Introduzione all'inferenza statistica, III lezione

Carla Rampichini

Dipartimento di Statistica "Giuseppe Parenti" - Firenze - Italia

carla@ds.unifi.it - www.ds.unifi.it/rampi/

Funzione di verosimiglianza

Sia $p(\mathbf{X} \mid \theta)$ la funzione di distribuzione congiunta del campione casuale $\mathbf{X}$ di dimensione n, con X_i i.i.d. e $X \sim f(X;\theta)$:

- *prima* di estrarre il campione, $p(\mathbf{X} \mid \theta) = P_\theta(\mathbf{X} = \mathbf{x})$ (nel caso discreto)
- *dopo* aver osservato $\mathbf{X} = \mathbf{x}$, $p(\mathbf{X} \mid \theta) = \mathcal{L}(\theta \mid \mathbf{x})$

$$\mathcal{L}(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta)$$

è detta funzione di verosimiglianza.

▶ Per $\mathcal{L}(\theta \mid \mathbf{x}) > 0$ si considera di solito la funzione di log-verosimiglianza:

$$\log \mathcal{L}(\theta \mid \mathbf{x}) = \sum_{i=1}^{n} f(x_i;\theta)$$

Principio di verosimiglianza

$\mathcal{L}(\theta \mid \mathbf{x})$ sintetizza le informazioni in $\mathbf{X}$ dato il modello $X \sim f(X;\theta)$ e indica la plausibilità dei valori $\theta \epsilon \Theta$ in base all'evidenza empirica $\mathbf{X} = \mathbf{x}$. Le inferenze derivate da $\mathcal{L}(\theta \mid \mathbf{x})$ sono conseguenza dei due principi generali seguenti:

- **PRINCIPIO DEBOLE DI VEROSIMIGLIANZA**

Data $X \sim f(X;\theta)$, se $\mathbf{X} = \mathbf{x}$ e $\mathbf{X} = \mathbf{y}$ sono tali che $\mathcal{L}(\theta \mid \mathbf{x}) \propto \mathcal{L}(\theta \mid \mathbf{y})$, allora $\mathbf{x}$ e $\mathbf{y}$ devono fornire la **stessa inferenza** su θ.

- **PRINCIPIO FORTE DI VEROSIMIGLIANZA**

Dati $X \sim F(X;\theta)$, $\mathbf{X} = \mathbf{x}$ con $\mathcal{L}_F(\theta \mid \mathbf{x})$ e $Y \sim G(Y;\theta)$, $\mathbf{Y} = \mathbf{y}$ con $\mathcal{L}_G(\theta \mid \mathbf{y})$ se $\mathcal{L}_F(\theta \mid \mathbf{x}) \propto \mathcal{L}_G(\theta \mid \mathbf{y})$, allora le inferenze su θ devono essere le **stesse** per i due campioni.

★ La maggior parte dei metodi inferenziali aderisce almeno al principio debole di verosimiglianza.

Si può derivare il principio di verosimiglianza?

Si può mostrare che il principo formale di verosimiglianza consegue necessariamente da altri due principi generali:

- il **principio di sufficienza formale**: si consideri l'*esperimento* $E = (\mathbf{X}, \theta, \{f(\mathbf{x} \mid \theta\})$ e si supponga che $T(\mathbf{X})$ sia una statistica sufficiente per θ. Se $\mathbf{x}$ e $\mathbf{y}$ sono due punti campionari che soddisfano $T(\mathbf{x}) = T(\mathbf{y})$, allora $Ev(E, \mathbf{x}) = Ev(E, \mathbf{y})$.
  Ev è l'*evidenza* su θ derivante dall'esperimento E e dal punto campionario $\mathbf{x}$.
- il **principio di condizionalità**

Principio di condizionalità

Si considerino due esperimenti $E_1 = (\mathbf{X}_1, \theta, \{f(\mathbf{x_1} \mid \theta\})$ e $E_2 = (\mathbf{X}_2, \theta, \{f(\mathbf{x_2} \mid \theta\})$ che possono avere in comune anche solo il parametro incognito θ. Si consideri l'*esperimento misto* seguente: (*i*) si osserva la v.c. J con $P(J = 1) = P(J = 2) = 1/2$ (indipendente da $\theta, \mathbf{x}_1, \mathbf{x}_2$) (*ii*) si effettua l'esperimento E_j. Formalmente si effettua l'esperimento $E^* = (\mathbf{X}^*, \theta, \{f^*(\mathbf{x}^* \mid \theta)\})$ dove $\mathbf{X}^* = (j, \mathbf{X})$ e $f^*(\mathbf{x}^* \mid \theta) = \frac{1}{2} f_j(\mathbf{x}_j \mid \theta)$ allora

$$Ev(E^*, (j, \mathbf{x}_j) = Ev(E_j, \mathbf{x}_j)$$

Il principio di condizionalità afferma che se si sceglie a caso tra i due esperimenti e si effettua l'esperimento scelto, ottenendo i dati $\mathbf{x}$, l'informazione su θ *dipende solo dell'esperimento effetuato*.

- si ha la stessa informazione che si sarebbe avuta decidendo fin dall'inizio (non a caso) di fare questo epserimento e si fossero osservati i dati $\mathbf{x}$.

Principio di verosimiglianza formale

Si considerino due esperimenti $E_1 = (\mathbf{X}_1, \theta, \{f(\mathbf{x_1} \mid \theta\})$ e $E_2 = (\mathbf{X}_2, \theta, \{f(\mathbf{x_2} \mid \theta\})$ per i quali si ha lo stesso parametro incognito θ. Siano $\mathbf{x}_1^*$ e $\mathbf{x}_2^*$ i punti campionari osservati rispettivamente da E_1 e E_2 tali che

$$\mathcal{L}(\theta \mid \mathbf{x}_2^*) = c\mathcal{L}(\theta \mid \mathbf{x}_1^*)$$

per tutti i θ e per qualche costante c, che può dipendere da $\mathbf{x}_1^*$ e $\mathbf{x}_2^*$ ma non da θ. Allora:

$$Ev(E_1, \mathbf{x}_1^*) = Ev(E_2, \mathbf{x}_2^*)$$

Il principio di verosimiglianza formale è diverso da quelli enunciati in precedenza, perchè riguarda 2 esperimenti diversi. Il principio debole di verosimiglianza può essere derivato dal principio formale di verosimiglianza considerando E_2 come una replica di E_1.

Conseguenze del Principio di verosimiglianza formale

Corollario al Principio di verosimiglianza formale
Se $E = (\mathbf{X}, \theta, \{f(\mathbf{x} \mid \theta\})$ è un esperimento, allora $Ev(E, \mathbf{x})$ dipende da E e da **x** solo attraverso $\mathcal{L}(\mathbf{x} \mid \theta)$.

Teorema di Birnbaum
Il Principio di verosimiglianza formale segue dal Principio di sufficienza formale e dal principio di condizionalità. Vale anche il contrario.

▶ Molte procedure di uso corrente violano il principio di verosimiglianza formale, il che comporta che violino sia il principio di sufficienza formale che il principio di condizionalità (per esempio analisi dei residui per validazione modello).

$\rightsquigarrow$ Prima di considerare il principio di sufficienza - e anche il principio di verosimiglianza - bisogna avere *fiducia nel modello*.

Funzione score

Data la log-verosimiglianza $\log \mathcal{L}(\theta \mid \mathbf{x}) = l(\theta \mid \mathbf{x}) = \sum_{i=1}^{n} f(x_i; \theta)$, si definisce funzione **SCORE** la derivata prima di $l(\theta \mid \mathbf{x})$:

$$S(\theta) = \frac{\partial l(\theta|\mathbf{x})}{\partial \theta} = \frac{\mathcal{L}'(\theta|\mathbf{x})}{\mathcal{L}(\theta|\mathbf{x})}$$

- $\mathbb{E}[S(\theta)] = 0$
- $V[S(\theta)] = \mathbb{E}[S(\theta)^2] = -\mathbb{E}[S'(\theta)] = I_n(\theta)$
- $I_n(\theta) = nI(\theta)$

★ $I_n(\theta)$ è detta **matrice di informazione** di Fisher:

$$I_n(\theta) = -n\mathbb{E}[\frac{\partial^2}{\partial\theta} l(\theta)] = n\mathbb{E}[\frac{\partial}{\partial\theta} l(\theta)]^2$$

Stima puntuale

- Se una popolazione è descritta da una *pdf* o *pmf* $f(x \mid \theta)$, conoscere θ significa conoscere completamente la popolazione!
- è quindi naturale cercare un metodo che consenta di trovare uno **stimatore** di θ
- spesso θ ha un significato suo (p.e. la media) e quindi si è direttamente interessati a stimare θ
- uno **stimatore puntuale** di θ è una qualunque funzione $W(X_1, \ldots, X_n)$ del campione, ossia ogni statistica è uno stimatore puntuale.
- una **stima** è la realizzazione, cioè il valore, dello stimatore che si ottiene quando il campione viene estratto.

Scelta di uno stimtore

- è utile conoscere qualche **metodo** per trovare stimatori
  - metodo dei momenti
  - massima verosimiglianza
  - stimatori bayesiani
  - algoritmo EM
- sarà necessario valutare le **proprietà** degli stimatori
  - errore quadratico medio
  - stimatori corretti
  - efficienza
  - proprietà asintotiche

Metodo dei momenti

- Sia $X_1, \ldots, X_n$ un campione da una popolazione $X \sim f(X \mid \theta_1, \ldots, \theta_K)$
- gli stimatori MM si trovano eguagliando i primi K momenti campionari ai corrispondenti momenti della popolazione

$$m_k = \mu_k(\theta_1, \ldots, \theta_K),\ k = 1, \ldots, K$$

e risolvendo il sistema di equazioni simultanee rispetto ai momenti campionari $m_1, \ldots, m_K$

  - il MM è uno dei metodi più *antichi* (Pearson fine '800) per la ricerca di stimatori puntuali
  - spesso ci sono stimatori migliori però ...
  - ... è un metodo semplice e fornisce sempre una qualche stima $\Rightarrow$ buon punto di partenza
  - esempio: stimatori MM per popolazione normale con media e varianza incognite

Come scegliere uno stimatore?

La sufficienza non consente di scegliere un unico stimatore

- **Criteri di scelta degli stimatori**

  - proprietà finite
    - correttezza
    - efficienza
  - proprietà asintotiche
    - correttezza asintotica
    - consistenza
    - efficienza asintotica

Correttezza

Uno stimatore $T(\mathbf{X})$ si dice **non distorto** (o *corretto*) per θ se:

$$\mathbb{E}(T(\mathbf{X})) = \theta$$

La **distorsione** (*bias*) di uno stimatore è definita da:

$$b(T(\mathbf{X})) = \mathbb{E}(T(\mathbf{X})) - \theta$$

$\rightsquigarrow$ Se uno stimatore è *non distorto* la sua distribuzione campionaria è *centrata* su θ.

Esempi:

- media campionaria
- varianza campionaria corretta

Caratteristiche degli stimatori corretti

- in una data situazione, possono esistere *più stimatori corretti* per θ;
- se $T(\mathbf{X})$ è corretto per θ, in generale $f(T(\mathbf{X}))$ *non* è corretto per $f(\theta)$;
- uno stimatore distorto con distorsione nota che non dipende da θ è equivalente ad uno stimatore corretto, perchè è facile compensare la distorsione;
- nell'approccio classico, la correttezza è spesso utilizzata per limitare la classe degli stimatori all'interno della quale cercare uno stimatore *ottimo*. P.e. nel caso del metodo dei *minimi quadrati* (least squares) o delle *statistiche d'ordine*.

Errore quadratico medio

Il valore atteso della v.c $(T(\mathbf{X}) - \theta)^2$ si definisce **Errore Quadratico Medio** (EQM) dello stimatore $T(\mathbf{X})$ di θ:

$$EQM = \mathbb{E}(T(\mathbf{X}) - \theta)^2$$

- EQM tiene conto sia della *varianza* che della *distorsione* di $T(\mathbf{X})$:

$$\mathbb{E}(T(\mathbf{X}) - \theta)^2 = Var(T(\mathbf{X})) + b^2(T(\mathbf{X}))$$

- $T_1(\mathbf{X})$ è *più efficiente* di $T_2(\mathbf{X})$ se

$$EQM(T_1(\mathbf{X})) < EQM(T_2(\mathbf{X})).$$

Come scegliere tra stimatori?

- **EFFICIENZA RELATIVA**

$$eff(T_1/T_2) = \frac{EQM(T_2(\mathbf{X}))}{EQM(T_1(\mathbf{X}))}.$$

Se $\mathbb{E}(T_1(\mathbf{X})) = \mathbb{E}(T_2(\mathbf{X})) = \theta$,

$$eff(T_1/T_2) = \frac{V(T_2(\mathbf{X}))}{V(T_1(\mathbf{X}))}$$

- **STIMATORE UMVU** E' possibile trovare uno stimatore non distorto a varianza minima (UMVUE) in base alla seguente diseguaglianza:

DISEGUAGLIANZA DI CRAMER-RAO Dato un campione casuale $\mathbf{X} = \mathbf{x}$, con X_i i.i.d e $X \sim F(x;\theta)$, per ogni $T(\mathbf{X})$ non distorto di θ vale la seguente diseguaglianza:

$$V(T(\mathbf{X})) \geq \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}$$

Teoremi rilevanti

TEOREMA Se esiste $T(\mathbf{X})$, non distorto, che raggiunge il limite di Cramèr-Rao allora esso è UNICO.

TEOREMA Condizione necessaria e sufficiente affinchè esista uno stimatore efficiente e non distorto per θ è che si abbia:

$$S(\theta) = \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta \mid \mathbf{X}) = I_n(\theta)[T(\mathbf{X}) - \theta]$$

Esempio: media campionaria per $X \sim N(\theta, 1)$

- La **famiglia esponenziale**
  - è l'unica ad ammettere stimatori sufficienti di dimensione pari a quella dello spazio parametrico
  - gli stimatori sono sufficienti minimali e funzioni della statistica $T(\mathbf{X}) = (\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i))$
  - Se esite $T(\mathbf{X})$ non distorto e efficiente per $\theta \Leftrightarrow F(X; \theta)$ appartiene alla famiglia esponenziale.