

11 Il metodo della massima verosimiglianza

Supponiamo di avere di fronte a noi due urne, una con 90 palline bianche e 10 nere, e l'altra con 10 nere e 90 bianche; identifichiamo le urne con la proporzione di palline bianche che contengono (denotata con p), in modo tale che la prima urna corrisponda a $p = .9$ e la seconda a $p = .1$. Viene estratta una singola pallina e non sappiamo da quale urna venga; osserviamo però che la pallina è bianca. Dovendo cercare di risalire all'urna di provenienza, sembra naturale scegliere quella che rende più probabile osservare quello che abbiamo effettivamente osservato, cioè l'urna che contiene la maggiore proporzione di palline bianche; in altre parole, dall'osservazione "è stata estratta una pallina bianca" sembra naturale far descendere lo stimatore $\hat{p} = 0.9$. Supponiamo ora che invece di una sola pallina ne siano estratte 5, e che risultino essere 3 bianche e 2 nere; immaginiamo altresì che l'alternativa non sia solo tra due urne, ma tra un continuo di urne con tutte le possibili proporzioni di palline bianche tra 0 ed 1, identificate sempre con p . In questo caso, la probabilità di osservare 3 bianche e 2 nere è data da una legge binomiale:

$$Pr\{3 \text{ bianche e } 2 \text{ nere}\} = \binom{5}{3} p^3 (1-p)^2,$$

ed è facilmente verificabile che questa probabilità è massimizzata prendendo $\hat{p} = \frac{3}{5}$.

Spiegazione Semplice

L'idea fondamentale della Massima Verosimiglianza (MLE)

Questo paragrafo introduce l'idea più intuitiva di tutta la statistica. Il metodo della Massima Verosimiglianza (Maximum Likelihood Estimation) risponde alla domanda: "Dato che ho osservato questi dati, qual è il valore del parametro (θ) che rende i miei dati i più probabili (verosimili) possibile?"

- **Esempio 1 (Urne):** Osservo 1 pallina bianca.

- Se l'urna fosse $p = 0.1$, la probabilità di osservare il mio dato (bianca) era del 10%.
- Se l'urna fosse $p = 0.9$, la probabilità di osservare il mio dato (bianca) era del 90%.

Scelgo $\hat{p} = 0.9$ perché è il valore del parametro che *massimizza la probabilità* di ciò che ho visto.

- **Esempio 2 (Binomiale):** Osservo 3 Bianche e 2 Nere. Cerco il valore p (tra 0 e 1) che massimizza la funzione $f(p) = \binom{5}{3} p^3 (1-p)^2$. Usando l'analisi matematica (facendo la derivata rispetto a p e ponendola uguale a 0), si scopre che il massimo si ha proprio a $\hat{p} = 3/5$.

Lo stimatore di massima verosimiglianza (\hat{p}) è il valore del parametro che meglio "spiega" i dati che abbiamo osservato.

Questo esempio molto semplice dovrebbe aiutare a capire l'idea che è alla base degli stimatori di massima verosimiglianza - si tratta di costruire la funzione di probabilità (o di densità, nel caso continuo) relativa ad un certo campione aletorario, e vederla quindi non più come funzione del campione, ma come funzione (aleatoria) dei parametri, prendendo come date le osservazioni. In altre parole, si ha la seguente definizione:

Definition 92 (Funzione di verosimiglianza). *Sia X_1, \dots, X_n un campione aleatorio con legge (funzione di probabilità o densità, a seconda del caso discreto o continuo) f_θ , $\theta \in \mathbb{R}^p$. La funzione di verosimiglianza $L : \mathbb{R}^p \rightarrow \mathbb{R}$ è definita da*

$$L(\theta; X_1, \dots, X_n) := \prod_{i=1}^n f(X_i; \theta) = f(X_1, \dots, X_n; \theta).$$

Spiegazione Semplice

Funzione di Densità $f(x)$ vs. Funzione di Verosimiglianza $L(\theta)$

Questa è la differenza più importante da capire. Entrambe usano la stessa formula, $f(x; \theta)$, ma la

"leggono" in due modi opposti:

- **Funzione di Densità/Probabilità, $f(x; \theta)$:**
 - **Fisso:** θ (il parametro, es. $p = 0.5$ per una moneta).
 - **Variabile:** x (i dati, es. $k = 3$ teste).
 - **Domanda:** "Qual è la probabilità di osservare x , dato θ ?"
- **Funzione di Verosimiglianza, $L(\theta; x)$:**
 - **Fisso:** x (i dati che ho già osservato, es. $k = 3$ teste).
 - **Variabile:** θ (il parametro, es. p è incognito).
 - **Domanda:** "Quanto era 'verosimile' un certo θ , dato che ho osservato x ?"

Dato che le X_i sono indipendenti, la probabilità congiunta (di osservarle tutte insieme) è il prodotto (\prod) delle singole probabilità.

Remark 93. Con un leggero abuso di notazione, a sinistra della prima uguaglianza abbiamo scritto f sia per la legge univariata che per quella congiunta; inoltre usiamo $f(\cdot)$ sia per il caso discreto che per quello continuo. Il punto più importante è però un altro: la verosimiglianza è funzione del parametro prendendo i valori di X_1, \dots, X_n come dati, mentre la densità congiunta prende il parametro θ come dato ed è funzione dei possibili valori delle osservazioni x_1, \dots, x_n .

Remark 94. Ovviamente si ha (nel caso continuo)

$$\int_{\mathbb{R}^n} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n = 1;$$

non c'è però nessun motivo di aspettarsi l'uguaglianza

$$\int_{\mathbb{R}^p} L(\theta; X_1, \dots, X_n) d\theta \stackrel{??}{=} 1.$$

anzi in generale questo integrale può divergere. La funzione di verosimiglianza pertanto NON identifica una densità di probabilità.

Example 95 (Valor medio nel caso Gaussiano). Sia X_1, \dots, X_n un campione aleatorio estratto da una legge Gaussiana $N(\mu, 1)$; è immediato verificare che la funzione di verosimiglianza prende la forma

$$L(\mu; X_1, \dots, X_n) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2\right\}.$$

In questo caso (e in quelli a seguire) è conveniente lavorare con il logaritmo della funzione di verosimiglianza; trattandosi di una trasformazione monotona crescente, il valore che massimizza la funzione (cioè l'argmax) non cambia, e pertanto nemmeno lo stimatore. Si ha in particolare

$$\begin{aligned} l(\mu; X_1, \dots, X_n) &:= \log L(\mu; X_1, \dots, X_n) \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2 \end{aligned}$$

funzione che è evidentemente massimizzata in

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$$

Spiegazione Semplice

Perché usare il Logaritmo? (Log-Likelihood)

Massimizzare $L(\theta)$ è lo stesso che massimizzare $\log(L(\theta))$, perché il logaritmo è una funzione crescente (se $A > B$, allora $\log(A) > \log(B)$).

Si usa il logaritmo ($l(\theta) = \log L(\theta)$) perché trasforma i prodotti (che sono difficili da derivare) in somme (che sono facili da derivare).

$$L(\theta) = \prod f(X_i; \theta) \xrightarrow{\log} l(\theta) = \sum \log f(X_i; \theta)$$

Esempio Gaussiano: Vogliamo massimizzare $l(\mu) = C - \frac{1}{2} \sum (X_i - \mu)^2$. Massimizzare questa espressione è equivalente a minimizzare $\sum (X_i - \mu)^2$.

Questo è il famoso **Metodo dei Minimi Quadrati!** Troviamo il μ che minimizza la somma delle distanze al quadrato dai punti X_i . Usando l'analisi (derivata rispetto a μ posta uguale a 0), si scopre che questo valore è esattamente la media campionaria \bar{X} .

Conclusione: Per dati Gaussiani, lo stimatore di Massima Verosimiglianza per la media μ è la media campionaria \bar{X} .

Example 96 (Variabili Bernoulliane). Sia X_1, \dots, X_n un campione aleatorio di variabili $Ber(p)$ la probabilità di k successi per $k = 0, \dots, n$ evidentemente data da $p^k(1-p)^{n-k}$. e la verosimiglianza è pertanto

$$L(p; X_1, \dots, X_n) = p^{\sum_{i=1}^n X_i} (1-p)^{n-\sum_{i=1}^n X_i}.$$

La massimizzazione della log-verosimiglianza ci fa ottenere facilmente

$$\log L(p; X_1, \dots, X_n) = \left(\sum_{i=1}^n X_i \right) \log p + \left(n - \sum_{i=1}^n X_i \right) \log(1-p)$$

$$\frac{\partial}{\partial p} \log L(p; \dots) = \frac{\sum_{i=1}^n X_i}{p} - \frac{(n - \sum_{i=1}^n X_i)}{1-p}$$

imponendo che la derivata sia pari a zero, si vede facilmente che

$$\hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$$

Spiegazione Semplice

Esempio Bernoulli (Moneta Truccata)

- **Dati:** n lanci di una moneta (1 per Testa, 0 per Croce).
- **Parametro:** p , la probabilità (ignota) di Testa.
- **Log-Verosimiglianza:** La formula $l(p) = (\text{Num. Teste}) \log(p) + (\text{Num. Croci}) \log(1-p)$.
- **Massimizzazione:** Facciamo la derivata di $l(p)$ rispetto a p e la poniamo uguale a 0.
- **Risultato:** La soluzione è $\hat{p} = \frac{\sum X_i}{n} = \frac{\text{Num. Teste}}{\text{Num. Lanci}}$.

Ancora una volta, lo stimatore di Massima Verosimiglianza è la media campionaria, che è esattamente quello che l'intuito (e il Metodo dei Momenti) ci aveva già suggerito.

Remark 97. Si sarebbe potuto scrivere la verosimiglianza come

$$L(p; X_1, \dots, X_n) = \binom{n}{\sum_{i=1}^n X_i} p^{\sum_{i=1}^n X_i} (1-p)^{n-\sum_{i=1}^n X_i}.$$

Quale è la versione "giusta" della verosimiglianza? Come sarebbero cambiate le stime utilizzando questa seconda versione? Si tratta di un risultato generale o c'è un principio generale sottostante?

11.0.1 Consistency and Asymptotic Gaussianity

To study consistency, we need first to introduce the Kullback-Leibler distance between two probability densities/probability mass function. This is defined by

Definition 98. Siano f, g funzioni di densità o funzioni di probabilità. La distanza di Kullback-Leibler tra f è definita da

$$D(f, g) = \int \log\left(\frac{f(x; \theta)}{g(x; \theta)}\right) f(x; \theta) dx$$

se f è assolutamente continua rispetto a g altrimenti la distanza viene posta pari a $+\infty$. Analogamente, nel caso discreto

$$D(p, q) = \sum_{x_i} \log\left(\frac{p(x_i; \theta)}{q(x_i; \theta)}\right) p(x_i; \theta)$$

Spiegazione Semplice

Cos'è la "Distanza" di Kullback-Leibler (KL)?

La distanza KL (o "divergenza KL") è un concetto teorico fondamentale. Misura quanto due distribuzioni di probabilità, f e g , sono diverse tra loro.

- $D(f, g) = 0$: Se f e g sono identiche.
- $D(f, g) > 0$: Se f e g sono diverse.

È una "distanza" asimmetrica: la distanza da f a g non è uguale alla distanza da g a f . Ci dice, in media, "quanta informazione perdiamo" se usiamo la distribuzione g per approssimare la distribuzione f .

Remark 99. E' facilmente verificabile che la distanza di Kullback-Leibler è sempre non-negativa; infatti, utilizzando la diseguaglianza di Jensen

$$\begin{aligned} D(f, g) &= E_f[\log \frac{f(X)}{g(X)}] = -E_f[\log \frac{g(X)}{f(X)}] \\ &\geq -\log(E_f[\frac{g(X)}{f(X)}]) = -\log\left(\int \frac{g(x; \theta)}{f(x; \theta)} f(x; \theta) dx\right) = -\log\left(\int g(x; \theta) dx\right) = 0. \end{aligned}$$

D'altra parte la distanza di Kullback-Leibler non identifica una metrica, ad esempio perché non è simmetrica negli argomenti (e non vale la diseguaglianza triangolare).

Quale è la relazione tra funzione di verosimiglianza e distanza di Kullback-Leibler? Focalizziamoci come al solito sulla log-verosimiglianza; considerando che la stimatore è invariante se la funzione è trasformata linearmente con coefficienti che non dipendono dal parametro, possiamo passare da $\log L = \sum_{i=1}^n \log f(X_i; \theta)$ a

$$M_n(\theta) := \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta) - \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta_0) = \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i; \theta)}{f(X_i; \theta_0)}$$

dove θ_0 denota il "vero" valore del parametro. Sotto alcune delle condizioni di regolarità che abbiamo visto precedentemente affinché valga la legge dei grandi numeri, abbiamo che, ad un valore di θ fissato:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i; \theta)}{f(X_i; \theta_0)} &\rightarrow_p E_{\theta_0}[\log \frac{f(X_1; \theta)}{f(X_1; \theta_0)}] \\ &= \int \log \frac{f(x; \theta)}{f(x; \theta_0)} f(x; \theta_0) dx \\ &= -D(f_{\theta_0}, f_{\theta}). \end{aligned}$$

L'idea euristica dietro alla dimostrazione della consistenza è dunque la seguente: $M_n(\theta)$ converge a $-D(f_{\theta_0}, f_{\theta})$, che è sempre negativa tranne quando $\theta = \theta_0$, dove vale zero. Poiché lo stimatore di massima

verosimiglianza è definito come il valore di θ che massimizza $M_n(\theta)$, a sua volta esso convergerà al valore che massimizza $-D(f_{\theta_0}, f_\theta)$, cioè $\theta = \theta_0$. Nel seguito, scriviamo per brevità $M(\theta) = -D(f_{\theta_0}, f_\theta)$ abbiamo la seguente proposizione:

Proposition 100 (Consistenza). *Assumiamo che valgano le seguenti condizioni di regolarità:*

a) (*Legge dei grandi numeri uniforme*) *Si ha che*

$$\sup_\theta |M_n(\theta) - M(\theta)| = o_p(1) \text{ per } n \rightarrow \infty;$$

b) (*Convessità*) *Si ha che per ogni $\epsilon > 0$ esiste $\eta_\epsilon > 0$ tale che $|\theta - \theta_0| > \epsilon \Rightarrow M(\theta_0) - M(\theta) > \eta_\epsilon$*

Allora $\hat{\theta}_{MLE;n} \rightarrow_p \theta_0$.

Remark 101. L'ipotesi a) è più forte delle leggi dei grandi numeri che abbiamo dimostrato in precedenza, perché richiede che la convergenza sia uniforme in θ , il che non è garantito anche se la convergenza avviene puntualmente per ogni valore fisso di θ . Le leggi dei grandi numeri uniformi sono ampiamente trattate nei corsi della magistrale (Teoremi di Glivenko-Cantelli) ma non fanno parte del programma di questo corso. L'ipotesi b) è essenzialmente una condizione di identificabilità: se esistesse un insieme di valori diversi di θ a distanza di Kullback-Leibler 0 da θ_0 il problema della consistenza sarebbe evidentemente irrisolvibile (e lo stimatore potrebbe non essere nemmeno ben definito).

Dimostrazione. L'idea è di mostrare che per ogni $\epsilon > 0$ fissato si ha $Pr\{M(\theta_0) - M(\hat{\theta}_n) > \eta_\epsilon\} \rightarrow 0$; grazie alla condizione di convessità abbiamo

$$Pr\{|\theta_0 - \hat{\theta}_n| > \epsilon\} \leq Pr\{M(\theta_0) - M(\hat{\theta}_n) > \eta_\epsilon\}$$

e conseguentemente la proposizione sarà dimostrata. A questo proposito notiamo che, per definizione di stimatore di massima verosimiglianza

$$M_n(\hat{\theta}_n) - M_n(\theta_0) > 0$$

e quindi

$$\begin{aligned} M(\hat{\theta}_n) - M(\theta_0) &= M_n(\theta_0) - M(\hat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \\ &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + M(\theta_0) - M_n(\theta_0). \end{aligned}$$

Segue quindi che

$$\begin{aligned} Pr\{M(\theta_0) - M(\hat{\theta}_n) > \eta_\epsilon\} &\leq Pr\{|M(\theta_0) - M_n(\theta_0)| > \frac{\eta_\epsilon}{2}\} + Pr\{|M_n(\hat{\theta}_n) - M(\hat{\theta}_n)| > \frac{\eta_\epsilon}{2}\} \\ &\rightarrow 0, \text{ per } n \rightarrow \infty, \end{aligned}$$

usando per il primo termine la legge dei grandi numeri standard e per il secondo quella uniforme (che implica ovviamente quella standard). \square

11.0.2 La funzione punteggio

Definition 102 (Funzione Punteggio). *La funzione punteggio (score function in inglese) è definita da*

$$s_n(\theta; X_1, \dots, X_n) = \frac{\partial}{\partial \theta} \log L(\theta; X_1, \dots, X_n),$$

assumendo ovviamente che la derivata esista. In genere, si tratta di una funzione da \mathbb{R}^p in \mathbb{R}^p .

Spiegazione Semplice

Cos'è la "Funzione Punteggio" (Score)?

È semplicemente la **derivata prima** della log-verosimiglianza $l(\theta) = \log L(\theta)$.

Ricorda che per trovare il massimo di una funzione (il nostro $\hat{\theta}_{MLE}$), dobbiamo trovare il punto in cui la sua derivata (lo "score") è uguale a zero.

$\hat{\theta}_{MLE}$ è la soluzione dell'equazione $s_n(\theta) = 0$.

Lo "score" $s_n(\theta)$ è un vettore (il gradiente) che "punta" nella direzione in cui la verosimiglianza cresce più velocemente.

Nella discussione che segue inizieremo dal caso $p = 1$. Notiamo innanzitutto che la funzione punteggio è essa stessa una variabile aleatoria, per ogni valore di θ fissato. Ha quindi senso domandarsi quale sia il suo valor medio e la sua varianza, ammesso che esistano. Iniziamo quindi a imporre alcune condizioni di regolarità.

Definition 103 (Condizioni di regolarità). *Valgono le seguenti ipotesi:*

- *Le condizioni di consistenza*
- *Le osservazioni X_1, \dots, X_n sono indipendenti ed identicamente distribuite*
- *La log-verosimiglianza ammette due derivate continue, $\frac{\partial^2}{\partial \theta^2} \log L_n \in C^2$*
- *La derivata seconda ha momento secondo finito, $E[|\frac{\partial^2}{\partial \theta^2} \log L_n|] < \infty$*
- *Il vero valore del parametro θ_0 appartiene all'interno dell'insieme dei valori ammissibili, $\theta_0 \in \text{Int}(\Theta)$*
- *E' possibile scambiare due volte la derivata con l'integrale nel valor medio della log-verosimiglianza*

Remark 104. Queste condizioni sono più forti del necessario per quello che riguarda i risultati sulla funzione punteggio, ma preferiamo enunciarle una volta per tutte in funzione del risultato sulla asintotica Gaussianità degli stimatori ML.

Lemma 105. *Sotto le condizioni di regolarità, abbiamo che*

$$E_{\theta_0}[s_n(\theta_0; X_1, \dots, X_n)] = 0.$$

Dimostrazione. Per definizione abbiamo che

$$\begin{aligned} E_{\theta_0}[s_n(\theta_0; X_1, \dots, X_n)] &= \int \{\log' L(\theta_0; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta_0) dx_1 \dots dx_n \\ &= \int \frac{\frac{\partial f(x_1, \dots, x_n; \theta)}{\partial \theta}|_{\theta=\theta_0}}{f(x_1, \dots, x_n; \theta_0)} f(x_1, \dots, x_n; \theta_0) dx_1 \dots dx_n \\ &= \int \frac{\partial f(x_1, \dots, x_n; \theta)}{\partial \theta}|_{\theta=\theta_0} dx_1 \dots dx_n. \end{aligned}$$

Poiché per ipotesi possiamo scambiare derivate ed integrali, abbiamo che

$$\begin{aligned} \int \frac{\partial f(x_1, \dots, x_n; \theta)}{\partial \theta}|_{\theta=\theta_0} dx_1 \dots dx_n &= \frac{\partial}{\partial \theta} \int f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n|_{\theta=\theta_0} \\ &= \frac{\partial}{\partial \theta} 1|_{\theta=\theta_0} = 0. \end{aligned}$$

□

Spiegazione Semplice

Prima Proprietà dello Score: La Media è Zero

Questo lemma ci dice che se calcoliamo lo "score" (la derivata prima) nel punto θ_0 (il *vero* valore del parametro), il suo valore medio (atteso) è 0.

Questo è intuitivo: il θ_0 vero è il valore che massimizza la funzione di verosimiglianza "teorica" (la distanza KL). In un punto di massimo, la derivata (lo score) deve essere 0.

Poiché lo stimatore $\hat{\theta}_{MLE}$ è il punto dove lo score $s_n(\theta)$ è *effettivamente* 0, questo ci dice che il nostro stimatore sta cercando di trovare il punto (θ_0) dove lo score *dovrebbe essere* 0 in media.

Example 106. Se consideriamo il caso della verosimiglianza Gaussiana, abbiamo facilmente (assumendo varianza nota e pari a 1, X_i iid $N(\mu_0, 1)$)

$$s_n(\mu; X_1, \dots, X_n) = \sum_{i=1}^n (X_i - \mu),$$

ed è evidentemente che il valor medio si annulla prendendo $\mu = \mu_0$:

$$E[s_n(\mu_0; X_1, \dots, X_n)] = n(\mu_0 - \mu_0) = 0.$$

(Nota: il testo originale riportava $E[s_n(\mu; \dots)] = n(\mu - \mu_0)$. Calcolando in μ_0 si ottiene 0.)

Lemma 107. Sotto le condizioni di regolarità, abbiamo che

$$\begin{aligned} E\left[\frac{\partial^2 \log L(\theta; x_1, \dots, x_n)}{\partial \theta^2}\Big|_{\theta=\theta_0}\right] &= \int \{\log'' L(\theta_0; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta_0) dx_1 \dots dx_n \\ &= - \int \{\log' L(\theta_0; x_1, \dots, x_n)\}^2 f(x_1, \dots, x_n; \theta_0) dx_1 \dots dx_n \\ &= -E\left[\left\{\frac{\partial \log L(\theta; x_1, \dots, x_n)}{\partial \theta}\Big|_{\theta=\theta_0}\right\}^2\right] \end{aligned}$$

Dimostrazione. Si noti innanzitutto che, per qualsiasi valore di θ

$$H(\theta) := \int \{\log' L(\theta; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \equiv 0.$$

Di nuovo per le condizioni di regolarità possiamo allora considerare

$$\begin{aligned} 0 &= \frac{\partial H(\theta)}{\partial \theta} = H'(\theta) \\ &= \frac{\partial}{\partial \theta} \int \{\log' L(\theta; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \\ &= \int \{\log'' L(\theta; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \\ &\quad + \int \{\log' L(\theta; x_1, \dots, x_n)\} f'(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \\ &= \int \{\log'' L(\theta; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \\ &\quad + \int \{\log' L(\theta; x_1, \dots, x_n)\} \frac{f'(x_1, \dots, x_n; \theta)}{f(x_1, \dots, x_n; \theta)} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \\ &= \int \{\log'' L(\theta; x_1, \dots, x_n)\} f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \\ &\quad + \int \{\log' L(\theta; x_1, \dots, x_n)\}^2 f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \end{aligned}$$

ed il risultato è dimostrato. \square

Remark 108. È importante notare che in $\theta = \theta_0$

$$\begin{aligned} &\int \{\log' L(\theta_0; x_1, \dots, x_n)\}^2 f(x_1, \dots, x_n; \theta_0) dx_1 \dots dx_n \\ &= E[(s_n(\theta_0))^2] = \text{Var}(s_n(\theta_0)). \end{aligned}$$

La varianza della funzione punteggio calcolata nel valore vero del parametro è pertanto uguale al reciproco della derivata seconda della log-verosimiglianza nello stesso punto.

Definition 109 (Informazione di Fisher). La varianza della funzione punteggio è nota come informazione di Fisher e si indica con $I_n(\theta_0)$. In generale, si tratta di una matrice di dimensione $p \times p$.

Spiegazione Semplice

Cos'è l' "Informazione di Fisher"?

Questo è uno dei concetti più importanti in statistica.

- **Definizione 1:** È la **varianza dello score** (la derivata prima). Misura "quanto balla" la derivata prima attorno alla sua media (che è 0).
- **Definizione 2 (dal Lemma 107):** È anche l'opposto (–) della **media della derivata seconda**.

Interpretazione intuitiva: La derivata seconda misura la *curvatura* della funzione di log-verosimiglianza nel punto di massimo θ_0 .

- **Picco "appuntito" (Alta curvatura):** La derivata seconda è un numero negativo grande. L'Informazione di Fisher I_n è *grande*. Questo significa che i dati contengono *molta informazione*: se ci spostiamo anche di poco da θ_0 , la verosimiglianza crolla. La nostra stima $\hat{\theta}$ sarà molto precisa.
- **Picco "piatto" (Bassa curvatura):** La derivata seconda è un numero negativo piccolo. L'Informazione di Fisher I_n è *piccola*. I dati contengono *poca informazione*: molti valori di θ sono quasi altrettanto "verosimili" quanto θ_0 . La nostra stima $\hat{\theta}$ sarà poco precisa (avrà alta varianza).

L'Informazione di Fisher $I_n(\theta)$ è la misura di "quanta informazione" il campione n contiene sul parametro θ .

11.0.3 L'asintotica Gaussianità degli stimatori MLE

Theorem 110. *Sotto le condizioni di regolarità di cui sopra, si ha che*

$$\sqrt{n}\{\hat{\theta}_n - \theta_0\} \xrightarrow{d} N(0, I_1^{-1}(\theta_0)).$$

Equivalentemente, abbiamo che

$$I_n^{1/2}(\theta_0)\{\hat{\theta}_n - \theta_0\} \xrightarrow{d} N(0, Id_p),$$

dove Id_p indica la matrice di identità di ordine p .

Spiegazione Semplice

La "Pagella" Completa dello Stimatore MLE

Questo è il secondo grande risultato (dopo la consistenza) e completa la "pagella" dello stimatore di Massima Verosimiglianza (MLE).

Ci dice che l'MLE non solo è consistente (Proposizione 100), ma è anche **asintoticamente Gaussiano**.

L'errore "zoomato" dello stimatore ($\sqrt{n}(\hat{\theta}_n - \theta_0)$) converge a una curva a campana (Gaussiana) $N(0, \dots)$.

E qual è la varianza di questa curva? È $I_1^{-1}(\theta_0)$, cioè l'**inverso dell'Informazione di Fisher** (calcolata per una singola osservazione, I_1).

Questo collega tutto:

- Tanta informazione (I_1 grande) \implies Picco "appuntito" \implies Varianza dell'errore I_1^{-1} *piccola* \implies Stima **molto precisa**.
- Poca informazione (I_1 piccola) \implies Picco "piatto" \implies Varianza dell'errore I_1^{-1} *grande* \implies Stima **poco precisa**.

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

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

da cui

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\log L'(\theta_0)/\sqrt{n}}{\log L''(\bar{\theta}_n)/n}$$

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

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

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 Slutsky otteniamo

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

Combinando i due risultati ed usando di nuovo Slutsky si arriva all'enunciato del Teorema. \square

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

