Varianza di stimatori maximum-likelihood

di Luigi Paparella

presentazione didattica per il corso di

Tecniche di trattamento dei dati

Anno Accademico 2011/2012

docente: Prof. A. Pompili

Stimatori

Supponiamo di avere una variabile aleatoria x distribuita secondo la pdf f(x).

Il problema fondamentale della statistica consiste nel dedurre le proprietà a partire da un campione di n osservazioni sperimentali $x_1, \ldots, x_n$.

Le espressioni che permettono di ottenere da un campione casuale tratto da una popolazione la migliore stima dei parametri sono detti stimatori.

Lo stimatore per una certa quantità θ è denotato di solito con $\hat{\theta}$.

Varianza, bias

Uno stimatore è funzione delle random variables $x_1, \ldots, x_n$,

per cui è anch'esso una RV.

Si possono definire valore d'aspettazione e varianza di $\hat{\theta}$:

$$E[\hat{\theta}(\vec{x})] = \int \hat{\theta}\, g(\hat{\theta}, \theta)\, d\hat{\theta} = \int \ldots \int \hat{\theta}(\vec{x})\, f(x_1; \theta) \cdots f(x_n; \theta)\, dx_1 \cdots dx_n.$$

$$V[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2] = E[\hat{\theta}^2] - E^2[\hat{\theta}]$$

Si definisce <u>bias</u> la quantità:

$$b = E[\hat{\theta}] - \theta$$

Si preferisce in genere avere stimatori non biassati (b=0) e con piccola varianza. Altra proprietà di pregio è la <u>coerenza</u>, ossia $\hat{\theta}$ converge a θ nel limite di qrande n.

Metodo della massima verosimiglianza

La procedura per la determinazione delle espressioni degli stimatori, detta inferenza statistica, è basata su diversi metodi; tra questi i più usati sono il metodo della massima verosimiglianza e il metodo dei minimi quadrati.

Il principio di *massima verosimiglianza* (*Maximum Likelihood*, ML) assume che la migliore stima dei parametri della distribuzione è quella che rende massima la probabilità di ottenere i valori osservati (x_1 , …, x_n).

Il metodo consiste nel trovare il valore del parametro θ che massimizza la funzione di verosimiglianza

$$L(\theta)=\prod_{i=1}^{n} f(x_i,\theta) \qquad (1)$$

che è proprio la pdf congiunta per le x_i, trattata come funzione del parametro θ.

Varianza degli stimatori ML: metodo Monte Carlo

Il calcolo analitico della varianza di uno stimatore ML richiede spesso il calcolo di integrali troppo complicati da risolvere analiticamente. In questi casi si può fornire una stima della varianza utilizzando un metodo Monte Carlo.

Il metodo può essere schematizzato in questi tre punti:

- Si simula un grande numero N di esperimenti di n misure indipendenti;
- Per ogni esperimento si calcola la stima $\hat{\theta}$ del parametro θ (si hanno così N stime indipendenti);
- Si calcola la deviazione standard campionaria

$$s(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^{N} (\theta_i - \bar{\theta})^2}{N-1}}$$

Esempio: applicazione metodo MC

Supponiamo di avere una RV distribuita esponenzialmente con media $\tau=1,0$.

$$f(t;\tau)=\frac{1}{\tau}e^{-\frac{t}{\tau}} \qquad (2)$$

Si può dimostrare che lo stimatore ML di τ è:

$$\hat{\tau}=\frac{1}{n}\sum_{i=1}^{n}t_i \quad (3) \quad ; \quad E[\hat{\tau}]=\tau \quad (4) \quad ; \quad V[\hat{\tau}]=\frac{\tau^2}{n} \quad (5)$$

Si supponga che la stima di $\hat{\tau}$ su un ampione di n=50 misure abbia dato come risultato $\hat{\tau}=1,062$

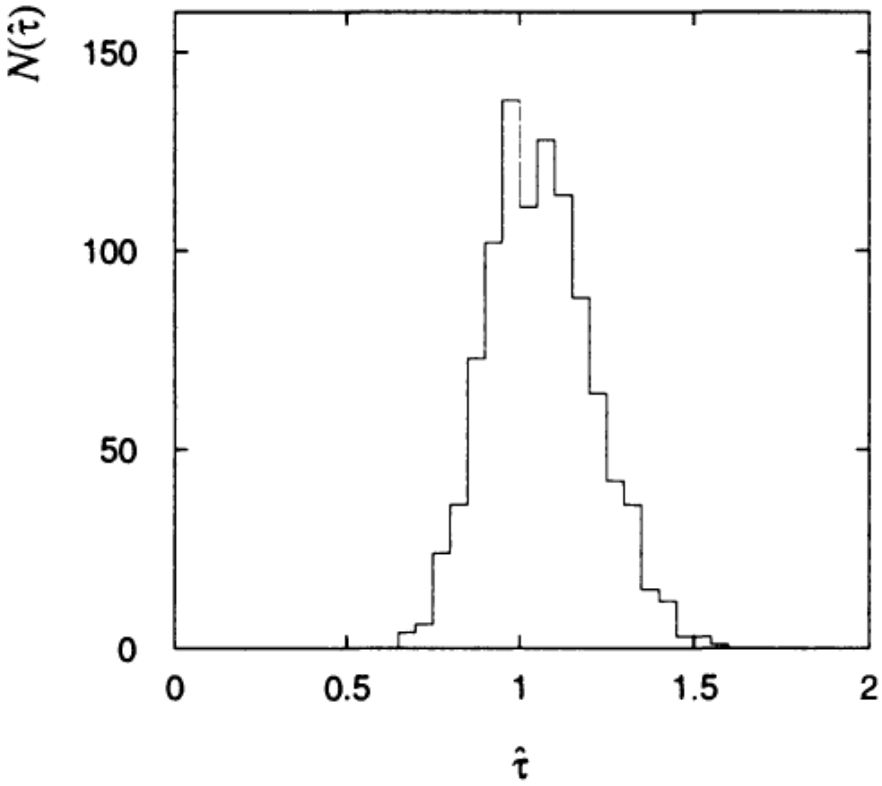
Si vuole calcolare la varianza di $\hat{\tau}$ con il metodo MC.

Esempio: applicazione metodo MC

Sono stati simulati 1000 esperimenti di 50 misure e per ciascuno è stata calcolata la media $\hat{\tau}$.

La deviazione standard campionaria dei 1000 esperimenti è s=0,151 molto vicino al valore teorico

$$\sigma_{\hat{\tau}} = \frac{\hat{\tau}}{n} = \frac{1,062}{\sqrt{50}} = 0,150$$



Si noti che la distribuzione di $\hat{\tau}$ è approssimativamente gaussiana.

Questa è una proprietà generale degli stimatori ML nel limite di grande campione, nota come normalità asintotica.

Disuguaglianza di Rao-Cramer-Frechet

La disuguaglianza di Rao-Cramer fornisce un limite inferiore alla varianza di uno stimatore:

$$V[\hat{\theta}] \geq \frac{\left(1+\frac{\partial b}{\partial \theta}\right)^2}{E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right]} \qquad (6)$$

Essa vale in generale, anche per stimatori non ML.

Se vale il segno di uguaglianza (cioè minima varianza), tale stimatore è detto efficiente.

Esempio

Nel caso di una distribuzione esponenziale con media τ si ha:

$$\log L(\tau) \overset{(1)}{=} \log \prod_{i=1}^{n} f(t_i, \tau) \overset{(2)}{=} \sum_{i=1}^{n} \log\left(\frac{1}{\tau} e^{-\frac{t_i}{\tau}}\right) = \sum_{i=1}^{n} \left(\log\frac{1}{\tau} - \frac{t_i}{\tau}\right) =$$

$$= -n\log\tau - \frac{1}{\tau}\sum_{i=1}^{n} t_i \overset{(3)}{=} n\left(-\log\tau - \frac{\hat{\tau}}{\tau}\right)$$

$$\frac{\partial logL}{\partial \tau} = n\left(-\frac{1}{\tau} + \frac{\hat{\tau}}{\tau^2}\right) \qquad \frac{\partial^2 logL}{\partial \tau^2} = n\left(\frac{1}{\tau^2} - \frac{2\hat{\tau}}{\tau^3}\right) = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

$\frac{\partial b}{\partial \tau} = 0$ essendo $\hat{\tau}$ unbiased per la (4).

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)\right]} = \frac{1}{-\frac{n}{\tau^2}\left(1 - \frac{2E[\hat{\tau}]}{\tau}\right)} \overset{(4)}{=} \frac{1}{-\frac{n}{\tau^2}\left(1 - \frac{2\tau}{\tau}\right)} = \frac{\tau^2}{n}$$

Poichè per la (5) vale il segno di uguaglianza, possiamo affermare che $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ è uno stimatore efficiente di τ.

Informazione, Score

Al fine di dimostrare la disuguaglianza di Rao-Cramer è utile introdurre le seguenti due quantità:

- Informazione:

$$I_{\vec{x}}(\theta)=E\left[\left(\frac{\partial \ln L(\vec{x},\theta)}{\partial \theta}\right)^2\right] \qquad (7)$$

- Score:

$$S(\vec{x},\theta)=\frac{\partial \ln L(\vec{x},\theta)}{\partial \theta} \qquad (8)$$

($\vec{x}$ è un set di n osservazioni della RV x)

- Proprietà:

$$\left.\begin{aligned} I_{\vec{x}}(\theta)&=E[S(\vec{x},\theta)^2] \quad (9)\\ E[S(\vec{x},\theta)]&=0 \quad (10)\\ I_{\vec{x}}(\theta)&=-E\left[\frac{\partial S(\vec{x},\theta)}{\partial \theta}\right] \quad (11)\end{aligned}\right\} \rightarrow I_{\vec{x}}(\theta)=V[S(\vec{x},\theta)] \quad (12)$$

la (10) e la (11) valgono con ipotesi molto generali sulla regolarità di $L(\vec{x},\theta)$

Dimostrazione disuguaglianza RCF

$$E[\hat{\theta}\,S(\vec{x},\theta)]=\int\ldots\int\hat{\theta}\left[\frac{\partial}{\partial\theta}\ln L(\vec{x},\theta)\right]L(\vec{x},\theta)\,dx_1\cdots dx_n=$$

$$\int\ldots\int\hat{\theta}\left[\frac{1}{L(\vec{x},\theta)}\frac{\partial}{\partial\theta}L(\vec{x},\theta)\right]L(\vec{x},\theta)\,dx_1\cdots dx_n=$$

$$\int\ldots\int\hat{\theta}\left[\frac{\partial}{\partial\theta}L(\vec{x},\theta)\right]dx_1\cdots dx_n=\int\ldots\int\hat{\theta}\frac{\partial}{\partial\theta}\left[\prod_{i=1}^{n}f(x_i,\theta)\,dx_i\right]=$$

$$\int\ldots\int\frac{\partial}{\partial\theta}\left[\hat{\theta}\prod_{i=1}^{n}f(x_i,\theta)\,dx_i\right]$$

dove l'ultimo passaggio segue per il fatto che $\hat{\theta}$ è una statistica, perciò non dipende da θ.

Assumendo di poter scambiare l'ordine di differenziazione e integrazione

$$E[\hat{\theta}\,S(\vec{x},\theta)]=\frac{\partial}{\partial\theta}\int\ldots\int\hat{\theta}\prod_{i=1}^{n}[f(x_i,\theta)\,dx_i]=\frac{\partial}{\partial\theta}E[\hat{\theta}]=$$

$$\frac{\partial}{\partial\theta}[\theta+b(\hat{\theta})]=1+\frac{\partial}{\partial\theta}b(\hat{\theta})$$

Dimostrazione disuguaglianza RCF

Calcoliamo il coefficiente di correlazione delle RV $S(\vec{x},\theta)$ e $\hat{\theta}$.

$$cov[S(\vec{x},\theta),\hat{\theta}(\vec{x})]=E[S(\vec{x},\theta)\hat{\theta}(\vec{x})]-E[S(\vec{x},\theta)]E[\hat{\theta}(\vec{x})]\overset{(10)}{=}$$
$$E[S(\vec{x},\theta)\hat{\theta}(\vec{x})]=1+\frac{\partial}{\partial\theta}b(\theta)$$

$$\rho^2_{S,\hat{\theta}}=\frac{\{cov[S,\hat{\theta}]\}^2}{V[S]V[\hat{\theta}]}\overset{(12)}{=}\frac{\left[1+\frac{\partial}{\partial\theta}b(\hat{\theta})\right]^2}{I(\theta)V[\hat{\theta}]}$$

Essendo $\rho^2\leq 1$ se ne deduce che:

$$\sigma^2(\hat{\theta})=V[\hat{\theta}]\geq\frac{\left[1+\frac{\partial}{\partial\theta}b(\hat{\theta})\right]^2}{I(\theta)}$$

che è proprio la (6), data la definizione di informazione (7).

In caso di stimatore non biassato la disuguaglianza diventa: $\sigma^2(\hat{\theta})\geq\frac{1}{I(\theta)}$

ossia la minima varianza è inversamente proporzionale all'informazione.

Per questo motivo la disuguaglianza RCF è detta anche *information inequality*.

Disuguaglianza RCF a m parametri

Generalizziamo il limite RCF al caso di m parametri $\vec{\theta}=(\theta_{1,}\cdots,\theta_m)$

Assumendo stimatori efficienti e zero-bias, si ha per la matrice di covarianza $V_{ij}=cov[\hat{\theta}_i,\hat{\theta}_j]$ la seguente relazione:

$$(V^{-1})_{ij}=E\left[-\frac{\partial^2 \log L}{\partial\theta_i\partial\theta_j}\right]$$

Esplicitando il valore d'aspettazione l'equazione diventa:

$$(V^{-1})_{ij}=\int\ldots\int -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\sum_{k=1}^{n}\log f(x_k;\vec{\theta})\right)\prod_{l=1}^{n} f(x_l;\vec{\theta})\,dx_l=$$

$$n\int -f(x,\vec{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x;\vec{\theta})\,dx$$

dove $f(x;\vec{\theta})$ è la pdf per la RV x, per la quale si hanno n misure.

Disuguaglianza RCF a m parametri

In alcuni casi questo integrale può essere difficile da risolvere analiticamente.

Se il campione di dati è sufficientemente grande si può stimare V^{-1} valutando la derivata seconda con la stimaML $\hat{\theta}$:

$$(\hat{V^{-1}})_{ij} = -\left[\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right]_{\theta=\hat{\theta}} \qquad (13)$$

Per un singolo parametro questa diventa:

$$\hat{\sigma^2}_{\hat{\theta}} = -\frac{1}{\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]_{\theta=\hat{\theta}}} \qquad (14)$$

Metodo grafico

Si consideri il caso di un singolo parametro θ e si espandi la funzione log-likelihood in serie di Taylor attorno alla stima $\hat{\theta}$:

$$\log L(\theta)=\log L(\hat{\theta})+\left[\frac{\partial \log L}{\partial \theta}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta})+\frac{1}{2!}\left[\frac{\partial^2 \log L}{\partial \theta^2}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2+\ldots$$

Per come è definita la stima ML si ha:

$$\log L(\hat{\theta})=\log L_{MAX} \qquad \text{e} \qquad \left[\frac{\partial \log L}{\partial \theta}\right]_{\theta=\hat{\theta}}=0$$

Utilizzando queste due equazioni e la (14) si ha:

$$\log L(\theta)=\log L_{MAX}-\frac{(\theta-\hat{\theta})^2}{2\hat{\sigma}^2_{\hat{\theta}}} \quad \text{o} \quad \log L(\hat{\theta}\pm\hat{\sigma}_{\hat{\theta}})=\log L_{MAX}-\frac{1}{2} \qquad (15)$$

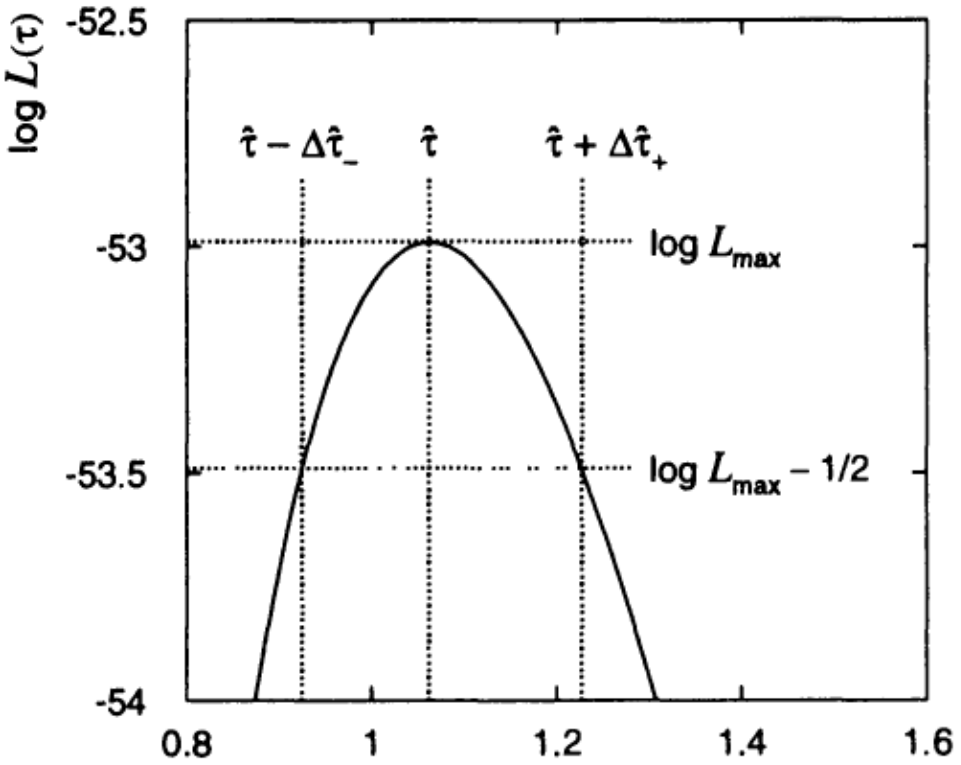
Si può dimostrare che la funzione log-likelihood diventa una parabola nel limite di grande campione (e quindi L diventa una gaussiana). Anche se logL non è paralabolica si può adottare l'equazione (14) come definizione dell'errore statistico.

Esempio

Riprendiamo l'esempio della distribuzione esponenziale.

Si è costruito un plot della funzione di likelihood in funzione del parametro τ. Sono stati individuati I punti $\hat{\tau}-\Delta\hat{\tau}$ e $\hat{\tau}+\Delta\hat{\tau}$ dove la funzione si riduce di 0,5 rispetto al valore massimo.

Si è trovato $\hat{\sigma}_{\hat{\tau}} \approx \Delta\tau_+ \approx \Delta\tau_- \approx 0{,}15$ in perfetto accordo con il calcolo analitico.



Esempio: ML con due parametri

Supponiamo di una RV x $(-1 \le x \le 1)$ distribuita come una pdf dipendente da due parametri α e β .

$$f(x;\alpha,\beta)=\frac{1+\alpha x+\beta x^2}{2+2\beta/3} \qquad (16)$$

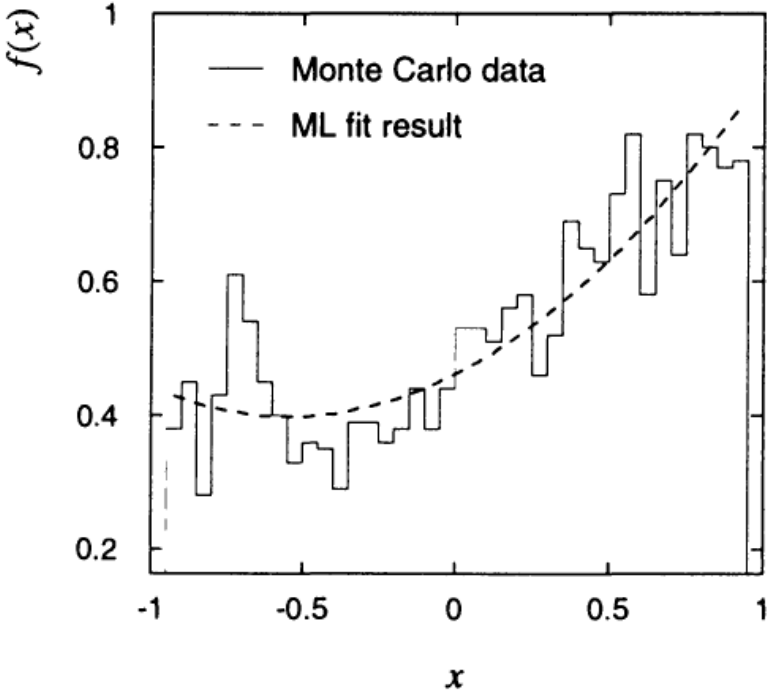
Ad esempio questa può essere la distribuzione angolare dell'angolo di scattering (x=cosθ) per la reazione $e^+ e^- \to \mu^+ \mu^-$ (con α=0 , β=1).

Generalizzando la (16) a un intervallo arbitrario $-x_{min} \le x \le x_{max}$ si ottiene, ricalcolando la costante di normalizzazione:

$$f(x;\alpha,\beta)=\frac{1+\alpha x+\beta x^2}{(x_{max}-x_{min})+\frac{\alpha}{2}(x_{max}^2-x_{min}^2)+\frac{\beta}{3}(x_{max}^3-x_{min}^3)}$$

Esempio: RCF bound

È stato simulato un esperimento di 2000 eventi con I parametri $\alpha=0,5$; $\beta=0,5$; $x_{min}=-0,95$; $x_{max}=0,95$.



Massimizzando numericamente la funzione di likelihood si è trovato

$$\hat{\alpha}=0,508\pm0,052$$
$$\hat{\beta}=0,47\pm0,11$$

dove gli errori statistici sono le radci delle varianze.

Quest'ultime sono state stimate calcolando numericamente la matrice delle derivate seconde della funzione log-likelihhod e invertendo per ottenere la matrice delle covarianze (equazione 13).

Si è trovato inoltre, $\widehat{cov}[\hat{\alpha},\hat{\beta}]=0,0026$; $r=0,46$

cioè α e β sono positivamente correlate.

Esempio: metodo Monte Carlo

Questo risulta più evidente se si gauarda lo scatter plot delle due RV $\hat{\alpha}$ *e* $\hat{\beta}$.

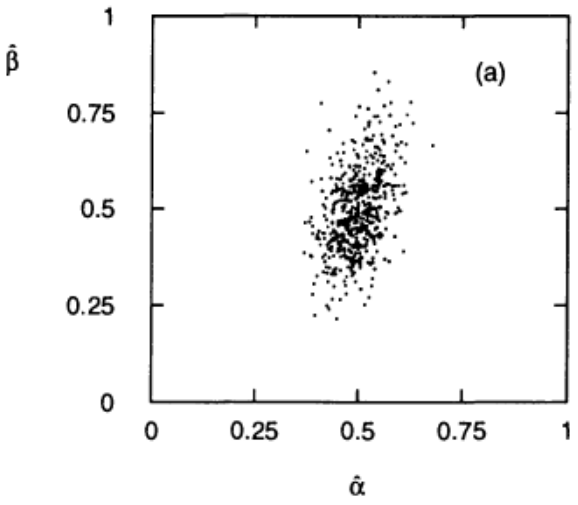
Esso è stato ottenuto simulando 500 esperimenti di 2000 eventi. Si è ottenuto:

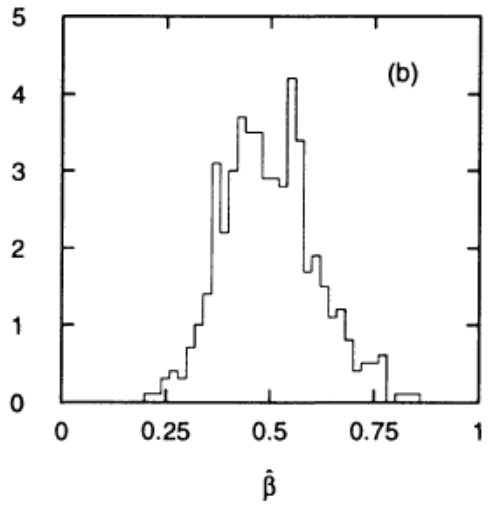
$$\overline{\hat{\alpha}} = 4{,}99 \qquad \overline{\hat{\beta}} = 0{,}498$$

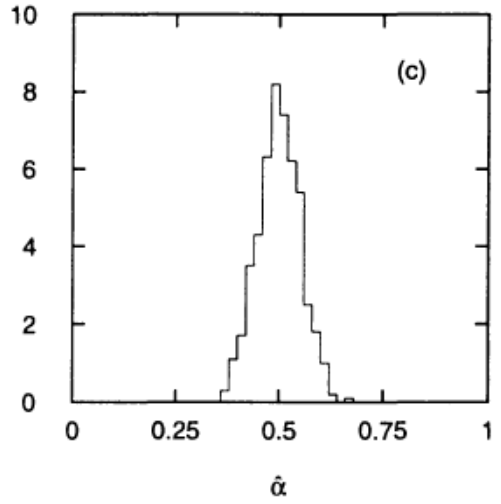
$$s_{\hat{\alpha}} = 0{,}051 \qquad s_{\hat{\beta}} = 0{,}111$$

$$\widehat{cov}[\hat{\alpha}, \hat{\beta}] = 0{,}0024 \qquad r = 0{,}42$$

in buon accordo con i valori veri e con i limiti RCF stimati.







Esempio: metodo grafico

Nel limite di grande campione la funzione log-likelihood assume la forma

$$\log L(\alpha,\beta)=\log L_{max}-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)^2+\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)^2-2\rho\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)\right]$$

dove ρ è il coefficiente di correlazione per le RV $\hat{\alpha}$ *e* $\hat{\beta}$.

Il luogo geometrico descritto da $\log L=\log L_{Mmax}-\frac{1}{2}$
è quindi dato da:

$$\frac{1}{(1-\rho^2)}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)^2+\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)^2-2\rho\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)\right]=1$$

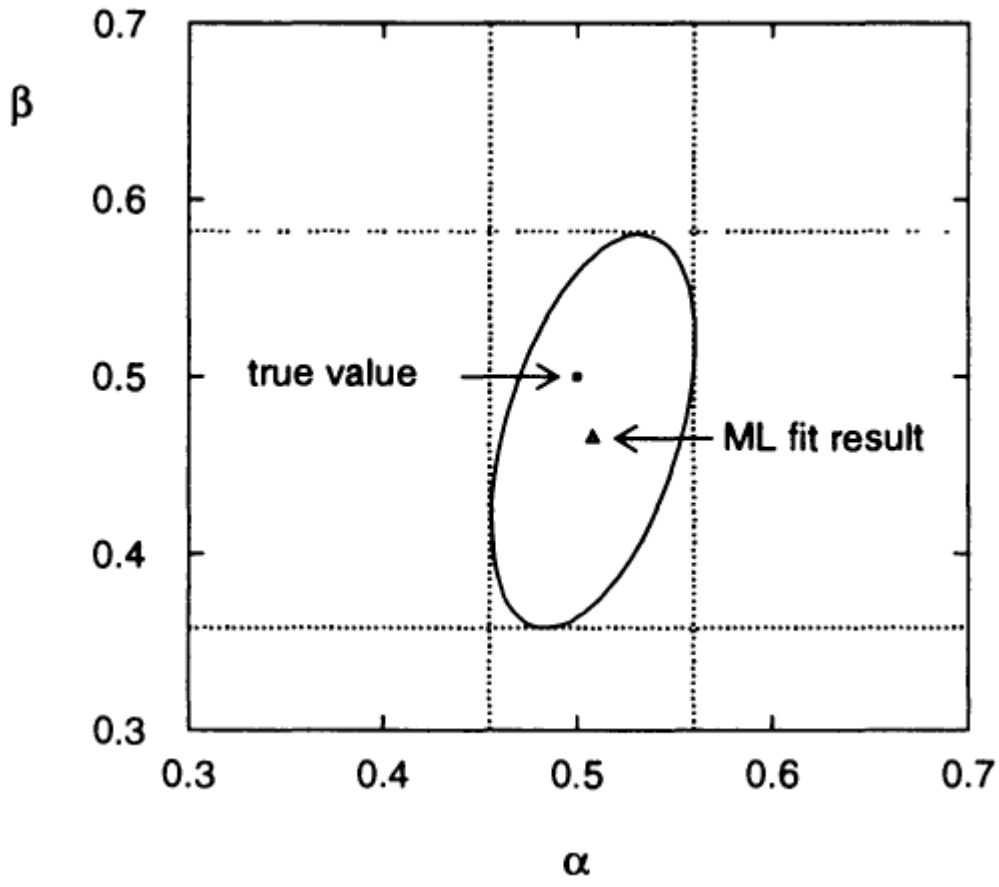
che è l'equazione di un ellisse nel piano $\alpha\beta$ centrata sulle stime ML $(\hat{\alpha},\hat{\beta})$ e formante un angolo ϕ con l'asse α dato da:

$$\tan 2\phi=\frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2-\sigma_{\hat{\beta}}^2}$$

Esempio: metodo grafico

Le tangenti all'ellisse sono a $\alpha = \hat{\alpha} \pm \sigma_{\hat{\alpha}}$, $\beta = \hat{\beta} \pm \sigma_{\hat{\beta}}$.

Se gli stimatori sono correlati, allora cambiando un parametro di una deviazione standard si avrà in generale una diminuzione della funzione di likelihood di più di 1/2. Se uno dei parametri, ad esempio β, fosse noto, allora la deviazione standard di $\hat{\alpha}$ sarebbe in qualche misura più piccola, poichè questa sarebbe data da una riduzione di ½ di logL(α).



Bibliografia

[1] Glen Cowan, Statistical Data Analysis
Oxford University Press, 1998
[2] W. J. Metzger, Statistical Methods in Data Analysis
[3] Gaetano Cannelli, Metodologie sperimentali in fisica
EdiSES, 2005