5. Stima parametrica

Indice

Consideriamo un campione aleatorio $X_1, \ldots, X_n$ estratto da una distribuzione F_θ che dipende da un vettore di parametri incogniti θ (e.g. una distribuzione di Poisson di cui non è noto il valore di λ). A differenza del calcolo delle probabilità, in cui è normale supporre che le distribuzioni in gioco siano note, in statistica il problema centrale è quello di dire qualcosa (ovvero, **fare inferenza**) sui parametri sconosciuti, a partire dai dati osservati.

Stimatori di massima verosimiglianza

Una qualunque statistica il cui scopo sia quello di dare una stima di un parametro θ si dice **stimatore** di θ (gli stimatori sono quindi variabili aleatorie). Il valore deterministico assunto da uno stimatore è detto invece **stima**.

Siano $X_1, \ldots, X_n$ variabili aleatorie, la cui distribuzione congiunta sia nota a meno di un parametro incognito θ. Un problema di interesse consiste quindi nello stimare θ usando i valori che vengono assunti da queste variabili aleatorie.

Esiste una particolare classe di stimatori, detti **stimatori di massima verosimiglianza**, che è spesso utilizzata per rispondere a questo tipo di problematiche. Denotiamo con $f(x_1, \ldots, x_n|\theta)$ la funzione di massa (o di densità) congiunta di $X_1, \ldots, X_n$ (a seconda che siano variabili aleatorie discrete o continue). Se interpretiamo $f(x_1, \ldots, x_n|\theta)$ come la **verosimiglianza** (o plausibilità) che si realizzi la $n-$upla di dati $x_1, \ldots, x_n$ quando θ è il vero valore assunto dal parametro, sembra ragionevole adottare come stima di θ quel valore che rende massima la verosimiglianza per i dati osservati. In altri termini, la stima di massima verosimiglianza $\hat{\theta}$ è definita come il valore di θ che rende massima $f(x_1, \ldots, x_n|\theta)$, quando i valori osservati sono $x_1, \ldots, x_n$.

Nota. La funzione $f(x_1, \ldots, x_n|\theta)$ è detta funzione di **likelihood**

Nel calcolare il valore di θ che massimizza f, conviene spesso utilizzare il fatto che le due funzioni $f(x_1, \ldots, x_n|\theta)$ e $\log\left[f(x_1, \ldots, x_n|\theta)\right]$ assumono il massimo in corrispondenza dello stesso valore di θ. Quindi è possibile calcolare θ anche massimizzando $\log\left[f(x_1, \ldots, x_n|\theta)\right]$

Nota. La funzione $\log[f(x_1, \ldots, x_n|\theta)]$ è detta funzione di **log-likelihood**

Intervalli di confidenza

Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione normale di media incognita μ e varianza nota σ^2 . $\bar{X} := \sum_i X_i/n$ è lo stimatore di massima verosimiglianza è per μ. Ciò non significa che possiamo aspettarci che la media campionaria sia esattamente uguale a μ, ma solo che le sarà "vicina". Perciò, rispetto ad uno stimatore puntuale, è a volte preferibile potere produrre un intervallo per il quale abbiamo in certo livello di fiducia (confidenza) che il parametro μ vi appartenga. Per ottenere un tale **intervallo di confidenza**, dobbiamo fare uso della distribuzione di probabilità dello stimatore puntuale.

Per calcolare un intervallo di confidenza al 95% per una distribuzione gaussiana, è fondamentale partire dalla media campionaria $\bar{x}$, che rappresenta la media dei dati osservati nel campione. Se la deviazione standard della popolazione σ è nota, si utilizza direttamente nel calcolo; altrimenti, si adopera la deviazione standard campionaria s come stima di σ. La dimensione del campione n indica il numero totale di osservazioni raccolte.

Se σ è nota, l'intervallo di confidenza si calcola utilizzando la distribuzione normale standard:

$$IC = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

In questo caso, z è il valore critico corrispondente al livello di confidenza desiderato. Per un livello di confidenza del 95%, z è pari a 1,96. Questo valore deriva dalle proprietà della distribuzione normale standard, in cui il 95% dell'area sotto la curva è compreso tra $-1,96$ e $+1,96$. Matematicamente, ciò si esprime come:

$$P(-1,96 \leq Z \leq 1,96) = 0,95$$

Questo significa che c'è una probabilità del 95% che una variabile casuale normale standardizzata Z assuma valori tra $-1,96$ e $+1,96$.

Se σ non è nota, si utilizza la distribuzione t di Student con $n-1$ gradi di libertà:

$$IC = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

Qui, t è il valore critico ottenuto dalla distribuzione t per il livello di confidenza desiderato e per $n-1$ gradi di libertà. La distribuzione t tiene conto della variabilità aggiuntiva introdotta dalla stima di σ tramite s, specialmente per campioni di piccola dimensione.

Il processo di calcolo dell'intervallo di confidenza comprende i seguenti passaggi:

Calcolo della media campionaria. $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$.

Calcolo dell'errore standard. In particolare:

- Se σ è noto: $SE = \frac{\sigma}{\sqrt{n}}$
- Se σ non è noto: $SE = \frac{s}{\sqrt{n}}$

Determinazione del valore critico. Per la distribuzione normale standard (quando σ è noto), $z = 1,96$ per un livello di confidenza del 95%.

Calcolo dei limiti dell'intervallo di confidenza. $IC = (\bar{x} - z \cdot SE, \quad \bar{x} + z \cdot SE)$

Per chiarire perché il valore critico z è 1,96 per un livello di confidenza del 95%, consideriamo la distribuzione normale standard, che ha media $\mu = 0$ e deviazione standard $\sigma = 1$. L'area totale sotto la curva è pari a 1, o al 100% della probabilità. Per un livello di confidenza del 95%, vogliamo trovare il valore di z tale che il 95% dell'area sia compreso tra $-z$ e $+z$. Poiché la distribuzione è simmetrica, l'area rimanente del 5% si divide equamente tra le due code, con il 2,5% a sinistra di $-z$ e il 2,5% a destra di $+z$.

Matematicamente, cerchiamo z tale che

$P(Z \leq z) = 0,975$

Consultando le tavole della distribuzione normale standard, troviamo che $z = 1,96$ soddisfa questa condizione.