



Stima puntuale

Stimatori e stime · media, varianza e proporzione campionaria · proprietà degli stimatori · metodi dei momenti e della massima verosimiglianza · informazione di Fisher · disuguaglianza di Cramér-Rao.

IN QUESTO FORMULARIO

2 Media campionaria

4 Proporzione campionaria

6 Metodi di ricerca

8 Cramér-Rao

3 Varianza campionaria

5 Proprietà degli stimatori

7 Informazione di Fisher

A Proprietà E e VAR

Stimatori e stime

Come si passa dai dati campionari ai parametri della popolazione.

Per stimare i parametri di una popolazione si usano i dati campionari: in genere si ricorre a funzioni dei dati (medie, proporzioni, varianze, ...). Queste funzioni campionarie si chiamano **statistiche campionarie** o **stimatori**.

DEFINIZIONE

Stimatore

Una statistica usata per stimare il valore di un parametro ignoto di una popolazione. Essendo funzione di variabili casuali, è a sua volta una variabile casuale con la propria distribuzione campionaria.

$$T = f(X_1, X_2, \dots, X_n)$$

DEFINIZIONE

Stima

Ogni particolare valore assunto da uno stimatore: è un numero.

Gli stimatori più rilevanti sono tre:

- **Media campionaria** $\bar{X}$ — per stimare la media della popolazione;
- **Varianza campionaria** — per stimare la varianza della popolazione;
- **Proporzione campionaria** p — per stimare la frazione di elementi con una data caratteristica.

SEZIONE 2 · MEDIA CAMPIONARIA

Media campionaria

Lo stimatore della media della popolazione.

Dato un campione casuale $(X_1, X_2, \dots, X_n)$ di dimensione n estratto da una popolazione qualsiasi con media μ e varianza σ^2 , la media campionaria è definita come:

FORMULA 1 · MEDIA CAMPIONARIA

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

I suoi primi due momenti valgono, per qualunque popolazione:



FORMULA 2 · VALORE ATTESO

$$E(\bar{X}) = \mu$$

FORMULA 3 · VARIANZA

$$VAR(\bar{X}) = \frac{\sigma^2}{n}$$

Se il campione proviene da una popolazione normale $N(\mu, \sigma^2)$, anche la media campionaria è normale, con la stessa media e varianza σ^2/n :

FORMULA 4 · DISTRIBUZIONE (POPOLAZIONE NORMALE)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Se la popolazione **non** è normale, per campioni «grandi» vale l'approssimazione del teorema del limite centrale: la media campionaria standardizzata converge a una normale standard.

FORMULA 5 · STANDARDIZZAZIONE (TLC)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n \rightarrow \infty} N(0, 1)$$

i La distribuzione della popolazione ha influenza pressoché nulla sulla media campionaria per campioni con numerosità superiore a 50.

SEZIONE 3 · VARIANZA CAMPIONARIA

Varianza campionaria

Due stimatori: non corretto e corretto.

Esistono due stimatori della varianza. Dato il campione $(X_1, X_2, \dots, X_n)$ con media μ e varianza σ^2 :

Varianza campionaria non corretta

FORMULA 6 · DEFINIZIONE

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



i Se la media μ è nota, al posto di $\bar{X}$ si sostituisce direttamente μ .

FORMULA 7 · VALORE ATTESO

$$E(\tilde{s}^2) = \frac{n-1}{n} \sigma^2$$

μ_4 momento centrato quarto σ^2 varianza della popolazione

FORMULA 8 · VARIANZA

$$VAR(\tilde{s}^2) = \frac{(n-1)^2}{n^3} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)$$

Se la popolazione è normale, la varianza non corretta moltiplicata per n/σ^2 si distribuisce come un chi-quadro:

FORMULA 9 · DISTRIBUZIONE (POPOLAZIONE NORMALE)

$$\frac{n}{\sigma^2} \tilde{s}^2 \sim \chi_{n-1}^2$$

Varianza campionaria corretta

FORMULA 10 · DEFINIZIONE

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

La si può riscrivere in forma operativa come:

FORMULA 11 · FORMA OPERATIVA

$$s^2 = \frac{n}{n-1} \left(\overline{x^2} - \bar{x}^2 \right)$$

Questo stimatore è corretto: il suo valore atteso coincide con la varianza vera.

FORMULA 12 · VALORE ATTESO

$$E(s^2) = \sigma^2$$



FORMULA 13 · VARIANZA

$$VAR(s^2) = \frac{\sigma^4}{n} \left(\gamma + 2 \frac{n}{n-1} \right)$$

γ curtosi (eccesso di curtosi)

Se la popolazione è normale:

FORMULA 14 · DISTRIBUZIONE (POPOLAZIONE NORMALE)

$$\frac{n-1}{\sigma^2} s^2 \sim \chi_{n-1}^2 \quad \Longleftrightarrow \quad s^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$$

SEZIONE 4 · PROPORZIONE CAMPIONARIA

Proporzione campionaria

Lo stimatore della frazione di elementi con una caratteristica.

Una popolazione può essere divisa in due gruppi in base al possesso di una determinata caratteristica. Si chiama **proporzione** il rapporto tra il numero di unità che possiedono la caratteristica e l'intera popolazione:

FORMULA 15 · PROPORZIONE (PARAMETRO)

$$\pi = \frac{\text{n° unità con la caratteristica}}{N}$$

Per definire lo stimatore si introduce una variabile che vale 1 sulle unità che possiedono la caratteristica e 0 altrimenti. Lo stimatore **frazione campionaria** è:

FORMULA 16 · FRAZIONE CAMPIONARIA

$$p = \frac{n_c}{n}$$

n_c n° di unità con la caratteristica n dimensione del campione

FORMULA 17 · VALORE ATTESO

$$E(p) = \pi$$



FORMULA 18 · VARIANZA

$$VAR(p) = \frac{\pi(1 - \pi)}{n}$$

Per n grande la frazione campionaria si distribuisce normalmente:

FORMULA 19 · DISTRIBUZIONE ASINTOTICA

$$p \xrightarrow{n \rightarrow \infty} N\left(\pi, \frac{\pi(1 - \pi)}{n}\right)$$

SEZIONE 5 · PROPRIETÀ DEGLI STIMATORI

Proprietà degli stimatori

Correttezza, consistenza, efficienza.

Ogni stimatore T di un parametro θ ha una certa variabilità, e ogni parametro può avere più stimatori possibili. Spesso è utile stabilire quale sia più attendibile. Le proprietà da valutare sono tre:

correttezza (non distorsione), **consistenza** ed **efficienza**.

Correttezza (non distorsione)

Lo stimatore T è corretto per θ se il suo valore atteso coincide con il parametro da stimare:

FORMULA 20 · CORRETTEZZA

$$E(T) = \theta \implies T \text{ corretto}$$

FORMULA 21 · DISTORSIONE

$$BIAS(T) = E(T) - \theta$$

Nel caso in cui il parametro sia la media:

FORMULA 22 · CORRETTO VS DISTORTO

$$E(T) = \mu \text{ (corretto)} \quad E(T) \neq \mu \text{ (distorto)}$$

Se lo stimatore non è corretto, si può verificare se lo è almeno asintoticamente:

FORMULA 23 · CORRETTEZZA ASINTOTICA

$$\lim_{n \rightarrow \infty} E(T) = \theta \quad (\text{correttezza asintotica})$$



Consistenza

Lo stimatore di θ è consistente se la successione degli stimatori converge in probabilità al parametro: all'aumentare della dimensione campionaria la distribuzione si concentra sempre più attorno al valore vero.

FORMULA 24 · CONDIZIONI SUFFICIENTI (TEOREMA)

$$\lim_{n \rightarrow \infty} E(T_n) = \theta \quad \lim_{n \rightarrow \infty} VAR(T_n) = 0$$

Efficienza

Per valutare la prossimità di T a θ si usa l'**errore quadratico medio** (MSE):

FORMULA 25 · MSE

$$MSE(T) = E[(T - \theta)^2]$$

Una forma più comoda da usare, che separa varianza e distorsione, è:

FORMULA 26 · SCOMPOSIZIONE MSE

$$MSE(T) = VAR(T) + BIAS^2(T)$$

Dati due stimatori T_1 e T_2 , è più efficiente quello con MSE inferiore:

FORMULA 27 · CONFRONTO DI EFFICIENZA

$$MSE(T_1) < MSE(T_2) \implies T_1 \text{ più efficiente di } T_2$$

- i** Uno stimatore non distorto si dice **efficiente** se la sua varianza raggiunge il limite inferiore di Cramèr-Rao (vedi Sezione 8): $VAR(T) = \frac{1}{n I(\theta)}$ (stimatore efficiente).

ESEMPIO · EFFICIENZA

Confronto di due stimatori

Quale dei due è più efficiente?

Si consideri un campione i.i.d. di ampiezza $n = 4$, estratto da una popolazione con media μ e varianza σ^2 , e i due stimatori:



Quale stimatore è più efficiente?

Gli stimatori da confrontare:

- primo stimatore

$$T_1 = \bar{X}$$

- secondo stimatore

$$T_2 = \frac{X_1 + 2X_2 + 3X_3 + X_4}{4}$$

1) Valore atteso di entrambi:

$$E(T_1) = E(\bar{X}) = \mu$$

$$E(T_2) = \frac{1}{4}\mu + \frac{1}{2}\mu + \frac{3}{4}\mu + \frac{1}{4}\mu = \frac{7}{4}\mu$$

Il primo è corretto, il secondo no. Le distorsioni:

$$BIAS(T_1) = 0 \quad BIAS(T_2) = \frac{7}{4}\mu - \mu = \frac{3}{4}\mu$$

2) Varianza di entrambi (i termini usano i coefficienti al quadrato):

$$VAR(T_1) = \frac{\sigma^2}{n} = \frac{\sigma^2}{4}$$

$$VAR(T_2) = \frac{1}{16}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{9}{16}\sigma^2 + \frac{1}{16}\sigma^2 = \frac{15}{16}\sigma^2$$

3) Errore quadratico medio:

$$MSE(T_1) = \frac{\sigma^2}{4} \quad MSE(T_2) = \frac{15}{16}\sigma^2 + \frac{9}{16}\mu^2$$

Poiché $MSE(T_1) = \sigma^2/4$ è sempre minore di $MSE(T_2)$, lo stimatore $T_1 = \bar{X}$ è più efficiente di T_2 .

Metodi di ricerca degli stimatori

Come si costruisce uno stimatore.

I metodi statistici che permettono di trovare stimatori sono:

- **metodo dei momenti**;
- **metodo della massima verosimiglianza**;
- **metodo dei minimi quadrati** (vedi regressione).

Metodo dei momenti

Per stimare i parametri si uguagliano i **momenti teorici** ai **momenti osservati** sul campione. Data una v.a. θ -dipendente, assolutamente continua con densità $f(x, \theta)$, e un campione casuale, si definiscono:

FORMULA 28 · MOMENTO TEORICO R-ESIMO

$$\mu_r = \int_{-\infty}^{+\infty} x^r f(x, \theta) dx$$

FORMULA 29 · MOMENTO CAMPIONARIO R-ESIMO

$$M_r = \frac{\sum_{i=1}^n X_i^r}{n}$$

La stima si ottiene risolvendo il sistema di k equazioni in k incognite ottenuto uguagliando i primi k momenti teorici ai corrispondenti momenti campionari.

Metodo della massima verosimiglianza

Si sceglie il valore del parametro che rende **massima la probabilità** di osservare il campione, cioè la **funzione di verosimiglianza**. Sia $(X_1, X_2, \dots, X_n)$ un campione casuale con realizzazione osservata $x_1, \dots, x_n$:

FORMULA 30 · VEROSIMIGLIANZA

$$L(\theta, \underline{X}) = \prod_{i=1}^n f(x_i; \theta)$$

- i** Nella verosimiglianza i dati sono **fissati**, mentre il parametro incognito varia: $L(\theta)$ indica quanto è plausibile uno specifico valore del parametro, una volta osservato il campione.

Convieni lavorare con la log-verosimiglianza:



FORMULA 31 · LOG-VEROSIMIGLIANZA

$$l(\theta, \underline{X}) = \log L(\theta, \underline{X})$$

La stima di massima verosimiglianza $\hat{\theta}$ è la soluzione dell'equazione di verosimiglianza, che nel punto di massimo soddisfa la condizione del secondo ordine:

FORMULA 32 · EQUAZIONE DI VEROSIMIGLIANZA

$$\frac{\partial}{\partial \theta} \log L(\theta, \underline{X}) = 0$$

FORMULA 33 · CONDIZIONE DI MASSIMO

$$\left. \frac{\partial^2}{\partial \theta^2} \log L(\theta, \underline{X}) \right|_{\theta=\hat{\theta}} < 0$$

SEZIONE 7 · INFORMAZIONE DI FISHER

Informazione attesa di Fisher

Quanta informazione sul parametro porta il campione.

Si definisce **informazione attesa di Fisher** la funzione di θ data dalla varianza della funzione di score:

FORMULA 34 · INFORMAZIONE DI FISHER

$$I_n(\theta) = VAR(u(\theta, \underline{X}))$$

FORMULA 35 · FUNZIONE DI SCORE

$$u(\theta, \underline{X}) = \frac{\partial \log L(\theta, \underline{X})}{\partial \theta}$$

Poiché per definizione il valore atteso dello score è nullo, $E(u(\theta, \underline{X})) = 0$, l'informazione si può scrivere in due forme equivalenti:

FORMULA 36 · FORME EQUIVALENTI

$$I_n(\theta) = E \left[\left(\frac{\partial}{\partial \theta} \log L \right)^2 \right] = -E \left(\frac{\partial^2}{\partial \theta^2} \log L \right)$$

L'informazione di Fisher gode della proprietà di **additività**, che semplifica i calcoli nel caso di campionamento casuale:



FORMULA 37 · ADDITIVITÀ

$$I_n(\theta) = \sum_{i=1}^n I_i(\theta) = n I(\theta)$$

FORMULA 38 · INFORMAZIONE DELLA SINGOLA OSSERVAZIONE

$$I(\theta) = E \left[\left(\frac{\partial}{\partial \theta} \log f(\theta, x_i) \right)^2 \right] = -E \left(\frac{\partial^2}{\partial \theta^2} \log f(\theta, x_i) \right)$$

ESEMPIO · FISHER

Informazione di Fisher — Bernoulli

Calcolo dell'informazione per una Bernoulliana.

Sia θ il parametro di una variabile Bernoulliana e $(X_1, X_2, \dots, X_n)$ il suo campionamento casuale. Calcolare l'informazione di Fisher.



Informazione di Fisher di una Bernoulli

Densità, verosimiglianza e log-verosimiglianza:

$$p(x, \theta) = \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}, \theta \in (0, 1)$$

$$L(\theta, \underline{X}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$$

$$l(\theta, \underline{X}) = \sum x_i \log \theta + \left(n - \sum x_i \right) \log(1 - \theta)$$

Funzione di score (derivata prima) e derivata seconda:

$$\frac{\partial}{\partial \theta} \log L = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta}$$

$$\frac{\partial^2}{\partial \theta^2} \log L = -\frac{\sum x_i}{\theta^2} - \frac{n - \sum x_i}{(1 - \theta)^2}$$

Per l'informazione si cambia segno al valore atteso della derivata seconda ($E(x_i) = \theta$):

$$I_n(\theta) = \frac{\sum E(x_i)}{\theta^2} + \frac{n - \sum E(x_i)}{(1 - \theta)^2} = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1 - \theta)^2} = \frac{n}{\theta(1 - \theta)}$$

In alternativa, sfruttando l'addittività sulla singola osservazione:

$$I(\theta) = \frac{1}{\theta(1 - \theta)} \quad \Rightarrow \quad I_n(\theta) = n I(\theta) = \frac{n}{\theta(1 - \theta)}$$

SEZIONE 8 · CRAMÈR-RAO

Diseguaglianza di Cramèr-Rao

Un limite inferiore alla varianza degli stimatori corretti.

Considerato un problema regolare di stima, e sia T uno stimatore **non distorto** per θ , vale la diseguaglianza di Cramèr-Rao:

FORMULA 39 · LIMITE DI CRAMÈR-RAO

$$\text{VAR}(T) \geq \frac{1}{n I(\theta)}$$

Il risultato si generalizza alla stima di una funzione $\tau(\theta)$ del parametro (ad esempio $\tau(\theta) = \theta^2$):



FORMULA 40 · FORMA GENERALIZZATA

$$VAR(T) \geq \frac{[\tau'(\theta)]^2}{n I(\theta)}$$

i Uno stimatore non distorto la cui varianza **raggiunge** questo limite è detto efficiente.

APPENDICE

Proprietà di valore atteso e varianza

Regole usate nei calcoli. Valide per un campione i.i.d.

A1 · LINEARITÀ DEL VALORE ATTESO

$$E(a + bX) = a + bE(X)$$

A2 · VARIANZA DI UNA TRASFORMAZIONE LINEARE

$$VAR(a + bX) = b^2 VAR(X)$$

A3 · CAMPIONE I.I.D.

$$E(X_i) = \mu \quad VAR(X_i) = \sigma^2$$

A4 · VALORE ATTESO DI UNA SOMMA

$$E\left(\sum X_i\right) = \sum E(X_i)$$

A5 · VARIANZA DI UNA SOMMA (V.C. INDIPENDENTI)

$$VAR\left(\sum X_i\right) = \sum VAR(X_i) \quad (\text{indipendenti})$$

