Massima verosimiglianza

Francesco Lagona
Roma Tre e Max Planck Institute

1 Introduzione

I dati provenienti da rilevazioni per il monitoraggio ambientale suscitano in genere domande alle quali non possiamo dare risposta mediante semplici metodi descrittivi. Ad esempio, il grafico riportato in Fig. 1 mostra la serie storica delle concentrazioni medie di ozono rilevate a New York nel periodo maggio-settembre 1973. Per riprodurre il grafico in R, possiamo digitare

```
data(airquality)
plot(airquality[,1],type="l",xlab="",ylab="concentrazione")
title(main="Ozono: medie giornaliere",sub="New-York:
maggio-settembre 1973")
```

Si tratta di una tipica serie storica ambientale che può dare adito a diverse questioni:

- è possibile ricostruire la serie stimando i valori mancanti?
- è possibile prolungare la serie in modo da poter prevedere l'andamento dell'ozono nel futuro?
- se x_c è un valore critico oltre il quale la concentrazione di ozono può provocare danni alla salute, è possibile stimare quante volte tale soglia sarà superata nel corso di un anno?
- è sufficientemente lunga la serie riportata nella Fig. 1 per rispondere alle domande precedenti?

Per rispondere a domande come queste sarebbe necessaria una completa e certa conoscenza del **processo generatore** dei dati. Se ad esempio fosse nota una funzione deterministica del tempo, diciamo $x_t = f(t)$, che per ogni istante t produca la concentrazione di ozono x_t, tutte le questioni enunciate (e molte altre) troverebbero immediata risposta.

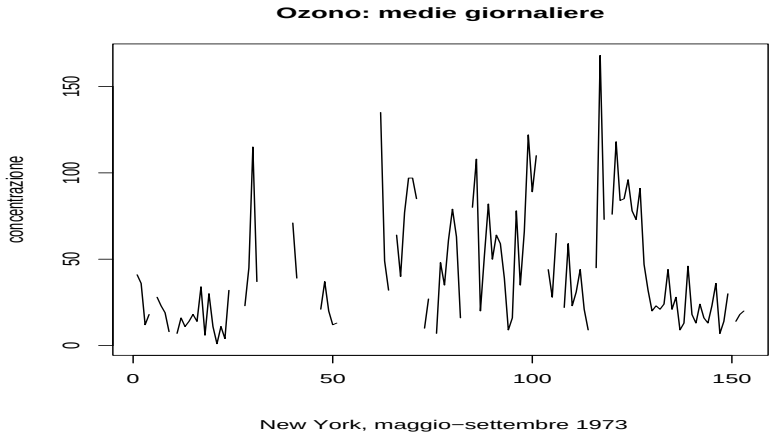

Figura 1: : serie storica delle concentrazioni medie giornaliere dell'ozono da maggio a settembre 1973 a New York (fonte: dataset R `airquality`); si noti la presenza di dati mancanti.

Una funzione deterministica è raramente disponibile. In casi particolarmente fortunati, le nostre informazioni a priori sul fenomeno oggetto di studio ci mettono tuttalpiù in grado di affermare che le osservazioni sono generate da una famiglia $\mathcal{F}$ di funzioni deterministiche ed il processo generatore dei dati può essere definito come quella funzione $f \in \mathcal{F}$ che riproduce la serie osservata.

Nella maggior parte delle applicazioni, comunque, le informazioni a nostra disposizione consentono di specificare una famiglia di **processi aleatori**, poichè la variabilità dei dati supera di gran lunga la variabilità che siamo in grado di spiegare in modo deterministico.

Un processo aleatorio è un oggetto matematico in grado di riprodurre osservazioni con la variabilità desiderata. Guardare ai dati osservati come ad una serie generata da un processo aleatorio può sembrare un'inutile complicazione del problema di individuazione del processo generatore dei dati. In realtà, invece, l'uso dei processi aleatori comporta un'enorme semplificazione: invece di andare alla ricerca di una funzione deterministica che riproduca esattamente le osservazioni, ci accontentiamo di identificare un processo aleatorio che genera serie di dati che non riproducono fedelmente le osservazioni , ma che condividono con la serie osservata alcune caratteristiche d'interesse.

L'idea può essere illustrata dal seguente semplice esempio. Supponiamo di aver lanciato una moneta 10 volte, ottenendo la serie di osservazioni

$$(T, C, C, T, C, T, T, T, C, C).$$

Supponiamo inoltre che siamo interessati a determinare l'onestà della moneta, ovvero desideriamo sapere se la moneta ha uguale propensione a generare teste e croci. Una risposta certa a tale domanda prevede una conoscenza deterministica del processo che la moneta usa per generare teste e croci. Se abbandoniamo la speranza di trovare un tale processo, possiamo porre il problema nei seguenti termini.

Supponiamo di poter disporre di un meccanismo, controllato da un parametro θ, in grado di simulare la successione di teste e croci generata da una moneta con le seguenti caratteristiche:

- la successione di teste e croci è completamente casuale, nel senso che qualunque sia la lunghezza della successione simulata, la nostra capacità di prevedere il risultato di lanci successivi rimane costante;
- la frequenza relativa è sempre meno variabile all'aumentare della lunghezza della successione ed è convergente al valore assunto dal parametro θ.

In R, un meccanismo di tale tipo è fornito dal comando

```
x<-rbinom(n,1,theta)
```

che genera una successione binaria x di lunghezza n, dove gli 0 possono essere interpretati come le croci e gli 1 come teste e dove il parametro θ controlla la propensione della moneta a generare teste (se $\theta = 0.5$, la moneta è onesta). Ripetendo il comando un numero di volte, ci si accorge che esso estrae casualmente ogni volta una successione dall'insieme $\{0,1\}^n$. Usando il comando con n elevato (per esempio $n > 50$), ci si accorge che la frequenza relativa degli uno è ben approssimata dal valore di θ che è stato inserito come input.

Un meccanismo del genere può essere rappresentato dalla famiglia parametrica $\mathcal{F} = \{f_\theta, \theta \in (0,1)\}$, dove $f_\theta =$`rbinom(n,1,theta)`. Se assumiamo che la serie che abbiamo osservato è stata generata da un membro della famiglia $\mathcal{F}$, la riposta al quesito sull'onestà della moneta si riduce al problema di **stimare** il valore di θ sulla base dei dati osservati.

L'enorme semplicazione del problema è la conseguenza dell'uso della famiglia $\mathcal{F}$, che fornisce una rappresentazione del processo generatore dei dati, distinguendo la componente d'interesse per rispondere al quesito sull'onestà della moneta (il parametro θ) da tutte le altre caratteristiche della serie, di cui bisogna certo tener conto, ma che non costituiscono l'oggetto diretto dell'analisi. Tali caratteristiche includono ad esempio l'ordine con cui si susseguono le teste e le croci, la comparsa di gruppi di teste e croci successivi, ... in una parola, la **variabilità intrinseca** della serie osservata.

La famiglia $\mathcal{F}$ è spesso chiamata **modello probabilistico** e la sua specificazione dipende strettamente dal problema di analisi con cui ci confrontiamo. Non esistono ricette facili da seguire per la sua definizione: ogni ricercatore acquisisce negli anni l'esperienza necessaria a capire come modellare i dati che via via gli si presentano.

Se però il modello probabilistico è stato definito, esistono ricette elaborate dalla statistica matematica per risolvere il problema dell'uso ottimale delle osservazioni ai fini dell'individuazione del membro della famiglia $\mathcal{F}$ che ha generato i dati.

Tra le diverse metodologie disponibili in letteratura, il più popolare criterio di estrazione delle informazioni contenute nei dati si basa sul **principio di verosimiglianza** che forma l'oggetto di queste note.

Nel seguito, illustreremo l'applicazione di tale principio ad un problema estremamente semplice (par.2), con riferimento alla serie dell'ozono riportata nella Fig. 1, dove otterremo uno stimatore di massima verosimiglianza per la frequenza annuale dei superamenti di una soglia critica. Le principali caratteristiche dello stimatore di massima verosimiglianza verranno dunque illustrate nelle sezioni successive con riferimento a tale esempio. Seguirà un paragrafo sulla stima di massima verosimiglianza nel caso generale.

2 Un modello parametrico per il superamento della soglia critica

Continuando a far riferimento alle concentrazioni di ozono, supponiamo di essere interessati al superamento di una soglia critica x_c nel corso di un anno. Più precisamente, supponiamo di essere interessati a stimare la percentuale θ del numero di giorni in cui tale soglia viene superata in un anno.

Se avessimo le osservazioni relative all'intero anno solare, il valore di θ potrebbe essere valutato senza errore. Se tuttavia la nostra serie storica è parziale, il problema della stima di θ può essere tradotto in termini inferenziali secondo il seguente procedimento:

- si individua una famiglia di processi $\mathcal{F}$ che potrebbero aver generato la serie osservata
- si usa la serie osservata per scegliere un membro di tale famiglia

In questa sezione ci occupiamo di individuare una semplice famiglia di processi, mentre nella sezione successiva ci occuperemo di individuare un criterio di scelta.

Per ogni giorno t, sia X_t una variabile aleatoria binaria che assume il valore 0 con probabilità $P(X_t = 0) = 1 - \theta$ ed il valore 1 con probabilità pari a θ, $\theta \in \Theta = (0, 1)$. In altri termini

$$p_t(x) = P(X_t = x) = \theta^x (1 - \theta)^{1-x}. \tag{1}$$

Con 0 abbiamo qui codificato l'evento "non superamento della soglia", mentre con l'evento 1 si è codificato l'evento "superamento della soglia". La distribuzione (1) è nota come distribuzione bernoulliana: il valore atteso di X_t è dato da $\mathbb{E}X_t = \theta$ mentre la sua varianza è $\mathbb{V}X_t = \theta(1 - \theta)$.

Sia la successione $(X_t, t \geq 1)$ il processo generatore dei superamenti di soglia durante una sequenza di giorni. Se assumiamo che (ipotesi semplificatrice) il superamento della soglia al giorno t non dipende dagli eventi accaduti nei giorni precedenti, allora la probabilità di osservare una sequenza binaria $\boldsymbol{x} = (x_1 \dots x_T)$ è semplicemente data dal prodotto

$$\begin{aligned} p_{1\dots T}(\boldsymbol{x}) &= \prod_{t=1}^{T} \theta^x (1 - \theta)^{1-x} \\ &= \theta^{T\bar{x}} (1 - \theta)^{T - T\bar{x}} \end{aligned}$$

dove $\bar{x} = T^{-1} \sum_{t=1}^{T} x_t$ indica la frequenza relativa dei superamenti di soglia nel periodo $(1, T)$.

La terna

$$(\{0, 1\}^T, p(\boldsymbol{x}; \theta), (0, 1)) \tag{2}$$

si chiama modello probabilistico parametrico. L'insieme prodotto $\{0, 1\}^T$ si chiama **spazio campionario** e contiene tutte le possibili sequenze $\boldsymbol{x}$ osservabili con probabilità $p(\boldsymbol{x}; \theta)$, la quale dipende dal parametro θ di interesse. Infine $(0, 1)$ è lo **spazio di ammissibilità** parametrico e raccoglie i valori che potrebbe assumere θ. Si osservi che, assumendo un intervallo aperto $(0, 1)$, stiamo ipotizzando che $\theta \neq 0, 1$, ovvero che siamo certi che ci saranno almeno due osservazioni differenti durante l'anno. La terna (2) descrive una famiglia di processi generatori dei dati indicizzata dal parametro di interesse e formalizza l'ipotesi che i dati siano generati da un processo stocastico noto a meno del valore θ, da stimare sulla base dei dati.

3 Massima verosimiglianza

La formulazione di un modello parametrico riduce il problema inferenziale alla stima del valore assunto θ. Per scegliere tale valore sulla base dei dati

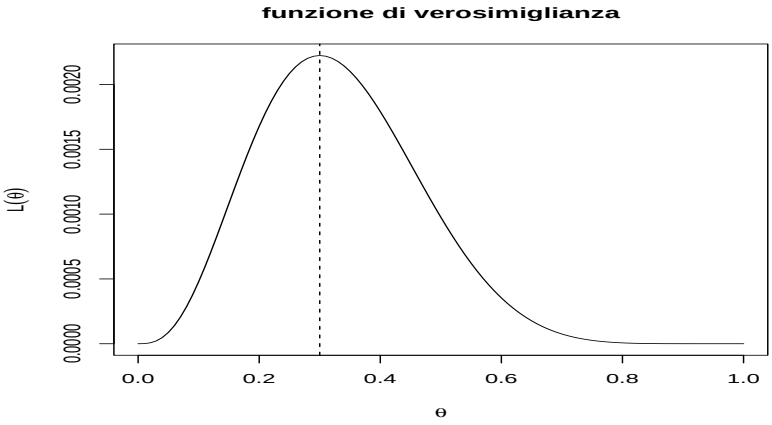

Figura 2: : funzione di verosimiglianza relativa ad una sequenza osservata di 10 giorni con 3 superamenti di soglia

osservati, possiamo ragionare come segue. Supponiamo, per fissare le idee, che

$$\boldsymbol{x}_0 = (0, 1, 0, 1, 0, 0, 1, 0, 0, 0)$$

sia la sequenza osservata. La probabilità di osservare ciò che è stato effettivamente osservato è data da

$$p(\boldsymbol{x}_0) = \theta^3(1-\theta)^7.$$

Se guardiamo a tale probabilità come ad una funzione di θ, ci accorgiamo che si tratta di una funzione positiva con un unico punto di massimo raggiunto in corrispondenza di 0.3 (Fig.2). Il grafico può essere riprodotto in R mediante i seguenti comandi:

```
theta<-seq(0,1,by=0.01) plot(theta,theta^(3)*(1-theta)^(7),

+ xlab=expression(theta),ylab=expression(L(theta)),type="l")
abline(v=0.3,lty=2) title(main="funzione di verosimiglianza")
```

Il fatto che tale funzione non si annulli mai in $(0, 1)$ indica che sulla base della sequenza osservata, nessun valore di θ può essere a priori escluso. Tuttavia, la funzione indica che esistono valori di θ che rendono la probabilità di osservare la sequenza osservata estremamente bassa, mentre altri valori (nell'intorno di 0.3) rendono tale probabilità estremamente alta: questi ultimi sono allora valori "più verosimili" e 0.3 è il valore più verosimile tra tutti. La funzione di θ ottenuta considerando la probabilità di osservare i

dati effettivamente osservati si chiama **funzione di verosimiglianza** perchè attribuisce un peso di verosimiglianza a ogni valore di θ, sulla base delle osservazioni.

Se pensiamo che quello illustrato sia un ragionamento efficace, possiamo allora enunciare il seguente principio:

> tutte le informazioni contenute nella sequenza osservata e utili per diminuire la nostra incertezza su θ sono contenute nella funzione di verosimiglianza e la stima di massima verosimiglianza di θ è il punto di massimo della funzione di verosimiglianza.

Nel nostro esempio, se $\boldsymbol{x}$ è un campione di T osservazioni, la funzione di verosimiglianza è data da

$$L(\theta) = \theta^{T\bar{x}}(1-\theta)^{T-T\bar{x}}.$$

Per motivi esclusivamente tecnici che saranno via via più chiari, è conveniente lavorare con la funzione di **log-verosimiglianza**:

$$l(\theta) = T\bar{x}\log\theta + (T - T\bar{x})\log(1-\theta).$$

Essendo il logaritmo una funzione monotona (e sotto l'ipotesi che $\theta \in (0,1)$), il punto di massimo di $l(\theta)$ coincide con quello di $L(\theta)$. Tale punto di massimo è lo stimatore di massima verosimiglianza ed è una funzione dei dati campionari:

$$\begin{aligned}\hat{\theta}(\boldsymbol{x}) &= \underset{\theta\in\Theta}{\operatorname{argmax}} L(\theta) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \log L(\theta) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} l(\theta)\end{aligned}$$

In altre parole, $\hat{\theta}(\boldsymbol{x})$ è la soluzione dell'equazione

$$\frac{\partial}{\partial\theta} l(\theta) = 0$$

e nel nostro esempio coincide con la frequenza relativa $\bar{x}$ dei superamenti di soglia osservati nel campione:

$$\frac{\partial}{\partial\theta} l(\theta) = \frac{T\bar{x}}{\theta} - \frac{(T - T\bar{x})}{(1-\theta)} = 0 \Rightarrow \hat{\theta}(\boldsymbol{x}) = \bar{x}$$

Nell'esempio che abbiamo considerato, il principio di massima verosimiglianza ci porta a considerare uno stimatore piuttosto intuitivo e così naturale da mettere in dubbio l'efficacia dell'impalcatura formale che abbiamo introdotto. Tuttavia, è bene precisare subito che, soprattutto quando abbiamo a che fare con modelli parametrici più complessi, la forma dello stimatore è spesso tutt'altro che ovvia.

4 Informazione e attendibilità

Nei paragrafi precedenti, il principio di veosimiglianza è stato giustificato sul piano intuitivo. È tuttavia importante cercare di dotarsi di strumenti che aiutano a valutare l'attendibilità dello stimatore che tale principio produce.

Le Figure 3 e 4 mostrano come varia la funzione di log-verosimiglianza al variare di θ e della dimensione campionaria. I due grafici sono stati prodotti mediante i seguenti comandi:

```
#log-verosimiglianza relativa al variare di T

theta<-seq(0,1,by=0.01)
log.lik<-function(theta,T){T*0.3*log(theta)+(T*0.7)*log(1-theta)}
sequenza<-seq(10,100,by=10) colori<-rainbow(100)
plot(theta,log.lik(theta,10)-log.lik(0.3,10),type="l",
col=colori[10],xlim=c(0,1.5),ylab="log-verosimiglianza relativa",
xlab=expression(theta))

for(i in
sequenza){lines(theta,log.lik(theta,i)-log.lik(0.3,i),col=colori[i])}
legend(1.1,-5,legend=paste("T=",sequenza),lty=1,col=rainbow(sequenza))
title(main=expression(l(theta)-l(hat(theta))))

#log-verosimiglianza relativa al variare della stima

log.like<-function(theta,theta.hat){10*theta.hat*log(theta)
(10*(1-theta.hat))*log(1-theta)} sequenza<-seq(0.1,0.9,by=0.1)
colori<-rainbow(100)
plot(theta,log.like(theta,0.1)-log.like(0.1,0.1),type="l",
col=colori[10],xlim=c(0,1.5),ylab="log-verosimiglianza relativa",
xlab=expression(theta))

for(i in
sequenza){lines(theta,log.like(theta,i)-log.like(i,i),col=colori[i*100])}
legend(1.1,-5,legend=sequenza,lty=1,col=rainbow(sequenza*100))
title(main=expression(l(theta)-l(hat(theta))))
```

È evidente che lo stimatore $\hat{\theta}(\boldsymbol{x})$ è tanto più attendibile quanto più alta è la velocità di caduta della funzione $l(\theta)$ man mano che ci si allontana dal punto di massimo. Di conseguenza, per valutare l'attendibilità di $\hat{\theta}$, è opportuno calcolare la curvatura (preceduta dal segno negativo) della log-

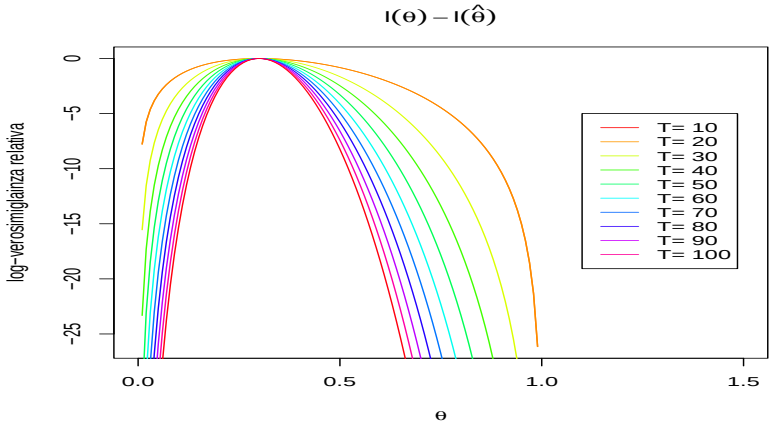

Figura 3: log-verosimiglianza relativa per diverse numerosità campionarie e $\hat{\theta} = 0.3$

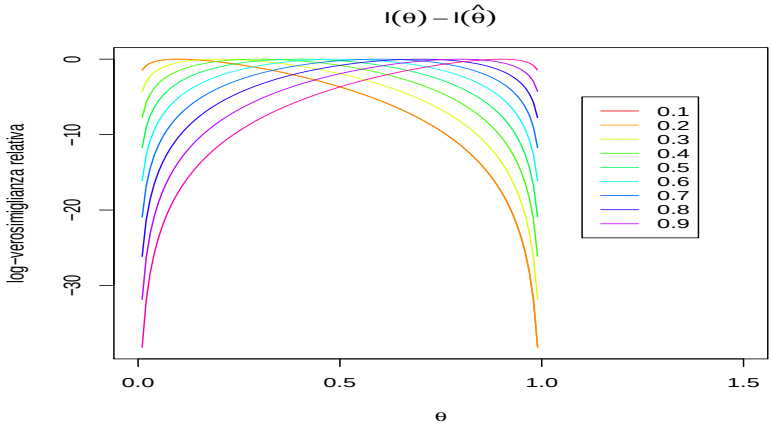

Figura 4: log-verosimiglianza relativa per $T = 10$ e diversi valori dello stimatore $\hat{\theta}$

verosimiglianza nel suo punto di massimo

$$i(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} l(\theta)_{\theta=\hat{\theta}}.$$

La quantità $i(\hat{\theta})$ è nota con il nome di **informazione di fisher osservata** e misura l'attendibilità dello stimatore di massima verosimiglianza. La presenza del segno negativo ha il solo scopo di aver a che fare con una quantità positiva, dato che (sotto condizioni di regolarità) la funzione $l(\theta)$ è concava e possiede derivata negativa nel suo punto di massimo.

Il fatto che la stima $\hat{\theta}$ dovrebbe sempre essere accompagnata dall'infomazione $i(\hat{\theta})$ è anche giustificata dal fatto che la funzione $l(\theta)$ può essere approssimata dalla polinomio di Taylor di secondo grado

$$l(\theta) \approx l(\hat{\theta}) + \frac{1}{2} i(\hat{\theta})(\theta - \hat{\theta})^2,$$

dove il termine di primo grado è pari a zero, avendo come coefficiente la derivata prima calcolata nella stima di massima verosimiglianza. Ne consegue che i numeri $\hat{\theta}$ e $i(\hat{\theta})$ possono essere considerati come una sintesi della forma assunta dalla funzione $l(\theta)$.

5 Distribuzione campionaria

In quanto funzione dei dati campionari $\boldsymbol{x}$, anche lo stimatore $\hat{\theta}(\boldsymbol{x})$ è una variabile aleatoria ed il suo valore varia al variare del campione. La valutazione della qualità di uno stimatore non può prescindere dalla considerazione di tale variabilità. In effetti, ogni volta che si estrae un campione $\boldsymbol{x}$, e si stima il parametro d'interesse θ mediante $\hat{\theta}(\boldsymbol{x})$, si commette l'errore

$$\epsilon(\boldsymbol{x}; \theta) = \hat{\theta}(\boldsymbol{x}) - \theta.$$

Tale errore dipende naturalmente sia dal campione (più o meno "fortunato") sia dal "vero" valore θ assunto dal parametro d'interesse. È anche ovvio che, poichè il valore assunto da θ è incognito, non abbiamo alcuna speranza di calcolare l'errore che commettiamo quando utilizziamo il campione per produrre la nostra stima.

Un approccio ragionevole in tale situazione consiste nel valutare la qualità dello stimatore considerando il suo comportamento rispetto a tutti i campioni possibili. In altre parole, si considera la **distribuzione campionaria** dello stimatore e si valuta la sua forma.

Nell'esempio dei superamenti di soglia, è possibile calcolare con esattezza la distribuzione campionaria dello stimatore $\hat{\theta}(\boldsymbol{x}) = \bar{x}$:

$$P(\hat{\theta}(\boldsymbol{x}) = \hat{\theta}) = \binom{T}{T\hat{\theta}} \theta^{T\hat{\theta}} (1-\theta)^{(T-T\hat{\theta})}.$$

Tale distribuzione è nota come **distribuzione binomiale** di parametri T e θ.

Nel caso specifico, tale distribuzione ci mette in grado di calcolare la probabilità di un qualunque errore di stima, per ogni valore di θ. In particolare, è semplice dimostrare che, se si distribuisce secondo la binomiale sopra riportata, il valore atteso di $\hat{\theta}(\boldsymbol{x}) = \bar{x}$ è dato da $\mathbb{E}\hat{\theta}(\boldsymbol{x}) = \theta$ e la sua varianza è data da $\mathbb{V}\hat{\theta}(\boldsymbol{x}) = \frac{\theta(1-\theta)}{T}$. Di conseguenza, qualunque sia il valore assunto da θ, possiamo affermare che lo stimatore $\hat{\theta}(\boldsymbol{x}) = \bar{x}$ commette in media un errore nullo

$$\mathbb{E}(\hat{\theta}(\boldsymbol{x}) - \theta) = \mathbb{E}\hat{\theta}(\boldsymbol{x}) - \theta = \theta - \theta = 0$$

ed inoltre, l'errore quadratico descresce all'aumentare della numerosità campionaria

$$\mathbb{V}\hat{\theta}(\boldsymbol{x}) = \mathbb{E}(\hat{\theta}(\boldsymbol{x}) - \theta)^2 = \frac{\theta(1-\theta)}{T}.$$

Si noti che tale varianza non è costante al variare di θ: a parità di numerosità campionaria, la probabilità di commettere errori rilevanti è più alta quando θ si trova in un intorno di 0.5, mentre decresce quando θ si trova vicino a 0 o a 1.

In generale, diciamo che uno stimatore $\hat{\theta}(\boldsymbol{x})$ è **non distorto** se il valore atteso della sua distribuzione campionaria coincide con il valore assunto dal parametro di interesse, ovvero se, in media, commette un errore nullo. In particolare, la **distorsione** di uno stimatore

$$\mathbb{B}\hat{\theta}(\boldsymbol{x}) = \mathbb{E}(\hat{\theta}(\boldsymbol{x}) - \theta)$$

misura il suo grado di "accuratezza" nel processo di stima. Inoltre, se due stimatori $\hat{\theta}_1(\boldsymbol{x})$ e $\hat{\theta}_2(\boldsymbol{x})$ sono non distorti, è abbastanza naturale preferire quello con la varianza minore, che in tal caso viene denominato come il più **efficiente**. L'efficienza di uno stimatore, misurata dalla sua varianza, è indice del grado di "precisione" dello stimatore.

Una misura sintetica di accuratezza e precisione di uno stimatore è data dal suo **errore quadratico medio**

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}(\boldsymbol{x})) &= \mathbb{E}(\hat{\theta}(\boldsymbol{x}) - \theta)^2 \\ &= \mathbb{V}\hat{\theta}(\boldsymbol{x}) + \mathbb{B}^2\hat{\theta}(\boldsymbol{x}). \end{aligned}$$

Se in un applicazione sono disponibili due stimatori, sceglieremo quello con il valore di MSE minore. Se uno stimatore è non distorto, il suo MSE coincide con la sua varianza.

La distribuzione campionaria di uno stimatore di massima verosimiglianza varia secondo l'applicazione che si sta considerando ed è strettamente legata alla distribuzione di probabilità del campione stesso. Tuttavia, al crescere della numerosità campionaria, la distribuzione di $\mathbb{V}\hat{\theta}(\boldsymbol{x})$ può essere approssimata da una densità gaussiana di media θ e varianza pari all'inverso dell'informazione di Fisher osservata

$$f(\mathbb{V}\hat{\theta}(\boldsymbol{x})) \approx N(\theta, \frac{1}{i(\hat{\theta})})$$

e tale approssimazione migliora all'aumentare della dimensione campionaria. Questo risultato, valido sotto opportune condizioni di regolarità, ci permette di dire che lo stimatore di massima verosimiglianza è **asintoticamente non distorto** e che la sua varianza è bene approssimabile dall'inverso dell'informazione osservata di Fisher.

6 Il caso generale

Supponiamo di aver a che fare con un campione $\boldsymbol{x} = (x_1 \dots x_i \dots x_n)$ di osservazioni indipendenti e identicamente distribuite secondo una variabile aleatoria X di distribuzione $f(x; \boldsymbol{\theta})$, nota a meno di un vettore di parametri reali $\boldsymbol{\theta} = (\theta_1 \dots \theta_q \dots \theta_Q) \in \Theta \subseteq \mathbb{R}^Q$ e definita su un supporto $\mathbb{X}$. Il supporto $\mathbb{X}$ è il sottoinsieme (proprio o improprio) per il quale

$$\forall x \in \mathbb{X} \quad f(x; \boldsymbol{\theta}) \geq 0$$
$$\int_{\mathbb{X}} f(x; \boldsymbol{\theta}) dx = 1$$

In particolare, supponiamo che il supporto non dipenda dal valore assunto da $\boldsymbol{\theta}$. In questo caso, il modello parametrico è dato dalla terna

$$(\mathbb{X}^n, \boldsymbol{x}; \boldsymbol{\theta}, \Theta)$$

dove

- $\mathbb{X}^n$ è il prodotto di n copie di $\mathbb{X}$, e
- la distribuzione del campione è data dal prodotto

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

La funzione $f(\boldsymbol{x};\boldsymbol{\theta})$ dipende sia dai dati campionati che dal valore assunto dai parametri. Fissato un campione estratto, essa fornisce pesi di verosimiglianza ai parametri e prende il nome di funzione di verosimiglianza. La funzione di log-verosimiglianza è data da

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i;\boldsymbol{\theta})$$

e lo stimatore di massima verosimiglianza (MLE, *Maximum Likelihood Estimator*) $\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = (\hat{\theta}_1 \dots \hat{\theta}_Q)$ è il punto-vettore di massimo di $l(\boldsymbol{\theta})$. In casi regolari, $l(\boldsymbol{\theta})$ è una funzione concava e lo stimatore ML esiste ed è l'unica radice del sistema di Q equazioni

$$\begin{aligned} \frac{\partial}{\partial\theta_1} l(\boldsymbol{\theta}) &= 0 \\ &\dots \\ \frac{\partial}{\partial\theta_Q} l(\boldsymbol{\theta}) &= 0 \end{aligned}$$

dette equazioni "punteggio" (*score*). Tale sistema di equazioni può essere brevemente indicato come segue

$$\frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \nabla l(\boldsymbol{\theta}) = \mathbf{0}$$

dove $\nabla l(\boldsymbol{\theta})$ è il vettore gradiente della log-verosimiglianza. Quindi lo stimatore ML è semplicemente il punto di Θ dove il gradiente della log-verosimiglianza si annulla.

La funzione $l(\boldsymbol{\theta})$ può essere approssimata dalla funzione quadratica

$$l(\boldsymbol{\theta}) \approx l(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' I(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

che non è altro che un polinomio di Taylor arrestato al secondo ordine calcolato intorno allo stimatore ML (il termine di primo ordine è ovviamente uguale a 0 avendo come coefficiente il gradiente della log-verosimiglianza calcolato in $\hat{\boldsymbol{\theta}}$). La matrice $Q \times Q$ $I(\hat{\boldsymbol{\theta}})$ è nota come informazione di Fisher osservata e non è altro che la matrice hessiana della funzione di log-verosimiglianza, preceduta dal segno negativo e calcolata nel punto $\hat{\boldsymbol{\theta}}$. Il generico elemento (p,q)-mo di $I(\hat{\boldsymbol{\theta}})$ è pertanto dato da

$$I_{pq}(\hat{\boldsymbol{\theta}}) = -\frac{\partial}{\partial\theta_p \partial\theta_q} l(\boldsymbol{\theta})_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Nel caso di osservazioni indipendenti e identicamente distribuite, tale derivata seconda è semplicemente data dalla somma

$$I_{pq}(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} \frac{\partial}{\partial\theta_p \partial\theta_q} \log f(x_i;\boldsymbol{\theta})_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

e, in molti casi regolari, la derivata seconda di $f(x;\boldsymbol{\theta})$, non dipende dal valore x. In tal caso, i valori dell'informazione di Fisher assumono la semplice forma

$$I_{pq}(\hat{\boldsymbol{\theta}}) = n i_{pq}$$

dove

$$i_{pq} = -\frac{\partial^2}{\partial\theta_p \partial\theta_q} f(x;\boldsymbol{\theta})_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Quando n cresce la distribuzione campionaria di $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ può essere approssimata dalla normale a Q dimensioni

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) \sim N(\boldsymbol{\theta}, I^{-1}(\hat{\boldsymbol{\theta}})) = (2\pi)^{-n/2} |I(\hat{\boldsymbol{\theta}})|^{1/2} \exp(-\frac{1}{2}\left((\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})' I(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\right)$$

dove, nei casi regolari sopra accennati, la matrice di varianze e covarianze della normale assume la forma

$$I^{-1}(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} H^{-1}(\hat{\boldsymbol{\theta}})$$

dove H è la matrice hessiana di $f(x;\boldsymbol{\theta})$.

Dalla distribuzione congiunta di $\hat{\boldsymbol{\theta}}$, si può dedurre che la distribuzione di ogni singolo stimatore $\hat{\theta}_q$ è ben approssimata da una normale univariata

$$\hat{\theta}_q(\boldsymbol{x}) \sim N(\theta_q, I^{(qq)}(\hat{\boldsymbol{\theta}}))$$

dove con $I^{(qq)}$ si è indicato l'elemento q-mo della diagonale dell'inversa dell'informazione osservata di Fisher. Nei citati casi regolari, tale varianza assume la semplice forma

$$\mathbb{V}\hat{\theta}_q = I^{(qq)}(\boldsymbol{\theta}) = -\frac{1}{n} h^{(qq)}(\hat{\boldsymbol{\theta}})$$

dove $h^{(qq)}(\hat{\boldsymbol{\theta}})$ è il q-mo elemento della diagonale dell'inversa dell'hessiana di $f(x;\boldsymbol{\theta})$.

Possiamo utilizzare tale risultato per costruire un intervallo di confidenza per ogni singolo parametro al livello $1-\alpha$:

$$\theta_q - z_{\alpha/2}\text{s.e.}(\hat{\theta}_q), \theta_q + z_{\alpha/2}\text{s.e.}(\hat{\theta}_q)$$

dove $z_{\alpha/2}$ è il percentile di ordine $1-\alpha/2$ della normale standardizzata (calcolabile in R dalla funzione `qnorm(level)`, dove `level`$= 1-\alpha/2$), e dove s.e.$(\hat{\theta}_q)$ è l'errore standard della stima del parametro q-mo e non è altro che la radice quadrata della varianza di stima $\mathbb{V}\hat{\theta}_q$. Il livello $1-\alpha \in (0,1)$ indica la confidenza dell'intervallo. Più precisamente, α è la probabilità di estrarre un campione che produce un intervallo non contenente il vero valore del parametro.

Se l'intervallo contiene lo zero, diciamo che la stima $\hat{\theta}_q$ non è significativamente diversa da zero (al livello $1-\alpha$).

Si osservi che il raggio dell'intervallo di confidenza dipende sia dalla scelta di α, che dalla numerosità campionaria. All'aumentare di n, l'intervallo si restringe a parità di α. Al diminuire di α, l'intervallo si allarga, a parità di dimensione campionaria.

7 Un esempio con R

La distribuzione log-normale è utilizzata spesso nelle applicazioni di statistica ambientale. Ad esempio è noto che la misurazione della precipitazione piovana segue spesso tale distribuzione. La distribuzione log-normale di parametri μ e σ è la densità di probabilità che si ottiene per la variabile $Y = e^X$, dove $X \sim N(\mu, \sigma)$ ed assume la forma

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left(-\frac{1}{2}\frac{(\log y - \mu)^2}{\sigma^2}\right).$$

In R, è possibile generare un campione $\boldsymbol{y} = (y_1 \dots y_n)$ di osservazioni indipendenti e identicamente distribuite secondo una distribuzione log-normale di parametri μ e σ. Ad esempio, il comando seguente genera un vettore di 100 osservazioni da una log-normale di parametri $\mu = 5$ e $\sigma = 0.5$:

```
> y <- rlnorm(100, 5, 0.5)
```

Per dare un'occhiata ai primi 10 di tali valori, è sufficiente digitare

```
> y[1:10]
```

```
 [1] 163.6633 162.8772 116.8939 107.5579 138.6956 279.5716 176.2567 161.6856
 [9] 173.8772 201.6113
```

La Fig. 5 mostra l'accostamento tra i dati simulati e la distribuzione teorica. Il grafico è stato creato mediante la funzione `dlnorm` che fornisce il valore

```
> hist(y, main = "accostamento dati simulati", freq = FALSE)
> lines(seq(0, max(y), by = 0.05), dlnorm(seq(0, max(y), by = 0.05),
+     meanlog = 5, sdlog = 0.5))
```

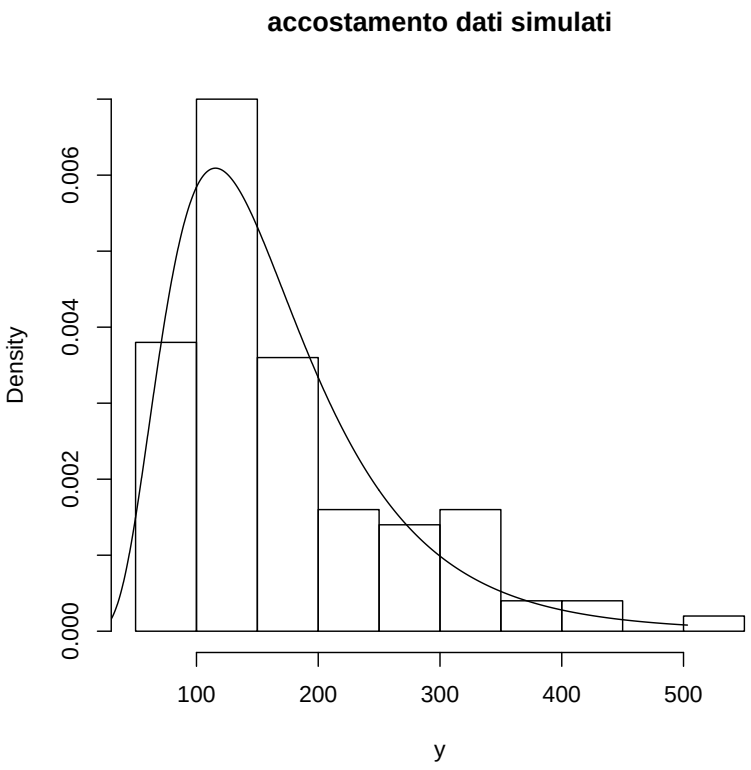

Figura 5: accostamento dati simulati all densità log-normale

della densità log-normale, calcolata in corrispondenza di una sequenza di valori `seq(0,max(y),by=0.05)` crescenti da 0 al massimo dei dati generati.

Supponendo ora che i dati simulati siano reali, desideriamo scoprire fino a che punto il metodo della massima verosimiglianza riesce a ricostruire il valore dei parametri usati nella simulazione.

Procederemo nel seguente modo: definiremo la funzione di log-verosimiglianza, preceduta dal segno meno e quindi chiederemo ad R di trovarne il punto di minimo. Nel nostro caso, la funzione di log-verosimiglianza è data da

$$\begin{aligned} l(\mu, \sigma) &= -\sum_{i=1}^{100} \left(\log \sigma + \log y_i + \frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right) \\ &= -(A + B) \end{aligned}$$

e in R la funzione $-l(\mu, \sigma) = A + B$ può essere definita nel seguente modo:

```
> neg.log.like <- function(theta, data) {
+     mu <- theta[1]
+     sigma <- theta[2]
+     A <- sum(log(sigma) + log(data))
+     B <- sum((1/(2 * sigma^(2))) * (log(data) - mu)^(2))
+     f <- A + B
+     f
+ }
```

I seguenti grafici (Fig. 6 e 7) mostrano l'andamento della funzione $-l(\mu, \sigma)$ al variare dei parametri. Per costruirli, è necessario valutare i valori assunti dalla funzione in una griglia:

```
> mu. <- seq(4, 6, by = 0.05)
> sigma. <- seq(0.4, 0.6, by = 0.05)
> z <- matrix(rep(0, length(mu.) * length(sigma.)), length(mu.),
+     length(sigma.))
> for (i in 1:length(mu.)) {
+     for (j in 1:length(sigma.)) {
+         z[i, j] <- neg.log.like(c(mu.[i], sigma.[j]), data = y)
+     }
+ }
```

per poi creare i grafici desiderati (Figure 6 e 7).

La ricerca numerica delle stime di massima verosimiglianza può essere realizzata mediante il comando

```
> persp(mu., sigma., z, theta = 30, phi = 30, expand = 0.5,
+     col = "lightblue")
```

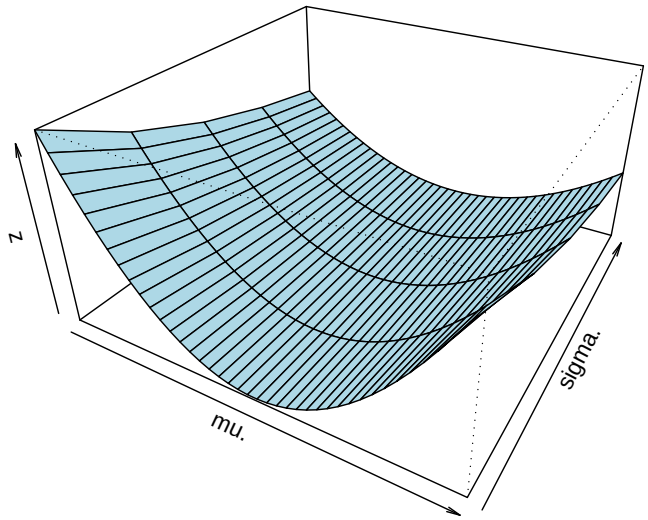

Figura 6: la curva $-l(\mu, \sigma)$ in un intorno del suo punto di minimo

```
> contour(mu., sigma., z, xlab = "mu", ylab = "sigma")
```

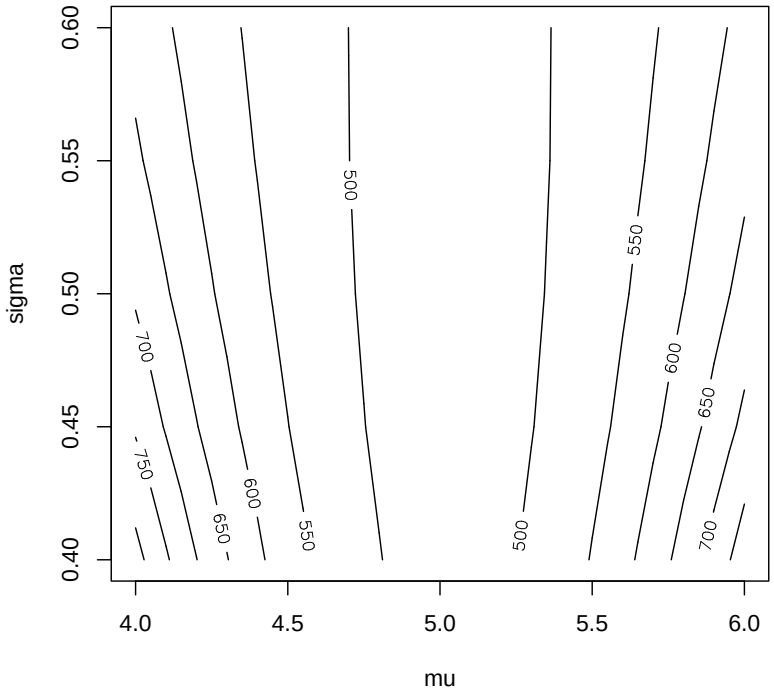

Figura 7: curve di livello per la funzione $-l(\mu, \sigma)$

```
> mle <- optim(c(1, 1), neg.log.like, data = y, hessian = TRUE)
> mle$par

[1] 5.0322506 0.4834105

> mle$hessian

            [,1]        [,2]
[1,] 427.9251747   0.1195184
[2,]   0.1195184 854.9780025
```

dove abbiamo richiesto le stime `par` e la matrice hessiana `hessian`.

Per costruire intervalli di confidenza al livello $1-\alpha = 0.95$ per i parametri μ e σ:

$$\hat{\mu} - z_{\alpha/2}\mathrm{sd}(\hat{\mu}), \hat{\mu} + z_{\alpha/2}\mathrm{sd}(\hat{\mu})$$
$$\hat{\sigma} - z_{\alpha/2}\mathrm{sd}(\hat{\sigma}), \hat{\sigma} + z_{\alpha/2}\mathrm{sd}(\hat{\sigma})$$

dobbiamo prima valutare $z_{\alpha/2}$:

```
> z.alpha <- qnorm(0.975)
> z.alpha

[1] 1.959964
```

In secondo luogo, troviamo gli errori standard, invertendo la matrice hessiana

```
> var <- solve(mle$hessian)
```

e trovando gli errori standard delle due stime

```
> s.e.mu <- sqrt(var[1, 1])
> s.e.sigma <- sqrt(var[2, 2])
```

L'intervallo di confidenza al livello 95% per μ è dunque dato da

```
> lower.mu <- mle$par[1] - z.alpha * s.e.mu
> upper.mu <- mle$par[1] + z.alpha * s.e.mu
> c(lower.mu, upper.mu)

[1] 4.937504 5.126997
```

mentre quello per σ è dato da

```
> lower.sigma <- mle$par[2] - z.alpha * s.e.sigma
> upper.sigma <- mle$par[2] + z.alpha * s.e.sigma
> c(lower.sigma, upper.sigma)

[1] 0.4163803 0.5504407
```

8 Esercizio

1. studiate attentamente l'esempio numerico (naturalmente, i vostri dati simulati saranno diversi e di conseguenza anche grafici e stime non saranno riproducibili esattamente)

2. generate ed esibite un campione $\boldsymbol{x}$ di 100 osservazioni indipendenti e identicamente distribute secondo una normale $N(5, 0.5)$, dove 0.5 indica la deviazione standard (usate il comando `rnorm`)

3. controllate graficamente che l'istogramma dei dati simulati si adatta alla densità teorica (usate i comandi `hist` e `lines`)

4. calcolate analiticamente lo stimatore di massima verosimiglianza dei parametri della normale, annullando il gradiente della funzione di log-verosimiglianza

5. definite in R l'opposto della funzione di log-verosimiglianza in R; in questo caso, la funzione è data da:

$$-l(\mu, \sigma) = \sum_{i=1}^{100} \left\{ \log(\sigma^2) + \frac{1}{\sigma^2}(x_i - \mu)^2 \right\}$$

6. trovate le stime di massima verosimiglianza dei parametri μ e σ mediante il comando `optim`

7. esaminate graficamente l'andamento dell'opposto della funzione di log-verosimiglianza in un intorno del suo punto di minimo (usate i comandi `persp` e `contour`)

8. trovate gli intervalli di confidenza per le due stime, al livello $1-\alpha = 0.95$ (usate il comando `solve`)

9. preparate un breve rapporto contenente i risultati e i comandi che avete usato in R per produrli (lo stile dovrebbe essere somigliante alla sezione precedente) e inviatelo a `lagona@uniroma3.it` con oggetto "IEAT PROVA MOD STAT-nome studente"