Lezione n.2: Il modello statistico

Roma, 3 marzo 2003

Brunero Liseo
Dipartimento di studi geoeconomici, linguistici, statistici e storici
per l'analisi regionale
Università di Roma "La Sapienza"
Rome, Italy
brunero.liseo@uniroma1.it
tel. 06-49766110

Ingredienti di m.s.

Un modello statistico è composto da

$$\mathcal{E} = (\mathcal{Z}, \mathcal{F}, \mathcal{P})$$

dove

- $\mathcal{Z}$ è l'insieme delle possibili osservazioni, in genere uno spazio misurabile, dotato di
- $\mathcal{F}$ una σ-algebra di sottoinsiemi
- $\mathcal{P}$ è una famiglia di distribuzioni di probabilità

In genere si scrive

$$\mathcal{P} = \{P_\theta(\cdot) : \theta \in \Theta\}$$

Θ si chiama **spazio dei parametri**; se $\Theta \subset \mathrm{R}^k$ il modello statistico si chiama **parametrico**

Ad ogni valore di $\theta \in \Theta$ corrisponde una diversa descrizione del fenomeno: noi assumiamo che

esista un valore $\theta^* \in \Theta$ che rappresenta il "vero valore" di θ.

Obiettivo: determinare il vero valore di θ osservando una realizzazione $z \in \mathcal{Z}$ con legge $P_\theta()$.

Non è necessario ma quasi sempre la realizzazione osservata è quella di n replicazioni di una stessa v.a.

$$X \sim P_\theta(\cdot)$$

con θ incognito.

Se per ciascuna osservazione noi assumiamo

$$(X, p(x \mid \theta), \theta \in \Theta)$$

dove la legge $p(x \mid \theta)$ può essere di tipo discreto o assolutamente continuo, il modello statistico associato al campione di n osservazioni sarà

$$\left(\mathcal{X}^{(n)}, \prod_{i=1}^{n} p(x_i \mid \theta), \theta \in \Theta \right)$$

Esempio

$X \sim Be(\theta)$

cioè X assume i valori 0 e 1 con prob. $1-\theta$ e θ, oppure

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}, \qquad x = 0, 1$$

In tal caso

- $\mathcal{X}^{(n)}$ è l'insieme di tutte le 2^n n-ple composte da 0 e 1
- $\prod_{i=1}^{n} p(x_i \mid \theta) = \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}$
- $\theta \in [0, 1]$

Esempio: errori di misurazione

$X = \theta + \epsilon_i, \quad i = 1, \cdots, n$

θ è la misura di un oggetto sottoposto a n diverse misurazioni indipendenti, $\epsilon \sim N(0, \sigma^2)$. Ne segue che

$$X_1, \cdots, X_n \text{i.i.d.} \sim N(\theta, \sigma^2)$$

e

$$p(x \mid \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\theta)^2, \quad x \in \mathrm{R}$$

In tal caso

- $\mathcal{X}^{(n)} = \mathrm{R}^n$
- $\prod_{i=1}^n p(x_i \mid \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_i^n (x_i - \theta)^2\}$
- $\theta \in \mathrm{R}, \sigma^2 > 0.$

Altri esempi

- Modello non parametrico
- Modello semi parametrico
- Campionamento inverso
- Dati censurati
- Campionamento non bernoulliano
  n palline estratte in blocco da un'urna che ne contiene $N\theta$ bianche e $N(1-\theta)$ nere: ogni estrazione vale 0 oppure 1

In quest'ultimo caso avremo

- $\mathcal{X}^{(n)}$: tutte le possibili n-ple di 0 e 1
-

$$p(x \mid \theta) = \frac{[N\theta]_{\sum x_i}[N(1-\theta)]_{n-\sum x_i}}{N_n}$$

Qui le osservazioni non sono indipendenti....nemmeno per θ fissato.

I problemi dell'inferenza

In un problema di inferenza ci sono due tipi di incertezza

- intorno al valore di $\theta \in \Theta$
- relativa al particolare $z \in \mathcal{Z}$ osservato

Nell'impostazione bayesiana ma non solo ci si concentra sull'incertezza intorno a θ in quanto l'incertezza intorno a z scompare una volta osservato $z_0 \in \mathcal{Z}$ (anche se non sappiamo da quale legge....)

- Problemi ipotetici: stima puntuale, verifica di ipotesi
- Problemi previsivi

Nel primo caso ci si concentra sull'inferenza su θ, sia attraverso una stima precisa (stimatore) sia attraverso un intervallo. Oppure si vuole verificare se il dato z_0 è compatibile con un dato sottoinsiemi di valori di Θ.

Nel secondo caso ci si concentra sulla previsione di futuri risultati sperimentali ovvero si cerca di rispondere alla seguente esigenza

$$P\left(X_{n+1} \in B \mid Z = z_0\right) = P\left(X_{n+1} \in B \mid X_1 = x_1, \cdots, X_n = x_n\right)$$

La teoria classica

Ripercorriamo la teoria classica attraverso un esempio di estrema semplicità .

Si osservano $X_1, \cdots, X_n$ i.i.d. $\sim Be(\theta)$

Si vuole stimare θ

Lo stimatore di massima verosimiglianza (SMV) e lo stimatore UMVUE è

$$\hat{\theta} = \frac{\sum X_i}{n} \neq \bar{X}$$

Tale stimatore è scelto perchè basato sulla statistica sufficiente $\sum X_i$ e perché non distorto cioè

$$\mathrm{E}_\theta\left(\bar{X}\right) = \theta, \qquad \forall \theta \in \Theta$$

Stesso esempio: diverso campionamento

Si osservano $X_1, X_2, \cdots$, i.i.d. $\sim Be(\theta)$ fino a quando non si arrivi a k successi.

Si vuole stimare θ

(Esercizio: scrivere il modello statistico)

Lo stimatore di massima verosimiglianza (SMV) è

$$\hat{\theta} = \frac{\sum X_i}{n} = \bar{X}$$

mentre lo stimatore UMVUE è , in questo caso,

$$\hat{\theta} = \frac{\sum X_i - 1}{n - 1} = \bar{X}$$

Il contrasto

In un laboratorio medico arriva **solo il risultato finale** dell'esperimento cioè

$$10 \text{ osservazioni}, \quad 4 \text{ successi}$$

Qual è la stima *giusta* per θ ?

Secondo la teoria classica, se il campionamento è diretto, $\hat{\theta} = 4/10 = 0.4$

Se il campionamento è inverso, $\hat{\theta} = 3/9 = 0.333$

La funzione di verosimiglianza invece, nei due casi, è

$$L_{dir}(\theta) = \theta^4(1-\theta)^6$$

$$L_{inv}(\theta) \propto \theta^3(1-\theta)^6\theta = L_{dir}(\theta)$$

e perciò coincidono anche le due stime di MV....

Esempio gaussiano

Siano $X_1, X_2 \sim N(\theta, 1)$ indipendenti per θ fissato. Vogliamo stimare θ utilizzando uno dei due stimatori

$$T_1(X_1, X_2) = \frac{X_1 + X_2}{2} \qquad T_2(X_1, X_2) = X_1$$

Il criterio classico di scelta è basato sull' EQM (errore quadratico medio):

$$\mathrm{E}_\theta\left((T - \theta)^2\right)$$

Calcoliamo

$$\mathrm{E}_\theta\left((T_1 - \theta)^2\right) = \frac{1}{2} \qquad \mathrm{E}_\theta\left((T_2 - \theta)^2\right) = 1$$

Dunque, prima di osservare i dati, si è propensi a scegliere T_1.

Effettuiamo ora l'esperimento e osserviamo $x_1 = 4, x_2 = 1$ ovvero

$$T_1 = 2.5, \qquad T_2 = 4$$

Sulla base del dato osservato, l'errore (condizionatamente a x) che si commette è allora

$$(\theta - T_1)^2 = (\theta - 2.5)^2, \qquad (\theta - T_2)^2 = (\theta - 4)^2$$

Si vede che, per $\theta > \frac{13}{4}$, è preferibile usare T_2. Il problema è che non conosciamo θ....

Dati uniformi

Siano $X_1, X_2, \cdots, X_n$ i.i.d. con legge uniforme tra $\theta - 1/2$ e $\theta + /2$. Si vuole costruire un intervallo di confidenza per θ.

Metodo classico Basato sulle statistiche d'ordine $X_{(1)}$ e $X_{(n)}$, sufficienti per θ. Più precisamente si determina la legge di

$$T = \frac{X_{(1)} + X_{(n)}}{2}$$

e si vede che

$$P\left(T + \frac{\alpha^{1/n} - 1}{2} < \theta < T - \frac{\alpha^{1/n} - 1}{2}\right) = 1 - \alpha$$

Dunque è possibile costruire un intervallo di confidenza per qualunque valore di α, che $100(1-\alpha)$ volte su 100, conterrà il valore di θ, qualunque esso sia.

Supponiamo che $n = 25$, $\alpha = 0.05$, $X_{(1)} = 3$ e $X_{(25)} = 3.96$. Si avrà allora

$$(3.424, 3.536)$$

Ma un ragionamento assolutamente deduttivo conduce a dire che, **certamente**,

$$X_{(n)} - \frac{1}{2} < \theta < X_{(1)} + \frac{1}{2}$$

che coi nostri dati diventa

$$(3.46, 3.50)$$

La semplice osservazione di un vincolo deterministico ci ha condotto ad un intervallo più preciso rispetto a quello costruito mediante la teoria classica

Un esempio di Savage

Consideriamo le tre seguenti situazioni:

[S1] Tizio sostiene di essere in grado di riconoscere se un brano musicale è stato scritto da Mozart oppure da Beethoveen dopo appena quattro note. Gli sottoponiamo allora gli incipit di dieci brani scelti a caso dal repertorio dei due autori e verifichiamo le sue capacità .

[S2] La signora Bianchi sostiene che bevendo una tazza di tè al latte, è in grado di stabilire se è stato versato prima il latte oppure il tè : anche in questo caso sottoponiamo la signora a un test di 10 prove.

[S3] Il signor Rossi sostiene di possedere capacità soprannaturali e di essere in grado di prevedere il risultato di un lancio di una moneta regolare; lo stesso, effettuiamo 10 prove sperimentali.

Formalmente le 3 situazioni non differiscono. In tutti i casi si hanno n v.a. $Be(\theta)$ che assumono il valore 1 con prob. (incognita) θ.

In tutti i casi si avrà come risultato un vettore di dati osservati (stringa di valori 0 e 1).

Supponiamo che in tutti gli esperimenti si abbia $k = 7$ successi su $n = 10$ prove.

Allora la funzione di verosimigianza nei tre casi sarà comunque la stessa, così come le stime puntuali del parametro incognito θ: valuteremo pari a 0.7 sia la probabilità dell'esperto di musica di riconoscere un brano sia la capacità del presunto sensitivo di prevedere il futuro. Allo stesso modo, l'incertezza relativa a tale stima, è espressa da

$$\mathrm{Var}\left(\hat{\theta}\right) = \hat{\theta}(1-\hat{\theta})/n = 0.21/10 = 0.021.$$

Analisi bayesiana

Assumiamo di essere in grado di determinare per il parametro θ una legge di probabilità *iniziale*. In altri termini, dotiamo Θ di una struttura probabilistica, con una σ-algebra $\mathcal{B}(\Theta)$ e su tale σ-algebra deponiamo una misura di probabilità H, con densità

$$h(\theta), \quad \theta \in \Theta.$$

Il nucleo dell'impostazione bayesiana è rappresentato dall'aggiornamento della legge H mediante l'osservazione sperimentale, ovvero la realizzazione z dell'esperimento $(\mathcal{Z}, \mathcal{F}, \mathcal{P})$

Tale aggiornamento avviene mediante applicazione del Teorema di Bayes nella sua versione continua.

Teorema di Bayes

Sotto condizioni molto generali sul modello statistico e sulla legge iniziale H

$$h(\theta \mid z) = \frac{h(\theta)p(z \mid \theta)}{\int_\Theta p(z \mid \theta)h(\theta)d\theta},$$

oppure

$$h(\theta \mid z) \propto h(\theta)p(z \mid \theta).$$

in quanto il denominatore è solo una costante, determinabile ex-post e spesso (ma non sempre!!) non importante da deteminare.