Stima insiemistica

Consideriamo un modello statistico $(\Omega, \mathcal{A}, \{P^\theta\}_{\theta\in D})$ e, su di esso, un campione statistico di taglia N e legge $\{\mathcal{L}(\theta)\}_{\theta\in D}$. La teoria degli stimatori puntuali permette di approssimare il valore del parametro sconosciuto $\psi(\theta)$, introducendo un opportuno stimatore $T = t(X_1, \dots, X_N)$ ed utilizzando le informazioni raccolte $x_1, \dots x_N$, mediante il numero $t(x_1, \dots, x_N)$. Ad esempio, come abbiamo visto, se il campione segue una legge gaussiana la sua media empirica $\overline{X}$ è uno stimatore corretto della media sconosciuta μ, la quale quindi viene stimata dal numero $\frac{1}{N}\sum_{i=1}^{N} x_i$.

Spesso però non interessa tanto il valore di un parametro, quanto invece avere la sicurezza che esso si trovi al di sotto, o al di sopra, di una certa soglia critica. Per esempio, se il campione si riferisce alla temperatura di un certo componente delicato di un macchinario, che potrebbe usurarsi a temperature troppo alte, saremo interessati a sapere se la temperatura di esso si trovi al di sotto di un certo valore. In questi casi non ci occorre uno stimatore puntuale, ma piuttosto una procedura che determini un <u>insieme di fiducia</u>, cioè un insieme nel quale possiamo stimare che il nostro parametro appartenga.

<u>Definizione</u> Sia $(\Omega, \mathcal{A}, \{P^\theta\}_{\theta\in D})$ un modello statistico e sia $\psi: D \to E$, $E = \psi(D) \subseteq \mathbb{R}$. Un'applicazione $S: \Omega \to \mathcal{P}(E)$ si chiama

insieme aleatorio se vale

$$\{\psi(\theta)\in S\} := \{\omega\in\Omega : \psi(\theta)\in S(\omega)\} \in \mathcal{A},$$

cioè è un evento. In tal caso, la funzione

$$\theta \longmapsto P^{\theta}(\psi(\theta)\in S)$$

si chiama curva di fiducia di S.

Esempio: Se T e T' sono due statistiche su $(\Omega, \mathcal{A}, \{P^{\theta}\}_{\theta\in D})$, allora l'applicazione

$$S(\omega) = [T(\omega), T'(\omega)]$$

è un insieme aleatorio: infatti

$$\{\psi(\theta)\in S\} = \{T \le \psi(\theta) \le T'\} = \{\psi(\theta)\le T'\} \cap \{\psi(\theta) < T\}^c,$$

e questi ultimi due insiemi sono eventi.

Similmente, le applicazioni

$$S_+(\omega) = [T(\omega), +\infty[, \quad S_-(\omega) =]-\infty, T(\omega)],$$

e, per $\delta>0$,

$$S(\omega) = [T(\omega)-\delta, T(\omega)+\delta]$$

sono insiemi aleatori.

Definizione Sia $\alpha\in]0,1[$, sia S un insieme aleatorio. Si dice che S è un insieme di fiducia di livello $1-\alpha$ per $\psi(\theta)$, se per ogni $\theta\in D$ si ha

$$P^{\theta}(\psi(\theta)\in S) \ge 1-\alpha.$$

Se S è un intervallo, diremo che S è un intervallo di fiducia; esso sarà bilatero se è limitato, unilatero se è illimitato.

Dunque, un intervallo di fiducia di livello $1-\alpha$ per $\psi(\theta)$ è un insieme aleatorio la cui curva di fiducia sta sempre al di sopra del livello $1-\alpha$.

In pratica, assegnato un campione statistico di taglia N, se $T = t(X_1, \dots X_N)$ è uno stimatore del parametro sconosciuto $\psi(\theta)$, e vogliamo trovare un intervallo di fiducia bilatero di $\psi(\theta)$, possiamo agire così: l'intervallo è della forma $[t(X_1, \dots X_N) - \delta, t(X_1, \dots, X_N) + \delta]$, quindi la condizione da richiedere è

$$P^{\theta}\left(|\psi(\theta) - t(X_1, \dots, X_N)| < \delta \right) \geq 1-\alpha .$$

A questo punto basta che sia nota, anche approssimativamente, la legge di T, per poter scegliere opportunamente δ, in modo da verificare la relazione precedente.

In definitiva, se le osservazioni del campione sono $x_1, \dots, x_N$, potremo dire che, a un livello di fiducia $1-\alpha$, il valore del parametro sconosciuto $\psi(\theta)$ appartiene all'intervallo reale

$$[t(x_1, \dots, x_N) - \delta, t(x_1, \dots, x_N) + \delta] .$$

Si noti che questo intervallo non ha niente di aleatorio: è vero che $P^{\theta}(\psi(\theta) \in [T-\delta, T+\delta]) \geq 1-\alpha$, ma per l'intervallo sopra scritto si può solo parlare di <u>fiducia</u> (di livello $1-\alpha$) nel fatto che $\psi(\theta)$ vi appartenga. Prima di introdurre le osservazioni, il problema era aleatorio, ma nell'intervallo sopra scritto si è persa l'aleatorietà.

<u>Esempio</u> (stima di una misura fisica). Supponiamo di poter effettuare una sola misura di una grandezza fisica con uno strumento. Detta σ la sensibilità dello strumento, si può vedere tale misura come un campione statistico X di taglia 1 e di legge $\mathcal{N}(\mu, \sigma^2)$, con μ sconosciuta. Indicando la nostra misura con $\bar{x}$, essa sarà una stima puntuale (mediocre) di μ. Vogliamo costruire un intervallo di fiducia bilatero per μ. Stimiamo la probabilità

$$P^\mu\left(|X-\mu| \leq \delta\right):$$

essendo X gaussiana, detta $Z = \frac{X-\mu}{\sigma}$, sarà

$$P^\mu\left(|X-\mu| \leq \delta\right) = P^\mu\left(|Z| \leq \frac{\delta}{\sigma}\right) = P^\mu\left(-\frac{\delta}{\sigma} \leq Z \leq \frac{\delta}{\sigma}\right) =$$

$$= \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(-\frac{\delta}{\sigma}\right) = 2\Phi\left(\frac{\delta}{\sigma}\right) - 1,$$

indipendente da μ. Scelto $\delta = k\sigma$, $k \in \mathbb{N}^+$, avremo

$$P^\mu\left(|X-\mu| \leq k\sigma\right) = 2\Phi(k) - 1,$$

e da qui vediamo che, per $k = 1, 2, 3$, i livelli di fiducia dei corrispondenti intervalli sono 0.68, 0.95, 0.99.

<u>Osservazione</u> Più δ è piccolo, più il parametro $\psi(\theta)$, a livello $1-\alpha$, è ben stimato dal numero $\tau = t(x_1, \dots, x_N)$. Chiamiamo <u>precisione relativa</u> del parametro $\psi(\theta)$ a livello $1-\alpha$ la quantità $\eta = \left|\frac{\tau - 2\delta}{\tau}\right| = \left|1 - \frac{2\delta}{\tau}\right|$. Nell'esempio precedente, le precisioni relative dei tre intervalli sono $\eta_k = \left|1 - \frac{2k\sigma}{\bar{x}}\right|$. Se, ad esempio,

$\bar{x} = 5.4$ e $\sigma = 0.5$, troviamo le seguenti precisioni relative:

- al livello 0.68 ($k=1$), $\eta_1 \cong 0.81$,
- al livello 0.95 ($k=2$), $\eta_2 \cong 0.63$,
- al livello 0.99 ($k=3$), $\eta_3 \cong 0.44$,

come è naturale aspettarsi: più sale il livello di fiducia richiesto, più è ampio l'intervallo, e dunque la precisione relativa è bassa.

Calcoliamo un intervallo di fiducia per la media μ e la varianza σ^2 di un campione gaussiano. Supponiamo dapprima che μ sia sconosciuta ma σ sia nota. Stimiamo μ con la media empirica $\bar{X}$, che ha legge $N\left(\mu, \frac{\sigma^2}{n}\right)$: allora $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ ha legge $N(0,1)$. Al livello $1-\alpha$ si avrà quindi

$$P^{\mu}\left(|\bar{X}-\mu| \leq \delta\right) = P^{\mu}\left(|Z| \leq \frac{\delta\sqrt{n}}{\sigma}\right) = 1-\alpha$$

se e solo se

$$2\,\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - 1 = 1-\alpha$$

ossia se e solo se

$$\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 1-\frac{\alpha}{2}, \quad \text{cioè} \quad \frac{\delta\sqrt{n}}{\sigma} = \phi_{1-\alpha/2}.$$

Perciò

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2}\right]$$

è un intervallo di fiducia per μ al livello $1-\alpha$.

Supponiamo adesso, più realisticamente, che σ non sia nota, e stimiamo μ e σ^2 con la media e la varianza empiriche $\bar{X}$ e S^2.

Poniamo $T = \frac{\bar{X}-\mu}{S}$, ossia sostituiamo, nella v.a.

precedente $Z = \frac{\bar{X}-\mu}{\sigma}$, la deviazione standard σ con quella empirica S. Allora, poiché $T\sqrt{N}$ ha legge di Student $t(N-1)$ per il teorema di Cochran, al livello $1-\alpha$ si ha

$$1-\alpha = P^{\mu,\sigma}\left(|T|\sqrt{N} \le \delta \right) = P^{\mu,\sigma}\left(\left|\frac{\bar{X}-\mu}{S}\right|\sqrt{N} \le \delta \right) =$$

$$= 2P^{\mu,\sigma}\left(\frac{\bar{X}-\mu}{S}\sqrt{N} \le \delta \right) - 1$$

se e solo se

$$P^{\mu,\sigma}\left(\frac{\bar{X}-\mu}{S}\sqrt{N} \le \delta \right) = 1-\alpha/2$$

cioè

$$\delta = t_{1-\alpha/2}(N-1).$$

Ne segue

$$P^{\mu,\sigma}\left(|\bar{X}-\mu| \le \frac{S}{\sqrt{N}}\, t_{1-\alpha/2}(N-1) \right) = 1-\alpha,$$

ossia

$$\left[\bar{X} - \frac{S}{\sqrt{N}}\, t_{1-\alpha/2}(N-1),\ \bar{X} + \frac{S}{\sqrt{N}}\, t_{1-\alpha/2}(N-1) \right]$$

è un intervallo di fiducia per μ al livello $1-\alpha$.

Per quanto riguarda σ^2, usiamo lo stimatore S^2 e ricordiamo che la v.a. $W = \frac{N-1}{\sigma^2} S^2$ ha legge $\chi^2(N-1)$. Allora si vede che

$$P^{\mu,\sigma}\left(W \le \chi^2_{\alpha/2}(N-1) \right) = \frac{\alpha}{2}, \quad P^{\mu,\sigma}\left(W \le \chi^2_{1-\alpha/2}(N-1) \right) = 1-\frac{\alpha}{2},$$

e dunque

$$P^{\mu,\sigma}\left(\chi^2_{\alpha/2}(N-1) \le W \le \chi^2_{1-\alpha/2}(N-1) \right) = 1-\frac{\alpha}{2}-\frac{\alpha}{2} = 1-\alpha.$$

Pertanto

$$1-\alpha = P^{\mu,\sigma}\left(\chi^2_{\alpha/2}(N-1) \le \frac{(N-1)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2}(N-1) \right) =$$

$$= P^{\mu,\sigma}\left(\frac{(N-1)S^2}{\chi^2_{1-\alpha/2}(N-1)} \le \sigma^2 \le \frac{(N-1)S^2}{\chi^2_{\alpha/2}(N-1)} \right),$$

ossia

$$\left[\frac{(N-1)S^2}{\chi^2_{1-\alpha/2}(N-1)} , \frac{(N-1)S^2}{\chi^2_{\alpha/2}(N-1)} \right]$$

è un intervallo di fiducia per σ^2 al livello $1-\alpha$.

Esempio Consideriamo un campione di taglia 14 estratto da una popolazione gaussiana, e misuriamo il peso dei suoi elementi: si ha una media empirica $\bar{X} = 13$ kg, con varianza empirica $S^2 = 2$ kg^2. Si vuole stimare il valore massimo del peso rilevabile in una misurazione, al livello di fiducia 0.90; idem per il valore minimo.

Dobbiamo trovare $\beta, \lambda \ge 0$ in modo che sia (con $1-\alpha = 0.90$)

$$P^{\mu,\sigma}\left(\max_{1\le i\le 14} X_i \le \lambda \right) \ge 0.90,$$

$$P^{\mu,\sigma}\left(\min_{1\le i\le 14} X_i \ge \beta \right) \ge 0.90.$$

Poiché le X_i sono indipendenti con legge $N(\mu, \sigma^2)$, si ha

$$P^{\mu,\sigma}\left(\max_{1\le i\le 14} X_i \le \lambda \right) = P^{\mu,\sigma}\left(\bigcap_{i=1}^{14} \{X_i \le \lambda\} \right) =$$

$$= \prod_{i=1}^{14} P^{\mu,\sigma}\{X_i \le \lambda\};$$

ponendo $Z_i = \frac{X_i - \mu}{\sigma}$ si ha

$$P^{\mu,\sigma}\left(\max_{1\le i\le 14} X_i \le \lambda\right) = \prod_{i=1}^{14} P^{0,1}\left(Z_i \le \frac{\lambda-\mu}{\sigma}\right) = \left[\Phi\left(\frac{\lambda-\mu}{\sigma}\right)\right]^{14} \geq 0.90.$$

Perciò, approssimando σ con la deviazione standard empirica $s=\sqrt{2}$, e μ con la media empirica $\bar{X}=13$,

$$\Phi\left(\frac{\lambda-13}{\sqrt{2}}\right) \geq (0.90)^{\frac{1}{14}} \simeq 0.99$$

ossia

$$\lambda = 13+\sqrt{2}\,\phi_{0.99} \cong 13+\sqrt{2}\cdot 2.33 \cong 16.295$$

Un intervallo di fiducia al livello 0.90 è allora

$$[13, 16.295].$$

Per il minimo degli X_i si trova, analogamente,

$$P^{\mu,\sigma}\left(\min_{1\le i\le 14} X_i \geq \beta\right) = \prod_{i=1}^{14} P^{\mu,\sigma}(X_i \geq \beta) =$$

$$= \prod_{i=1}^{14} P^{0,1}\left(Z_i \geq \frac{\beta-\mu}{\sigma}\right) = \prod_{i=1}^{14}\left(1-\Phi\left(\frac{\beta-\mu}{\sigma}\right)\right) =$$

$$= \left[1-\Phi\left(\frac{\beta-\mu}{\sigma}\right)\right]^{14} = \left[\Phi\left(\frac{\mu-\beta}{\sigma}\right)\right]^{14} \geq 0.90,$$

e con le approssimazioni $\mu \simeq 13$, $\sigma \simeq \sqrt{2}$, si ottiene

$$\Phi\left(\frac{13-\beta}{\sqrt{2}}\right) \geq (0.90)^{1/14} \cong 0.99,$$

$$\beta = 13-\sqrt{2}\cdot\phi_{0.99} = 13-\sqrt{2}\cdot 2.33 \cong 9.705,$$

e dunque un intervallo di fiducia al livello 0.90 è

$$[9.705, 13].$$

Esempio Un campione gaussiano di taglia N ha media empirica $\bar{X} = 23$ e varianza empirica $S^2 = 5$. Quanto grande deve essere N, affinché l'intervallo di fiducia per la media al livello di 0.98 abbia ampiezza minore di 10^{-3}?

Dalla teoria si sa che l'intervallo di fiducia per la media al livello $1-\alpha = 0.98$ è

$$\left[\bar{X} - \frac{S}{\sqrt{N}}\, t_{1-\alpha/2}(N-1),\ \bar{X} + \frac{S}{\sqrt{N}}\, t_{1-\alpha/2}(N-1)\right],$$

con $\bar{X} = 23$, $S = \sqrt{5}$ e $\alpha = 0.02$. La sua ampiezza è $\frac{2\sqrt{5}}{\sqrt{N}}\, t_{0.99}(N-1)$. Poiché, dalla tavola relativa alla legge $t(N)$ segue che

$$\lim_{N\to\infty} t_{0.99}(N) \simeq 2.33,$$

affinché sia

$$\frac{2\sqrt{5}}{\sqrt{N}}\, 2.33 < 10^{-3}$$

dovremo scegliere

$$\sqrt{N} > 1000 \cdot 2 \cdot \sqrt{5} \cdot 2.33 \simeq 10.420,$$

ossia

$$N > 108.576.400.$$