedente espressione si ha la seguente approssimazione

$$\mathrm{E}[\mathbf{1}_{\{0\}}(x - X)] \simeq \mathrm{E}[K_h(x - X)] \,.$$

In base al principio di corrispondenza, $\mathrm{E}[K_h(x - X)]$ può essere dunque stimato mediante $\widehat{f}_h(x)$.

Usualmente il nucleo K viene selezionato come una funzione di densità simmetrica. Questa assunzione assicura che $\widehat{f}_h$ sia a sua volta una funzione di densità. Una scelta comune per K è la funzione di densità di una variabile casuale Normale standard. Una selezione alternativa è rappresentata dalla funzione di densità della cosiddetta variabile casuale "biweight" standard, ovvero

$$K(x) = \frac{15}{16} (1 - x^2)^2 \, \mathbf{1}_{[-1,1]}(x) \,.$$

Il parametro h controlla la quantità di smorzamento applicata allo stimatore di nucleo. All'aumentare di h la stima risulta più "liscia", mentre al diminuire di h la stima diventa più "rugosa"

e tende alla funzione di densità empirica, ovvero alla distribuzione di probabilità che pone una probabilità pari a $1/n$ su ogni osservazione.

• **Esempio 5.1.1.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. La stima di nucleo viene ottenuta richiamando la libreria `sm` che permette di implementare metodi di smorzamento avanzati. In particolare, i grafici della stima di nucleo per $h = 1.00, 0.33, 0.05$ vengono ottenuti mediante i seguenti comandi:

```
> sm.density(Diameter, 1.00, yht = 2, xlim = c(-0.35, 2.65),
+    xlab = "Ball diameter (micron)")
> title(main = "Kernel density estimation (h = 1.00)")
> sm.density(Diameter, 0.33, yht = 2, xlim = c(-0.35, 2.65),
+    xlab = "Ball diameter (micron)")
> title(main = "Kernel density estimation (h = 0.33)")
> sm.density(Diameter, 0.05, yht = 2, xlim = c(-0.35, 2.65),
+    xlab = "Ball diameter (micron)")
> title(main = "Kernel density estimation (h = 0.05)")
```

I precedenti comandi forniscono i grafici della Figura 5.1.1. Risulta evidente come differenti scelte del parametro di smorzamento forniscano stime della funzione di densità molto differenti. □

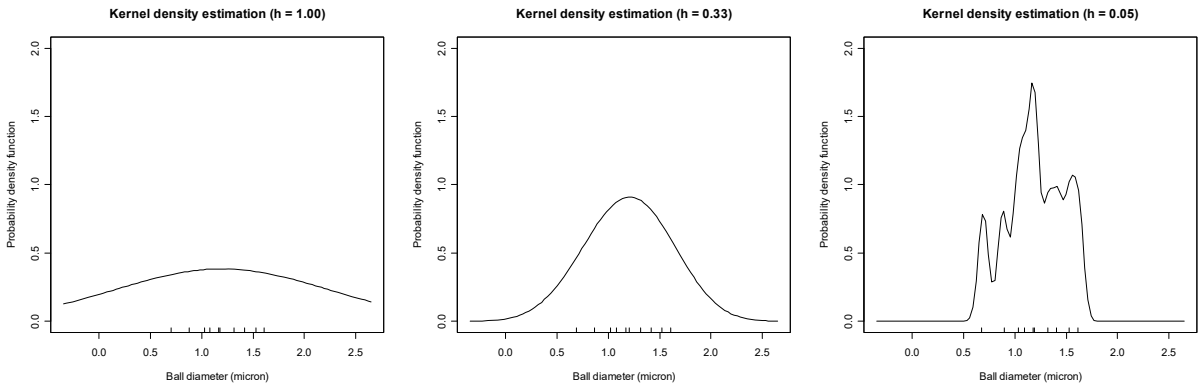


Figura 5.1.1.

La precisione di uno stimatore di nucleo $\widehat{f}_h$ nel punto x viene misurata attraverso l'errore quadratico medio, ovvero

$$\text{MSE}[\widehat{f}_h(x)] = \text{Bias}[\widehat{f}_h(x)]^2 + \text{Var}[\widehat{f}_h(x)] \, .$$

Dal momento che usualmente si richiede la stima sull'intero supporto della variabile casuale, una misura globale della precisione di $\widehat{f}_h$ è data dall'errore quadratico medio integrato, ovvero

$$\text{MISE}[\widehat{f}_h] = \int_{-\infty}^{\infty} \text{MSE}[\widehat{f}_h(x)] dx \, .$$

Si assuma che f'' esista per ogni x e sia continua ed integrabile e che per una generica funzione g

$$\mu_2(g) = \int_{-\infty}^{\infty} x^2 g(x) dx < \infty$$

e

$$R(g) = \int_{-\infty}^{\infty} g(x)^2 dx < \infty \, .$$

Sotto queste condizioni si può dimostrare che per $h \to 0$ si ha

$$\mathrm{E}[\widehat{f}_h(x)] \simeq f(x) + \frac{1}{2}\, h^2 f''(x)\mu_2(K)\,.$$

Quindi, $\widehat{f}_h(x)$ è uno stimatore distorto la cui distorsione tende a 0 quando $h \to 0$. Inoltre, per $h \to 0$ e $nh \to \infty$, si può dimostrare che

$$\mathrm{Var}[\widehat{f}_h(x)] \simeq \frac{1}{nh}\, R(K) f(x)\,.$$

Dunque, $\widehat{f}_h(x)$ è uno stimatore coerente di $f(x)$ se $h \to 0$ e $nh \to \infty$ quando $n \to \infty$. Tenendo presente le precedenti espressioni, il cosiddetto errore quadratico medio per grandi campioni risulta

$$\mathrm{AMSE}[\widehat{f}_h(x)] = \frac{1}{4}\, h^4 f''(x)^2 \mu_2(K)^2 + \frac{1}{nh}\, R(K) f(x)\,,$$

per cui l'errore medio quadratico integrato per grandi campioni è dato da

$$\mathrm{AMISE}[\widehat{f}_h] = \int_{-\infty}^{\infty} \mathrm{AMSE}[\widehat{f}_h(x)]dx = \frac{1}{4}\, h^4 \mu_2(K)^2 R(f'') + \frac{1}{nh}\, R(K)\,.$$

Risulta semplice verificare che AMISE è minimizzato quando

$$h = \left(\frac{R(K)}{\mu_2(K)^2 R(f'')}\right)^{1/5} n^{-1/5}\,,$$

mentre

$$\min_{h>0} \mathrm{AMISE}[\widehat{f}_h] \propto n^{-4/5}\,.$$

Si noti che, mentre la scelta del nucleo è quasi ininfluente nella stima della funzione di densità, risulta fondamentale la selezione del parametro di smorzamento. Quando questa selezione viene effettuata sulla base dei dati campionari, si ha una scelta automatica del parametro di smorzamento. Le quantità $\mathrm{MISE}[\widehat{f}_h]$ e $\mathrm{AMISE}[\widehat{f}_h]$ dipendono da f e quindi non è possibile adoperarle per la selezione ottima di h. Quindi, si deve adottare opportune stime di queste quantità per implementare selettori da adoperare in pratica.

Una prima classe di selettori del parametro di smorzamento è basata sulla minimizzazione di una opportuna stima di $\mathrm{MISE}[\widehat{f}_h]$. Il principale metodo basato su questo criterio è la cosiddetta "cross-validation". Una seconda classe di selettori del parametro di smorzamento è invece basata sulla minimizzazione di una opportuna stima di $\mathrm{AMISE}[\widehat{f}_h]$. Il principale metodo basato su questo criterio è il cosiddetto "plug-in". Questi metodi sono comunemente implementati nei principali pacchetti per l'elaborazione dei dati.

• **Esempio 5.1.2.** Si considera di nuovo i dati relativi ai diametri delle sfere dell'Esempio 4.3.4. I grafici della stima di nucleo con i selettori basati sui metodi "cross-validation" e "plug-in" si ottengono mediante i seguenti comandi:

```
> library(sm)
> sm.density(Diameter, hcv(Diameter, hstart = 0.01, hend = 1),
+    yht = 0.92, xlim = c(-0.35, 2.65),
+    xlab = "Ball diameter (micron)")
> title(main = "Kernel density estimation ('CV' h = 0.32)")
> sm.density(Diameter, hsj(Diameter), yht = 1.06,
+    xlim = c(-0.05, 2.35), xlab = "Ball diameter (micron)")
> title(main = "Kernel density estimation ('Plug-in' h = 0.23)")
```

I due grafici sono riportati rispettivamente nella Figura 5.1.2 e nella Figura 5.1.3. In questo caso, i due selettori forniscono stime simili della funzione di densità. □

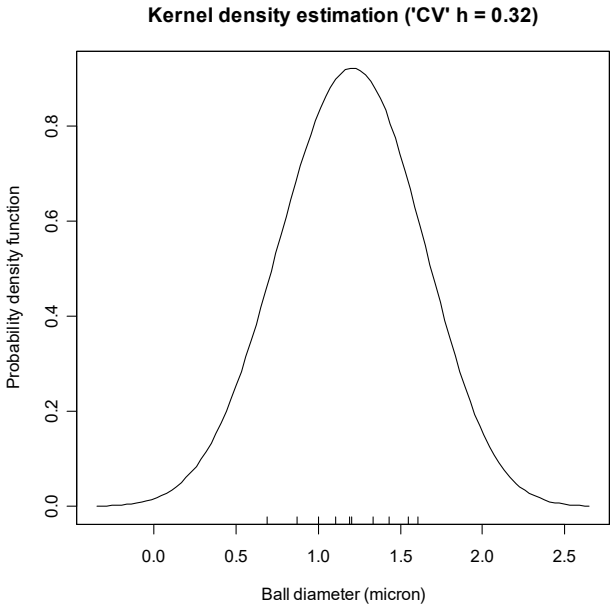


Figura 5.1.2.

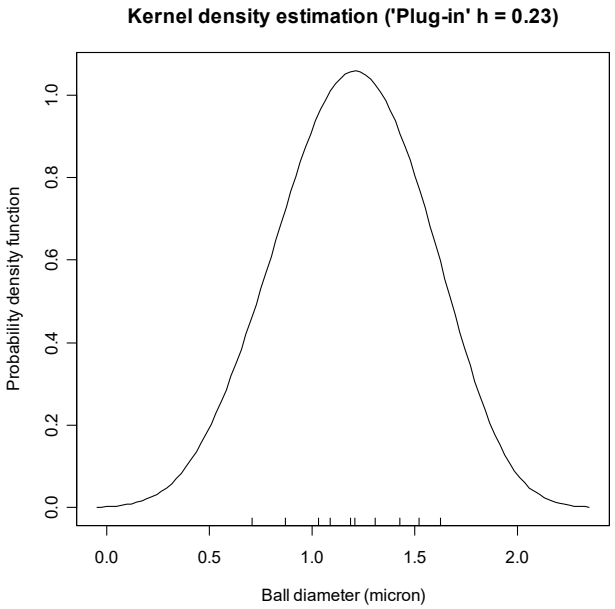


Figura 5.1.3.

• **Esempio 5.1.3.** Si dispone di un campione casuale delle durate delle eruzioni (in minuti) di un geyser nel parco nazionale di Yellowstone (Fonte: Silverman, B.W., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, p.8). I dati sono contenuti nel file `geyser.txt` e vengono resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\geyser.txt", header = T)
> attach(d)
```

I grafici della stima di nucleo con i selettori basati sui metodi della “cross-validation” e del “plug-in” si ottengono mediante i seguenti comandi:

```
> library(sm)
> sm.density(Duration, hcv(Duration, hstart = 0.01, hend = 1),
+    yht = 0.69, xlim = c(1.4, 5.1), xlab = "Waiting time(minutes)")
> title(main = "Kernel density estimation ('CV' h = 0.10)")
> sm.density(Duration, hsj(Duration), yht = 0.69,
+    xlim = c(1.4, 5.1), xlab = "Waiting time (minutes)")
> title(main = "Kernel density estimation ('Plug-in' h = 0.20)")
```

I due grafici sono riportati rispettivamente nella Figura 5.1.4 e nella Figura 5.1.5. I due selettori forniscono stime piuttosto differenti della funzione di densità, anche se i rispettivi valori del parametro di smorzamento non sono troppo dissimili. □

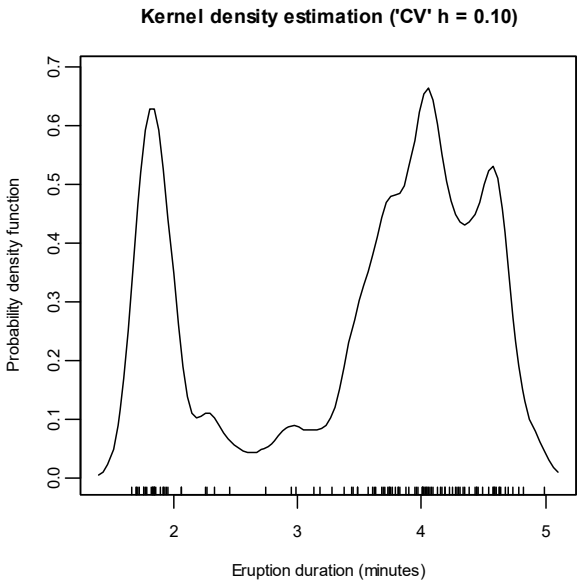


Figura 5.1.4.

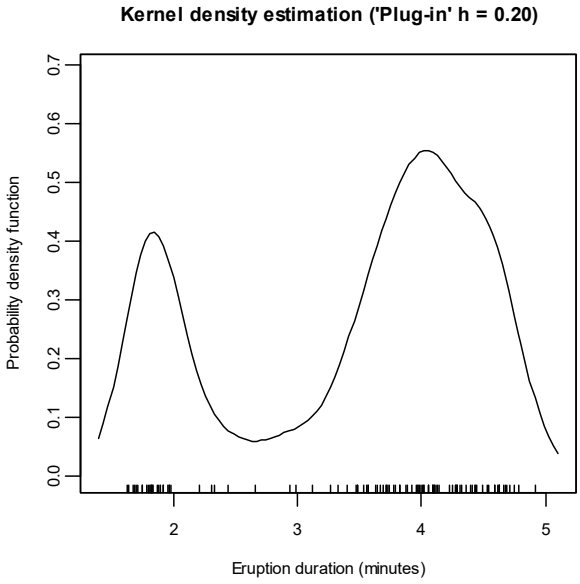


Figura 5.1.5.

Nel derivare le proprietà dello stimatore di nucleo per grandi campioni si è assunto che f'' sia continua. Tuttavia, è frequente che perfino f non sia continua. Ad esempio, molte funzioni di densità sono discontinue in un punto estremo del relativo supporto. Supponendo per semplicità (ma senza perdita di generalità) che il supporto di f sia $[0, \infty[$ e che la discontinuità si trovi nell'origine, se si vuole stimare $f(0)$ è facile verificare che $\widehat{f}_h(0)$ è distorto anche se $h = 0$. Al fine di evitare difficoltà di stima di questo tipo si preferisce (quando possibile) considerare una opportuna variabile casuale trasformata $t(X)$ con funzione di densità g e supporto $\mathbb{R}$, dove t è una trasformazione monotona. Dal momento che per le proprietà delle trasformazioni di variabili casuali si ha

$$f(x) = g[t(x)]t'(x) \, ,$$

si può stimare g sulla base delle osservazioni trasformate $t(X_1), \ldots, t(X_n)$ e lo stimatore di nucleo di f si riduce a

$$\widehat{f}_h(x) = \frac{t'(x)}{n} \sum_{i=1}^{n} K_h(t(x) - t(X_i)) \, .$$

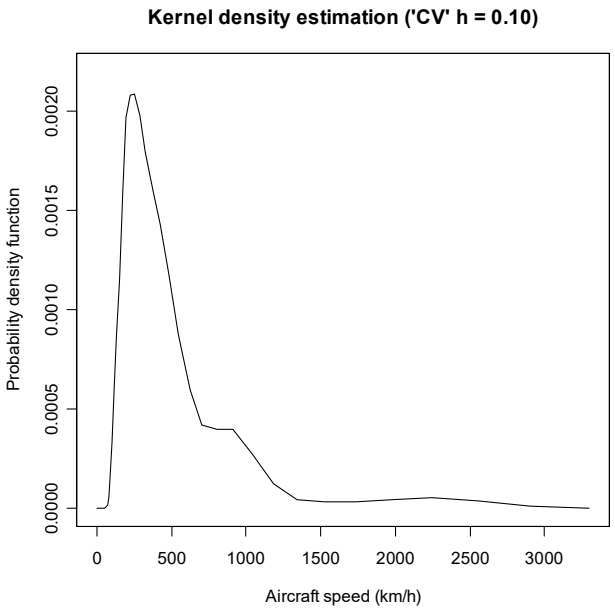


Figura 5.1.6.

• **Esempio 5.1.4.** Si dispone di un campione casuale di velocità massime in chilometri orari di aerei costruiti fra il 1914 e il 1984 (Fonte: Saviotti, P.P. e Bowman, A.W., 1984, Indicators of output of technology, in *Proceedings of the ICSSR/SSRC Workshop on Science and Technology in the 1980's*, M. Gibbons *et al.*, eds., Harvester Press, Brighton). I dati sono contenuti nel file `aircraft.txt` e vengono resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\aircraft.txt", header = T)
> attach(d)
```

Assumendo una trasformata logaritmica delle osservazioni, il grafico della stima di nucleo con il selettore basato sul metodo della "cross-validation" si ottiene mediante i seguenti comandi:

```
> library(sm)
> sm.density(Speed, hcv(log(Speed), hstart = 0.01, hend = 1),
+    yht = 0.0022, xlim = c(0, 3300),
+    xlab = "Aircraft speed (km/h)", rugplot = F, positive = T)
> title(main = "Kernel density estimation ('CV' h = 0.10)")
```

Il relativo grafico è riportato nella Figura 5.1.6. □

5.2. Lo stimatore di nucleo bivariato

Sia $(X_1, Y_1), \ldots, (X_n, Y_n)$ un campione casuale da una variabile casuale continua (X, Y) con funzione di densità congiunta f. In modo analogo al caso univariato, lo stimatore di nucleo nel punto (x, y) può essere costruito come

$$\widehat{f}_{h_1,h_2}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(x - X_i) K_{h_2}(y - Y_i) \,,$$

dove K è un nucleo, mentre $h_1, h_2 > 0$ sono due parametri di smorzamento. In una formulazione più generale si potrebbe adoperare anche una funzione di nucleo bivariata (con tre parametri di smorzamento) invece di un prodotto di due funzioni di nucleo marginali. La presente formulazione è tuttavia conveniente e sufficiente nelle applicazioni pratiche.

Le proprietà dello stimatore di nucleo bivariato si possono ottenere in modo analogo a quelle dello stimatore di nucleo univariato. Si tenga presente tuttavia che la precisione dello stimatore di nucleo bivariato diminuisce rispetto alla controparte univariata. Questo fenomeno, noto come "maledizione della dimensionalità", è dovuto al fatto che n osservazioni si rarefanno all'aumentare della dimensione dello spazio di riferimento. In effetti, si può dimostrare che per lo stimatore di nucleo multivariato in $\mathbb{R}^d$ il minimo di AMISE è proporzionale a $n^{-4/(d+4)}$, ovvero lo stimatore diventa rapidamente inefficiente all'aumentare di d.

• **Esempio 5.2.1.** Si dispone delle osservazioni relative ad alcune variabili per le guardie nel campionato professionistico di basket NBA nel 1992-93 (Fonte: Chatterjee, S., Handcock, M.S. e Simonoff, J.S., 1995, *A Casebook for a First Course in Statistics and Data Analysis*, Wiley, New York). Le variabili considerate sono state punti segnati per minuto giocato e assist per minuto giocato. I dati sono contenuti nel file `basket.txt` e vengono rese disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\basket.txt", header = T)
> attach(d)
```

I grafici (tridimensionale, per curve di livello e a toni di colori) della stima di nucleo bivariata si ottengono mediante i seguenti comandi:

```
> library(sm)
> sm.density(d[, c(1, 2)], hcv(d[, c(1, 2)]),
+    xlim = c(0, 0.9), ylim = c(0, 0.4), zlim = c(0, 20),
+    xlab = "Points per minute", ylab = "Assists per minute")
> title(main = "Kernel density estimation ('CV' h1 = 0.06,
+    h2 = 0.03)")
> plot(Score, Assist, xlim = c(0, 0.9), ylim = c(0, 0.4),
+    xlab = "Points per minute", ylab = "Assists per minute")
> sm.density(d[, c(1, 2)], hcv(d[, c(1, 2)]), display = "slice",
+    props = c(75, 50, 25, 2), add = T)
> title(main = "Kernel density estimation ('CV' h1 = 0.06,
+    h2 = 0.03)")
> sm.density(d[, c(1, 2)], hcv(d[, c(1, 2)]),
+    display = "image", xlim = c(0, 0.9), ylim = c(0, 0.4),
+    xlab = "Points per minute", ylab = "Assists per minute")
> title(main = "Kernel density estimation ('CV' h1 = 0.06,
+    h2 = 0.03)")
```

I tre grafici sono rispettivamente riportati nella Figura 5.2.1, nella Figura 5.2.2 e nella Figura 5.2.3. Da questi grafici si evidenzia una bimodalità della funzione di densità. □

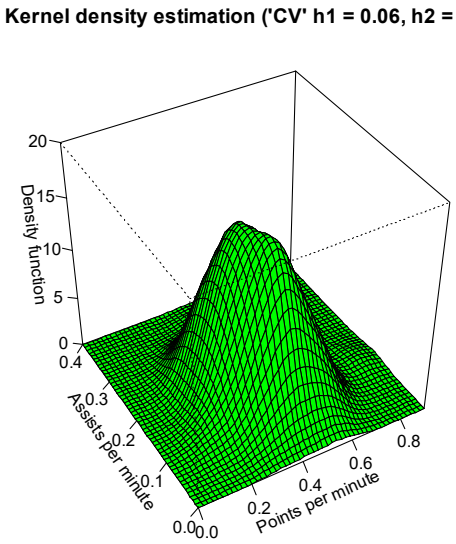


Figura 5.2.1.

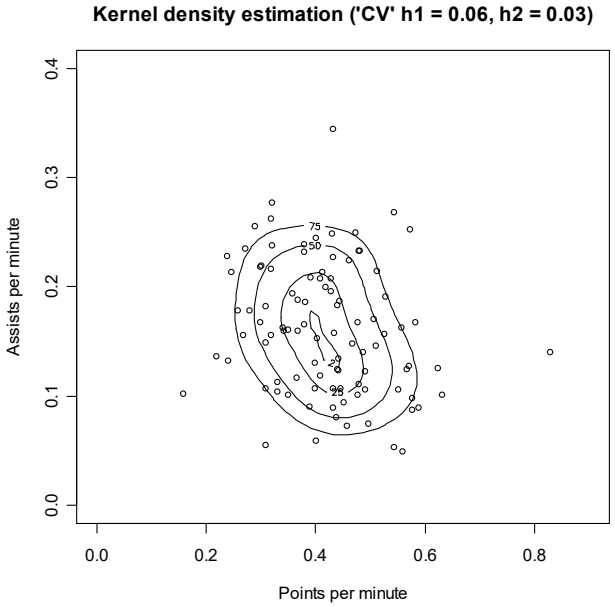


Figura 5.2.2.

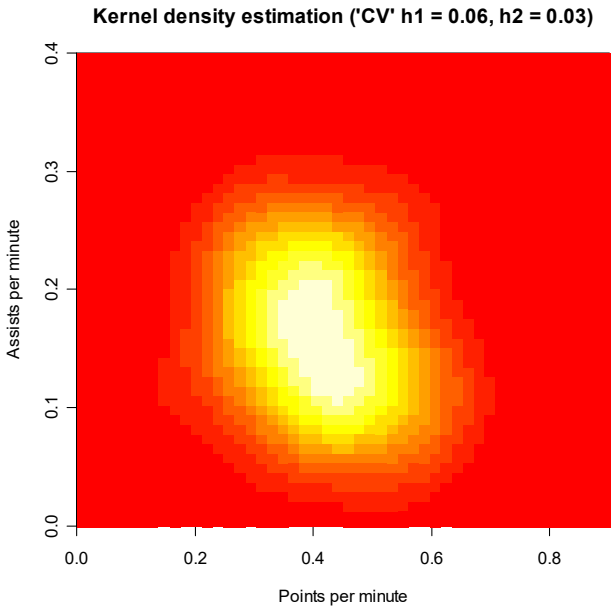


Figura 5.2.3.

• **Esempio 5.2.2.** Si dispone delle osservazioni relative alla larghezza e alla lunghezza della diagonale in millimetri dell'immagine contenuta in banconote svizzere per metà falsificate (Fonte: Flury, B. e Riedwyl, H., 1988, *Multivariate Statistics: a Practical Approach*, Chapman and Hall, London). I dati sono contenuti nel file `swissmoney.txt`. e vengono resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\swissmoney.txt", header = T)
> attach(d)
```

I grafici (tridimensionale, per curve di livello e a toni di colori) della stima di nucleo bivariata si ottengono mediante i seguenti comandi (le osservazioni relative alle banconote false vengono contrassegnate da punti in grassetto nel diagramma di dispersione):

```
> library(sm)
> sm.density(d[, c(2, 3)], hcv(d[, c(2, 3)]),
+    xlim = c(6, 14), ylim = c(137, 143), zlim = c(0, 0.2),
+    xlab = "Width (mm)", ylab = "Length (mm)")
> title(main = "Kernel density estimation ('CV' h1 = 0.35,
+    h2 = 0.25)")
> plot(d[1:100, 2], d[1:100, 3], xlim = c(6, 14),
+    ylim = c(137, 143), xlab = "Width (mm)", ylab = "Length (mm)")
> points(d[101:200, 2], d[101:200, 3], pch = 16)
> sm.density(d[, c(2, 3)], hcv(d[, c(2, 3)]),
+    display = "slice", props = c(75, 50, 25), add = T)
> title(main = "Kernel density estimation ('CV' h1 = 0.35,
+    h2 = 0.25)")
> sm.density(d[, c(2, 3)], hcv(d[, c(2, 3)]),
+    display = "image", xlim = c(6, 14), ylim = c(137, 143),
+    xlab = "Width (mm)", ylab = "Length (mm)")
> title(main = "Kernel density estimation ('CV' h1 = 0.35,
+    h2 = 0.25)")
```

I tre grafici sono rispettivamente riportati nella Figura 5.2.4, nella Figura 5.2.5 e nella Figura 5.2.6. Da questi grafici si evidenzia una multimodalità della funzione di densità. □

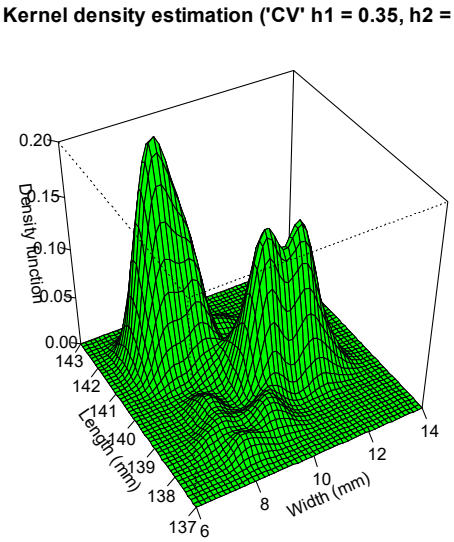


Figura 5.2.4.

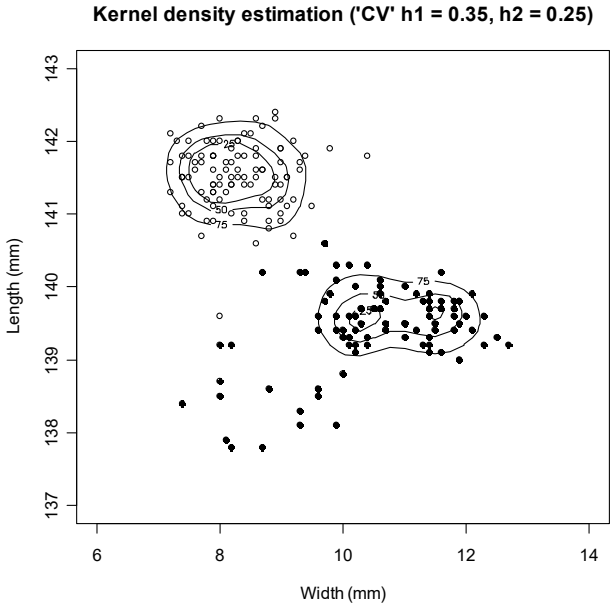


Figura 5.2.5.

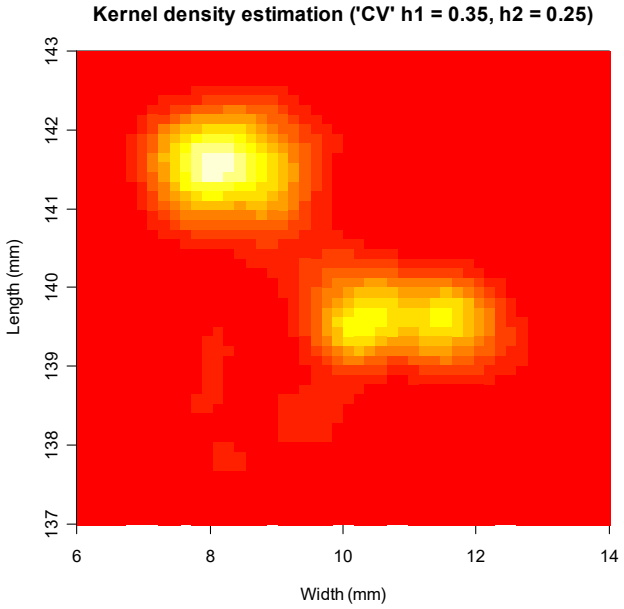


Figura 5.2.6.

5.3. La regressione lineare locale

Prima di adottare un modello di regressione per la relazione fra la variabile esplicativa e quella di risposta è conveniente indagare la natura del legame con metodi esplorativi. Un modo "distribution-free" per stimare la funzione di regressione è attraverso la regressione lineare locale. Se $Y_1, \ldots, Y_n$ sono le osservazioni della variabile di risposta per i livelli del regressore $x_1, \ldots, x_n$, il modello di regressione risulta

$$Y_i = m(x_i) + \mathcal{E}_i \,,$$

dove m è una funzione di regressione non nota, mentre $\mathrm{E}[\mathcal{E}_i] = 0$ e $\mathrm{Var}[\mathcal{E}_i] = \sigma^2$.

In generale, la funzione m non è lineare. Tuttavia, se m risulta abbastanza regolare, allora in un intorno di un punto x è approssimativamente lineare, ovvero si può assumere che $m(x) \simeq \beta_0 + \beta_1 x$ per valori prossimi ad x. La funzione obiettivo smorzata localmente nel punto x da adottare per il metodo dei minimi quadrati è data da

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} K_h(x_i - x)(y_i - \beta_0 - \beta_1(x_i - x))^2 \,,$$

dove la funzione K_h è definita analogamente allo stimatore di nucleo della funzione di densità. Senza perdita di generalità e per semplicità di notazione, i valori del regressore sono stati centrati rispetto al punto x. Minimizzando la funzione obiettivo si ottengono delle stime locali di β_0 e β_1 nel punto x, che forniscono di conseguenza il seguente stimatore di $m(x)$

$$\widehat{m}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{(s_{2,h}(x) - s_{1,h}(x)(x_i - x))K_h(x_i - x)Y_i}{s_{2,h}(x)s_{0,h}(x) - s_{1,h}(x)^2} \,,$$

dove

$$s_{r,h}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^r K_h(x_i - x)\ .$$

Anche in questo caso, il parametro h controlla il livello di smorzamento, ovvero quanto locale deve essere la stima di m. Per $h \to \infty$ la stima di m coincide con quella ottenuta con il metodo dei minimi quadrati quando si assume un modello lineare, mentre per $h = 0$ si ottiene una spezzata che congiunge i punti sul piano cartesiano.

• **Esempio 5.3.1.** Si dispone delle osservazioni per tre variabili misurate su alcuni motori a etanolo, ovvero la concentrazione di ossido di nitrogeno (in microgrammi/J), il rapporto di compressione e il rapporto di equivalenza che è una misura della ricchezza della miscela di aria e etanolo (Fonte: Brinkman, N.D., 1981, Ethanol fuel - a single-cylinder engine study of efficiency and exhaust emissions, *SAE Transactions* **90**, 1410-1424). La variabile di risposta è la concentrazione di ossido di nitrogeno, mentre il regressore è il rapporto di equivalenza. I dati sono contenuti nel file `ethanol.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\ethanol.txt", header = T)
> attach(d)
```

La stima della funzione di regressione viene ottenuta richiamando la libreria `sm`. In particolare, i grafici della stima della funzione di regressione per $h = 1.00, 0.05, 0.01$ vengono ottenuti mediante i seguenti comandi:

```
> library(sm)
> plot(Equivalence, NOx, xlab = "Equivalence ratio",
+    ylab = "Concentration of nitrogen oxides (micrograms/J)")
> sm.regression(Equivalence, NOx, h = 1.00, add = T)
> title(main = "Local linear regression (h = 1.00)")
> plot(Equivalence, NOx, xlab = "Equivalence ratio",
+    ylab = "Concentration of nitrogen oxides (micrograms/J)")
> sm.regression(Equivalence, NOx, h = 0.05, add = T)
> title(main = "Local linear regression (h = 0.05)")
> plot(Equivalence, NOx, xlab = "Equivalence ratio",
+    ylab = "Concentration of nitrogen oxides (micrograms/J)")
> sm.regression(Equivalence, NOx, h = 0.01, add = T)
> title(main = "Local linear regression (h = 0.01)")
```

I precedenti comandi forniscono i grafici della Figura 5.3.1. Risulta evidente come differenti scelte del parametro di smorzamento forniscano stime della funzione di regressione molto differenti. □

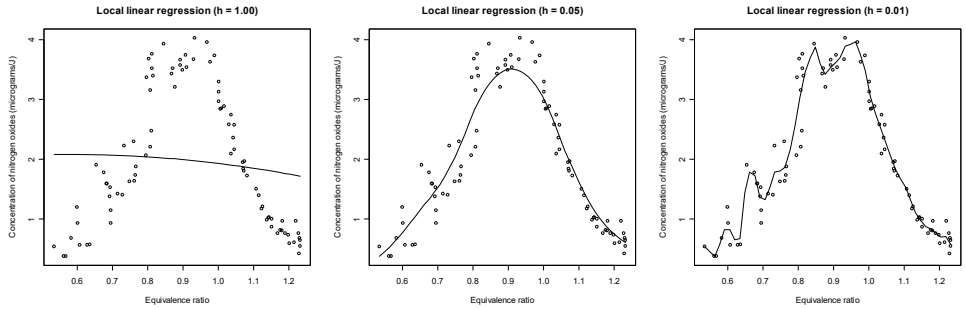


Figura 5.3.1.

Si assuma che m'' esista per ogni x, che i regressori siano generati da una variabile casuale continua con funzione di densità f e che valgano alcune opportune condizioni sulla disposizione dei regressori all'aumentare della numerosità campionaria. Si può dimostrare che per $h \to 0$ si ha

$$\mathrm{E}[\widehat{m}_h(x)] \simeq m(x) + \frac{1}{2} h^2 m''(x) \mu_2(K) \, .$$

Inoltre, per $h \to 0$ e $nh \to \infty$, si può anche dimostrare

$$\mathrm{Var}[\widehat{m}_h(x)] \simeq \frac{1}{nh} \frac{\sigma^2 R(K)}{f(x)} \, .$$

Dunque $\widehat{m}_h(x)$ è uno stimatore coerente di $m(x)$ se $h \to 0$ e $nh \to \infty$ quando $n \to \infty$. Anche per lo stimatore $\widehat{m}_h(x)$ si può definire l'errore quadratico medio integrato, ovvero

$$\mathrm{MISE}[\widehat{m}_h] = \int_{-\infty}^{\infty} \mathrm{MSE}[\widehat{m}_h(x)] dx$$

e si può ottenere l'errore medio quadratico integrato per grandi campioni, ovvero

$$\mathrm{AMISE}[\widehat{m}_h] = \frac{1}{4} h^4 \mu_2(K)^2 R(m'') + \frac{1}{nh} \frac{\sigma^2 R(K)}{f(x)} \, .$$

Al fine di stimare σ^2 è opportuno notare che la stima della funzione di regressione è lineare rispetto alle realizzazioni della variabile di risposta. Se la funzione di regressione viene stimata nei punti $x_1, \ldots, x_n$, si assuma che $\widehat{\mathbf{m}} = (\widehat{m}_h(x_1), \ldots, \widehat{m}_h(x_n))^{\mathrm{T}}$. Se $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, allora si può scrivere $\widehat{\mathbf{m}} = \mathbf{S}\mathbf{y}$, dove $\mathbf{S}$ è una matrice le cui righe contengono i pesi opportuni basati sui valori $s_{r,h}(x_i)$. In analogia con la regressione lineare multipla (vedi Capitolo 10), si può dunque definire lo stimatore

$$\widehat{\sigma}^2 = \frac{1}{\mathrm{df}_e} \sum_{i=1}^{n} (y_i - \widehat{m}_h(x_i))^2 \, ,$$

dove $\mathrm{df}_e = \mathrm{tr}(\mathbf{I} - \mathbf{S})$ rappresentano i gradi di libertà "approssimati" dell'errore.

In modo simile alla stima di nucleo della funzione di densità, la scelta del nucleo è quasi ininfluente, mentre risulta fondamentale la selezione del parametro di smorzamento. Di nuovo, la quantità $\mathrm{MISE}[\widehat{m}_h]$ dipende da m e quindi non è possibile adoperarla per la selezione ottima di h. Esistono comunque alcuni metodi per implementare selettori da adoperare in pratica. Una prima classe di selettori è basato sui gradi di libertà "approssimati". Una seconda classe di selettori è basata sulla minimizzazione di una opportuna stima di $\mathrm{MISE}[\widehat{f}_h]$, ovvero sul metodo "cross-validation".

• **Esempio 5.3.2.** Si considera di nuovo i dati relativi ai motori a etanolo dell'Esempio 5.3.1. I grafici della stima della funzione di regressione con i selettori basati sui metodi dei gradi di libertà "approssimati" e della "cross-validation" si ottengono mediante i seguenti comandi:

```
> library(sm)
> plot(Equivalence, NOx, xlab = "Equivalence ratio",
+    ylab = "Concentration of nitrogen oxides")
> sm.regression(Equivalence, NOx, method = "df", add = TRUE)
> plot(Equivalence, NOx, xlab = "Equivalence ratio",
+    ylab = "Concentration of nitrogen oxides")
> sm.regression(Equivalence, NOx, method = "cv", add = TRUE)
```

I precedenti comandi forniscono i grafici della Figura 5.3.2 e della Figura 5.3.3. □

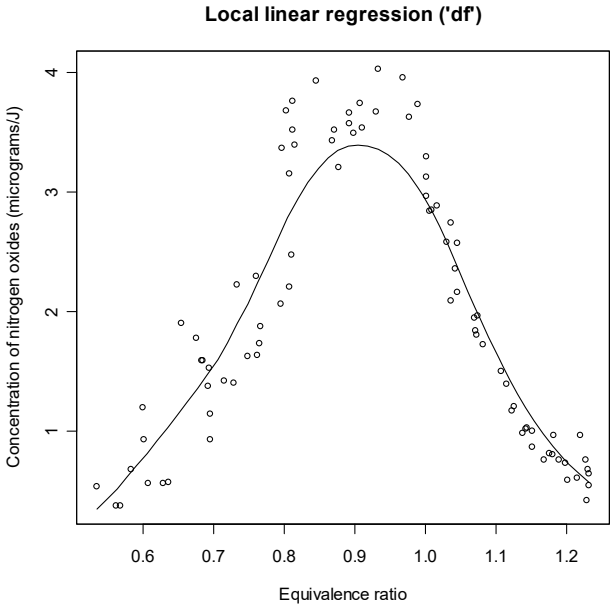


Figura 5.3.2.

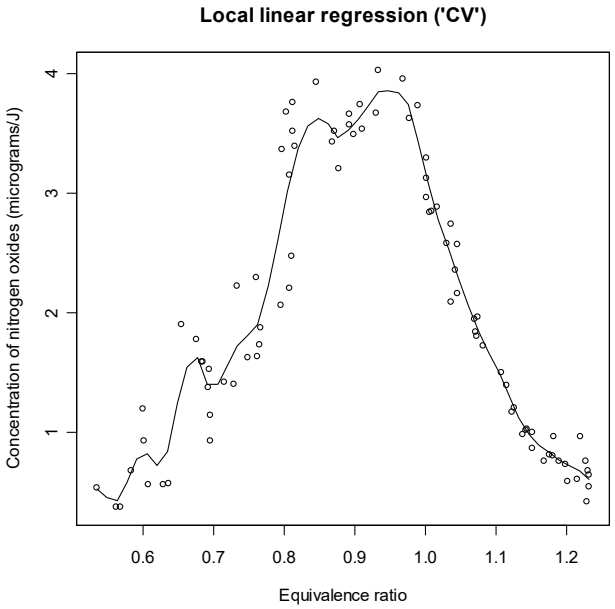


Figura 5.3.3.

Un approccio alternativo alla regressione lineare locale è basato su un parametro di smorzamento variabile per ogni punto x. Più esattamente, si considera la minimizzazione della funzione criterio basata su una funzione di nucleo con parametro di smorzamento variabile del tipo

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} K_{d_k(x_i)}(x_i - x)(y_i - \beta_0 - \beta_1(x_i - x))^2 ,$$

dove $d_k(x_i)$ è la distanza di x_i dal k-esimo vicino più prossimo dei restanti valori del regressore. Questo metodo è detto `loess`. Il metodo `loess` evita la scelta di un selettore e si limita a richiedere la specificazione del parametro k. Il parametro k è evidentemente legato alla proporzione del campione che contribuisce al peso attribuito per ogni punto x. Una scelta grossolana di questo parametro è solitamente sufficiente e l'usuale scelta di compromesso risulta $k = \lfloor 0.5n \rfloor$, dove $\lfloor x \rfloor$ rappresenta la funzione di troncamento.

● **Esempio 5.3.3.** Si considera di nuovo i dati relativi ai motori a etanolo dell'Esempio 5.3.1. Il grafico della stima della funzione di regressione con il metodo `loess` si può ottenere mediante i seguenti comandi:

```
> plot(Equivalence, NOx, xlab = "Equivalence ratio",
+    ylab = "Concentration of nitrogen oxides (micrograms/J)")
> od <- d[order(Equivalence), 1:3]
> lines(od[, 3], fitted.values(loess(od[, 1] ~ od[, 3],
+    span = 0.5)))
> title(main = "Local linear regression ('loess')")
```

I precedenti comandi forniscono il grafico della Figura 5.3.4. □

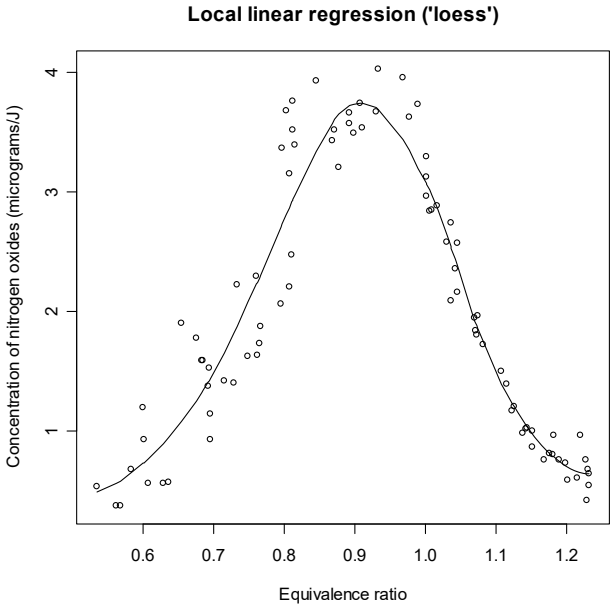


Figura 5.3.4.

5.4. Riferimenti bibliografici

Bowman, A.W. e Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, Oxford.

Chacón, J.E. e Duong, T. (2018) *Multivariate Kernel Smoothing and its Applications*, CRC Press, Boca Raton.

Cleveland, W.S. (1993) *Visualizing Data*, Hobart Press, Summit.

Efromovich, S. (1999) *Nonparametric Curve Estimation*, Springer, New York.

Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Klemelä, J. (2014) *Multivariate Nonparametric Regression and Visualization*, Wiley, New York.

Loader, C. (1999) *Local Regression and Likelihood*, Springer, New York.

Scott, D.W. (2015) *Multivariate Density Estimation*, seconda edizione, Wiley, New York.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Simonoff, J.S. (1996) *Smoothing Methods in Statistics*, Springer, New York.

Wand, M.P. e Jones, M.C. (1995) *Kernel Smoothing*, Chapman and Hall, London.

Pagina intenzionalmente vuota

Capitolo 6

La verifica di ipotesi e gli intervalli di confidenza

6.1. La verifica di ipotesi

Sulla base del campione osservato si è interessati a stabilire se il vero valore del parametro appartiene ad un certo sottoinsieme dello spazio parametrico Θ, ovvero l'insieme di tutti i valori plausibili per il parametro θ. Dato un modello statistico, se gli insiemi Θ_0 e Θ_1 costituiscono una partizione di Θ, la verifica di ipotesi consiste in un procedimento decisionale di scelta fra l'ipotesi di base $H_0 : \theta \in \Theta_0$ e l'ipotesi alternativa $H_1 : \theta \in \Theta_1$. L'insieme delle ipotesi ammissibili e la sua partizione in H_0 e H_1 è detto sistema di ipotesi.

• **Esempio 6.1.1.** Nella semplice situazione in cui si ha un campione casuale da una variabile casuale X, il tipico modello classico assume che $X \sim N(\mu, \sigma^2)$ e l'usuale sistema di ipotesi consiste nel verificare $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$. Dal momento che lo spazio parametrico in questo caso è dato da

$$\Theta = \{(\mu, \sigma^2) : \mu \in]-\infty, \infty[, \sigma^2 \in]0, \infty[\} ,$$

risulta $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 \in]0, \infty[\}$ e $\Theta_1 = \{(\mu, \sigma^2) : \mu \neq \mu_0, \sigma^2 \in]0, \infty[\}$. Al contrario, in un approccio "distribution-free", si assume solo che X sia una variabile casuale continua con funzione di ripartizione F e mediana λ, mentre il sistema di ipotesi risulta $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$. In questo caso, lo spazio "parametrico" è dato da

$$\Theta = \{(\lambda, F) : \lambda \in]-\infty, \infty[, F \in \mathcal{R}_c\} ,$$

dove $\mathcal{R}_c$ è lo spazio delle funzioni di ripartizione continue. Si ha $\Theta_0 = \{(\lambda, F) : \lambda = \lambda_0, F \in \mathcal{R}_c\}$, mentre $\Theta_1 = \{(\lambda, F) : \lambda \neq \lambda_0, F \in \mathcal{R}_c\}$. □

Lo strumento statistico che sulla base del campione consente di concludere in favore dell'una o dell'altra ipotesi è il test statistico. Scelta una opportuna statistica T definita sull'insieme $\mathcal{T}$, si dice test basato su T la funzione

$$D : \mathcal{T} \to \{H_0, H_1\} ,$$

mentre T è detta statistica test. Il test basato su T è una regola decisionale che suddivide $\mathcal{T}$ negli insiemi complementari $\mathcal{T}_0$ e $\mathcal{T}_1$, in modo tale che si accetta H_0 se la realizzazione t di T è tale che $t \in \mathcal{T}_0$, mentre si accetta H_1 se $t \in \mathcal{T}_1$. L'insieme $\mathcal{T}_0$ è detto regione di accettazione, mentre l'insieme $\mathcal{T}_1$ è detto regione critica del test basato su T. Un test è detto "distribution-free" se è basato su una statistica test "distribution-free".

• **Esempio 6.1.2.** In un approccio classico, si consideri un campione casuale $X_1, \ldots, X_n$ da $X \sim N(\mu, 1)$ e il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$, dove μ_0 è una quantità nota. Se si suppone che la statistica test sia la media campionaria delle osservazioni $\bar{X}$, risulta $\mathcal{T} = \mathbb{R}$. Una possibile scelta per $\mathcal{T}_0$ potrebbe essere data da

$$\mathcal{T}_0 = \{\overline{x} : |\overline{x} - \mu_0| < a\} ,$$

con a costante. Di conseguenza, la regione critica del test basato su $\overline{X}$ risulta

$$\mathcal{T}_1 = \{\overline{x} : |\overline{x} - \mu_0| \geq a\} .$$

Questa scelta di $\mathcal{T}_1$ appare logica, in quanto più la realizzazione della media campionaria differisce dal valore ipotizzato μ_0 per la media, più si è propensi ad accettare l'ipotesi alternativa. Rimane aperto il problema della scelta della costante a. Alternativamente, in un approccio "distribution-free", si consideri un campione casuale da una variabile casuale continua X con mediana pari a λ. Si vuole verificare il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$. Il test basato sulla statistica

$$B = \sum_{i=1}^{n} \mathbf{1}_{]0,\infty[}(X_i - \lambda_0) ,$$

è "distribution-free" essendo una trasformata di variabili casuali segno (vedi Esempio 3.3.1). Si ha $\mathcal{T} = \{b : b = 0, 1, \ldots, n\}$, mentre è ragionevole assumere $\mathcal{T}_0 = \{b : b = a + 1, a + 2, \ldots, n - a\}$, dove a è un intero tale che $a < n/2$. In effetti, sotto ipotesi di base si attende che circa la metà delle osservazioni campionarie siano maggiori della mediana, ovvero una realizzazione b di B prossima a $n/2$. Di nuovo, si deve decidere il valore della costante a. □

Uno strumento per misurare la capacità discriminatoria del test basato su una statistica T è la funzione potenza. La funzione potenza del test basato su T è data da

$$P_T(\theta) = P(T \in \mathcal{T}_1) ,$$

dove la probabilità è indotta dalla distribuzione specificata dal modello quando il valore del parametro è pari a θ. Per ogni $\theta \in \Theta_0$ la funzione potenza $P_T(\theta)$ fornisce la probabilità di respingere H_0 quando questa è vera, ovvero la probabilità di commettere il cosiddetto errore di I specie. Analogamente, per ogni $\theta \in \Theta_1$ la quantità $(1 - P_T(\theta))$ fornisce la probabilità di accettare H_0 quando è vera H_1, ovvero la probabilità di commettere il cosiddetto errore di II specie. Per ogni $\theta \in \Theta_1$, la funzione potenza $P_T(\theta)$ fornisce la probabilità di accettare H_1 quando questa è vera. Si dice infine che il test basato su T è al livello di significatività α se

$$\sup_{\theta \in \Theta_0} P_T(\theta) = \alpha .$$

Il livello di significatività α rappresenta dunque la massima probabilità di commettere un errore di I specie.

• **Esempio 6.1.3.** Dato un campione casuale da $X \sim N(\mu, 1)$, si consideri il sistema di ipotesi $H_0 : \mu \leq 0$ contro $H_1 : \mu > 0$. Se si suppone che la statistica test sia $\overline{X}$, dato che $\mathcal{T} = \mathbb{R}$ si può scegliere

$$\mathcal{T}_0 =]-\infty, z_{1-\alpha}/\sqrt{n}[$$

e

$$\mathcal{T}_1 = [z_{1-\alpha}/\sqrt{n}, \infty[.$$

Questa selezione di $\mathcal{T}_1$ appare logica, in quanto più la realizzazione della media campionaria è elevata, più si è propensi ad accettare l'ipotesi alternativa. Risulta $\overline{X} \sim N(\mu, 1/n)$ per ogni $\mu \in \mathbb{R}$ e la funzione potenza è data da

$$P_{\overline{X}}(\mu) = \Phi(\sqrt{n}\mu - z_{1-\alpha}) .$$

Dal momento che $P_{\bar{X}}(\mu)$ è crescente e che

$$\sup_{\mu \leq 0} P_{\bar{X}}(\mu) = \alpha \, ,$$

il test basato su $\bar{X}$ è al livello di significatività α. Alternativamente, in un approccio "distribution-free", si consideri un campione casuale da una variabile casuale continua X con mediana pari a λ e funzione di ripartizione $F(x) = G(x - \lambda)$ con $G \in \mathcal{R}_0$. Si consideri inoltre il sistema di ipotesi $H_0 : \lambda \leq 0$ contro $H_1 : \lambda > 0$. Se si suppone che la statistica test sia data da $B = \sum_{i=1}^{n} \mathbf{1}_{]0,\infty[}(X_i)$, dal momento che $\mathcal{T} = \{b : b = 0, 1, \ldots, n\}$ si può scegliere

$$\mathcal{T}_0 = \{b : b = 0, 1, \ldots, b_{n,1-\alpha}\}$$

e

$$\mathcal{T}_1 = \{b : b = b_{n,1-\alpha} + 1, \ldots, n\} \, ,$$

dove $b_{n,\alpha}$ rappresenta il quantile di ordine α della distribuzione Binomiale $Bi(n, 1/2)$. Di nuovo, la scelta di $\mathcal{T}_1$ è evidente, in quanto più il numero di osservazioni maggiori di zero è elevato, più si è propensi ad accettare l'ipotesi alternativa. Se è vera l'ipotesi alternativa, risulta

$$P(X > 0) = 1 - G(-\lambda) = G(\lambda) \, ,$$

per cui B ha distribuzione Binomiale $Bi(n, G(\lambda))$. La funzione potenza è dunque data da

$$P_B(\lambda) = \sum_{b=b_{n,1-\alpha}+1}^{n} \binom{n}{b} G(\lambda)^b (1 - G(\lambda))^{n-b} \, .$$

Essendo $G(0) = 1/2$, allora risulta $P_B(0) = \alpha$ per ogni $G \in \mathcal{R}_0$. Si può dimostrare che $P_B(\lambda)$ è crescente e si deve concludere quindi che il test è al livello di significatività α. Si osservi che il test basato su B ha livello di significatività α per ogni modello statistico con campionamento casuale da una variabile casuale continua, mentre il test basato su $\bar{X}$ non mantiene in generale questo livello se viene a mancare l'assunzione di Normalità. □

Dal momento che non si può rendere contemporaneamente pari a zero gli errori di I e II specie, si deve stabilire un insieme di proprietà desiderabili per un test. Una prima proprietà opportuna per un test è quella della correttezza. Un test basato su T al livello di significatività α con funzione potenza $P_T(\theta)$ è detto corretto al livello di significatività α se

$$P_T(\theta) \geq \alpha \, , \forall \theta \in \Theta_1 \, .$$

La proprietà della correttezza permette di controllare l'errore di I specie e al tempo stesso assicura che la probabilità di accettare H_1 quando è vera risulta maggiore dell'errore di I specie.

Una seconda proprietà riguarda il comportamento per grandi campioni del test. Il test al livello di significatività α è detto coerente se

$$\lim_{n} P_T(\theta) = 1 \, , \forall \theta \in \Theta_1 \, .$$

La proprietà della coerenza assicura che la probabilità di commettere un errore di II specie tende a zero quando si dispone di grandi campioni.

• **Esempio 6.1.4.** Dato un campione casuale da $X \sim N(\mu, 1)$, si consideri il sistema di ipotesi $H_0 : \mu \leq 0$ contro $H_1 : \mu > 0$. Il test basato su $\bar{X}$ è corretto in quanto si ha $P_{\bar{X}}(\mu) \leq \alpha$ per ogni $\mu \leq 0$, mentre $P_{\bar{X}}(\mu) > \alpha$ per ogni $\mu > 0$. Dal momento che la successione di funzioni $(\Phi(\sqrt{n}\mu - z_{1-\alpha}))_{n \geq 1}$ converge uniformemente ad una funzione costante pari ad 1 per $\mu > 0$, allora si ha

$$\lim_n P_{\bar{X}}(\mu) = 1 \, , \forall \mu > 0 \, ,$$

ovvero il test è coerente. Alternativamente, in un approccio "distribution-free", sotto le assunzioni dell'Esempio 6.1.3, si consideri inoltre il sistema di ipotesi $H_0 : \lambda \leq 0$ contro $H_1 : \lambda > 0$. Il test basato su B è corretto in quanto $P_B(\lambda) \leq \alpha$ per ogni $\lambda \leq 0$, mentre $P_B(\lambda) > \alpha$ per ogni $\lambda > 0$. Sulla base del Teorema Fondamentale del Limite, si può anche dimostrare che il test è anche coerente. Queste proprietà sono valide per ogni modello statistico con campionamento casuale da una variabile casuale continua. □

Anche se si desidera che $P_T(\theta)$ sia più elevata possibile quando $\theta \in \Theta_1$ e più piccola possibile quando $\theta \in \Theta_0$, questi requisiti sono conflittuali tra loro. Un possibile modo di procedere è quello di fissare il livello di significatività α e scegliere quel test che ha più alta potenza per ogni $\theta \in \Theta_1$. Un test basato su T al livello di significatività α con funzione potenza $P_T(\theta)$ è detto uniformemente più potente al livello di significatività α se

$$P_T(\theta) \geq P_U(\theta) \, , \forall \theta \in \Theta_1 \, ,$$

per ogni altro test basato su una qualsiasi statistica U al livello di significatività α. I test uniformente più potenti esistono in alcuni casi solamente quando si considera un approccio classico. Al contrario, quando si considera un test "distribution-free" in generale non è possibile determinare il test uniformemente più potente.

Se $X_1, \ldots, X_n$ è un campione casuale da una variabile casuale X con funzione di ripartizione F, si consideri il sistema di ipotesi $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, dove θ è un parametro reale. Se si vuole analizzare le prestazioni di due statistiche test T e U per questo sistema di ipotesi, si può considerare la cosiddetta efficienza relativa. Siano $N_T(\alpha, \beta, \theta, F)$ e $N_U(\alpha, \beta, \theta, F)$ le numerosità campionarie necessarie ai due test per raggiungere la potenza β al livello di significatività α quando si considera l'alternativa $\theta \neq \theta_0$. L'efficienza relativa è data dalla quantità

$$e_{T,U}(\alpha, \beta, \theta, F) = \frac{N_U(\alpha, \beta, \theta, F)}{N_T(\alpha, \beta, \theta, F)} \, .$$

Se si ha $e_{T,U}(\alpha, \beta, \theta, F) > 1$, allora il test basato su T è più efficiente di quello basato su U, mentre se $e_{T,U}(\alpha, \beta, \theta, F) < 1$ è vero l'opposto. L'efficienza relativa è una misura locale dell'efficienza, nel senso che dipende dalle quantità α, β, θ e F. Al fine di eliminare la dipendenza da α, β, θ e quindi di ottenere una misura più globale dell'efficienza, si adotta la cosiddetta efficienza asintotica relativa $\text{ARE}_{T,U}$, che è data dal limite del rapporto delle numerosità campionarie necessarie ad ottenere la medesima potenza dei test T e U per una successione di alternative che converge a θ_0 ad un livello di significatività costante. In effetti, si ha $\text{ARE}_{T,U} = \text{ARE}_{T,U}(F)$, ovvero l'efficienza asintotica relativa dipende comunque dalla struttura funzionale della funzione di ripartizione F. Se $\text{ARE}_{T,U} > 1$ il test basato su T è preferibile a quello basato su U, mentre se $\text{ARE}_{T,U} < 1$ è vero l'opposto.

• **Esempio 6.1.5.** Dato un campione casuale da una variabile casuale continua X con funzione di ripartizione F, si consideri il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$, dove λ rappresenta la mediana. La Tavola 6.1.1 riporta $\text{ARE}_{B,T}$ per alcune distribuzioni quando si confronta il classico test T di Student e il test B considerato nell'Esempio 6.1.2. Anche se il test basato su B dimostra scarse prestazioni per una distribuzione a code leggere come l'Uniforme, il test diventa infinitamente efficiente per una distribuzione a code pesanti come la Cauchy. Si può verificare inoltre che le prestazioni del test dei segni risultano generalmente superiori (talvolta in modo molto marcato) a quelle del test T di Student quando si considerano distribuzioni asimmetriche. In generale, si può dimostrare che $\text{ARE}_{B,T} \geq 1/3$. Dunque, il test T di Student può avere un comportamento pessimo quando viene a mancare l'ipotesi di Normalità. □

Tavola 6.1.1.

distribuzione	$\text{ARE}_{B,T}$
$U(\lambda, \lambda+\delta)$	$1/3 \simeq 0.3333$
$N(\mu, \sigma^2)$	$2/\pi \simeq 0.6366$
$C(\lambda, \delta)$	∞

Un modo per costruire test ottimali quando si considera un approccio classico è mediante il cosiddetto rapporto delle verosimiglianze. Il test del rapporto delle verosimiglianze è basato sulla statistica test R, la cui realizzazione è data da

$$r = \frac{\max\limits_{\theta \in \Theta_0} L(\theta)}{\max\limits_{\theta \in \Theta} L(\theta)} .$$

Questa statistica test ha un'interpretazione intuitiva, nel senso che se si sta confrontando la "plausibilità" di un valore θ rispetto ad un altro sulla base di un campione, siamo portati a scegliere quel valore che fornisce la verosimiglianza più alta. Se non esiste un valore θ che fornisce una verosimiglianza sensibilmente più alta in Θ rispetto alla verosimiglianza massima in Θ_0, siamo propensi ad accettare H_0. Ovviamente, per una realizzazione r di R si ha $0 \leq r \leq 1$. Se la realizzazione r è prossima ad 1 si è più propensi ad accettare H_0, mentre se la realizzazione r è prossima a 0 si è più propensi ad accettare H_1. Di conseguenza, si può scegliere come regione critica al livello di significatività α l'insieme $\mathcal{T}_1 = \{r : r \leq r_\alpha\}$, dove r_α è il quantile di ordine α della distribuzione di R.

• **Esempio 6.1.6.** Dato un campione casuale dalla variabile casuale $X \sim N(\mu, 1)$, si consideri il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$. La determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{e^{-\frac{1}{2}n(s_x^2 + (\overline{x} - \mu_0)^2)}}{e^{-\frac{1}{2}n s_x^2}} = e^{-\frac{1}{2}n(\overline{x} - \mu_0)^2} .$$

Dal momento che r è una funzione biunivoca di $(\overline{x} - \mu_0)^2$, la regione critica indotta da R è la stessa di quella indotta da $(\overline{X} - \mu_0)^2$ e quindi i relativi test sono equivalenti. Se è vera H_0 si ha $\sqrt{n}(\overline{X} - \mu_0) \sim N(0, 1)$ e dunque $n(\overline{X} - \mu_0)^2 \sim \chi^2_1$. Tenendo presente che r è una funzione monotona decrescente di $(\overline{x} - \mu_0)^2$ e dunque si respinge H_0 per realizzazioni elevate di $(\overline{X} - \mu_0)^2$, allora la regione critica del test basato sul rapporto delle verosimiglianze è data da

$$\mathcal{T}_1 = \{\overline{x} : n(\overline{x} - \mu_0)^2 \geq \chi^2_{1,1-\alpha}\} .$$

Essendo $z_{1-\alpha/2} = \sqrt{\chi^2_{1,1-\alpha}}$ per la relazione fra le distribuzioni $N(0, 1)$ e χ^2_1, la regione critica può essere anche espressa in modo equivalente come

$$\mathcal{T}_1 = \{\overline{x} : \sqrt{n}\,|\overline{x} - \mu_0| \geq z_{1-\alpha/2}\} . \qquad \square$$

• **Esempio 6.1.7.** Si consideri un campionamento casuale da $X \sim N(\mu, \upsilon)$ e si supponga di voler verificare il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$. In questo caso si ha

$$\Theta = \{(\mu, \upsilon) : (\mu, \upsilon) \in \mathbb{R} \times \mathbb{R}^+\}$$

e

$$\Theta_0 = \{(\mu, \upsilon) : \mu = \mu_0,\ \upsilon \in \mathbb{R}^+\} .$$

Tenendo presente l'Esempio 4.3.1, si ottiene

$$\max_{(\mu,\upsilon)\in\Theta} l(\mu,\upsilon) = l(\overline{x}, s_x^2) \, .$$

Inoltre, dalla disuguaglianza dell'Esempio 4.3.1 con $d = s_x^2 + (\overline{x} - \mu_0)^2$, si ha

$$\begin{aligned} l(\mu_0,\upsilon) &= \log c - \frac{n}{2}\log\upsilon - \frac{n}{2\upsilon}\left(s_x^2 + (\overline{x}-\mu_0)^2\right) \leq \log c - \frac{n}{2}\log(s_x^2 + (\overline{x}-\mu_0)^2) - \frac{n}{2} \\ &= l(\mu_0, s_x^2 + (\overline{x}-\mu_0)^2) = \max_{(\mu,\upsilon)\in\Theta_0} l(\mu,\upsilon) \, . \end{aligned}$$

La determinazione campionaria di R è dunque data da

$$r = \frac{e^{-\frac{n}{2}\log(s_x^2+(\overline{x}-\mu_0)^2)}}{e^{-\log s_x^2}} = \left(1 + \frac{n(\overline{x}-\mu_0)^2}{(n-1)s_{c,x}^2}\right)^{-\frac{1}{2}n} = \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}n}$$

dove

$$t = \frac{\sqrt{n}(\overline{x}-\mu_0)}{s_{c,x}} \, .$$

Se è vera H_0, t è la determinazione campionaria di una statistica T distribuita come una t di Student con $(n-1)$ gradi di libertà, dal momento che

$$T = \frac{\sqrt{n}(\overline{X}-\mu_0)}{S_{c,x}} = \frac{\sqrt{n}(\overline{X}-\mu_0)/\sqrt{\upsilon}}{\sqrt{((n-1)S_{c,x}^2/\upsilon)/(n-1)}}$$

e che inoltre le statistiche $\sqrt{n}(\overline{X}-\mu_0)/\sqrt{\upsilon} \sim N(0,1)$ e $(n-1)S_{c,x}^2/\upsilon \sim \chi_{n-1}^2$ sono indipendenti. Inoltre, r è una funzione monotona decrescente di $|t|$ e quindi i test costruiti su R e $|T|$ sono equivalenti. Tenendo presente la relazione fra le statistiche T e $|T|$ e dal momento che rifiutare H_0 per piccole determinazioni di R è equivalente a rifiutare H_0 per determinazioni elevate di $|T|$, la regione critica del test risulta

$$\mathcal{T}_1 = \{t : t \leq -t_{n-1,1-\alpha/2}, t \geq t_{n-1,1-\alpha/2}\} \, . \qquad \square$$

Per quanto riguarda la distribuzione per grandi campioni del rapporto delle verosimiglianze, se k è il numero di parametri nel modello e q è il numero di parametri da stimare sotto ipotesi di base, assumendo alcune ipotesi di regolarità, la statistica test $-2\ln R$ converge in distribuzione ad una variabile casuale con distribuzione χ_{k-q}^2 per $n \to \infty$. La regione critica per grandi campioni del test basato sul rapporto delle verosimiglianze è dunque data da

$$\mathcal{T}_1 = \{r : -2\ln r \geq \chi_{k-q,1-\alpha}^2\} \, .$$

• **Esempio 6.1.8.** Dato un campione casuale dalla variabile casuale $X \sim E(0,\sigma)$, si consideri il sistema di ipotesi $H_0 : \sigma = \sigma_0$ contro $H_1 : \sigma \neq \sigma_0$. La determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{e^{-n\log\sigma_0 - n\overline{x}/\sigma_0}}{e^{-n\log\overline{x} - n}} = e^{n\log(\overline{x}/\sigma_0) - n\overline{x}/\sigma_0 + n} \, ,$$

ovvero r è una funzione non biunivoca di $\overline{x}$. Dunque, risulta laborioso determinare la distribuzione per campioni finiti del rapporto delle verosimiglianze. Tuttavia, se è vera H_0 risulta

$$-2\log R = -2n\log\left(\frac{\overline{X}}{\sigma_0}\right) + \frac{2n\overline{X}}{\sigma_0} - 2n\,,$$

che converge in distribuzione ad una variabile casuale con distribuzione χ^2_1 per $n \to \infty$. Di conseguenza, la regione critica per grandi campioni del test basato sul rapporto delle verosimiglianze è data da

$$\mathcal{T}_1 = \left\{\overline{x} : -2n\log\left(\frac{\overline{x}}{\sigma_0}\right) + \frac{2n\overline{x}}{\sigma_0} - 2n \geq \chi^2_{1,1-\alpha}\right\}. \qquad \square$$

• **Esempio 6.1.9.** Dato un campione casuale dalla variabile casuale $X \sim Bi(1,\pi)$, si consideri il sistema di ipotesi $H_0 : \pi = \pi_0$ contro $H_1 : \pi \neq \pi_0$. La determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{e^{n\overline{x}\log\pi_0 + (n-n\overline{x})\log(1-\pi_0)}}{e^{n\overline{x}\log\overline{x} + (n-n\overline{x})\log(1-\overline{x})}} = e^{n\overline{x}\log(\pi_0/\overline{x}) + (n-n\overline{x})\log((1-\pi_0)/(1-\overline{x}))}\,,$$

ovvero r è anche in questo caso una funzione non biunivoca di $\overline{x}$ e quindi non è semplice ottenere la relativa distribuzione per campioni finiti. Tuttavia, se è vera H_0 risulta

$$-2\log R = -2n\overline{X}_n\log\left(\frac{\pi_0}{\overline{X}_n}\right) - 2(n-n\overline{X}_n)\log\left(\frac{1-\pi_0}{1-\overline{X}_n}\right),$$

che converge in distribuzione ad una variabile casuale con distribuzione χ^2_1 per $n \to \infty$. Di conseguenza, la regione critica per grandi campioni del test basato sul rapporto delle verosimiglianze è data da

$$\mathcal{T}_1 = \left\{\overline{x} : -2n\overline{x}\log\left(\frac{\pi_0}{\overline{x}}\right) - 2(n-n\overline{x})\log\left(\frac{1-\pi_0}{1-\overline{x}}\right) \geq \chi^2_{1,1-\alpha}\right\}. \qquad \square$$

Nel presente approccio, l'ipotesi di base e l'ipotesi alternativa vengono trattate in modo non simmetrico. In effetti, usualmente H_0 costituisce una affermazione privilegiata e si preferisce controllare il livello di significatività del test (ovvero l'errore di I specie) che comporta l'erroneo rifiuto di questa ipotesi privilegiata. Anche se per sviluppare la teoria è necessario fissare il livello di significatività α, quando si lavora operativamente non esiste nessuna regola per stabilirne la scelta. Questa considerazione porta al concetto di livello di significatività osservato o valore-P.

Se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \geq c\}$, per un determinato valore campionario t si dice significatività osservata la quantità

$$\alpha_{oss} = \sup_{\theta\in\Theta_0} P(T \geq t)\,,$$

mentre, se la regione critica è data da $\mathcal{T}_1 = \{t : t \leq c\}$, allora si dice significatività osservata la quantità

$$\alpha_{oss} = \sup_{\theta\in\Theta_0} P(T \leq t)\,.$$

Quando invece la statistica test T ha una distribuzione simmetrica, se la regione critica del test basato su T è data da $\mathcal{T}_1 = \{t : t \leq c_1, t \geq c_2\}$, si dice significatività osservata la quantità

$$\alpha_{oss} = 2\min\{\sup_{\theta\in\Theta_0} P(T \leq t), \sup_{\theta\in\Theta_0} P(T \geq t)\}\,.$$

La significatività osservata rappresenta la probabilità di ottenere, quando H_0 è vera, un valore campionario t di T estremo (nella appropriata direzione) almeno quanto quello osservato. Dunque, la significatività osservata fornisce una misura su quanto l'ipotesi di base risulta compatibile con i dati campionari. Una significatività osservata bassa porta a ritenere poco compatibile con i dati campionari l'ipotesi di base, mentre con una significatività osservata elevata è vera l'affermazione contraria. In una verifica di ipotesi si può semplicemente riportare la significatività osservata, oppure si può arrivare ad una decisione sull'accettazione di H_0 fissando un livello di significatività α. Se il livello di significatività osservato è minore o uguale ad α, allora si respinge H_0, altrimenti si accetta H_0. Il livello di significatività osservato diventa in questo caso il più elevato livello di significatività per cui si accetta H_0. In questo caso il livello di significatività osservato diventa non solo uno strumento per la decisione nella verifica di ipotesi, ma anche una misura quantitativa di questa decisione.

6.2. Gli intervalli di confidenza

Piuttosto che selezionare sulla base del campione un unico valore come nella stima per punti, può essere utile dal punto di vista operativo ottenere un insieme di valori plausibili del parametro. Considerato per semplicità il caso in cui si ha un singolo campione casuale $X_1, \ldots, X_n$, sia $Q = Q(X_1, \ldots, X_n; \theta)$ una quantità pivotale, ovvero una trasformata che dipende dal parametro ma con una distribuzione non dipende dal parametro stesso. Se c_1 e c_2 sono due valori tali che

$$P(c_1 < Q(X_1, \ldots, X_n; \theta) < c_2) = 1 - \alpha \,,$$

con $0 < \alpha < 1$, e se $L = L(X_1, \ldots, X_n)$ e $U = U(X_1, \ldots, X_n)$ sono statistiche tali che per ogni θ

$$\{(x_1, \ldots, x_n) : c_1 < Q(x_1, \ldots, x_n; \theta) < c_2\} \Leftrightarrow \{(x_1, \ldots, x_n) : L(x_1, \ldots, x_n) < \theta < U(x_1, \ldots, x_n)\} \,,$$

allora l'intervallo casuale (L, U) è detto intervallo di confidenza di θ al livello di confidenza $(1 - \alpha)$. Se la quantità pivotale Q è "distribution-free", ovvero se la sua distribuzione rimane invariata per un modello "distribution-free", allora il relativo intervallo di confidenza è detto "distributon-free".

La nozione di intervallo di confidenza deve essere adoperata con cautela, ovvero non si deve affermare che il vero valore del parametro è contenuto in un intervallo con probabilità pari a $(1 - \alpha)$. In termini rigorosi di probabilità, una volta che l'intervallo di confidenza è stato determinato sul campione, questo contiene il vero valore con probabilità 0 o 1. Si può affermare invece che l'intervallo di confidenza è la determinazione di un procedimento casuale che seleziona intervalli in modo tale che la probabilità di ottenere un intervallo contenente il vero valore del parametro è pari a $(1 - \alpha)$.

• **Esempio 6.2.1.** Se si considera un campione casuale da $X \sim N(\mu, 1)$, una possibile quantità pivotale è data da

$$Q(X_1, \ldots, X_n; \mu) = \sqrt{n}(\bar{X} - \mu) \,.$$

Questa variabile casuale è in effetti una quantità pivotale, dal momento che la sua distribuzione non dipende da μ essendo $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$. Scelto un livello di confidenza $(1 - \alpha)$, risulta

$$P(-z_{1-\alpha/2} < \sqrt{n}(\bar{X} - \mu) < z_{1-\alpha/2}) = 1 - \alpha \,.$$

L'insieme

$$\{(x_1, \ldots, x_n) : -z_{1-\alpha/2} < \sqrt{n}(\bar{x} - \mu) < z_{1-\alpha/2}\}$$

è equivalente all'insieme

$$\{(x_1, \ldots, x_n) : \bar{x} - z_{1-\alpha/2}/\sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2}/\sqrt{n}\}$$

e si deve dunque concludere che

$$(\overline{x} - z_{1-\alpha/2}/\sqrt{n}, \overline{x} + z_{1-\alpha/2}/\sqrt{n})$$

è un intervallo di confidenza per μ al livello di confidenza $(1-\alpha)$. □

• **Esempio 6.2.2.** Se si consideri un campione casuale da una variabile casuale continua X con mediana λ, una quantità pivotale è data da

$$Q(X_1, \ldots, X_n; \lambda) = B(X_1 - \lambda, \ldots, X_n - \lambda) = \sum_{i=1}^{n} \mathbf{1}_{]0,\infty[}(X_i - \lambda) \, .$$

La quantità pivotale $B(X_1 - \lambda, \ldots, X_n - \lambda)$ è in effetti tale ed è "distribution-free", dal momento che la sua distribuzione non dipende da λ e $B(X_1 - \lambda, \ldots, X_n - \lambda) \sim Bi(n, 1/2)$. Dunque, scelto un livello di confidenza $(1-\alpha)$, si ha

$$P(b_{n,\alpha/2} < B(X_1 - \lambda, \ldots, X_n - \lambda) < n - b_{n,\alpha/2}) = 1 - \alpha \, .$$

Si osservi che la quantità pivotale $B(X_1 - \lambda, \ldots, X_n - \lambda)$ rappresenta il numero di osservazioni campionarie maggiori di λ. Se $(X_{(1)}, \ldots, X_{(n)})$ è la statistica ordinata, si ha che l'insieme

$$\{(x_1, \ldots, x_n) : B(x_1 - \lambda, \ldots, x_n - \lambda) < n - b_{n,\alpha/2}\}$$

è equivalente all'insieme

$$\{(x_1, \ldots, x_n) : x_{(b_{n,\alpha/2}+1)} < \lambda\} \, .$$

Analogamente, l'insieme

$$\{(x_1, \ldots, x_n) : b_{n,\alpha/2} < B(x_1 - \lambda, \ldots, x_n - \lambda)\}$$

è equivalente all'insieme

$$\{(x_1, \ldots, x_n) : \lambda < x_{(n-b_{n,\alpha/2})}\} \, .$$

Dunque, si ha $L = X_{(b_{n,\alpha/2}+1)}$ e $U = X_{(n-b_{n,\alpha/2})}$, ovvero $(X_{(b_{n,\alpha/2}+1)}, X_{(n-b_{n,\alpha/2})})$ è un intervallo di confidenza per λ "distribution-free" al livello di confidenza $(1-\alpha)$. □

Esiste una stretta connessione tra il problema della stima per intervalli e quello della verifica di ipotesi. Questa è anche la ragione per cui la stima per intervalli viene analizzata successivamente alla teoria relativa alla verifica di ipotesi. Più esattamente, se si considera il sistema di ipotesi $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, esiste una equivalenza tra la regione critica della statistica test e l'intervallo di confidenza di θ. Si consideri un test corretto al livello di significatività α. Se C_0 è la regione di accettazione del test nello spazio campionario, ovvero l'insieme delle realizzazioni campionarie $(x_1, \ldots, x_n)$ che portano all'accettazione dell'ipotesi di base, è immediato verificare che C_0 dipende dal valore prefissato θ_0 di θ. Dunque risulta $C_0 = C_0(\theta)$. Inversamente, per un data realizzazione campionaria $(x_1, \ldots, x_n)$ nello spazio campionario deve esistere un insieme $E(x_1, \ldots, x_n)$ di valori di θ per i quali si accetta l'ipotesi di base. I due insiemi sono equivalenti, ovvero

$$\{(x_1, \ldots, x_n) : (x_1, \ldots, x_n) \in C_0(\theta)\} \Leftrightarrow \{(x_1, \ldots, x_n) : \theta \in E(x_1, \ldots, x_n)\} \, .$$

Dal momento che

$$\{(x_1, \ldots, x_n) : (x_1, \ldots, x_n) \in C_0(\theta)\} = \{(x_1, \ldots, x_n) : \mathbf{1}_{C_0(\theta)}(x_1, \ldots, x_n) = 1\}$$

e

$$P((X_1,\ldots,X_n)\in C_0(\theta)) = P(\mathbf{1}_{C_0(\theta)}(X_1,\ldots,X_n)=1) = 1-\alpha\,,$$

se come quantità pivotale si considera la variabile casuale $\mathbf{1}_{C_0(\theta)}(X_1,\ldots,X_n)$, allora $E(x_1,\ldots,x_n)$ è un insieme di confidenza al livello di confidenza $(1-\alpha)$. Quindi, l'insieme di tutti i valori θ per cui si accetta l'ipotesi di base costituisce un insieme di confidenza per θ. Nel caso che $E(x_1,\ldots,x_n)$ sia un intervallo, con la precedente procedura si ottiene ovviamente un intervallo di confidenza e questa situazione è frequente con i modelli usualmente adoperati in pratica.

Partendo dunque da un test opportuno è possibile costruire un intervallo di confidenza al livello di confidenza scelto. Inoltre, più il test prescelto ha funzione potenza elevata più l'intervallo di confidenza risultante è desiderabile, dal momento che *a priori* dal campionamento la probabilità che contenga un qualunque valore θ diverso da θ_0 è bassa. È dunque buona regola costruire intervalli di confidenza a partire da un test che gode di buone proprietà. Queste considerazioni consentono di costruire un intervallo di confidenza per un dato parametro partendo da un opportuno sistema di ipotesi. Nel caso di un modello classico, l'intervallo di confidenza viene usualmente costruito a partire dal test del rapporto delle verosimiglianze. L'intervallo di confidenza risulta "distribution-free" se è costruito a partire da un test "distribution-free". Esiste anche una connessione fra test e stima per punti. In effetti, la stima per punti risulta usualmente il valore centrale dell'intervallo di confidenza.

• **Esempio 6.2.3.** Dato un campione casuale da $X\sim U(0,\delta)$, si vuole costruire un intervallo di confidenza per δ. Se si considera il sistema di ipotesi $H_0:\delta=\delta_0$ contro $H_1:\delta\neq\delta_0$, tenendo presente l'Esempio 4.3.2, la determinazione campionaria del rapporto delle verosimiglianze è data da

$$r = \frac{\delta_0^{-n}\mathbf{1}_{(0,1)}(x_{(n)}/\delta_0)}{x_{(n)}^{-n}} = \left(\frac{x_{(n)}}{\delta_0}\right)^n \mathbf{1}_{(0,1)}\left(\frac{x_{(n)}}{\delta_0}\right),$$

Si noti che r è funzione monotona crescente di $x_{(n)}/\delta_0$ e quindi il test costruito sulla statistica $X_{(n)}/\delta_0$ è equivalente a quello del rapporto delle verosimiglianze. Quando H_0 è vera, la statistica $X_{(n)}/\delta_0$ ha distribuzione $Be(0,1;n,1)$ (vedi Esempio 4.3.2). Dal momento che il quantile di ordine α di una $Be(0,1;n,1)$ è dato da $\alpha^{1/n}$, la regione di accettazione per il test del rapporto delle verosimiglianze risulta

$$\mathcal{T}_0 = \{(x_1,\ldots,x_n) : \alpha^{1/n} < x_{(n)}/\delta < 1\}\,,$$

che è equivalente all'insieme

$$\{(x_1,\ldots,x_n) : x_{(n)} < \delta < x_{(n)}\alpha^{-1/n}\}\,.$$

Di conseguenza, $(x_{(n)}, x_{(n)}\alpha^{-1/n})$ è un intervallo di confidenza per δ al livello di confidenza $(1-\alpha)$.

6.3. Riferimenti bibliografici

Boos, D.D. e Stefanski, L.A. (2013) *Essential Statistical Inference*, Springer, New York.

Cox, D.R. e Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.

Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.

Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.

Lauritzen, S. (2023) *Fundamentals of Mathematical Statistics*, Chapman & Hall/CRC Press, Boca Raton.

Lehmann, E.L. (1999) *Elements of Large Sample Theory*, Springer, New York.

Lehmann, E.L. e Romano, J.P. (2022) *Testing Statistical Hypothesis*, quarta edizione, Springer, Switzerland.

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, seconda edizione, Wiley, New York.

Shao, J. (2003) *Mathematical Statistics*, seconda edizione, Springer, New York.

Wilks, S.S. (1962) *Mathematical Statistics*, Wiley, New York.

Pagina intenzionalmente vuota

Capitolo 7

L'inferenza con una variabile

7.1. L'inferenza per un parametro di tendenza centrale

Il modello statistico più semplice assume un campionamento da una singola variabile, in modo tale che l'obiettivo è quello di produrre inferenza per un parametro di tendenza centrale. In un tipico approccio classico, si consideri inizialmente un campione casuale $X_1, \ldots, X_n$ da una variabile casuale $X \sim N(\mu, \sigma^2)$. Gli stimatori di massima verosimiglianza di μ e σ^2 risultano $\widehat{\mu} = \bar{X}$ e $\widehat{\sigma}^2 = S_x^2$. Se si considera il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$ (o una ipotesi direzionale), il test del rapporto delle verosimiglianze fornisce il test di t di Student basato sulla statistica test

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S_{c,x}} ,$$

che sotto ipotesi di base è distribuita come $T \sim t_{n-1}$. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \mu \neq \mu_0$, si rifiuta H_0 per realizzazioni basse o elevate di T. Nel caso dell'ipotesi alternativa direzionale $H_1 : \mu > \mu_0$ ($H_1 : \mu < \mu_0$), si rifiuta H_0 per realizzazioni elevate (basse) di T. Infine, l'intervallo di confidenza per μ al livello di confidenza $(1 - \alpha)$ basato sul test T risulta

$$(\bar{X} - t_{n-1,1-\alpha/2} S_{c,x}/\sqrt{n}, \bar{X} + t_{n-1,1-\alpha/2} S_{c,x}/\sqrt{n}) .$$

Il test basato su T è "distribution-free" per grandi campioni dal momento che T converge in distribuzione a una variabile casuale con distribuzione $N(0, 1)$ per $n \to \infty$, se $\sigma^2 < \infty$. In questo caso, l'intervallo di confidenza per grandi campioni per μ al livello di confidenza $(1 - \alpha)$ risulta

$$(\bar{X} - z_{1-\alpha/2} S_{c,x}/\sqrt{n}, \bar{X} + z_{1-\alpha/2} S_{c,x}/\sqrt{n}) .$$

Il test t di Student è quindi "robusto" rispetto all'assunzione di Normalità.

• **Esempio 7.1.1.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. Dal momento che l'azienda che produce le sfere desidera produrre sfere con un diametro standard di un micron, il sistema di ipotesi risulta $H_0 : \mu = 1$ contro $H_1 : \mu \neq 1$. Il comando `t.test` fornisce l'implementazione del test t di Student e il relativo intervallo di confidenza:

```
> t.test(Diameter, alternative = "two.sided", mu = 1)

        One Sample t-test

data:  Diameter
t = 2.1178, df = 9, p-value = 0.06327
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 0.9867741 1.4012259
sample estimates:
mean of x
    1.194
```

La significatività osservata porta dunque ad ipotizzare che la produzione sia leggermente fuori controllo. □

In un approccio "distribution-free", si consideri un campione casuale da una variabile casuale continua X con funzione di ripartizione $F(x-\lambda)$, dove λ rappresenta la mediana. Si osservi che in questo caso l'inferenza è relativa alla mediana, invece che alla media, per non restringere eccessivamente il modello statistico. In effetti, esistono ampie classi di variabili casuali continue per cui la media non è finita o non è definita.

Se si considera il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$ (o una ipotesi direzionale), il test dei segni è basato sulla statistica test

$$B = \sum_{i=1}^{n} \mathbf{1}_{]0,\infty[}(X_i - \lambda_0) \, ,$$

che sotto ipotesi di base è tale che $B \sim Bi(n, 1/2)$. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda \neq \lambda_0$, si rifiuta H_0 per realizzazioni basse o elevate di B. Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda > \lambda_0$ ($H_1 : \lambda < \lambda_0$), si rifiuta H_0 per realizzazioni elevate (basse) di B. Lo stimatore di λ basato sul test B è la mediana campionaria $\widetilde{X}_{0.5}$. Inoltre, se $b_{n,\alpha}$ rappresenta il quantile di ordine α di B, l'intervallo di confidenza per λ al livello di confidenza $(1-\alpha)$ basato sul test B risulta

$$(X_{(b_{n,\alpha/2}+1)}, X_{(n-b_{n,\alpha/2})}) \, .$$

• **Esempio 7.1.2.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4 e il sistema di ipotesi $H_0 : \lambda = 1$ contro $H_1 : \lambda \neq 1$. Anche se non esiste un comando specifico per il test dei segni, si può equivalentemente adottare il comando `binom.test` per verifiche di ipotesi sulle proporzioni:

```
> binom.test(length(Diameter[Diameter > 1]), length(Diameter),
+    p = 1/2, alternative = "two.sided")

        Exact binomial test

data:  length(Diameter[Diameter > 1]) and length(Diameter)
number of successes = 8, number of trials = 10, p-value = 0.1094
alternative hypothesis: true probability of success is not equal to
0.5
95 percent confidence interval:
 0.4439045 0.9747893
sample estimates:
probability of success
                   0.8
```

Inoltre, la stima per punti e gli estremi dell'intervallo di confidenza per λ possono essere ottenuti mediante i seguenti comandi:

```
> median(Diameter)
[1] 1.185
> sd <- sort(Diameter)
> sd[qbinom(0.025, length(Diameter), 1 / 2)]
[1] 0.88
> sd[qbinom(0.975, length(Diameter), 1 / 2)]
[1] 1.42
```

Si osservi che in questo caso la significatività osservata e l'ampiezza dell'intervallo di confidenza sono maggiori di quelli ottenuti nell'Esempio 7.1.1. Questo fenomeno deriva dal fatto che le assunzioni fatte per il test dei segni sono molto più blande rispetto a quelle fatte per il test t di Student e quindi vi è una maggiore incertezza nel processo inferenziale. □

Di nuovo con un approccio "distribution-free", si consideri un campione casuale da una variabile casuale continua X, simmetrica rispetto alla mediana λ e con funzione di ripartizione $F(x-\lambda)$. La variabile casuale X è simmetrica rispetto alla mediana se la distribuzione di $(X-\lambda)$ coincide con quella di $(\lambda - X)$. Quindi, rispetto alle assunzioni fatte per il test dei segni, si assume questa ulteriore ipotesi.

Se si considera il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$ (o una ipotesi direzionale), il test di Wilcoxon è basato sulla statistica test

$$W^+ = \sum_{i=1}^{n} \mathbf{1}_{]0,\infty[}(X_i - \lambda_0) R_i^+ \,,$$

dove $R_1^+, \ldots, R_n^+$ rappresentano i ranghi assegnati alle trasformate $|X_1 - \lambda_0|, \ldots, |X_n - \lambda_0|$. La distribuzione di W^+ sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda \neq \lambda_0$, si rifiuta H_0 per realizzazioni basse o elevate di W^+. Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda > \lambda_0$ ($H_1 : \lambda < \lambda_0$), si rifiuta H_0 per realizzazioni elevate (basse) di W^+.

Se $k = n(n+1)/2$, siano $W_1, \ldots, W_k$ le medie di Walsh, ovvero tutte le possibili k semisomme distinte delle n osservazioni. La stima di λ basata sul test W^+ è la mediana delle medie di Walsh (la cosiddetta pseudomediana). Se $w^+_{n,\alpha}$ rappresenta il quantile di ordine α di W^+ e se $W_{(1)}, \ldots, W_{(k)}$ è la statistica ordinata relativa alle medie di Walsh, allora l'intervallo di confidenza per λ al livello di confidenza $(1-\alpha)$ basato sul test di Wilcoxon risulta

$$(W_{(w^+_{n,\alpha/2}+1)}, W_{(k-w^+_{n,\alpha/2})}) \,.$$

• **Esempio 7.1.3.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4 e il sistema di ipotesi $H_0 : \lambda = 1$ contro $H_1 : \lambda \neq 1$. Il comando `wilcox.test` fornisce l'implementazione del test di Wilcoxon:

```
> wilcox.test(Diameter, alternative = "two.sided",
+     mu = 1, conf.int = TRUE)

        Wilcoxon signed rank test

data:  Diameter
V = 46, p-value = 0.06445
alternative hypothesis: true location is not equal to 1
95 percent confidence interval:
 0.985 1.405
sample estimates:
(pseudo)median
          1.19
```

In questo caso, la significatività osservata e l'ampiezza dell'intervallo di confidenza sono quasi identici a quelli ottenuti nell'Esempio 7.1.1. Dunque, l'ipotesi di simmetria aumenta l'efficienza della procedura "distribution-free" in modo sostanziale. □

Con un approccio "distribution-free", si consideri un campione scambiabile $X_1, \ldots, X_n$ da una variabile casuale continua e simmetrica X con funzione di ripartizione $F(x - \lambda)$, dove λ rappresenta di nuovo la mediana. Un campione è detto scambiabile se la distribuzione di $X_1, \ldots, X_n$ coincide con quella di $X_{i_1}, \ldots, X_{i_n}$ per qualsiasi permutazione $i_1, \ldots, i_n$ nell'insieme $\mathcal{S}_n$ delle permutazioni dei primi n interi. Dunque, un campione scambiabile non è generalmente un campione casuale, in quanto le osservazioni campionarie possono essere dipendenti.

Se si considera il sistema di ipotesi $H_0 : \lambda = \lambda_0$ contro $H_1 : \lambda \neq \lambda_0$ (o una ipotesi direzionale), condizionatamente alla realizzazione $x_1, \ldots, x_n$ del campione, sotto ipotesi di base i valori $-|x_i - \lambda_0|$ e $+|x_i - \lambda_0|$ sono ugualmente probabili. Il test di permutazione è basato sulle 2^n (ugualmente probabili) permutazioni dei segni delle osservazioni trasformate $|x_1 - \lambda_0|, \ldots, |x_n - \lambda_0|$. Se $S_1, \ldots, S_n$ rappresenta un vettore scambiabile di variabili casuali di Bernoulli ognuna con parametro $1/2$ e supporto $\{-1, 1\}$, il test di permutazione dei segni è basato sulla statistica test

$$T = \sum_{i=1}^{n} |x_i - \lambda_0| \, S_i \, .$$

La distribuzione della statistica test sotto ipotesi di base può essere calcolata, anche se non è possibile esprimerla in forma chiusa. Tuttavia, per ogni differente realizzazione campionaria, si deve effettuare un nuovo calcolo, che comunque è proibitivo anche per numerosità campionarie modeste. Quindi, la distribuzione della statistica test è usualmente approssimata mediante il metodo Monte Carlo. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda \neq \lambda_0$, si rifiuta H_0 per realizzazioni basse o elevate di T. Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda > \lambda_0$ ($H_1 : \lambda < \lambda_0$), si rifiuta H_0 per realizzazioni elevate (basse) di T.

Si deve osservare che l'approccio basato sui test di permutazione è radicale, nel senso che l'inferenza è completamente condizionata al campione osservato. Questo paradigma può essere evidentemente in contrasto con un approccio come quello Bayesiano.

• **Esempio 7.1.4.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. Richiamando la libreria `exactRankTests`, il comando `wilcox.test` fornisce l'implementazione del test di permutazione dei segni. Le osservazioni sono state moltiplicate per cento al fine di ottenere valori interi per ottenere una elaborazione più rapida. Di conseguenza, anche il sistema di ipotesi risulta $H_0 : \lambda = 100$ contro $H_1 : \lambda \neq 100$. Il test di permutazione viene implementato mediante i seguenti comandi:

```
> library(exactRankTests)
> perm.test(round(100 * Diameter), paired = FALSE,
+         alternative = "two.sided", mu = 100)

        1-sample Permutation Test

data:  round(100 * Diameter)
T = 237, p-value = 0.07031
alternative hypothesis: true mu is not equal to 100
```

La significatività osservata risulta sono quasi identica a quella ottenuti nell'Esempio 7.1.1. Tuttavia, in questo approccio non è stato fatta neanche l'assunzione di indipendenza delle osservazioni rispetto al test classico. □

Con un approccio "distribution-free", si supponga che $x_1, \ldots, x_n$ sia la determinazione di un campione casuale da una variabile casuale X con media μ. Di nuovo in un approccio condizionato alla realizzazione del campione $x_1, \ldots, x_n$, la cosiddetta distribuzione "bootstrap" della media campionaria può essere ottenuta considerando tutti i campioni con ripetizione di ordine n dai valori $x_1, \ldots, x_n$. Se $B_1, \ldots, B_n$ rappresenta un vettore di variabili casuali con distribuzione Binomiale

(ognuna con parametri n e $1/n$) e tali che $\sum_{i=1}^{n} B_i = n$, allora la media campionaria bootstrap è data dalla statistica test

$$\bar{X}_{\text{BOOT}} = \frac{1}{n} \sum_{i=1}^{n} x_i B_i \, .$$

Se si considera il sistema di ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$ (o una ipotesi direzionale), il test bootstrap e il relativo intervallo di confidenza bootstrap della media possono essere basati sulla statistica $\bar{X}_{\text{BOOT}}$. La distribuzione bootstrap della statistica può essere calcolata, anche se ovviamente non è possibile esprimerla in forma chiusa. Anche in questo caso, la distribuzione bootstrap della statistica è usualmente approssimata mediante il metodo Monte Carlo. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \mu \neq \mu_0$, si rifiuta H_0 per realizzazioni basse o elevate della statistica test. Nel caso dell'ipotesi alternativa direzionale $H_1 : \mu > \mu_0$ ($H_1 : \mu < \mu_0$), si rifiuta H_0 per realizzazioni elevate (basse) della statistica test.

• **Esempio 7.1.5.** Si considerano di nuovo i dati relativi alle sfere di acciaio dell'Esempio 4.3.4. La distribuzione bootstrap mediante il metodo Monte Carlo, e il relativo istogramma, della media campionaria si possono ottenere mediante i seguenti comandi:

```
> Boot.mean <- numeric(10000)
> for (i in 1:10000) Boot.mean[i] <- mean(sample(Diameter,
+    replace = T))
> hist(Boot.mean, xlab = "Bootstrap sample mean",
+    ylab = "Density", main = "Histogram")
```

Il precedente comando fornisce il grafico della Figura 7.1.1.

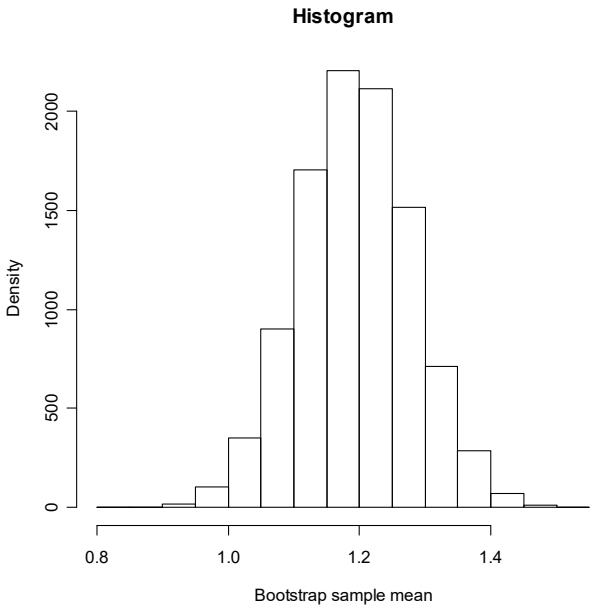


Figura 7.1.1.

Inoltre, la significatività osservata del test bootstrap può essere ottenuta mediante il seguente comando

```
> 2 * length(Boot.mean[Boot.mean < 1]) / 10000
[1] 0.0218
```

Richiamando la libreria `exactRankTests`, il comando `boot` fornisce l'implementazione della stima bootstrap della distorsione e della varianza della media campionaria:

```
> library(boot)
> m <- function(x, w) sum(x$Diameter * w)
> boot(d, m, R = 9999, stype = "w")

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = d, statistic = m, R = 9999, stype = "w")

Bootstrap Statistics :
    original       bias    std. error
t1*    1.194 0.0006592659  0.08684963
```

Inoltre, Il comando `boot.ci` fornisce l'implementazione dell'intervallo di confidenza bootstrap per la media con quattro metodi differenti:

```
> boot.ci(boot(d, m, R = 9999, stype = "w"), conf = 0.95,
+         type = c("norm", "basic", "perc", "bca"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 9999 bootstrap replicates

CALL :
boot.ci(boot.out = boot(d, m, R = 9999, stype = "w"), conf = 0.95,
    type = c("norm", "basic", "perc", "bca"))

Intervals :
Level      Normal              Basic
95%   ( 1.025,  1.363 )   ( 1.028,  1.363 )

Level     Percentile            BCa
95%   ( 1.025,  1.360 )   ( 1.021,  1.356 )
Calculations and Intervals on Original Scale
```

Dunque, la significatività osservata e l'ampiezza dell'intervallo di confidenza con l'approccio bootstrap sono minori di quelli ottenuti nell'Esempio 7.1.1 con un test classico. □

Si considerino le osservazioni relative a n soggetti su cui è stata osservata una certa variabile prima e dopo un trattamento, ovvero si hanno le osservazioni $x_{11}, \ldots, x_{n1}$ prima del trattamento e le osservazioni $x_{12}, \ldots, x_{n2}$ dopo il trattamento. L'obiettivo è quello di valutare l'efficacia del trattamento e i dati di questo tipo sono detti appaiati. Al fine di analizzare queste osservazioni si costruiscono le differenze delle osservazioni $d_1, \ldots, d_n$, dove $d_i = x_{i2} - x_{i1}$. Supponendo che queste differenze siano realizzazioni di un campione casuale proveniente da una variabile casuale da $D \sim N(\mu, \sigma^2)$, la verifica dell'efficacia del trattamento si riduce a considerare il sistema di ipotesi sulla media $H_0 : \mu = 0$ contro $H_1 : \mu \neq 0$ (o una ipotesi direzionale). Alternativamente, assumendo D come una variabile casuale con funzione di ripartizione (non nota) $F(x - \lambda)$ dove λ rappresenta la mediana, si può considerare il sistema di ipotesi sulla mediana $H_0 : \lambda = 0$ contro $H_1 : \lambda \neq 0$ (o una ipotesi direzionale). In questo caso, è sufficiente applicare le procedure di verifica di ipotesi viste in precedenza.

• **Esempio 7.1.6.** Su 8 pazienti con anemia cronica grave è stato misurato l'indice di infarto (in ml/battito/m^2) prima e dopo un trattamento medico (Fonte: Bhatia, M.L., Manchanda, S.C. e Roy, S.B., 1969, Coronary haemodinamic studies in chronic severe anaemia, *British Heart Journal* **31**, 365-374). I dati sono contenuti nel file `stroke.txt` e vengono letti e resi disponibili mediante i seguenti comandi:

```
> d <- read.table("c:\\Rwork\\examples\\stroke.txt", header = T)
> attach(d)
> Difference <- Post - Pre
```

La stima di nucleo della funzione di densità viene ottenuta mediante i seguenti comandi:

```
> library(sm)
> sm.density(Difference, hcv(Difference, hstart = 0.01, hend = 100),
+    yht = 0.02, xlim = c(-70, 30),
+    xlab = "Stroke index difference")
> title(main = "Kernel density estimation ('CV' h = 18.272)")
```

I precedenti comandi forniscono il grafico della Figura 7.1.2.

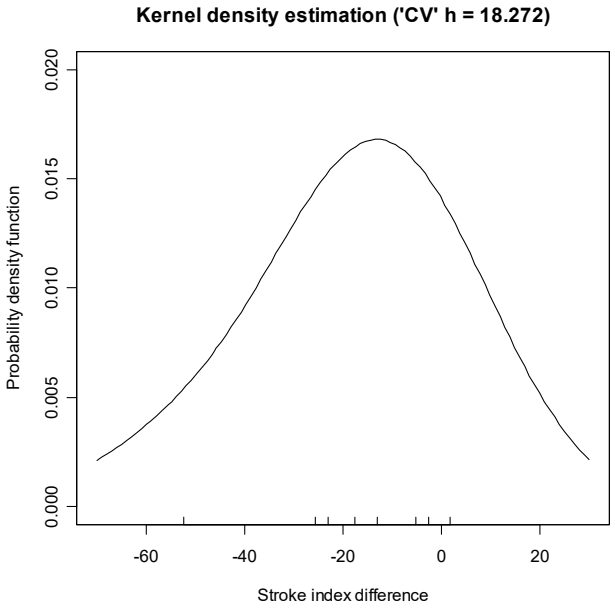


Figura 7.1.2.

Il diagramma quantile-quantile per verificare empiricamente l'ipotesi di normalità viene ottenuto mediante i seguenti comandi:

```
> qqnorm(Difference)
> qqline(Difference)
```

I precedenti comandi forniscono il grafico della Figura 7.1.3.

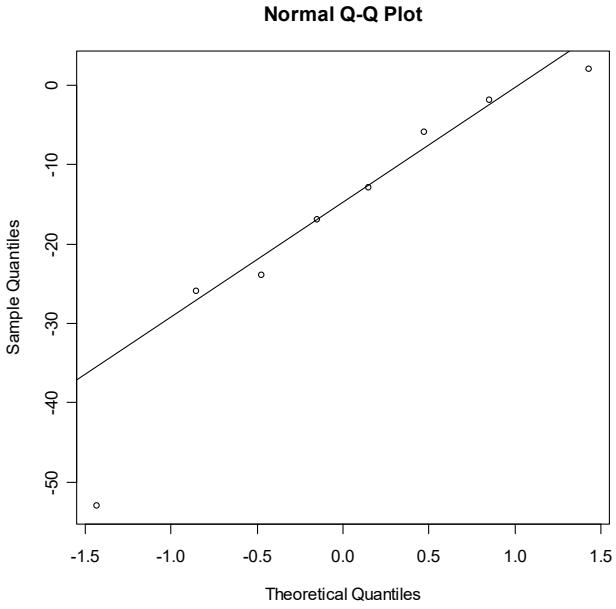


Figura 7.1.3.

Il sistema di ipotesi $H_0 : \mu = 0$ contro $H_1 : \mu < 0$ può essere verificato con il test t di Student mediante il seguente comando:

```
> t.test(Difference, alternative = "less", mu = 0)

        One Sample t-test

data:  Difference
t = -2.8055, df = 7, p-value = 0.01316
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
     -Inf -5.64162
sample estimates:
mean of x
  -17.375
```

Il sistema di ipotesi $H_0 : \lambda = 0$ contro $H_1 : \lambda < 0$ può essere verificato con il test di Wilcoxon mediante il seguente comando:

```
> wilcox.test(Difference, alternative = "less", mu = 0,
+    conf.int = F)

        Wilcoxon signed rank test with continuity correction

data:  Difference
V = 1.5, p-value = 0.01244
alternative hypothesis: true location is less than 0
```

Le significatività osservate ottenute con l'approccio classico e “distribution-free” sono dunque molto simili e portano a ritenere il trattamento efficace, in quanto l'indice di infarto tende a diminuire dopo il trattamento. □

7.2. I test funzionali

Assumendo un modello statistico per un campionamento da una singola variabile, l'obiettivo è spesso quello di produrre inferenza per l'intera distribuzione. Questi metodi inferenziali sono anche detti test per la bontà d'adattamento. Si consideri inizialmente un campione casuale da una variabile casuale continua X con funzione di ripartizione F. Si desidera verificare il sistema di ipotesi funzionale $H_0 : F(x) = F_0(x)$ per ogni x, contro $H_1 : F(x) \neq F_0(x)$ per qualche x, dove F_0 è una funzione di ripartizione completamente specificata. Il test si basa sulla statistica di Kolmogorov data da

$$D = \sup_x |\widehat{F}(x) - F_0(x)| \, .$$

Si può dimostrare che la distribuzione di D non dipende da F e quindi il test è "distribution-free". Inoltre, la distribuzione della statistica test sotto ipotesi di base è nota e può essere tabulata. Se la funzione di ripartizione empirica si discosta molto da quella ipotizzata, allora si hanno valori elevati di D che conseguentemente portano a respingere l'ipotesi di base in favore dell'ipotesi alternativa. Per un calcolo pratico, la statistica D può essere convenientemente espressa come

$$D = \max_{1 \leq i \leq n} (\max(|i/n - F_0(X_{(i)})|, |(i-1)/n - F_0(X_{(i)})|)) \, .$$

• **Esempio 7.2.1.** Sono stati determinati i carichi da applicare a un campione di fibre di poliestere al fine di provocarne il cedimento (Fonte: Quesenberry, C.P. e Hales, C., 1980, Concentration bands for uniformity plots, *Journal of Statistical Computation and Simulation* **11**, 41-53). Si sospetta che la distribuzione dei carichi segua una distribuzione Log-Normale. Le osservazioni originali sono state ricalcolate mediante una trasformazione che conduce all'uniformità, ovvero il nuovo campione deve provenire da una variabile casuale con distribuzione $U(0, 1)$ se l'ipotesi di log-normalità è vera. In questo caso si ha

$$F_0(x) = x\mathbf{1}_{]0,1[}(x) + \mathbf{1}_{[1,\infty[}(x) \, .$$

I dati sono contenuti nel file `tensile.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\tensile.txt", header = T)
> attach(d)
```

Il grafico della funzione di ripartizione empirica e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> plot(ecdf(Strength), do.points = F, verticals = T,
+    xlab = "Transformed tensile strength", ylab = "Probability",
+    main = "Distribution function")
> rug(Strength)
> plot(function(x) punif(x), -0.05, 1.05, lty = 3,
+    ylab = "Probability", add = T)
> legend(0.6, 0.3, c("Empirical", "Theorical"), lty = c(1, 3))
```

I precedenti comandi forniscono il grafico della Figura 7.2.1.

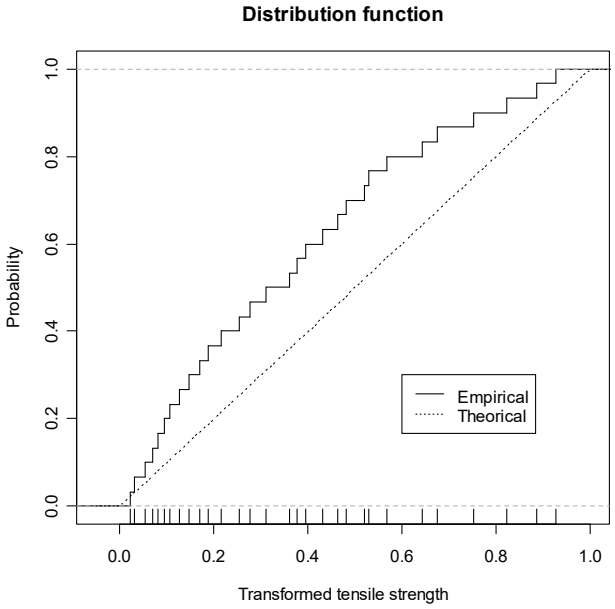


Figura 7.2.1.

Il test di Kolmogorov viene eseguito mediante il seguente comando:

```
> ks.test(Strength, "punif", 0, 1)

        One-sample Kolmogorov-Smirnov test

data:  Strength
D = 0.2377, p-value = 0.05644
alternative hypothesis: two-sided
```

La significatività osservata porta ad ipotizzare che l'uniformità della distribuzione considerata sia questionabile, anche se non è abbastanza bassa per permettere un rifiuto netto dell'ipotesi di base. □

Se si considera un campionamento casuale da una variabile casuale discreta a supporto finito o da una variabile qualitativa X, allora le osservazioni campionarie possono essere espresse attraverso le frequenze osservate $n_1, \ldots, n_r$ delle realizzazioni distinte $c_1, \ldots, c_r$ della variabile in analisi. Se la funzione di probabilità di X è data da $p(c_j) = \pi_j$ per $j = 1, \ldots, r$, le quantità $n\pi_1, \ldots, n\pi_r$ sono dette frequenze attese. Si è interessati a verificare il sistema di ipotesi $H_0 : \pi_j = \pi_{0j}$ per ogni j, contro $H_1 : \pi_j \neq \pi_{0j}$ per qualche j. Dal momento che le probabilità π_j specificano completamente la distribuzione di X, la precedente ipotesi è a tutti gli effetti una ipotesi funzionale. Per verificare questo sistema di ipotesi si adotta la statistica test Chi-quadrato per la bontà d'adattamento data da

$$\chi^2 = \sum_{j=1}^{r} \frac{(n_j - n\pi_{0j})^2}{n\pi_{0j}} .$$

La distribuzione per grandi campioni di χ^2 non dipende dai valori π_{0j} e quindi il test è "distribution-free" per grandi campioni. In effetti, sotto ipotesi di base, per $n \to \infty$ la statistica test χ^2 converge in distribuzione ad una variabile casuale con distribuzione χ^2_{r-1}. L'approssimazione è valida per campioni finiti se $n > 30$ e se tutte le frequenze attese sono maggiori di uno. Se le frequenze osservate si discostano molto dalle frequenze attese, si ottengono determinazioni elevate della statistica test che portano a respingere l'ipotesi di base.

• **Esempio 7.2.2.** È stata osservata la prima cifra dei numeri contenuti in un volume della rivista Reader's Digest scelto casualmente (Fonte: Benford, F., 1938, The law of anomalous numbers, *Proceedings of the American Philosophical Society* **78**, 551-572). Un modello teorico per questi dati è la cosiddetta distribuzione anomala con funzione di probabilità

$$p_0(x) = \log_{10}\left(1 + \frac{1}{x}\right) \mathbf{1}_{\{1,\ldots,9\}}(x) \, .$$

I dati sono contenuti nel file `benford.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\benford.txt", header = T)
> attach(d)
```

Le probabilità teoriche vengono calcolate mediante i seguenti comandi:

```
> Theory.Probs <- numeric(9)
> for (i in 1:9) Theory.Probs[i] <- logb(1 + 1 / i, 10)
```

Il grafico della distribuzione di probabilità osservata e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> h <- list(Digit = c(Digit, Digit), Type = c(rep("Theorical", 9),
+     rep("Observed", 9)),
+     Probs = c(Theory.Probs, Counts / sum(Counts)))
> class(Table <- xtabs(Probs ~ ., h))
[1] "xtabs" "table"
> barplot(t(Table), beside = T, legend = colnames(Table),
+     xlab = "Digit", ylab = "Probability",
+     main = "Probability distribution")
```

I precedenti comandi forniscono il grafico della Figura 7.2.2.

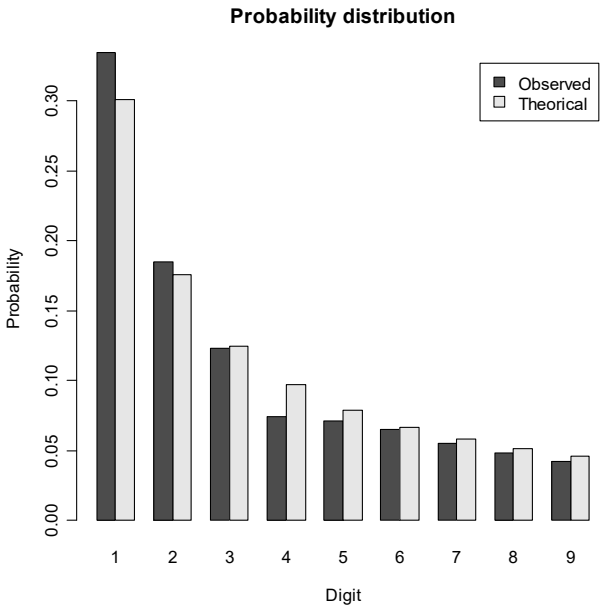


Figura 7.2.2.

Il test χ^2 viene ottenuto mediante il seguente comando:

```
> chisq.test(xtabs(Counts ~ ., d), p = Theory.Probs)

        Chi-squared test for given probabilities

data:  xtabs(Counts ~ ., d)
X-squared = 3.2735, df = 8, p-value = 0.916
```

Dal momento che la significatività osservata è molto elevata, si può accettare l'ipotesi di base. □

• **Esempio 7.2.3.** In un esperimento di genetica sono stati considerati ibridi di pomodoro con un rapporto atteso di quattro fenotipi pari a $9:3:3:1$ ottenendo le frequenze del numero di piante generate per ogni fenotipo (Fonte: McArthur, J.W., 1931, Linkage studies with the tomato III. Fifteen factors in six groups, *Transaction of the Royal Canadian Institute* **18**, 1-19. Si vuole verificare sperimentalmente i risultati della teoria genetica, ovvero l'ipotesi di base $H_0 : \pi_1 = 9/16$, $\pi_2 = 3/16, \pi_3 = 3/16, \pi_4 = 1/16$. I dati sono contenuti nel file `tomato.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\tomato.txt", header = T)
> attach(d)
```

Le probabilità teoriche vengono calcolate mediante i seguenti comandi:

```
> Theory.Probs <- c(9 / 16, 3 / 16, 3 / 16, 1 / 16)
```

Il grafico delle distribuzioni di probabilità osservata e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> h <- list(Phenotype = c(Phenotype, Phenotype),
+    Type = c(rep("Theoretical", 4), rep("Observed", 4)),
+    Probs = c(Theory.Probs, Counts / sum(Counts)))
> class(Table <- xtabs(Probs ~ ., h))
[1] "xtabs" "table"
> barplot(t(Table), beside = T, legend = colnames(Table),
+    names.arg = c("Tall cut-leaf", "Tall potato-leaf",
+    "Dwarf cut-leaf", "Dwarf potato-leaf"),
+    xlab = "Phenotype", ylab = "Probability",
+    main = "Probability distribution")
```

I precedenti comandi forniscono il grafico della Figura 7.2.3. Il test χ^2 viene implementato mediante il seguente comando:

```
> chisq.test(xtabs(Counts ~ ., d), p = Theory.Probs)

        Chi-squared test for given probabilities

data:  xtabs(Counts ~ ., d)
X-squared = 1.4687, df = 3, p-value = 0.6895
```

Dal momento che la significatività osservata è elevata, si può accettare l'ipotesi di base. □

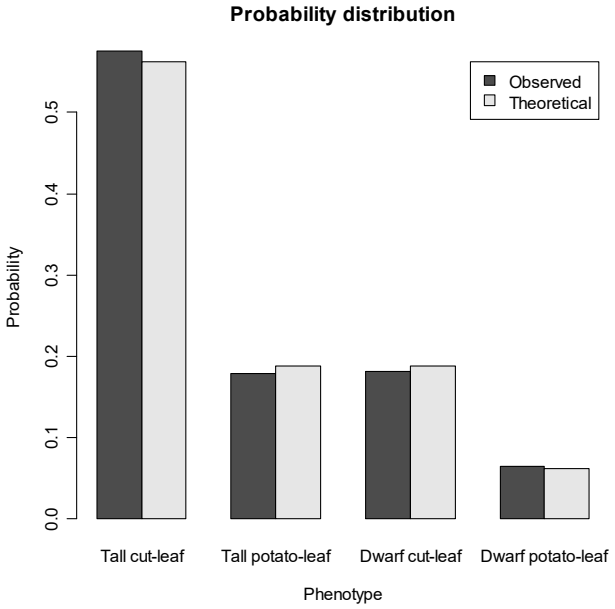


Figura 7.2.3.

Se la funzione di probabilità di X dipende da un insieme di k parametri (non noti) θ, ovvero se $p(c_j) = \pi_j(\theta)$, il sistema di ipotesi diventa $H_0 : \pi_j = \pi_{0j}(\theta)$ per ogni j, contro $H_1 : \pi_j \neq \pi_{0j}(\theta)$ per qualche j. Si assuma l'esistenza uno stimatore $\widehat{\Theta}$ per θ che sia coerente, efficiente per grandi campioni e distribuito normalmente per grandi campioni. Lo stimatore $\widehat{\Theta}$ viene usualmente ottenuto con il metodo della massima verosimiglianza e le quantità $n\pi_1(\widehat{\Theta}), \ldots, n\pi_r(\widehat{\Theta})$ sono dette frequenze attese stimate. Per verificare questo sistema di ipotesi si adotta una opportuna modifica della statistica test Chi-quadrato data da

$$\chi^2 = \sum_{j=1}^{r} \frac{(n_j - n\pi_{0j}(\widehat{\Theta}))^2}{n\pi_{0j}(\widehat{\Theta})} .$$

La distribuzione per grandi campioni di χ^2 non dipende dai valori π_j e quindi il test è "distribution-free" per grandi campioni. Sotto ipotesi di base, per $n \to \infty$ la statistica test χ^2 converge in distribuzione ad una variabile casuale con distribuzione χ^2_{r-k-1}. L'approssimazione è valida per campioni finiti se $n > 30$ e se tutte le frequenze attese stimate sono maggiori di uno. Se le frequenze osservate si discostano molto dalle frequenze attese stimate, si ottengono determinazioni elevate della statistica test che portano a respingere l'ipotesi di base.

• **Esempio 7.2.4.** È stato osservato il numero di taxi arrivati in ogni intervalli di un minuto alla stazione di Euston a Londra fra le 9.00 e le 10.00 di una mattina del 1950 (Fonte: Kendall, D.G., 1951, Some problems in the theory of queues, *Journal of the Royal Statistical Society* **B13**, 151-185). Se gli arrivi sono casuali, per la teoria dei processi stocastici, le osservazioni provengono da $X \sim Po(\mu)$. La distribuzione di Poisson è definita sui numeri naturali, e quindi si devono raggruppare le osservazioni maggiori di un predeterminato valore (in questo caso 5) in una unica classe. I dati sono contenuti nel file `taxi.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\taxi.txt", header = T)
> attach(d)
```

Le probabilità teoriche stimate vengono calcolate mediante i seguenti comandi:

```
> mu <- sum(Taxi * Counts) / sum(Counts)
> Theory.Probs <- numeric(6)
> for (i in 0:4) Theory.Probs[i + 1] <- dpois(i, mu)
> Theory.Probs[6] <- 1 - sum(Theory.Probs)
```

Il grafico delle distribuzioni di probabilità osservata e di quella teorica viene ottenuto mediante i seguenti comandi:

```
> h <- list(Taxi = c(Taxi, Taxi), Type = c(rep("Estimated", 6),
+     rep("Observed", 6)),
+     Probs = c(Theory.Probs, Counts / sum(Counts)))
> class(Table <- xtabs(Probs ~ ., h))
[1] "xtabs" "table"
> barplot(t(Table), beside = T, legend = colnames(Table),
+     names.arg = c("0", "1", "2", "3", "4", "+5"),
+     xlab = "Taxi", ylab = "Probability",
+     main = "Probability distribution")
```

I precedenti comandi forniscono il grafico della Figura 7.2.3.

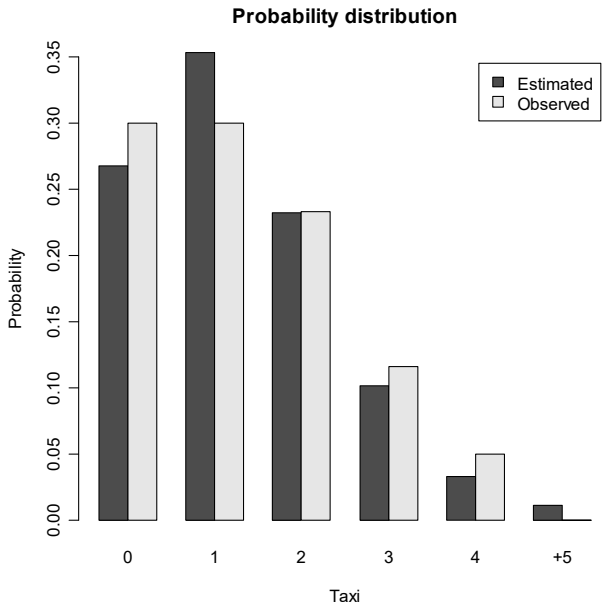


Figura 7.2.4.

Il test χ^2 viene ottenuto mediante il seguente comando:

```
> 1 - pchisq(sum((Counts - sum(Counts) * Theory.Probs)^2 /
+         (sum(Counts) * Theory.Probs)), 4)
[1] 0.7380024
```

Dal momento che la significatività osservata è elevata, si può accettare l'ipotesi di base. □

7.3. Riferimenti bibliografici

Agresti, A. (2019) *An Introduction to Categorical Data Analysis*, terza edizione, Wiley, New York.
Chernick, M.R. (2008) *Bootstrap Methods*, seconda edizione, Wiley, New York.

Davison, A.C. e Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.

Efron, B. e Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, London.

Good (2005) *Permutation, Parametric and Bootstrap Test of Hypotheses*, terza edizione, Springer, New York.

Hájek, J. (1969) *Nonparametric Statistics*, Holden Day, San Francisco.

Hájek, J. e Šidák, Z. (1967) *Theory of Rank Tests*, Academic Press, New York.

Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.

Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.

Lehmann, E.L. e Romano, J.P. (2022) *Testing Statistical Hypothesis*, quarta edizione, Springer, Switzerland.

Shao, J. e Tu, D. (1995) *The jackknife and bootstrap*, Springer, New York.

Pagina intenzionalmente vuota

Capitolo 8

L'inferenza con due variabili

8.1. L'inferenza per i parametri di tendenza centrale

Un modello statistico più elaborato di quello visto nel precedente Capitolo assume un campionamento da due variabili. Se la prima variabile è qualitativa e sotto controllo dello sperimentatore (ovvero è un fattore), mentre l'altra variabile è quantitativa, si ha la tipica situazione usualmente descritta come analisi con due campioni (se il fattore assume due determinazioni) o analisi della varianza (se il fattore assume più determinazioni). Ovviamente, queste strutture campionarie possono essere analizzate sia con un approccio classico che con un approccio "distribution-free". L'analisi della dipendenza è finalizzata invece all'indagine dell'esistenza di un legame fra le due variabili, che possono essere entrambe quantitative o entrambe qualitative.

Risulta importante avere cautela per quanto riguarda l'uso della notazione nel caso in cui si analizzi una variabile qualitativa e una quantitativa. Ad esempio, quando si considera l'analisi con due campioni, viene assunta l'esistenza di due campioni casuali di numerosità n_1 e n_2, tali che $n = n_1 + n_2$, provenienti da due differenti variabili casuali. Tuttavia, si dovrebbe affermare più correttamente che si dispone di n_1 osservazioni della variabile quantitativa Y al livello c_1 della variabile qualitativa X e di n_2 osservazioni della variabile quantitativa Y al livello c_2 della variabile qualitativa X. Dunque, si dovrebbe adoperare la notazione $(Y \mid X = c_1)$ e $(Y \mid X = c_2)$ per specificare le due variabili casuali da cui si sta campionando. Per parsimonia e con un lieve abuso di notazione, queste due variabili casuali vengono denotate con Y_{c_1} e Y_{c_2} nel seguito. Analogamente, se il fattore X assume r livelli $c_1, \ldots, c_r$ (ovvero quando si considera l'analisi della varianza) si adotta la notazione $Y_{c_1}, \ldots, Y_{c_r}$. In questi casi, si ha comunque un'analisi statistica relativa a due variabili, ovvero una variabile qualitativa X sotto controllo e una variabile quantitativa Y su cui si desidera effettuare l'inferenza.

In un tipico approccio classico, si considerino inizialmente due campioni casuali indipendenti di numerosità n_1 e n_2, con $n = n_1 + n_2$, da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che $Y_{c_1} \sim N(\mu_1, \sigma^2)$ e $Y_{c_2} \sim N(\mu_2, \sigma^2)$. Il modello statistico è quindi indicizzato dai parametri μ_1, μ_2 e σ^2. Quando si assume l'omogeneità delle varianze di Y_{c_1} e Y_{c_2} come in questo caso, si ha la cosiddetta ipotesi di omoschedasticità. Si indichi con $\overline{Y}_{c_1}$ e $S^2_{c_1}$ la media campionaria e la varianza campionaria delle osservazioni provenienti da Y_{c_1}. Analogamente, si indichi con $\overline{Y}_{c_2}$ e $S^2_{c_2}$ la media campionaria e la varianza campionaria delle osservazioni provenienti da Y_{c_2}. Si denoti inoltre con

$$S^2_w = \frac{1}{n}\left(n_1 S^2_{c_1} + n_2 S^2_{c_2}\right)$$

la media ponderata delle varianze campionarie. Si può facilmente dimostrare che $\overline{Y}_{c_1}$, $\overline{Y}_{c_2}$ e S^2_w sono gli stimatori di massima verosimiglianza di μ_1, μ_2 e σ^2. Se si considera il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ (o una ipotesi direzionale), il test del rapporto delle verosimiglianze fornisce il test t di Student a due campioni basato sulla statistica test

$$T = \sqrt{\frac{n_1 n_2 (n-2)}{n^2}}\,\frac{\overline{Y}_{c_2} - \overline{Y}_{c_1}}{S_w}\,,$$

che sotto ipotesi di base è distribuita come $T \sim t_{n-2}$. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \mu_1 \neq \mu_2$, si rifiuta H_0 per realizzazioni basse o elevate di T. Nel caso dell'ipotesi alternativa direzionale $H_1 : \mu_1 > \mu_2$ ($H_1 : \mu_1 < \mu_2$), si rifiuta H_0 per realizzazioni basse (elevate) di T. Il test basato su T è "distribution-free" per grandi campioni dal momento che T converge in distribuzione a una variabile casuale con distribuzione $N(0,1)$ per $n \to \infty$. Il test t di Student a due campioni è quindi "robusto" rispetto all'assunzione di normalità.

Si è evidenziato che il test t di Student a due campioni assume l'omoschedasticità. Al fine di verificare l'omogeneità delle varianze è opportuno assumere $Y_{c_1} \sim N(\mu_1, \sigma_1^2)$ e $Y_{c_2} \sim N(\mu_2, \sigma_2^2)$ e verificare il sistema di ipotesi $H_0 : \sigma_1^2 = \sigma_2^2$ contro $H_1 : \sigma_1^2 \neq \sigma_2^2$. In questo caso, il test del rapporto delle verosimiglianze fornisce il test di Fisher basato sulla statistica test

$$F = \frac{S_{c_1}^2}{S_{c_2}^2},$$

che sotto ipotesi di base si distribuisce come $F \sim F_{n_1-1,n_2-1}$. L'ipotesi di base viene rifiutata per realizzazioni elevate di F. Nel caso che l'ipotesi di omoschedasticità venga respinta, per verificare l'omogeneità delle medie si può comunque adoperare la statistica test

$$T = \frac{\overline{Y}_{c_2} - \overline{Y}_{c_1}}{\sqrt{S_{c_1}^2/n_1 + S_{c_2}^2/n_2}},$$

che converge in distribuzione a una variabile casuale con distribuzione $N(0,1)$ per $n \to \infty$ ed è quindi "distribution-free" per grandi campioni.

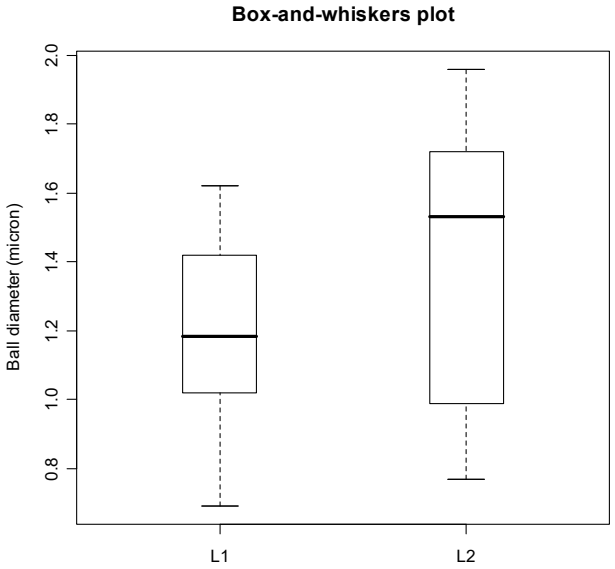


Figura 8.1.1.

• **Esempio 8.1.1.** Si dispone delle osservazioni di due campioni casuali di diametri di sfere misurate in micron provenienti da due differenti linee di produzione (Fonte: Romano, A., 1977, *Applied Statistics for Science and Industry*, Allyn and Bacon, Boston). I dati sono contenuti nel file `ball2.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\ball2.txt", header = T)
> attach(d)
```

I dati originali vengono ripartiti nei due campioni mediante i comandi:

```
> Diameter.1 <- split(Diameter, Line)[[1]]
> Diameter.2 <- split(Diameter, Line)[[2]]
```

Il diagramma a scatola e baffi condizionato viene ottenuto mediante il comando:

```
> boxplot(Diameter ~ Line, boxwex = 0.3,
+    ylab = "Ball diameter (micron)",
+    main = "Box-and-whiskers plot")
```

Il precedente comando fornisce il grafico della Figura 8.1.1. Le stime di nucleo delle funzioni di densità vengono ottenute mediante i comandi:

```
> library(sm)
> hcv(Diameter.1, hstart = 0.01, hend = 1)
> sm.density(Diameter.1, hcv(Diameter, hstart = 0.01, hend = 1),
+    yht = 0.92, xlim = c(-0.35, 2.65),
+    xlab = "Line 1 ball diameter (micron)")
> title(main = "Kernel density estimation ('CV' h = 0.32)")
> hcv(Diameter.2, hstart = 0.01, hend = 1)
> sm.density(Diameter.2, hcv(Diameter, hstart = 0.01, hend = 1),
+    yht = 0.92, xlim = c(-0.35, 2.65),
+    xlab = "Line 2 ball diameter (micron)")
> title(main = "Kernel density estimation ('CV' h = 0.37)")
```

I precedenti comandi forniscono il grafico della Figura 8.1.2.

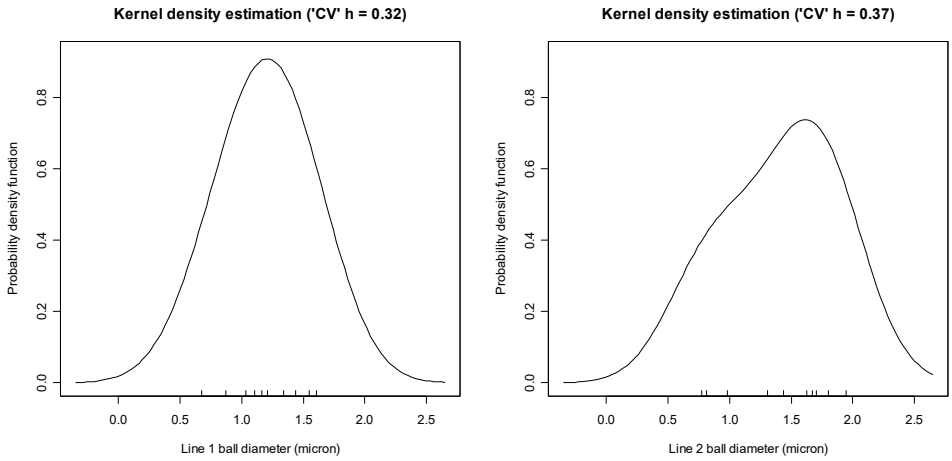


Figura 8.1.2.

Il sistema di ipotesi $H_0 : \sigma_1^2 = \sigma_2^2$ contro $H_1 : \sigma_1^2 \neq \sigma_2^2$ può essere verificato mediante il comando `var.test` che fornisce l'implementazione del test F di Fisher:

```
> var.test(Diameter.1, Diameter.2, paired = F,
+    alternative = "two.sided")
```

```
        F test to compare two variances

data:  Diameter.1 and Diameter.2
F = 0.4574, num df = 9, denom df = 9, p-value = 0.2595
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1136199 1.8416221
sample estimates:
ratio of variances
         0.4574329
```

Il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ può essere verificato mediante il comando `t.test` che fornisce l'implementazione del test t di Student per due campioni:

```
> t.test(Diameter.1, Diameter.2, var.equal = T,
+    alternative = "two.sided")

        Two Sample t-test

data:  Diameter.1 and Diameter.2
t = -1.2965, df = 18, p-value = 0.2112
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5555277  0.1315277
sample estimates:
mean of x mean of y
    1.194     1.406
```

Il test per l'omogeneità delle medie senza assunzione di omoschedasticità viene ottenuto mediante i seguenti comandi:

```
> 2 * pnorm((mean(Diameter.1) - mean(Diameter.2)) /
+    (variance(Diameter.1) / length(Diameter.1) +
+    variance(Diameter.2) / length(Diameter.2))^(1 / 2))
[1] 0.1717296
```

Le significatività osservate portano dunque ad accettare l'omoschedasticità e successivamente l'omogeneità delle medie. Anche il test senza l'ipotesi di omoschedasticità permette di accettare l'ipotesi di omogeneità delle medie. □

In un approccio "distribution-free", si consideri di nuovo due campioni casuali indipendenti da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che Y_{c_1} ha funzione di ripartizione $F(y - \lambda_1)$ e Y_{c_2} ha funzione di ripartizione $F(y - \lambda_2)$, mentre λ_1 e λ_2 rappresentano le rispettive mediane. Si indichi come campione misto l'insieme di tutte le n osservazioni senza considerare l'effetto del fattore. Inoltre, si assuma che $R_1, \ldots, R_{n_1}$ siano i ranghi assegnati alle osservazioni campionarie da Y_{c_1} nel campione misto, mentre siano $R_{n_1+1}, \ldots, R_n$ i ranghi assegnati alle osservazioni campionarie da Y_{c_2} nel campione misto. Se si considera il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2$ contro $H_1 : \lambda_1 \neq \lambda_2$ (o una ipotesi direzionale), il test di Mann-Withney è basato sulla statistica test

$$W = \sum_{i=1}^{n_1} R_i \, .$$

La distribuzione di W sotto ipotesi di base può essere tabulata, anche se non può essere espressa in forma chiusa. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda_1 \neq \lambda_2$, si rifiuta H_0 per realizzazioni

basse o elevate di W. Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda_1 > \lambda_2$ ($H_1 : \lambda_1 < \lambda_2$), si rifiuta H_0 per realizzazioni basse (elevate) di W.

• **Esempio 8.1.2.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. Il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2$ contro $H_1 : \lambda_1 \neq \lambda_2$ può essere verificato mediante il comando `wilcox.test` che fornisce l'implementazione del test di Mann-Whitney:

```
> wilcox.test(Diameter ~ Line, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  Diameter by Line
W = 32.5, p-value = 0.1986
alternative hypothesis: true location shift is not equal to 0
```

Il test di Mann-Whitney produce una significatività simile a quella del test t di Student e si può dunque accettare l'ipotesi di omogeneità delle mediane. □

In un approccio "distribution-free", si consideri due campioni scambiabili da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che Y_{c_1} ha funzione di ripartizione $F(y - \lambda_1)$ e Y_{c_2} ha funzione di ripartizione $F(y - \lambda_2)$, mentre λ_1 e λ_2 rappresentano le rispettive mediane. Se si considera il sistema di ipotesi $H_0 : \lambda_1 = \lambda_2$ contro $H_1 : \lambda_1 \neq \lambda_2$ (o una ipotesi direzionale), condizionatamente alla realizzazione del campione misto, sotto ipotesi di base ogni partizione del campione misto in due gruppi di numerosità n_1 e n_2 è ugualmente probabile. In questo caso, un test di permutazione è basato sulle $\binom{n_1+n_2}{n_1}$ (ugualmente probabili) permutazioni di livelli del fattore assegnati al campione misto. Se $S_1, \ldots, S_n$ rappresenta un vettore scambiabile di variabili casuali ognuna con supporto $\{-1, 1\}$, tali che $P(S_i = -1) = P(S_i = 1) = 1/2$ e $\sum_{i=1}^{n} S_i = n_1 - n_2$, il test di permutazione dei segni è basato sulla statistica test

$$T = \sum_{i=1}^{n} x_i S_i \, .$$

La distribuzione della statistica test T è usualmente approssimata mediante il metodo Monte Carlo. Nel caso dell'ipotesi alternativa bilaterale $H_1 : \lambda_1 \neq \lambda_2$, si rifiuta H_0 per realizzazioni basse o elevate di T. Nel caso dell'ipotesi alternativa direzionale $H_1 : \lambda_1 > \lambda_2$ ($H_1 : \lambda_1 < \lambda_2$), si rifiuta H_0 per realizzazioni basse (elevate) di T.

• **Esempio 8.1.3.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. Richiamando la libreria `exactRankTests`, il comando `wilcox.test` fornisce l'implementazione del test di permutazione (le osservazioni sono state moltiplicate per cento al fine di ottenere valori interi per ottenere una elaborazione più rapida):

```
> perm.test(round(100 * Diameter.1), round(100 * Diameter.2),
+    paired = F, alternative = "two.sided")

        2-sample Permutation Test

data:  round(100 * Diameter.1) and round(100 * Diameter.2)
T = 1194, p-value = 0.2105
alternative hypothesis: true mu is not equal to 0
```

Il test di permutazione produce una significatività simile a quella del test di Mann-Whitney e si può di nuovo accettare l'ipotesi di omogeneità delle mediane. □

In un approccio "distribution-free", si consideri due campioni casuali da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che Y_{c_1} ha funzione di ripartizione $F(y-\mu_1)$ e Y_{c_2} ha funzione di ripartizione $F(y-\mu_2)$, dove μ_1 e μ_2 rappresentano le rispettive medie. Condizionatamente alla realizzazione del campione misto, la distribuzione bootstrap della differenza delle medie campionarie indicata con T_{BOOT} può essere ottenuta considerando tutti i campioni con ripetizione di ordine n dal campione misto che vengono successivamente ripartiti in due campioni di numerosità n_1 e n_2. Se si considera il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ (o una ipotesi direzionale), il test bootstrap e il relativo intervallo di confidenza bootstrap può essere basato sulla statistica test T_{BOOT}. Anche in questo caso, la distribuzione della statistica test T_{BOOT} è usualmente approssimata mediante il metodo Monte Carlo.

• **Esempio 8.1.4.** Si considera di nuovo i dati relativi alle sfere di acciaio. La significatività osservata del test bootstrap può essere ottenuta mediante i seguenti comandi:

```
> Boot.meandif <- numeric(10000)
> Boot.sample <- numeric(length(Diameter))
> for (i in 1:10000) {Boot.sample <- sample(Diameter, replace = T);
+    Boot.diameter1 <- Boot.sample[c(1:length(Diameter.1))];
+    Boot.diameter2 <-
+    Boot.sample[c((length(Diameter.1) + 1):length(Diameter))];
+    Boot.meandif[i] <- mean(Boot.diameter1) - mean(Boot.diameter2)}
> 2 * length(Boot.meandif[Boot.meandif <
+    mean(Diameter.1) - mean(Diameter.2)]) / 10000
[1] 0.191
```

Il test bootstrap produce a sua volta una significatività simile a quella del test di Mann-Whitney e si può accettare l'ipotesi di omogeneità delle medie. □

8.2. I test funzionali

In un approccio "distribution-free", si consideri due campioni casuali indipendenti, di numerosità n_1 e n_2 con $n = n_1 + n_2$, da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore e tali che Y_{c_1} ha funzione di ripartizione F_{c_1} e Y_{c_2} ha funzione di ripartizione F_{c_2}. Si assuma che $\widehat{F}_{c_1}$ e $\widehat{F}_{c_2}$ siano rispettivamente le funzioni di ripartizione empiriche relative alle osservazioni provenienti da Y_{c_1} e Y_{c_2}. Se si considera il sistema di ipotesi $H_0 : F_{c_1}(y) = F_{c_2}(y)$ per ogni y, contro $H_1 : F_{c_1}(y) \neq F_{c_2}(y)$ per qualche y, il test di Kolmogorov-Smirnov è basato sulla statistica test

$$D = \sup_y |\widehat{F}_{c_1}(y) - \widehat{F}_{c_2}(y)| .$$

Si può dimostrare che la distribuzione di D non dipende da F_{c_1} e F_{c_2} sotto ipotesi di base e quindi il test è "distribution-free". La distribuzione della statistica test sotto ipotesi di base è nota e può essere tabulata. Evidentemente, se le due funzioni di ripartizione empiriche si discostano molto fra loro, allora si hanno valori elevati di D che conseguentemente portano a respingere l'ipotesi di base in favore dell'ipotesi alternativa. Per un calcolo pratico, se $Y_{(1)}, \ldots, Y_{(n)}$ è la statistica ordinata relativa al campione misto, la statistica D può essere espressa come

$$D = \max_{1 \leq i \leq n} |\widehat{F}_{c_1}(Y_{(i)}) - \widehat{F}_{c_2}(Y_{(i)})| .$$

• **Esempio 8.2.1.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. I grafici delle due funzioni di ripartizione empiriche sono ottenute mediante i seguenti comandi:

```
> plot(ecdf(Diameter.1), do.points = F, verticals = T,
+    xlim = c(0.5, 2.0), lty = 1,
+    xlab = "Ball diameter (micron)", ylab = "Probability",
+    main = "Empirical distribution function")
> plot(ecdf(Diameter.2), do.points = F, verticals = T,
+    lty = 3, add = T)
> legend(1.5, 0.3, c("Line 1", "Line 2"), lty = c(1, 3))
```

I precedenti comandi forniscono il grafico della Figura 8.2.1.

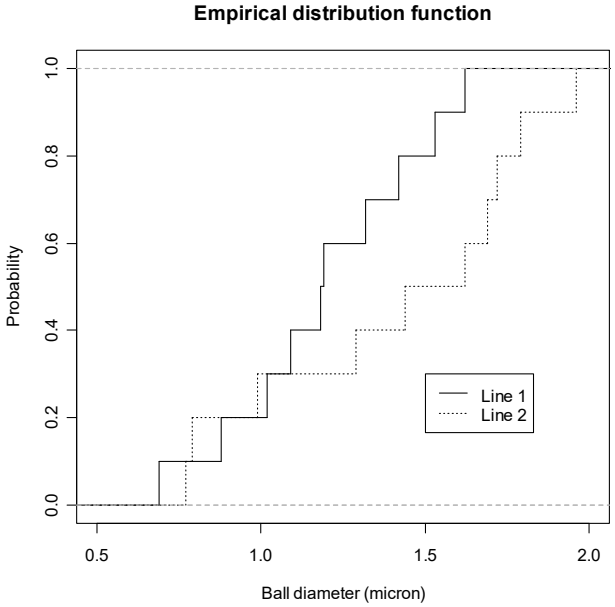


Figura 8.2.1.

Il sistema di ipotesi $H_0 : F_{c_1}(y) = F_{c_2}(y)$ per ogni y, contro $H_1 : F_{c_1}(y) \neq F_{c_2}(y)$ per qualche y, può essere verificato mediante il seguente comando:

```
> ks.test(Diameter.1, Diameter.2)

        Two-sample Kolmogorov-Smirnov test

data:  Diameter.1 and Diameter.2
D = 0.4, p-value = 0.4005
alternative hypothesis: two-sided
```

Il test di Kolmogorov-Smirnov produce porta dunque ad accettare l'omogeneità delle distribuzioni. Si osservi che la significatività del test è più alta rispetto a quella che è stata ottenuta nel caso dei test per l'omogeneità della tendenza centrale. Questo fatto è dovuto alla generalità del test di Kolmogorov-Smirnov rispetto alla specificità dei test visti nella precedente Sezione. □

8.3. L'analisi della varianza

In un tipico approccio classico, si consideri r campioni casuali indipendenti, ciascuno di numerosità n_j e tali che $\sum_{j=1}^{r} n_j = n$, da una variabile casuale Y a r livelli differenti $c_1, \ldots, c_r$ di un fattore e tali che $Y_{c_j} \sim N(\mu_j, \sigma^2)$. Siano $\overline{Y}_{c_j}$ e $S^2_{c_j}$ la media campionaria e la varianza campionaria delle osservazioni provenienti da $\overline{Y}_{c_j}$, mentre siano

$$S_b^2 = \frac{1}{n} \sum_{j=1}^{r} n_j (\overline{Y}_{c_j} - \overline{Y})^2$$

e

$$S_w^2 = \frac{1}{n} \sum_{j=1}^{r} n_j S^2_{c_j}$$

la cosiddetta varianza "between" (ovvero fra i gruppi) e la cosiddetta varianza "within" (ovvero all'interno dei gruppi). Si può verificare che gli stimatori di massima verosimiglianza di $\mu_1, \ldots, \mu_r$ e σ^2 sono dati rispettivamente da $\overline{Y}_{c_1}, \ldots, \overline{Y}_{c_r}$ e S_w^2. Se si considera il sistema di ipotesi $H_0 : \mu_1 = \ldots = \mu_r$ contro $H_1 : \mu_j \neq \mu_l$ per qualche (j, l), il test del rapporto delle verosimiglianze fornisce il test F di Fisher per l'analisi della varianza basato sulla statistica test

$$F = \frac{(n-m) S_b^2}{(m-1) S_w^2},$$

che sotto ipotesi di base si distribuisce come $F \sim F_{r-1, n-r}$. L'ipotesi di base viene rifiutata per realizzazioni elevate di F.

Quando l'ipotesi di base viene rifiutata, si desidera conoscere da quale coppia di medie dipende il rifiuto. Non è opportuno effettuare singoli test per la verifica dell'omogeneità di coppie di medie, in quanto le statistiche test sono dipendenti e quindi la significatività globale non può essere calcolata a partire da quella dei singoli test. In questo caso, la procedura di Tukey genera un insieme di intervalli di confidenza simultanei per le $r(r-1)/2$ possibili differenze $(\mu_j - \mu_l)$. Questa procedura è "distribution-free" per grandi campioni. Data la dualità esistente fra intervallo di confidenza e test statistico, analizzando gli intervalli di confidenza simultanei che non contengono il valore zero, si può risalire alle coppie di medie che hanno causato il rifiuto dell'ipotesi di base nell'analisi della varianza.

• **Esempio 8.3.1.** In un famoso esperimento di Michelson e Morley sono stati fatti 5 esperimenti di 20 prove ognuno per determinare la velocità della luce (Fonte: Weekes, A.J., 1986, *A Genstat Primer*, Arnold, London). Queste misurazioni riportano solo le ultime 3 cifre (senza decimali) della velocità della luce (in km/sec). Si noti che la moderna misurazione risulta 299,792.458 (km/sec). I dati sono contenuti nel file `light.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\light.txt", header = T)
> attach(d)
```

I dati originali vengono ripartiti nei cinque campioni mediante i comandi:

```
> Speed.1 <- split(Speed, Trial)[[1]]
> Speed.2 <- split(Speed, Trial)[[2]]
> Speed.3 <- split(Speed, Trial)[[3]]
> Speed.4 <- split(Speed, Trial)[[4]]
> Speed.5 <- split(Speed, Trial)[[5]]
```

Il diagramma a scatola e baffi condizionato viene ottenuto mediante il comando:

```
> boxplot(Speed ~ Trial, main = "Box-and-whiskers plot",
+     xlab = "Experiment Number")
```

Il precedente comando fornisce il grafico della Figura 8.3.1.

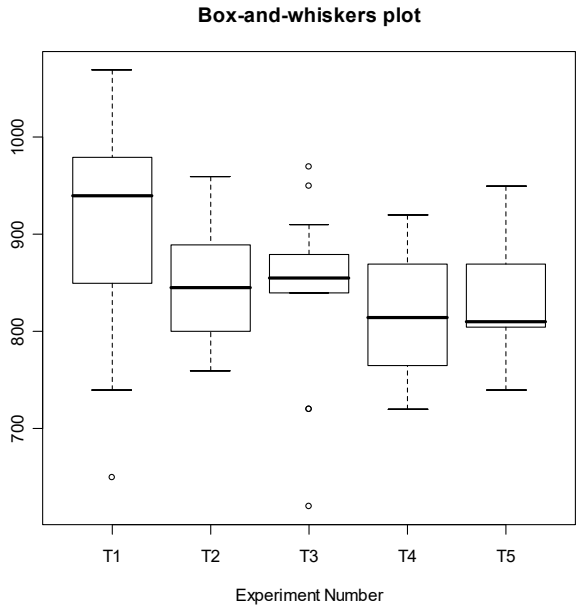


Figura 8.3.1.

Le stime di nucleo delle funzioni di densità vengono ottenute mediante i comandi:

```
> library(sm)
> par(mfrow = c(3, 2))
> sm.density(Speed.1, hnorm(Speed.1), yht = 0.008,
+     xlim = c(580, 1120), xlab = "Ligth speed (Trial 1)")
> title(main = "Kernel density estimation")
> sm.density(Speed.2, hnorm(Speed.2), yht = 0.008,
     xlim = c(580, 1120), xlab = "Ligth speed (Trial 2)")
> title(main = "Kernel density estimation")
> sm.density(Speed.3, hnorm(Speed.3), yht = 0.008,
+     xlim = c(580, 1120), xlab = "Ligth speed (Trial 3)")
> title(main = "Kernel density estimation")
> sm.density(Speed.4, hnorm(Speed.4), yht = 0.008,
+     xlim = c(580, 1120), xlab = "Ligth speed (Trial 4)")
> title(main = "Kernel density estimation")
> sm.density(Speed.5, hnorm(Speed.5), yht = 0.008,
+     xlim = c(580, 1120), xlab = "Ligth speed (Trial 5)")
> title(main = "Kernel density estimation")
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 8.3.2.

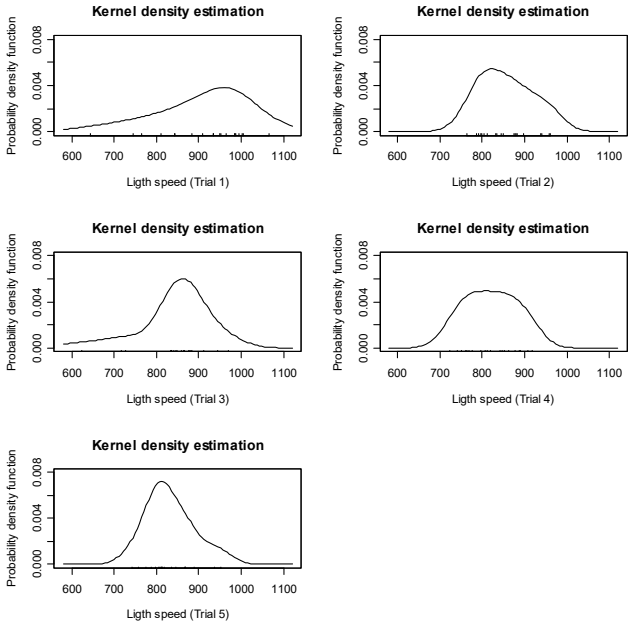


Figura 8.3.2.

L'analisi della varianza viene implementata mediante il comando:

```
> summary(aov(Speed ~ Trial))
            Df Sum Sq Mean Sq F value   Pr(>F)
Trial        4  94514   23629  4.2878 0.003114 **
Residuals   95 523510    5511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La procedura di Tukey può essere implementata mediante il seguente comando:

```
> TukeyHSD(aov(Speed ~ Trial), "Trial")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Speed ~ Trial)

$Trial
       diff        lwr        upr     p adj
T2-T1 -53.0 -118.28006  12.280058 0.1679880
T3-T1 -64.0 -129.28006   1.280058 0.0574625
T4-T1 -88.5 -153.78006 -23.219942 0.0025733
T5-T1 -77.5 -142.78006 -12.219942 0.0115793
T3-T2 -11.0  -76.28006  54.280058 0.9899661
T4-T2 -35.5 -100.78006  29.780058 0.5571665
T5-T2 -24.5  -89.78006  40.780058 0.8343360
T4-T3 -24.5  -89.78006  40.780058 0.8343360
T5-T3 -13.5  -78.78006  51.780058 0.9784065
T5-T4  11.0  -54.28006  76.280058 0.9899661
```

L'analisi della varianza porta a rifiutare l'ipotesi di omogeneità delle medie. Inoltre, la procedura di Tukey evidenzia che il rifiuto risulta principalmente da $\mu_1 > \mu_4$ e $\mu_1 > \mu_5$. □

In un approccio "distribution-free", si consideri r campioni casuali indipendenti, ciascuno di numerosità n_j e tali che $\sum_{j=1}^{r} n_j = n$, da una variabile casuale Y a r livelli differenti $c_1, \ldots, c_r$ di un fattore, tali che Y_{c_j} ha funzione di ripartizione $F(y - \lambda_j)$ e λ_j rappresenta la rispettiva mediana. Si assuma che R_{c_j} sia la somma dei ranghi assegnati alle osservazioni provenienti da Y_{c_j} nel campione misto. Se si considera il sistema di ipotesi $H_0 : \lambda_1 = \ldots = \lambda_r$ contro $H_1 : \lambda_j \neq \lambda_l$ per qualche (j, l), il test di Kruskal-Wallis è basato sulla statistica test

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{r} n_j (R_{c_j}/n_j - (n+1)/2)^2 \, .$$

La distribuzione di H sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa. Inoltre, per grandi campioni H converge in distribuzione ad una variabile casuale con distribuzione χ^2_{r-1}. Evidentemente, valori elevati della realizzazione di F portano al rifiuto dell'ipotesi di base.

• **Esempio 8.3.2.** Si considerano di nuovo i dati all'esperimento sulla velocità della luce dell'Esempio 8.3.1. Il test di Kruskal-Wallis viene implementato mediante il seguente comando:

```
> kruskal.test(Speed ~ Trial)

        Kruskal-Wallis rank sum test

data:  Speed by Trial
Kruskal-Wallis chi-squared = 15.0221, df = 4, p-value = 0.004656
```

Il test di Kruskal-Wallis produce a sua volta una significatività simile a quella ottenuta nell'analisi della varianza e si rifiuta quindi l'ipotesi di omogeneità delle medie. □

8.4. L'inferenza per l'associazione

In un tipico approccio classico, si consideri un campione casuale $(X_1, Y_1), \ldots, (X_n, Y_n)$ da una variabile casuale Normale bivariata (X, Y). Per una proprietà di caratterizzazione della Normale bivariata, esclusivamente per questa distribuzione si ha che la nullità del coefficiente di correlazione, ovvero

$$\rho_{xy} = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \, ,$$

implica l'indipendenza delle componenti marginali X e Y. Dunque la verifica dell'indipendenza si riduce alla verifica dell'ipotesi $H_0 : \rho_{xy} = 0$ contro $H_1 : \rho_{xy} \neq 0$. Il test del rapporto delle verosimiglianze fornisce il test basato sul rapporto di correlazione campionario $R_{xy} = S_{xy}/(S_x S_y)$, ovvero

$$T = \sqrt{n-2} \, \frac{R_{xy}}{\sqrt{1 - R^2_{xy}}} \, ,$$

che sotto ipotesi di base si distribuisce come $T \sim t_{n-2}$. Si noti inoltre che $F = T^2 \sim F_{1,n-2}$. Evidentemente, valori elevati della realizzazione di F portano al rifiuto dell'ipotesi di base.

• **Esempio 8.4.1.** Si sono considerate le misure (in metri) fatte nel lancio del peso e del giavellotto dalle 25 atlete partecipanti alla gara di eptathlon femminile alle Olimpiadi del 1988 (Fonte: Lunn,

A.D. e McNeil, D.R., 1991, *Computer-interactive Data Analysis*, Wiley, New York). I dati sono contenuti nel file `heptathlon.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\heptathlon.txt", header = T)
> attach(d)
```

Il diagramma di dispersione con la stima di nucleo della funzione di densità bivariata viene ottenuto mediante i comandi:

```
> library(sm)
> plot(Shot, Javelin, xlab = "Shot (meters)",
+    ylab = "Javelin (meters)")
> sm.density(d[, c(1, 2)], hcv(d[, c(1, 2)]), display = "slice",
+    props = c(75, 50, 25), add = T)
> title(main = "Kernel density estimation
+    ('CV' h1 = 1.11, h2 = 2.64)")
```

I precedenti comandi forniscono il grafico della Figura 8.4.1.

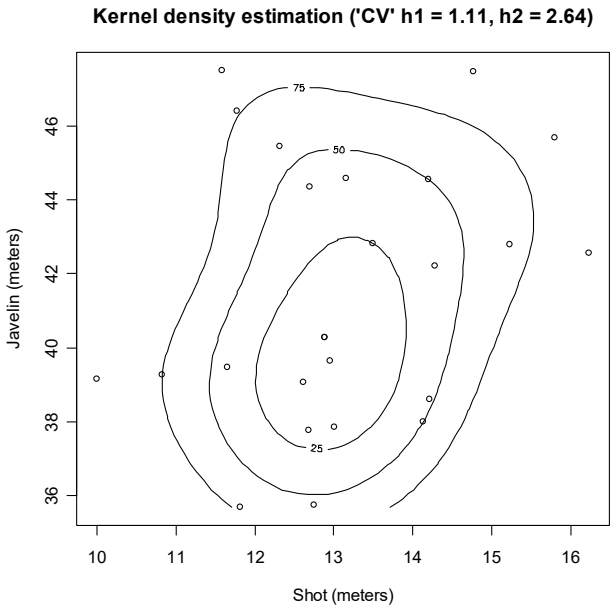


Figura 8.4.1.

Il sistema di ipotesi $H_0 : \rho_{xy} = 0$ contro $H_1 : \rho_{xy} \neq 0$ può essere verificato mediante il comando `cor.test` che fornisce l'implementazione del test per l'indipendenza:

```
> cor.test(Shot, Javelin, method = "pearson")

        Pearson's product-moment correlation

data:  Shot and Javelin
t = 1.3394, df = 23, p-value = 0.1935
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1411436  0.6003148
sample estimates:
      cor
0.2689888
```

Sulla base della significatività osservata si accetta dunque l'ipotesi di indipendenza fra i risultati nel lancio del peso e del giavellotto. □

In un ambito generale non è possibile verificare l'ipotesi di indipendenza basandosi su ipotesi relative ad un singolo parametro che descrive la dipendenza fra le componenti della coppia di variabili casuali. In questo caso, ci si deve limitare a verificare la presenza o l'assenza di associazione, ovvero dell'esistenza di una relazione di dipendenza diretta o inversa tra le variabili. Si consideri un campione casuale $(X_1, Y_1), \ldots, (X_n, Y_n)$ da una variabile casuale bivariata (X, Y). Il coefficiente di correlazione campionario di Spearman è il coefficiente di correlazione campionario calcolato sui ranghi relativi a $X_1, \ldots, X_n$ e sui ranghi relativi a $Y_1, \ldots, Y_n$. Il coefficiente di correlazione di Spearman può essere ottenuto semplicemente ordinando rispetto alle realizzazioni di $Y_1, \ldots, Y_n$ e successivamente assegnando i ranghi $R_1, \ldots, R_n$ alle realizzazioni di $X_1, \ldots, X_n$. Il coefficiente di correlazione di Spearman risulta dunque

$$\rho_S = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} iR_i - \frac{3(n+1)}{n-1} .$$

La statistica ρ_S è basata sui ranghi e quindi è "distribution-free". La statistica ρ_S gode ovviamente di tutte le proprietà di un coefficiente di correlazione campionario. Dunque, valori di ρ_S intorno allo zero denotano mancanza di associazione, mentre valori vicino ad 1 o -1 denotano presenza di associazione monotona diretta e inversa. La statistica test ρ_S può essere dunque adottata per la verifica dell'ipotesi di associazione. In questo caso sotto ipotesi di base viene verificata la mancanza di associazione. La distribuzione di ρ_S sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa.

• **Esempio 8.4.2.** Si considera di nuovo i dati dell'eptathlon dell'Esempio 8.4.1. Il comando `cor.test` fornisce l'implementazione del test di Spearman:

```
> cor.test(Shot, Javelin, method = "spearman")

        Spearman's rank correlation rho

data:  Shot and Javelin
S = 2062.793, p-value = 0.3217
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2066179
```

Sulla base della significatività osservata si accetta dunque l'ipotesi di mancanza di associazione fra i risultati nel lancio del peso e del giavellotto. □

Si consideri di nuovo un campione casuale $(X_1, Y_1), \ldots, (X_n, Y_n)$ da una variabile casuale bivariata (X, Y). Il cosiddetto coefficiente di correlazione di Kendall è dato dalla percentuale di coppie campionarie concordanti (ovvero coppie campionarie con lo stesso segno), ovvero

$$\tau = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{segn}(X_j - X_i)\text{segn}(Y_j - Y_i) ,$$

dove $\text{segn}(x) = 2\mathbf{1}_{]0,\infty[}(x) - 1$. La statistica τ gode delle proprietà di un indice di dipendenza. Evidentemente, valori intorno a 0 denotano mancanza di associazione, mentre valori vicino ad 1 o -1 denotano presenza di associazione monotona diretta e inversa. La statistica test τ può essere

dunque adottata per la verifica dell'ipotesi di associazione. La distribuzione di τ sotto ipotesi di base può essere tabulata anche se non può essere espressa in forma chiusa.

• **Esempio 8.4.3.** Si considera di nuovo i dati dell'eptathlon dell'Esempio 8.4.1. Il comando `cor.test` fornisce l'implementazione del test di Kendall:

```
> cor.test(Shot, Javelin, method = "kendall")

        Kendall's rank correlation tau

data:  Shot and Javelin
z = 1.0515, p-value = 0.293
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.1505017
```

Il test basato sul coefficiente di correlazione di Kendall fornisce una significatività simile a quella del test basato sul coefficiente di correlazione campionario di Spearman e si accetta dunque l'ipotesi di mancanza di associazione fra i risultati nel lancio del peso e del giavellotto. □

Se si considera un campionamento casuale da una variabile qualitativa bivariata (X, Y), allora le osservazioni campionarie forniscono le frequenze osservate congiunte n_{jl} delle realizzazioni distinte (c_j, d_l) della variabile. Si assuma inoltre che la funzione di probabilità congiunta di (X, Y) sia data da $p_{X,Y}(c_j, d_l) = \pi_{jl}$, dove $j = 1, \ldots, r, l = 1, \ldots, s$, mentre la distribuzione marginale di probabilità di X sia data da $p_X(c_j) = \pi_{j+}$, dove $j = 1, \ldots, r$ con $\pi_{j+} = \sum_{l=1}^{s} \pi_{jl}$, e quella di Y sia data da $p_Y(d_l) = \pi_{+l}$, $l = 1, \ldots, s$ con $\pi_{+l} = \sum_{j=1}^{r} \pi_{jl}$. Si è interessati a verificare l'indipendenza di X e Y, ovvero il sistema di ipotesi $H_0 : \pi_{jl} = \pi_{j+}\pi_{+l}$ per ogni (j, l), contro $H_1 : \pi_{jl} \neq \pi_{j+}\pi_{+l}$ per qualche (j, l). Per verificare questo sistema di ipotesi si adotta la statistica test Chi-quadrato per l'indipendenza data da

$$\chi^2 = \sum_{j=1}^{r}\sum_{l=1}^{s} \frac{(n_{jl} - n_{j+}n_{+l}/n)^2}{n_{j+}n_{+l}/n} .$$

Le quantità $n_{j+}n_{+l}/n$ sono le frequenze attese stimate sotto ipotesi d'indipendenza. La distribuzione per grandi campioni di χ^2 non dipende dai valori π_{jl} e quindi il test è "distribution-free" per grandi campioni. Sotto ipotesi di base, per $n \to \infty$ la statistica test χ^2 converge in distribuzione ad una variabile casuale con distribuzione $\chi^2_{(r-1)(s-1)}$. L'approssimazione è valida per campioni finiti se $n > 30$ e se tutte le frequenze attese stimate sono maggiori di uno. Se le frequenze osservate si discostano molto dalle frequenze attese stimate, si ottengono determinazioni elevate della statistica test che portano a respingere l'ipotesi di base.

• **Esempio 8.4.4.** Durante uno studio della malattia di Hodgkin sono stati considerati 538 malati, ognuno dei quali è stato classificato per tipologie istologiche (indicate con le sigle LP, NS, MC e LD) e per la risposta al trattamento dopo tre mesi di cura (Fonte: Dunsmore, I.R. e Daly, F., 1987, *M345 Statistical Methods, Unit 9: Categorical Data*, The Open University, Milton Keynes). I dati sono contenuti nel file `hodgkin.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\hodgkin.txt", header = T)
> attach(d)
```

La tabella a doppia entrata viene ottenuta mediante il comando:

```
> xtabs(Count ~ Type + Response)
    Response
Type None Part Pos
  LD   44   10  18
  LP   12   18  74
  MC   58   54 154
  NS   12   16  68
```

Il diagramma a nastri condizionato viene ottenuto mediante il comando:

```
> library(lattice)
> barchart(xtabs(Count ~ Type + Response),
+    ylab = "Hystological type",
+    auto.key = list(title = "Response", cex = 0.8))
```

I precedenti comandi forniscono il grafico della Figura 8.4.2.

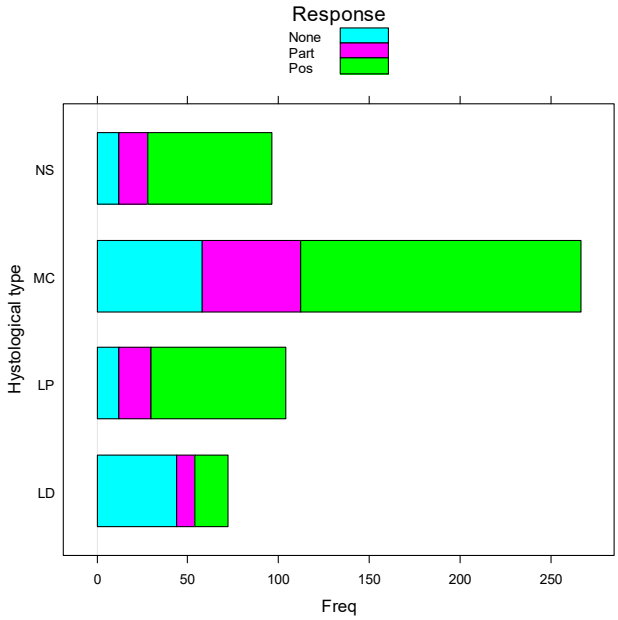


Figura 8.4.2.

Il comando `chisq.test` fornisce l'implementazione del test χ^2 per l'indipendenza:

```
> chisq.test(xtabs(Count ~ Type + Response))

        Pearson's Chi-squared test

data:  xtabs(Count ~ Type + Response)
X-squared = 75.8901, df = 6, p-value = 2.517e-14
```

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di indipendenza delle tipologie istologiche dagli esiti del trattamento. □

Considerando un test di permutazione, condizionatamente alla realizzazione del campione, sotto ipotesi di base (ovvero di indipendenza delle variabili) ogni partizione delle osservazioni relative alla seconda variabile in gruppi di numerosità $n_{1+}, \ldots, n_{r+}$ è ugualmente probabile. Dunque, si può

costruire un test di permutazione basato sulle $\binom{n}{n_{1+}\dots n_{r+}}$ (ugualmente probabili) permutazioni di livelli della prima variabile rispetto alla seconda variabile. Il test esatto di Fisher è basato sulla statistica test χ^2 di permutazione che si ottiene calcolando la statistica χ^2 sulle $\binom{n}{n_{1+}\dots n_{r+}}$ tabelle a doppia entrata, ognuna relativa ad una delle possibili permutazioni di gruppi. La distribuzione della statistica test sotto ipotesi di base può essere anche in questo caso approssimata mediante un metodo Monte Carlo. Determinazioni elevate della statistica test portano a respingere l'ipotesi di base.

• **Esempio 8.4.5.** In un famoso studio sono state considerate coppie di gemelli ognuna delle quali è stata classificata per tipologia (ovvero se la coppia è costituita da gemelli omozigoti o eterozigoti) e per propensione alla criminalità (ovvero se entrambi i gemelli sono stati detenuti in prigione) (Fonte: Fisher, R. A., 1970, *Statistical Methods for Research Workers*, quattordicesima edizione, Oliver & Boyd, London). I dati sono contenuti nel file `twins.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\twins.txt", header = T)
> attach(d)
```

La tabella a doppia entrata viene ottenuta mediante il comando:

```
> xtabs(Count ~ Type + Conviction)
             Conviction
Type          No Yes
  Dizygotic   15   2
  Monozygotic  3  10
```

Il diagramma a nastri condizionato viene ottenuto mediante i comandi:

```
> library(lattice)
> barchart(xtabs(Count ~ Type + Conviction), ylab = "Type",
+          auto.key = list(title = "Conviction", cex = 0.8))
```

I precedenti comandi forniscono il grafico della Figura 8.4.3.

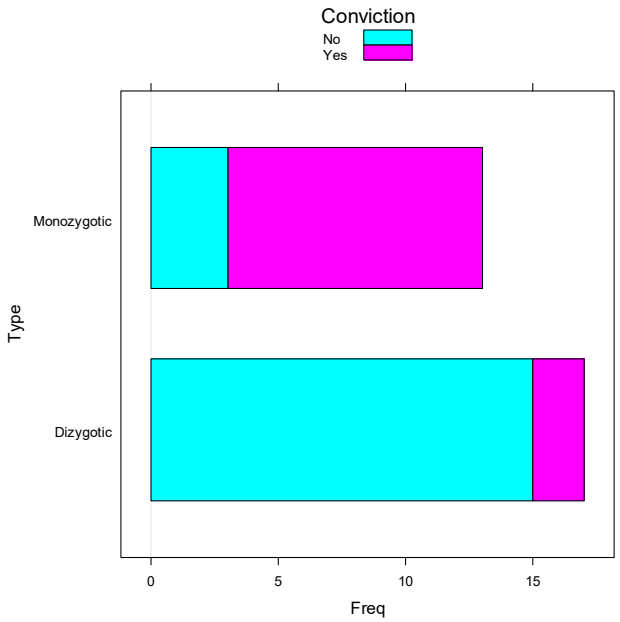


Figura 8.4.3.

Il comando `fisher.test` fornisce l'implementazione del test esatto di Fisher:

```
> fisher.test(xtabs(Count ~ Type + Conviction))

        Fisher's Exact Test for Count Data

data:  xtabs(Count ~ Type + Conviction)
p-value = 0.0005367
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   2.753438 300.682787
sample estimates:
odds ratio
  21.30533
```

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di indipendenza fra la tipologia dei gemelli e la propensione alla criminalità. □

• **Esempio 8.4.6.** Per evidenziare la relazione fra il test χ^2 e il test esatto di Fisher, si assuma per semplicità che $r = s = 2$. A priori dal campionamento i dati possono essere equivalentemente organizzati nella tabella a doppia entrata di Tavola 8.4.1.

Tavola 8.4.1

	d_1	d_2	
c_1	N_{11}	N_{12}	N_{1+}
c_2	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	N

Se si assume che N_{11}, N_{12}, N_{21} e N_{22} sono variabili casuali indipendenti, rispettivamente con distribuzione di Poisson $Po(\lambda_{11})$, $Po(\lambda_{12})$, $Po(\lambda_{21})$ e $Po(\lambda_{22})$, la funzione di probabilità congiunta del vettore casuale $(N_{11}, N_{12}, N_{21}, N_{22})$ è data da

$$p_{N_{11},N_{12},N_{21},N_{22}}(n_{11}, n_{12}, n_{21}, n_{22}) = \prod_{j,l=1}^{2} e^{-\lambda_{jl}} \frac{\lambda_{jl}^{n_{jl}}}{n_{jl}!} \mathbf{1}_{\mathbb{N}}(n_{jl}) \, .$$

Per le proprietà della distribuzione di Poisson, N ha distribuzione di Poisson $Po(\lambda)$ dove $\lambda = \sum_{j,l=1}^{2} \lambda_{jl}$. Inoltre, si ha la seguente funzione di probabilità condizionata

$$p_{N_{11},N_{12},N_{21},N_{22}}(n_{11}, n_{12}, n_{21}, n_{22} \mid N = n) = \binom{n}{n_{11}\, n_{12}\, n_{21}\, n_{22}} \prod_{j,l=1}^{2} \pi_{jl}^{n_{jl}} \mathbf{1}_S(n_{11}, n_{12}, n_{21}, n_{22}) \, ,$$

dove si è posto $\pi_{jl} = \lambda_{jl}/\lambda$ e

$$S = \{(n_{11}, n_{12}, n_{21}, n_{22}) : n_{jl} \in \{0, 1, \ldots, n\}, \sum_{j,l=1}^{2} n_{jl} = n\} \, .$$

Dal momento che $\pi_{jl} \geq 0$ e $\sum_{j,l=1}^{2} \pi_{jl} = 1$, è evidente che questa distribuzione condizionata è Multinomiale $M_4(n, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))$. Inoltre, la distribuzione di N_{1+} condizionata all'evento $\{N = n\}$ è Binomiale $Bi(n, \pi_{1+})$ con $\pi_{1+} = \pi_{11} + \pi_{12}$, mentre la distribuzione di N_{+1} condizionata all'evento $\{N = n\}$ è Binomiale $Bi(n, \pi_{+1})$ con $\pi_{+1} = \pi_{11} + \pi_{21}$. Quindi, si hanno le seguenti funzioni di probabilità condizionate

$$p_{N_{1+}|N=n}(n_{1+}) = \binom{n}{n_{1+}} \pi_{1+}^{n_{1+}} (1-\pi_{1+})^{n-n_{2+}} \mathbf{1}_{\{0,1,\ldots,n\}}(n_{1+}) = \binom{n}{n_{1+}} \pi_{1+}^{n_{1+}} \pi_{2+}^{n_{2+}} \mathbf{1}_{\{0,1,\ldots,n\}}(n_{1+})$$

e

$$p_{N_{+1}|N=n}(n_{+1}) = \binom{n}{n_{+1}} \pi_{+1}^{n_{+1}} (1-\pi_{+1})^{n-n_{+2}} \mathbf{1}_{\{0,1,\ldots,n\}}(n_{+1}) = \binom{n}{n_{+1}} \pi_{+1}^{n_{+1}} \pi_{+2}^{n_{+2}} \mathbf{1}_{\{0,1,\ldots,n\}}(n_{+1}) \, .$$

Supponendo l'indipendenza delle distribuzioni marginali, ovvero assumendo che $\pi_{jl} = \pi_{j+}\pi_{+l}$ per $j, l = 1, 2$, queste due distribuzioni condizionate sono indipendenti. In questo caso, si ha la funzione di probabilità condizionata

$$\begin{aligned} p_{N_{11}}(n_{11} \mid N = n, N_{1+} = n_{1+}, N_{+1} = n_{+1}) &= \frac{p_{N_{11},N_{12},N_{21},N_{22}}(n_{11}, n_{12}, n_{21}, n_{22} \mid N = n)}{p_{N_{1+}|N=n}(n_{1+}) p_{N_{+1}|N=n}(n_{+1})} \\ &= \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} \mathbf{1}_A(x) \, , \end{aligned}$$

dove $A = \{\max(0, n_{+1} - n + n_{1+}), \ldots, \min(n_{+1}, n_{1+})\}$. La distribuzione di N_{11} condizionata all'evento

$$\{N = n, N_{1+} = n_{1+}, N_{+1} = n_{+1}\} = \{N_{1+} = n_{1+}, N_{2+} = n_{2+}, N_{+1} = n_{+1}, N_{+2} = n_{+2}\}$$

è dunque data dalla legge Ipergeometrica $I(n_{+1}, n_{1+}, n)$. Quindi, partendo dallo schema probabilistico di Poisson nella tabella a doppia entrata e condizionando sul totale $N = n$, si ottiene lo schema probabilistico Multinomiale. Il test χ^2 per l'indipendenza viene sviluppato nello schema probabilistico Multinomiale. Partendo invece dallo schema probabilistico Multinomiale nella tabella a doppia entrata e condizionando sui totali $N_{1+} = n_{1+}, N_{2+} = n_{2+}, N_{+1} = n_{+1}, N_{+2} = n_{+2}$, si ottiene lo schema probabilistico Ipergeometrico. Il test esatto di Fisher viene sviluppato nello schema probabilistico Ipergeometrico. Il test possiede questa denominazione in quanto può essere effettivamente basato sulla statistica test N_{11} che, avendo distribuzione Ipergeometrica, ha funzione di probabilità esplicita. Al contrario, la statistica test χ^2 è basata su una trasformazione del vettore casuale $(N_{11}, N_{12}, N_{21}, N_{22})$ con una funzione di probabilità che non può essere espressa in forma semplice, ma con una distribuzione nota per grandi campioni, ovvero la distribuzione Chi-quadrato. Si deve sottolineare che i due approcci derivano da paradigmi inferenziali differenti. □

8.5. Riferimenti bibliografici

Bretz, F., Hothorn, T. e Westfall, P. (2011) *Multiple Comparisons Using R*, Chapman & Hall/CRC Press, Boca Raton.

Chernick, M.R. (2008) *Bootstrap Methods*, seconda edizione, Wiley, New York.

Davison, A.C. e Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.

Dickhaus, T. (2014) *Simultaneous Statistical Inference*, Springer, New York.

Efron, B. e Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, London.

Fisher, R.A. (1970) *Statistical Methods for Research Workers*, quattordicesima edizione, Oliver & Boyd, London.

Hájek, J. (1969) *Nonparametric Statistics*, Holden Day, San Francisco.

Hájek, J. e Šidák, Z. (1967) *Theory of Rank Tests*, Academic Press, New York.

Hettmansperger, T.P. e McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*, seconda edizione, Chapman & Hall/CRC Press, Boca Raton.

Hollander, M., Wolfe, D.A. e Chicken, E. (2014) *Nonparametric Statistical Methods*, terza edizione, Wiley, New York.

Lehmann, E.L. e Romano, J.P. (2022) *Testing Statistical Hypothesis*, quarta edizione, Springer, Switzerland.

Shao, J. e Tu, D. (1995) *The Jackknife and Bootstrap*, Springer, New York.

Pagina intenzionalmente vuota

Capitolo 9

La regressione

9.1. La regressione lineare

Nella sua versione più semplice, ovvero quando si dispone di un regressore e di una variabile di risposta, il modello statistico di regressione analizza la struttura di dipendenza fra le due variabili. Il modello di regressione lineare assume ovviamente un legame lineare fra le variabili. In questo caso, si consideri il modello

$$Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i \, ,$$

dove $\mathcal{E}_1, \ldots, \mathcal{E}_n$ sono variabili casuali indipendenti (detti errori) tali che $\mathrm{E}[\mathcal{E}_i] = 0$ e $\mathrm{Var}[\mathcal{E}_i] = \sigma^2$. La formulazione alternativa del modello di regressione lineare è data dalle relazioni

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$$

e

$$\mathrm{Var}[Y_i] = \sigma^2 \, .$$

In questo modello di regressione esistono quindi tre parametri.

Le stime di β_0 e β_1 ottenute con il metodo dei minimi quadrati (che coincidono con le stime di massima verosimiglianza assumendo la Normalità degli errori) risultano

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

e

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} \, .$$

Indicando per semplicità stima e stimatore con lo stesso simbolo (un abuso di notazione che è usualmente adottato, anche se può produrre confusione), gli stimatori $\widehat{\beta}_0$ e $\widehat{\beta}_1$ sono corretti con varianze

$$\mathrm{Var}[\widehat{\beta}_0] = \frac{\sigma^2}{n s_x^2} \, (s_x^2 + \overline{x}^2)$$

e

$$\mathrm{Var}[\widehat{\beta}_1] = \frac{\sigma^2}{n s_x^2} \, .$$

Questi risultati saranno discussi in generale nel caso della regressione lineare multipla presentata nel prossimo Capitolo. I valori stimati risultano inoltre

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \, ,$$

mentre le quantità $(y_i - \widehat{y}_i)$ sono dette residui. Il parametro σ^2 può essere stimato in modo corretto mediante la varianza correstta dei residui, ovvero

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \, .$$

Le stime di $\mathrm{Var}[\widehat{\beta}_0]$ e $\mathrm{Var}[\widehat{\beta}_1]$ possono essere infine ottenute sostituendo $\widehat{\sigma}^2$ al posto di σ^2.

La variabilità totale delle osservazioni relative alla variabile di risposta può essere scomposta come

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \, ,$$

dove la prima componente rappresenta la variabilità degli errori e la seconda componente la variabilità dovuta al modello lineare. Di conseguenza, la quantità

$$r^2_{xy} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

rappresenta la percentuale di variabilità spiegata dalla relazione lineare rispetto alla variabilità totale ed è il cosiddetto coefficiente di determinazione lineare. Ovviamente, risulta $r^2_{xy} \in [0, 1]$ e il valore uno indica la presenza di linearità perfetta. Si dimostra facilmente che $r^2_{xy} = s^2_{xy}/(s^2_x s^2_y)$, ovvero il coefficiente di determinazione lineare è il quadrato del coefficiente di correlazione lineare. Si noti che può esistere una relazione (non lineare) perfetta fra variabile di risposta e regressore per cui $r^2_{xy} = 0$.

• **Esempio 9.1.1.** Si dispone delle osservazioni di distanze lineari e stradali fra località a Sheffield (in km) (Fonte: Gilchrist, W., 1984, *Statistical Modelling*, Wiley, New York, p.5). La variabile di risposta è la distanza stradale, mentre il regressore è la distanza lineare. I dati sono contenuti nel file `roaddistance.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\roaddistance.txt",
+    header = T)
> attach(d)
```

Le stime dei parametri del modello di regressione lineare vengono ottenute mediante il seguente comando:

```
> lm(Road ~ Linear)

Call:
lm(formula = Road ~ Linear)

Coefficients:
(Intercept)       Linear
     0.3791       1.2694
```

Il diagramma di dispersione con retta di regressione stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Linear, Road, xlab = "Linear distance (km)",
+    ylab = "Road distance (km)", main = "Scatter plot")
> abline(lm(Road ~ Linear))
```

I precedenti comandi forniscono il grafico della Figura 9.1.1.

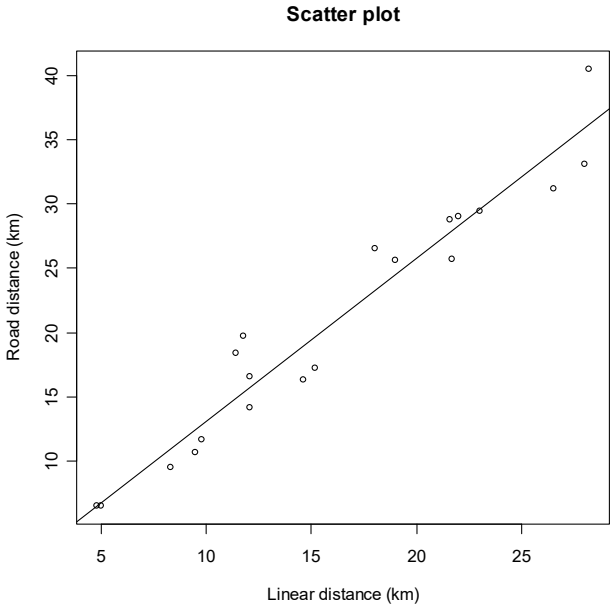


Figura 9.1.1.

Il diagramma di dispersione evidenzia l'adeguatezza del modello lineare. □

Oltre alla costruzione dell'indice di determinazione, i residui permettono di considerare le diagnostiche grafiche per esaminare la validità della relazione lineare. In particolare, il diagramma di Anscombe, che fornisce il diagramma cartesiano dei residui rispetto ai valori stimati, dovrebbe presentare una disposizione casuale dei punti se effettivamente il modello lineare è adeguato. Questo grafico viene anche riportato con le radici dei valori assoluti dei residui standardizzati. Il diagramma quantile-quantile, che fornisce il diagramma cartesiano dei residui standardizzati e ordinati rispetto ai quantili della distribuzione normale standardizzata, dovrebbe avere una disposizione dei punti lungo la bisettrice se l'ipotesi di normalità per gli errori è valida. Il diagramma con le distanze di Cook consente di verificare l'impatto della rimozione di ogni singola osservazione sulle stime dei parametri e quindi l'influenza delle singole osservazioni.

• **Esempio 9.1.2.** Si considerano di nuovo i dati relativi alle distanze stradali dell'Esempio 9.1.1. I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook vengono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Road ~ Linear), which = c(1:4), add.smooth = F)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 9.1.2. I punti che hanno maggiore influenza possono essere evidenziati sul diagramma di dispersione mediante i seguenti comandi:

```
> plot(Linear, Road, xlab = "Linear distance (km)",
+         ylab = "Road distance (km)", main = "Scatter plot")
> abline(lm(Road ~ Linear))
> text(x = Linear[9] + 0.3, y = Road[9], labels = "9", adj=0)
> text(x = Linear[11] + 0.3, y = Road[11], labels = "11", adj=0)
> text(x = Linear[17] - 0.5, y = Road[17] - 1, labels = "17", adj=0)
```

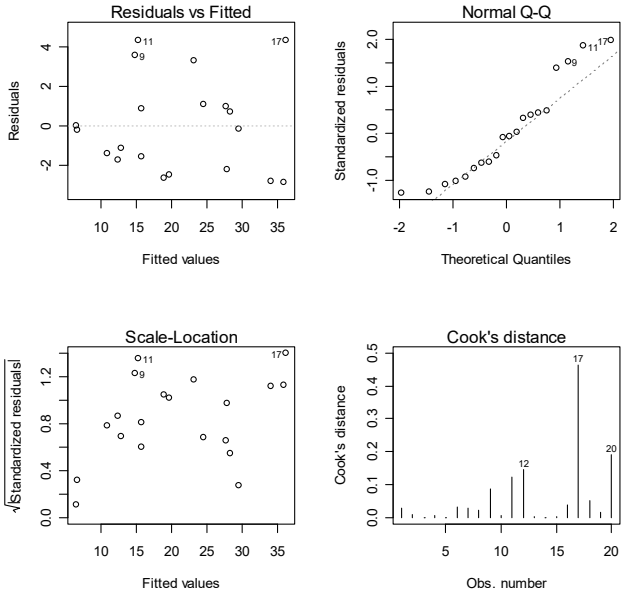


Figura 9.1.2.

I precedenti comandi forniscono il grafico della Figura 9.1.3.

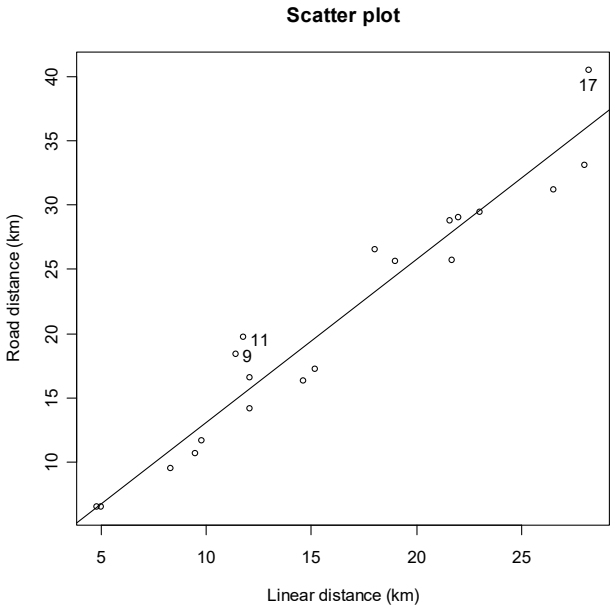


Figura 9.1.3.

Infine, il grafico della stima della funzione di regressione basata sulla regressione lineare locale è ottenuto mediante i seguenti comandi:

```
> library(sm)
> plot(Linear, Road, xlab = "Linear distance (km)",
+          ylab = "Road distance (km)", main = "Scatter plot")
> sm.regression(Linear, Road, method = "df", add = T)
```

I precedenti comandi forniscono il grafico della Figura 9.1.4.

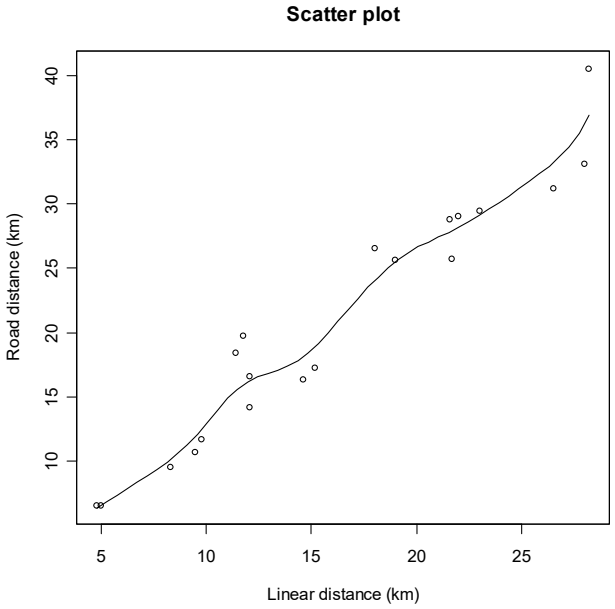


Figura 9.1.4.

Il precedente grafico conferma l'adeguatezza del modello lineare. □

Per quanto riguarda la verifica delle ipotesi, supponendo la normalità degli errori e di conseguenza la normalità di $Y_1, \ldots, Y_n$, la validità del modello lineare viene verificata considerando il sistema di ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Questo sistema di ipotesi viene verificato attraverso la statistica test fornita dal rapporto delle verosimiglianze

$$F = (n-2)\,\frac{R_{xy}^2}{1-R_{xy}^2}\,,$$

dove R_{xy} è il rapporto di correlazione campionario. La statistica test F si distribuisce come una variabile casuale di Snedecor $F_{1,n-2}$. Una simile verifica di ipotesi può essere condotta anche su β_0.

• **Esempio 9.1.3.** Si considerano di nuovo i dati relativi alle distanze stradali dell'Esempio 9.1.1. L'analisi relativa alla verifica delle ipotesi viene implementata mediante il seguente comando:

```
> summary(lm(Road ~ Linear))
Call:
lm(formula = Road ~ Linear)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8231 -1.8604 -0.2011  1.0263  4.3416

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.37908    1.34401   0.282    0.781
Linear       1.26943    0.07617  16.665 2.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.436 on 18 degrees of freedom
Multiple R-Squared: 0.9391,     Adjusted R-squared: 0.9358
F-statistic: 277.7 on 1 and 18 DF,  p-value: 2.187e-12
```

L'analisi relativa alla verifica delle ipotesi con il modello senza intercetta viene implementata mediante il seguente comando:

```
> summary(lm(Road ~ -1 + Linear))

Call:
lm(formula = Road ~ -1 + Linear)

Residuals:
   Min     1Q Median     3Q    Max
-2.994 -1.728 -0.097  1.029  4.489

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Linear  1.28907    0.03012    42.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.376 on 19 degrees of freedom
Multiple R-Squared: 0.9897,     Adjusted R-squared: 0.9892
F-statistic:  1832 on 1 and 19 DF,  p-value: < 2.2e-16
```

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di base, ovvero si accetta l'esistenza di un forte legame lineare fra le variabili. Inoltre, il modello lineare senza intercetta si adegua quasi perfettamente ai dati. □

Anche se la relazione fra variabile di risposta e regressore non è lineare, ci si può spesso ricondurre alla linearità mediante trasformazioni monotone opportune. Se G e H sono trasformazioni monotone, il modello di regressione lineare per le osservazioni trasformate risulta

$$G(Y_i) = \beta_0 + \beta_1 H(x_i) + \mathcal{E}_i \, .$$

In questo caso, la relazione fra la variabile di risposta e il regressore è data quindi dalla funzione $y = G^{-1}(\beta_0 + \beta_1 H(x))$.

• **Esempio 9.1.4.** Si dispone delle osservazioni della percentuale di ricordi nel tempo (in minuti) relativi ad un esperimento psicometrico su un soggetto (Fonte: Mosteller, F., Rourke, R.E.K. e Thomas, G.B. (1970) *Probability with Statistical Applications*, seconda edizione, Addison-Wesley, Reading, p.383). La variabile di risposta è la percentuale di ricordi, mentre il regressore è il tempo. I dati sono contenuti nel file `memory.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\memory.txt", header = T)
> attach(d)
```

Da una analisi del diagramma di dispersione contenuto nella Figura 9.1.5 è evidente che una trasformazione logaritmica sulle osservazioni relative al regressore può essere opportuna. L'analisi relativa al modello di regressione lineare

$$Y_i = \beta_0 + \beta_1 \log x_i + \mathcal{E}_i$$

viene ottenuta mediante il seguente comando:

```
> summary(lm(Memory ~ log(Time)))
```

```
Call:
lm(formula = Memory ~ log(Time))

Residuals:
      Min        1Q    Median        3Q       Max
-0.036077 -0.015330 -0.006415  0.017967  0.037799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
log(Time)   -0.079227   0.002416  -32.80 2.53e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02339 on 11 degrees of freedom
Multiple R-Squared: 0.9899,     Adjusted R-squared: 0.989
F-statistic:  1076 on 1 and 11 DF,  p-value: 2.525e-12
```

Il diagramma di dispersione con la relativa funzione di regressione si può ottenere mediante i seguenti comandi:

```
> plot(Time, Memory, xlab = "Time (minutes)",
+    ylab = "Memory retention", main = "Scatter plot")
> lines(seq(1, 11000, 1), predict(lm(Memory ~ log(Time)),
+    data.frame(Time = seq(1, 11000, 1))))
```

I precedenti comandi forniscono il grafico della Figura 9.1.5.

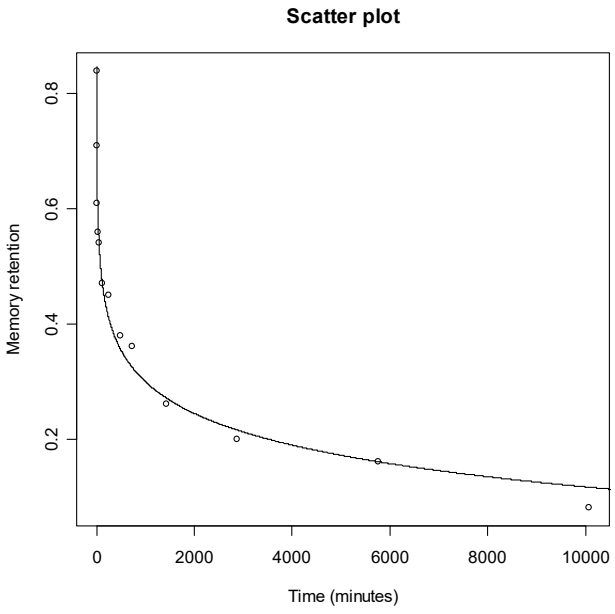


Figura 9.1.5.

Sulla base della significatività osservata si deve dunque rifiutare l'ipotesi di base, ovvero si accetta l'esistenza di un forte legame logaritmico fra le variabili. □

La struttura inferenziale basata su due campioni casuali indipendenti (di numerosità n_1 e n_2 con $n = n_1 + n_2$) da una variabile casuale Y a due livelli differenti c_1 e c_2 di un fattore, tali che $\mathrm{E}[Y_{c_1}] = \mu_1$, $\mathrm{E}[Y_{c_2}] = \mu_2$ e $\mathrm{Var}[Y_{c_1}] = \mathrm{Var}[Y_{c_2}] = \sigma^2$, può essere riportata ad un modello di

regressione lineare. In effetti, se $Y_1, \ldots, Y_n$ rappresentano le osservazioni relative al campione misto, allora si può scrivere il modello lineare con

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$$

e

$$\mathrm{Var}[Y_i] = \sigma^2 \,,$$

dove $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$, mentre $x_1, \ldots, x_n$ sono i valori assunti da un regressore binario che vale uno se l'osservazione è relativa al secondo livello del fattore e zero altrimenti. L'ipotesi di omogeneità delle medie, supponendo la normalità di $Y_1, \ldots, Y_n$, viene verificata considerando il sistema di ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Di conseguenza, le tecniche viste in precedenza possono essere applicate anche in questo caso.

• **Esempio 9.1.5.** Si considera di nuovo i dati relativi alle sfere di acciaio dell'Esempio 8.1.1. Il sistema di ipotesi $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_1 \neq \mu_2$ può essere verificato mediante il seguente comando:

```
> summary(lm(Diameter ~ Line))

Call:
lm(formula = Diameter ~ Line)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6360 -0.2090  0.0150  0.2915  0.5540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1940     0.1156  10.327 5.43e-09 ***
LineL2        0.2120     0.1635   1.297    0.211
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3656 on 18 degrees of freedom
Multiple R-Squared: 0.08541,    Adjusted R-squared: 0.0346
F-statistic: 1.681 on 1 and 18 DF,  p-value: 0.2112
```

L'omogeneità delle medie viene ovviamente accettata con lo stesso livello di significatività osservata dell'Esempio 8.1.1. Tuttavia, in questo caso si ha una maggiore informazione sulle due medie. In effetti, la stima del coefficiente β_1 è positiva, anche se non significativa, e dunque la media nel secondo gruppo tende ad essere più elevata rispetto a quella nel primo gruppo. □

9.2. I modelli lineari generalizzati

Una generalizzazione del modello di regressione lineare può essere ottenuta assumendo una distribuzione non Normale per $Y_1, \ldots, Y_n$ e considerando una opportuna funzione legame g tale che

$$g(\mathrm{E}[Y_i]) = \beta_0 + \beta_1 x_i \,.$$

La classe dei modelli lineari generalizzati è basata appunto su questa relazione. Il modello di regressione lineare è un caso particolare di questa classe quando le variabili di risposta hanno una

distribuzione Normale e g è la funzione identità. Il modello lineare generalizzato consente di gestire in modo più appropriato variabili di risposta che sono asimmetriche, discrete o binarie.

Per ogni distribuzione che si assume per la variabile di risposta esiste una funzione legame canonica, ovvero una parametrizzazione naturale del modello. In generale, non è possibile ottenere in forma chiusa le stime di massima verosimiglianza $\widehat{\beta}_0$ e $\widehat{\beta}_1$ di β_0 e β_1 e delle relative varianze. Quindi si deve ricorrere a procedure numeriche per ottenere queste stime. L'analisi per la validità del modello viene verificata di nuovo mediante il sistema di ipotesi $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$. Questo sistema di ipotesi è basato sulla devianza, che è una statistica test basata sul metodo del rapporto delle verosimiglianze. La statistica test si distribuisce per grandi campioni come una variabile casuale con distribuzione Chi-quadrato con opportuni gradi di libertà. La devianza si può scomporre in devianza sotto ipotesi di base e devianza residua. Una valore elevato della devianza residua rispetto ai rispettivi gradi di libertà è indice di super dispersione, ovvero della presenza di una variabilità più accentuata delle stime rispetto a quella prevista dal modello lineare generalizzato. Questo fenomeno può essere gestito mediante l'uso di tecniche basate sulla quasi-verosimiglianza per la stima della variabilità degli stimatori di β_0 e β_1.

Al fine di spiegare la scelta della funzione legame, si consideri una variabile casuale Y che appartiene alla famiglia Esponenziale, ovvero con funzione di densità o probabilità data da

$$f_Y(y;\theta,\phi) = e^{\frac{1}{a(\phi)}(\theta y - b(\theta)) + c(y,\phi)} \, \mathbf{1}_S(y) \, ,$$

dove θ e ϕ sono parametri, mentre a, b e c sono funzioni specifiche e $S \subseteq \mathbb{R}$ è un insieme opportuno che non dipende dai parametri. Inoltre, si assume che b sia una funzione differenziabile al secondo ordine. Se ϕ è un parametro noto, allora la famiglia è detta Esponenziale con parametro canonico θ. Per un dato valore y la log-verosimiglianza risulta

$$l(\theta,\phi;y) = \log f_Y(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \, .$$

Dunque, si ha

$$\frac{dl}{d\theta} = \frac{y - b'(\theta)}{a(\phi)} \, , \, \frac{d^2 l}{d\theta^2} = -\frac{b''(\theta)}{a(\phi)} \, ,$$

da cui per le proprietà della funzione punteggio risulta

$$\mathrm{E}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{\mathrm{E}[Y] - b'(\theta)}{a(\phi)} = 0$$

e

$$\mathrm{Var}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{\mathrm{Var}[Y]}{a(\phi)^2} = \frac{b''(\theta)}{a(\phi)} \, ,$$

ovvero

$$\mathrm{E}[Y] = b'(\theta)$$

e

$$\mathrm{Var}[Y] = a(\phi) b''(\theta) \, .$$

Si assuma che $\mu = \mu(\theta) = \mathrm{E}[Y]$ e $V(\mu) = \mathrm{Var}[Y]$. La funzione inversa $\theta = \mu^{-1}$ fornisce la funzione legame canonica, mentre la funzione $V(\mu)$ è detta funzione di varianza.

• **Esempio 9.2.1.** La distribuzione Normale $N(\mu, \sigma^2)$ fa parte della famiglia Esponenziale, in quanto si può scrivere

$$f_Y(y;\theta,\phi) = e^{\frac{1}{\sigma^2}(y\mu-\mu^2/2)-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)} ,$$

per cui $\theta = \mu$ e $\phi = \sigma^2$, mentre

$$a(\phi) = \phi ,$$

$$b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$$

e

$$c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) .$$

Dunque, si ha

$$\mathrm{E}[Y] = b'(\theta) = \theta .$$

In questo caso, essendo $\mathrm{E}[Y] = \mu$, si ha $\mu = \theta$ e $V(\mu) = \sigma^2$, ovvero la funzione legame canonica è la funzione identità, mentre la funzione di varianza è costante. □

• **Esempio 9.2.2.** La distribuzione di Poisson $P(\lambda)$ fa parte della famiglia Esponenziale, in quanto si può scrivere

$$f_Y(y;\theta,\phi) = e^{y\log(\lambda)-\lambda-\log(y!)}\mathbf{1}_{\mathbb{N}}(y) ,$$

per cui $\theta = \log(\lambda)$, ovvero $\lambda = e^\theta$, e $\phi = 1$, mentre

$$a(\phi) = 1 ,$$

$$b(\theta) = \lambda = e^\theta$$

e

$$c(y,\phi) = -\log(y!) .$$

Dunque, si ha

$$\mathrm{E}[Y] = b'(\theta) = e^\theta .$$

In questo caso, essendo $\mu = \mathrm{E}[Y] = \lambda$, si ha $\theta = \log(\lambda)$ e $V(\mu) = \mu$, ovvero la funzione legame canonica è la funzione logaritmo, mentre la funzione di varianza è la funzione identità. □

• **Esempio 9.2.3.** La distribuzione Binomiale $B(1,p)$ fa parte della famiglia Esponenziale, in quanto si può scrivere

$$f_Y(y;\theta,\phi) = e^{y\log\left(\frac{p}{1-p}\right)+\log(1-p)}\mathbf{1}_{\{0,1\}}(y) ,$$

per cui $\theta = \log\left(\frac{p}{1-p}\right)$, ovvero $p = \frac{e^\theta}{1+e^\theta}$, e $\phi = 1$, mentre

$$a(\phi) = 1 ,$$

$$b(\theta) = -\log(1-p) = \log(1+e^\theta)$$

e

$$c(y, \phi) = 0 \, .$$

Dunque, si ha

$$\mathrm{E}[Y] = b'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}} \, .$$

In questo caso, essendo $\mu = \mathrm{E}[Y] = p$, si ha $\theta = \log\left(\frac{\mu}{1-\mu}\right)$ e $V(\mu) = \mu(1 - \mu)$, ovvero la funzione legame canonica è la funzione logit. □

Quando le variabili di risposta prendono valori sugli interi positivi è conveniente assumere che $Y_1, \ldots, Y_n$ siano variabili casuali di Poisson. Questo caso particolare del modello lineare generalizzato costituisce la regressione di Poisson. Tenendo presente l'Esempio 9.2.2, la funzione legame canonica è la funzione logaritmo, ovvero si ha

$$\log(\mathrm{E}[Y_i]) = \beta_0 + \beta_1 x_i \, .$$

• **Esempio 9.2.4.** Si dispone delle osservazioni relative alla lunghezza di pezzi di stoffa (in metri) e del relativo numero di difetti (Fonte: Bissell, A.F., 1972, A negative binomial model with varying element sizes, *Biometrika* **59**, 435-441). La variabile di risposta è il numero di difetti riscontrati, mentre il regressore è la lunghezza. I dati sono contenuti nel file `clothes.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\clothes.txt", header = T)
> attach(d)
```

L'analisi relativa al modello di regressione di Poisson viene ottenuta mediante il seguente comando:

```
> summary(glm(Defects ~ Length, poisson))

Call:
glm(formula = Defects ~ Length, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.74127  -1.13312  -0.03904   0.66179   3.07446

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.9717506  0.2124693   4.574 4.79e-06 ***
Length      0.0019297  0.0003063   6.300 2.97e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 103.714  on 31  degrees of freedom
Residual deviance:  61.758  on 30  degrees of freedom
AIC: 189.06

Number of Fisher Scoring iterations: 4
```

L'analisi della devianza residua indica la presenza di super dispersione e quindi è conveniente impiegare un metodo di quasi-verosimiglianza, ovvero:

```
> summary(glm(Defects ~ Length, quasipoisson))

Call:
glm(formula = Defects ~ Length, family = quasipoisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.74127  -1.13312  -0.03904   0.66179   3.07446

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9717506  0.3095033   3.140 0.003781 **
Length      0.0019297  0.0004462   4.325 0.000155 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.121965)

    Null deviance: 103.714  on 31  degrees of freedom
Residual deviance:  61.758  on 30  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

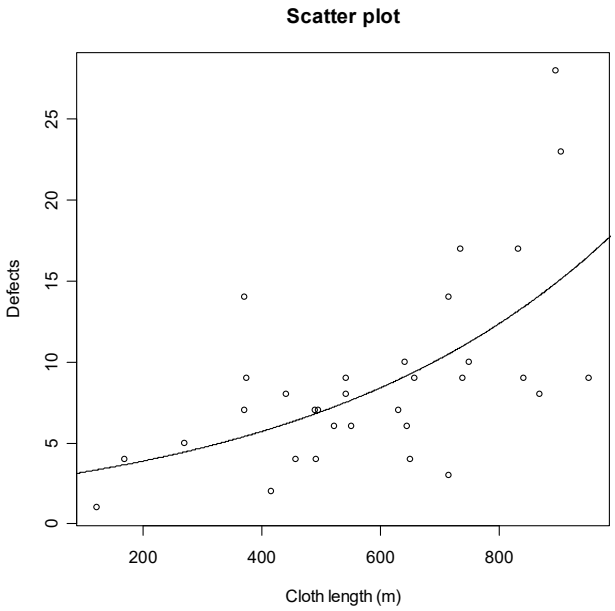


Figura 9.2.1.

La precedente elaborazione indica la validità del modello. Il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Length, Defects, xlab = "Cloth length (m)", ylab = "Defects",
+    main = "Scatter plot")
> lines(seq(0, 1000, 1),
+    exp(predict(glm(Defects ~ Length, quasipoisson),
+    data.frame(Length = seq(0, 1000, 1)))))
```

I precedenti comandi forniscono il grafico della Figura 9.2.1. □

Quando le variabili di risposta sono binarie è conveniente assumere che $Y_1, \ldots, Y_n$ siano variabli casuali di Bernoulli. Questo caso particolare del modello lineare generalizzato costituisce la regressione logistica. Dall'Esempio 9.2.3, la funzione legame canonica è la funzione logit, ovvero si ha

$$\log\left(\frac{\mathrm{E}[Y_i]}{1-\mathrm{E}[Y_i]}\right) = \beta_0 + \beta_1 x_i \,.$$

• **Esempio 9.2.5.** Si dispone delle osservazioni relative alla presenza di danneggiamento dei pannelli di protezione e delle temperature (in gradi Fahrenheit) per alcuni voli di shuttle (Fonte: Dalal, S.R., Fowlkes, E.B. e Hoadley, B. (1989) Risk analysis of the space shuttle: pre-*challenger* prediction of failure, *Journal of the American Statistical Association* **84**, 945-957). La variabile di risposta è la presenza di danneggiamento, mentre il regressore è la temperatura. I dati sono contenuti nel file `shuttle.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\shuttle.txt", header = T)
> attach(d)
```

L'analisi relativa al modello di regressione logistica viene ottenuta mediante il seguente comando:

```
> summary(glm(Failure ~ Temp, binomial))

Call:
glm(formula = Failure ~ Temp, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temp         -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

La precedente elaborazione indica una certa adeguatezza del modello. Il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Temp, Failure, xlab = "Temperature (°F)", ylab = "Failure",
+    main = "Scatter plot")
> lines(seq(50, 90, 0.1),
+    predict(glm(Failure ~ Temp, binomial),
+    data.frame(Temp = seq(50, 90, 0.1)), type = "response"))
```

I precedenti comandi forniscono il grafico della Figura 9.2.2. □

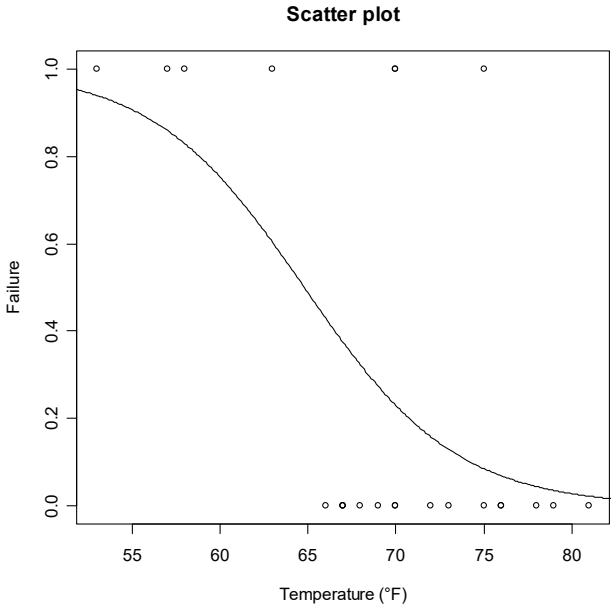


Figura 9.2.2.

La regressione logistica può essere applicata anche quando esistono solo $c_1, \ldots, c_r$ livelli distinti del regressore (che potrebbero essere anche i valori centrali di classi opportune) per i due livelli $d_1 = 0$ e $d_2 = 1$ della variabile di risposta e quindi le osservazioni possono essere organizzate in una tabella a doppia entrata con r righe e 2 colonne. In questo caso, la regressione logistica può essere applicata considerando le variabili di risposta $n_{12}/n_{1+}, \ldots, n_{r2}/n_{r+}$, ovvero le proporzioni della variabile di risposta per ogni livello del regressore, ottenendo il modello

$$\log\left(\frac{\mathrm{E}[n_{j2}/n_{j+}]}{1-\mathrm{E}[n_{j2}/n_{j+}]}\right) = \beta_0 + \beta_1 c_j \,.$$

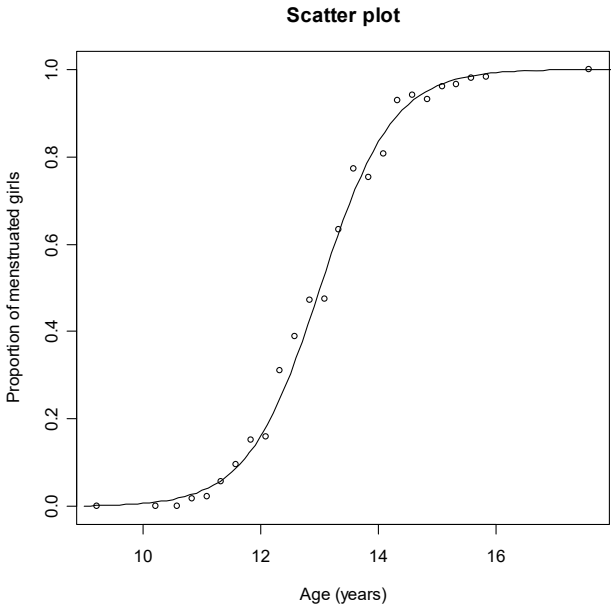


Figura 9.2.3.

● **Esempio 9.2.6.** Si dispone delle osservazioni relative alla percentuale di un gruppo di adolescenti polacche con menarca per vari livelli d'età (Fonte: Morgan, B.J.T. (1989) *Analysis of Quantal Response Data*, Chapman and Hall, London, p.7). La variabile di risposta è la percentuale di un gruppo di adolescenti polacche che hanno avuto il menarca, mentre il regressore è l'età. I dati sono contenuti nel file `menarche.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\menarche.txt", header = T)
> attach(d)
```

L'analisi relativa al modello di regressione logistica viene ottenuta mediante il seguente comando:

```
> Proportion <- cbind(Menarche, Total - Menarche)
> summary(glm(Proportion ~ Age, binomial))

Call:
glm(formula = Proportion ~ Age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0363  -0.9953  -0.4900   0.7780   1.3675

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.22639    0.77068  -27.54   <2e-16 ***
Age           1.63197    0.05895   27.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: 114.76

Number of Fisher Scoring iterations: 4
```

La precedente elaborazione indica la validità del modello. Il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> plot(Age, Menarche / Total, xlab = "Age (years)",
+    ylab = "Proportion of menstruated girls",
+    main = "Scatter plot")
> lines(seq(9, 18, 0.1),
+    predict(glm(Proportion ~ Age, binomial),
+    data.frame(Age = seq(9, 18, 0.1)), type = "response"))
```

I precedenti comandi forniscono il grafico della Figura 9.2.3. □

9.3. Riferimenti bibliografici

Agresti, A. (2013) *Categorical Data Analysis*, terza edizione, Wiley, New York.
Agresti, A. (2019) *An Introduction to Categorical Data Analysis*, terza edizione, Wiley, New York.

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*, Wiley, New York.
Atkinson, A. e Riani, M. (2000) *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York.
Belsley, D.A., Kuh, E. e Welsch, R.E. (1980) *Regression Diagnostics*, Wiley, New York.
Chatterjee, S. e Simonoff, J.S. (2013) *Handbook of Regression Analysis*, Wiley, New York.
Cook, R.D. e Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*, Wiley, New York.
Dunn, P.K. e Smyth, G.K. (2018) *Generalized Linear Models with Examples in R*, Springer, New York.
James, G., Witten, D., Hastie T. e Tibshirani, R. (2013) *An Introduction to Statistical Learning*, Springer, New York.
McCullagh, P. e Nelder, J.A. (1989) *Generalized Linear Models*, seconda edizione, Chapman and Hall, London.
Sheather, S.J. (2009) *A Modern Approach to Regression with R*, Springer, New York.
Simonoff, J.S. (2003) *Analyzing Categorical Data*, Springer, New York.
Weisberg, S. (2014) *Applied Linear Regression*, Wiley, New York.

Capitolo 10

La regressione multipla

10.1. La regressione lineare multipla

Il modello di regressione multipla analizza la struttura di dipendenza fra un insieme di regressori e una variabile di risposta. Nella versione più semplice, il modello di regressione multipla assume un legame lineare fra le variabili e questo modello include come caso particolare anche l'analisi della varianza. Quando la variabile di risposta dipende da p regressori si consideri dunque il modello

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \mathcal{E}_i \,,$$

dove $\mathcal{E}_1, \ldots, \mathcal{E}_n$ sono variabili casuali indipendenti tali che $\mathrm{E}[\mathcal{E}_i] = 0$ e $\mathrm{Var}[\mathcal{E}_i] = \sigma^2$. La formulazione alternativa del modello di regressione lineare multipla è data quindi dalle relazioni

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

e

$$\mathrm{Var}[Y_i] = \sigma^2 \,.$$

In questo modello esistono dunque $(p+2)$ parametri.

Sia $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ il vettore dei parametri di ordine $(p+1)$ e sia $\mathbf{X} = (x_{ij})$ la matrice di ordine $(n \times (p+1))$ delle osservazioni relative ai regressori, in cui la prima colonna è composta da unità (ovvero, si è aggiunto in pratica un ulteriore regressore che assume valori pari all'unità). Inoltre, sia $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ il vettore delle determinazioni del vettore casuale $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Infine, sia $\boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)^{\mathrm{T}}$ con vettore medio $\mathrm{E}[\boldsymbol{\mathcal{E}}] = \mathbf{0}$ e matrice di varianza-covarianza $\mathrm{Var}[\boldsymbol{\mathcal{E}}] = \sigma^2 \mathbf{I}_n$. In questo caso, il modello di regressione lineare multipla può essere espresso in notazione matriciale come

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \,,$$

o alternativamente mediante le due relazioni

$$\mathrm{E}[\boldsymbol{Y}] = \mathbf{X}\boldsymbol{\beta}$$

e

$$\mathrm{Var}[\boldsymbol{Y}] = \sigma^2 \mathbf{I}_n \,.$$

Al fine di stimare il vettore dei parametri $\boldsymbol{\beta}$, Il metodo dei minimi quadrati consiste nel minimizzare la somma degli scarti al quadrato dei valori osservati dai valori teorici della variabile di risposta, ovvero nel minimizzare la forma quadratica

$$\varphi(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} \,,$$

dove la matrice $\mathbf{V} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ è evidentemente definita positiva. A questo fine, si ha

$$\frac{\partial \varphi}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{V}\boldsymbol{\beta} .$$

Uguagliando questa quantità a $\mathbf{0}$, si ottiene l'equazione vettoriale

$$\mathbf{V}\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y} .$$

Se si suppone che il rango di $\mathbf{X}$ sia $p+1 < n$, ovvero che non vi siano relazioni lineari fra le colonne di $\mathbf{X}$ e che la numerosità campionaria sia maggiore del numero dei regressori (per evitare relazioni esatte fra i regressori), si può invertire la matrice $\mathbf{V}$, ottenendo in questo modo la stima di $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = \mathbf{V}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} .$$

La matrice hessiana

$$\frac{\partial^2 \varphi}{\partial \boldsymbol{\beta}^{\mathrm{T}} \partial \boldsymbol{\beta}} = 2\mathbf{V}$$

è definita positiva, per cui φ è una funzione convessa e quindi $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$ è effettivamente un punto di minimo. Con un abuso in notazione comune il letteratura, lo stimatore $\widehat{\boldsymbol{\beta}} = \mathbf{V}^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$ viene indicato con lo stesso simbolo della stima.

Si può verificare che lo stimatore $\widehat{\boldsymbol{\beta}}$ è corretto e la relativa matrice di varianza-covarianza è data da $\mathrm{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2\mathbf{V}^{-1}$. Se il vettore dei valori stimati è denotato con $\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, la somma dei quadrati dei residui è data dalla forma quadratica

$$S = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^{\mathrm{T}}(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) .$$

Dal momento che si ha $\mathrm{E}[S] = (n-p-1)\sigma^2$, uno stimatore corretto di σ^2 è dato da

$$\widehat{\sigma}^2 = \frac{S}{n-p-1} .$$

Indicando con s la realizzazione di S, la stima di $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ è dunque data da

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] = \frac{s}{n-p-1}\,\mathbf{V}^{-1} .$$

Anche per il modello di regressione multipla il coefficiente di determinazione viene definito come

$$r^2_{xy} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2} ,$$

dove $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_p x_{ip}$. L'interpretazione di r^2_{xy} è analoga a quella dello stesso indice considerato nel modello lineare con un singolo regressore. Infine, se si suppone che $\boldsymbol{\mathcal{E}}$ abbia distribuzione Normale $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, essendo $\boldsymbol{Y}$ e $\widehat{\boldsymbol{\beta}}$ trasformate lineari di $\boldsymbol{\mathcal{E}}$, allora $\boldsymbol{Y}$ ha distribuzione Normale $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, mentre $\widehat{\boldsymbol{\beta}}$ ha distribuzione Normale $N_{p+1}(\boldsymbol{\beta}, \sigma^2\mathbf{V}^{-1})$. Inoltre, si può verificare che S/σ^2 ha distribuzione Chi-quadrato χ^2_{n-p-1}.

In alternativa al metodo dei minimi quadrati, sotto le assunzioni fatte per la distribuzione del vettore casuale $\boldsymbol{Y}$, si può considerare il metodo della massima verosimiglianza. Posto per semplicità di notazione $\upsilon = \sigma^2$, si ottiene la seguente funzione di verosimiglianza

$$L(\boldsymbol{\beta}, \upsilon) = c\, e^{-\frac{1}{2\upsilon}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} ,$$

da cui si ha la seguente funzione di log-verosimiglianza

$$l(\boldsymbol{\beta}, \upsilon) = \log c - \frac{n}{2} \log \upsilon - \frac{1}{2\upsilon} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \, .$$

Derivando opportunamente la log-verosimiglianza si hanno le equazioni

$$\frac{\partial l}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = \frac{1}{\upsilon} \mathbf{X}^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

e

$$\frac{\partial l}{\partial \upsilon} = -\frac{n}{2\upsilon} + \frac{1}{2\upsilon^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \, .$$

Dalle precedenti relazioni si ottengono le stime di massima verosimiglianza date da

$$\widehat{\boldsymbol{\beta}} = \mathbf{V}^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y}$$

e

$$\widehat{\upsilon} = \frac{1}{n} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{s}{n} \, .$$

Dunque, lo stimatore di $\boldsymbol{\beta}$ coincide con quello ottenuto con il metodo dei minimi quadrati, mentre lo stimatore di σ^2 è distorto. Inoltre, la matrice hessiana è data da

$$\begin{pmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\beta}^{\mathrm{T}} \partial \boldsymbol{\beta}} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}^{\mathrm{T}} \partial \upsilon} \\ \frac{\partial^2 l}{\partial \boldsymbol{\beta}^{\mathrm{T}} \partial \upsilon} & \frac{\partial^2 l}{\partial \upsilon^2} \end{pmatrix} = -\frac{1}{\upsilon} \begin{pmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\upsilon} \end{pmatrix} ,$$

che è definita negativa e quindi l è una funzione concava sullo spazio parametrico. Dunque, le stime ottenute sono effettivamente dei punti di massimo. La matrice di varianza-covarianza per grandi campioni degli stimatori di massima verosimiglianza si ottiene mediante la matrice d'informazione di Fisher $\mathbf{I}(\boldsymbol{\beta}, \upsilon)$, che coincide con la precedente matrice hessiana. Dunque, la matrice di varianza-covarianza per grandi campioni risulta

$$\mathrm{Var}[\widehat{\boldsymbol{\beta}}, \widehat{\upsilon}] = -\mathbf{I}(\boldsymbol{\beta}, \upsilon)^{-1} = \upsilon \begin{pmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\upsilon}{n} \end{pmatrix} ,$$

che può essere stimata sostituendo $\widehat{\upsilon}$ al posto di υ. Per le proprietà del metodo della massima verosimiglianza gli stimatori $\widehat{\boldsymbol{\beta}}$ e $\widehat{\upsilon}$ sono asintoticamente corretti e con distribuzione Normale per grandi campioni. Le distribuzioni per piccoli campioni sono analoghe a quelle evidenziate in precedenza per il metodo dei minimi quadrati.

Per quanto riguarda la verifica delle ipotesi, supponendo la Normalità del vettore casuale $\boldsymbol{Y}$, la validità del modello lineare viene basata sul sistema di ipotesi $H_0 : \beta_1 = \ldots = \beta_p = 0$ contro $H_1 : \beta_j \neq 0$ per qualche j. Questo sistema di ipotesi viene verificato attraverso la statistica test fornita dal rapporto delle verosimiglianze

$$F = \frac{n-p}{p-1} \frac{R_{xy}^2}{1 - R_{xy}^2} \, ,$$

che ha distribuzione di Snedecor $F_{p-1,n-p}$. La verifica di ipotesi su uno specifico parametro β_j, ovvero $H_0 : \beta_j = 0$ contro $H_1 : \beta_j \neq 0$, viene effettuata con la statistica test del rapporto delle verosimiglianze

$$T = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{v_{jj}}},$$

dove v_{jj} è il j-esimo elemento sulla diagonale della matrice $\mathbf{V}^{-1}$, e che si distribuisce come $T \sim t_{n-p}$.

Quando si è verificato che alcuni parametri del modello sono nulli, si cerca di costruire un modello più semplice (ovvero con un numero minore di regressori) di quello iniziale. Se si eliminano alcuni parametri, il modello di regressione diventa più semplice, ma la sua adeguatezza ai dati diminuisce. In generale, quando si ipotizza un modello con k parametri $\theta_1, \ldots, \theta_k$, il criterio di Akaike è dato da

$$\text{AIC} = -2(\log L(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) - k)$$

dove $L(\theta_1, \ldots, \theta_k)$ è la funzione di verosimiglianza relativa al modello. Il criterio di Akaike è in effetti il massimo della log-verosimiglianza penalizzato dal numero di parametri presenti nel modello stesso (il tutto moltiplicato da una costante negativa), ovvero un indice finalizzato a valutare il compromesso fra adeguatezza e semplicità del modello. Quando si deve dunque scegliere fra più modelli, il modello che si preferisce è quindi quello che fornisce il valore minimo del criterio di Akaike.

• **Esempio 10.1.1.** Si dispone delle osservazioni del numero di specie di uccelli in "isole" di vegetazione nel nord delle Ande (Fonte: Vuilleumier, F., 1970, Insular biogeography in continental regions. I. The northern Andes of South America, *American Naturalist* **104**, 373-388). I regressori sono l'area dell'isola di vegetazione (in migliaia di km quadrati), la sua elevazione (in km), la sua distanza dall'equatore (in km) e la sua distanza dall'isola di vegetazione più vicina (in km). I dati sono contenuti nel file `paramo.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\paramo.txt", header = T)
> attach(d)
```

La matrice dei diagrammi di dispersione viene ottenuta mediante il seguente comando:

```
> pairs(d, main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 10.1.1.

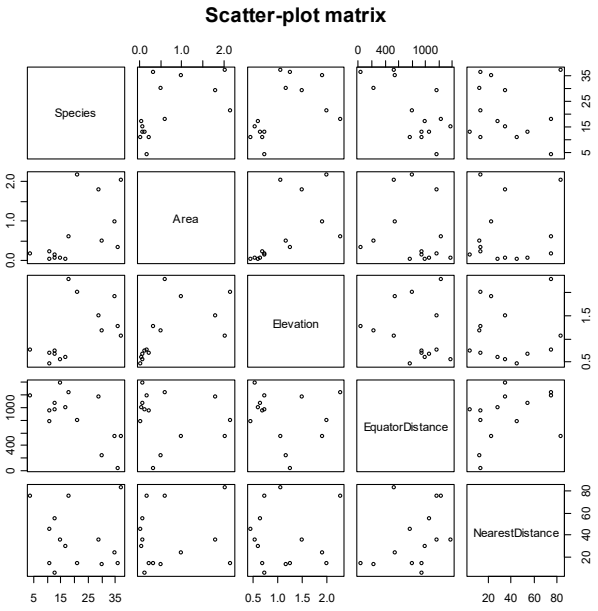


Figura 10.1.1.

L'analisi del modello lineare con tutti i regressori viene implementata mediante il seguente comando:

```
> summary(lm(Species ~ Area + Elevation +
+    EquatorDistance + NearestDistance))

Call:
lm(formula = Species ~ Area + Elevation +
EquatorDistance + NearestDistance)

Residuals:
      Min        1Q    Median        3Q       Max
-10.66596  -3.40900   0.08345   3.55920   8.23565

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     27.889386   6.181843   4.511  0.00146 **
Area             5.153864   3.098074   1.664  0.13056
Elevation        3.075136   4.000326   0.769  0.46175
EquatorDistance -0.017216   0.005243  -3.284  0.00947 **
NearestDistance  0.016591   0.077573   0.214  0.83541
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.705 on 9 degrees of freedom
Multiple R-Squared: 0.7301,     Adjusted R-squared: 0.6101
F-statistic: 6.085 on 4 and 9 DF,  p-value: 0.01182
```

La semplificazione automatica del modello mediante il criterio AIC viene implementata mediante il seguente comando:

```
> summary(step(lm(Species ~ Area + Elevation + EquatorDistance +
+    NearestDistance)))

Start:  AIC=57.09
Species ~ Area + Elevation + EquatorDistance + NearestDistance

                  Df Sum of Sq    RSS    AIC
- NearestDistance  1      2.06 406.65  55.16
- Elevation        1     26.57 431.15  55.98
<none>                       404.59  57.09
- Area             1    124.41 529.00  58.85
- EquatorDistance  1    484.71 889.30  66.12

Step:  AIC=55.16
Species ~ Area + Elevation + EquatorDistance

                  Df Sum of Sq    RSS    AIC
- Elevation        1     26.06 432.71  54.03
<none>                       406.65  55.16
- Area             1    133.51 540.15  57.14
- EquatorDistance  1    537.39 944.04  64.96

Step:  AIC=54.03
Species ~ Area + EquatorDistance
```

```
                  Df Sum of Sq    RSS    AIC
<none>                         432.71  54.03
- Area             1    342.64 775.35  60.20
- EquatorDistance  1    557.23 989.94  63.62

Call:
lm(formula = Species ~ Area + EquatorDistance)

Residuals:
    Min      1Q  Median      3Q     Max
-10.637  -4.396   0.899   4.084   7.273

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     30.797969   4.648155   6.626 3.73e-05 ***
Area             6.683038   2.264403   2.951  0.01318 *
EquatorDistance -0.017057   0.004532  -3.764  0.00313 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.272 on 11 degrees of freedom
Multiple R-Squared: 0.7113,     Adjusted R-squared: 0.6588
F-statistic: 13.55 on 2 and 11 DF,  p-value: 0.001077
```

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello ridotto vengono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Species ~ Area + EquatorDistance),
+    which = c(1:4), add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.2. □

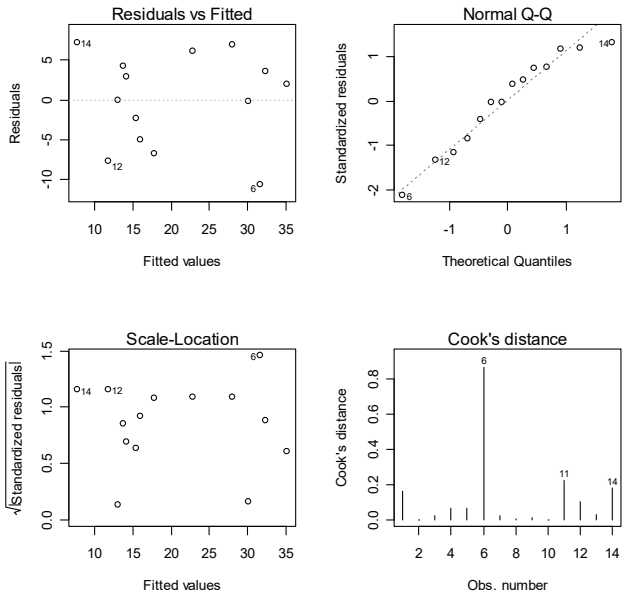


Figura 10.1.2.

Al fine di migliorare l'adeguatezza del modello iniziale di regressione multipla, può essere utile considerare interazioni o ulteriori trasformate dei regressori originali. Ad esempio, un modello iniziale con due regressori

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

potrebbe essere esteso considerando l'interazione fra le variabili

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

e introducendo di fatto un nuovo regressore (dato dal prodotto dei due regressori originali). Il modello iniziale potrebbe essere ulteriormente esteso considerando effetti non lineari dei regressori come nel seguente modello che introduce dipendenze quadratiche

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2$$

considerando in effetti due nuovi regressori (dati dai quadrati dei due regressori originali). Evidentemente, interazioni e dipendenze non lineari potrebbero essere introdotte congiuntamente

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2$$

considerando tre nuovi regressori. I modelli estesi costruiti in questo modo sono comunque lineari nei parametri e possono essere trattati come visto in precedenza. Tuttavia, si osservi che i nuovi modelli perdono la semplicità del modello originale e possono introdurre difficoltà di interpretazione dei parametri. Modelli più complessi devono dunque essere adottati con cautela.

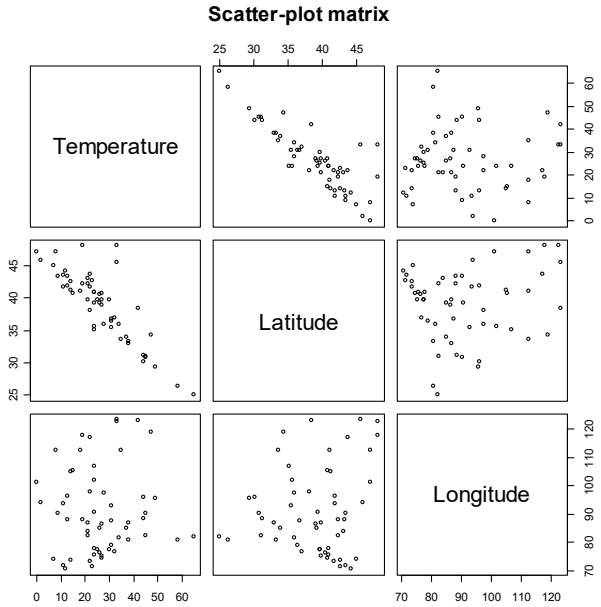


Figura 10.1.3.

• **Esempio 10.1.2.** Si dispone delle osservazioni delle temperature medie del mese di gennaio (in °F) nel periodo 1931-1960 in alcune città degli Stati Uniti (Fonte: Peixoto, J.L., 1990, A property of well-formulated polynomial regression models, *American Statistician* **44**, 26-30). I regressori sono la latitudine e la longitudine della città. I dati sono contenuti nel file `temperature.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\temperature.txt",
+    header = T)
> attach(d)
```

La matrice dei diagrammi di dispersione viene ottenuta mediante il seguente comando:

```
> pairs(d, main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 10.1.3. L'analisi del modello lineare con i regressori originali viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude + Longitude))

Call:
lm(formula = Temperature ~ Latitude + Longitude)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9983  -3.8957   0.5577   3.7330  22.0113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.64523    8.32708  11.846   <2e-16 ***
Latitude    -2.16355    0.17570 -12.314   <2e-16 ***
Longitude    0.13396    0.06314   2.122   0.0386 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.935 on 53 degrees of freedom
Multiple R-Squared: 0.7411,     Adjusted R-squared: 0.7314
F-statistic: 75.88 on 2 and 53 DF,  p-value: 2.792e-16
```

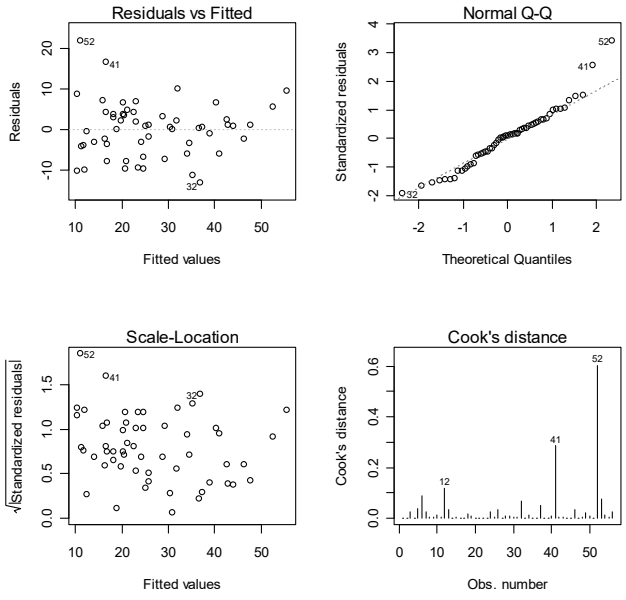


Figura 10.1.4.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello vengono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude + Longitude), which = c(1:4),
+    add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.4. L'analisi del modello lineare con i regressori originali senza l'osservazione anomala viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude + Longitude,
+     subset = (1:length(Temperature) != 52)))

Call:
lm(formula = Temperature ~ Latitude + Longitude,
subset = (1:length(Temperature) != 52))

Residuals:
       Min         1Q     Median         3Q        Max
-12.092772  -3.678680   0.001197   3.505167  19.667543

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.30043    7.84331  13.808   <2e-16 ***
Latitude     -2.28584    0.15992 -14.294   <2e-16 ***
Longitude     0.07522    0.05837   1.289    0.203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.184 on 52 degrees of freedom
Multiple R-Squared: 0.7971,     Adjusted R-squared: 0.7893
F-statistic: 102.2 on 2 and 52 DF,  p-value: < 2.2e-16
```

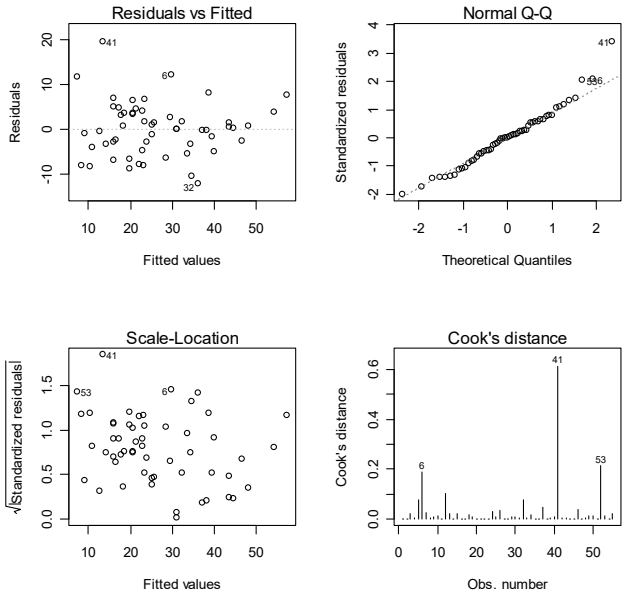


Figura 10.1.5.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello con i regressori originali e senza l'osservazione anomala sono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude + Longitude,
+     subset = (1:length(Temperature) != 52)),
+     which = c(1:4), add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.5. L'analisi del modello lineare con i regressori originali e l'interazione viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude * Longitude))

Call:
lm(formula = Temperature ~ Latitude * Longitude)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6738  -2.8165  -0.1268   3.4107  15.0605

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       259.48952   44.71515   5.803 3.93e-07 ***
Latitude           -6.07039    1.08235  -5.609 7.94e-07 ***
Longitude          -1.61025    0.48139  -3.345 0.001533 **
Latitude:Longitude  0.04220    0.01156   3.649 0.000611 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.247 on 52 degrees of freedom
Multiple R-Squared: 0.7939,     Adjusted R-squared: 0.782
F-statistic: 66.77 on 3 and 52 DF,  p-value: < 2.2e-16
```

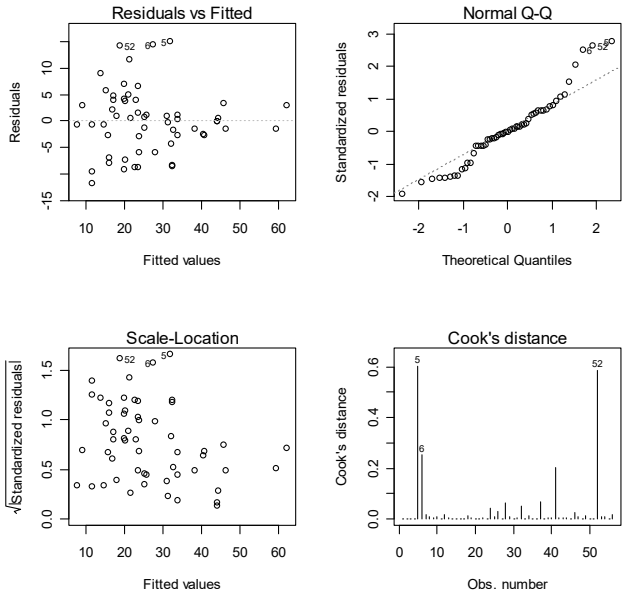


Figura 10.1.6.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello con interazione sono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude * Longitude), which = c(1:4),
+    add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.6. L'analisi del modello lineare con i regressori originali, l'interazione e una dipendenza cubica dalla longitudine viene implementata mediante il seguente comando:

```
> summary(lm(Temperature ~ Latitude * Longitude + I(Longitude^3)))

Call:
lm(formula = Temperature ~ Latitude * Longitude + I(Longitude^3))

Residuals:
     Min       1Q   Median       3Q      Max
-7.30440 -2.85850  0.04342  2.49406  9.06776

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.908e+02  3.151e+01  12.402  < 2e-16 ***
Latitude           -5.891e+00  6.773e-01  -8.698 1.20e-11 ***
Longitude          -3.632e+00  3.750e-01  -9.687 3.77e-13 ***
I(Longitude^3)      8.064e-05  8.912e-06   9.050 3.47e-12 ***
Latitude:Longitude  3.656e-02  7.261e-03   5.035 6.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.908 on 51 degrees of freedom
Multiple R-Squared: 0.9209,     Adjusted R-squared: 0.9147
F-statistic: 148.5 on 4 and 51 DF,  p-value: < 2.2e-16
```

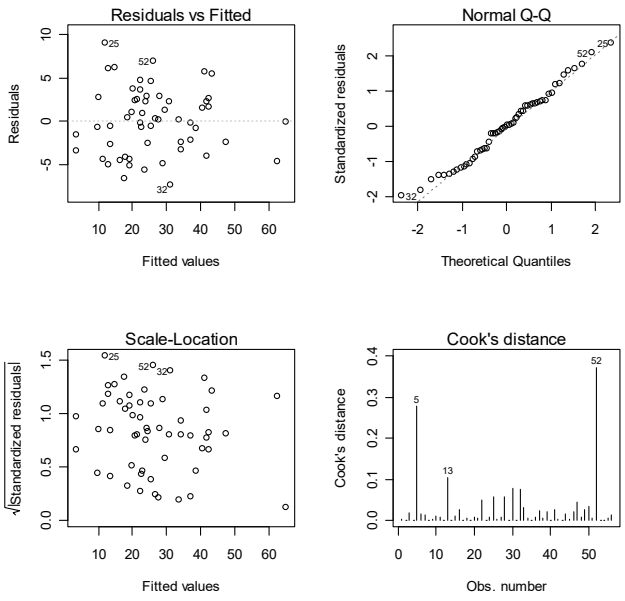


Figura 10.1.7.

I diagrammi di Anscombe, il diagramma quantile-quantile e il diagramma con le distanze di Cook del modello con interazione e dipendenza cubica sono ottenuti mediante i seguenti comandi:

```
> par(mfrow = c(2, 2))
> plot(lm(Temperature ~ Latitude * Longitude + I(Longitude^3)),
+    which = c(1:4), add.smooth = FALSE)
> par(mfrow = c(1, 1))
```

I precedenti comandi forniscono il grafico della Figura 10.1.7. Il modello esteso con interazione e dipendenza cubica dalla longitudine non può essere semplificato:

```
> summary(step(lm(Temperature ~ Latitude * Longitude +
+    I(Longitude^3))))
Start:  AIC=157.41
Temperature ~ Latitude * Longitude + I(Longitude^3)

                     Df Sum of Sq     RSS     AIC
<none>                             778.71  157.41
- Latitude:Longitude  1    387.05 1165.76  178.00
- I(Longitude^3)      1   1250.42 2029.12  209.04

Call:
lm(formula = Temperature ~ Latitude * Longitude + I(Longitude^3))

Residuals:
     Min       1Q   Median       3Q      Max
-7.30440 -2.85850  0.04342  2.49406  9.06776

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.908e+02  3.151e+01  12.402  < 2e-16 ***
Latitude           -5.891e+00  6.773e-01  -8.698 1.20e-11 ***
Longitude          -3.632e+00  3.750e-01  -9.687 3.77e-13 ***
I(Longitude^3)      8.064e-05  8.912e-06   9.050 3.47e-12 ***
Latitude:Longitude  3.656e-02  7.261e-03   5.035 6.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.908 on 51 degrees of freedom
Multiple R-Squared: 0.9209,     Adjusted R-squared: 0.9147
F-statistic: 148.5 on 4 and 51 DF,  p-value: < 2.2e-16
```

Anche se questo modello fornisce il migliore adeguamento, si tenga presente che il significato dei regressori non è facilmente interpretabile. □

La struttura inferenziale dell'analisi della varianza basata su r campioni casuali indipendenti, ciascuno di numerosità n_j e tali che $\sum_{j=1}^{r} n_j = n$, da una variabile casuale Y a r livelli $c_1, \ldots, c_r$ di un fattore, con $\mathrm{E}[Y_{c_j}] = \mu_j$ e $\mathrm{Var}[Y_{c_j}] = \sigma^2$, può essere riportata ad un modello di regressione lineare. Se $Y_1, \ldots, Y_n$ sono le osservazioni relative al campione misto, allora si può scrivere il modello lineare

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{r-1} x_{i,r-1}$$

e

$$\mathrm{Var}[Y_i] = \sigma^2 \,,$$

dove $\beta_0 = \mu_1$ e $\beta_j = \mu_{j+1} - \mu_1$, mentre x_{ij} è il valore assunto sull'i-esima unità da un regressore binario che vale uno se l'osservazione è relativa al $(j+1)$-esimo livello del fattore e zero altrimenti. L'ipotesi di omogeneità delle medie, supponendo la normalità di $Y_1, \ldots, Y_n$, viene verificata considerando il sistema di ipotesi $H_0 : \beta_1 = \ldots = \beta_{r-1} = 0$ contro $H_1 : \beta_j \neq 0$ per qualche j. Di conseguenza, le tecniche viste in precedenza possono essere applicate anche in questo caso. Il

vantaggio di questo approccio è quello di individuare da quale livello del fattore dipende l'eventuale rifiuto dell'ipotesi di base.

• **Esempio 10.1.3.** Si considera di nuovo i dati relativi alla velocità della luce dell'Esempio 8.3.1. L'ipotesi di omogeneità delle medie può essere verificato mediante il seguente comando:

```
> summary.lm(lm(Speed ~ Trial))

Call:
lm(formula = Speed ~ Trial)

Residuals:
    Min      1Q  Median      3Q     Max
-259.00  -42.62    2.25   41.75  161.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   909.00      16.60  54.762  < 2e-16 ***
TrialT2       -53.00      23.47  -2.258 0.026251 *
TrialT3       -64.00      23.47  -2.726 0.007627 **
TrialT4       -88.50      23.47  -3.770 0.000283 ***
TrialT5       -77.50      23.47  -3.301 0.001356 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.23 on 95 degrees of freedom
Multiple R-Squared: 0.1529,     Adjusted R-squared: 0.1173
F-statistic: 4.288 on 4 and 95 DF,  p-value: 0.003114
```

L'omogeneità delle medie viene ovviamente respinta con lo stesso livello di significatività osservata dell'Esempio 8.3.1. Analizzando le stime dei coefficienti, si nota che il rifiuto è principalmente dovuto ai coefficienti β_3 e β_4 le cui stime sono elevate in valore assoluto e con segno negativo, ovvero il rifiuto risulta principalmente da $\mu_1 > \mu_4$ e $\mu_1 > \mu_5$, analogamente all'Esempio 8.3.1. □

L'analisi della varianza può essere estesa anche al caso che si abbiano due o più fattori, per cui si ha la cosiddetta analisi della varianza a due o più criteri. Anche in questo caso l'analisi può essere riportata ad un modello di regressione lineare. Per descrivere questo modello, si consideri per semplicità una analisi della varianza a due criteri, dove ogni fattore assume due livelli. Dunque, si hanno 4 campioni, ovvero un campione per ogni combinazione di livelli. Assumendo la normalità delle osservazioni, se $Y_1, \ldots, Y_n$ sono le osservazioni relative al campione misto, allora si può scrivere il modello lineare

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

e

$$\mathrm{Var}[Y_i] = \sigma^2 \,,$$

dove x_{i1} è il valore assunto sull'i-esima unità da un regressore binario che vale uno se l'osservazione è relativa al secondo livello del primo fattore e zero altrimenti, mentre x_{i2} è il valore assunto sull'i-esima unità da un regressore binario che vale uno se l'osservazione è relativa al secondo livello del secondo fattore e zero altrimenti. Evidentemente, in questo caso la verifica d'ipotesi risulta più complessa. In effetti, si può voler verificare l'effetto marginale del primo fattore (ovvero l'ipotesi $H_0 : \beta_1 = 0$), l'effetto marginale del secondo fattore (ovvero l'ipotesi $H_0 : \beta_2 = 0$) o l'effetto dell'interazione dei due fattori (ovvero l'ipotesi $H_0 : \beta_3 = 0$). In una maniera simile anche se con

crescente complessità in notazione, il caso generale dell'analisi della varianza a più criteri può essere riportata ad un modello di regressione. Il vantaggio di questi approcci è quello di individuare da quale combinazioni di livelli dei fattori dipende la media delle osservazioni.

● **Esempio 10.1.4.** Si dispone delle osservazioni dei pesi di topi (in grammi) con quattro differenti genotipi per la genitrice e quattro differenti genotipi per la nidiata (Fonte: Scheffé, H., 1959, *Analysis of Variance*, Wiley, New York, p.140). I dati sono contenuti nel file `foster.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\foster.txt", header = T)
> attach(d)
```

Il seguente comando permette di ottenere le medie per tutte le combinazioni dei fattori:

```
> tapply(Weight, list(Mother, Litter), mean)
       A      B        I        J
A 63.680 52.325 47.10000 54.35000
B 52.400 60.640 64.36667 56.10000
I 54.125 53.925 51.60000 54.53333
J 48.960 45.900 49.43333 49.06000
```

L'analisi della varianza a due criteri può essere implementata mediante il seguente comando:

```
> summary(aov(Weight ~ Litter * Mother))
              Df  Sum Sq Mean Sq F value   Pr(>F)
Litter         3   60.16   20.05  0.3697 0.775221
Mother         3  775.08  258.36  4.7632 0.005736 **
Litter:Mother  9  824.07   91.56  1.6881 0.120053
Residuals     45 2440.82   54.24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il seguente comando che semplifica il modello di regressione è equivalente alla precedente analisi della varianza:

```
> summary(step(lm(Weight ~ Litter * Mother)))
Start:  AIC=257.04
Weight ~ Litter * Mother

                Df Sum of Sq    RSS    AIC
- Litter:Mother  9     824.1 3264.9  256.8
<none>                       2440.8  257.0

Step:  AIC=256.79
Weight ~ Litter + Mother

         Df Sum of Sq    RSS    AIC
- Litter  3      63.6 3328.5  252.0
<none>                3264.9  256.8
- Mother  3     775.1 4040.0  263.8

Step:  AIC=251.96
Weight ~ Mother
```

```
         Df Sum of Sq    RSS    AIC
<none>                 3328.5  252.0
- Mother  3     771.6 4100.1  258.7

Call:
lm(formula = Weight ~ Mother)

Residuals:
   Min     1Q Median     3Q    Max
-19.10  -5.90   1.50   5.32  12.80

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   55.400      1.910  28.999   <2e-16 ***
MotherB        3.300      2.797   1.180   0.2429
MotherI       -2.038      2.702  -0.754   0.4539
MotherJ       -6.720      2.746  -2.447   0.0175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.642 on 57 degrees of freedom
Multiple R-Squared: 0.1882,     Adjusted R-squared: 0.1455
F-statistic: 4.405 on 3 and 57 DF,  p-value: 0.007433
```

Dunque, si rifiuta l'ipotesi di base ed il rifiuto è dovuto al genotipo della genitrice. L'analisi con il modello di regressione evidenzia in particolare che il primo genotipo della genitrice tende a produrre nidiate di tipo maggiore rispetto all'ultimo genotipo. □

10.2. I modelli additivi generalizzati

In molti casi risulta difficile specificare la struttura di dipendenza della variabile di risposta dai regressori e questo fatto preclude la costruzione di un modello di regressione lineare (o comunque riconducibile a tale). Un approccio "distribution-free" alla regressione multipla può essere ottenuto considerando il modello additivo generalizzato con

$$\mathrm{E}[Y_i] = \beta_0 + m_1(x_{i1}) + \ldots + m_p(x_{ip})$$

e

$$\mathrm{Var}[Y_i] = \sigma^2 \,,$$

dove $m_1, \ldots, m_p$ sono funzioni non note. Dunque, il modello additivo generalizzato suppone che la variabile di risposta sia dipendente da una somma di trasformazioni non note dei regressori. Le funzioni $m_1, \ldots, m_p$ e la quantità β_0 vengono stimate con una procedura simile al caso della regressione lineare locale, ottenendo i valori stimati

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{m}_1(x_{i1}) + \ldots + \widehat{m}_p(x_{ip}) \,.$$

L'adeguatezza del modello può essere valutata mediante la devianza che è la somma dei quadrati dei residui dei valori osservati da quelli stimati, ovvero

$$D = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \,.$$

• **Esempio 10.2.1.** Si considera di nuovo i dati relativi alle temperature delle città degli Stati Uniti. La stima del modello viene ottenuta richiamando la libreria `mgcv`. L'analisi del modello additivo generalizzato viene implementata come segue:

```
> library(mgcv)
> summary(gam(Temperature ~ s(Latitude) + s(Longitude)))

Family: gaussian
Link function: identity

Formula:
Temperature ~ s(Latitude) + s(Longitude)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.5179     0.3487   76.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Est.rank      F  p-value
s(Latitude)  6.579        9 119.85  < 2e-16 ***
s(Longitude) 4.867        9  31.31 3.46e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.962   Deviance explained =   97%
GCV score = 8.7557   Scale est. = 6.8098    n = 56
```

I grafici delle funzioni $\widehat{m}_1(x_1)$ e $\widehat{m}_2(x_2)$ possono essere ottenuti mediante il comando

```
> plot(gam(Temperature ~ s(Latitude) + s(Longitude)))
```

Il precedente comando fornisce i grafici della Figura 10.2.1 e della Figura 10.2.2.

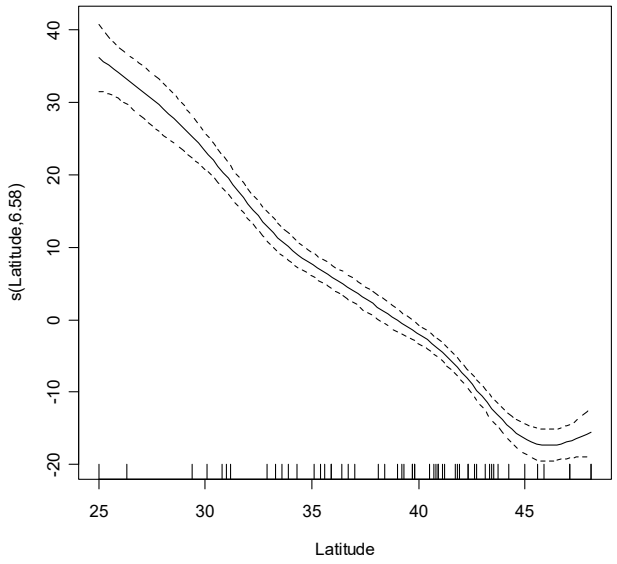


Figura 10.2.1.

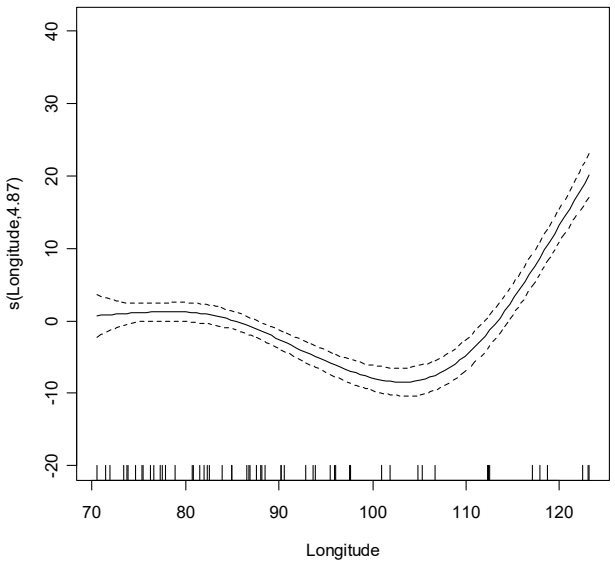


Figura 10.2.2.

Dai precedenti grafici si osserva che l'effetto della latitudine è sostanzialmente lineare quando lo spostamento avviene verso nord, mentre l'effetto della longitudine è curvilineo quando lo spostamento avviene da est verso ovest. □

10.3. I modelli lineari generalizzati multipli

Il modello lineare generalizzato multiplo viene ottenuto assumendo una distribuzione non Normale per $Y_1, \ldots, Y_n$ e considerando una opportuna funzione legame g tale che

$$g(\mathrm{E}[Y_i]) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \, .$$

Il modello di regressione lineare multiplo è un caso particolare di questa classe quando le variabili di risposta hanno distribuzione Normale e g è la funzione identità. In modo simile al caso univariato, per ogni distribuzione che si assume per la variabile di risposta esiste una funzione legame canonica, ovvero una parametrizzazione naturale del modello.

Quando si assume che $Y_1, \ldots, Y_n$ siano variabili casuali di Poisson, si ottiene la regressione di Poisson multipla e la funzione legame canonica è la funzione logaritmo, ovvero si ha

$$\log(\mathrm{E}[Y_i]) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \, .$$

Inoltre, se si assume che $Y_1, \ldots, Y_n$ siano variabili casuali di Bernoulli, si ha la regressione logistica multipla e la funzione legame canonica è la funzione logit, ovvero si ha

$$\log\left(\frac{\mathrm{E}[Y_i]}{1 - \mathrm{E}[Y_i]}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \, .$$

In modo analogo al caso univariato, se si dispone delle frequenze relative ai vari livelli dei regressori, la regressione logistica può essere applicata considerando le proporzioni della variabile di risposta per ogni livello dei regressori.

I modelli additivi possono essere adattati anche ai modelli lineari generalizzati introducendo una relazione del tipo

$$g(\mathrm{E}[Y_i]) = \beta_0 + m_1(x_{i1}) + \ldots + m_p(x_{ip}) \,,$$

dove $m_1, \ldots, m_p$ sono funzioni non note. Questo tipo di modello risulta molto flessibile e particolarmente adatto per le applicazioni reali.

• **Esempio 10.3.1.** Si dispone dei dati relativi ai fallimenti di industrie degli Stati Uniti che hanno operato nelle telecomunicazioni fra il 2000 e il 2002 (Fonte: Simonoff, J.S., 2003, *Analyzing Categorical Data*, Springer, New York, p.381). I regressori sono il capitale da lavoro rapportato all'attivo totale, il guadagno rapportato all'attivo totale, il guadagno al netto di interessi e tasse rapportato all'attivo totale, le vendite rapportate all'attivo totale e il valore dell'azienda rapportato alla passività. I dati sono contenuti nel file `bankruptcy.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\bankruptcy.txt", header = T)
> attach(d)
```

La matrice dei diagrammi di dispersione viene ottenuta mediante il seguente comando:

```
> pairs(d[,2:7], main = "Scatter-plot matrix")
```

Il precedente comando fornisce il grafico della Figura 10.3.1.

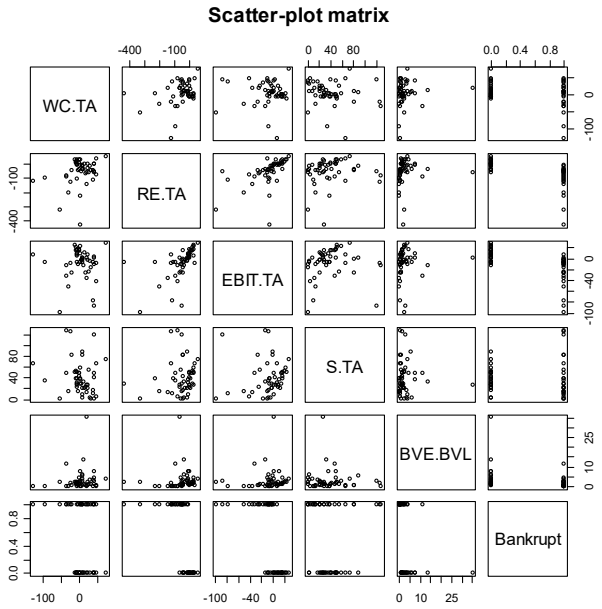


Figura 10.3.1.

L'analisi del modello di regressione logistica viene effettuata sui primi quattro regressori, dato che il quinto non sembra influenzare la probabilità di fallimento. La semplificazione automatica di questo modello mediante il criterio AIC viene implementata mediante il seguente comando:

```
> summary(step(glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA +
+    S.TA, binomial)))
```

```
Start:  AIC=32.92
Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA

          Df Deviance    AIC
- S.TA     1   22.968 30.968
- WC.TA    1   24.493 32.493
<none>         22.923 32.923
- RE.TA    1   26.347 34.347
- EBIT.TA  1   28.529 36.529

Step:  AIC=30.97
Bankrupt ~ WC.TA + RE.TA + EBIT.TA

          Df Deviance    AIC
- WC.TA    1   24.532 30.532
<none>         22.968 30.968
- RE.TA    1   26.407 32.407
- EBIT.TA  1   28.594 34.594

Step:  AIC=30.53
Bankrupt ~ RE.TA + EBIT.TA

          Df Deviance    AIC
<none>         24.532 30.532
- EBIT.TA  1   28.697 32.697
- RE.TA    1   35.979 39.979

Call:
glm(formula = Bankrupt ~ RE.TA + EBIT.TA, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66450  -0.33579  -0.01778   0.28999   1.97243

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.95801    0.81637  -2.398   0.0165 *
RE.TA       -0.04405    0.01989  -2.214   0.0268 *
EBIT.TA     -0.10761    0.06445  -1.670   0.0950 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.315  on 49  degrees of freedom
Residual deviance: 24.532  on 47  degrees of freedom
AIC: 30.532

Number of Fisher Scoring iterations: 7
```

Un modello più semplice basato sul solo primo regressore potrebbe essere ottenuto notando che esiste un valore anomalo, ovvero la prima osservazione. L'analisi di questo modello viene implementata come segue:

```
> summary(glm(Bankrupt ~ RE.TA, binomial,
+     subset = (1:length(Bankrupt) != 1)))
```

```
Call:
glm(formula = Bankrupt ~ RE.TA, family = binomial,
subset = (1:length(Bankrupt) != 1))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.07214  -0.30405  -0.03211   0.24542   2.07304

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05107    1.08850  -2.803  0.00506 **
RE.TA       -0.08277    0.02591  -3.194  0.00140 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.908  on 48  degrees of freedom
Residual deviance: 24.256  on 47  degrees of freedom
AIC: 28.256

Number of Fisher Scoring iterations: 8
```

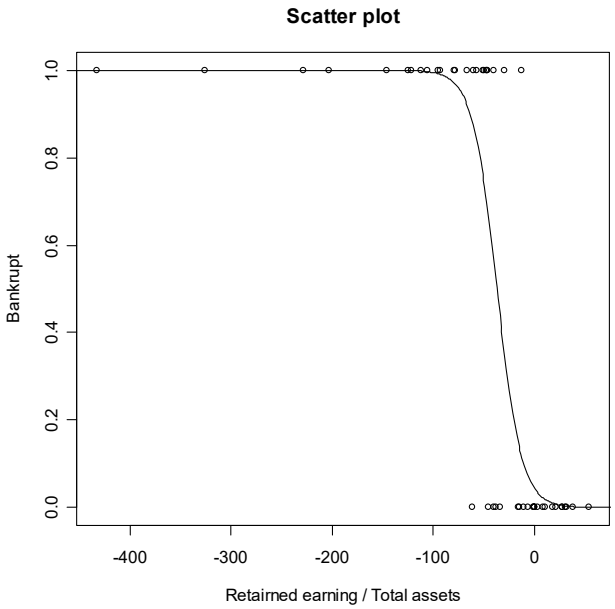


Figura 10.3.2.

Nell'ultimo modello, il diagramma di dispersione con la relativa funzione legame stimata viene ottenuto mediante i seguenti comandi:

```
> Bankrupt.h <- Bankrupt[2:50]
> RE.TA.h <- RE.TA[2:50]
> plot(RE.TA.h, Bankrupt.h,
+    xlab = "Retairned earning / Total assets",
+    ylab = "Bankrupt", main = "Scatter plot")
> lines(seq(-500, 100, 1),
+    predict(glm(Bankrupt.h ~ RE.TA.h, binomial),
+    data.frame(RE.TA.h = seq(-500, 100, 1)), type = "response"))
```

I precedenti comandi forniscono il grafico della Figura 10.3.2. Dall'analisi si questo grafico si osserva che la probabilità di fallimento sale drasticamente da 0 a 1 quando il guadagno rapportato all'attivo totale tende a zero, un risultato facilmente intuibile. □

I modelli lineari generalizzati possono essere applicati anche all'analisi delle tabelle a due o più entrate. Questi modelli sono detti log-lineari. Per descrivere questo modello si consideri per semplicità una tabella a doppia entrata, dove ogni fattore assume due livelli. Supponendo che le quattro frequenze osservate N_{ij} siano delle variabili casuali di Poisson (e dunque si considera in effetti lo schema probabilistico di Poisson descritto nell'Esempio 8.4.6), il modello log-lineare è dato da

$$\log(\mathrm{E}[N_{jl}]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \,,$$

dove x_1 è un regressore binario che vale uno se la frequenza è relativa al secondo livello del primo fattore e zero altrimenti, mentre x_2 è un regressore binario che vale uno se la frequenza è relativa al secondo livello del secondo fattore e zero altrimenti. La funzione legame canonica è evidentemente la funzione logaritmo per le assunzioni fatte. In questo caso, si può voler verificare l'effetto marginale del primo fattore (ovvero l'ipotesi $H_0 : \beta_1 = 0$), l'effetto marginale del secondo fattore (ovvero l'ipotesi $H_0 : \beta_2 = 0$) o l'effetto dell'interazione dei due fattori (ovvero l'ipotesi $H_0 : \beta_3 = 0$). In una maniera simile anche se con complessità in notazione, il caso generale dell'analisi delle tabelle a più entrate può essere riportata ad un modello log-lineare.

• **Esempio 10.3.2.** Durante uno studio sull'uso di droghe da parte degli studenti nell'università di Dayton (Ohio) nel 1992 sono stati considerati un gruppo di studenti, ognuno dei quali è stato classificato per l'uso o meno di alcolici, di sigarette e di marijuana (Agresti, A., 1990, *Categorical data analysis*, prima edizione, Wiley, New York, p.152). I dati sono contenuti nel file `drug.txt` e vengono letti e resi disponibili mediante i comandi:

```
> d <- read.table("c:\\Rwork\\examples\\drug.txt", header = T)
> attach(d)
```

La tabella a più entrate viene ottenuta mediante il comando:

```
> xtabs(Count ~ Alcohol + Cigarette + Marijuana)
, , Marijuana = No

       Cigarette
Alcohol  No Yes
    No  279  43
    Yes 456 538

, , Marijuana = Yes

       Cigarette
Alcohol  No Yes
    No    2   3
    Yes  44 911
```

Il diagramma a nastri condizionato viene ottenuto mediante il comando:

```
> library(lattice)
> barchart(xtabs(Count ~ Alcohol + Cigarette + Marijuana),
+    ylab = "Alcohol use",
+    auto.key = list(title = "Marijuana use", cex = 0.8),
+    strip = strip.custom(strip.names = T, strip.levels = T))
```

I precedenti comandi forniscono il grafico della Figura 10.3.3.

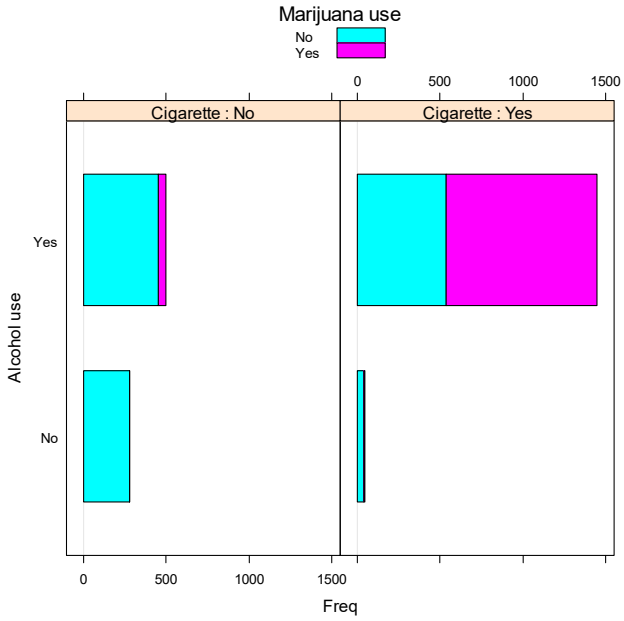


Figura 10.3.3.

Il comando `chisq.test` fornisce l'implementazione del test χ^2 per l'indipendenza:

```
> chisq.test(xtabs(Count ~ Alcohol + Cigarette + Marijuana))

        Chi-squared test for given probabilities

data:  xtabs(Count ~ Alcohol + Cigarette + Marijuana)
X-squared = 2676.337, df = 7, p-value < 2.2e-16
```

Una volta rifiutata l'indipendenza, il modello log-lineare può essere analizzato mediante il seguente comando:

```
> summary(step(glm(Count ~ Alcohol * Cigarette * Marijuana,
+    poisson)))

Start:  AIC=65.04
Count ~ Alcohol * Cigarette * Marijuana

                              Df  Deviance    AIC
- Alcohol:Cigarette:Marijuana  1     0.374 63.417
<none>                           -2.92e-13 65.043

Step:  AIC=63.42
Count ~ Alcohol + Cigarette + Marijuana + Alcohol:Cigarette +
    Alcohol:Marijuana + Cigarette:Marijuana

                      Df Deviance    AIC
<none>                      0.37  63.42
- Alcohol:Marijuana    1    92.02 153.06
- Alcohol:Cigarette    1   187.75 248.80
- Cigarette:Marijuana  1   497.37 558.41
```

```
Call:
glm(formula = Count ~ Alcohol + Cigarette + Marijuana +
Alcohol:Cigarette + Alcohol:Marijuana + Cigarette:Marijuana,
family = poisson)

Deviance Residuals:
       1         2         3         4         5         6         7
 0.02044  -0.02658  -0.09256   0.02890  -0.33428   0.09452   0.49134
       8
-0.03690

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                5.63342    0.05970  94.361  < 2e-16 ***
AlcoholYes                 0.48772    0.07577   6.437 1.22e-10 ***
CigaretteYes              -1.88667    0.16270 -11.596  < 2e-16 ***
MarijuanaYes              -5.30904    0.47520 -11.172  < 2e-16 ***
AlcoholYes:CigaretteYes    2.05453    0.17406  11.803  < 2e-16 ***
AlcoholYes:MarijuanaYes    2.98601    0.46468   6.426 1.31e-10 ***
CigaretteYes:MarijuanaYes  2.84789    0.16384  17.382  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2851.46098  on 7  degrees of freedom
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.417

Number of Fisher Scoring iterations: 4
```

L'analisi del modello porta dunque a concludere che vi è una forte dipendenza fra le coppie di sostanze, mentre non esiste una dipendenza fra l'uso contemporaneo delle tre sostanze. □

10.4. Riferimenti bibliografici

Atkinson, A.C., Riani, M. e Cerioli, A. (2004) *Exploring Multivariate Data with the Forward Search*, Springer, New York.

Agresti, A. (2013) *Categorical Data Analysis*, terza edizione, Wiley, New York.

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*, Wiley, New York.

Bretz, F., Hothorn, T. e Westfall, P. (2011) *Multiple Comparisons using R*, Chapman & Hall/CRC Press, Boca Raton.

Dickhaus, T. (2014) *Simultaneous Statistical Inference*, Springer, New York.

Dunn, P.K. e Smyth, G.K. (2018) *Generalized Linear Models with Examples in R*, Springer, New York.

Everitt, B. e Hothorn, T. (2011) *An Introduction to Applied Multivariate Analysis with R*, Springer, New York.

Fieller, N. (2016) *Basics of Matrix Algebra for Statistics with R*, CRC Press, Boca Raton.

Friendly, M. e Meyer, D. (2019) *Discrete Data Analysis with R*, Chapman and Hall, London.

Hastie, T., Tibshirani, R. e Friedman, J.H. (2009) *The Elements of Statistical Learning*, seconda edizione, Springer, New York.

Hastie, T., Tibshirani, R. e Wainwright, M. (2015) *Statistical Learning with Sparsity*, Chapman and Hall/CRC Press, Boca Raton.

Harezlak, J., Ruppert, D. e Wand, M.P. (2018) *Semiparametric Regression with R*, Springer, New York.

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2014) *An Introduction to Statistical Learning*, Springer, New York.

Klemelä, J. (2014) *Multivariate Nonparametric Regression and Visualization*, Wiley, New York.

McCullagh, P. e Nelder, J.A. (1989) *Generalized Linear Models*, seconda edizione, Chapman and Hall, London.

Simonoff, J.S. (2003) *Analyzing Categorical Data*, Springer, New York.

Seber, G.A.F. (2015) *The Linear Model and Hypothesis*, Springer, New York.

Weisberg, S. (2014) *Applied Linear Regression*, Wiley, New York.

Wood, S.N. (2017) *Generalized Additive Models*, seconda edizione, Chapman and Hall, London.

Xu, J. (2023) *Modern Applied Regressions*, CRC Press, Boca Raton.