Università di Firenze
Corso di Laurea in Statistica
Statistica

Leonardo Grilli

Stima

Cicchitelli cap. 16

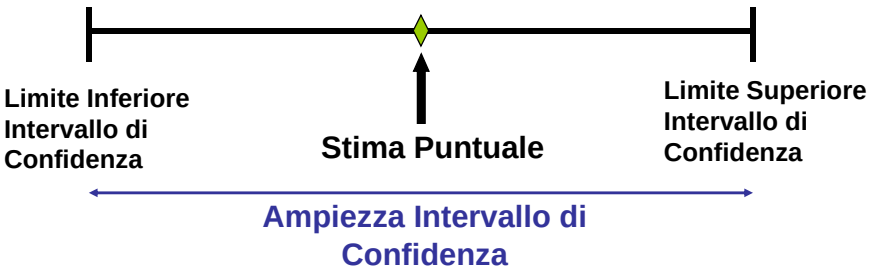
A.A. 2010/2011

Argomenti

- Definizione di stimatore
- Proprietà degli stimatori (campioni finiti):
  - Non distorsione
  - Efficienza relativa
- Margine di errore
- Proprietà asintotiche degli stimatori:
  - Non distorsione asintotica
  - Consistenza
- Stimatori di massima verosimiglianza

Stima Puntuale e per Intervallo

- Una **stima puntuale** è un unico valore
- Un **intervallo di confidenza** è un insieme di valori e fornisce ulteriori informazioni circa la variabilità



Inferenza

- Tipicamente l'inferenza riguarda alcuni **parametri**, cioè indici relativi alla distribuzione del carattere di interesse nella popolazione, es. la media, la mediana, la deviazione standard
- Alcuni dei metodi di inferenza che vedremo assumono che la distribuzione del carattere nella popolazione sia ben approssimata da una v.a. appartenente ad una certa famiglia parametrica (es. la Normale) → in tal caso tutto ciò che è incognito sono i **parametri della v.a.** (per la Normale: la media μ e la deviazione std. σ)
- I parametri sono supposti:
  - Incogniti
  - Fissati (non sono v.a. – approccio frequentista)

Statistica e stimatore /1

- **Statistica (campionaria)**: qualsiasi funzione delle v.a. che compongono il campione

$$T = t(X_1, X_2, \ldots, X_n)$$

- In particolare, una statistica che viene utilizzata per stimare un parametro θ viene detta **stimatore di θ**
- Es. la media campionaria è una statistica (definita dalla funzione che prescrive di sommare i valori e dividere per *n*); quando la si impiega per stimare la media della popolazione μ allora viene detta stimatore di μ

Statistica e stimatore /2

- POPOLAZIONE (numerosità *N*) → parametri (μ, σ, ...)
- CAMPIONE (numerosità $n<N$) → statistica ($\bar{X}$, S , ...)

Ogni quantità della popolazione (parametro) ha un suo analogo nel campione (statistica). Ad es. al parametro

> media del carattere nella popolazione, indicata con la lettera greca μ

corrisponde la statistica

> media del carattere nel campione, detta "media campionaria" e indicata con la lettera latina $\bar{X}$

E' naturale quindi cercare di stimare un parametro di interesse (es. μ_X) con la corrispondente statistica (cioè $\bar{X}$). Quando una statistica viene usata a fini inferenziali per stimare un parametro viene detta ***stimatore*** (es. $\bar{X}$ è uno stimatore di μ_X)

Stimare la media

- Tipicamente il parametro di interesse primario è la media
- Disponendo di un campione, lo stimatore naturale della media della popolazione è la **media campionaria**

Parametro:	Stimatore:	Stima:
μ: media della popolazione	$\bar{X}$: variabile aleatoria media campionaria (stimatore di μ_X)	$\bar{x}$: valore assunto dalla media campionaria nel particolare campione osservato (stima di μ_X)

Paradigma

- Le proprietà degli stimatori che vedremo d'ora in avanti fanno riferimento al paradigma dell'**inferenza parametrica** basata su **campione casuale**:

$$X \sim f(x;\theta) \quad e \quad X_1, \ldots, X_n \sim iid - X$$

Inferenza parametrica: f è una funzione (di massa o di densità di probabilità) nota a meno di un parametro θ (che può essere uno scalare, es. nella Bernoulli $\theta = p$, oppure un vettore, es, nella Normale $\theta = (\mu, \sigma)$)

Campione casuale: gli elementi campionari sono v.a. indipendenti e identicamente distribuite come X

Stimatori non distorti

- Non si può valutare se una specifica **stima** è buona o no
- Tuttavia si possono valutare le proprietà dello **stimatore**: in generale, cioè considerando tutti i possibili campioni, lo stimatore come si comporta?
- Una prima proprietà da valutare è la ***non distorsione*** (detta anche ***correttezza***): uno stimatore si dice ***non distorto*** o ***corretto*** quando il valore atteso dell'errore di stima (= errore di stima medio nell'insieme dei possibili campioni) è nullo
- *Qualunque sia la distribuzione del carattere nella popolazione*, la media campionaria è uno stimatore corretto della media della popolazione → in alcuni campioni sovrastima, in altri sottostima, ma nell'insieme dei campioni sovrastime e sottostime si compensano, per cui lo stimatore non ha una tendenza sistematica né alla sovrastima né alla sottostima

Stimatori non distorti

Formalmente la proprietà di non distorsione si scrive come

$$E(\bar{X} - \mu_X) = 0 \qquad \text{qualunque sia il valore di } \mu_X$$

In alternativa, per le proprietà del valore atteso si può scrivere anche come

$$E(\bar{X}) = \mu_X \qquad \text{qualunque sia il valore di } \mu_X$$

A parole: uno stimatore è non distorto quando il suo valore atteso coincide con il parametro di interesse (qualunque sia il suo valore)

Osservazione 1: $E(\ )$ è il valore atteso della distribuzione campionaria, cioè è il valore medio nell'insieme dei possibili campioni

Osservazione 2: la precisazione "qualunque sia il valore del parametro" è cruciale perché in pratica il valore del parametro è ignoto

Stimatori non distorti

Oltre alla media campionaria altri stimatori non distorti sono:

Proporzione campionaria (in realtà è un caso speciale di media campionaria)

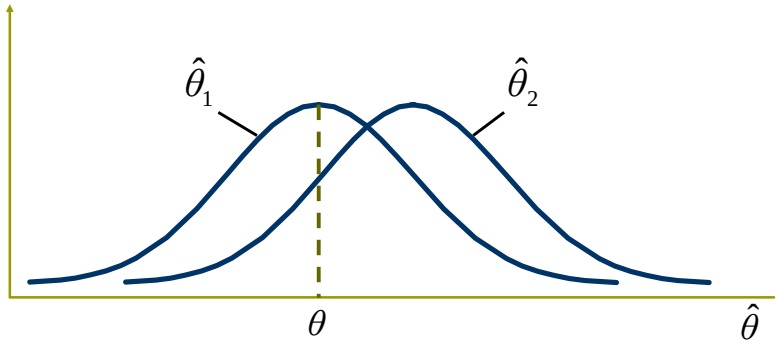
$$E(\hat{P}) = p \qquad \text{qualunque sia il valore di } p$$

Varianza campionaria (divisore *n*-1)

$$E(S^2) = \sigma_X^2 \qquad \text{qualunque sia il valore di } \sigma_X^2$$

Stimatori non distorti

- $\hat{\theta}_1$ è uno stimatore non distorto, $\hat{\theta}_2$ è distorto:



Distorsione

- Sia $\hat{\theta}$ uno stimatore per θ
- La distorsione di $\hat{\theta}$ è definita come la differenza tra la sua media e θ

$$\mathrm{D}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$$

- Es. per la varianza della popolazione

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad E\left(\tilde{S}^2\right) = \frac{n-1}{n}\sigma^2 \quad D\left(\tilde{S}^2\right) = -\frac{1}{n}\sigma^2$$

- La distorsione di uno stimatore non distorto è 0

Errore quadratico medio

- Dato uno stimatore T del parametro θ
  - Errore quadratico medio

$$EQM_\theta(T) = E\left[(T-\theta)^2\right] = Var_\theta(T) + \left[D_\theta(T)\right]^2$$

  - Errore standard

$$ES_\theta(T) = \sqrt{E\left[(T-\theta)^2\right]}$$

Efficienza relativa /1

- Consideriamo due stimatori T_1 e T_2 del parametro θ basati sullo stesso numero di osservazioni campionarie
- Lo stimatore T_1 si dice più efficiente dello stimatore T_2 se ha EQM minore, cioè

$$EQM_\theta(T_1) \le EQM_\theta(T_2)$$

per ogni θ (e vale < per almeno un valore di θ)

- Non esiste uno stimatore più efficiente in assoluto, cioè con EQM inferiore a tutti i possibili stimatori per tutti i possibili valori di θ
- In alcuni casi si può trovare lo stimatore più efficiente in una classe ristretta di stimatori, come gli stimatori non distorti o gli stimatori lineari

Efficienza relativa /2

- Dovendo scegliere tra due stimatori non distorti si preferisce quello più efficiente (= a varianza più bassa)
  - Es. se la distribuzione del carattere nella popolazione è Normale di media qualunque μ_X e varianza qualunque σ_X^2 sia la media campionaria $\bar{X}$ che la mediana campionaria *M* sono stimatori non distorti di μ_X (infatti nella Normale μ_X è sia la media che la mediana). Tuttavia la media campionaria è preferibile in quanto più efficiente: infatti si dimostra che

$$EQM(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} \qquad EQM(M) = \sigma_M^2 \sim 1.57\frac{\sigma_X^2}{n}$$

$$\text{Efficienza Relativa} = \frac{EQM(M)}{EQM(\bar{X})} \sim 1.57$$

La mediana ha un EQM che supera quello della media del 57%

Errore di stima e margine di errore

- Stima della media di una popolazione tramite un campione casuale di dimensione *n*
  - Se *X*~Normale oppure se *n* è grande l'*errore di stima* ha la seguente distribuzione (esatta oppure approssimata)

$$\bar{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

  - Allora l'errore di stima è limitato in senso probabilistico:

$$P\left(\left|\bar{X} - \mu\right| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$ *margine di errore*

Errore standard vs margine di errore

- Supponiamo μ = temperatura media di fusione di una nuova lega metallica, espressa in C°
- Media su un campione casuale di 25 esemplari
- Ipotizziamo σ = 1.5 C°
- **Errore standard**

$$ES(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{25}} = 0.3$$

Mediamente si commette un errore di 0.3 C°

- **Margine di errore al 95%**

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96\frac{1.5}{\sqrt{25}} = 0.588$$

Nel 95% dei campioni si commette un errore inferiore a 0.588 C°

Proprietà: campioni finiti vs asintotiche

- La non distorsione e l'efficienza sono proprietà che valgono a prescindere dalla dimensione campionaria, cioè valgono *per qualunque n* (proprietà per campioni finiti)
- Altre proprietà valgono in campioni infinitamente grandi, cioè per $n \to \infty$ (proprietà asintotiche)
  - Non distorsione asintotica
  - Efficienza asintotica
  - Consistenza
- Le proprietà asintotiche riguardano una successione di stimatori per cui si usa il pedice *n*

$$T_n = t_n(X_1, X_2, \ldots, X_n)$$

Non distorsione asintotica

- Uno stimatore T_n del parametro θ si dice **asintoticamente non distorto** se la sua distorsione tende a 0 per $n \to \infty$

$$\lim_{n\to\infty} D(T_n) = 0 \quad \forall\theta \qquad \lim_{n\to\infty} E(T_n) = \theta \quad \forall\theta$$

- Ovviamente uno stimatore non distorto è anche asintoticamente non distorto (es. la media campionaria, la varianza campionaria)
- La varianza della popolazione è uno stimatore distorto in campioni finiti ma non asintoticamente

$$\lim_{n\to\infty} D\left(\tilde{S}_n^2\right) = \lim_{n\to\infty} -\frac{1}{n}\sigma^2 = 0 \quad \forall\sigma^2$$

Consistenza /1

- La consistenza è una **proprietà asintotica**, cioè riguarda il comportamento di uno stimatore al crescere della dimensione campionaria
- Se la popolazione ha dimensione finita *N* e si campiona senza ripetizione, quando l'ampiezza campionaria *n* raggiunge *N* il campione coincide con la popolazione → ogni stimatore "sensato" stima alla perfezione il parametro di interesse (es. quando $n=N$ → media campionaria = media della popolazione)
- Cosa accade se la popolazione è infinita ($N=\infty$)? In tal caso *n* non può raggiungere *N* (il campione non può avere ampiezza infinita) e quindi non è possibile avere stime perfette
- Tuttavia se lo stimatore è consistente si possono avere stime quasi perfette, perché al crescere dell'ampiezza campionaria gli errori di stima diventano sempre più piccoli

Consistenza /2

- Uno stimatore (più precisamente, una successione di stimatori) del parametro θ

$$T_n = t_n(X_1, X_2, \ldots, X_n)$$

si dice **consistente** se per ogni $\varepsilon > 0$, piccolo a piacere, la probabilità che l'errore di stima in valore assoluto sia inferiore a ε tende a 1 al crescere della dimensione campionaria *n*

$$\lim_{n\to\infty} P\left(\left|T_n - \theta\right| < \varepsilon\right) = 1 \qquad \forall\, \theta \in \Theta$$

Consistenza /3

- Uno stimatore è consistente quando l'errore di stima può essere reso piccolo a piacere aumentando la dimensione campionaria
- In altri termini: più grande è il campione, più preciso è lo stimatore
- La consistenza è un requisito minimo di uno stimatore: uno stimatore non consistente non dovrebbe essere usato

Consistenza /4

- Affinché uno stimatore T_n di un parametro θ sia consistente è *sufficiente* valgano entrambe le seguenti condizioni:
  1. T_n è asintoticamente non distorto $\lim_{n\to\infty} E(T_n) = \theta \quad \forall\, \theta$
  2. T_n ha varianza campionaria che tende a zero al crescere di *n*

$$\lim_{n\to\infty} Var(T_n) = 0 \quad \forall\, \theta$$

Questo significa che al crescere di *n* la distribuzione di *T* diviene sempre più concentrata attorno al parametro di interesse θ, quindi lo stimatore è sempre più preciso.

La media campionaria è uno stimatore consistente perché: 1. è non distorto per qualunque *n* e quindi anche per $n\to\infty$ 2. la sua varianza è σ_X^2/n e quindi tende a 0 per $n\to\infty$

Metodi per definire gli stimatori

- Uno stimatore è una funzione t degli elementi campionari che si utilizza per stimare un parametro θ: $T = t(X_1,\ldots,X_n)$
- Come scegliere la funzione t in modo da stimare bene θ? (bene significa che lo stimatore ha buone proprietà: non distorto, consistente, efficiente ecc.)
- Vi sono infinite possibilità di scegliere la funzione t ma le migliori di solito sono fornite dai metodi:
  - Metodo dell'analogia tra stimatore e parametro
  - Metodo dei minimi quadrati
  - Metodo dei momenti
  - Metodo della massima verosimiglianza

Analogia stimatore/parametro

- Gli stimatori definiti fino ad ora sono derivati dall'analogia tra lo stimatore stesso e il parametro di interesse
  - Per stimare la media della popolazione si usa la media campionaria
  - Per stimare la proporzione della popolazione si usa la proporzione campionaria
- Attenzione: talvolta l'analogo campionario non è la scelta migliore
  - Per stimare la varianza della popolazione si può usare la varianza descrittiva calcolata sul campione (divisore n) ma tale stimatore è distorto → è preferibile usare la varianza con il divisore n-1
- Attenzione: talvolta esistono più analoghi campionari
  - Per stimare il parametro μ di una distribuzione Normale si può usare sia la media campionaria che la mediana campionaria (ma abbiamo visto che la media è più efficiente della mediana)