Politecnico di Milano
Appunti delle lezioni del corso di Statistica (2L) per gli allievi INF e TEL, AA 2008/2009*

Teoria della stima puntuale†

Ilenia Epifani

9 marzo 2009

Prima edizione AA 2003/2004; Quarta edizione AA 2008/2009

†Gli sperabili miglioramenti della versione AA 2005/2006 derivano da quelli sicuramente apportati dal Prof. Barchielli e dalla Dott.ssa Salvati alla versione inglese

Indice

Nelle lezioni precedenti abbiamo trattato il problema della stima puntuale e intervallare per il modello gaussiano. Nelle prossime affronteremo il problema della stima puntuale in generale cioè per modelli non necessariamente gaussiani. In particolare, ci porremo i seguenti problemi:

1. descrivere una teoria degli stimatori "ottimali" in una classe di stimatori che soddisfano opportune proprietà;

2. fornire metodi di ricerca degli stimatori.

A tal fine abbiamo bisogno di qualche definizione preliminare. Per maggiore ordine e semplicità espositiva, riprenderemo quanto già ampiamente descritto nelle lezioni precedenti.

D'altro canto, la presente nota è intesa a integrazione del libro di testo (Pestman 1998); di conseguenza, per le parti di programma sviluppate in classe ma qui non esplicitamente trattate si rimanda a Pestman 1998 e agli appunti presi in classe.

1 Definizioni e notazioni

Sia X una variabile aleatoria che ha funzione di ripartizione (f.d.r.) F e densità di probabilità f. Supponiamo che f non sia completamente incognita, ma sia nota a meno di un parametro θ m-dimensionale a valori in Θ sottoinsieme di $\mathbb{R}^m$; per esempio: $\Theta = \mathbb{R}$ o $\Theta = [0,1]$ o $\Theta = (0,\infty)$, o $\Theta = \mathbb{R}\times(0,\infty)$, ... L'insieme Θ di tutti i possibili valori che il parametro θ può assumere è detto *spazio parametrico*. Per indicare che la densità f dipende da un parametro incognito θ, scriviamo f come funzione della realizzazione x (di X) e di θ come $f(x,\theta)$. Quindi, al variare di θ in Θ, abbiamo la *famiglia di densità di probabilità* $\{f(\cdot,\theta), \theta \in \Theta\}$. Di seguito useremo lo stesso simbolo $f(x,\theta)$ sia per le densità di variabili aleatorie discrete che assolutamente continue.

Definizione 1.1 Siano $X_1,\ldots,X_n$ n variabili aleatorie indipendenti e identicamente distribuite con comune funzione di densità $f(x,\theta)$, $\theta \in \Theta$. Diremo che $X_1,\ldots,X_n$ è un *campione causale* di dimensione n estratto dalla popolazione di densità $f(x,\theta)$.

Siccome un campione casuale $X_1,\ldots,X_n$ è costituito da variabili aleatorie indipendenti tutte aventi la stessa densità marginale $f(\cdot,\theta)$, segue che la densità congiunta di $X_1,\ldots,X_n$ è il prodotto delle densità marginali $\prod_{i=1}^n f(x_i,\theta)$.

Definizione 1.2 Una *statistica* è una variabile aleatoria T funzione del campione: $T = g(X_1,\ldots,X_n)$. La distribuzione (o legge) di una statistica T è detta *distribuzione (o legge) campionaria*.

Badate bene che una statistica T non dipende MAI dai parametri incogniti θ; mentre, la distribuzione campionaria di T dipenderà in generale da θ, dal momento che T è funzione di $X_1,\ldots,X_n$ e la comune densità delle X_i dipende da θ.

Definizione 1.3 Una funzione $\kappa : \Theta \to \mathbb{R}$ è detta *caratteristica della popolazione*. Se κ è costante su Θ, κ è una *caratteristica banale* (*trivial characteristic*).

Definizione 1.4 Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x,\theta), \theta \in \Theta$, e $\kappa(\theta)$ una caratteristica della popolazione. Uno *stimatore* di $\kappa(\theta)$, basato sul campione $X_1, \ldots, X_n$, è una statistica $T = g(X_1, \ldots, X_n)$ usata per stimare $\kappa(\theta)$. Il valore assunto da uno stimatore T di $\kappa(\theta)$ è detto *stima* di $\kappa(\theta)$.

Esemplifichiamo ora le nozioni di caratteristica di una popolazione e stimatore in qualche modello statistico "notevole".

Di seguito, come usuale, $\bar{X}$ indicherà la statistica *media campionaria*, $\bar{X} = \sum_{i=1}^n X_i/n$, e S^2 la *varianza campionaria*, $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$.

Esempio 1.5 (Modello gaussiano) Consideriamo un campione casuale gaussiano con media μ e varianza σ^2 entrambe incognite: $X_1, \ldots, X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$. Allora, $\theta = (\mu, \sigma^2)$ è un parametro bidimensionale a valori nello spazio parametrico $\Theta = \mathbb{R} \times (0, \infty)$. Esempi di caratteristiche della popolazione gaussiana sono la media della popolazione $\kappa_1(\mu, \sigma^2) = \mu$, la varianza $\kappa_2(\mu, \sigma^2) = \sigma^2$, il momento secondo $\kappa_3(\mu, \sigma^2) = \mathrm{E}(X_1^2) = \sigma^2 + \mu^2$, la probabilità $\kappa_4(\mu, \sigma^2) = P_\theta(X_1 \leq 2) = \Phi\left((2-\mu)/\sigma\right)$.

La media campionaria $\bar{X}$ fornisce uno stimatore per μ e la varianza campionaria S^2 è uno stimatore di σ^2. Infine,

$$U := \Phi\left(\frac{2 - \bar{X}}{\sqrt{(n-1)S^2/n}}\right)$$

è una statistica che può essere usata come stimatore di $\kappa_4(\mu, \sigma^2)$.

Esempio 1.6 (Modello bernoulliano) Sia $(X_1, \ldots, X_n)$ un campione casuale estratto dalla popolazione bernoulliana di parametro θ, cioè $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x,\theta)$ con

$$f(x,\theta) = \theta \mathbf{1}_{\{1\}}(x) + (1-\theta)\mathbf{1}_{\{0\}}(x) = \theta^x (1-\theta)^{1-x} \mathbf{1}_{\{0,1\}}(x), \quad \theta \in \Theta = [0,1]$$

Nel modello bernoulliano la caratteristica media coincide con il parametro θ e la caratteristica varianza è data dalla funzione $\theta(1-\theta)$. La media campionaria $\bar{X}$ è uno stimatore di θ , mentre, per quanto concerne la varianza, potremmo stimarla con la statistica $\bar{X}(1-\bar{X})$ oppure con la varianza campionaria S^2. In realtà, è facile dimostrare che per il modello bernoulliano, la varianza campionaria S^2 differisce dallo stimatore $\bar{X}(1-\bar{X})$ solo per il fattore moltiplicativo $n/(n-1)$; infatti, i soli valori che ciascuna osservazione X_i può assumere sono $0, 1$ e quindi: $\sum_{j=1}^n X_j^2/n = \sum_{j=1}^n X_j/n = \bar{X}$, da cui segue che

$$S^2 = \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{n-1} = \frac{n}{n-1}\Big(\frac{\sum_{j=1}^n X_j^2}{n} - \bar{X}^2\Big) = \frac{n}{n-1}\bar{X}(1-\bar{X})$$

Esempio 1.7 (Modello di Poisson) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x,\theta)$ con

$$f(x,\theta) = \frac{\mathrm{e}^{-\theta}\theta^x}{x!}\mathbf{1}_{\{0,1,2,\ldots\}}(x), \quad \theta \in \Theta = (0,\infty) \tag{1}$$

Le caratteristiche media e varianza sono uguali e coincidono con il parametro θ. Segue che le statistiche $\bar{X}$ e S^2 costituiscono due diversi stimatori per θ. Un'altra caratteristica di interesse è $\kappa(\theta) := P_\theta(X_1 > 0)$ data da $\kappa(\theta) = 1 - \mathrm{e}^{-\theta}$ che uno potrebbe stimare con la statistica $1 - \mathrm{e}^{-\bar{X}}$ oppure con la statistica $1 - \mathrm{e}^{-S^2}$.

Esempio 1.8 (Modello esponenziale) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta)$ con

$$f(x, \theta) = \frac{1}{\theta} \mathrm{e}^{-x/\theta} \mathbf{1}_{(0,\infty)}(x), \quad \theta \in \Theta = (0, \infty) \tag{2}$$

Nella parametrizzazione (2) del modello esponenziale la caratteristica media coincide con il parametro θ, mentre la caratteristica varianza è data dalla funzione θ^2.

Esempio 1.9 (Modello uniforme) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta)$ con

$$f(x, \theta) = \frac{1}{\theta} \mathbf{1}_{(0,\theta)}(x), \quad \theta \in \Theta = (0, \infty)$$

In questo modello la caratteristica media è $\theta/2$ e due suoi possibili stimatori sono $\bar{X}$ e, come vedremo in seguito, il massimo delle osservazioni $X_{(n)} := \max\{X_1, \ldots, X_n\}/2$.

Concludiamo l'introduzione con una osservazione sulle notazioni. Sia $T = g(X_1, \ldots, X_n)$ uno stimatore di $\kappa(\theta)$ che ha media $\mathrm{E}(T)$ per ogni θ; $\mathrm{E}(T)$ in generale dipende da θ; per esempio, nel caso di una popolazione assolutamente continua di densità $f(x, \theta)$ abbiamo:

$$\mathrm{E}(T) = \mathrm{E}(g(X_1, \ldots, X_n)) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \theta) \cdots f(x_n, \theta)\, dx_1 \cdots dx_n$$

Per sottolineare questa dipendenza, useremo in seguito l'operatore E_θ anzicché semplicemente E. Analogamente, anche la media di una qualunque funzione $Y = h(T)$ di T sarà $\mathrm{E}_\theta(h(T))$.

1.1 Errore quadratico medio

Appare evidente dai precedenti esempi che in un problema di stima, possiamo trovarci a dover scegliere fra stimatori diversi di una caratteristica della popolazione $\kappa(\theta)$. Chiaramente, fra diverse opzioni, usiamo come criterio di scelta la concentrazione (o di contro dispersione) dello stimatore intorno a $\kappa(\theta)$. Poiché ogni stimatore T è una variabile aleatoria, potremmo misurare la concentrazione di T intorno a $\kappa(\theta)$ per esempio come $P(|T - \kappa(\theta)| < \epsilon)$, per qualche $\epsilon > 0$. Alternativamente, possiamo dare una misura "media" della prossimità di T a $\kappa(\theta)$ in termini di $\mathrm{E}_\theta[(T - \kappa(\theta))^2]$. Ovviamente, preferireremo (cioè sceglieremo) lo stimatore più concentrato intorno a $\kappa(\theta)$.

Definizione 1.10 Se T è uno stimatore di $\kappa(\theta)$ tale che $\mathrm{E}_\theta[(T - \kappa(\theta))^2] < \infty$ per ogni $\theta \in \Theta$, allora $\mathrm{E}_\theta[(T - \kappa(\theta))^2]$ è detto *errore quadratico medio* di T rispetto a $\kappa(\theta)$ ed è indicato con l'acronimo MSE che sta per Mean Squared Error (o Mean Square Error).

Il MSE di T rappresenterà il nostro criterio di scelta fra stimatori e fornirà una misura dell'efficienza dello stimatore scelto.

Il MSE di uno stimatore T esiste se e solo se T ha (media e) varianza finite, o, equivalentemente, se e solo se ha momento secondo finito. Infatti, usando la diseguaglianza $(a + b)^2 \leq 2(a^2 + b^2)$, valida per ogni $a, b \in \mathbb{R}$[1], otteniamo

$$\mathrm{E}_\theta \left[\big(T - \kappa(\theta)\big)^2 \right] \leq 2\, \mathrm{E}_\theta[T^2 + \kappa(\theta)^2] = 2\, \mathrm{E}_\theta[T^2] + 2\kappa(\theta)^2$$

[1] $(|a| - |b|)^2 \geq 0 \iff a^2 - 2|a||b| + b^2 \geq 0 \iff 2|a||b| \leq a^2 + b^2$, da cui $(a+b)^2 \leq (|a| + |b|)^2 = a^2 + 2|a||b| + b^2 \leq 2(a^2 + b^2)$.

pertanto $\mathrm{E}_\theta[T^2] < \infty$ implica $\mathrm{E}_\theta\left[(T-\kappa(\theta))^2\right] < \infty$.

Viceversa,

$$\mathrm{E}_\theta[T^2] = \mathrm{E}_\theta\left[\left(T-\kappa(\theta)+\kappa(\theta)\right)^2\right] \leq 2\,\mathrm{E}_\theta\left[\left(T-\kappa(\theta)\right)^2+\kappa(\theta)^2\right] = 2\,\mathrm{E}_\theta\left[\left(T-\kappa(\theta)\right)^2\right] + 2\kappa(\theta)^2$$

e quindi $\mathrm{E}_\theta\left[\left(T-\kappa(\theta)\right)^2\right] < \infty$ implica $\mathrm{E}_\theta[T^2] < \infty$.

Alla luce di quanto appena discusso, in ciò che segue ci ridurremo a considerare la classe degli stimatori a varianza finita.

Osservazione 1.11 Per calcolare l'errore quadratico medio è utile decomporlo nel seguente modo:

$$\mathrm{E}_\theta[(T-\kappa(\theta))^2] = \mathrm{E}_\theta[((T-\mathrm{E}_\theta(T)) + (\mathrm{E}_\theta(T)-\kappa(\theta)))^2] = \mathrm{E}_\theta[(T-\mathrm{E}_\theta(T))^2 + (\mathrm{E}_\theta(T)-\kappa(\theta))^2]$$

cioè

$$\mathrm{E}_\theta[(T-\kappa(\theta))^2] = \mathrm{Var}_\theta(T) + [\mathrm{E}_\theta(T)-\kappa(\theta)]^2 \tag{3}$$

Osservate che tutti i passaggi sono leciti perché abbiamo assunto la finitezza di media e varianza di T. L'Equazione (3) è molto utile sia dal punto di vista teorico che nelle applicazioni ed esercizi per calcolare l'errrore quadratico medio; questa decomposizione tornerà utile più avanti nello studio della consistenza in media quadratica (cfr. Sezione 4).

La quantità $(\mathrm{E}_\theta(T)-\kappa(\theta))$ nella decomposizione (3) è detta *distorsione* (*bias*) di T.

Sulla base dell'errore quadratico medio, *preferiremo* T_1 a T_2 se

- $\mathrm{E}_\theta[(T_1-\kappa(\theta))^2] \leq \mathrm{E}_\theta](T_2-\kappa(\theta))^2]$ $\forall\theta\in\Theta$ e
- $\mathrm{E}_\theta[(T_1-\kappa(\theta))^2] < \mathrm{E}_\theta[(T_2-\kappa(\theta))^2]$ per qualche $\theta\in\Theta$

In linea teorica, fra tutti gli stimatori di $\kappa(\theta)$ a varianza finita, ci piacerebbe scegliere quello che minimizza il MSE qualunque sia θ, cioè ci piacerebbe scegliere uno stimatore T_0 tale che

$$\mathrm{E}_\theta[(T_0-\kappa(\theta))^2] \leq \mathrm{E}_\theta[(T-\kappa(\theta))^2] \quad \forall\theta\in\Theta \quad \forall T$$

Ma un tale T_0 non esiste: supponete infatti che $\kappa(\theta)=\theta\in\mathbb{R}$. Forse è insensato, comunque potreste decidere di stimare $\kappa(\theta)$ con la statistica costante $\widetilde{T}=5$, il cui MSE in $\theta=5$ vale $\mathrm{E}_5[(\widetilde{T}-5)^2] = \mathrm{E}_5[(5-5)^2] = 0$; cioè sebbene la scelta di $\widetilde{\theta}=5$ sia assurda, per $\theta=5$ la costante 5 si comporta meglio di qualunque altro stimatore.

La ragione di ciò è che la classe di tutti gli stimatori con MSE finito è troppo grande; per esempio contiene anche stimatori banali (per esempio $\widetilde{T}=5$) "buoni" solo in un singolo punto. Allora, per uscire dall'impaccio è sufficiente restringerci nella ricerca a una sottoclasse di stimatori. La sottoclasse ci è suggerita dalla decomposizione (3) del MSE, nella quale leggiamo che minimizzare MSE equivale a minimizzare contemporaneamente varianza e distorsione di T. Allora, aggiustiamo il tiro e restringiamo la nostra ricerca degli stimatori ottimali alla classe di quelli che hanno distorsione nulla, detti *non distorti* o anche *corretti*. L'errore quadratico medio di uno stimatore non distorto di una caratteristica $\kappa(\theta)$ coincide con la sua varianza e il programma di ricerca di uno stimatore ottimale fra quelli non distorti diventa la ricerca di uno stimatore che abbia varianza più piccola per ogni $\theta\in\Theta$.

2 Stimatori non distorti

In questa sezione investigheremo la classe degli stimatori non distorti; quindi torneremo al problema della ricerca degli stimatori ottimali nella successiva.

Definizione 2.1 Una statistica T che ammette media per ogni θ in Θ è detta *stimatore non distorto* o corretto (*unbiased*) della caratteristica $\kappa(\theta)$ se

$$\mathrm{E}_\theta(T) = \kappa(\theta) \qquad \forall \theta \in \Theta \tag{4}$$

Esempio 2.2 Se $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta)$, $\theta \in \Theta$, ed $\mathrm{E}_\theta(X_1)$ esiste qualunque sia θ, allora $\mathrm{E}_\theta(\bar{X}) = \mathrm{E}_\theta(X_1),\ \forall \theta$ e quindi:

La media campionaria $\bar{X}$ è stimatore non distorto della media "teorica" $\mathrm{E}_\theta(X_1)$.

Se inoltre esiste anche $\mathrm{Var}_\theta(X_1),\ \forall \theta \in \Theta$, allora $\mathrm{E}_\theta(S^2) = \mathrm{Var}_\theta(X_1),\ \forall \theta \in \Theta$ e quindi

la varianza campionaria S^2 è stimatore non distorto della varianza "teorica" $\mathrm{Var}_\theta(X_1)$.

Osservazione 2.3 Media e varianza campionarie continuano a essere stimatori non distorti rispettivamente di media e varianza teoriche anche se le osservazioni $X_1, \ldots, X_n$ costituiscono un campione casuale estratto da una f.d.r. F completamente incognita.

Esempio 2.4 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla f.d.r. F. Allora $\mathbf{1}_{(-\infty,x]}(X_1)$ è una statistica che ha densità bernoulliana di parametro $F(x)$; infatti

$$P(\mathbf{1}_{(-\infty,x]}(X_1) = 1) = P(X_1 \leq x) = F(x)$$

Quindi $\mathbf{1}_{(-\infty,x]}(X_1)$ è uno stimatore non distorto della f.d.r. $F(x)$. Un altro stimatore non distorto di $F(x)$ è dato dalla media campionaria del "campione casuale" $\mathbf{1}_{(-\infty,x]}(X_1), \ldots, \mathbf{1}_{(-\infty,x]}(X_n)$, cioè

$$F_n(x) := \frac{\sum_{j=1}^n \mathbf{1}_{(-\infty,x]}(X_j)}{n}$$

La funzione definita da $x \mapsto F_n(x)$, è detta *f.d.r. empirica* ed è usata per stimare F quando non si conosce nulla di F, neppure la forma. Per ogni x reale, semplicemente $F_n(x)$ indica la frequenza relativa delle osservazioni di valore al più pari a x. Torneremo sulla f.d.r. empirica nell'ultima parte del corso dedicata alla statistica non parametrica.

La proprietà di non distorsione di uno stimatore traduce in qualche senso la richiesta di non commettere errori sistematici di sottostima o sovrastima nell'approssimazione di una caratteristica incognita della popolazione. La non distorsione è un modo per formalizzare la richiesta che il campione sia rappresentativo della popolazione. Se infatti questo è il caso, un buon stimatore T mediamente riprodurrà la vera caratteristica incognita.

Ma il programma di ricerca di stimatori non distorti ha qualche limite. Infatti *uno stimatore non distorto potrebbe non esistere* (si confronti l'Esempio 2.5), oppure se esiste *potrebbe non essere unico* (si confronti l'Esempio 2.6). Infine, *può succedere che esista un unico stimatore non distorto di $\kappa(\theta)$, ma sia* insensato (si confronti l'Esempio 2.8).

Esempio 2.5 Sia $X_1 \sim \mathbf{Bi}(n, \theta)$, $\theta \in (0,1)$ con n noto e $\kappa(\theta) = 1/\theta$. Allora $T = g(X_1)$ è uno stimatore non distorto di $1/\theta$ se e solo se

$$\mathrm{E}_\theta(T) = \sum_{k=0}^{n} g(k) \binom{n}{k} \theta^k (1-\theta)^{n-k} = \frac{1}{\theta} \quad \forall \theta \in (0,1)$$

Ma ciò è impossibile perché $\mathrm{E}_\theta(T)$ è sempre un polinomio di grado $n \geq 1$ in θ, qualunque sia la scelta della funzione g.

Esempio 2.6 Se $X_1, \ldots, X_n$ è un campione casuale estratto dalla popolazione di Poisson di parametro θ, allora $T_1 = \bar{X}$ e $T_2 = S^2$ sono due (diversi) stimatori entrambi non distorti di θ che rappresenta sia la media che la varianza (teoriche).

Osservazione 2.7 Se esistono due stimatori non distorti di $\kappa(\theta)$, allora ne esistono infiniti; infatti, se T_1 e T_2 sono due stimatori non distorti di $\kappa(\theta)$, allora, per ogni $a \in \mathbb{R}$, $T_a := aT_1 + (1-a)T_2$ è uno stimatore di $\kappa(\theta)$ che ha media e

$$\mathrm{E}_\theta(T_a) = a\,\mathrm{E}_\theta(T_1) + (1-a)\,\mathrm{E}_\theta(T_2) = a\kappa(\theta) + (1-a)\kappa(\theta) = \kappa(\theta) \quad \forall \theta \in \Theta$$

Esempio 2.8 Supponiamo che il numero di chiamate a un numero verde all'ora di punta del giorno i-esimo si possa modellare con una variabile aleatoria discreta X_i di Poisson di parametro θ e assumiamo l'indipendenza fra chiamate in giorni diversi. Conosciamo il numero di telefonate X_1 arrivate il primo giorno e vogliamo stimare in modo non distorto la probabilità che non arrivi nessuna telefonata nei due giorni successivi. In altre parole, stiamo cercando uno stimatore non distorto $T = g(X_1)$ della caratteristica $\kappa(\theta) = P_\theta(X_2 + X_3 = 0)$. Poiché X_2, X_3 *i.i.d.* $\sim \mathcal{P}(\theta)$ implica che $X_2 + X_3 \sim \mathcal{P}(2\theta)$, allora

$$\kappa(\theta) = \mathrm{e}^{-2\theta} = \mathrm{e}^{-\theta} \sum_{k=0}^{\infty} (-1)^k \frac{\theta^k}{k!}$$

Inoltre,

$$\mathrm{E}_\theta(T) = \sum_{k=0}^{\infty} g(k) \mathrm{e}^{-\theta} \frac{\theta^k}{k!}$$

e la condizione di non distorsione $\mathrm{E}_\theta(T) = \kappa(\theta)$ diventa:

$$\sum_{k=0}^{\infty} g(k) \frac{\theta^k}{k!} = \sum_{k=0}^{\infty} (-1)^k \frac{\theta^k}{k!}, \quad \forall \theta > 0 \tag{5}$$

Nell'ultima equazione ci sono due serie di potenze in θ che coincidono se e solo se i corrispondenti coefficenti coincidono, cioè se e solo se $g(k) = (-1)^k$, $\forall k = 0, 1, 2, \ldots$. Segue che l'unico stimatore non distorto di $\mathrm{e}^{-2\theta}$ è $T = (-1)^{X_1}$. Ora, secondo voi, ha senso stimare una probabilità strettamente positiva con una statistica le cui realizzazioni sono soltanto $-1, 1$?

3 Stimatori UMVUE

Riprendiamo ora il filo della ricerca di stimatori ottimali. Riassumendo: abbiamo un campione casuale $X_1, \ldots, X_n$ estratto da $\{f(\cdot, \theta), \theta \in \Theta\}$ e stiamo cercando "lo" stimatore T^* "ottimo" per $\kappa(\theta)$, cioè stiamo cercando "lo" stimatore T^* che soddisfa le seguenti proprietà:

1. T^* è non distorto per $\kappa(\theta)$;

2. $\mathrm{Var}_\theta(T^*) \leq \mathrm{Var}_\theta(T)$ per ogni θ e per ogni stimatore T non distorto e a varianza finita.

Definizione 3.1 Uno stimatore T^* che gode delle proprietà 1. e 2. è detto *stimatore non distorto a varianza uniformemente minima* (*Uniform Minimum Variance Unbiased Estimator*).

In letteratura un tale stimatore T^* è indicato con l'acronimo UMVUE.

Vediamo ora alcune proprietà (e non proprietà) degli stimatori UMVUE, e cioè *unicità, simmetria, "nosense"*.

Proposizione 3.2 (Unicità dell'UMVUE) *Se uno stimatore UMVUE per $\kappa(\theta)$ esiste, allora esso è essenzialmente unico, cioè se T_1, T_2 sono entrambi UMVUE, allora $P_\theta(T_1 = T_2) = 1\ \forall \theta \in \Theta$.*

Dimostrazione Innanzitutto osserviamo che $\mathrm{E}(T_1 - T_2) = \kappa(\theta) - \kappa(\theta) = 0$. Rimane da dimostrare che $\mathrm{Var}(T_1 - T_2) = 0$. Infatti, per le proprietà della varianza abbiamo che $\mathrm{E}(T_1 - T_2) = 0$ e $\mathrm{Var}(T_1 - T_2) = 0$ implicano $P_\theta(T_1 - T_2 = 0) = 1$. Chiamiamo v il comune valore di $\mathrm{Var}(T_1), \mathrm{Var}(T_2)$. Allora

$$\mathrm{Var}(T_1 - T_2) = \mathrm{Var}(T_1) + \mathrm{Var}(T_2) - 2\,\mathrm{Cov}(T_1, T_2) = 2v\,\mathrm{Cov}(T_1, T_2) \geq 0$$

se e solo se $\mathrm{Cov}(T_1, T_2) \leq v$. D'altro canto, siccome $(T_1 + T_2)/2$ è stimatore non distorto di $\kappa(\theta)$, allora la sua varianza è maggiore di quella dei due stimatori UMVUE, cioè abbiamo

$$\mathrm{Var}((T_1 + T_2)/2) = (1/4)\,(\mathrm{Var}(T_1) + \mathrm{Var}(T_2) + 2\,\mathrm{Cov}(T_1, T_2)) = v/2 + \mathrm{Cov}(T_1, T_2)/2 \geq v\ .$$

Segue che $\mathrm{Cov}(T_1, T_2) \geq v$. Ma, $\mathrm{Cov}(T_1, T_2) \leq v$ e $\mathrm{Cov}(T_1, T_2) \geq v$ se e solo se $\mathrm{Cov}(T_1, T_2) = v$, da cui deduciamo $\mathrm{Var}(T_1 - T_2) = 0$. ∎

Non a caso, se esiste parlo "dello" stimatore UMVUE, e non "di uno" stimatore UMVUE.

Proposizione 3.3 (Simmetria dell'UMVUE) *Sia $T^* = g(X_1, \ldots, X_n)$ UMVUE, allora*

$$P_\theta(g(X_1, \ldots, X_n) = g(X_{\pi(1)}, \ldots, X_{\pi(n)})) = 1 \quad \forall \theta \in \Theta \tag{6}$$

per ogni permutazione π di $\{1, \ldots, n\}$.

Dimostrazione Si rimanda al Pestman (1998) pagina 113 (Teorema II.9.6). ∎

In sostanza, la Proposizione 3.3 ci dice che se cambia l'ordine con cui le osservazioni arrivano, il valore dello stimatore UMVUE non cambia.

Uno stimatore che soddisfa la proprietà (6) è detto *essenzialmente simmetrico*. Praticamente, uno stimatore è simmetrico quando tutte le osservazioni hanno pari dignità (peso). Per esempio, la media campionaria $\bar{X}$ e la varianza campionaria S^2 sono stimatori simmetrici. Sulla classe degli stimatori simmetrici non diremo altro.

Nosense. Infine notiamo che lo stimatore UMVUE potrebbe esistere ma essere insensato. Torniamo all'Esempio 2.8 in cui, in un modello di Poisson, abbiamo ottenuto $T = (-1)^{X_1}$ come unico stimatore non distorto di $\mathrm{e}^{-2\theta}$. Ovviamente $T = (-1)^{X_1}$ ha varianza, quindi $T = (-1)^{X_1}$ è l'unico UMVUE per $\mathrm{e}^{-2\theta}$. Abbiamo già osservato l'inutilità di T.

Per completare l'analisi degli UMVUE, rimarrebbe ora da discutere l'esistenza e i metodi di ricerca degli stimatori UMVUE, cioè vedere operativamente come gli UMVUE si costruiscono. Dovremmo quindi introdurre la teoria delle statistiche "sufficienti e complete" e almeno i Lemmi di Rao Blackwell e Lehmann-Scheffé. Ma questi risultati richiedono lo sforzo teorico probabilistico di introdurre le medie condizionate. Pigramente, abbiamo deciso di non trattare questo capitolo in questo corso. Si rimanda l'allievo interessato a Rohatgi e Saleh (1999). La sufficienza è anche trattata in Pestman (1998).

Comunque, riusciremo a stabilire in alcuni casi particolari ma rilevanti se uno stimatore è UMVUE, lavorando sul confine inferiore della varianza degli stimatori non distorti in modelli statistici che soddisfano opportune condizioni di regolarità. Prima di approfondire questo argomento noto come "Disuguaglianza di Fréchet-Cramer-Rao", concludiamo la sezione con qualche accenno alle proprietà asintotiche degli stimatori.

4 Alcune proprietà asintotiche degli stimatori

Supponiamo di poter ripetere l'esperimento un numero arbitrariamente grande di volte e sempre nelle stesse condizioni. Abbiamo ora una successione di variabili aleatorie $X_1, \ldots, X_n, \ldots$ con X_n che descrive il risultato (aleatorio) dell'n-esima prova.
Cosa succede all'aumentare del numero di osservazioni a nostra disposizione? Ci aspettiamo che un campione più numeroso sia più rappresentativo della f.d.r. sottostante al campione e quindi, se T è un "buon" stimatore della caratteristica $\kappa(\theta)$, allora, all'aumentare di n, T sarà più prossimo a $\kappa(\theta)$. Maggiore prossimità di T a $\kappa(\theta)$, all'aumentare di n, può per esempio essere intesa nelle seguenti due accezioni matematiche:

Definizione 4.1 Sia $X_1, \ldots, X_n, \ldots$ una successione di variabili aleatorie i.i.d. con comune funzione di densità $f(x, \theta)$, $\theta \in \Theta$ e sia T_n uno stimatore di $\kappa(\theta)$ che è funzione delle prime n osservazioni. La successione $\{T_n\}_n$ è *asintoticamente non distorta* per $\kappa(\theta)$ se

$$\lim_{n\to\infty} \mathrm{E}_\theta(T_n) = \kappa(\theta) \quad \forall \theta \in \Theta$$

Definizione 4.2 Sia $X_1, \ldots, X_n, \ldots$ una successione di variabili aleatorie i.i.d. con comune funzione di densità $f(x, \theta)$, $\theta \in \Theta$ e sia T_n uno stimatore di $\kappa(\theta)$ che è funzione delle prime n osservazioni. La successione $\{T_n\}_n$ è *consistente in media quadratica* per $\kappa(\theta)$ se

$$\lim_{n\to\infty} \mathrm{E}[(T_n - \kappa(\theta))^2] = 0 \quad \forall \theta \in \Theta$$

Osservazione 4.3 Sappiamo dall'Equazione (3), che il MSE di (T_n) si decompone nella somma di varianza e quadrato della distorsione, che sono entrambi termini non negativi. Segue allora che la consistenza in media quadratica di una successione di stimatori $\{T_n\}_n$ è equivalente a non distorsione asintotica unita a varianza asintoticamente nulla.
Allora, dal punto di vista pratico, per verificare la consistenza di $\{T_n\}_n$ per $\kappa(\theta)$ procederemo a verificare *a*) $\lim_{n\to\infty} \mathrm{E}_\theta(T_n) = \kappa(\theta)$ e *b*) $\lim_{n\to\infty} \mathrm{Var}_\theta(T_n) = 0$.

Infine, la proprietà di gaussianità asintotica riguarda la f.d.r. limite di una sequenza di stimatori $\{T_n\}$.

Definizione 4.4 Sia $X_1, \ldots, X_n, \ldots$ una successione di variabili aleatorie i.i.d. con comune funzione di densità $f(x, \theta)$, $\theta \in \Theta$ e sia T_n una statistica funzione soltanto delle prime n osservazioni. La successione $\{T_n\}_n$ è *asintoticamente gaussiana* (o normale) con *media asintotica* $\mu_n(\theta)$ e *varianza asintotica* $\sigma_n^2(\theta)$ se

$$\lim_{n\to\infty} P\left(\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \leq z\right) = \Phi(z), \qquad \forall z \in \mathbb{R}$$

La proprietà di normalità asintotica è utile in presenza di grandi campioni per approssimare la distribuzione di uno stimatore T_n di una caratteristica $\kappa(\theta)$. Infatti, se la statistica T_n è usata come stimatore di una caratteristica $\kappa(\theta)$ e la successione $\{T_n\}$ è sia asintoticamente non distorta che asintoticamente gaussiana, allora $\mu_n(\theta) = \kappa(\theta)$ e approssimativamente la f.d.r. di T_n è gaussiana di media la caratteristica da stimare e varianza un'opportuna quantità σ_n^2.

5 Funzione di verosimiglianza

In questa sezione introduciamo la nozione di funzione di verosimiglianza e la esemplifichiamo per alcuni modelli notevoli.

Siano $X_1, \ldots, X_n$ n osservazioni i.i.d. estratte dalla popolazione di densità $f(\cdot, \theta), \theta \in \Theta$. Poiché le osservazioni sono indipendenti, la densità congiunta $f(x_1, \ldots, x_n)$ del vettore $(X_1, \ldots, X_n)$ coincide con il prodotto delle densità di ciascuna osservazioni, cioè $f(x_1, \ldots, x_n) = f(x_1, \theta) \cdots f(x_n, \theta)$.

Definizione 5.1 La *funzione di verosimiglianza* (*Likelihood function*) di n variabili aleatorie $X_1, \ldots, X_n$ è data dalla funzione di densità congiunta di $X_1, \ldots, X_n$, considerata come funzione di θ. Se $X_1, \ldots, X_n$ è un campione casuale estratto dalla densità $f(x, \theta)$, $\theta \in \Theta$, la funzione di verosimiglianza è

$$\theta \mapsto L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, \theta)$$

Leggiamo nella definizione che, per data realizzazione campionaria $x_1, \ldots, x_n$, la funzione di verosimiglianza è la densità congiunta del campione casuale $X_1, \ldots, X_n$ calcolata in $x_1, \ldots, x_n$, ma letta come funzione di θ. Nell'*impostazione frequentista* (che noi stiamo seguendo), nella funzione di verosimiglianza $L_\theta(x_1, \ldots, x_n)$ è sintetizzata tutta l'informazione sperimentale necessaria allo statistico per fare inferenza sul vero, unico ma incognito valore di θ o di alcune "caratteristiche" del fenomeno $\kappa(\theta)$. Praticamente, nell'impostazione frequentista lo statistico per fare inferenza ha a disposizione soltanto le realizzazioni campionarie $x_1, \ldots, x_n$ e la funzione di verosimiglianza, $L_\theta(x_1, \ldots, x_n)$.

La situazione cambia nell'*impostazione bayesiana*, quando ignoranza (e incertezza) sul parametro θ è tradotta matematicamente modellando θ come una variabile aleatoria (vettore aleatorio nel caso multidimensionale.) In questo corso non ci occuperemo di questa impostazione, ma trovate un breve cenno alla statistica Bayesiana anche nel vostro libro di testo, Pestman (1998), pagg. 97-103.

Determiniamo ora la funzione di verosimiglianza per qualche modello statistico "notevole".

Esempio 5.2 (Modello gaussiano) Consideriamo un campione casuale gaussiano con media e varianza entrambe incognite: $X_1, \ldots, X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$. Allora, $\theta = (\mu, \sigma^2)$ è un

parametro bidimensionale a valori nello spazio parametrico $\Theta = \mathbb{R} \times (0, \infty)$. La funzione di verosimiglianza $L_{\mu,\sigma^2}(x_1, \ldots, x_n)$ è data da

$$\begin{aligned} L_{\mu,\sigma^2}(x_1, \ldots, x_n) &= \prod_{j=1}^{n} f(x_j, \mu, \sigma^2) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{\sum_{j=1}^{n}(x_j - \mu)^2}{2\sigma^2}\Big\} \\ &= \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}\Big\} \\ &= \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}\Big\} \end{aligned} \tag{7}$$

dove $s^2 = \sum_{j=1}^{n}(x_j - \bar{x})^2/(n-1)$. Notate che la verosimiglianza L_{μ,σ^2} dipende dalle osservazioni solo tramite le realizzazioni di media e varianza campionarie.

Esempio 5.3 (Modello bernoulliano) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta)$ con

$$f(x, \theta) = \theta \mathbf{1}_{\{1\}}(x) + (1 - \theta)\mathbf{1}_{\{0\}}(x) = \theta^x (1 - \theta)^{1-x} \mathbf{1}_{\{0,1\}}(x) \quad \theta \in \Theta = [0, 1]$$

La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, \theta) = \theta^{\sum_{j=1}^{n} x_j} (1 - \theta)^{n - \sum_{j=1}^{n} x_j} = \theta^{n\bar{x}} (1 - \theta)^{n - n\bar{x}} \quad x_1, \ldots, x_n = 0, 1$$

Esempio 5.4 (Modello esponenziale) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{E}(\theta)$, $\theta > 0$ ($\mathcal{E}(\theta)$ = esponenziale di parametro θ con la parametrizzazione in (2)). La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, \theta) = \frac{1}{\theta^n} \mathrm{e}^{-\frac{\sum_{j=1}^{n} x_j}{\theta}} = \frac{1}{\theta^n} \mathrm{e}^{-\frac{n\bar{x}}{\theta}}, \quad x_1, \ldots, x_n > 0$$

Notate che nel modello bernoulliano e nel modello esponenziale L_θ dipende dalle osservazioni solo tramite la realizzazione della media campionaria $\bar{x}$.

Esercizio 5.5 (Modello gamma) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità $\Gamma(a, \beta)$, cioè

$$f(x, a, \beta) = \frac{1/\beta^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-x/\beta} \mathbf{1}_{(0,+\infty)}(x) \quad (a, \beta) \in \Theta = (0, +\infty)^2$$

Determinate la funzione di verosimiglianza.

Esempio 5.6 (Modello di Poisson) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità di Poisson di parametro θ incognito; la densità è

$$f(x, \theta) = \frac{\mathrm{e}^{-\theta} \theta^x}{x!} \mathbf{1}_{\{0,1,2,\ldots\}}(x) \quad \theta \in \Theta = (0, \infty)$$

La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, \theta) = \frac{\mathrm{e}^{-n\theta} \theta^{\sum_{j=1}^{n} x_j}}{\prod_{j=1}^{n} x_j!} = \frac{\mathrm{e}^{-n\theta} \theta^{n\bar{x}}}{\prod_{j=1}^{n} x_j!} \quad x_1, \ldots, x_n \in \{0, 1, 2, \ldots\} \tag{8}$$

Esempio 5.7 (Modello uniforme) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta)$ con

$$f(x, \theta) = \frac{1}{\theta} \mathbf{1}_{(0,\theta)}(x) \quad \theta \in \Theta = (0, \infty)$$

La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, \theta) = \frac{1}{\theta^n} \mathbf{1}_{(0,\theta)}(x_1) \cdots \mathbf{1}_{(0,\theta)}(x_n) = \frac{1}{\theta^n} \mathbf{1}_{(0,\theta)}(x_{(n)}) \quad x_{(1)} > 0$$

$$= \frac{1}{\theta^n} \mathbf{1}_{(x_{(n)}, +\infty)}(\theta) \quad x_1, \ldots, x_n > 0 \tag{9}$$

dove $x_{(n)} = \max\{x_1, \ldots, x_n\}$. Notate che in questo caso la verosimiglianza dipende dalle realizzazioni solo tramite il massimo delle osservazioni $x_{(n)}$.

Esercizio 5.8 (Modelli uniformi)

(*a*) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta)$ con

$$f(x, \theta) = -\frac{1}{\theta} \mathbf{1}_{(\theta,0)}(x) \quad \theta \in \Theta = (-\infty, 0)$$

Determinate la funzione di verosimiglianza.

(*b*) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x, \theta_1, \theta_2)$ con

$$f(x, \theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} \mathbf{1}_{(\theta_1,\theta_2)}(x) \quad (\theta_1, \theta_2) \text{ t.c. } -\infty < \theta_1 < \theta_2 < +\infty$$

Determinate la funzione di verosimiglianza.

6 Disuguaglianza di Fréchet-Cramer-Rao

In questa sezione torniamo ad affrontare il problema della ricerca di uno stimatore ottimale. Abbiamo già introdotto l'errore quadratico medio come misura della bontà di uno stimatore (quanto minore è l'MSE, tanto migliore è lo stimatore) e abbiamo visto come nella classe degli stimatori non distorti il MSE si riduce alla varianza cosicché il problema della minimizzazione del MSE diventa il problema della minimizzazione della varianza. Ci chiediamo ora se esiste un confine inferiore (lower bound) della varianza nella classe di tutti gli stimatori non distorti, che sia funzione soltanto della caratteristica da stimare $\kappa(\theta)$ e del modello statistico mediante la verosimiglianza L_θ. Inoltre, se questo confine inferiore per la varianza esiste, ci chiediamo se sia possibile costruire uno stimatore che abbia varianza coincidente con esso. Se ciò è possibile, allora quello stimatore realizzerà il programma di ricerca dell'UMVUE.

Entrambe le domande hanno risposta positiva se sono soddisfatte opportune condizioni di regolarità fissate nel teorema sulla disuguaglianza di Fréchet-Cramer-Rao, o dell'informazione di Fisher.

Per semplicità espositiva in questa sezione presentiamo la disuguaglianza di Fréchet-Cramer-Rao nel caso di uno spazio parametrico Θ unidimensionale. Qualche accenno alla versione multidimensionale della disuguaglianza di Fréchet-Cramer-Rao sarà dato nella Sezione 6.2.

Teorema 6.1 *Siano $X_1, \dots, X_n$ variabili aleatorie i.i.d. con comune densità $f(x,\theta)$, $\theta \in \Theta \subset \mathbb{R}$ e sia $T = g(X_1, \dots, X_n)$ uno stimatore non distorto della caratteristica $\kappa(\theta)$ a varianza finita. Assumiamo che le seguenti condizioni di regolarità siano soddisfatte:*

(i) Θ *è un intervallo aperto di* $\mathbb{R}$*;*

(ii) $S = \{x : f(x,\theta) > 0\}$ *è indipendente da* θ*;*

(iii) $\theta \mapsto f(x,\theta)$ *è derivabile su* Θ*,* $\forall x \in S$*;*

(iv) $\mathrm{E}_\theta\left(\dfrac{\partial}{\partial\theta}\log f(X_1,\theta)\right) = 0 \quad \forall\theta \in \Theta$*;*

(v) $0 < \mathrm{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta}\log f(X_1,\theta)\right)^2\right] < \infty \quad \forall\theta \in \Theta$*;*

(vi) $\kappa : \Theta \to \mathbb{R}$ *è derivabile su* Θ *e*

$$\kappa'(\theta) = \mathrm{E}_\theta\left(T\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n)\right) \quad \forall\theta \in \Theta$$

Allora

$$\mathrm{Var}_\theta(T) \geq \frac{(\kappa'(\theta))^2}{nI(\theta)} \quad \forall\theta \in \Theta \tag{10}$$

dove

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right)^2\right]$$

Inoltre, l'uguaglianza in (10) *vale se e solo se esiste una funzione* $a(n,\theta)$ *tale che*

$$P_\theta\left(\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n) = a(n,\theta)(T-\kappa(\theta))\right) = 1 \quad \forall\theta \in \Theta \tag{11}$$

Osservazione 6.2 Tutte le ipotesi del Teorema 6.1 sono condizioni di regolarità sulla famiglia di densità $\{f(x,\theta), \theta \in \Theta\}$; in particolare l'ipotesi (ii) insieme a opportune condizioni di derivazione sotto segno di integrale di $f(x,\theta)$ garantiscono il soddisfacimento delle ipotesi (iv) e (vi). Cerchiamo di spiegare il perché sviluppando i conti nel caso di un modello assolutamente continuo. I conti per il caso discreto procedono analogamente con le somme/serie al posto degli integrali.

Partiamo dalla condizione di normalizzazione

$$1 = \int_{\mathbb{R}} f(x,\theta)\,dx = \int_S f(x,\theta)\,dx$$

Supponiamo ora che la funzione (in θ) $\int_S f(x,\theta)\,dx$ sia derivabile sotto segno di integrale, cioè

$$\frac{\partial}{\partial\theta}\int_S f(x,\theta)\,dx = \int_S \frac{\partial}{\partial\theta} f(x,\theta)\,dx \tag{12}$$

Allora,

$$\frac{\partial}{\partial\theta}\int_S f(x,\theta)\,dx = 0$$

Quindi

$$0 = \frac{\partial}{\partial\theta}\int_S f(x,\theta)\,dx = \int_S \frac{\partial}{\partial\theta} f(x,\theta)\,dx = \int_S \frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)} f(x,\theta)\,dx \quad [\text{perché } f(x,\theta) > 0 \text{ su } S]$$
$$= \int_S \frac{\partial}{\partial\theta}\log f(x,\theta) f(x,\theta)\,dx = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right)$$

cioè (12) implica (iv).

Analogamente, l'ipotesi (vi) può essere tradotta in termini di derivazione sotto segno di integrale della funzione $\int_{S^n} g(x_1,\dots,x_n)L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n$ come segue:

$$(13)\quad \frac{\partial}{\partial\theta}\int_{S^n} g(x_1,\dots,x_n)L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n =$$
$$= \int_{S^n} g(x_1,\dots,x_n)\frac{\partial}{\partial\theta}L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n$$

Infatti, per la non distorsione abbiamo

$$\kappa(\theta) = \int_{S^n} g(x_1,\dots,x_n)\prod_{j=1}^n f(x_j,\theta)\,dx_1\cdots dx_n = \int_{S^n} g(x_1,\dots,x_n)L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n$$

quindi, se (13) è vera, allora

$$\begin{aligned}
\kappa'(\theta) &= \frac{\partial}{\partial\theta}\int_{S^n} g(x_1,\dots,x_n)L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n \\
&= \int_{S^n} g(x_1,\dots,x_n)\frac{\partial}{\partial\theta}L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n \\
&= \int_{S^n} g(x_1,\dots,x_n)\frac{\frac{\partial}{\partial\theta}L_\theta(x_1,\dots,x_n)}{L_\theta(x_1,\dots,x_n)}L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n \\
&= \int_{\mathbb{R}^n} g(x_1,\dots,x_n)\Big[\frac{\partial}{\partial\theta}\log L_\theta(x_1,\dots,x_n)\Big]\Big(\prod_{i=1}^n f(x_i,\theta)\Big)\,dx_1\cdots dx_n \\
&= \mathrm{E}_\theta\Big[T\frac{\partial}{\partial\theta}\log L_\theta(X_1,\dots,X_n)\Big]
\end{aligned}$$

Nota 6.3 La dimostrazione del Teorema 6.1 si basa su alcune proprietà della covarianza che qui riprendiamo. La *covarianza* di due variabili aleatorie X, Y a varianza finita è data da $\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))]$ e si può calcolare come $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$. La covarianza gode delle seguenti proprietà:

(j) $|\mathrm{Cov}(X,Y)| \le \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$;

(jj) $|\mathrm{Cov}(X,Y)| = \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$ se e solo se esistono $a, b \in \mathbb{R}$ tali che $P(Y = aX+b) = 1$. Inoltre, se $\mathrm{Var}(X) > 0$ allora $a = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$ e $b = \mathrm{E}(Y) - a\,\mathrm{E}(X)$

(jjj) se $\mathrm{E}(Y) = 0$ allora $b = -a\,\mathrm{E}(X)$ e $P(Y = a(X-\mathrm{E}(X))) = 1$

Dimostrazione del Teorema 6.1 Per maggiore semplicità notazionale, introduciamo le variabili aleatorie $Y_1, \ldots, Y_n$ definite da

$$Y_j = \frac{\partial}{\partial\theta} \log f(X_j, \theta), \quad \forall j = 1, \ldots, n$$

$Y_1, \ldots, Y_n$ sono variabili aleatorie i.i.d. a media nulla e varianza finita $I(\theta)$ $\forall\theta$. Infatti

$$\mathrm{E}_\theta(Y_j) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta} \log f(X_j, \theta)\right) = 0 \qquad \text{[per l'ipotesi } (iv)\text{]}$$

e

$$\mathrm{Var}_\theta(Y_j) = \mathrm{E}_\theta(Y_j^2) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(X_j, \theta)\right)^2\right] = I(\theta) \in (0, \infty) \quad \text{[per l'ipotesi } (v)\text{]}$$

Inoltre

$$\frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n) = \frac{\partial}{\partial\theta} \log \prod_{j=1}^n f(X_j, \theta) = \sum_{j=1}^n \frac{\partial}{\partial\theta} \log f(X_j, \theta) = \sum_{j=1}^n Y_j$$

da cui deriviamo che

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)\right) = \sum_{j=1}^n \mathrm{E}_\theta(Y_j) = 0 \tag{14}$$

e

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)\right) = \sum_{j=1}^n \mathrm{Var}_\theta(Y_j) = nI(\theta) \tag{15}$$

L'ipotesi (vi) e l'equazione (14) insieme forniscono

$$\kappa'(\theta) = \mathrm{E}_\theta\left(T \frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)\right) = \mathrm{Cov}\left(T, \frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)\right)$$

-0 -0

cosicché

$$\begin{aligned}
(\kappa'(\theta))^2 &= \left(\mathrm{Cov}\left(T, \frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)\right)\right)^2 \\
&\leq \mathrm{Var}_\theta(T)\, \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)\right) \qquad \text{[per la proprietà } (j) \text{ della covarianza]} \\
&= \mathrm{Var}_\theta(T) nI(\theta) \qquad \text{[per l'equazione (15)]}
\end{aligned}$$

La disuguaglianza (10) segue in virtù dell'ipotesi (v).

Infine, per verificare che la condizione (11) è necessaria e sufficiente perché valga l'uguaglianza in (10), applichiamo la proprietà (jjj) della covarianza alle variabili $X = T$ e $Y = \frac{\partial}{\partial\theta} \log L_\theta(X_1, \ldots, X_n)$. ∎

Definizione 6.4 $I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right)^2\right]$ è detta *informazione di Fisher su* θ basata su una singola osservazione e $I_n(\theta) = \left[\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right)^2\right]$ è l'*informazione di Fisher su* θ fornita dall'intero campione.

Nella dimostrazione del Teorema 6.1 abbiamo fatto vedere che l'informazione di Fisher di un campione casuale su θ è n volte quella di ogni singola osservazione, cioè $I_n(\theta) = nI(\theta)$.

Osservazione 6.5 Il Teorema 6.1 ci dice che se il modello statistico è "regolare" allora ogni stimatore T di θ non distorto ha varianza non inferiore al reciproco dell'informazione di Fisher del campione $nI(\theta)$. Se poi esiste uno stimatore T^* non distorto la cui varianza raggiunge il confine inferiore di Fréchet-Cramer-Rao, allora quanto maggiore sarà $I(\theta)$ tanto minore sarà la varianza di T^* e quindi maggiori informazioni si avranno su θ. Da qui il nome di informazione dato a $I(\theta)$.

Definizione 6.6 Uno stimatore T^* non distorto la cui varianza raggiunge il confine inferiore di Fréchet-Cramer-Rao è detto *efficiente*.

È importante ribadire che:

a) *Se uno stimatore T efficiente esiste,* ovviamente *esso è anche UMVUE.*

b) *Con un campione casuale, la varianza dello stimatore efficiente è inversamente proporzionale al numero di osservazioni nel campione.*

Esempio 6.7 Sia $X_1,\ldots,X_n$ un campione casuale estratto dalla popolazione esponenziale $\mathcal{E}(\theta)$, $\theta > 0$, ($\mathcal{E}(\theta)$ = esponenziale di parametro θ con la parametrizzazione (2)). Le condizioni $(i)-(iii)$ del Teorema 6.1 sono banalmente soddisfatte perché $\Theta = \{x : f(x,\theta) > 0\} = (0,\infty)$ e $f(x,\theta) = (1/\theta)\mathrm{e}^{-x/\theta}$. Essendo

$$\frac{\partial}{\partial\theta}\log f(X_1,\theta) = -\frac{1}{\theta} + \frac{X_1}{\theta^2}$$

allora

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right) = \mathrm{E}_\theta\left(-\frac{1}{\theta} + \frac{X_1}{\theta^2}\right) = -\frac{1}{\theta} + \frac{\theta}{\theta^2} = 0$$

e

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X_1,\theta)\right)^2\right] = \mathrm{Var}_\theta\left(-\frac{1}{\theta} + \frac{X_1}{\theta^2}\right) = \frac{\mathrm{Var}_\theta(X_1)}{\theta^4} = \frac{1}{\theta^2} \in (0,\infty) \quad \forall\theta > 0$$

e quindi anche le ipotesi (iv) e (v) sono soddisfatte. Il confine di Fréchet-Cramer-Rao per la caratteristica $\kappa_1(\theta) = \theta$ è $(\kappa_1'(\theta))^2/(nI(\theta)) = \theta^2/n$ che coincide con la varianza di $\bar{X}$. Già sappiamo che $\bar{X}$ è stimatore non distorto di θ. Inoltre,

$$\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n) = \sum_{j=1}^{n}\left(-\frac{1}{\theta} + \frac{X_j}{\theta^2}\right) = -\frac{n}{\theta} + \frac{n\bar{X}}{\theta^2}$$

quindi

$$\mathrm{E}_\theta\left(\bar{X}\left(\frac{\partial}{\partial\theta}\log L_\theta(X_1,\ldots,X_n)\right)\right) = -n + \frac{n}{\theta^2}\left(\frac{\theta^2}{n} + \theta^2\right) = 1 = \kappa_1'(\theta)$$

cosicché lo stimatore $\bar{X}$ soddisfa l'ipotesi (vi). Concludiamo che la media campionaria è stimatore efficiente per la media teorica nel modello esponenziale.

Se invece la caratteristica da stimare è la varianza θ^2, allora il confine di Fréchet-Cramer-Rao per $\kappa_2(\theta) = \theta^2$ è

$$\frac{(\kappa_2(\theta)')^2}{nI(\theta)} = \frac{(2\theta)^2}{n/\theta^2} = \frac{4\theta^4}{n}$$

e dalla seconda parte del Teorema 6.1 sappiamo che esso è raggiungibile da uno stimatore non distorto T se e solo se esiste $a(n,\theta)$ tale che

$$P_\theta\left(-\frac{n}{\theta} + \frac{n\bar{X}}{\theta^2} = a(n,\theta)(T-\theta^2)\right) = 1$$

Ma la precedente equazione è soddisfatta se e solo se $a(n,\theta) = n/\theta^3$ e simultaneamente $a(n,\theta) \propto b(n)/\theta^2$: assurdo! Concludiamo che uno stimatore efficiente per la varianza non esiste.

Fidatevi che $(n+1)\bar{X}^2/n$ è stimatore UMVUE di θ^2 (non lo dimostreremo.)

Esercizio 6.8 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità di Poisson di parametro θ incognito.

1. Dimostrate che $\bar{X}$ è stimatore efficiente della media.
2. Verificate che non esiste lo stimatore efficiente per $\kappa(\theta) = P(X_1 = 0)$.

Le ipotesi di regolarità richieste nel Teorema 6.1 sono tutte necessarie. Se qualcuna di esse non è soddisfatta non possiamo concludere sul confine inferiore delle varianze della classe degli stimatori non distorti (a varianza finita). Anzi, come mostra il seguente esempio su un modello unifome continuo, potremmo trovare uno stimatore non distorto che ha varianza tendente a zero come $1/n^2$ anzicché $1/n$.

Esempio 6.9 Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim f(x;\theta) = \mathcal{U}(0,\theta)$, $\theta > 0$ cioè

$$f(x,\theta) = \frac{1}{\theta}\mathbf{1}_{(0,\infty)}(x)\mathbf{1}_{[x,\infty]}(\theta), \quad \theta > 0,$$

Questo modello viola alcune condizioni di regolarità; per esempio osserviamo che $\{x : f(x,\theta) > 0\} = (0,\theta)$ e quindi la condizione (ii) del Teorema 6.1 non vale; ancora:

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = -\frac{1}{\theta}$$

da cui segue che

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right) = -\frac{1}{\theta} \neq 0 \qquad \forall\theta > 0,$$

e quindi anche la condizione (iv) è violata.

Siano $X_{(n)} = \max\{X_1, \ldots, X_n\}$ e $T = (n+1)X_{(n)}/n$. La f.d.r. di $X_{(n)}$ è

$$F_{X_{(n)}}(x,\theta) = P_\theta(X_{(n)} \leq x) = P_\theta(X_1 \leq x, X_2 \leq x, \cdots, X_n \leq x) = \prod_{j=1}^{n} P_\theta(X_j \leq x)$$

$$= P_\theta^n(X_1 \leq x) = F_{X_1}^n(x,\theta) = \begin{cases} 0 & \text{se } x \leq 0 \\ \left(\frac{x}{\theta}\right)^n & \text{se } 0 < x < \theta \\ 1 & \text{se } x \geq \theta \end{cases}$$

Inoltre,

$$\frac{\partial F_{X_{(n)}}(x,\theta)}{\partial x} = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{se } x \in (0,\theta) \\ 0 & \text{se } x < 0 \text{ oppure } x > \theta \end{cases}$$

e quindi la funzione di densità di $X_{(n)}$ è

$$f_{X_{(n)}}(x,\theta) = \frac{nx^{n-1}}{\theta^n}\mathbf{1}_{(0,\theta)}(x)$$

Segue che

$$\mathrm{E}_\theta(X_{(n)}) = \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n}\, dx = \frac{n\theta}{n+1}$$

e

$$\begin{aligned}\mathrm{Var}_\theta(X_{(n)}) &= \mathrm{E}_\theta(X_{(n)}^2) - \left(\frac{n\theta}{n+1}\right)^2 = \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n}\, dx - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}\end{aligned}$$

Deduciamo che T è stimatore non distorto di θ e ha varianza

$$\mathrm{Var}_\theta(T) = \frac{(n+1)^2}{n^2} \times \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$$

Abbiamo così trovato uno stimatore non distorto T la cui varianza per n grande converge a zero come $1/n^2$, quindi molto più velocemente dell'"usuale" tasso $1/n$.

6.1 Famiglia esponenziale. Accenni

Abbiamo visto nell'Esempio 6.7 un modello statistico regolare (modello esponenziale), per cui lo stimatore UMVUE di una caratteristica (la varianza) esiste ma la sua varianza non raggiunge il confine di Fréchet-Cramer-Rao. Il fatto che esista l'UMVUE ma non l'efficiente non è l'eccezione. D'altro canto, l'esempio visto non era patologico... Notate inoltre che per stabilire l'inesistenza dello stimatore efficiente abbiamo usato il Teorema di Fréchet-Cramer-Rao. Effettivamente questo teorema è uno strumento potente e operativo per investigare gli stimatori efficienti. A leggerlo bene, esso fornisce indicazioni strette sulle famiglie di densità $\{f(x,\theta), \theta \in \Theta\}$ e sulle caratteristiche $\kappa(\theta)$ per cui uno stimatore efficiente esiste. Questa è l'ultima parte del teorema. Infatti, se integriamo l'equazione (11) in θ, otteniamo che $\forall \theta \in \Theta$ con probabilità 1

$$\log L_\theta(X_1,\ldots,X_n) = \int_0^\theta a(n,u)(g(X_1,\ldots,X_n) - \kappa(u))\, du + c(X_1,\ldots,X_n)$$

Equivalentemente stiamo dicendo che $\forall \theta \in \Theta$ con probabilità 1 $L_\theta(X_1,\ldots,X_n)$ è della forma

$$L_\theta(X_1,\ldots,X_n) = C(X_1,\ldots,X_n)\exp\{A(n,\theta)g(X_1,\ldots,X_n) + B(n,\theta)\}$$

Definizione 6.10 La famiglia di densità $\{f(x,\theta), \theta \in \Theta \subset \mathbb{R}\}$ definita da

$$f(x,\theta) = C(x)\exp\{A(\theta)g(x) + B(\theta)\} \tag{16}$$

è detta *famiglia esponenziale*.

Si può dimostrare in modo rigoroso che

Metateorema 6.11 *Gli stimatori efficienti esistono solo per le caratteristiche $\kappa(\theta)$ della famiglia esponenziale $f(x,\theta) = C(x)\exp\{A(\theta)g(x)+B(\theta)\}$ del tipo $\kappa(\theta) = -cB'(\theta)/A'(\theta)+d$. Inoltre questi stimatori hanno forma $(c/n)\sum_{j=1}^n g(X_j) + d$, $c, d \in \mathbb{R}$.*

(Metateorema perché manca qualche ipotesi per poterlo enunciare e dimostrare rigorosamente).

Esercizio* 6.12 (difficile? se si, saltate...) Dimostrate il Metateorema 6.11.

Esempio 6.13 (Esercizio 2.6 pag. 44 in Silvey (1975)) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x,\theta) = \frac{x+1}{\theta(\theta+1)}\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x),\ \theta > 0.$$

Determiniamo uno stimatore efficiente della caratteristica $\kappa(\theta) = (3+2\theta)(2+\theta)/(\theta+1)$.

Osserviamo che

$$\log L_\theta(x_1,\ldots,x_n) = \log\prod_{i=1}^n f(x_i,\theta) = \log\prod_{i=1}^n (x_i+1) - n\log[\theta(\theta+1)] - \frac{\sum_{j=1}^n x_j}{\theta}$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L_\theta(x_1,\ldots,x_n) &= -n\frac{2\theta+1}{\theta(\theta+1)} + \frac{\sum_{j=1}^n x_j}{\theta^2} = \frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n x_j}{n} - \frac{\theta(2\theta+1)}{\theta+1}\right) = \\ &= \frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n x_j}{n} \mp \kappa(\theta) - \frac{\theta(2\theta+1)}{\theta+1}\right) = \frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n (x_j+6)}{n} - \kappa(\theta)\right)\end{aligned} \tag{17}$$

Infatti $\kappa(\theta) - \theta(2\theta+1)/(\theta+1) = (3+2\theta)(2+\theta)/(\theta+1) - \theta(2\theta+1)/(\theta+1) = 6$. Inoltre

$$\mathrm{E}_\theta\left(\frac{\sum_{j=1}^n (X_j+6)}{n}\right) = \mathrm{E}_\theta(X_1) + 6 = \int_0^\infty \frac{x(x+1)}{\theta(\theta+1)}\mathrm{e}^{-x/\theta}\,dx + 6 = \kappa(\theta)$$

Segue da (17) che $T = (1/n)\sum_{j=1}^n (x_j+6)$ è lo stimatore efficiente per $\kappa(\theta)$.

Osservate che $f(x,\theta)$ appartiene alla famiglia esponenziale (16) con $g(x) = x$, $A(\theta) = -1/\theta$ e $B(\theta) = -\log(\theta(\theta+1))$ e quindi $-B'(\theta)/A'(\theta) + 6 = \kappa(\theta)$.

6.2 Disuguaglianza di Fréchet-Cramer-Rao nel caso multivariato

La disuguaglianza di Fréchet-Cramer-Rao può essere estesa al caso di un parametro m-dimensionale $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_m \end{pmatrix}$ con $m \geq 2$. Nel caso m-dimensionale, l'informazione di Fisher diventa una matrice $m \times m$ e al posto della varianza dello stimatore unidimensionale viene considerata la matrice di covarianza di uno stimatore m-dimensionale $T = \begin{pmatrix} T_1 \\ \vdots \\ T_m \end{pmatrix}$ del parametro θ. Si richiede quindi che ogni componente di T sia stimatore non distorto della corrispondente componente di θ, cioè $\mathrm{E}_\theta(T_j) = \theta_j \ \forall j = 1, \ldots$. Le condizioni di regolarità $(i) - (vi)$ sono facilmente estese al caso m-dimensionale: al posto della derivata prima $\frac{\partial}{\partial \theta} \log f(X_1, \theta)$ ora vi sarà il vettore delle derivate prime parziali

$$\begin{pmatrix} \frac{\partial}{\partial \theta_1} \log f(X_1, \theta) \\ \vdots \\ \frac{\partial}{\partial \theta_m} \log f(X_1, \theta) \end{pmatrix}$$

e la *matrice di informazione di Fisher* $I(\theta)$ sarà la matrice di covarianza di questo vettore. Se chiamiamo C_T la matrice di covarianza del vettore T, la disuguaglianza diventerà

$$\text{“}C_T - \frac{I(\theta)^{-1}}{n} \geq 0\text{”} \text{ nel senso che } C_T - \frac{I(\theta)^{-1}}{n} \text{ è matrice semidefinita positiva.}$$

L'estensione della disuguaglianza di Fréchet-Cramer-Rao al caso multivariato è utile per esempio per investigare l'efficienza degli stimatori $\bar{X}, S^2$ di media e varianza della popolazione gaussiana. Nel caso gaussiano abbiamo

$$I((\mu, \sigma^2)) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^2) \end{bmatrix}$$

che ha inversa

$$I^{-1}((\mu, \sigma^2)) = 2\sigma^6 \begin{bmatrix} 1/(2\sigma^4) & 0 \\ 0 & 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

mentre

$$C_{(\bar{X}, S^2)} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{bmatrix}$$

Concludiamo che $\bar{X}$ è efficiente per la media ma S^2 non è efficiente per σ^2. Non lo dimostreremo, ma S^2 è UMVUE per σ^2.

7 Metodi di ricerca degli stimatori

Nelle prossime due sezioni descriviamo due metodi di ricerca di stimatori: il metodo dei momenti e il metodo di massima verosimiglianza.

7.1 Metodo dei momenti

Il metodo dei momenti è il più antico metodo generale di ricerca di stimatori. Fu proposto da Karl Pearson nel 1894.

Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità $f(x, \theta)$ con $\theta = (\theta_1, \ldots, \theta_m) \in \Theta \subset \mathbb{R}^m$. Supponiamo che per ogni θ X_1 ammette i primi m *momenti* definiti da $\mu_1(\theta) := \mathrm{E}_\theta(X_1), \ldots, \mu_m(\theta) := \mathrm{E}_\theta(X_1^m)$. Per esempio nel caso continuo abbiamo che se $\int_{\mathbb{R}} |x^r| f(x, \theta)\, dx < \infty$ allora il momento r-esimo $\mu_r(\theta) = \mathrm{E}_\theta(X_1^r)$ esiste ed è dato da $\mu_r(\theta) = \int_{\mathbb{R}} x^r f(x, \theta)\, dx$. Introduciamo poi le m statistiche $M_1, \ldots, M_m$ date dai primi m *momenti campionari*

$$M_r = \frac{1}{n} \sum_{j=1}^{n} X_j^r \qquad r = 1, \ldots, m$$

In particolare, il primo momento campionario è la media campionaria: $M_1 = \bar{X}$. Consideriamo poi il sistema di m equazioni

$$\begin{cases} \mu_1(\theta) = M_1 \\ \ldots \\ \mu_m(\theta) = M_m \end{cases} \tag{18}$$

nelle m incognite $\theta_1, \ldots, \theta_m$. Se il sistema (18) ammette soluzione $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$, allora $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$ sono statistiche, perché dipendono soltanto dai momenti campionari $M_1, \ldots, M_m$. Possiamo quindi usare $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$ come stimatori rispettivamente di $\theta_1, \ldots, \theta_m$.

$\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ è lo stimatore di $\theta = (\theta_1, \ldots, \theta_m)$ ottenuto con il metodo dei momenti.

Ovviamente,

se $\mu_r(\theta)$ esiste allora M_r è stimatore non distorto della caratteristica $\kappa(\theta) = \mu_r(\theta)$.

Infatti:

$$\mathrm{E}_\theta(M_r) = \mathrm{E}_\theta\left(\frac{1}{n}\sum_{j=1}^{n} X_j^r\right) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}_\theta(X_j^r) = \mu_r(\theta).$$

Esempio 7.1 (Modello gaussiano) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$, $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Stimiamo μ e σ^2 con il metodo dei momenti. Essendo $\mu_1(\mu, \sigma^2) = \mu$ e $\mu_2(\mu, \sigma^2) = \mathrm{E}(X_1^2) = \mathrm{Var}(X_1) + \mathrm{E}^2(X_1) = \sigma^2 + \mu^2$, allora gli stimatori $\widehat{\mu}, \widehat{\sigma^2}$ ottenuti con il metodo dei momenti sono le soluzioni del sistema di equazioni

$$\begin{cases} \mu = \bar{X} \\ \sigma^2 + \mu^2 = M_2 \end{cases}$$

date da

$$\widehat{\mu} = \bar{X} \quad \text{e} \quad \widehat{\sigma^2} = M_2 - \bar{X}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2 = \frac{(n-1)S^2}{n}$$

Per inciso osservate che lo stimatore dei momenti per la varianza è distorto, infatti $\mathrm{E}(\widehat{\sigma^2}) = (n-1)\sigma^2/n$.

Esempio 7.2 (Modelli Uniformi)

(*a*) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{U}(0, \theta)$, con $\theta > 0$. Essendo $\mathrm{E}(X_1) = \theta/2$, allora lo stimatore dei momenti di θ è $\widehat{\theta} = 2\bar{X}$.

(*b*) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{U}(\theta_1, \theta_2)$, con $-\infty < \theta_1 < \theta_2 < \infty$. Essendo $\mathrm{E}(X_1) = (\theta_1 + \theta_2)/2$ e $\mathrm{Var}(X_1) = (\theta_2 - \theta_1)^2/12$, allora $\widehat{\theta}_1, \widehat{\theta}_2$ sono le soluzioni del sistema di equazioni

$$\begin{cases} \frac{\theta_1+\theta_2}{2} = M_1 \\ \frac{(\theta_2-\theta_1)^2}{12} + \frac{(\theta_2+\theta_1)^2}{4} = M_2 \end{cases} \iff \begin{cases} \frac{\theta_1+\theta_2}{2} = M_1 \\ (\theta_2-\theta_1)^2 = 12(M_2 - M_1^2) = \frac{12(n-1)S^2}{n} \end{cases}$$

che, considerato il vincolo $\theta_2 > \theta_1$, ha unica soluzione

$$\begin{cases} \widehat{\theta}_1 = \bar{X} - \sqrt{\frac{3(n-1)S^2}{n}} \\ \widehat{\theta}_2 = \bar{X} + \sqrt{\frac{3(n-1)S^2}{n}} \end{cases}$$

Osservazione 7.3 *Il metodo dei momenti presenta un grosso difetto di arbitrarietà* come esemplificato di seguito.

(*c*) Siano $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{U}(-\theta, \theta)$, con $\theta > 0$. Allora $\mathrm{E}_\theta(X_1) = 0\ \forall\theta$. Seguirebbe che lo stimatore dei momenti non esiste. A meno che non scegliamo di stimare θ a partire dall'equazione

$$\mathrm{E}_\theta(X_1^2) = \frac{\theta^2}{3} = M_2$$

da cui otteniamo $\widehat{\theta} = \sqrt{3M_2}$. Ma allora perché non partire da una qualunque altra equazione del tipo

$$\mathrm{E}_\theta(X_1^{2r}) = \frac{\theta^{2r}}{2r+1} = M_{2r}$$

per $r = 2, 3, \ldots$? Per ogni scelta di r ovviamente otteniamo un diverso stimatore di θ. Quindi, in casi come questo, il metodo non fornisce un'unica strada per scegliere quale equazione risolvere. Il suggerimento di Rohatgi e Saleh (1999) per evitare questo tipo di ambiguità è di scegliere gli stimatori che coinvolgono momenti di ordine inferiore.

7.2 Metodo di massima verosimiglianza

Un altro metodo utile e generale per costruire stimatori "ragionevoli" delle caratteristiche della popolazione è il *metodo di massima verosimiglianza*, proposto da Ronald Fisher nel 1921 (cfr. Fisher 1992).

Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità $f(x, \theta)$ con $\theta = (\theta_1, \ldots, \theta_m) \in \Theta \subset \mathbb{R}^m$ e sia $\theta \mapsto L_\theta$ la funzione di verosimiglianza. Per introdurre in modo intuitivo il metodo supponiamo inizialmente che $f(x, \theta)$ sia una densità di probabilità discreta. Se θ è noto e pari per esempio a θ_0, e se $f(x, \theta)$ è una densità di probabilità discreta, allora $L_{\theta_0}(x_1, \ldots, x_n) = P_{\theta_0}(X_1 = x_1, \ldots, X_n = x_n)$ è la probabilità di osservare $x_1, \ldots, x_n$. D'altro canto, se invece θ è incognito e $x_1, \ldots, x_n$ è l'effettiva realizzazione campionaria, allora è sensato scegliere il valore di θ in Θ che individua la densità $f(x, \theta)$ da cui è più verosimile provenga quanto effettivamente osservato, cioè $x_1, \ldots, x_n$. Altrimenti detto, scegliamo $\widehat{\theta}$ tale che $L_{\widehat{\theta}}(x_1, \ldots, x_n) = \max_{\theta\in\Theta} L_\theta(x_1, \ldots, x_n)$. Se tale $\widehat{\theta}$ esiste, esso è funzione solo di $x_1, \ldots, x_n$ e quindi può essere usato come una stima di θ basata su $x_1, \ldots, x_n$: $\widehat{\theta}$ è una stima di massima verosimiglianza di θ.

Diamo ora una definizione generale di stimatore di massima verosimiglianza, valida anche per modelli statistici continui.

Definizione 7.4 Siano $X_1, \ldots, X_n$ un campione casuale con funzione di verosimiglianza L_θ, $\theta \in \Theta$, $x_1, \ldots, x_n$ una realizzazione campionaria e $g(x_1, \ldots, x_n)$ un valore in Θ tale che

$$L_{g(x_1,\ldots,x_n)}(x_1, \ldots, x_n) = \max_{\theta \in \Theta} L_\theta(x_1, \ldots, x_n)$$

La statistica $\widehat{\theta} = g(X_1, \ldots, X_n)$ è detta *stimatore di massima verosimiglianza di* θ. Per indicare $\widehat{\theta}$ useremo l'acronimo ML (che sta per *Maximum Likelihood*) o MLE (*Maximum Likelihood Estimator*).

Se θ è unidimensionale, la funzione di verosimiglianza $L_\theta(x_1, \ldots, x_n)$ è continua (in θ) e Θ è un insieme chiuso e limitato, allora una stima ML di θ esiste. Ma, se questa condizione (sufficiente) non è soddisfatta, uno stimatore ML potrebbe anche non esistere perché la funzione di verosimiglianza non ha massimo (cfr Esempio 7.9). Può anche accadere che la stima di massima verosimiglianza non sia unica perché la funzione di verosimiglianza ha più punti di massimo (cfr Esempio 7.10). Per determinare un punto di massimo della funzione di verosimiglianza userete le tecniche standard. In particolare, potrà risultare utile notare che L_θ e il suo logaritmo $\log L_\theta$ ammettono gli stessi punti di massimo (dove L_θ è nulla, $\log L_\theta = -\infty$).

Considerazioni analoghe valgono se θ è un vettore m-dimensionale, con $m \geq 2$; nel caso di funzioni di verosimiglianza sufficientemente regolari, risolveremo il seguente sistema delle equazioni di verosimiglianza (*likelihood equations*):

$$\frac{\partial}{\partial \theta_j} \log L_\theta(x_1, \ldots, x_n) = 0\,, \qquad j = 1, \ldots, m$$

7.2.1 Proprietà di invarianza degli stimatori ML

Se siamo interessati a stimare una caratteristica della popolazione $\kappa(\theta)$, possiamo partire dalla *funzione di verosimiglianza indotta da* $\kappa(\theta)$ definita da

$$L^*_\kappa(x_1, \ldots, x_n) := \sup_{\{\theta \in \Theta\,:\,\kappa(\theta)=\kappa\}} L_\theta(x_1, \ldots, x_n)$$

In modo naturale definiamo *stimatore di massima verosimiglianza di* $\kappa(\theta)$ una statistica che massimizza la funzione di verosimiglianza indotta da $\kappa(\theta)$ L^*_κ.

Se $\hat{\theta}(x_1, \ldots, x_n)$ è uno stimatore ML di θ, allora

(*a*) $\sup_{\{\theta \in \Theta\,:\,\kappa(\theta)=\kappa\}} L_\theta(x_1, \ldots, x_n) \leq \sup_{\theta \in \Theta} L_\theta(x_1, \ldots, x_n)$, dal momento che $\{\theta \in \Theta\ :\ \kappa(\theta) = \kappa\}$ è un sottoinsieme di Θ,

(*b*) $L_{\hat{\theta}}(x_1, \ldots, x_n) = \sup_{\{\theta \in \Theta\,:\,\kappa(\theta)=\kappa(\hat{\theta}(x_1,\ldots,x_n))\}} L_\theta(x_1, \ldots, x_n)$, dal momento che $\hat{\theta}$ massimizza L_θ e $\hat{\theta} \in \{\theta \in \Theta\ :\ \kappa(\theta) = \kappa(\hat{\theta}(x_1, \ldots, x_n))\}$

Pertanto, abbiamo

$$L^*_\kappa(x_1, \ldots, x_n) = \sup_{\{\theta \in \Theta\,:\,\kappa(\theta)=\kappa\}} L_\theta(x_1, \ldots, x_n) \leq \sup_{\theta \in \Theta} L_\theta(x_1, \ldots, x_n) = L_{\hat{\theta}}(x_1, \ldots, x_n) =$$
$$= \sup_{\{\theta \in \Theta\,:\,\kappa(\theta)=\kappa(\hat{\theta}(x_1,\ldots,x_n))\}} L_\theta(x_1, \ldots, x_n) = L^*_{\kappa(\hat{\theta})}(x_1, \ldots, x_n)$$

La disuguaglianza appena dimostrata ci dice che se $\hat{\theta}(x_1, \ldots, x_n)$ è uno stimatore ML di θ, allora $\kappa(\hat{\theta})$ è uno stimatore ML di $\kappa(\theta)$. Questa proprietà va sotto il nome di *proprietà di invarianza degli stimatori di massima verosimiglianza*.

Nota 7.5 Operativamente la proprietà di invarianza di uno stimatore ML è molto utile: Se $\kappa(\theta)$ è la caratteristica della popolazione da stimare e $\widehat{\theta}$ è uno stimatore ML di θ, allora $T = \kappa(\widehat{\theta})$ è uno stimatore di massima verosimiglianza di $\kappa(\theta)$.

7.3 Esempi

Esempio 7.6 (Modello gaussiano con media e varianza entrambe incognite) Sia $X_1, \ldots, X_n$ *i.i.d.* $\sim N(\mu, \sigma^2)$, $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Stimiamo μ e σ^2 con il metodo di massima verosimiglianza. Sappiamo dall'equazione (7) che la funzione di verosimiglianza è data da

$$L_{\mu,\sigma^2}(x_1, \ldots, x_n) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big\{-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\Big\}$$

dove $s^2 = \sum_{j=1}^n (x_j - \bar{x})^2/(n-1)$. $L_{\mu,\sigma^2}(x_1 \ldots, x_n)$ è strettamente positiva $\forall\, x_1 \ldots, x_n, \mu, \sigma^2$ cosicché $\log L_{\mu,\sigma^2}(x_1 \ldots, x_n)$ è sempre ben definita e data da

$$\log L_{\mu,\sigma^2}(x_1 \ldots, x_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}$$

Inoltre L_{μ,σ^2} è massima se e solo se $\log L_{\mu,\sigma^2}$ è massima. Ma studiare l'ultima funzione è più semplice. I punti stazionari di $\log L_{\mu,\sigma^2}$ (cioè quelli che annullano ogni derivata parziale di $\log L_{\mu,\sigma^2}$) sono le soluzioni del sistema

$$\begin{cases} \frac{\partial}{\partial\mu}\log L_{\mu,\sigma^2}(x_1 \ldots, x_n) & = \frac{n(\bar{x}-\mu)}{\sigma^2} = 0 \\ \frac{\partial}{\partial\sigma^2}\log L_{\mu,\sigma^2}(x_1 \ldots, x_n) & = -\frac{n}{2\sigma^2} + \frac{(n-1)s^2+n(\bar{x}-\mu)^2}{2\sigma^4} = 0 \end{cases} \iff \begin{cases} \mu = \bar{x} \\ \sigma^2 = \frac{n-1}{n}s^2 \end{cases}$$

Poiché $L_{\mu,\sigma^2}(x_1, \ldots, x_n) \to 0$, per $\mu \to -\infty$ e $\mu \to +\infty$, allora $\bar{x}$ massimizza $L_{\mu,\sigma^2}(x_1, \ldots, x_n)$, $\forall\sigma^2$. Inoltre, $L_{\bar{x},\sigma^2}(x_1, \ldots, x_n) \to 0$ per $\sigma^2 \to 0$ e $\sigma^2 \to +\infty$: concludiamo che $\bar{X}$ è lo stimatore ML di μ e $\frac{(n-1)S^2}{n}$ di σ^2; $\bar{X}$ è stimatore non distorto di μ mentre $\frac{(n-1)S^2}{n}$ è distorto per σ^2. Notiamo che stimatori dei momenti e di massima verosimiglianza delle caratteristiche media e varianza del modello statistico gaussiano (con osservazioni i.i.d.) coincidono.

Esercizio 7.7 (Modello gaussiano con media o varianza nota) (*a*) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione gaussiana di media μ incognita e varianza σ^2 nota. Verificate che $\bar{X}$ è stimatore ML di μ.

(*b*) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione gaussiana di media μ nota e varianza σ^2 incognita. Verificate che lo stimatore ML di σ^2 è $\dfrac{\sum_{j=1}^n (X_j - \mu)^2}{n}$.

Nelle applicazioni tipicamente non abbiamo una forma analitica degli stimatori ML e il metodo di ricerca di stimatori ML è numerico.

Esempio 7.8 Sia $X_1, \ldots, X_n$ un campione casuale dalla densità di Cauchy centrata in $\theta \in \mathbb{R}$, cioè

$$f(x, \theta) = \frac{1}{\pi[1 + (x-\theta)^2]} \qquad \theta \in \mathbb{R}$$

Allora

$$L_\theta(x_1,\dots,x_n) = \frac{1}{\pi^n}\prod_{j=1}^{n}\frac{1}{1+(x_j-\theta)^2}$$

$$\log L_\theta(x_1,\dots,x_n) = -n\log\pi - \sum_{j=1}^{n}\log[1+(x_j-\theta)^2]$$

$$\frac{\partial \log L_\theta}{\partial\theta} \propto \sum_{j=1}^{n}\frac{x_j-\theta}{1+(x_j-\theta)^2}$$

Segue che possiamo determinare gli stimatori ML solo numericamente.

Il metodo numerico più comunemente usato è il metodo iterativo di Newton-Raphson: a ogni passo calcoliamo

$$\widehat{\theta}^{(t+1)} = \widehat{\theta}^{(t)} - \left.\frac{\partial \log L_\theta}{\partial\theta}\right|_{\theta=\widehat{\theta}^{(t)}}\left[\left.\frac{\partial^2 \log L_\theta}{\partial\theta\partial\theta^T}\right|_{\theta=\widehat{\theta}^{(t)}}\right]^{-1}, \quad t=0,1,\dots$$

dove $\widehat{\theta}^{(0)}$ è un assegnato valore iniziale e $\frac{\partial^2 \log L_\theta}{\partial\theta\partial\theta^T}$ è la matrice hessiana (potremmo infatti avere $m \geq 1$ parametri da stimare, cioè $\theta^T = (\theta_1,\dots,\theta_m)$). La matrice hessiana si assume sia a rango pieno.

Se a ogni iterazione $\frac{\partial^2 \log L_\theta}{\partial\theta\partial\theta^T}$ è sostituita con la matrice dei valori attesi di ogni singola cella, il metodo è noto come "Fisher-scoring method".

Esempio 7.9 Non sapete se una data moneta sia truccata o equa. Ma sapete che c'è aleatorietà, cioè che la probabilità di ottenere testa non è né 0 né 1. Lanciate n volte la moneta truccata e registrate i risultati $x_1,\dots,x_n$. Determinate sulla base di $x_1,\dots,x_n$ lo stimatore di massima verosimiglianza della probabilità di ottenere in successivi 3 lanci (della stessa moneta) la sequenza croce, testa, croce.

Soluzione Chiamiamo θ la probabilità di ottenere testa; allora θ appartiene all'intervallo $(0,1)$ e $x_1,\dots,x_n$ è la realizzazione di un campione $X_1,\dots,X_n$ *i.i.d.* $\sim$ **Be**(θ), con $\theta \in (0,1)$. La caratteristica da stimare è

$$\begin{aligned}\kappa(\theta) &= P_\theta(X_{n+1}=0, X_{n+2}=1, X_{n+3}=0) = \\ &= P_\theta(X_{n+1}=0)P_\theta(X_{n+2}=1)P_\theta(X_{n+3}=0) = \theta(1-\theta)^2\end{aligned}$$

La funzione di verosimiglianza del modello è

$$L_\theta(x_1,\dots,x_n) = \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j} \quad \theta\in(0,1)$$

Per ogni $\theta \in (0,1)$ abbiamo $L_\theta \in (0,1)$ e quindi $\log L_\theta$ è ben definita. Massimizziamo $\log L_\theta = \sum_{j=1}^n x_j \log\theta + (n-\sum_{j=1}^n x_j)\log(1-\theta)$:

$$\frac{\partial}{\partial\theta}\log L_\theta = \frac{\sum_{j=1}^n x_j}{\theta} - \frac{n-\sum_{j=1}^n x_j}{1-\theta} \geq 0 \iff \theta \leq \bar{x} \text{ e } \frac{\partial}{\partial\theta}\log L_\theta = 0 \iff \theta = \bar{x}.$$

Segue che

(a) per tutti i campioni tali che almeno una volta è apparsa testa ed almeno una volta è apparsa croce $\bar{X}$ è lo stimatore di massima verosimiglianza;

(*b*) se osserviamo su n lanci sempre testa allora $L_\theta(1,\dots,1) = \theta^n$ che ha estremo superiore $\theta = 1$. Ma 1 non appartiene a $\Theta = (0,1)$ e quindi lo stimatore di massima verosimiglianza non esiste.

(*c*) Analogamente, la stima di massima verosimiglianza di θ non esiste se osserviamo $(0,\dots,0)$, cioè sempre croce, perché l'estremo superiore di $L_\theta(0,\dots,0) = (1-\theta)^n$ è raggiunto in $\theta = 0$ che non appartiene a $(0,1)$.

Infine, per la proprietà di invarianza, lo stimatore di massima verosimiglianza di $\theta(1-\theta)^2$ è $\bar{X}(1-\bar{X})^2$. ∎

Esempio 7.10 Siano $X_1,\dots,X_n$ *i.i.d.* $\sim \mathcal{U}[\theta - 1/2, \theta + 1/2]$, con $\theta \in \mathbb{R}$, cioè $f(x,\theta) = \mathbf{1}_{[\theta-1/2,\theta+1/2]}(x)$. La funzione di verosimiglianza è

$$\begin{aligned} L_\theta(x_1,\dots,x_n) &= \prod_{j=1}^{n} \mathbf{1}_{[\theta-1/2,\theta+1/2]}(x_j) = \mathbf{1}(\theta - 1/2 \le x_{(1)} < x_{(n)} \le \theta + 1/2) \\ &= \mathbf{1}_{[x_{(n)}-1/2,x_{(1)}+1/2]}(\theta) \end{aligned}$$

dove $x_{(1)} = \min\{x_1,\dots,x_n\}$ e $x_{(n)} = \max\{x_1,\dots,x_n\}$. La funzione $\theta \mapsto L_\theta$ è costante e pari a 1 sull'intervallo[2] $[x_{(n)} - 1/2, x_{(1)} + 1/2]$ e nulla altrove: quindi ogni punto dell'intervallo $[x_{(n)} - 1/2, x_{(1)} + 1/2]$ massimizza L_θ. Deduciamo da questo esempio che lo stimatore ML non è necessariamente unico.

Nel prossimo esempio vedremo che stimatori ottenuti con il metodo dei momenti e quelli di massima verosimiglianza possono essere diversi.

Esempio 7.11 Siano $X_1,\dots,X_n$ *i.i.d.* $\sim \mathcal{U}[0,\theta]$, con $\theta > 0$. La funzione di verosimiglianza del modello è

$$L_\theta(x_1\dots,x_n) = \frac{1}{\theta^n}\mathbf{1}_{[x_{(n)},+\infty)}(\theta)\,, \quad x_1,\dots,x_n > 0$$

(cfr. Equazione (9)). Poiché $\theta \mapsto L_\theta$ è strettamente decrescente sull'intervallo $[x_{(n)},\infty)$ allora $x_{(n)}$ è l'unico punto di massimo. Segue che $X_{(n)}$ è lo stimatore ML di θ.

Con il metodo dei momenti avevamo trovato lo stimatore $2\bar{X}$ di θ. Per scegliere fra $X_{(n)}$ e $2\bar{X}$, ricorriamo al calcolo dell'errore quadratico medio (MSE). Nell'Esempio 6.9 abbiamo calcolato media e varianza di $X_{(n)}$, ottenendo

$$\mathrm{E}_\theta(X_{(n)}) = \frac{n\theta}{n+1}, \qquad \mathrm{Var}_\theta(X_{(n)}) = \frac{n\theta^2}{(n+1)^2(n+2)}$$

e quindi

$$\begin{aligned} \mathrm{MSE}(X_{(n)}) &= \mathrm{Var}_\theta(X_{(n)}) + (\mathrm{E}_\theta(X_{(n)}) - \theta)^2 = \frac{n\theta^2}{(n+1)^2(n+2)} + \left(\frac{n}{n+1} - 1\right)^2 \theta^2 \\ &= \frac{2\theta^2}{(n+1)(n+2)} \end{aligned}$$

[2] $[x_{(n)} - 1/2, x_{(1)} + 1/2]$ è essenzialmente un intervallo in quanto $P_\theta(X_{(n)} - 1/2 < X_{(1)} + 1/2) = P_\theta(X_{(n)} - X_{(1)} < 1) = 1\ \forall \in \mathbb{R}$, perché per qualunque coppia di osservazioni (X_i, X_j) $(i \neq j)$ la distanza massima fra le due è 1.

Invece, $2\bar{X}$ è non distorto e ha varianza

$$\mathrm{Var}_\theta(2\bar{X}) = \frac{4\,\mathrm{Var}_\theta(X_1)}{n} = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}$$

da cui deduciamo

$$\mathrm{MSE}(2\bar{X}) = \mathrm{Var}_\theta(2\bar{X}) = \frac{\theta^2}{3n}$$

Se $n = 1$ allora $2X_1$ e $X_{(1)} = X_1$ hanno lo stesso MSE e scegliamo $2X_1$ che è non distorto. Mentre, per $n \geq 2$ preferiamo $X_{(n)}$ che ha MSE più piccolo uniformemente in $\theta > 0$.

Dovendo confrontare i due metodi di stima dei momenti e di massima verosimiglianza, ci aspettiamo che lo stimatore di massima verosimiglianza sia "migliore" di quello ottenuto con il metodo dei momenti in quanto il secondo non usa né tutte le informazioni provenienti dal campione né tutte quelle provenienti dalla f.d.r. sottostante alla popolazione. Stimando con i momenti, usiamo solo la sintesi del campione realizzata dai momenti campionari e la sintesi della f.d.r. teorica realizzata dai momenti teorici. Uno stimatore ML usa invece tutte le informazioni empiriche provenienti dai dati e teoriche riassunte nella verosimiglianza.

7.4 Proprietà degli stimatori di massima verosimiglianza

Lo stimatore di ML gode di alcune proprietà che lo rendono un "buon" stimatore. La prima proprietà che ricordiamo lega uno stimatore ML agli stimatori efficienti:

Proposizione 7.12 *Se le condizioni di regolarità necessarie perché la disuguaglianza di Fréchet-Cramer-Rao sussista sono soddisfatte e $\kappa(\theta)$ ammette uno stimatore T "efficiente" (cioè non distorto e la cui varianza raggiunge il confine inferiore di Fréchet-Cramer-Rao), allora esiste uno stimatore ML essenzialmente unico e coincide con T.*

Dimostrazione Sia T lo stimatore efficiente di $\kappa(\theta)$. Esso è essenzialmente unico e per il Teorema 6.1 abbiamo

$$P_\theta\Big(\frac{\partial}{\partial\theta}\log L_\theta(x_1\ldots,x_n) = a(\theta,n)(T-\theta)\Big) = 1 \quad \forall\theta\in\Theta \tag{19}$$

La funzione $a(\theta, n)$ è non nulla per ogni $\theta \in \Theta$ perché, in virtù di (19), abbiamo

$$0 < nI(\theta) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L_\theta(x_1\ldots,x_n)\right) = \mathrm{Var}_\theta(a(\theta,n)(T-\theta)) = a^2(\theta,n)\,\mathrm{Var}_\theta(T)$$

Inoltre, se valgono le condizioni di regolarità $(i)-(iii)$ del Teorema 6.1, per determinare uno stimatore $\widehat{\theta}$ ML dobbiamo risolvere l'equazione

$$\frac{\partial}{\partial\theta}\log L_\theta(x_1\ldots,x_n) = 0 \tag{20}$$

Tenuto conto che $a(\theta, n) \neq 0\ \forall\theta$ e sostituendo $a(\theta, n)(T - \theta)$ a $\frac{\partial}{\partial\theta}\log L_\theta(x_1\ldots,x_n)$ in (20), otteniamo che necessariamente $P_\theta(\widehat{\theta} = T) = 1\ \forall\theta \in \Theta$. ∎

Infine, sintetizziamo le proprietà asintotiche degli stimatori ML nella seguente proposizione.

Proposizione 7.13 *Sia $X_1, \ldots, X_n, \ldots$ una successione di variabili aleatorie i.i.d. con comune funzione di densità $f(x,\theta)$, $\theta \in \Theta$ e sia $\{T_n\}_n$ la successione degli stimatori ML di $\kappa(\theta)$. Se $f(x,\theta)$ soddisfa le condizioni di regolarità della disuguaglianza di Fréchet-Cramer-Rao e altre ancora (di esistenza, continuità e limitatezza delle derivate seconda e terza di $f(x,\theta)$ rispetto a θ), allora la successione $\{T_n\}_n$ è*

1. *asintoticamente non distorta per $\kappa(\theta)$,*
2. *consistente in media quadratica per $\kappa(\theta)$,*
3. *asintoticamente gaussiana con media asintotica $\kappa(\theta)$ e varianza asintotica $\dfrac{[\kappa'(\theta)]^2}{nI(\theta)}$, cioè*

$$\lim_{n\to\infty} P\left(\frac{T_n - \kappa(\theta)}{\sqrt{\frac{(\kappa'(\theta))^2}{nI(\theta)}}} \le z\right) = \Phi(z) \qquad \forall z \in \mathbb{R}$$

L'allievo interessato è rimandato alla Sezione §25 pagine 228-233 in Borovkov (1987) sia per vedere in dettaglio le ulteriori condizioni che il modello statistico deve soddisfare per la validità del risultato sia per una dimostrazione di esso.

La Proposizione 7.13 tra le altre cose ci dice che *uno stimatore ML è asintoticamente efficiente e quindi asintoticamente UMVUE.*

Le ipotesi della Proposizione 7.13 sono soddisfatte da quasi tutti i modelli statistici che voi vedrete. Praticamente gli unici modelli in cui ci imbatteremo e che non le soddisfano sono quelli con funzione di densità $f(x,\theta)$ il sui supporto $\{x : f(x,\theta) > 0\}$ dipende da θ.

Concludiamo con un paio di esercizi: il primo riguarda un modello statistico regolare e quindi le proprietà asintotiche degli stimatori ML valgono; di contro, il secondo esemplifica un caso di modello statistico NON regolare.

Esempio 7.14 Data la famiglia di densità:

$$f(x,\theta) = \frac{1}{\theta} x^{-\left(1+\frac{1}{\theta}\right)} I_{[1,+\infty]}(x), \qquad \theta > 0$$

1. determinare lo stimatore di massima verosimiglianza $\widehat{\theta}_n$ di θ basato su un campione casuale di dimensione n estratto da $f(x,\theta)$;
2. dedurre lo stimatore di massima verosimiglianza, chiamiamolo $\widehat{\tau}_n$, per $\tau(\theta) = 1/\theta$;
3. mostrare che $\widehat{\tau}_n$ è asintoticamente non distorto e consistente per $\tau(\theta)$;
4. determinare la f.d.r. asintotica di $\widehat{\tau}_n$.

Soluzione

$$\ell_n(\theta) := \log L_\theta(x_1, \ldots, x_n) = \log\left[\prod_{i=1}^n f(x_i,\theta)\right] = -n\log\theta - \left(1+\frac{1}{\theta}\right)\sum_{i=1}^n \log x_i$$

e $\dfrac{\partial \ell_n}{\partial \theta} = -\dfrac{n}{\theta} + \dfrac{1}{\theta^2}\sum_{i=1}^n \log x_i = 0$ ha soluzione $\theta = \dfrac{1}{n}\sum_{i=1}^n \log x_i$. Inoltre

$$\frac{\partial^2 \ell_n}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^n \log x_i < 0 \quad \Longleftrightarrow \quad n - \frac{2}{\theta}\sum_{i=1}^n \log x_i < 0$$

che è verificata in $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$. Segue che

1. $\widehat{\theta}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \log X_i$ è lo stimatore ML di θ;

2. $\widehat{\tau}_n = 1/\widehat{\theta}_n$ è lo stimatore ML di $\tau(\theta)$.

3. È facile mostrare che se X ha densità $f(x,\theta)$ allora $Y = \log X$ ha densità esponenziale di parametro θ. Infatti, per $y > 0$ abbiamo

$$F_Y(y,\theta) = P_\theta(Y \le y) = P_\theta(\log X \le y) = P_\theta\left(X \le \mathrm{e}^y\right) = \int_1^{\mathrm{e}^y} \frac{1}{\theta x^{1+1/\theta}}\, dx = -\frac{1}{x^{1/\theta}}\bigg|_1^{\mathrm{e}^y} = 1 - \mathrm{e}^{-\frac{y}{\theta}}$$

Quindi $\sum_{i=1}^{n} \log X_i$ ha densità $\Gamma(n,\theta)$ e $\widehat{\theta}_n \sim \Gamma(n, \frac{\theta}{n})$ da cui

$$\begin{aligned}\mathrm{E}(\hat{\tau_n}) &= \int_0^\infty \frac{1}{x} \cdot \frac{\left(\frac{n}{\theta}\right)^n}{\Gamma(n)} x^{n-1} e^{-x\frac{n}{\theta}} dx = \int_0^\infty \frac{n}{\theta} \frac{(\frac{n}{\theta})^{n-1}}{(n-1)\Gamma(n-1)} x^{n-1-1} e^{-x\frac{n}{\theta}} dx \\ &= \frac{n}{n-1} \cdot \frac{1}{\theta} \to \frac{1}{\theta} \qquad \text{per } \; n \to +\infty\end{aligned}$$

Segue che $\{\widehat{\tau}_n\}_n$ è asintoticamente non distorto per τ. Poiché

$$\begin{aligned}\mathrm{E}(\hat{\tau_n}^2) &= \int_0^\infty (\frac{n}{\theta})^2 \frac{(\frac{n}{\theta})^{n-2}}{(n-1)(n-2)\Gamma(n-2)} x^{n-2-1} e^{-x\frac{n}{\theta}} dx = \frac{n^2}{(n-1)(n-2)} \cdot \frac{1}{\theta^2} \\ \mathrm{Var}(\widehat{\tau}_n) &= \frac{n^2}{(n-1)(n-2)} \cdot \frac{1}{\theta^2} - \left(\frac{n}{n-1} \cdot \frac{1}{\theta}\right)^2 = \frac{n^2}{(n-1)^2(n-2)\theta^2} \to 0\end{aligned}$$

allora $\{\widehat{\tau}_n\}_n$ è consistente in media quadratica per τ.

4. Poiché $\log X_1 \sim \Gamma(1,\theta)$, allora

$$\begin{aligned}nI(\theta) = \mathrm{Var}_\theta\left(\frac{\partial l_n(X_1,\ldots,X_n)}{\partial\theta}\right) &= \mathrm{Var}_\theta\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}\log X_i\right) = \\ &= \mathrm{Var}_\theta\left(\frac{1}{\theta^2}\sum_{i=1}^{n}\log X_i\right) = \frac{n}{\theta^4}\,\mathrm{Var}_\theta(\log X_1) = \frac{n\theta^2}{\theta^4} = \frac{n}{\theta^2}\end{aligned}$$

Inoltre, $[\tau'(\theta)]^2 = (-1/\theta^2)^2 = \theta^{-4}$ e quindi

$$\frac{(\tau'(\theta))^2}{nI(\theta)} = \frac{1}{n\theta^2}$$

Essendo $\widehat{\tau}_n$ lo stimatore ML di un modello statistico regolare, abbiamo

$$\lim_{n\to\infty} P\left(\sqrt{n}\theta(\widehat{\tau}_n - 1/\theta) \le z\right) = \Phi(z) \qquad \forall z \in \mathbb{R}. \quad \blacksquare$$

Esempio 7.15 (Modello Uniforme. Continuazione) Siano $X_1,\ldots,X_n$ *i.i.d.* $\sim \mathcal{U}[0,\theta]$, con $\theta > 0$. Abbiamo trovato nell'Esempio 7.11 che lo stimatore ML di θ è il massimo delle osservazioni $X_{(n)}$. Discutiamo ora le proprietà asintotiche di $X_{(n)}$.

(a) $X_{(n)}$ è asintoticamente non distorto, infatti

$$\lim_{n\to\infty} \mathrm{E}_\theta(X_{(n)}) = \lim_{n\to\infty} \frac{n\theta}{n+1} = \theta$$

(b) $X_{(n)}$ è consistente in media quadratica, infatti

$$\lim_{n\to\infty} \mathrm{Var}_\theta(X_{(n)}) = \lim_{n\to\infty} \frac{n\theta^2}{(n+1)^2(n+2)} = 0$$

(c) Ma, $X_{(n)}$ *non è asintoticamente gaussiano.* Infatti se *standardizziamo* $X_{(n)}$ cioè costruiamo la variabile aleatoria

$$Z_n := \frac{X_{(n)} - E_\theta(X_{(n)})}{\sqrt{\mathrm{Var}_\theta(X_{(n)})}} = \frac{X_{(n)} - \frac{n\theta}{n+1}}{\sqrt{\frac{n\theta^2}{(n+1)^2(n+2)}}} = \sqrt{\frac{n+2}{n}} \left(\frac{n+1}{\theta} X_{(n)} - n \right)$$

allora si può dimostrare che la f.d.r. limite di Z_n è

$$F(z) = \begin{cases} \mathrm{e}^z & z \le 0 \\ 1 & z > 0 \end{cases}$$

cioè asintoticamente $-Z_n$ è esponenziale di parametro 1.

Seguono le *technicalities* per appassionati:

$$\begin{aligned} P_\theta(Z_n \le z) &= P_\theta\left(X_{(n)} \le \frac{\theta}{n+1}\Big(z\sqrt{\frac{n}{n+2}} + n\Big)\right) \\ &= \begin{cases} 0 & \text{se } \frac{\theta}{n+1}\left(z\sqrt{\frac{n}{n+2}} + n\right) \le 0 \\ \left(\frac{z\sqrt{\frac{n}{n+2}}+n}{n+1}\right)^n & \text{se } 0 < \frac{\theta}{n+1}\left(z\sqrt{\frac{n}{n+2}} + n\right) < \theta \\ 1 & \text{se } \frac{\theta}{n+1}\left(z\sqrt{\frac{n}{n+2}} + n\right) \ge \theta \end{cases} \quad \text{[si confronti Esempio 6.9]} \\ &\simeq \begin{cases} 0 & \text{se } \frac{\theta z}{n} + \theta \le 0 \\ \left(1 + \frac{z}{n}\right)^n & \text{se } 0 < \frac{\theta z}{n} + \theta < \theta \\ 1 & \text{se } \frac{\theta z}{n} + \theta \ge \theta \end{cases} \to \begin{cases} \mathrm{e}^z & \text{se } z \le 0 \\ 1 & \text{se } z > 0 \end{cases} \qquad \text{per } n \to \infty \end{aligned}$$

Riferimenti bibliografici

[1] Cifarelli, D.M. Conti, L., Regazzini, E. (1996) On the asymptotic distribution of a general measure of monotone dependence. *Ann. Statist.* **24**, 1386–1399

[2] Conover, W.J. (1999) *Practical Nonparametric Statistics 3ª Ed*, Wiley, New York

[3] Del Barrio, E. Cuesta-Albertos, J. A.; Matrán, C. (2000) Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. (With comments) *Test* **9**, 1–96

[4] Fisher, R. (1922) On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, A*, **222**, 309–368

[5] Fisher, R.A. (1924) The conditions under which χ^2 measures the discepancy between observation and hypohesis. *J. Roy. Statist. Soc.*, **87**, 442–450

[6] Mann, H.B. and Wald, A. (1942) On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math. Stat.*,**13**, 306–317

[7] Karl Pearson (1894) Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society A*, **185**, 71–110

[8] Pestman, Wiebe R. (1998) *Mathematical Statistics An Introduction* De Gruyter

[9] *R: A language and environment for statistical computing* R Development Core Team (2003) `http://www.R-project.org` , R Foundation for Statistical Computing Vienna, Austria

[10] Rohatgi, V.K e Saleh, A.K. Md. E. (1999) *An Introduction to Probability and Statistics* Wiley, New York

[11] Silvey, S.D (1975) *Statistical Inference* Chapman & Hall London