MODELLO STATISTICO PARAMETRICO

la famiglia $\{F(x;\theta) : \theta \in \Theta, x \in S\}$ di funzioni di ripartizione indicizzate dal **vettore di parametri** θ appartenente allo **spazio parametrico** Θ verrà chiamata **modello statistico**. L'insieme S è detto spazio campionario.

Alternativamenete il modello statistico può essere descritto attraverso la funzione di probabilità $\{p(x,\theta)\colon \theta \in \Theta, x \in S\}$ (variabili casuali discrete) o la funzione di densità di probabilità $\{f(x,\theta)\colon \theta \in \Theta, x \in S\}$(variabili casuali continue)

esempio: modello variabile casuale indicatore (binomiale con $n = 1$) :

$$\left\{\pi^x(1-\pi)^{1-x} : 0 < \pi < 1, x = 0, 1\right\}$$

esempio: modello di Poisson

$$\left\{\frac{1}{x!}\lambda^x e^{-\lambda} : \lambda > 0, x = 1, 2, ...\right\}$$

esempio: modello normale o di Gauss

$$\left\{ \begin{array}{c} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} : \\ \sigma > 0, -\infty < \mu < \infty, x \in \Re \end{array} \right\}$$

esempio: modello esponenziale (Gamma con $\alpha = 1$)

$$\left\{\lambda e^{-\lambda x} : \lambda > 0, x > 0\right\}$$

CAMPIONE CASUALE

Il vettore $[X_1, X_2, ..., X_n]$ di n variabili casuali indipendenti in probabilità e identicamente distribuite con funzione di ripartizione appartenente al modello statistico

$$M \equiv \{F(x;\theta) : \theta \in \Theta, x \in S\}$$

è chiamato campione casuale dal modello M.

Le variabili casuali $X_i, i = 1, 2, ..., n$, sono chiamate variabili casuali campionarie e servono per modellare altrettante osservazioni sperimentali.

Una determinazione sperimentale $[x_1, x_2, ..., x_n]$ di $[X_1, X_2, ..., X_n]$ è chiamata **realizzazione campionaria.**

Ad $[x_1, x_2, ..., x_n]$ ci si riferisce anche con i termini di dati sperimentali o dati campionari

Una operazione di inferenza statistica consiste nel fare congetture sull'incognito vettore di parametri $\theta = [\theta_1, \theta_2, ..., \theta_p]$ basandosi su una realizzazione campionaria

$$[x_1, x_2, ..., x_n] .$$

Vi sono tre forme fondamentali di inferenza statistica

- **Stima puntuale**: in base ad una realizzazione campionaria $[x_1, x_2, ..., x_n]$ i parametri sono valutati numericamente

- **Stima per intervalli:** la realizzazione campionaria $[x_1, x_2, ..., x_n]$ è utilizzata per valutare un parametro non attraverso un singolo numero ma mediante un intervallo all'interno del quale si confida si trovi il vero valore incognito del parametro.

- **Verifica di ipotesi statistiche:**spesso se uno o più parametri hanno opportuni valori predefiniti sono rispettati degli standard di produzione o sono soddisfatti certi obiettivi o condizioni di ottimalità. In questo contesto non interessa tanto valutare numericamente i parametri sconosciuti e la realizzazione campionaria è utilizzata per sapere se i parametri hanno proprio quei valori che soddisfano standard e obiettivi.

A questo proposito si osservi che:

- il modello statistico serve per rappresentare con alcuni dei suoi parametri le quantità fisiche di interesse (*parametri di interesse fisico*)

- il modello statistico serve per tener conto del fatto che si osserva un *fenomeno ripetibile con risultato incerto* cioè che i risultati non sono costanti da prova a prova (anche se ottenuti sotto le stesse condizioni) se non altro per via degli errori di misura. Il modello statistico quindi tiene conto della *variabilità dei risultati*. I parametri del modello che rispondono solo a questo scopo sono chiamati *parametri di interesse statistico*

La definizione di campione casuale da un modello statistico formalizza che **lo stesso fenomeno** ripetibile con risultato incerto viene osservato **n volte** (le **n variabili casuali campionarie con la stessa funzione di ripartizione**) e che le n osservazioni sono ottenute in **modo indipendente** (la condizione di **indipendenza in probabilità** delle n variabili casuali campionarie)

UN PARTICOLARE TIPO DI VARIABILI CASUALI: LE STATISTICHE

Definizione: una funzione

$$T = f(X_1, X_2, ..., X_n)$$

delle sole variabili casuali campionarie è chiamata *statistica*.

Il valore $t = f(x_1, x_2, ..., x_n)$ che la statistica assume in corrispondenza di una realizzazione campionaria $[x_1, x_2, ..., x_n]$ è chiamato realizzazione campionaria o valore sperimentale della statistica

Le statistiche sono uno degli strumenti di base utilizzati nell'ineferenza statistica.

Se usate in un problema di stima puntuale le statistiche prendono il nome di stimatori se usate in un problema di verifica di ipotesi vengono chiamate *statistiche test*.

esempi:

- media campionaria $M_n = \frac{\sum_{i=1}^{n} X_i}{n}$

- varianza campionaria $V_n^2 = \frac{\sum_{i=1}^{n} (X_i - M_n)^2}{n}$

- varianza campionaria corretta:

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

  (il perchè dell'aggettivo corretto verrà discusso in seguito)

- frequenza cumulata relativa campionaria: $C_n(x) = \frac{\sum_{i=1}^{n} I_{x_i \leq x}}{n}$

- frequenza relativa campionaria: $F_n(x) = \frac{\sum_{i=1}^{n} I_{x_i = x}}{n}$

Le seguenti variabili casuali non sono statistiche a meno che non siano noti i parametri da cui dipendono

Media campionaria standardizzata: $Z_n = \frac{M_n - \mu}{\sqrt{\sigma^2}}\sqrt{n}$

Media campionaria standardizzata di Student: $T_n = \frac{M_n - \mu}{\sqrt{S_n^2}}\sqrt{n}$

Nelle formule precedenti μ e σ^2 sono rispettivamente il valore atteso e la varianza delle variabili casuali campionarie

UN PONTE TRA NOTO E IGNOTO: LA FUNZIONE DI VEROSIMIGLIANZA (variabili campionarie discrete)

Sia $[X_1, X_2, ..., X_n]$ un campione casuale di variabili casuali discrete dal modello:

$$M \equiv \{p(x, \theta) \colon \theta \in \Theta, x \in S\}$$

la funzione:

$$V(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i, \theta)$$

è chiamata funzione di probabilità congiunta delle variabili casuali campionarie. La precedente funzione vista come funzione di θ (con le realizzazioni campionarie $x_1, x_2, ..., x_n$ considerate come parametri) è chiamata funzione di verosimiglianza (al variare di θ descrive quanto sono verosimili i dati sperimentali) Il logaritmo della funzione di verosimiglianza è chiamato funzione di log-verosimiglianza:

$$L(\theta) = \sum_{i=1}^{n} \ln p(x_i, \theta)$$

UN PONTE TRA NOTO E IGNOTO: LA FUNZIONE DI VEROSIMIGLIANZA (variabili campionarie continue)

Sia $[X_1, X_2, ..., X_n]$ un campione casuale di variabili casuali continue dal modello: $M \equiv \{f(x,\theta)\colon \theta \in \Theta, x \in S\}$ la funzione

$$V(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i, \theta)$$

è chiamata funzione di densità di probabilità congiunta delle variabili casuali campionarie. La precedente funzione vista come funzione di θ (con le realizzazioni campionarie $x_1, x_2, ..., x_n$ considerate come parametri) è chiamata funzione di verosimiglianza (al variare di θ descrive quanto sono verosimili i dati sperimentali) Il logaritmo della funzione di verosimiglianza è chiamato funzione di log-verosimiglianza:

$$L(\theta) = \sum_{i=1}^{n} \ln f(x_i, \theta)$$

ESEMPI DI LOG-VEROSIMIGLIANZA

modello per variabile casuale indicatore

$$L(\pi) = \sum_{i=1}^{n} x_i \ln \pi + (n - \sum_{i=1}^{n} x_i) \ln(1 - \pi)$$

modello di Poisson

$$L(\lambda) = -\sum_{i=1}^{n} \ln(x_i!) - n\lambda - \sum_{i=1}^{n} x_i \ln \lambda$$

modello Normale

$$L(\mu, \sigma) = -\frac{1}{2} \ln(2\pi\sigma) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

modello esponenziale

$$L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

STIMA PUNTUALE

L'inferenza mediante stima puntuale di un parametro scalare θ consiste in:

Nella individuazione di una statistica chiamata stimatore

$$T(X_1, X_2, ..., X_n)$$

Nell'ottenere una realizzazione campionaria

$$[x_1, x_2, ..., x_n]$$

Nell'assegnare al parametro incognito oggetto di stima il valore sperimentale $T(x_1, x_2, ..., x_n)$ dello stimatore

REQUISITI MINIMI DI UNO STIMATORE

Corretezza o non distorsione:

$$E(T(X_1, X_2, ..., X_n)) = \theta$$

per ogni θ,ovvero $E(\, T(X_1, X_2, ..., X_n) - \theta) = 0$ cioè il valore atteso dell'errore di stima deve essere nullo

Consitenza:

$$\lim_{n \to \infty} P(|\, T(X_1, X_2, ..., X_n) - \theta| > \varepsilon) = 0$$

per ogni ε. Ovvero la probabilità di un errore di stima grossolano deve tendere a zero al divergere della numerosità campionaria n.

Efficienza: dati due stimatori corretti

$$T_i(X_1, X_2, ..., X_n), i = 1, 2$$

di θ la loro varianza

$$\sigma_i^2(\theta) = E\left[(\, T(X_1, X_2, ..., X_n) - \theta)^2\right]$$

è una misura della loro precisione. Se

$$\sigma_1^2(\theta) \leq \sigma_2^2(\theta)$$

per ogni θ lo stimatore $T_1(X_1, X_2, ..., X_n)$ è più efficiente di $T_2(X_1, X_2, ..., X_n)$.

Un problema importante della statistica è la ricerca dello stimatore più efficiente tra tutti quelli corretti per un parametro.

Per gli stimatori corretti si può dimostrare che condizione sufficiente per la consistenza è che la loro varianza tenda a zero al divergere di n.

UNO STIMATORE CORRETTO FONDAMENTALE

Sia $[X_1, X_2, ..., X_n]$ un campione casuale con $E(X_i) = \mu, \quad Var(X_i) = \sigma^2, \quad i = 1, 2, ..., n.$ Per quanto detto a proposito del teorema di normalità asintotica la statistica $M_n = \frac{\sum_{i=1}^{n} X_i}{n}$ ha valore atteso μ e varianza $\frac{\sigma^2}{n}$.

Quindi la media campionaria è uno stimatore corretto e consistente di μ

La seguente tabella specifica il precedente risultato per vari modelli statistici.

COSA STIMA CORRETTAMENTE LA MEDIA CAMPIONARIA?

mod.	**param.**	$E(M_n)$	$Var(M_n)$
Binom.	$1, \pi$	π	$\frac{\pi(1-\pi)}{n}$
Poiss.	λ	λ	$\frac{\lambda}{n}$
Espon.	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2 n}$
Gamma	α, λ	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2 n}$
Weibull	β, λ	$\lambda^{-1/\beta}.$ $\cdot\Gamma(1+\frac{1}{\beta})$	$\frac{1}{n}\lambda^{-2/\beta}.$ $\cdot\left[\Gamma(1+\frac{2}{\beta})-\right.$ $\left.-\Gamma^2(1+\frac{1}{\beta})\right]$
Norm.	μ, σ	μ	$\frac{\sigma^2}{n}$

STIMATORE CORRETTO DELLA VARIANZA

Poichè se $[X_1, X_2, ..., X_n]$ è un campione casuale con $E(X_i) = \mu, \quad Var(X_i) = \sigma^2, \quad i = 1, 2, ..., n$,si ha che:

$$E\left(\sum_{i=1}^{n}(X_i - m)^2\right) = (n-1)\sigma^2$$

segue che:

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

è uno stimatore corretto di $Var(X_i) = \sigma^2$

COSA STIMA CORRETTAMENTE S_n^2?

mod.	**param.**	$E(S_n^2)$
Binom.	$1, \pi$	$\pi(1-\pi)$
Poiss.	λ	λ
Espon.	λ	$\frac{1}{\lambda^2}$
Gamma	α, λ	$\frac{\alpha}{\lambda^2}$
Weibull	β, λ	$\lambda^{-2/\beta}.$ $\cdot\left[\Gamma(1+\frac{2}{\beta})-\right.$ $\left.-\Gamma^2(1+\frac{1}{\beta})\right]$
Norm.	μ, σ	σ^2

Metodi automatici per generare stimatori puntuali

Metodo analogico: ad esempio se il parametro da stimare un valore atteso (varianza) si usa la media aritmetica (varianza) dei dati campionari

Metodo dei momenti: ad esempio nel caso di modelli statistici con due parametri gli stimatori degli stessi sono ottenuti risolvendo il sistema la cui prima equazione uguaglia la media campionaria al valore atteso delle variabili casuali campionarie mentre la seconda equazione eguaglia la varianza campionaria con la varianza delle variabil casuali campionarie

Metodo di massima verosimiglianza: Gli stimatori sono definiti come quei valori dei parametri in corrispondenza dei quali la log-verosimiglianza ha un massimo globale.