dove Id_p indica la matrice di identità di ordine p.

Proof. Consideriamo per semplicità il caso $p = 1$. Per il teorema del valor medio di Lagrange, esiste $\overline{\theta}_n$ intermedio tra $\widehat{\theta}_n$ e θ_0 tale per cui vale l'uguaglianza

$$0 = \log' L(\widehat{\theta}_n) = \log L'(\theta_0) + \log L''(\overline{\theta}_n)(\widehat{\theta}_n - \theta_0) \ ,$$

da cui

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{\log L'(\theta_0)/\sqrt{n}}{\log L''(\overline{\theta}_n)/n} \ .$$

Per il numeratore abbiamo una somma di variabili aleatorie IID con valor medio nullo e varianza finita; siamo quindi nel dominio di applicabilità del teorema del limite centrale ed otteniamo

$$\log L'(\theta_0)/\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f(X_i;\theta)}{\partial \theta} \right|_{\theta=\theta_0} \rightarrow_d N(0, I_1(\theta_0)) \ .$$

Per il denominatore abbiamo una somma di variabili IID con valor medio finito, quindi per la legge dei grandi numeri su variabili uniformemente integrabili ed il teorema di Slutzky otteniamo

$$\log L''(\overline{\theta}_n)/n = \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} \right|_{\theta=\overline{\theta}_n} \rightarrow_p I_1(\theta_0) \ .$$

Combinando i due risultati ed usando di nuovo Slutzky si arriva all'enunciato del Teorema. ■

Remark 111 *Questa dimostrazione rende evidente come la consistenza degli stimatori sia un prerequisito necessario perché abbia senso la domanda sulla loro asintotica Gaussianità.*

12 Il limite inferiore di Cramér-Rao

Un risultato notevole riguarda la determinazione della varianza minima degli stimatori non-distorti, sotto condizioni di regolarità; in particolare, emerge che tale varianza minima coincide con quella degli stimatori di massima verosimiglianza, sotto condizioni di regolarità. Per tale motivo, gli stimatori di massima verosimiglianza sotto opportune ipotesi (che coprono gran parte delle distribuzioni di uso comune, almeno nei casi più semplici) risultano essere non solo consistenti ed asinotitcamente Gaussiani, ma anche efficienti in senso assoluto.

Remark 112 *E' evidente che porsi la questione sulla varianza minima ha senso solo per stimatori che siano non-distorti; altrimenti qualsiasi stimatore con valore identicamente costante non potrebbe essere migliorato, avendo varianza identicamente pari a zero.*

Consideriamo ora uno stimatore generico di un parametro $\theta \in \mathbb{R}$,

$$W_n = W_n(X_1, ..., X_n)$$

e supponiamo che la funzione $\psi(\theta) := E_{f_\theta}[W(X_1, ..., X_n)]$ sia di classe C^1 per ogni n; assumiamo inoltre che la densità congiunta f_θ sia tale per cui, per ogni $h \in C^1$, $h : \mathbb{R}^n \to \mathbb{R}$ con $E_\theta[|h(X_1, ..., X_n)|] < \infty$ valga la scambiabilità

$$\frac{d}{d\theta} E_\theta[h(X_1, ..., X_n)] = \int_{\mathbb{R}^n} h(x_1, ..., x_n) \frac{\partial}{\partial \theta} f(x_1, ..., x_n; \theta) dx_1 ... dx_n \ .$$

Chiamiamo ora $b(\theta) = E[W_n] - \theta$ il bias della nostra statistica, e $b'(.)$ la sua derivata rispetto a θ. Abbiamo allora il seguente

Theorem 113 *(Cramér-Rao) Per $\theta \in \mathbb{R}$*

$$Var[W_n] \geq \frac{\{b'(\theta_0)\}^2}{I_n(\theta_0)} \ .$$

Remark 114 *Per semplicità, abbiamo enunciato e dimostreremo il teorema solo nel caso $p = 1$. La generalizzazione a p generico è comunque abbastanza semplice; consideriamo lo stimatore W_n come un vettore colonna $p \times 1$, con valor medio $\Psi(\theta)$ che ha matrice Jacobiana $p \times p$ denotata con $J\Psi()$. Allora la matrice di varianza e covarianza di W_n soddisfa la disuguaglianza*

$$Var[W_n] \geq J\Psi(\theta_0) I_n^{-1}(\theta_0) J\Psi(\theta_0)^T .$$

La disuguaglianza va intesa none solito senso di disuguaglianza tra matrici simmetriche e non-negative definite: la loro differenza deve essere non-negativa definita.

Proof. (caso $p = 1$). L'idea di fondo è utilizzare la disuguaglianza di Cauchy-Schwartz, scrivendola come

$$Var[Y] \geq \frac{Cov^2(X, Y)}{Var[X]} \ .$$

Prendiamo come X la funzione punteggio $\frac{d}{d\theta} \log L(\theta; X_1, ..., X_n)\big|_{\theta=\theta_0}$; sappiamo che $E[X] = 0$, da cui $Cov(X, Y) = E[XY]$. Prendiamo quindi $Y = W_n$, e osserviamo che

$$\begin{aligned}
E[XY] &= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \left. \frac{d}{d\theta} \log L(\theta; x_1, ..., x_n) \right|_{\theta=\theta_0} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \frac{\frac{d}{d\theta} f(\theta; x_1, ..., x_n)\big|_{\theta=\theta_0}}{f(x_1, ..., x_n; \theta_0)} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \left. \frac{d}{d\theta} f(\theta; x_1, ..., x_n) \right|_{\theta=\theta_0} dx_1 ... dx_n \\
&= \left. \frac{d}{d\theta} \int_{\mathbb{R}^n} W(x_1, ..., x_n) f(\theta; x_1, ..., x_n) dx_1 ... dx_n \right|_{\theta=\theta_0} \\
&= \left. \frac{d}{d\theta} E_\theta[W_n] \right|_{\theta=\theta_0} .
\end{aligned}$$

Questo conclude la dimostrazione nel caso con densità; il caso discreto è identico. ■

Remark 115 *E' interessante studiare cosa succede nel caso in cui le condizioni di regolarità, ed in particolare la possibilità di scambiare la derivata con l'integrale nel valor medio, non siano soddisfatte. Prendiamo ad esempio un campione di variabili i.i.d., uniformi in* $[0, \theta]$; *si verifica facilmente che la funzione di verosimiglianza prende la forma*

$$L(\theta; X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(X_i) = \frac{1}{\theta^n} I_{[0,\theta]}(X_{(n)}) \ ,$$

dove $X_{(n)}$ *indica la più grande delle n osservazioni; abbiamo inoltre*

$$\widehat{\theta}_{ML;n} = X_{(n)} \ .$$

Ora, si vede anche facilmente che, per ogni $\varepsilon > 0$

$$\begin{aligned} \Pr\left\{\theta_0 - X_{(n)} > \varepsilon\right\} &= \prod_{i=1}^{n} \Pr\left\{\theta_0 - X_i > \varepsilon\right\} \\ &= \left\{\Pr\left\{\theta_0 - \varepsilon > X_1\right\}\right\}^n \\ &= \left\{\frac{1}{\theta_0}(\theta_0 - \varepsilon)\right\}^n = (1 - \frac{\varepsilon}{\theta_0})^n. \end{aligned}$$

Poichè queste probabilità sono sommabili, abbiamo che lo stimatore è completamente convergente (e pertanto converge quasi certamente). Inoltre abbiamo che

$$\begin{aligned} \Pr\left\{n(\theta_0 - X_{(n)}) > \varepsilon\right\} &= \left\{\Pr\left\{\theta_0 - \frac{\varepsilon}{n} > X_1\right\}\right\}^n \\ &= (1 - \frac{\varepsilon}{n\theta_0})^n \to \exp(-\frac{\varepsilon}{\theta_0}) \end{aligned}$$

per $n \to \infty$: *abbiamo quindi una forma di superconsistenza, perché la convergenza avviene a velocità* n^{-1} *invece che* $n^{-1/2}$ *come nel teorema precedente. D'altra parte la distribuzione limite è esponenziale invece che Gaussiana. Questo esempio mostra come, quando le condizioni di regolarità vengono a mancare, non necessariamente le proprietà degli stimatori di massima verosimiglianza debbano peggiorare: possono addirittura essere superiori, sia come modalità che come velocità di convergenza.*

13 Statistiche sufficienti

Una domanda naturale che possiamo porci, specialmente in un momento in cui masse enormi di dati sono a disposizione, è la seguente: posso comprimere un campione di dati osservati senza perdere informazione sul parametro che mi interessa? Questa domanda ci porta alla nozione di statistiche sufficienti.