Stima Puntuale

Vers. 1.1.3

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Outline

Stima puntuale I

Abbiamo visto come i dati possono dare informazioni circa il parametro incognito di una popolazione e abbiamo individuato quale trasformazione dei dati contiene tutta l'informazione (statistica sufficiente).
Lo scopo di questo parte del corso è di trovare una stima puntuale dei parametri. Diamo qualche definizione

Definizione - Stimatore puntuale (o stimatore) per un parametro θ

Uno stimatore puntuale è qualsiasi funzione $W(X_1, \ldots, X_n)$ di un campione, usate par dire qualcosa su un parametro θ della distribuzuone di $\mathbf{X} = (X_1, \ldots, X_n)'$. Uno stimatore è una statistica.

Definizione - Stima

Una stima è il valore assunto dallo stimatore quando un campione è stato raccolto

per chiarire, la media campionaria $W(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i/n$ è uno stimatore, mentre $W(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i/n$ è la stima.

Stima puntuale II

Noi vogliamo usare uno stimatore per dire qualcosa su un parametro e quindi abbiamo bisogno di un modo per trovare “buoni” stimatori.

Esempio

Se i dati provengono da una normale e voglio dire qualcosa sulla media, la media campionaria potrebbe essere un buono stimatore, mentre il minimo probabilmente non lo è.

Per alcuni parametri ci sono dei candidati naturali, mentre per altri è più complicato, per esempio per i parametri della gamma.
Ci occuparemo di due aspetti

- Metodi per trovare stimatori (motodo dei momenti e di massima verosimiglianza);
- Metodi per valutare la bontà di uno stimatore.

Metodo dei Momenti I

Il primo metodo che vedremo è uno dei primi che è stato utilizzato, e sebbene ci siano altri metodi che danno soluzioni migliori, il metodo dei momenti è ancora molto usato perchè richiede solamente la conoscenza dei momenti.

Ricordiamo che se per la legge dei grandi numeri dato delle variabili aleatorie iid, con media μ e varianza finita, allora

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{P} \mu = \begin{cases} \int x f(x) dx & \text{se assolutamente continua} \\ \sum x p(x) & \text{se discreta} \\ \ldots & \text{casi misti e particolari} \end{cases}$$

Simili risultati valgono anche per momenti di ordine maggiore ("**Law of the unconscious statistician**")

- $\frac{\sum_{i=1}^{n} X_i^2}{n} \xrightarrow{P} E(X^2) = \mu_2$;
- $\frac{\sum_{i=1}^{n} X_i^3}{n} \xrightarrow{P} E(X^3) = \mu_3$

Metodo dei Momenti II

- ...

visto che se definiamo $Z = X^j$, il primo momento di Z corrisponde al $j-$esimo momento di X (assumendo che la varianza di Z sia finita).

Visto che i momenti dipendono dai parametri, viene naturale equiparare il momento campionario $j-$esimo $\sum_{i=i}^{n} X_i^j/n$ con il momento $j-$esimo μ_j, costruendo un sistema che ci permette di stimare i p parametri. Naturalmente mi servono p equazioni che dipendono dai p parametri

$$\frac{\sum_{i=i}^{n} X_i^1}{n} = \mu_1$$
$$\frac{\sum_{i=i}^{n} X_i^2}{n} = \mu_2$$
$$\dots$$
$$\frac{\sum_{i=i}^{n} X_i^p}{n} = \mu_p$$

Metodo dei Momenti III

generalmente si prendono i primi p momenti, se questi dipendono da tutti i parametri, e si risolve ripetto ai parametri. Le soluzioni che troviamo sono gli stimatori.

Esempio - Normale

Supponiamo di avere un campione iid di dimensione n da una normale $N(\mu, \sigma^2)$. Usiamo il metodo dei momenti per determinare degli stimatori di μ e σ.

Soluzione:

Il primo e secondo momento di una normale sono

$$\mathrm{E}(X) = \mu \quad \mathrm{E}(X^2) = \mu^2 + \sigma^2$$

Metodo dei Momenti IV

che dipendono dai parametri. Costruiamo il sistema

$$\begin{cases} \bar{X} & = \mu \\ \frac{\sum_{i=i}^{n} X_i^2}{n} & = \mu^2 + \sigma^2 \end{cases} \Longrightarrow \begin{cases} \bar{X} & = \mu \\ \frac{\sum_{i=i}^{n} X_i^2}{n} & = \bar{X}^2 + \sigma^2 \end{cases} \Longrightarrow \begin{cases} \bar{X} & = \mu \\ \frac{\sum_{i=i}^{n} X_i^2}{n} - \bar{X}^2 & = \sigma^2 \end{cases} \Longrightarrow$$

$$\begin{cases} \bar{X} & = \mu \\ \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} & = \sigma^2 \end{cases}$$

dove abbiamo usato la relazione

$$\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} = \frac{\sum_{i=i}^{n} X_i^2}{n} + \bar{X}^2 - 2\bar{X} \frac{\sum_{i=i}^{n} X_i}{n} = \frac{\sum_{i=i}^{n} X_i^2}{n} - \bar{X}^2$$

Quindi gli stimatore sono $\bar{X}$ per la media e $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$ per la varianza. (la notazione “cappello” si usa spesso per gli stimatori) $\square$

Metodo dei Momenti V

Esempio - Gamma

Supponiamo di avere un campione iid di dimensione n da una normale $G(a, b)$, con $\mathrm{E}(X) = a/b$ e $\mathrm{var}(X) = a/b^2$. Usiamo il metodo dei momenti per determinare degli stimatori di a e b.

Soluzione:

Dalla prima equazione del sistema abbiamo che

$$\bar{X} = \frac{a}{b} \Longrightarrow a = \bar{X}b$$

Metodo dei Momenti VI

Per la seconda equazione abbiamo

$$\frac{\sum_{i=i}^{n} X_i^2}{n} = \mathrm{E}^2(X) + \mathrm{Var}(X) = \frac{a(a+1)}{b^2} = \frac{\bar{X}b(\bar{X}b+1)}{b^2} = \frac{b\bar{X}^2 + \bar{X}}{b}$$

da cui segue

$$b\left(\frac{\sum_{i=i}^{n} X_i^2}{n} - \bar{X}^2\right) = \bar{X} \Longrightarrow b = \frac{\bar{X}}{\hat{\sigma}^2}$$

con $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$, e

$$a = \bar{X}b = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

che ci porta quindi ai due stimatori

$$\hat{a} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

e

$$\hat{b} = \frac{\bar{X}}{\hat{\sigma}^2}$$

Metodo dei Momenti VII

Esempio - Gamma

Supponiamo di avere un campione iid di dimensione n da una normale $\mathrm{Exp}(\lambda)$, con $\mathrm{E}(X) = 1/\lambda$. Usiamo il metodo dei momenti per determinare lo stimatore di λ.

Soluzione:

In questo caso abbiamo solo un parametro, quindi

$$\bar{X} = \frac{1}{\lambda}$$

da cui abbiamo $\hat{\lambda} = \frac{1}{\bar{X}}$ $\square$

Metodo dei Momenti VIII

Nel **“Chunk Metodo Momenti”** si possono trovare degli esempi sul metodo dei momenti.

Il problema del metodo dei momenti è che non c’è nessun vincolo circa i valori che può assumere lo stimatore, prendete il prossimo esempio

Esempio - Binomiale

Supponiamo di avere un campione iid di dimensione n da una normale $\text{Bin}(k, p)$. Usiamo il metodo dei momenti per determinare degli stimatori di k e p.

Soluzione:

Metodo dei Momenti IX

Otteniamo

$$\hat{p} = \frac{\bar{X}}{\hat{k}}$$

e

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

I passaggi li lascio come esercizio. $\square$

Dall'esempio sopra si vede come non ci sia niente che ci garantisca che $\hat{k}$ sia intero. Questo ne rendono l'uso non sempre appropriato.

Massima verosimiglianza I

Introduciamo il concetto di "**verosimiglianza**".

Definizione - Verosimiglianza

Sia $f(\mathbf{x}|\theta)$ la densità congiunta (pmf o pdf) di un campione $(X_1, \ldots, X_n)$, allora, assumendo di aver osservato $(\mathbf{X} = \mathbf{x})$ la funzione

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

è chiamata funzione di verosimiglianza.

Ci sono un paio di aspetti importanti

- La verosimiglianza (così come lo stimatore di massima verosimiglianza) non richiedono dati iid, ne dati indipendeniti, ne identicamente distribuiti, ma solo la congiunta
- $f(\mathbf{x}|\theta)$ è funzione del campione, mentre la verosimiglianza è funzione di θ, ne segue che $L(\theta|\mathbf{x})$ non integra a 1, come $f(\mathbf{x}|\theta)$, e non può essere vista come pdf o pmf.

Massima verosimiglianza II

Cerchiamo di interpretare la funzione di verosimiglianza assumendo che il campione $\mathbf{X}$ sia composto da variabili discrete e valutiamo la funzione di verosimiglianza usando due valori del parametro θ_1 e θ_2. Ipotizziamo adesso di trovare che

$$L(\theta_1|\mathbf{x}) = P_{\theta_1}(\mathbf{X} = \mathbf{x}) > P_{\theta_2}(\mathbf{X} = \mathbf{x}) = L(\theta_2|\mathbf{x})$$

che ci indica che la verosimiglianza di θ_1 è maggiore della verosimiglianza θ_2 se e solo se è la probabilità di osservare quel particolare campione da una distribuzione con parametro θ_1 ($P_{\theta_1}(\mathbf{X} = \mathbf{x})$) è maggiore di quella con θ_2 ($P_{\theta_2}(\mathbf{X} = \mathbf{x})$). La verosimiglianza ci dice allora quanto è verosimile che θ abbia generato $\mathbf{X}$, fate attenzione che θ non è variabile aleatoria e quindi non possiamo assegnargli una probabilità.

Alcuni punti su cui fare attenzione

Massima verosimiglianza III

- la differenza tra verosimiglianza e congiunta sta in cosa è fisso e cosa è variabile. Nella congiunta il parametro è fisso e consideriamo la probabilità che quel particolare campione sia osservato, mentre nella verosimiglianza fissiamo il campione e per quel particolare campione vediamo per quale parametro è più verosimile aver osservato quel campione
- θ non è una variabile aleatoria, e quindi non è possibile assegnare una probabilità all'evento $\theta = c$, ma solo quanto questo è verosimile. Per questo motivo il valore della verosimiglianza non è interpretabile e si ragiona sempre sul rapporto/confronto tra valori. Es. se
$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \frac{P_{\theta_1}(\mathbf{X}=\mathbf{x})}{P_{\theta_2}(\mathbf{X}=\mathbf{x})} = 2$$
diremo che θ_1 è più verosimile di θ_2 e , anche se non è comune, possiamo anche dire che è due volte più verosimile;

Massima verosimiglianza IV

- sebbene l'oggetto sia le verosimiglianza, vedremo che è più facile lavorare con la log-verosimiglianza, anche perchè il logaritmo è monotona crescente e quindi se un parametro è più verosimile in scala naturale lo è anche in logaritmica.

Possiamo dare una simile interpretazione nel caso continuo un intorno piccolo di $\mathbf{x}$ del tipo $(\mathbf{x} - \epsilon \mathbf{1}_n, \mathbf{x} + \epsilon \mathbf{1}_n)$ e considerando i rapporti

$$\frac{P_{\theta_1}(\mathbf{x} - \epsilon \mathbf{1}_n < \mathbf{X} < \mathbf{x} + \epsilon \mathbf{1}_n)}{P_{\theta_2}(\mathbf{x} - \epsilon \mathbf{1}_n < \mathbf{X} < \mathbf{x} + \epsilon \mathbf{1}_n)} \approx \frac{(2\epsilon)^n f(\mathbf{x}|\theta_1)}{(2\epsilon)^n f(\mathbf{x}|\theta_2)} = \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_2)} = \frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})}$$

Massima verosimiglianza V

Definizione - Principio di verosimiglianza

se $\mathbf{x}$ e $\mathbf{y}$ sono due campioni per cui le funzioni di verosimiglianza sono proporzionali

$$L(\theta|\mathbf{x}) = C(\mathbf{x},\mathbf{y})L(\theta|\mathbf{y}),\ \forall\theta \in \Theta$$

dove $C(\mathbf{x},\mathbf{y})$ è una costante che non dipende da θ, allora l'inferenza su θ basata sui due campioni dovrebbe essere la stessa

come detto prima, la verosimiglianza si valuta come rapporti tra diversi valori θ, e quindi la costante $C(\mathbf{x},\mathbf{y})$ si semplifica

$$\frac{L(\theta_1|\mathbf{y})}{L(\theta_2|\mathbf{y})} = \frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})}$$

E' quindi abbastanza ragionevole dire che il parametro che massimizza la verosimiglianza, è una buona stima del parametro vero, e quindi dobbiamo trovare la statistica funzione

Massima verosimiglianza VI

dei dati che porta al valore massimo dello verosimiglianza (o log-verosimiglianza), che è lo **stimatore di massima verosimiglianza**

Definizione - Stimatore di verosimiglianza

Per ogni possibile campione $\mathbf{x}$ definiamo $\hat{\theta}(\mathbf{x})$ (generalemnte si scrive $\hat{\theta}$) come un valore del parametro per cui $L(\theta|\mathbf{x})$ raggiunge il massimo come funzione di θ, con $\mathbf{x}$ fissato. Lo stimatore di massima verosimiglianza (MLE) del parametro θ basato su $\mathbf{X}$ è $\hat{\theta}(\mathbf{X})$.

Nel **"Chunk massima verosimiglianza"** potete vedere un esempio per capire meglio il concetto di verosimiglianza.
Assumiamo che il vettore dei parametri sia $\theta = (\theta_1, \ldots, \theta_p)$, In questo caso, se la funzione di verosimiglianza è differenziabile, allora possibili candidati sono i valori di $(\theta_1, \ldots, \theta_p)$ che soddisfano

$$\frac{\partial L(\theta|\mathbf{x})}{\partial \theta_i} = 0, \;\; i = 1, \ldots, k$$

Massima verosimiglianza VII

Fate attenzione che questi sono solo candidati perchè bisogna verificare che sono punti di massimo e non minimo, inoltre, bisogna anche verificare che il massimo non sia sui bordi del dominio.

Notate che, per costruzione, lo stimatore di massimaverosimiglianza appartiene sempre a Θ, perchè la congiunta è calcolata solo per valori del parametro in Θ, e quindi non soffre del problema dello stimatore dei momenti

Esempio - Normale

Assumiamo che $(X_1, \ldots, X_n)$ siano iid da una $N(\mu, 1)$ e di essere interessati a trovare lo stimatore di massima verosimiglianza di μ.

Soluzione:

Massima verosimiglianza VIII

Calcoliamo la verosimiglianza

$$L(\mu|\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right)$$

Calcolarne le derivate può essere complicato e è più facile farlo con la log-verosimiglianza

$$\log L(\mu|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} \propto -\sum_{i=1}^{n}(x_i-\mu)^2$$

Calcoliamo la derivata e imponiamola uguale a 0

$$\frac{d\log L(\mu|\mathbf{x})}{d\mu} = -2\sum_{i=1}^{n}(x_i-\mu) = 0$$

che da come soluzione

$$\mu = \bar{x}.$$

Massima verosimiglianza IX

che è chiaramente un massimo di $L(\mu|\mathbf{x})$ visto che la derivata seconda è

$$\frac{d^2 \log L(\mu|\mathbf{x})}{d\mu^2} = -n < 0$$

oltretutto il valore della verosimiglianza tende a 0 per $\mu \Rightarrow \infty$ e $\mu \Rightarrow -\infty$. Quindi lo stimatore di massima verosimiglianza è $W(\mathbf{X}) = \bar{X}$ e la stima è $\hat{\mu} = \bar{x}$. □

Esempio - Normale I

Assumiamo che $(X_1, \ldots, X_n)$ siano iid da una $N(\mu, 1)$ e di essere interessati a trovare lo stimatore di massima verosimiglianza di μ ma assumendo che $\Theta = (0, 1)$.

Massima verosimiglianza X

Soluzione:

La verosimiglianza è

$$L(\mu|\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right)$$

visto che Θ ha solo due valori, dobbiamo vedere quale dei due ha verosimiglianza più alta e se $L(\mu = 0|\mathbf{x}) > L(\mu = 1|\mathbf{x})$ lo stimatore di massima verosimiglianza è $W(\mathbf{X}) = 0$, se $L(\mu = 0|\mathbf{x}) < L(\mu = 1|\mathbf{x})$ lo stimatore di massima verosimiglianza è $W(\mathbf{X}) = 1$, se $L(\mu = 0|\mathbf{x}) = L(\mu = 1|\mathbf{x})$ sono entrambi stimatori di massima verosimiglianza. □

Massima verosimiglianza XI

Esempio - Normale III

Assumiamo che $(X_1, \dots, X_n)$ siano iid da una $N(\mu, \sigma^2)$ e di essere interessati a trovare lo stimatore di massima verosimiglianza di μ e σ^2.

Soluzione:

Calcoliamo la log-verosimiglianza

$$\log L(\mu, \sigma^2 | \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \propto -\frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

Massima verosimiglianza XII

Imponiamo pari a zero le derivate parziali

$$\frac{\partial \log L(\mu|\mathbf{x})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\frac{\partial \log L(\mu|\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Dovremmo adesso vedere le derivate seconde e lo Jacobiano, possiamo evitarlo facendo dei ragionamenti sulle derivate prime. Possiamo vedere che

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

porta a $\mu = \bar{x}$ indipendentemente da σ^2, e con ragionamenti simili a quelli dellìesercizio precedente possiamo determinare che è un massimo per qualsiasi valori di σ^2.

Massima verosimiglianza XIII

Il problema si riduce a un problema univariata e verificare che la soluzione $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ sia un massimo. Notate che abbiamo sotituito μ con $\bar{x}$ perchè abbiamo verificato che $\bar{x}$ è lo stimatore di μ. La verifica che sia un punto di massimo la lascio come esercizio e quindi lo stimatore della varianza è

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

La coppia $(\bar{X}, \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n})$ sono gli stimatori di massima verosimiglianza dei parametri della normale. $\square$

Massima verosimiglianza XIV

Esempio - Poisson

Assumiamo che $(X_1, \ldots, X_n)$ siano iid da una Pois(λ) e di essere interessati a trovare lo stimatore di massima verosimiglianza di λ.

Soluzione:

La log-verosimiglianza in questo caso è

$$\log L(\lambda|\mathbf{x}) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!) \propto -n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i$$

la cui derivata è

$$\frac{d \log L(\lambda|\mathbf{x})}{d\lambda} = -n\lambda + \sum_{i=1}^{n} x_i$$

Massima verosimiglianza XV

che porta alla soluzione $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$. Possiamo verificare che è un massimo vedendo che la derivata seconda

$$\frac{d^2 \log L(\lambda|\mathbf{x})}{d\lambda^2} = -n$$

è negativa e per $\lambda = 0$ e $\lambda \Rightarrow \infty$ la verosimiglinza è zero.
Lo stimatore di massima verosimiglianza è quindila media campionaria. $\square$

Lo stimatore di massima verosimiglianza è collegato alle statistiche sufficienti in un maniera molto intuitiva, e la cosa non dovrebbe sorprendere visto che se tutta l'informazione su un parametro che abbiamo nel campione è nella statistica sufficiente, allora il massimo della verosimiglianza deve essere funzione della statistica sufficiente.

Massima verosimiglianza XVI

Ricordiamo che per il teorema di fattorizzazione abbiamo che

$$f(\mathbf{x}|\theta) = h(x)g(T(\mathbf{x})|\theta)$$

ma visto che la congiunta è uguale alla verosimiglianza dobbiamo anche avere che

$$L(\theta|\mathbf{x}) = h(x)g(T(\mathbf{x})|\theta) \propto g(T(\mathbf{x})|\theta)$$

perchè $h(\mathbf{x})$ non dipende dal parametro per definizione. Possiamo andare anche oltre e ricordare che tra le tante fuznione $g(T(\mathbf{x})|\theta)$, potremmo usare anche la pfm/pdf della statistica sufficiente $q(T(\mathbf{x})|\theta)$. Adesso definiamo

$$L_T(\theta|T(\mathbf{x})) = q(T(\mathbf{x})|\theta)$$

che è la verosimiglianza basata sulal distribuzione delle statisrtica sufficiente, e abbiamo che

$$L(\theta|\mathbf{x}) \propto L_T(\theta|T(\mathbf{x}))$$

Massima verosimiglianza XVII

abbiamo quindi che possiamo massimizzare la verosimiglianza dell'intero campione, o quella della statistica sufficiente del parametro, arrivando alla stessa soluzione.

Per vederlo riprendiamo un esempio precedente e risolviamolo usando la statistica sufficiente

Esempio - Normale

Assumiamo che $(X_1, \ldots, X_n)$ siano iid da una $N(\mu, 1)$ e di essere interessati a trovare lo stimatore di massima verosimiglianza di μ.

Soluzione:
sappiamo in questo caso che la statistica sufficiente per μ è $\bar{x} \sim N(\mu, 1/n)$ e la sua log-verosimiglianza è

$$\log L(\theta|\mathbf{x}) = -\frac{1}{2}\log(2\pi/n) - \frac{n(\bar{x}-\mu)^2}{2} \propto -(\bar{x}-\mu)^2$$

Massima verosimiglianza XVIII

che per gli stessi ragionamenti fatti precedentemente ha il suo massimo a $\mu = \bar{x}$ e lo stimatore di massima verosimiglianza è quindi la media campionaria. Lo stesso stimatore trovato usando tutto il campione. $\square$

Introduciamo il concetto di invarianza dello stimatore di massima verosimiglianza

Esempio - Poisson

Se $\hat{\theta}$ è lo stimatore di massima verosimiglianza di θ, allora per ogni funzione $\tau(\theta)$, lo stimatore di massima verosimiglianza di $\tau(\theta)$ è $\tau(\hat{\theta})$.

La dimostrazione non la facciamo ma se volete la trovate sul Casella Berger. Il senso del teorema è intuitivo facendo qualche esempio.

Massima verosimiglianza XIX

Prendiamo una campione iid da una Poisson con parametro λ e ricordiamo che la log-verosimiglianza è

$$\log L(\lambda|\mathbf{x}) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\log(x_i!)$$

e che viene massimizzata con $\lambda = \bar{x}$. Se invece di λ io usassi un altro parametro $\eta = \tau(\lambda) = \log(\lambda)$, la nuova verosimiglianza è

$$\log L(\eta|\mathbf{x}) = -n\exp(\eta) + \log(\exp(\eta))\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\log(x_i!)$$

La verosimiglianza deve massimizzarsi quando $\exp(\eta) = \bar{x}$, perchè solo in questo caso avremmo $\lambda = \bar{x}$, e quindi espressa in termini di η il massimo si raggiunge a $\hat{\eta} = \log(\bar{x})$ e lo stimatore di massima verosimiglianza è $W(\mathbf{X}) = \log(\bar{X})$, come detto dal teorema.

Massima verosimiglianza XX

Per chiudere questa parte sugli stimatori di massima verosimiglianza, parliamo del problema maggiore di questo approccio: non è sempre possibile trovare uno stimatore di massima verosimiglianza, per esempio vedete il prossimo esercizio

Esempio - Gamma

Supponiamo di avere un campione iid di dimensione n da una normale $G(a, b)$, con $\mathrm{E}(X) = a/b$ e $\mathrm{var}(X) = a/b^2$. Trovare gli stimatori di massima verosimiglianza di a e b.

Soluzione:

la log-verosimiglianza è data da

$$\log L(a, b|\mathbf{x}) = \sum_{i=1}^{n} (a \log(b) - \log(\Gamma(a)) + (a-1)\log(x_i) - bx_i)$$

in questo semplice caso è anche complicato calcolare la derivata prima visto che un parametro è parte della funzione $\Gamma()$. $\square$

Massima verosimiglianza XXI

L'impossibilità di calcolare lo stimatore di massima verosimiglianza è abbastanza comune in problemi complessi, ma fortunatamente possiamo trovare il massimo numericamente usando R, mathematica, Julia, python etc, vedere **"Chunk ottimizzazione della massima verosimiglianza"**.

Distorsione I

Abbiamo visto che ci sono diversi modi per trovare stimatori, e adesso si pone il problema di quale tra i possibili stimatori sia il migliore. Ricordiamo che uno stimatore è una variabile aleatoria, e quindi possiamo richiedere qualcosa circa la sua distribuzione. La cosa più immediate e ragionevole è di chiedere che, se $W(\mathbf{X})$ è uno stimatore di θ, allora

$$\mathrm{E}(W(\mathbf{X})) = \theta$$

In altre parole, il valore atteso dello stimatore, deve essere uguale al parametro di stimare. Questo ci garantisce che mediamente (ripetendo l'esperimento), i risultati saranno giusti.

Definizione - Disorsione

Dato uno stimatore $W(\mathbf{X})$ usato per stimare θ, la sua distorsione è definita come

$$\mathrm{E}(W(\mathbf{X})) - \theta$$

Se la distorsione è uguale a 0, quindi $\mathrm{E}(W(\mathbf{X})) = \theta$, per ogni $\theta \in \Theta$, lo stimatore si dice **non distorto** (unbiased)

Distorsione II

Esempio - Non distorsione della media campionaria

La media campionaria è sempre non distorta se $\mathbf{X}$ è composto da variabili iid

Soluzione:

Questo è abbastanza facile e si può vedere che

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \mathrm{E}(X_i).$$

$\square$

Distorsione III

Esempio - Stimatore varianza

Nelle stesse situazioni dell'esempio precedente, verificare la distorsione di $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$ usato come stimatore di $\mathrm{Var}(\mathbf{X}) = \sigma^2$

Soluzione:

Calcoliamo il valore atteso

$$\mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}\right) = \frac{\sum_{i=1}^{n}\mathrm{E}\left((X_i-\mu)^2\right)}{n} = \sigma^2$$

quindi è non distorto $\square$

Distorsione IV

Notate i due esempi sopra sono validi indipendentemente dalla distribuzione di $\mathbf{X}$, abbiamo chiesto solo che fossero iid. Vediamo cosa succede se usiamo $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$ per stimare $\mathrm{Var}(\mathbf{X}) = \sigma^2$

Esempio - Stimatore varianza II

Nelle stesse situazioni dell'esempio precedente, verificare la distorsione di $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$ usato come stimatore di $\mathrm{Var}(\mathbf{X}) = \sigma^2$

Soluzione:

Distorsione V

Calcoliamo il valore atteso

$$\mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2}{n}\right) =$$

$$\mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \bar{X})^2 - 2(\mu - \bar{X})\sum_{i=1}^{n}(\mu - X_i)}{n}\right) =$$

$$\mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \bar{X})^2}{n}\right) = \frac{1}{n}\left(\sum_{i=1}^{n}\mathrm{E}((X_i - \mu)^2) - n\mathrm{E}((\mu - \bar{X})^2)\right)$$

e ricordando che Var$(\bar{X})\sigma^2/n$, allora

$$\mathrm{E}(\hat{\sigma}^2) = \frac{1}{n}\left(n\sigma^2 - n\sigma^2/n\right) = \frac{(n-1)}{n}\sigma^2$$

Quindi lo stimatore è distorto. $\square$

Distorsione VI

L'esempio precedente chiarisce perchè si usa come stimatore per la varianza

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

che visto che può essere scritto come

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2$$

il suo valore atteso è

$$\mathrm{E}(S^2) = \frac{n}{n-1}\mathrm{E}(\hat{\sigma}^2) = \frac{n}{n-1}\frac{(n-1)}{n}\sigma^2 = \sigma^2$$

che è quindi non distorto.

Errore quadratico medio (MSE) I

Sebbene la distorsione sia una proprietà desiderabile, non è la sola di cui dovremmo tener conto. Possiamo per esempio richiedere che la distanza tra $W(\mathbf{X})$ e θ sia la più piccola possibile, mediamente. Ci sono vari modi per definire questa distanza, ma il più usato è l'**MSE** (mean squared error - errore quadratico media)

$$\mathrm{E}(W(\mathbf{X})-\theta)^2$$

Invece del quadrato si potrebbe usare il valore assoluto, e epr esempio calcolare $\mathrm{E}|W(\mathbf{X})-\theta|$, ma l'MSE ha almeno due proprietà interessanti: la prima è che è matematicamente facile da trattere, mentre la seconda è che si può scrivere in un modo facilmente interpretabile

$$\mathrm{E}(W(\mathbf{X})-\theta)^2 = \mathrm{E}(W(\mathbf{X})-\mathrm{E}(W(\mathbf{X}))+\mathrm{E}(W(\mathbf{X}))-\theta)^2 =$$

$$\mathrm{E}(W(\mathbf{X})-\mathrm{E}(W(\mathbf{X})))^2+(\mathrm{E}(W(\mathbf{X}))-\theta)^2+2(\mathrm{E}(W(\mathbf{X}))-\theta)\mathrm{E}(W(\mathbf{X})-\mathrm{E}(W(\mathbf{X})))=$$

$$\mathrm{Var}(W(\mathbf{X}))+\mathrm{Bias}^2$$

Errore quadratico medio (MSE) II

Quindi l'MSE è composto da un termine che ci dice la distorsione dello stimatore, e uno che ci dice quanto è variabile il suo valore. Se lo stimatore è non distorto abbiamo

$$\mathrm{MSE}(W(\mathbf{X})) = \mathrm{Var}(W(\mathbf{X}))$$

L'MSE ci dice anche che per un dato θ, l'MSE di uno stimatore distorto potrebbe esser minore di uno non distorto. Per esempio anche se $\hat{\sigma}^2$ dell'esercizio precedente è distorto e S^2 non lo è, si può dimostrare che l'MSE di $\hat{\sigma}^2$ è minore.

Supponiamo di avere due stimatori $W_1(\mathbf{X})$ e $W_2(\mathbf{X})$, entrambi non distorti, e di aver determinato che $\mathrm{MSE}(W_1(\mathbf{X})) < \mathrm{MSE}(W_2(\mathbf{X}))$. Ci chiediamo adesso se è possibile trovare uno stimatore migliore di $\mathrm{MSE}(W_1(\mathbf{X}))$. Per esempio, lo stimatore

$$\mathrm{MSE}(W_3(\mathbf{X})) = a\mathrm{MSE}(W_1(\mathbf{X})) + (1-a)\mathrm{MSE}(W_2(\mathbf{X}))$$

con $a \in [0,1]$, che si dimostra essere non distorto. E anche se questo non fosse migliore, come facciamo a sapere se possiamo migliorare lo stimatore che abbiamo?

Errore quadratico medio (MSE) III

Definizione - best unbiased estimator

Uno stimatore $W^*(\mathbf{X})$ è il migliore stimatore non distorto di θ se $\mathrm{E}(W^*(\mathbf{X})) = \theta$ per tutti i θ e se per ogni altro stimatore non distorto $W(\mathbf{X})$ di θ abbiamo

$$\mathrm{Var}(W^*(\mathbf{X})) \leq \mathrm{Var}(W(\mathbf{X}))$$

Definiamo come

$$C_{\tau(\theta)} = \{W(\mathbf{X}) : \mathrm{E}(W(\mathbf{X})) = \tau(\theta)\}$$

per ogni $\theta \in \Theta$, la classe degli stimatore con la stessa distorsione $\tau(\theta) - \theta$. Allora, il confronto tra stimatori in questa classe si basa solamente sulla varianza degli stimatori e la ricerca del miglior stimatore in questa classe diventa la ricerca dello stimatore con varianza minima:

$$\mathsf{MSE}(W_1(\mathbf{X})) - \mathsf{MSE}(W_2(\mathbf{X})) = \mathrm{Var}(W_1(\mathbf{X})) - \mathrm{Var}(W_2(\mathbf{X}))$$

con $W_1(\mathbf{X}), W_2(\mathbf{X}) \in C_{\tau(\theta)}$

Teorema di Cramer Rao I

Il teorema di Cramer-rao è un risultato molto utile che per mette di trovare un limite inferiore per la varianze di uno stimatore $W(\mathbf{X})$ che abbia una specifica distorsione $\tau(\theta)$. Quindi, se troviamo uno stimatore che raggiunge il limite di Cramer-Rao, sappiamo che è il mogliore in quella classe.

Teorema di Cramer Rao II

Teorema - Disuguaglianze di Cramer-Rao

Assumiamo che $X_1, \ldots, X_n$ sia un campione con pmf o pdf uguale a $f(\mathbf{x}|\theta)$, con θ univariato, e sia $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ uno stimatore di θ con

$$\frac{dE(W(\mathbf{X}))}{d\theta} = \frac{d}{d\theta}\int_{\mathbf{x}\in\mathcal{X}} W(\mathbf{x})f(\mathbf{x}|\theta)d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} \frac{d}{d\theta}W(\mathbf{x})f(\mathbf{x}|\theta)d\mathbf{x}$$

nel caso $\mathbf{x}$ sia continua o

$$\frac{dE(W(\mathbf{X}))}{d\theta} = \frac{d}{d\theta}\sum_{\mathbf{x}\in\mathcal{X}} W(\mathbf{X})f(\mathbf{x}|\theta) = \sum_{\mathbf{x}\in\mathcal{X}} \frac{d}{d\theta}W(\mathbf{X})f(\mathbf{x}|\theta)$$

nel caso $\mathbf{x}$ sia discreto (cioè possiamo cambiare l'ordine di integrazione e derivazione), e inoltre

$$\mathrm{Var}(W(\mathbf{X})) < \infty$$

allora

$$\mathrm{Var}(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}E(W(\mathbf{X}))\right)^2}{\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right)}$$

Teorema di Cramer Rao III

Dimostrazione:

La dimostrazione la facciamo solo nel caso assolutamente continuo. Usiamo un semplice risultato

$$\mathrm{Cov}^2(Z,Y) = \mathrm{Var}(Z)\mathrm{var}(Y)\mathrm{Cor}^2(Z,Y) \leq \mathrm{Var}(Z)\mathrm{var}(Y)$$

e quindi

$$\mathrm{Var}(Z) \geq \frac{\mathrm{Cov}^2(Z,Y)}{\mathrm{Var}(Y)} \tag{1}$$

per adesso mettiamo da parte questo risultato e concentriamoci su altri aspetti.

Abbiamo che

$$\frac{dE(W(\mathbf{X}))}{d\theta} = \int_{\mathbf{x}\in\mathcal{X}} W(\mathbf{x})\frac{d}{d\theta}f(\mathbf{x}|\theta)d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} W(\mathbf{x})\frac{\frac{d}{d\theta}f(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)}f(\mathbf{x}|\theta)d\mathbf{x} =$$

$$\mathrm{E}\left(W(\mathbf{X})\frac{\frac{d}{d\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right)$$

Teorema di Cramer Rao IV

visto che abbiamo

$$\frac{d}{d\theta}\log f(\mathbf{X}|\theta) = \frac{\frac{d}{d\theta} f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}$$

allora

$$\frac{dE(W(\mathbf{X}))}{d\theta} = \mathrm{E}\left(W(\mathbf{X})\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)$$

definiamo $Z = W(\mathbf{X})$ e $Y = \frac{d}{d\theta}\log f(\mathbf{x}|\theta)$ e vediamo che dalla formula sopra se assumiamo $Z = 1$ allora

$$0 = \frac{dE(1)}{d\theta} = \mathrm{E}\left(Y\right) = \mathrm{E}\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)$$

Da cui abbiamo che

$$\mathrm{Cov}(Z, Y) = \mathrm{E}(ZY) = \mathrm{E}\left(W(\mathbf{X})\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right) = \frac{dE(W(\mathbf{X}))}{d\theta}$$

Teorema di Cramer Rao V

e abbiamo quindi trovato il numeratore di (1). Per il denominatore dobbiamo calcolare

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}^2(Y) = \mathrm{E}(Y^2) = \mathrm{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right)$$

che sostituendo in (1) dimostra la tesi. $\square$

Sebbene l'ipotesi di poter scambiare il segno di integrata e derivazione sia inusuale, questo è sempre possibile se l'insieme

$$S = \{\mathbf{x} : f(\mathbf{x}) > 0\}$$

non dipende da θ, il che è vero per la maggior parte delle distribuzioni che vengono usate in questo corso, ad eccezione dell'uniforme.

Teorema di Cramer Rao VI

Alcune punti importanti

- Notate come il numeratore dipende solo dalla distorsione, mentre il denominatore solo dalla densità (o pmf);
- Il denominatore del limite di CR $\mathcal{I}_n = \mathsf{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right)$ è chiamata informazione di Fisher, visto che nel caso in cui lo stimatore sia non distorto abbiamo che $\mathrm{Var}(W(\mathbf{X})) \geq \mathcal{I}_n^{-1}$;
- Il teorema si può estendere anche al caso in cui θ sia multidimensionale, e lo stimatore sia non distorto. Abbiamo in questo caso che il limite è $\mathrm{Var}(W(\mathbf{X})) \geq \mathcal{I}_n^{-1}$ con

$$[\mathcal{I}_n]_{ij} = \mathrm{E}\left(\frac{d}{d\theta_i}\log f(\mathbf{X}|\theta)\frac{d}{d\theta_j}\log f(\mathbf{X}|\theta)\right)$$

Teorema di Cramer Rao VII

- se n è abbastanza grande, allora (non lo dimostriamo) lo stimatore di massima verosimiglianza $\hat{\boldsymbol{\theta}}_n$ ha la seguente distribuzione

$$\hat{\boldsymbol{\theta}}_n \sim N(\boldsymbol{\theta}, \mathcal{I}_n^{-1})$$

  cioè è (asintoticamente) non distorto, e a varianza minima (si dice efficiente).

Introduciamo un corollario che può essere utile

Teorema - Disuguaglianze di Cramer-Rao per v.a.iid

Se all'ipotesi del teorema di Cramer-Rao aggiungiamo anche che il campione ha componenti iid. abbiamo che

$$\mathrm{Var}(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta} E(W(\mathbf{X}))\right)^2}{n\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right)}$$

dove con $f(X|\theta)$ si intende la pmf o pdf di una generica componente di $\mathbf{X}$, visto che sono iid non serve specificare quale

Teorema di Cramer Rao VIII

Dimostrazione:

Il numeratore è invariato rispetto al teorema precedente, e dobbiamo solo dimostrare come cambia il denominatore.

$$\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right) = \mathsf{E}\left(\left(\sum_{i=1}^{n}\frac{d}{d\theta}\log f(X_i|\theta)\right)^2\right) =$$

$$\sum_{i=1}^{n}\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(X_i|\theta)\right)^2\right) + \sum_{i\neq j}^{n}\mathsf{E}\left(\frac{d}{d\theta}\log f(X_i|\theta)\frac{d}{d\theta}\log f(X_j|\theta)\right)$$

Siccome i dati sono indipendenti abbiamo che

$$\mathsf{E}\left(\frac{d}{d\theta}\log f(X_i|\theta)\frac{d}{d\theta}\log f(X_j|\theta)\right) = \mathsf{E}\left(\frac{d}{d\theta}\log f(X_i|\theta)\right)\mathsf{E}\left(\frac{d}{d\theta}\log f(X_j|\theta)\right) = 0$$

Teorema di Cramer Rao IX

perchè nel primo teorema abbiamo dimostrato che $\mathsf{E}\left(\frac{d}{d\theta}\log f(X_i|\theta)\right) = 0$ nel caso in cui la variabile fosse un vettore e deve valere anche per il caso in cui è uno scalare. Inoltre, visto che le variabili sono iid allora

$$\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right) = \sum_{i=1}^{n}\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(X_i|\theta)\right)^2\right) = n\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right)$$

dimostrando la tesi. □

Lo stesso risultato vale anche per il caso in cui θ sia multivariato dove è $\mathrm{Var}(W(\mathbf{X})) \geq \mathcal{I}_n^{-1} = n\mathcal{I}^{-1}$ con

$$[\mathcal{I}]_{ij} = \mathrm{E}\left(\frac{d}{d\theta_i}\log f(X|\theta)\frac{d}{d\theta_j}\log f(X|\theta)\right)$$

Teorema di Cramer Rao X

Lemma - Disuguaglianze di Cramer-Rao

Se all'ipotesi del teorema di Cramer-Rao aggiungiamo anche che

$$\frac{d}{d\theta}\mathrm{E}\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right) = \int \frac{d}{d\theta}\left(\frac{d}{d\theta}\left(\log f(\mathbf{X}|\theta)\right) f(\mathbf{X}|\theta)\right) d\mathbf{x}$$

allora

$$\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right) = -\mathsf{E}\left(\frac{d^2}{d\theta^2}\log f(\mathbf{X}|\theta)\right)$$

Dimostrazione:

Abbiamo che

$$\mathrm{E}\left(\frac{d^2}{d\theta^2}\log f(\mathbf{X}|\theta)\right) = \mathrm{E}\left(\frac{d}{d\theta}\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right) = \mathrm{E}\left(\frac{d}{d\theta}\left(\frac{\frac{d}{d\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right)\right) =$$

Teorema di Cramer Rao XI

$$\mathrm{E}\left(\frac{f(\mathbf{X}|\theta)\frac{d^2}{d\theta^2}f(\mathbf{X}|\theta)-\left(\frac{d}{d\theta}f(\mathbf{X}|\theta)\right)^2}{\left(f(\mathbf{X}|\theta)\right)^2}\right)=\mathrm{E}\left(\frac{\frac{d^2}{d\theta^2}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right)-\mathrm{E}\left(\left(\frac{\frac{d}{d\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right)^2\right)$$

Prendiamo il primo termine

$$\mathrm{E}\left(\frac{\frac{d^2}{d\theta^2}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right)=\int\frac{\frac{d^2}{d\theta^2}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}f(\mathbf{X}|\theta)d\mathbf{x}=\int\frac{d}{d\theta}\left(\frac{d}{d\theta}f(\mathbf{X}|\theta)\right)d\mathbf{x}$$

e visto che

$$\frac{d}{d\theta}f(\mathbf{X}|\theta)=\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)f(\mathbf{X}|\theta)$$

allora per ipotesi

$$\int\frac{d}{d\theta}\left(\frac{d}{d\theta}f(\mathbf{X}|\theta)\right)d\mathbf{x}=\int\frac{d}{d\theta}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)f(\mathbf{X}|\theta)\right)d\mathbf{x}=$$

$$\frac{d}{d\theta}\mathrm{E}\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)=0$$

Teorema di Cramer Rao XII

Questo è uguale a 0 perchè lo abbiamo dimostrato in Cramer Rao $\left(\mathrm{E}\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right) = 0\right)$ allora

$$\mathrm{E}\left(\frac{d^2}{d\theta^2}\log f(\mathbf{X}|\theta)\right) = -\mathrm{E}\left(\left(\frac{\frac{d}{d\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right)^2\right)$$

ma per le proprietà del logaritmo abbiamo che

$$\frac{\frac{d}{d\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)} = \frac{d}{d\theta}\log f(\mathbf{X}|\theta)$$

e quindi

$$\mathrm{E}\left(\frac{d^2}{d\theta^2}\log f(\mathbf{X}|\theta)\right) = -\mathrm{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right)$$

dimostrando la tesi $\square$

Teorema di Cramer Rao XIII

L'ipotesi alla base di questo lemma sono soddisfatte per tutte le distribuzioni standard. Questo lemma è utile perchè è in generale più facile calcolare il valore atteso della derivata seconda piuttosto che il valore atteso del quadrato della derivata prima
Inoltre il lemma e il teorema precedente possono essere usati insieme, se le ipotesi sono soddisfatte, per poter scrivere

$$\mathsf{E}\left(\left(\frac{d}{d\theta}\log f(\mathbf{X}|\theta)\right)^2\right) = -n\mathsf{E}\left(\frac{d^2}{d\theta^2}\log f(X|\theta)\right)$$

Esempio - Poisson

Calcolare il limite di Cramer Rao per uno stimatore del parametro di una $P(\lambda)$, con $E(W(\mathbf{X})) = \frac{\lambda}{n}$ e campione di variabili di dimensione n iid.

Teorema di Cramer Rao XIV

Soluzione:

Calcoliamo il numeratore di CR:

$$\left(\frac{d}{d\lambda}E(W(\mathbf{X}))\right)^2 = \left(\frac{d}{d\lambda}\frac{\lambda}{n}\right)^2 = \frac{1}{n^2}$$

Calcoliamo il denominatore usando il lemma

$$-n\mathsf{E}\left(\frac{d^2}{d\lambda^2}\log f(X|\lambda)\right) = -n\mathsf{E}\left(\frac{d^2}{d\lambda^2}\left(-\lambda + X\log\lambda - \log(X!)\right)\right)$$

da cui abbiamo

$$\frac{d^2}{d\theta^2}\left(-\lambda + X\log\lambda - log(X!)\right) = -\frac{X}{\lambda^2}$$

e quindi

$$-n\mathsf{E}\left(\frac{d^2}{d\lambda^2}\log f(X|\theta)\right) = \frac{n}{\lambda}$$

Teorema di Cramer Rao XV

e il limite è $\frac{\lambda}{n^3}$ $\square$

Esempio - Poisson

Ipotizziamo di avere un campione di variabili iid di dimensione n da una $P(\lambda)$ e di vole stimare λ con la media campionaria. DImostrare che raggiunge il limite di Cramer-Rao

Soluzione:

Abbiamo detto che il denominatore di cramer rao non dipenden dalla distorsione ma solo dalla congiunta, quindi anche in questo caso è

$$\frac{n}{\lambda}$$

Teorema di Cramer Rao XVI

Abbiamo dimostrato che la media campionaria è non distorto per la media di variabili iid, quindi abbiamo

$$\mathrm{E}(\bar{X}) = \lambda$$

e il numeratore di CR è

$$\left(\frac{d}{d\lambda}E(W(\mathbf{X}))\right)^2 = \left(\frac{d}{d\lambda}\lambda\right)^2 = 1$$

e il limite di CR è

$$\frac{\lambda}{n}$$

Abbiamo anche visto precedentemente che la varianza delle media campionaria è $\mathrm{Var}(X)/n$ che in questo caso è

$$\frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}$$

e quindi la media campionaria è il miglior stimatore non distorto di λ. $\square$

Teorema di Cramer Rao XVII

Per concludere vediamo un altro modo per connettere gli stimatori e le statistiche sufficienti

Teorema - Rao-Blackwell

Se W è uno stimatore non distorto di θ e T è una statistica sufficiente per θ, allora se definiamo $\phi(T) = \mathrm{E}(W|T)$ come nuovo stimatore, questo sarà ancora non distorto e avrà varianza minore o uguale di W.

Dimostrazione:

La dimostrazione segue dal calcolo delle medie e varianze iterate visto che

$$\mathrm{E}(W) = \mathrm{E}(\mathrm{E}(W|T)) = \mathrm{E}(\phi(T)) = \theta$$

Teorema di Cramer Rao XVIII

quindi $\phi(T)$ è non distorto e

$$\mathrm{Var}(W) = \mathrm{Var}(\mathrm{E}(W|T)) + \mathrm{E}(\mathrm{Var}(W|T)) =$$

$$\mathrm{Var}(\phi(T)) + \mathrm{E}(\mathrm{Var}(W|T)) \geq \mathrm{Var}(\phi(T))$$

Sebbene non abbiamo mai usato la sufficienza, questa è richiesta altrimenti

$$\phi(T) = \mathrm{E}(W|T) = \int W(\mathbf{x}) f(\mathbf{x}|T(\mathbf{x})) d\mathbf{x}$$

dipenderà da θ e quindi non può essere uno stimatore visto che per definizione di statistiche sufficienti se $f(\mathbf{x}|T(\mathbf{x}))$ non dipende da θ allora $T(\mathbf{X})$ è sufficiente. $\square$