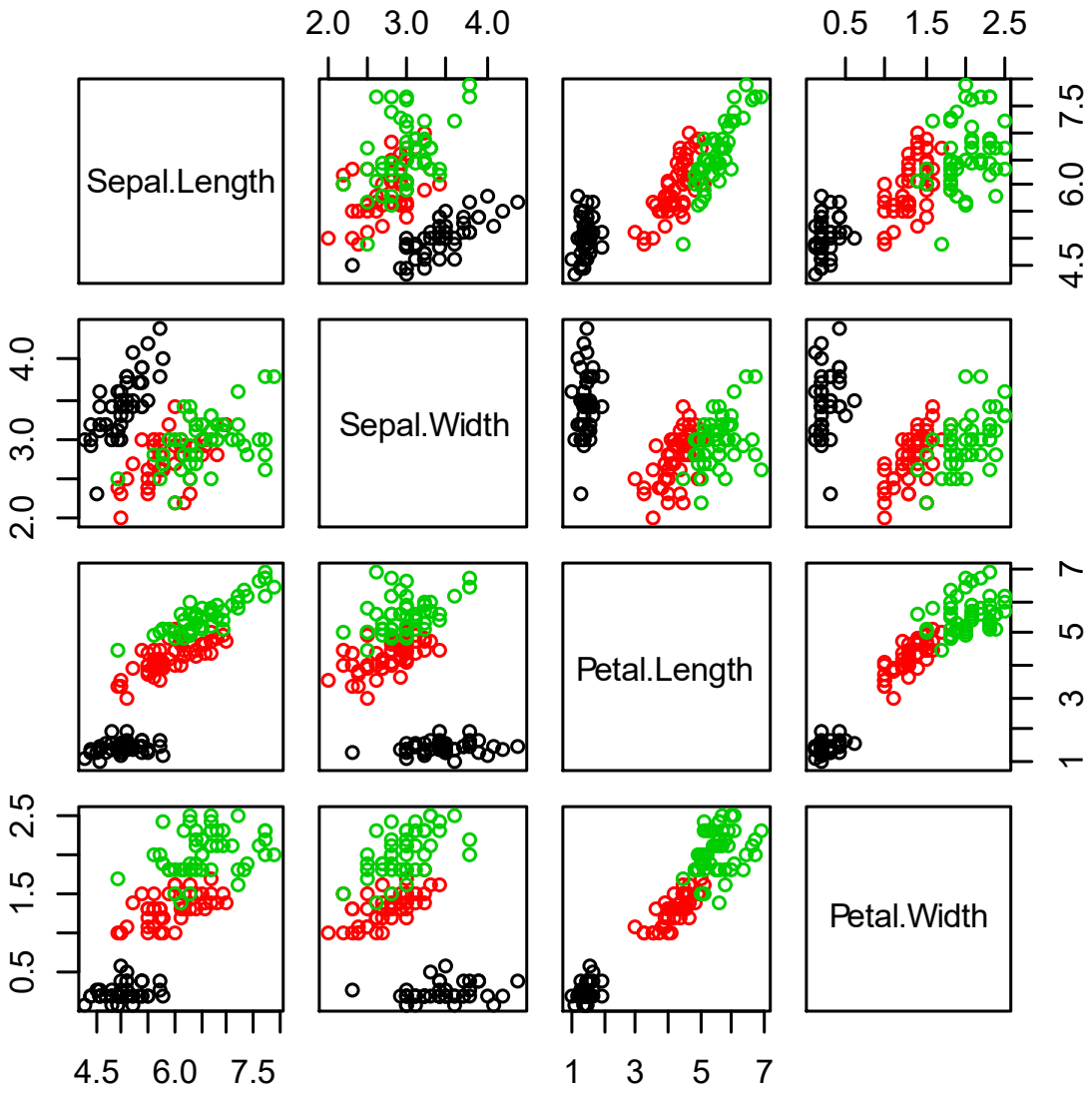
Elementi di Inferenza per Data Science

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 un 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.

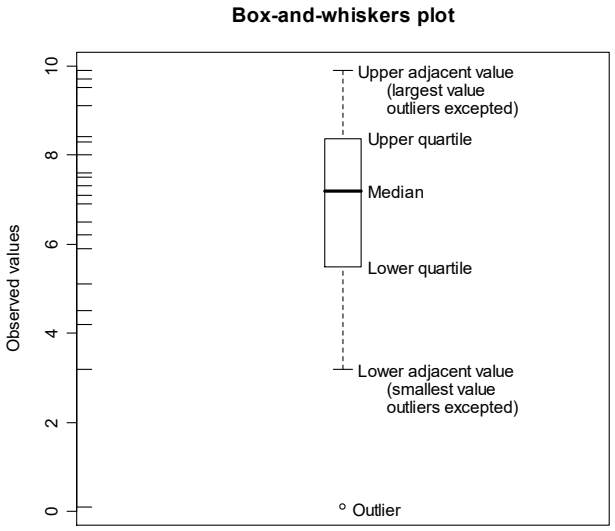


Figura 1.2.1.

• **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. □

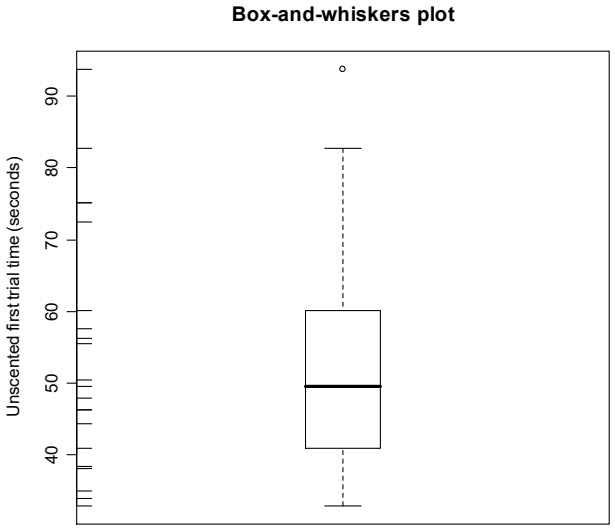


Figura 1.2.2.

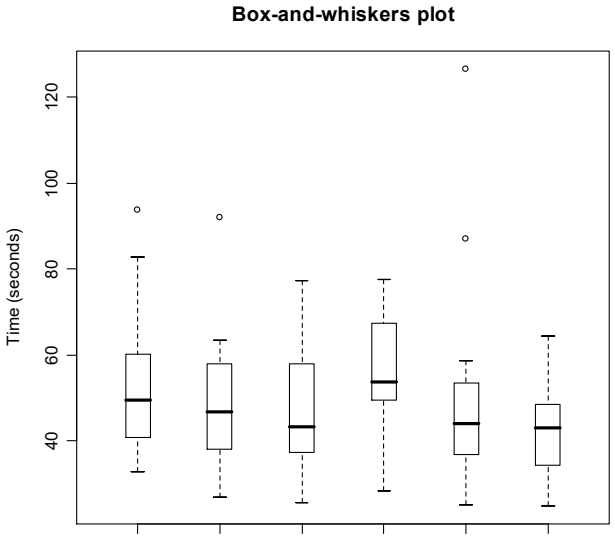


Figura 1.2.3.

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
```

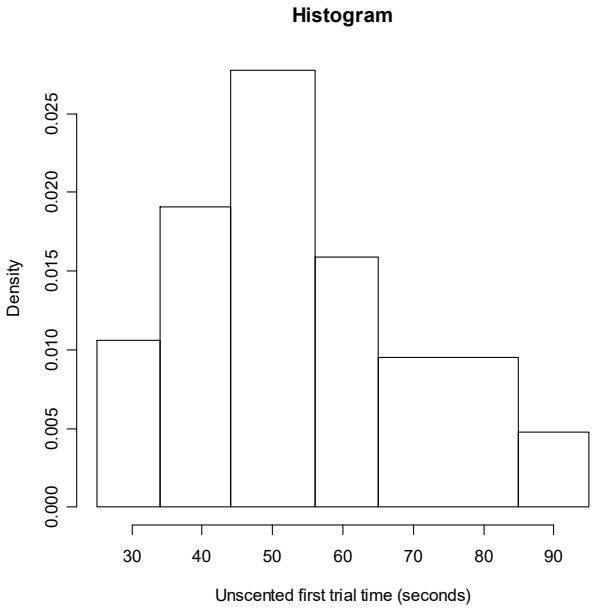


Figura 1.2.4.

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 spesso 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.

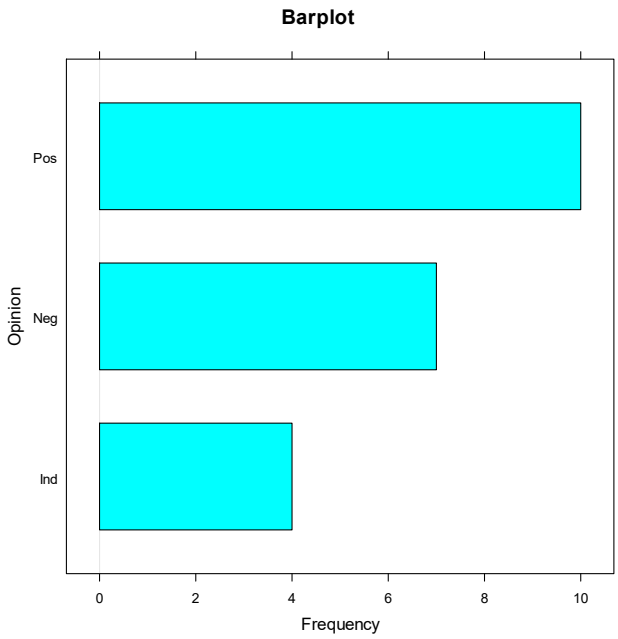


Figura 1.2.5.

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.

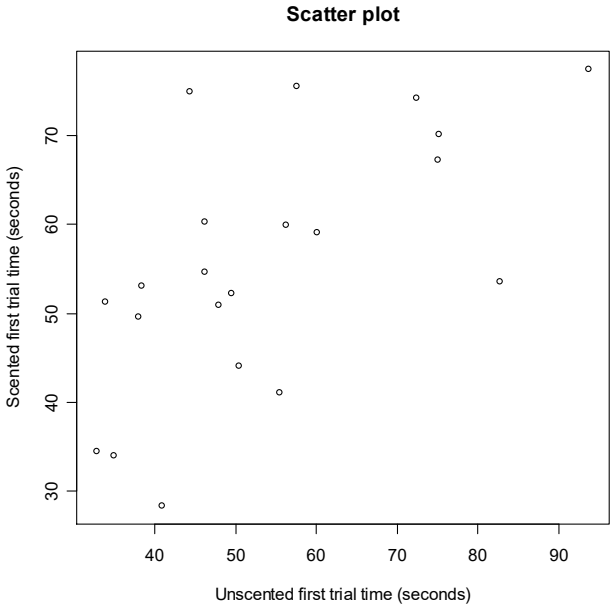


Figura 1.3.1.

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. □

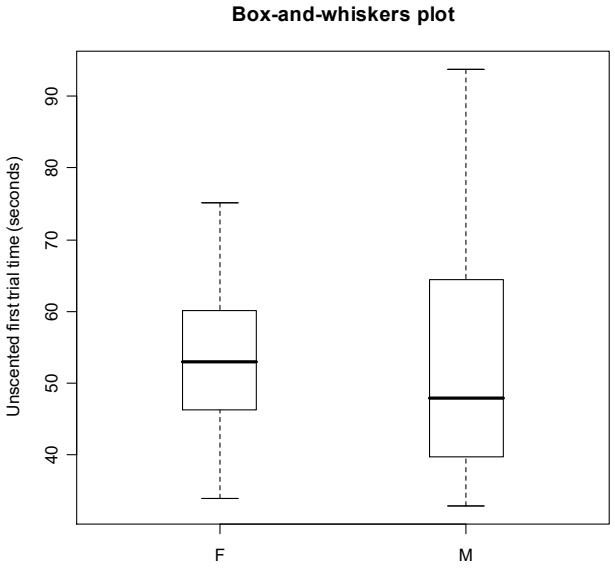


Figura 1.3.2.

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.

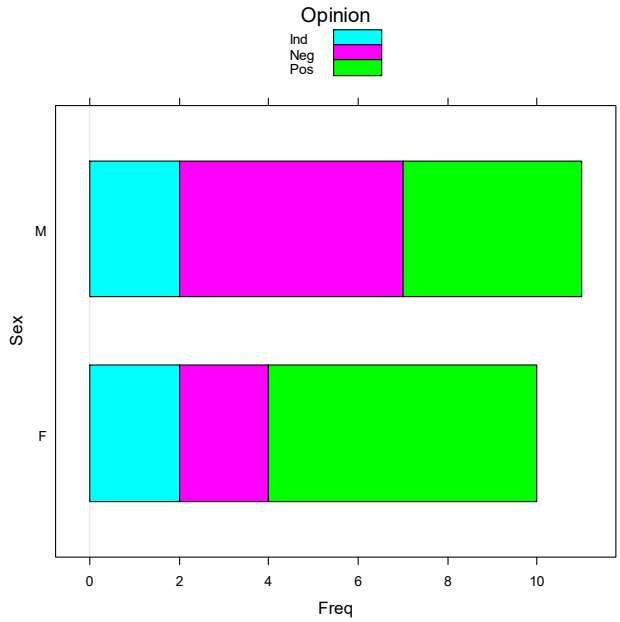


Figura 1.3.3.

• **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.

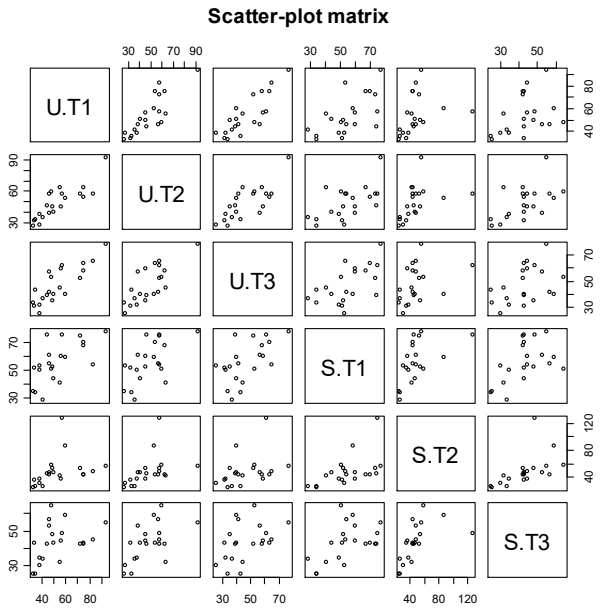


Figura 1.4.1.

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. □

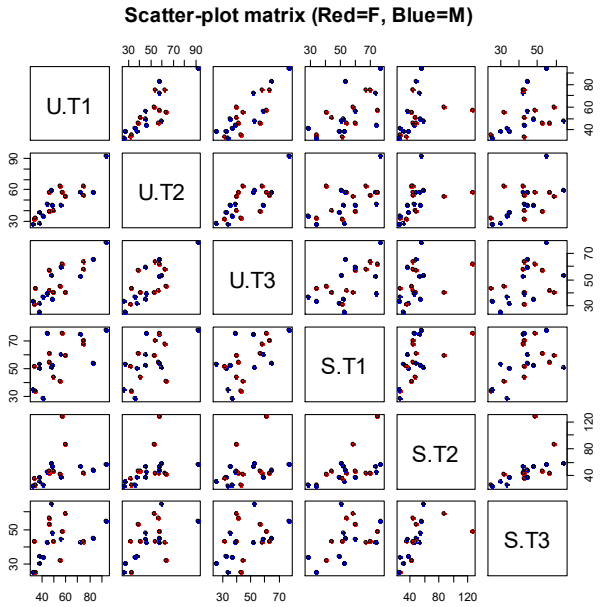


Figura 1.4.2.

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.

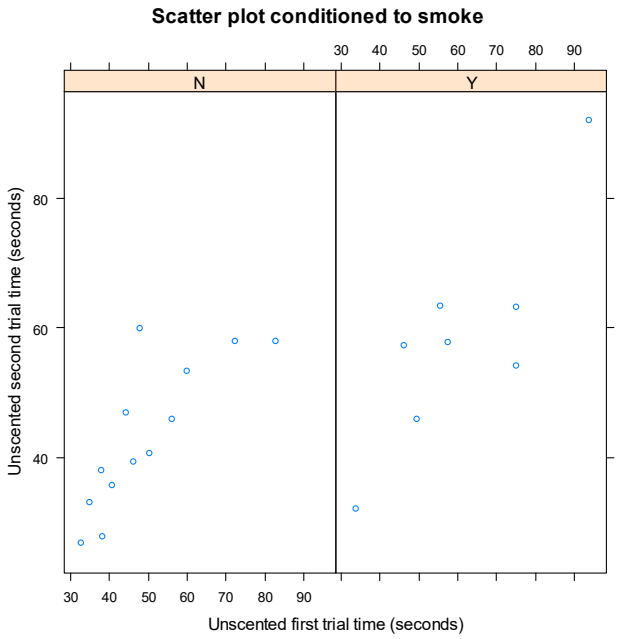


Figura 1.4.3.

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. □

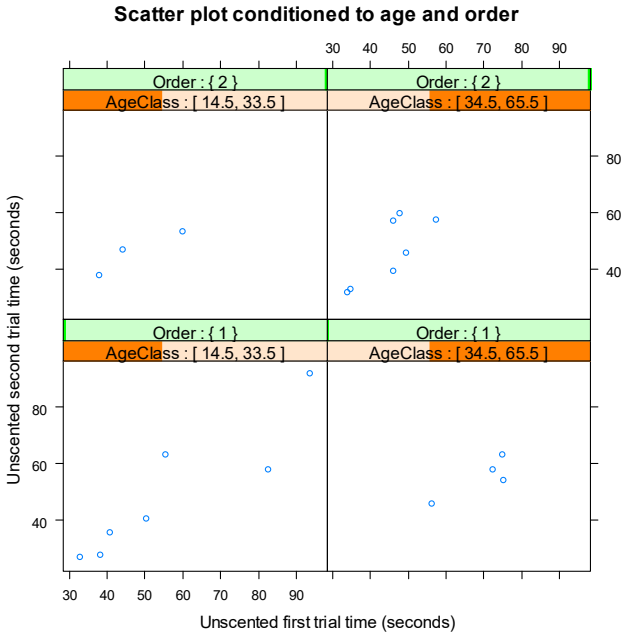


Figura 1.4.4.

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.

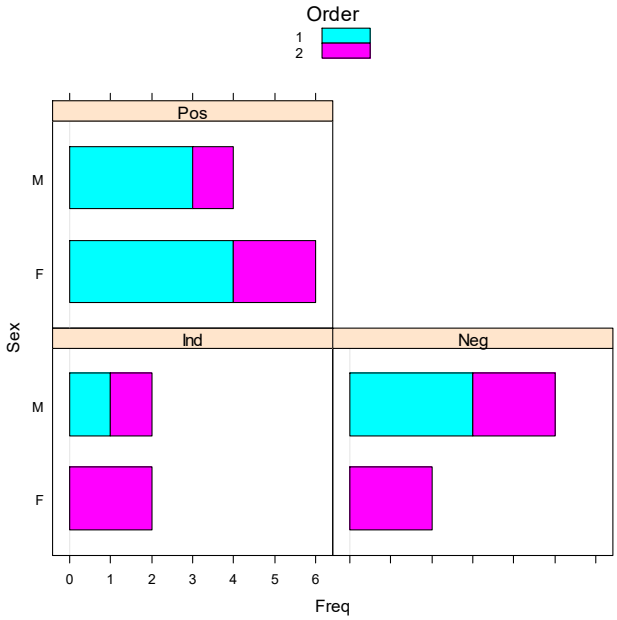


Figura 1.4.5. □

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.

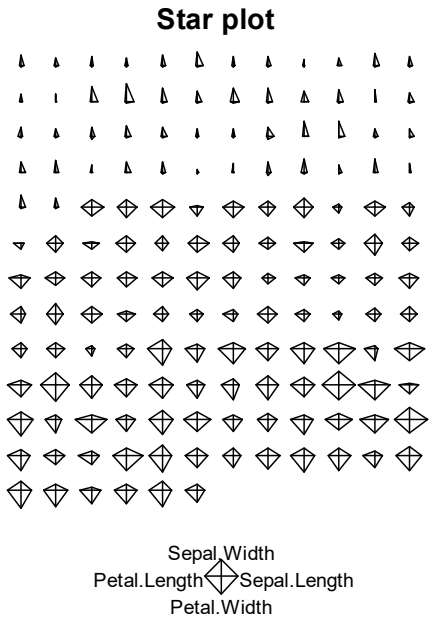


Figura 1.4.6.

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. □

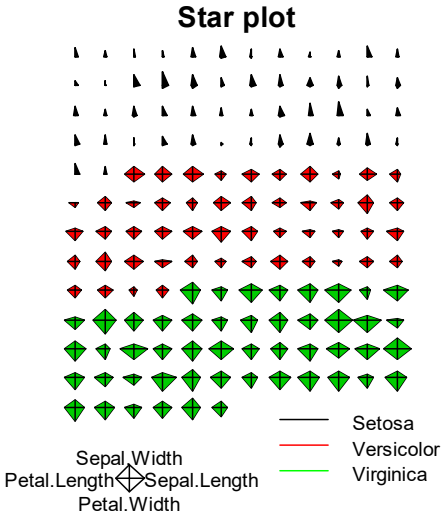


Figura 1.4.7.

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.

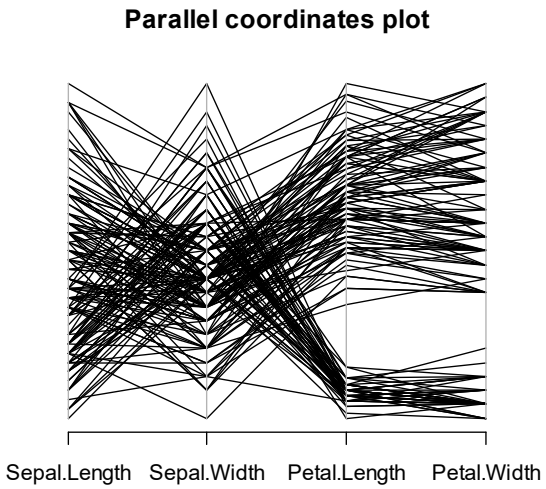


Figura 1.4.8.

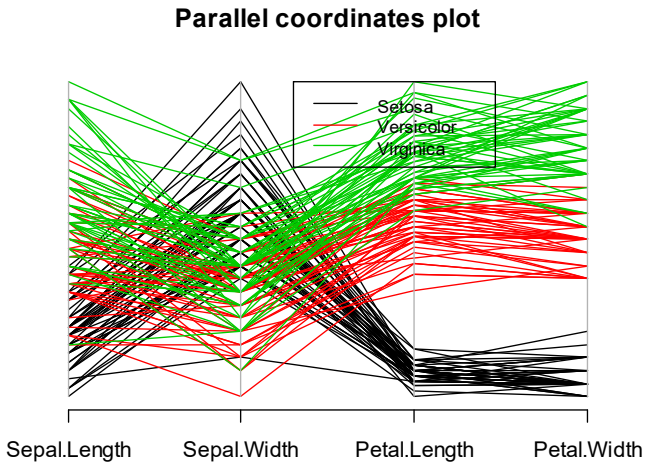


Figura 1.4.9.

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.

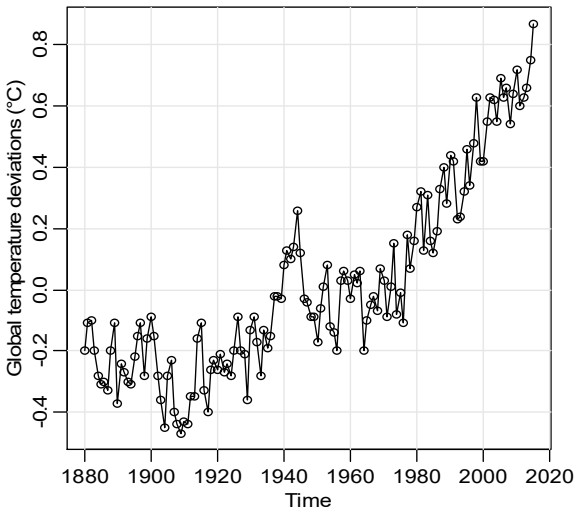


Figura 1.5.1.

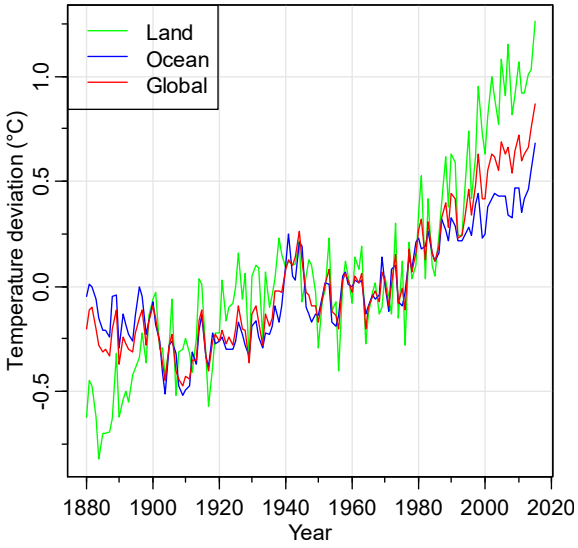


Figura 1.5.2.

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.

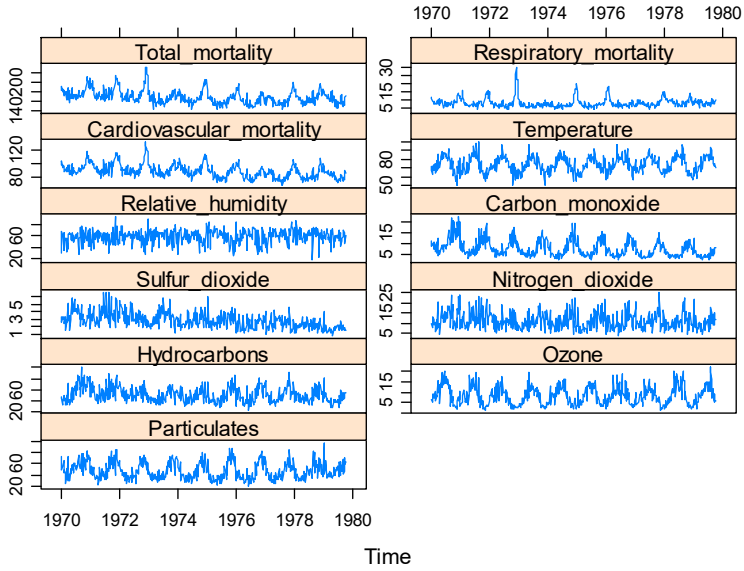


Figura 1.5.3.

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. □

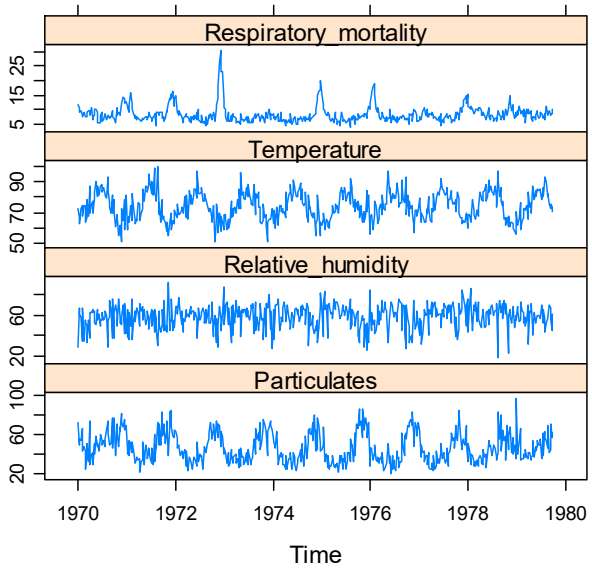


Figura 1.5.4.

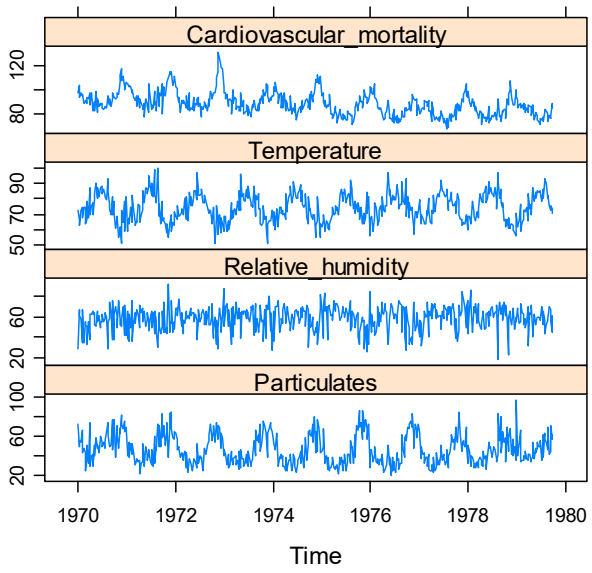


Figura 1.5.5.

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. □

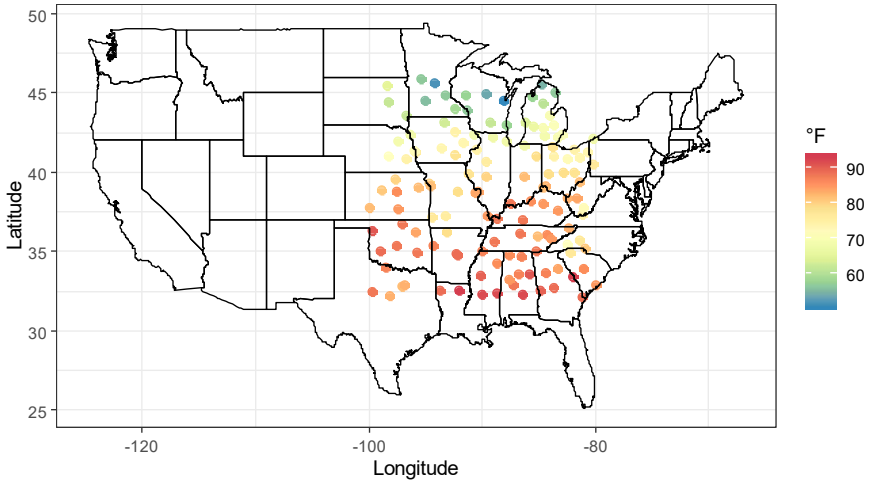


Figura 1.5.6.

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)
```

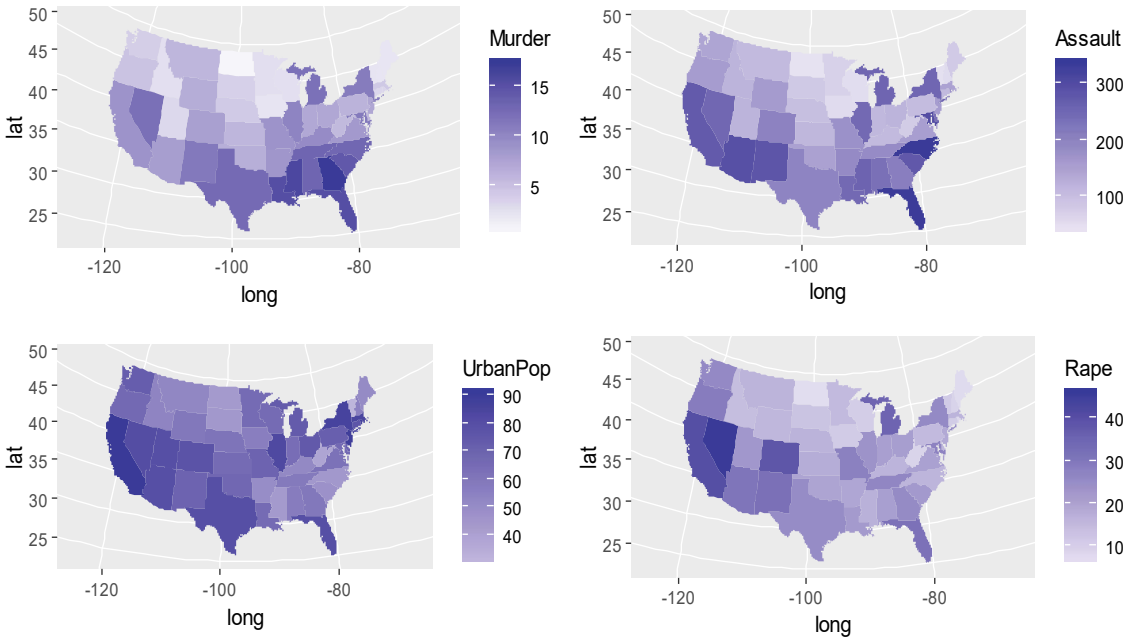


Figura 1.5.7.

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. □

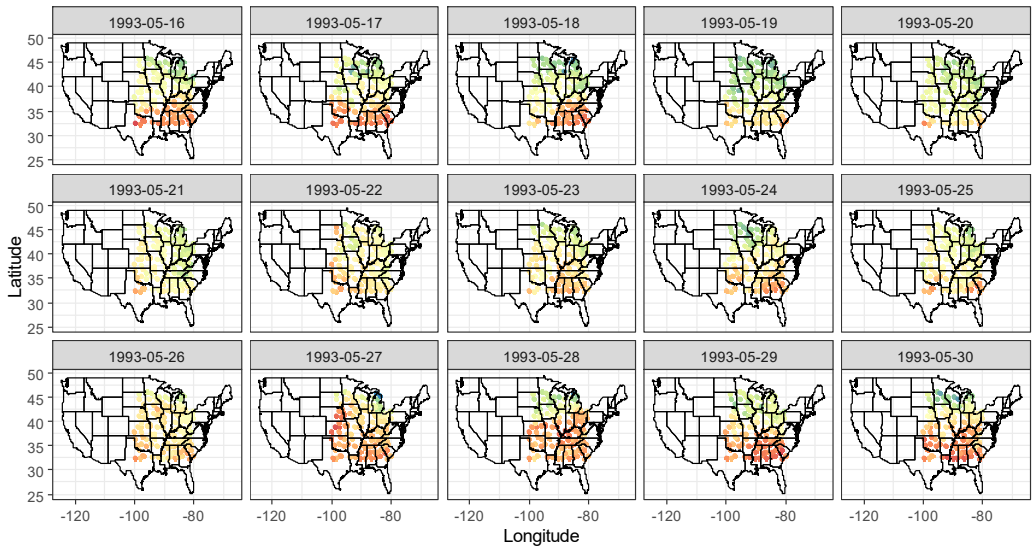


Figura 1.5.8.

● **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. □

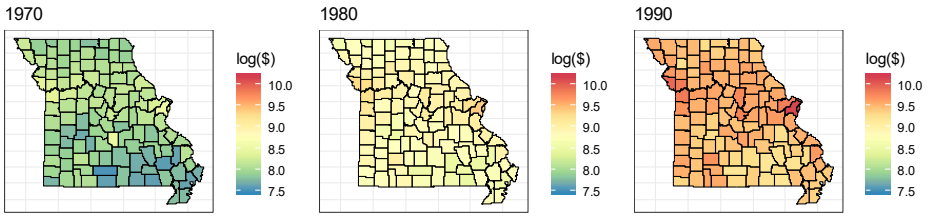


Figura 1.5.9.

1.6. Riferimenti bibliografici

Alexander, R. (2023) *Telling Stories with Data*, CRC Press, Boca Raton.

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.

Cremonini, M. (2025) *Data Visualization in R and Python*, Wiley, Hoboken.

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

Grant, R. (2019) *Data Visualization*, 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.

Irizarry, R.A. (2025) *Introduction to Data Science*, CRC Press, Boca Raton.

Maindonald, J.H., Braun, W.J. e Andrews, J.L. (2024) *A Practical Guide to Data Analysis Using R*, Cambridge University Press, Cambridge.

Jones, E., Harden, S. e Crawley, M.J. (2023) *The R Book*, terza edizione, Wiley, New York.

Kabacoff, R. (2024) *Modern Data Visualization with R*, 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.
Pearson, R.K. (2018) *Exploratory Data Analysis Using R*, CRC Press, Boca Raton.
Pebesma, E. e Bivand, R. (2023) *Spatial Data Science*, CRC Press, Boca Raton.
Pruim, R. (2018) *Foundations and Applications of Statistics*, seconda edizione, American Mathematical Society, Providence.
Rahlf, T. (2019) *Data Visualisation with R*, Springer, Cham.
Rathi, M. (2025) *Introduction to Statistical Computing and Visualization Using R*, 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.
Tulin, M. (2025) *Modern Statistics with R*, seconda edizione, CRC Press, Boca Raton.
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.
Wimberly, M.C. (2023) *Geographic Data Science with R*, CRC Press, Boca Raton.

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.

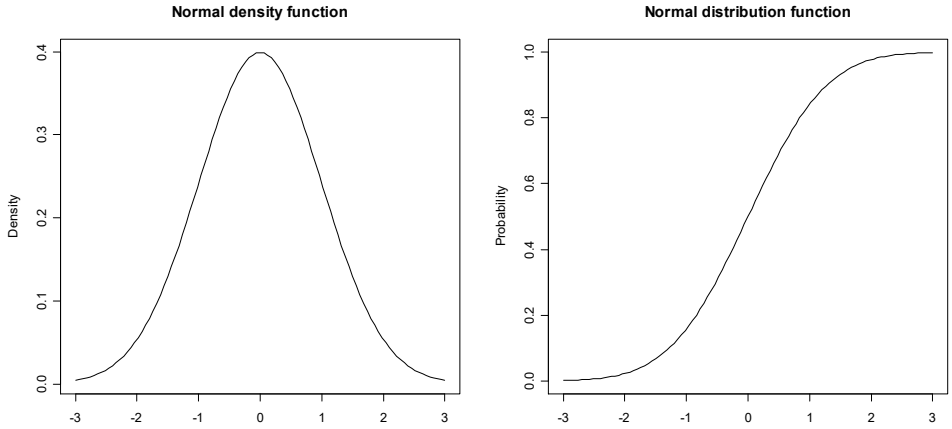


Figura 2.2.1.

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.

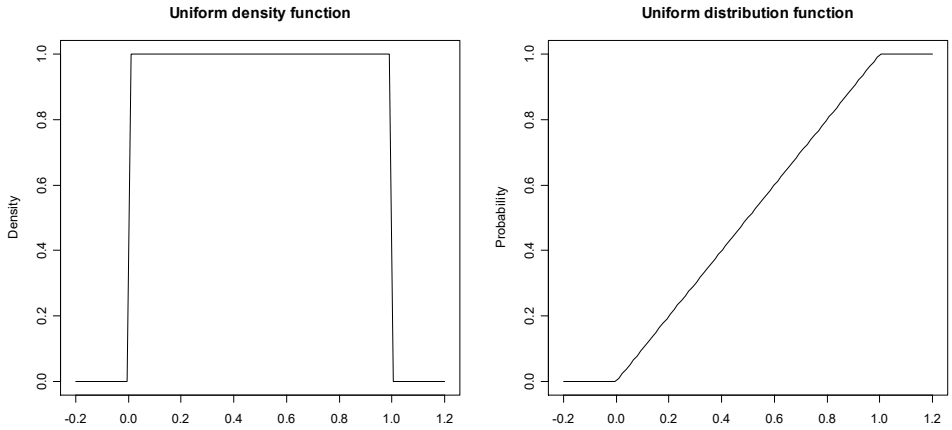


Figura 2.2.2.

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.

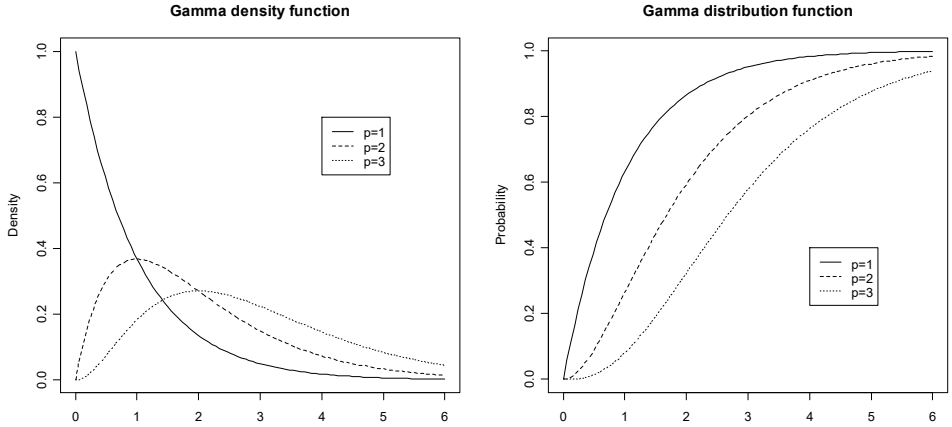


Figura 2.2.3.

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.

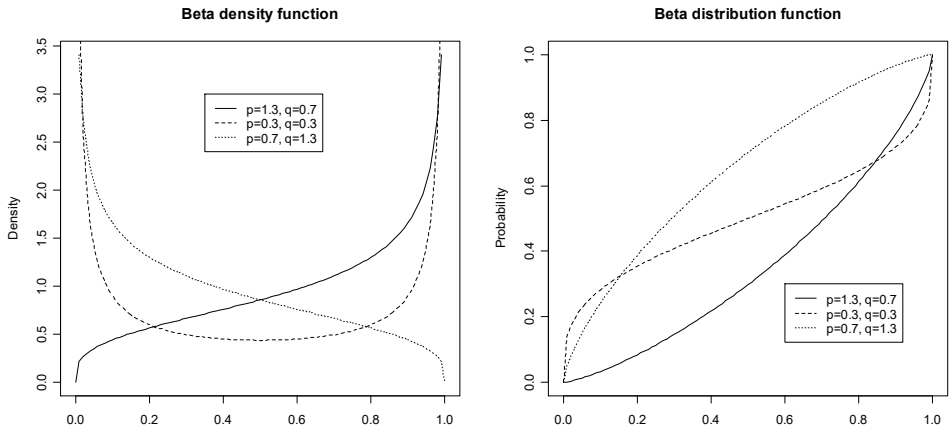


Figura 2.2.4.

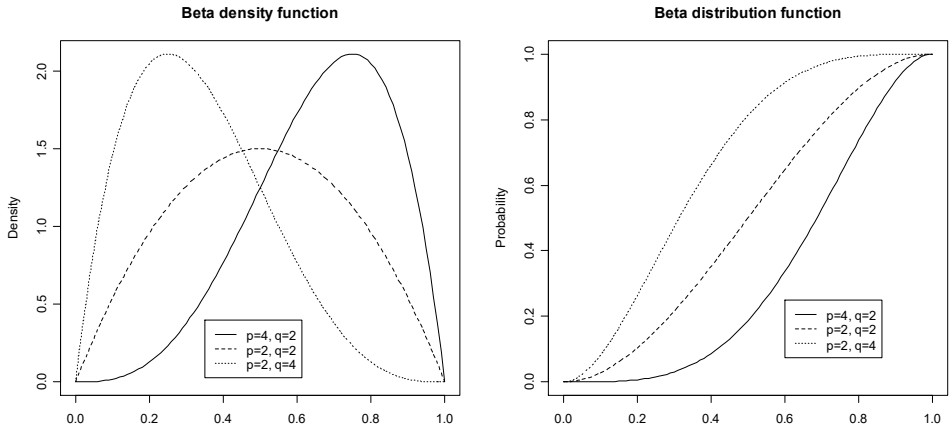


Figura 2.2.5.

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.

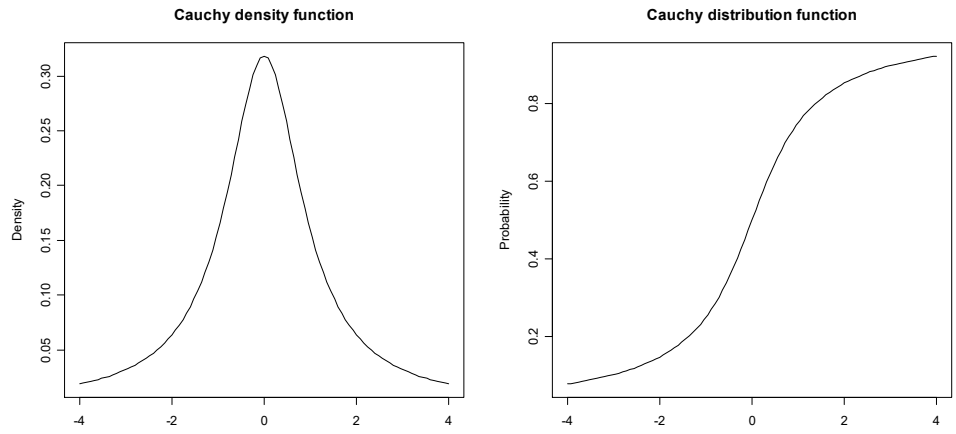


Figura 2.2.6.

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.

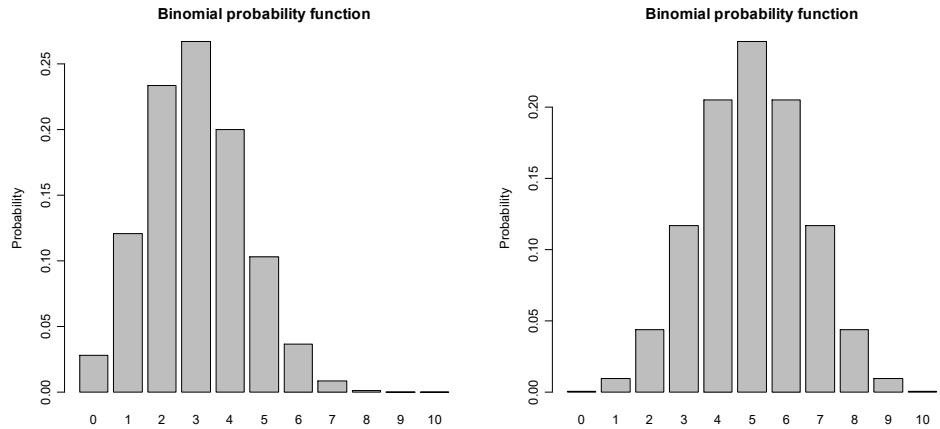


Figura 2.3.1.

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.

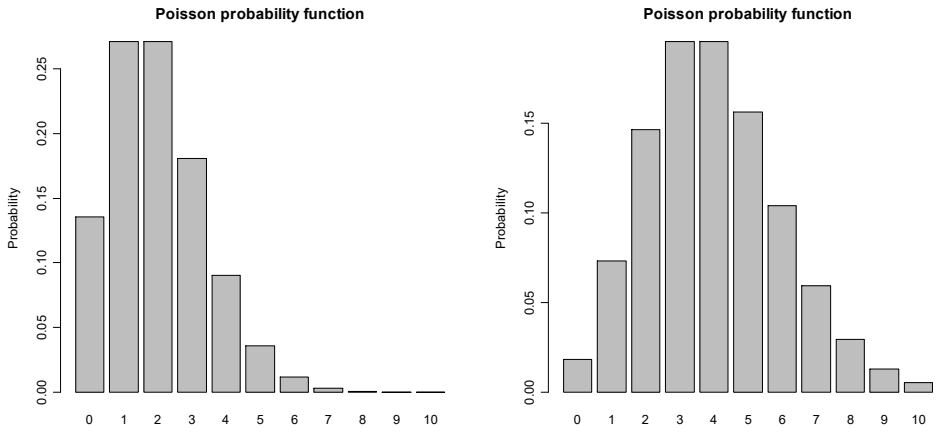


Figura 2.3.2.

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.

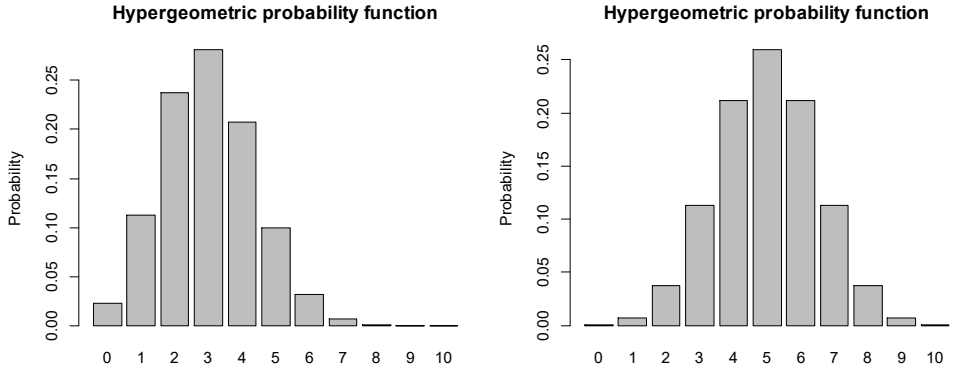


Figura 2.3.3.

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.

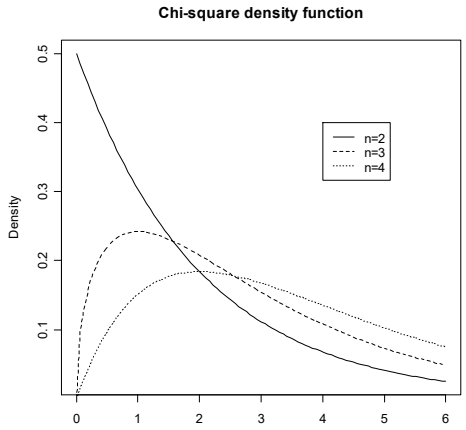


Figura 2.4.1.

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.

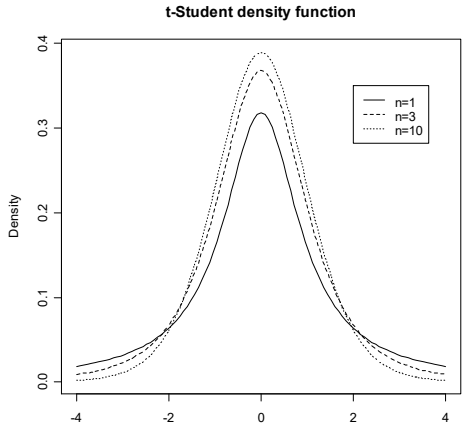


Figura 2.4.2.

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.

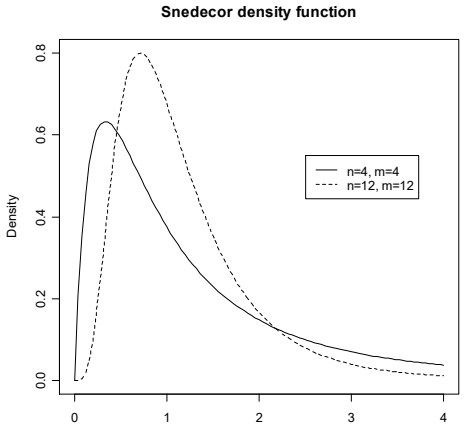


Figura 2.4.3.

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 $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.

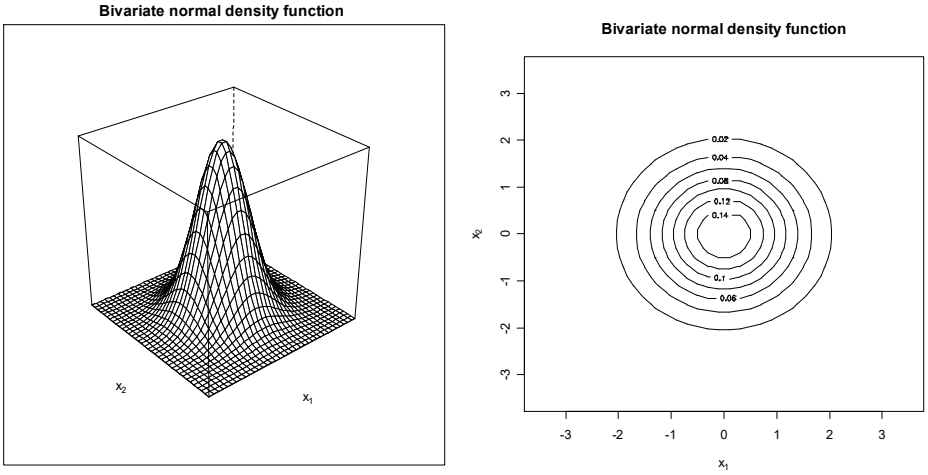


Figura 2.5.1.

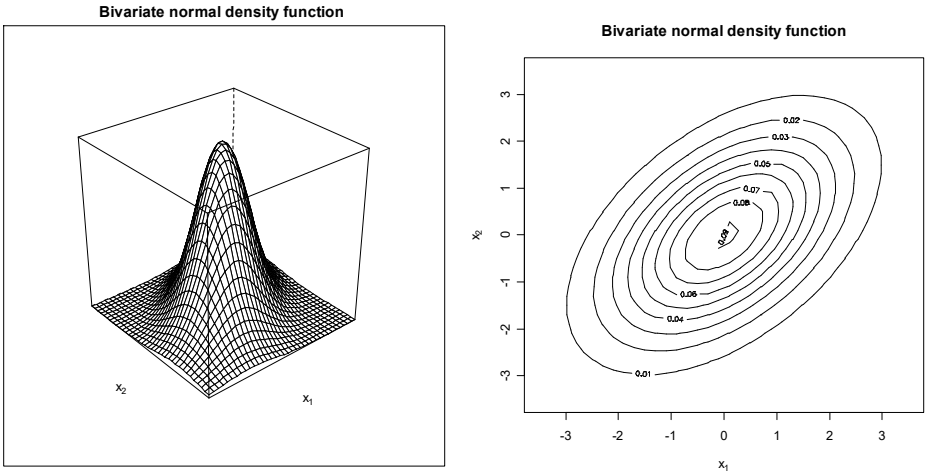


Figura 2.5.2.

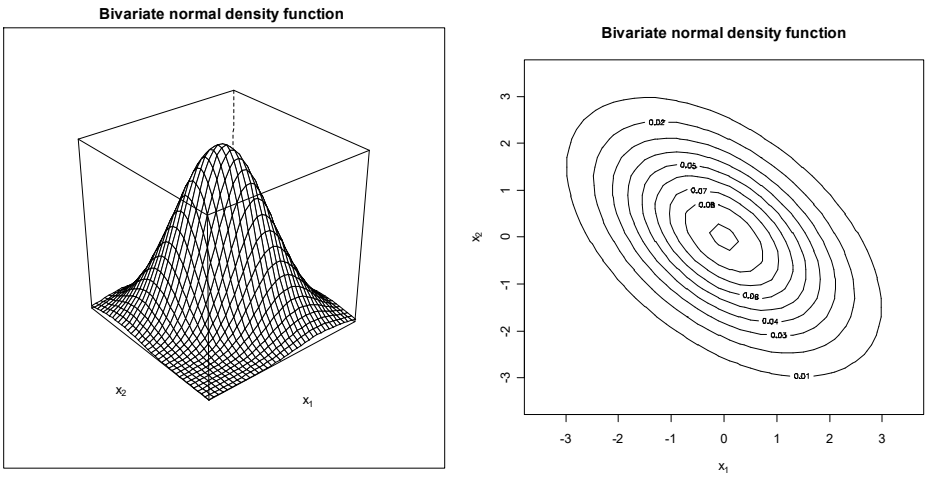


Figura 2.5.3.

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.

Deshmukh, S.R. e Kashikar, A.S. (2025) *Probability Theory*, CRC Press, Boca Raton.

Fang, K.T., Kotz, S. e 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.

Florescu, I. e Tudor, C. (2014) *Handbook of Probability*, Wiley, New York.

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

Foss, S., Korshunov, D. e Zachary, S. (2013) *An Introduction to Heavy-Tailed and Subexponential Distributions*, seconda edizione, Springer, 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.

Kroese, D.P. e Botev, Z.I. (2024) *An Advanced Course in Probability and Stochastic Processes*, CRC Press, Boca Raton.

Linde, W. (2025) *Probability Theory*, seconda edizione, de Gruyter, Berlin.

Nolan, J.P. (2020) *Univariate Stable Distributions*, Springer, Cham.

Stirzaker, D. (2003) *Elementary Probability*, Cambridge University Press, Cambridge.

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

Yao, W. e Xiang, S. (2024) *Mixture Models*, CRC Press, Boca Raton.

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

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

e

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

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 $\mathrm{E}[\bar{X}] = \sigma$ e $\mathrm{Var}[\bar{X}] = \sigma^2/n$. Dunque, risultano verificati i risultati generali visti in precedenza, in quanto per la variabile casuale Esponenziale $X \sim E(0, \sigma)$ si ha $\mathrm{E}[X] = \sigma$ e $\mathrm{Var}[X] = \sigma^2$. 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. □

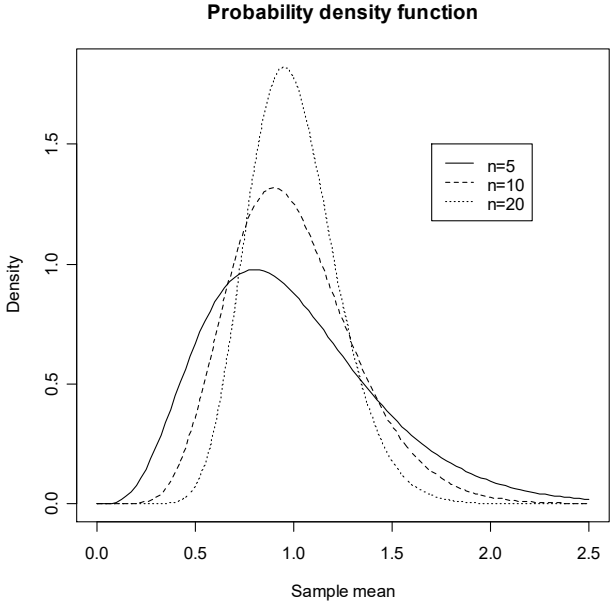


Figura 3.2.1.

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 $n \to \infty$. □

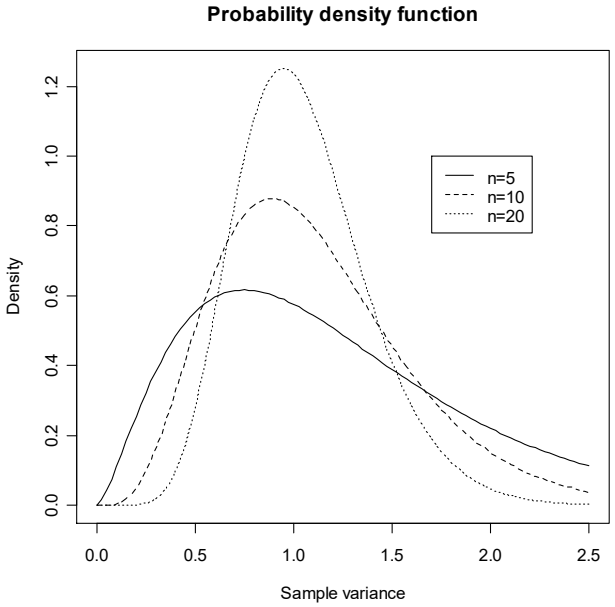


Figura 3.2.2.

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.

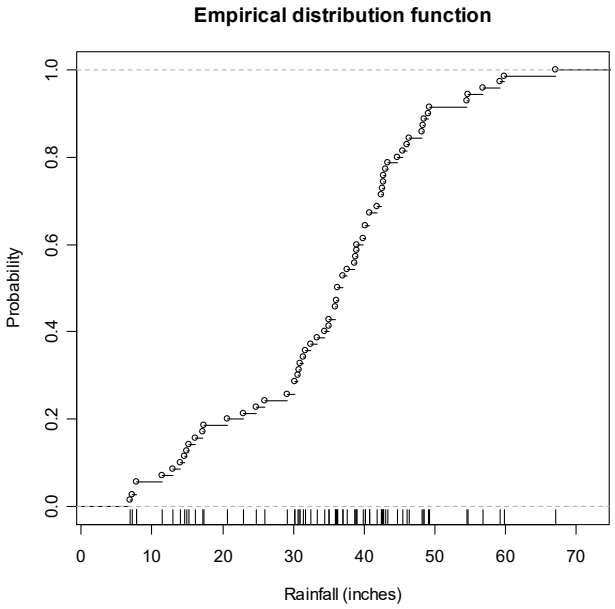


Figura 3.2.3.

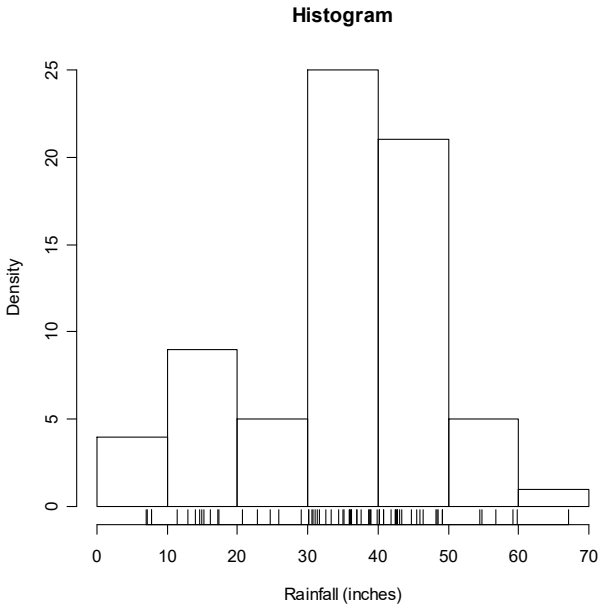


Figura 3.2.4.

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. □

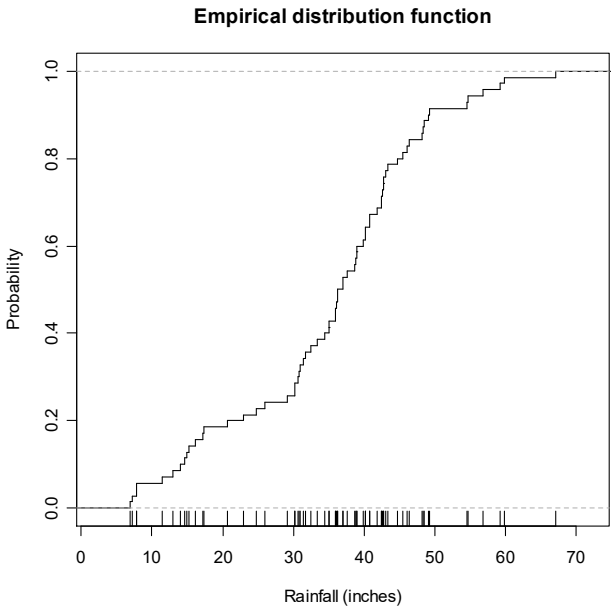


Figura 3.2.5.

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

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

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

Agresti, A. e Kateri, M. (2022) *Foundations of Statistics for Data Scientists*, CRC Press, Boca Raton.
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.
Chan, J.C.C. e Kroese, D.P. (2025) *Statistical Modeling and Computation*, seconda edizione, Springer, New York.
Cox, D.R. (2006) *Principles of Statistical Inference*, Cambridge University Press, Cambridge.
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.

Henze, N. (2024) *Asymptotic Stochastics*, Springer, Berlin.
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 $S^2_{c,x}$ è corretto per σ^2, dal momento che

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

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^2_x è uno stimatore coerente, essendo asintoticamente corretto e

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

Di nuovo, la proprietà di coerenza di S^2_x 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. □

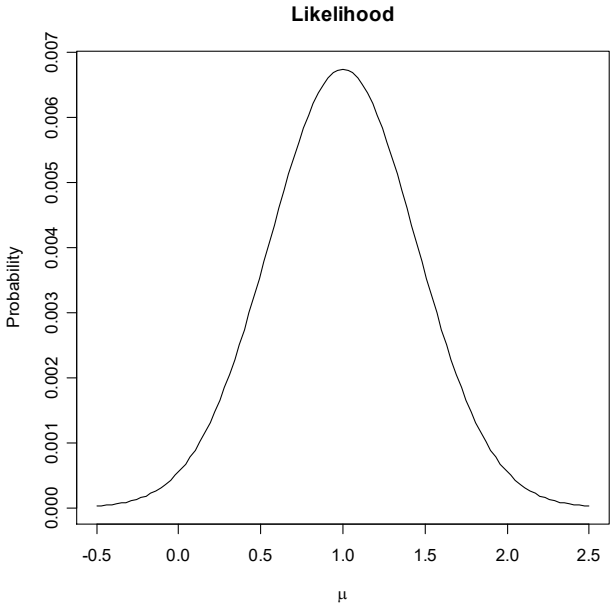


Figura 4.2.1.

• **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. □

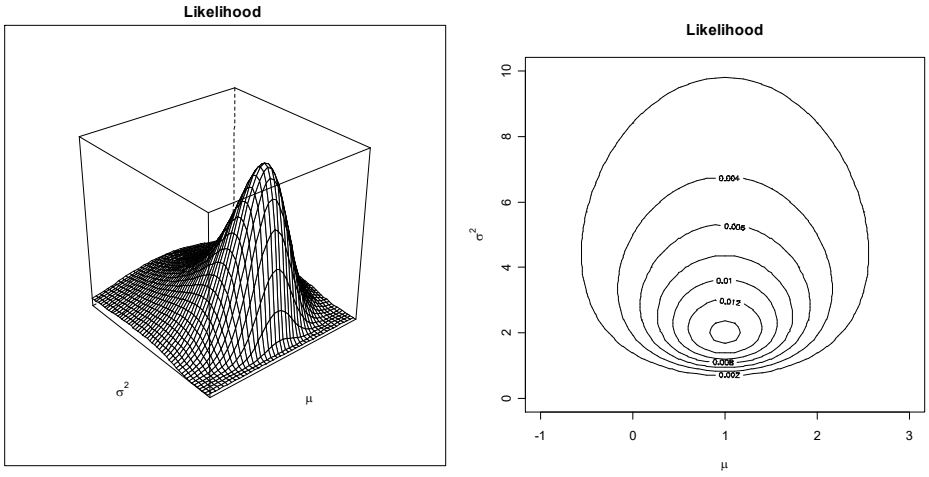


Figura 4.2.2.

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

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

Dunque, la verosimiglianza è data da

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

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

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

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 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. □

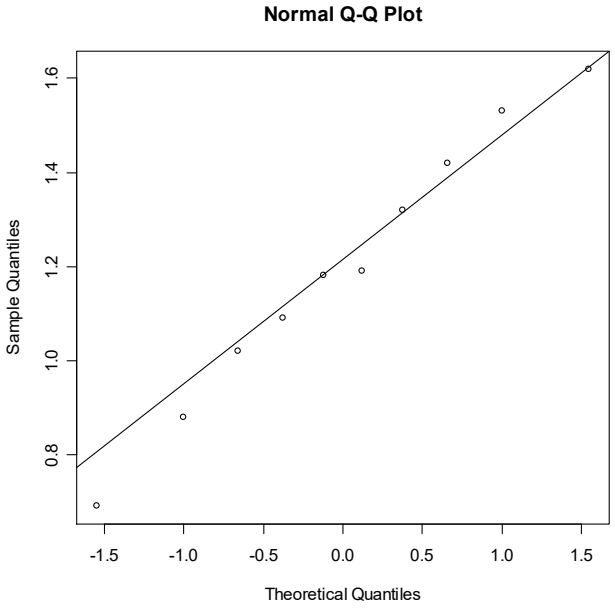


Figura 4.3.1.

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. □

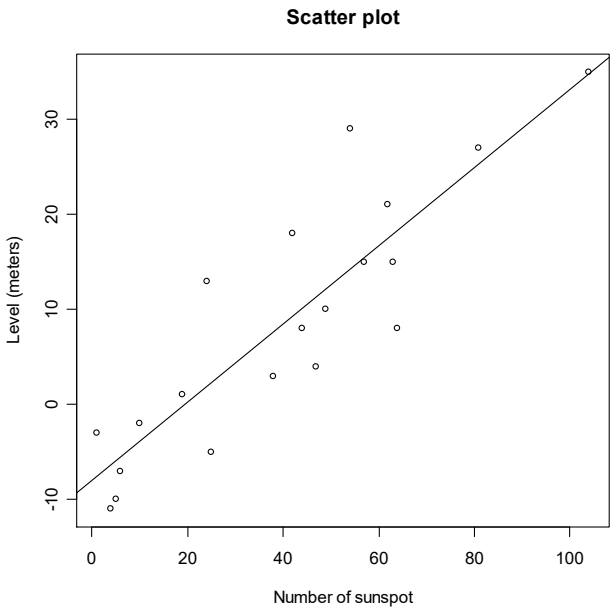


Figura 4.4.1.

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.

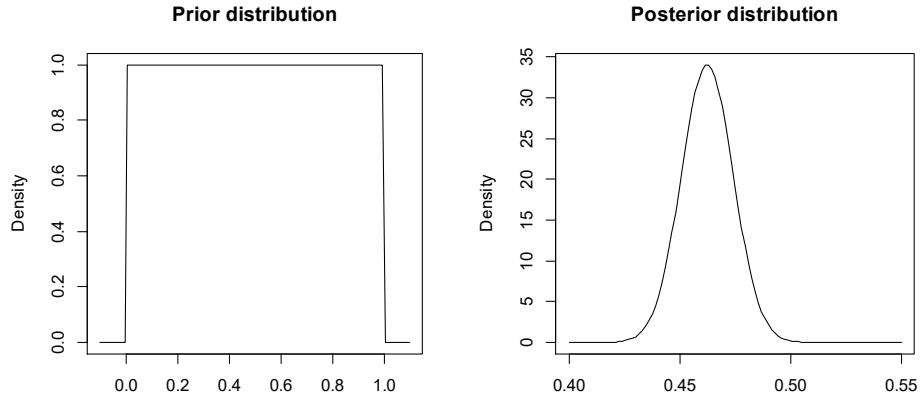


Figura 4.5.1.

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. □

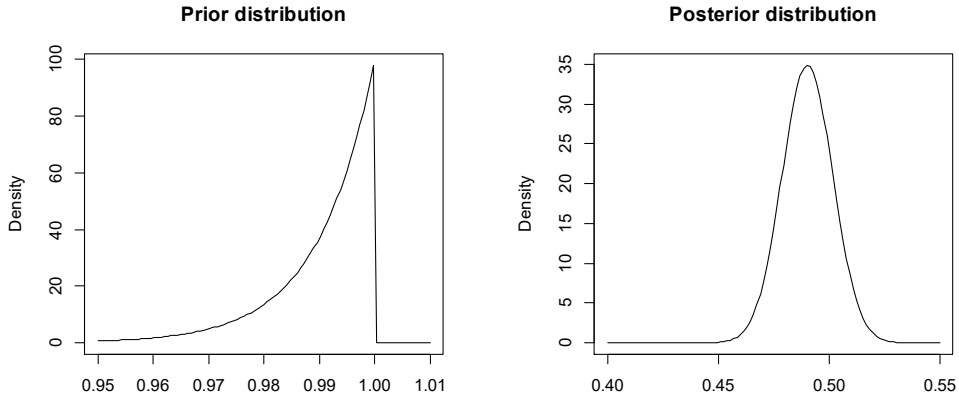


Figura 4.5.2.

4.6. Riferimenti bibliografici

Albert, J. (2009) *Bayesian Computation with R*, seconda edizione, Springer, New York.
Bhattacharya, P.K. e Burman, P. (2016) *Theory and Methods of Statistics*, Academic Press, Amsterdam.
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.
Held. L. e Bové, D.S. (2014) *Applied Statistical Inference*, Springer, Berlin.
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.
Samaniego, F.J. (2010) *A Comparison of the Bayesian and Frequentist Approaches to Estimation*, Springer, 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.
Watanabe, S. (2018) *Mathematical Theory of Bayesian Statistics*, CRC Press, Boca Raton.
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. □

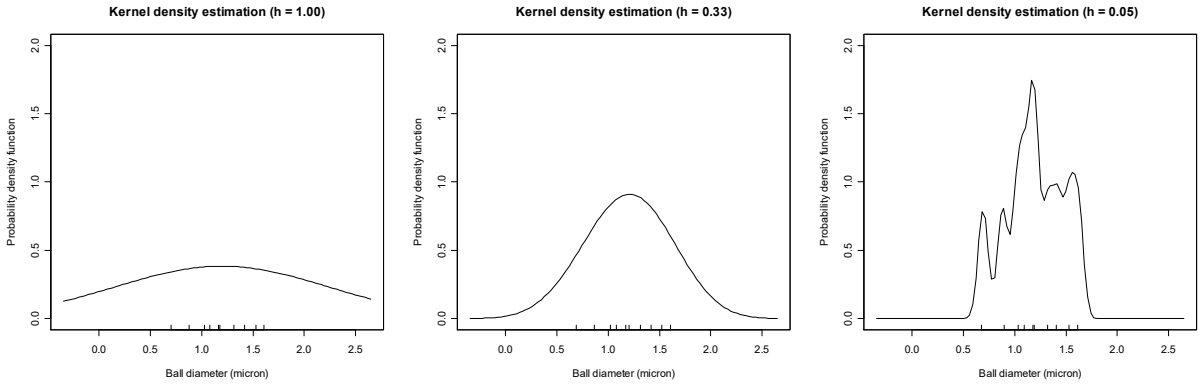


Figura 5.1.1.

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. □

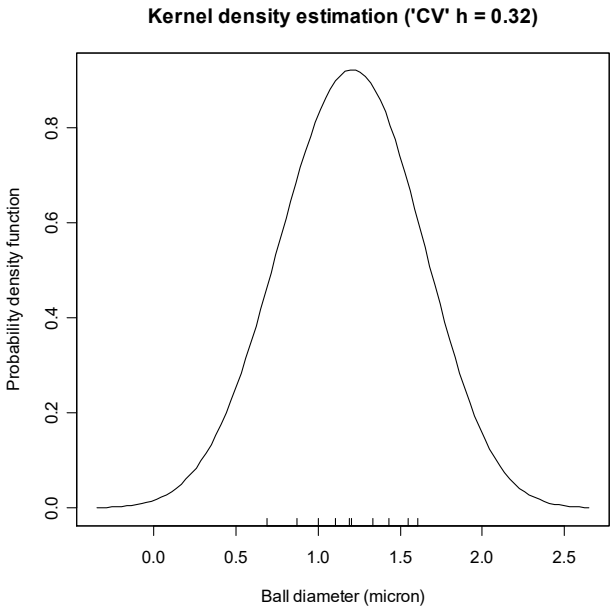


Figura 5.1.2.

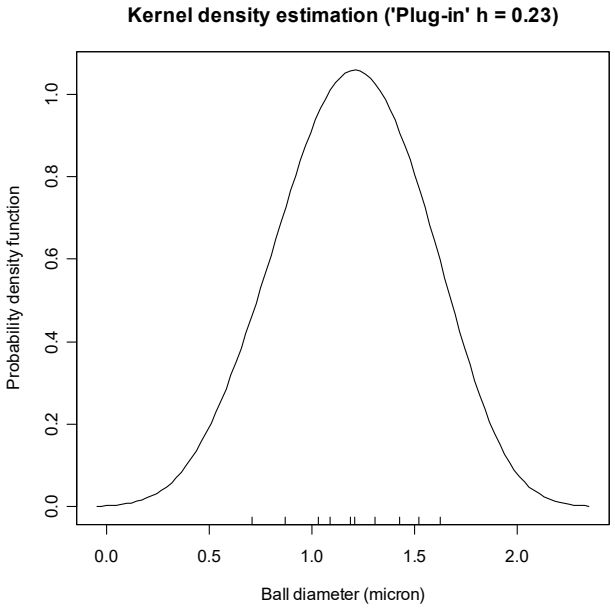


Figura 5.1.3.

• **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. □

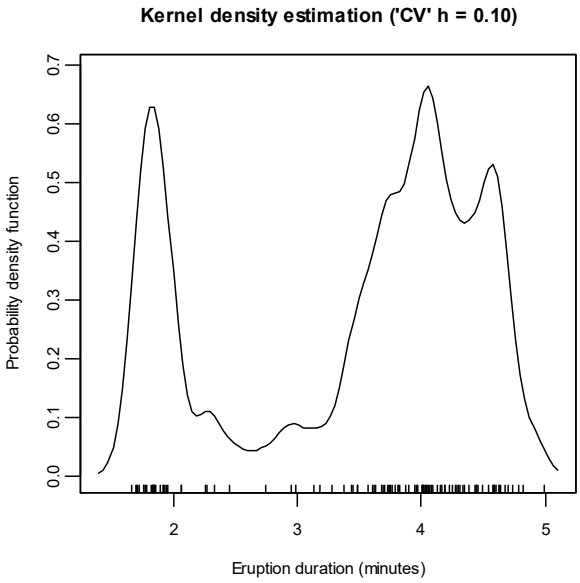


Figura 5.1.4.

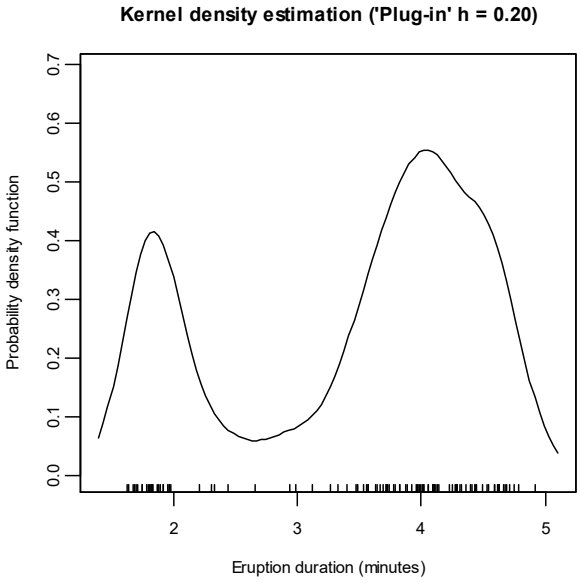


Figura 5.1.5.

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

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

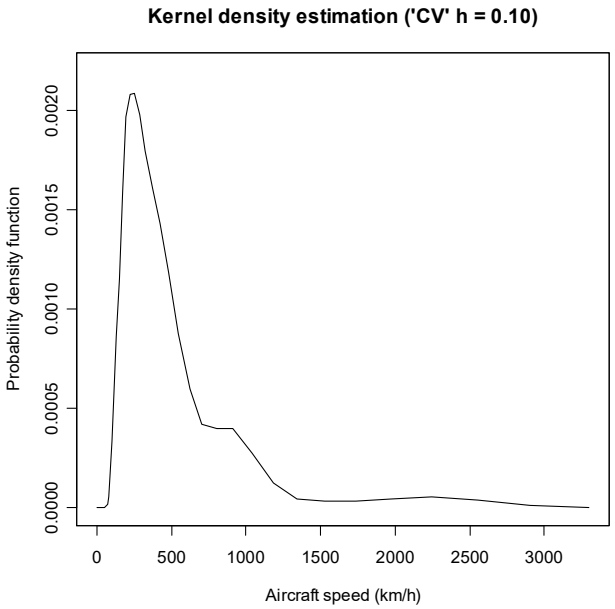


Figura 5.1.6.

• **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. □

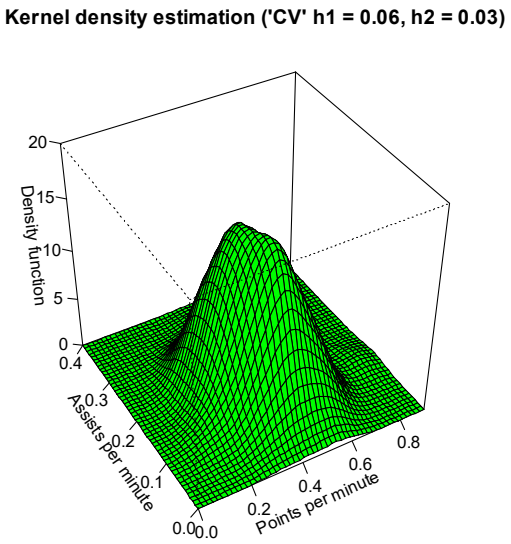


Figura 5.2.1.

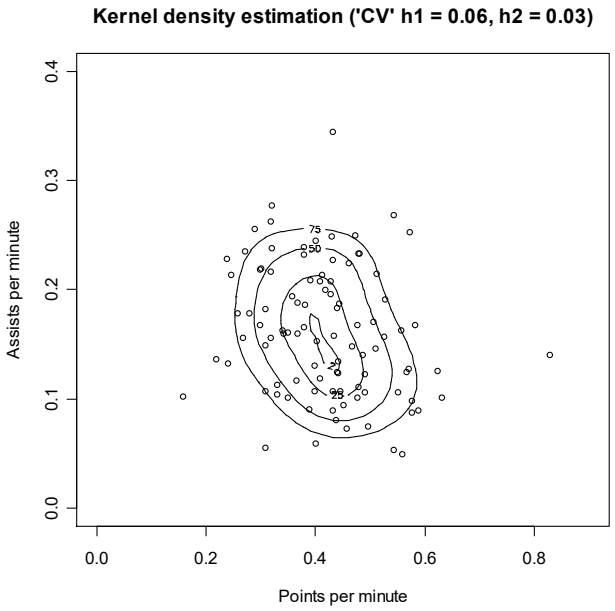


Figura 5.2.2.

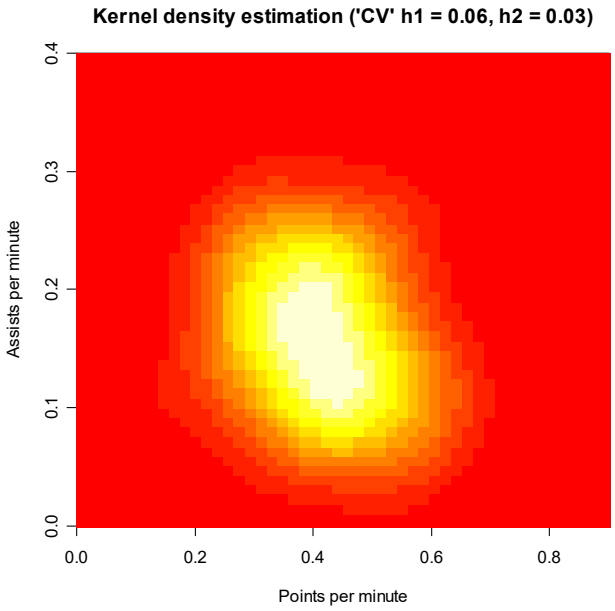


Figura 5.2.3.

• **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. □

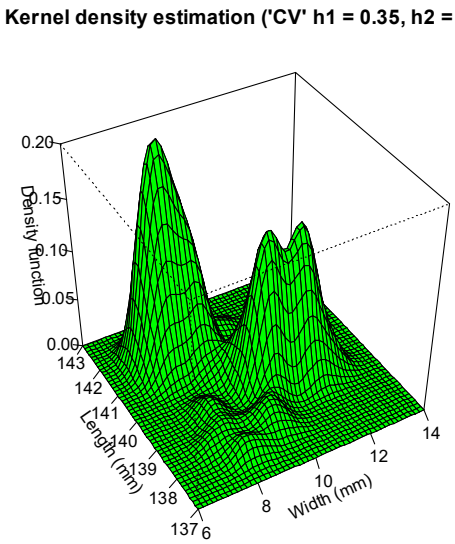


Figura 5.2.4.

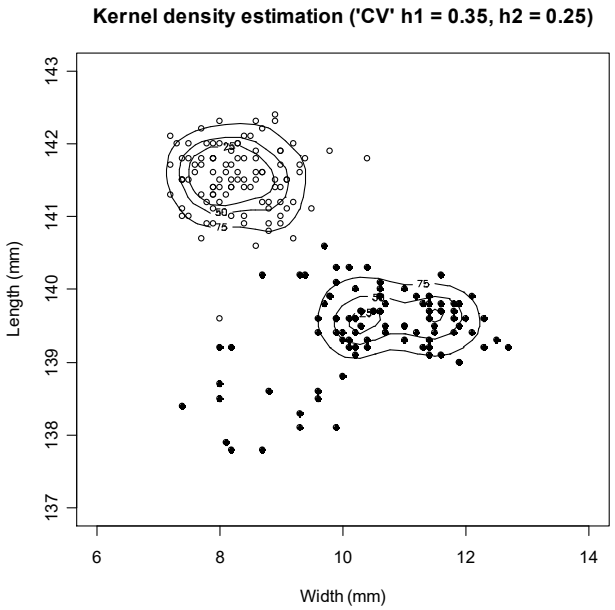


Figura 5.2.5.

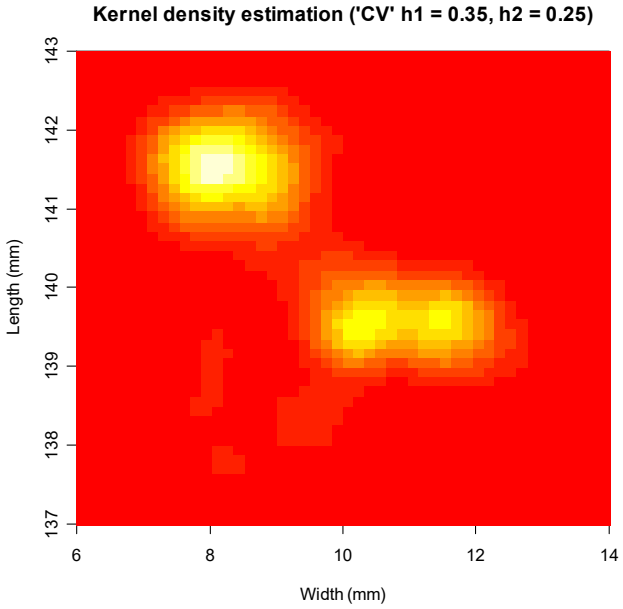


Figura 5.2.6.

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. □

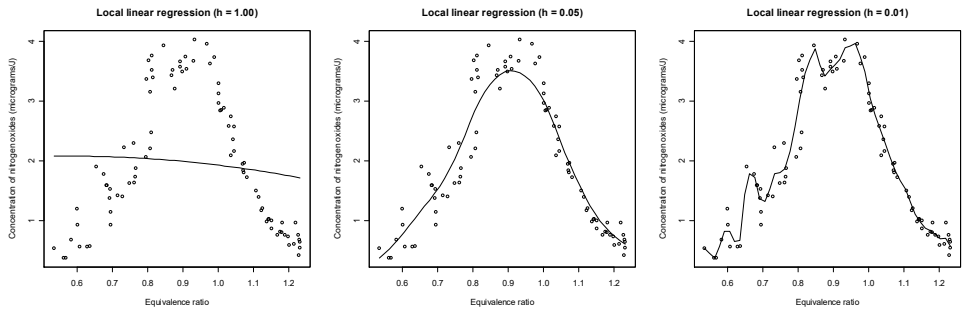


Figura 5.3.1.

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. □

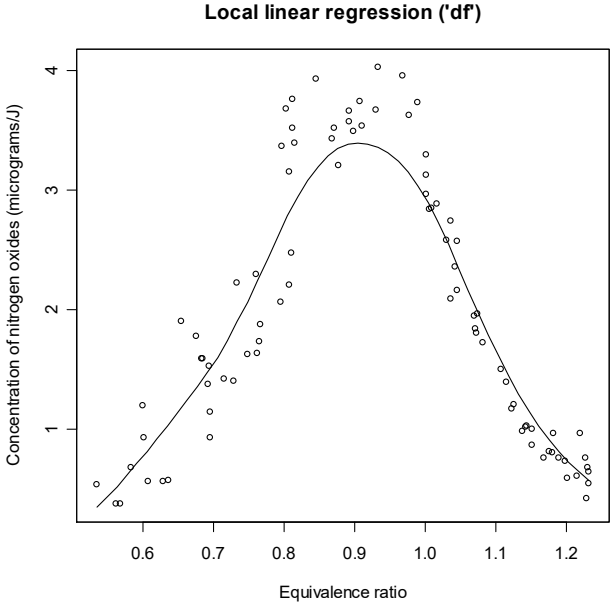


Figura 5.3.2.

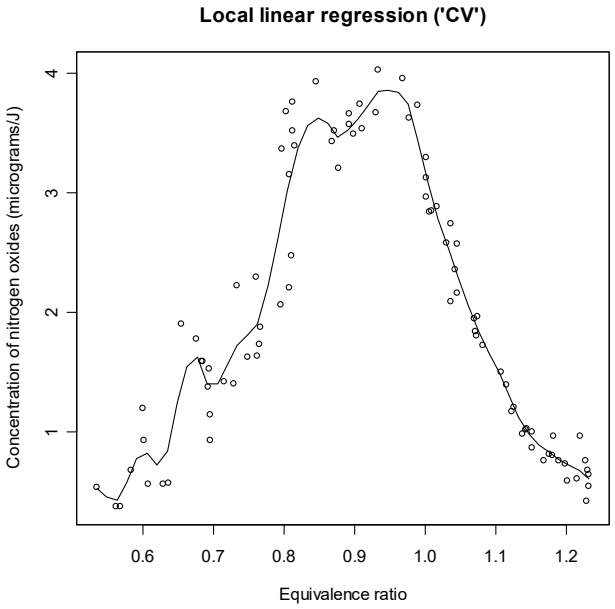


Figura 5.3.3.

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. □

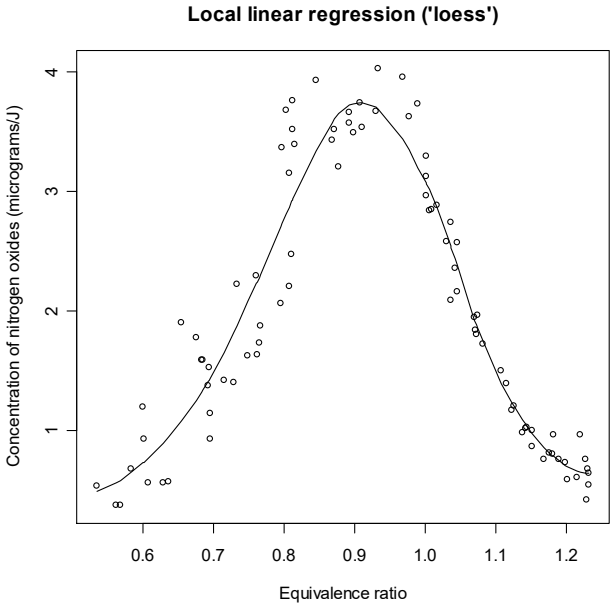


Figura 5.3.4.

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.

Fan, J. e Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, Chapman and Hall/CRC Press, Boca Raton.

Ghosh, S. (2018) *Kernel Smoothing*, Wiley, New York.

Gramacki, A. (2018) *Nonparametric Kernel Density Estimation and its Computational Aspects*, Springer, Cham.

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.

Pons, O. (2023) *Functional Estimation for Density, Regression Models and Processes*, seconda edizione, World Scientific Publishing, Singapore.
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.
Wang, Y. (2011) *Smoothing Splines*, CRC Press, Boca Raton.