

I modelli probabilistici

Finora abbiamo visto che esistono modelli probabilistici che possiamo utilizzare per *prevedere* gli esiti di esperimenti aleatori.

Naturalmente la *previsione* è di tipo probabilistico: diciamo con quanta probabilità si verificano certi esiti.

Ad esempio se per una moneta scegliamo il modello $\mathcal{B}(1/2)$, diciamo che la probabilità che esca testa in un lancio futuro vale $1/2$.

La statistica inferenziale

Con la statistica inferenziale si utilizzano i dati di più esperimenti aleatori per dire qualcosa sul modello probabilistico che si suppone incognito (almeno in parte).

Esempio: la moneta

Se abbiamo una moneta, essendo gli esiti possibili solamente due (testa e croce), il modello adatto è quello bernoulliano, ma p = probabilità che esca testa potrebbe anche **non** essere $= 1/2$. Come fare per capire quanto vale p ?

Idea: lanciamo molte volte (n) la moneta e contiamo il numero totale di teste (S_n). Dalla legge dei grandi numeri sappiamo che se n è grande sarà assai improbabile che tale S_n/n sia “molto diverso” da p .

Dall'esempio della moneta

Guidati dall'intuizione abbiamo

- 1 scelto un modello (Bernoulli) con un parametro incognito (p);
- 2 fatto n esperimenti con la stessa moneta e indipendenti (n lanci);
- 3 raccolto i dati e calcolato una funzione di questi dati (S_n/n);
- 4 usato questa funzione per stimare il parametro incognito.

I passaggi evidenziati in rosso sono quelli che si fanno in tutti i problemi di stima di parametri.

Introduciamo allora le definizioni necessarie.

Modello statistico

DEFINIZIONE DI MODELLO STATISTICO

Un **modello statistico parametrico** è una famiglia di leggi di v.a., dipendenti da uno o più parametri θ : lo indichiamo tramite la funzione di densità:

$$\{f(x; \theta) : \theta \in \Theta\}.$$

Θ è lo **spazio dei parametri**.

In altre parole, il modello statistico indica la scelta di un **“tipo” di v.a.** (ricordiamo inoltre che abbiamo già osservato che conoscere la legge di una v.a. equivale a conoscere la sua densità (quando quest’ultima esiste)).

Esempi di modelli statistici.

- Il modello $\mathcal{B}(p)$: $f(1; p) = p$, $f(0; p) = 1 - p$ e $f(x; p) = 0$ per $x \neq 0, 1$.
Il parametro incognito è p e lo spazio dei parametri è $[0, 1]$.
- Il modello $\mathcal{N}(\mu, \sigma^2)$: la densità è una funzione di x (la campana) che dipende da μ e σ^2 . Lo spazio dei parametri è $\mathbb{R} \times (0, +\infty)$ (infatti μ può essere qualsiasi numero reale, σ^2 deve essere > 0).

Il campione casuale

DEFINIZIONE DI CAMPIONE CASUALE

Un **campione casuale** di dimensione n è una n -upla di v.a. X_1, \dots, X_n i.i.d.

Il campione si dice estratto da una popolazione di legge $f(x; \theta)$ se ciascuna delle v.a. ha legge $f(x; \theta)$.

Il campione casuale è quindi il “prodotto” degli n esperimenti che ci accingiamo a fare.

DEFINIZIONE DI STATISTICA

Dato un campione casuale X_1, X_2, \dots, X_n , si dice **statistica** una v.a. T funzione del campione casuale che NON sia funzione di alcun parametro incognito.

In altre parole, esiste una funzione t per cui $T = t(X_1, X_2, \dots, X_n)$ e t NON dipende da parametri incogniti.

T è funzione dei dati che raccoglieremo dagli n esperimenti che ci accingiamo a fare.

Esempi.

$T = X_1 + X_3 - 5$, $W = X_4 X_5$ sono statistiche.

$Y = X_1 + pX_2$, se p è incognito, non è una statistica.

DEFINIZIONE DI STIMATORE

Si dice **stimatore** di $g(\theta)$ (con g funzione definita su Θ una statistica usata per stimare $g(\theta)$). Assegnata la statistica $T = t(X_1, X_2, \dots, X_n)$, una volta estratto un particolare campione (x_1, x_2, \dots, x_n) , il numero $\tau = t(x_1, x_2, \dots, x_n)$ si dice **stima** di $g(\theta)$.

Osservazione

Lo stimatore è una variabile aleatoria, mentre la stima è un numero reale.

Lo stimatore è ciò che ho PRIMA degli esperimenti (è il modello del mio esperimento), la stima è ciò che ho DOPO gli esperimenti (a posteriori, dopo aver “pescato” $\omega \in \Omega$, si ha $x_i := X_i(\omega)$ da cui $T(\omega) = t(X_1(\omega), \dots, X_n(\omega)) = \tau$).

Quando uno stimatore è “buono”?

A priori ogni funzione del campione casuale può essere usata per stimare un parametro, ma è chiaro che ci saranno stimatori migliori di altri.

Tra le tante proprietà per uno stimatore “desiderabili” ne segnaliamo due: la non distorsione e la consistenza.

Stimatori non distorti

DEFINIZIONE DI STIMATORE NON DISTORTO

Uno stimatore T di $g(\theta)$ si dice **non distorto**, se $\mathbb{E}_\theta(T) = g(\theta)$ per ogni $\theta \in \Theta$.

Perché \mathbb{E}_θ

La legge, e quindi anche il valore atteso, di T dipende da θ ; nel seguito spesso, per semplicità di notazione, sottintenderemo i pedici θ .

Significato della non distorsione

Se T è uno stimatore non distorto di $g(\theta)$, la v.a. T ha come “centro dei suoi possibili valori” proprio $g(\theta)$.

Stimatori consistenti

Spesso lo stimatore scelto dipende esplicitamente, oltre che dal campione casuale X_1, \dots, X_n , dal numero n (la numerosità del campione). Quindi di fatto abbiamo una successione di stimatori $\{T_n\}_{n \geq 0}$.

DEFINIZIONE DI STIMATORE CONSISTENTE

Una successione di stimatori $\{T_n\}_{n \geq 0}$ di $g(\theta)$ si dice **consistente** (in media quadratica) se

$$\lim_{n \rightarrow +\infty} \mathbb{E}_\theta \left((T_n - g(\theta))^2 \right) = 0, \quad \forall \theta \in \Theta.$$

$\mathbb{E}_\theta \left((T_n - g(\theta))^2 \right)$ è detto **rischio quadratico medio** di T_n .

Significato della consistenza

La successione $\{T_n\}_{n \geq 0}$ è consistente se il rischio quadratico medio

$$\mathbb{E}_\theta \left((T_n - g(\theta))^2 \right)$$

va a zero quando n va a infinito.

Ma il rischio quadratico medio rappresenta una **misura di quanto T_n sia “disperso” rispetto a $g(\theta)$** (ricordiamo che $g(\theta)$ è l'incognita che vogliamo stimare).

Essere consistenti significa quindi che questa dispersione diventa piccola quando n è grande.

Dunque la stima sarà tanto più precisa tanto più n è grande.

Stimatori non distorti e consistenti

Stimatori non distorti e consistenti

Se i T_n sono stimatori non distorti di $g(\theta)$, allora

$$\mathbb{E}_\theta \left((T_n - g(\theta))^2 \right) = \text{Var}(T_n).$$

Dunque in questo caso rischio quadratico medio = varianza. Quindi se so che uno stimatore è non distorto, per vedere se è consistente devo solo vedere se la varianza tende a zero per $n \rightarrow \infty$.

Più precisamente, per uno stimatore generico T di $g(\theta)$, vale

$$\mathbb{E}_\theta \left((T - g(\theta))^2 \right) = \text{Var}(T) + (\mathbb{E}_\theta(T) - g(\theta))^2.$$

Stimatore per $\mathbb{E}(X)$

Se abbiamo una v.a. X e vogliamo stimare $\mathbb{E}(X)$ come possiamo fare? (È la generalizzazione del problema della moneta...).

Prenderemo un campione casuale X_1, \dots, X_n in cui ogni X_i abbia la stessa legge di X , **lo stimatore naturale per $\mathbb{E}(X)$ è la media campionaria**

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Proprietà di \bar{X}_n

Basta ricordare i conti già visti per $\mathbb{E}(\bar{X}_n)$ e $\text{Var}(\bar{X}_n)$ per ottenere che

- è uno stimatore non distorto di $\mathbb{E}(X)$;
- è uno stimatore consistente di $\mathbb{E}(X)$.

\bar{X}_n è meglio di \bar{X}_m se $n > m$

Qualsiasi sia n (numerosità del campione) \bar{X}_n è uno stimatore non distorto di $\mathbb{E}(X)$. Quindi dal punto di vista della distorsione $\bar{X}_2 = \frac{X_1 + X_2}{2}$ non appare migliore o peggiore di $\bar{X}_{40} = \frac{X_1 + \dots + X_{40}}{40}$.

L'intuizione ci dice che fare 40 esperimenti porta in genere ad una stima migliore di quella fornita da 2 esperimenti.

\bar{X}_n è uno stimatore migliore se n è grande

È una conseguenza del fatto che il rischio quadratico medio (che coincide con la varianza) diminuisce all'aumentare di n : $\text{Var}(\bar{X}_n) = \text{Var}(X)/n$.

Stimatori per $\text{Var}(X)$

Ricordiamo la definizione di varianza.

DEFINIZIONE DI VARIANZA DI UNA V.A.

Data una variabile aleatoria X la sua varianza è il numero reale:

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

Ricordiamo anche che un buon stimatore di $\mathbb{E}(X)$ è \bar{X}_n (la media aritmetica del campione).

Idea per costruire uno stimatore

Si prende l'espressione teorica del parametro da stimare e si sostituiscono le medie teoriche (\mathbb{E}) con le medie aritmetiche.

Stimiamo $\text{Var}(X)$

Supporremo che la media teorica $\mathbb{E}(X)$ sia incognita: esiste uno stimatore della varianza per i casi in cui $\mathbb{E}(X)$ sia nota, ma di fatto nelle situazioni reali sia $\mathbb{E}(X)$ che $\text{Var}(X)$ sono incognite (quindi quello stimatore è poco usato ai fini pratici).

Sostituiamo nella formula della varianza la \mathbb{E} con le medie aritmetiche:

$$\begin{aligned}\mathbb{E}((X - \mathbb{E}(X))^2) &\implies \mathbb{E}((X - \bar{X}_n)^2) \\ &\implies \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2\end{aligned}$$

Da cui sembra naturale proporre come stima per $\text{Var}(X)$ lo stimatore

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

W_n è distorto

Se calcolassimo il valore atteso di W_n scopriremmo che

$$\mathbb{E}(W_n) = \frac{n-1}{n} \text{Var}(X).$$

La cosa non è irrimediabile: basta moltiplicare W_n per $\frac{n}{n-1}$ e si ottiene lo stimatore S_n^2 che è non distorto. Altri calcoli mostrerebbero che è anche consistente.

Stimatore di $\text{Var}(X)$

Per ogni n

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

è uno stimatore non distorto di $\text{Var}(X)$ e $\{S_n^2\}_{n \geq 1}$ è una successione di stimatori consistente in media quadratica. Questo stimatore è detto **varianza campionaria**.

La formula

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

costringe a calcolare n differenze da elevare al quadrato. Si potrebbe mostrare che questa formula equivale alla seguente.

Formula alternativa per la varianza campionaria

$$s_n^2 = \left(\frac{1}{n-1} \sum_{i=1}^n (X_i)^2 \right) - \frac{n}{n-1} (\bar{X}_n)^2.$$

Questa formula risulta particolarmente utile quando abbiamo la somma dei dati e la somma dei loro quadrati.