La stima puntuale

Argomenti trattati:

- Stima puntuale e stimatore
- Proprietà degli stimatori
- Stima puntuale della media della popolazione e sua distribuzione
- Stima puntuale di una proporzione e sua distribuzione
- Stima puntuale della varianza della popolazione e sua distribuzione

Contenuti del Capitolo 11 del libro di testo Borra-Di Ciaccio

INTRODUZIONE

Nella prima parte sono stati forniti tutti gli strumenti per completare ora tutti i passaggi alle procedure di INFERENZA STATISTICA:

i principali modelli che descrivono i più importanti tipi di fenomeni statistici (gaussiano, bernoulliano-binomiale, poissoniano, ecc.) e i principali strumenti statistici (media, varianza, mediana di una distribuzione, ecc.).

In questa parte:

si assumerà che parte del modello è incognito e interesserà ricostruirlo sulla base delle informazioni provenienti da un campione estratto dalla Popolazione descritta dal modello.

La procedura di inferire dai dati campionari il valore assunto da alcune caratteristiche della Popolazione viene chiamato appunto "inferenza statistica".

ESEMPIO

Potrebbe interessare la stima della domanda **media** settimanale di una particolare marca di un certo prodotto; la stima della **proporzione** dei dipendenti di un'azienda a favore di un nuovo piano di incentivi; etc

Attraverso la cosiddetta **stima puntuale** si ha l'obiettivo di stimare un parametro incognito (una media, una proporzione, la varianza, in generale una quantità caratteristica) della Popolazione attraverso un singolo valore a partire dalle informazioni (***risultanze***) provenienti da un campione casuale.

Ogni **INFERENZA** relativa ad una popolazione è basata su *statistiche campionarie* ossia funzioni dei dati campionari.

La scelta delle statistiche più opportune dipende dal parametro di cui occorre produrre una stima.

Il valore del parametro di P non è noto ed uno degli obiettivi del campionamento è appunto quello di riuscire a stimarne il valore

Sia *X* una v.c. che rappresenta un carattere osservato su una popolazione e sia **θ** un parametro incognito della Popolazione.

Sia $X_1, X_2, \ldots X_N$ un campione casuale di dimensione *N* e $x_1, x_2, \ldots x_n$ il corrispondente campione osservato di dimensione *h*.

Ricordiamo che:

Le X_i sono variabili i.i.d. con la stessa distribuzione di probabilità di X, cioè n copie di X: fino a che non verrà estratto un campione queste X_i possono assumere uno qualsiasi dei valori assumibili dalla variabile X. Una volta estratto il campione avremo a disposizione n valori numerici. Se X rappresenta l'altezza degli individui in una P, allora (X_1,…,X_n) sono le n altezze possibili che possiamo rilevare su n individui. Una volta selezionati n individui otteniamo n valori numerici (x_1,…x_n) che sono le altezze di quegli n individui.

Obiettivo:

Ottenere attraverso un'opportuna funzione (statistica) delle osservazioni

$$t = t(x_1, x_2, \ldots x_n)$$

una stima di **θ**

E' importante la differenza tra Stima (puntuale) e Stimatore

La stima $t = t(x_1, \ldots, x_n)$ può essere considerata come una realizzazione della variabile casuale $T = t(X_1, \ldots, X_n)$ chiamata **stimatore** di θ

Variabili casuali $X_1, \ldots, X_n$ *campione casuale*	Variabile casuale $T = t(X_1, \ldots, X_n)$ *stimatore*
↓	↓
Realizzazioni $x_1, \ldots, x_n$ *campione osservato*	Realizzazione $t = t(x_1, \ldots, x_n)$ *stima*
(Insieme di n valori: i *dati del campione*)	(Singolo valore: la stima di θ)

Esempio: campione osservato (2,5,3,6,4,4,1,2,2,5)

Parametro: $\theta = \mu$ media della popolazione.

Stimatore: media campionaria $\overline{X} = \frac{1}{10} \sum_{i=1}^{10} X_i$

Stima: $\overline{x} = 3{,}4$

esempio

Consideriamo le vendite medie settimanali di un prodotto di una certa marca:

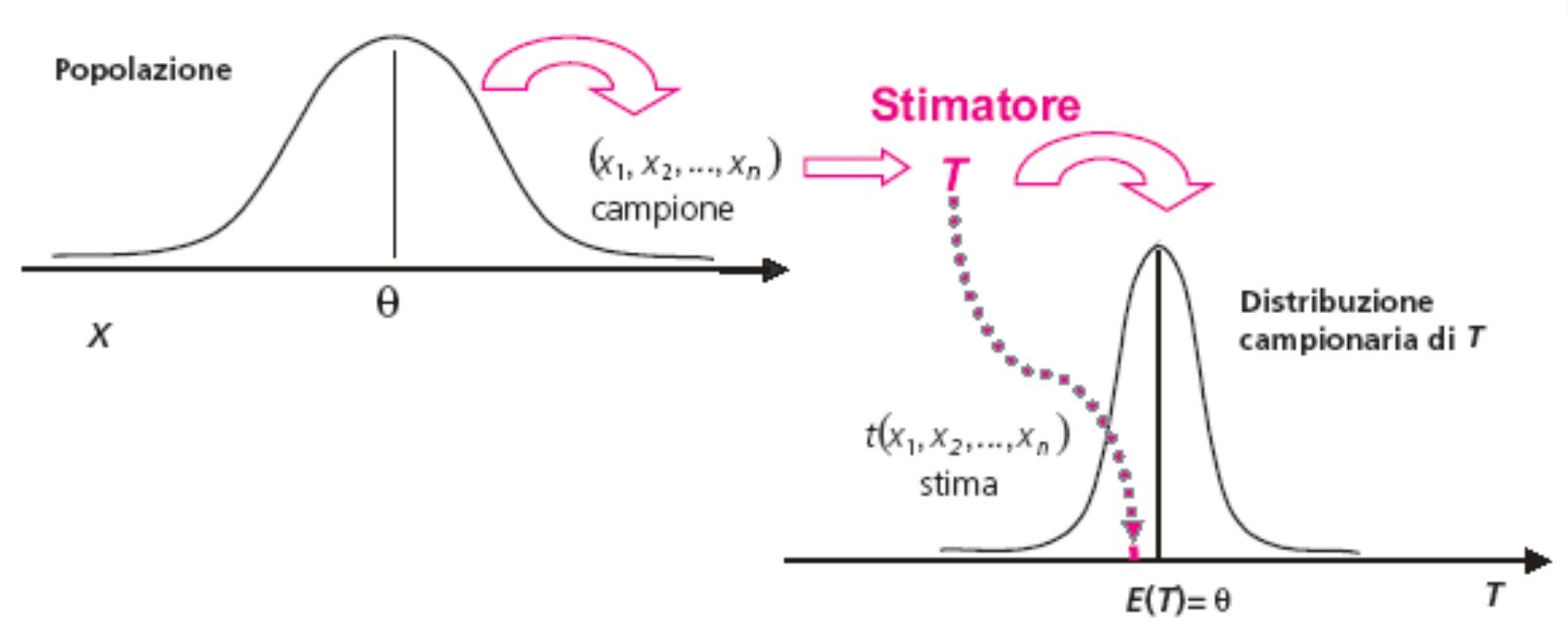
*un possibile stimatore della media della suddetta P è la media campionaria: se si rileva che la media di un campione casuale di vendite settimanali è pari a **xxxx**, tale valore è una **stima** delle vendite medie settimanali in P. Un altro possibile **stimatore** potrebbe essere la mediana campionaria.*

→ Una stima puntuale è il valore che uno stimatore assume in corrispondenza di una particolare realizzazione campionaria

Es. la media campionaria $\bar{X}$ è uno stimatore puntuale della media μ della popolazione; il valore che $\bar{X}$ assume in corrispondenza di un certo campione viene detto stima puntuale $\bar{x}$

Stima puntuale e stimatore

Lo stimatore, dipendendo dal campione, è una variabile casuale e quindi possiede una **distribuzione campionaria** la cui conoscenza permette di capire se lo stimatore scelto produrrà con elevata probabilità stime "vicine" al valore vero del parametro.



Non esiste infatti un unico meccanismo per determinare lo stimatore puntuale “migliore”: esistono però alcuni criteri che permettono di valutare gli stimatori.

Ad esempio anche la mediana campionaria è uno stimatore puntuale per la media della popolazione μ, tuttavia generalmente la mediana non è lo stimatore migliore.

Le tre più importanti proprietà per valutare gli stimatori sono: **non distorsione, consistenza ed efficienza**

Proprietà degli stimatori

Per valutare la "bontà" di uno stimatore T si può guardare alle sue proprietà:

Proprietà per n finito:

- Correttezza
- Efficienza

Proprietà per $n \to +\infty$ (asintotiche):

- Consistenza
- Correttezza asintotica

Proprietà degli stimatori

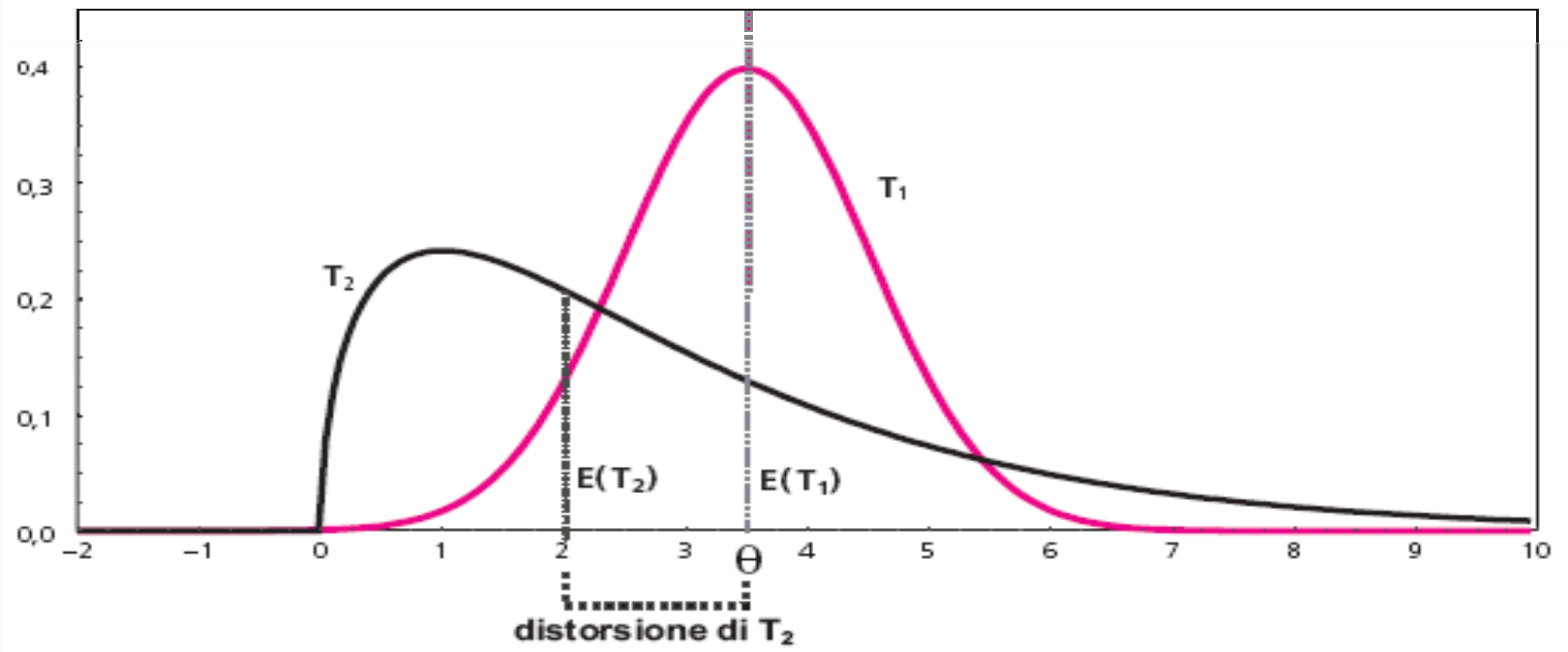
Correttezza

Lo stimatore T è uno stimatore **corretto** di θ se

$$E(T) = \theta$$

per tutti i possibili valori di θ

Se uno stimatore è distorto, la **distorsione** di uno stimatore è misurata da: $B(T) = E(T) - \theta$



ATTENZIONE:

Non distorsione non significa che un particolare valore di T debba essere uguale a θ, anzi T talvolta potrà sovrastimare e altre volte sottostimare θ.

Piuttosto **uno stimatore non distorto ha la capacità di stimare correttamente in media un parametro di P: uno stimatore non distorto è non distorto in media**.

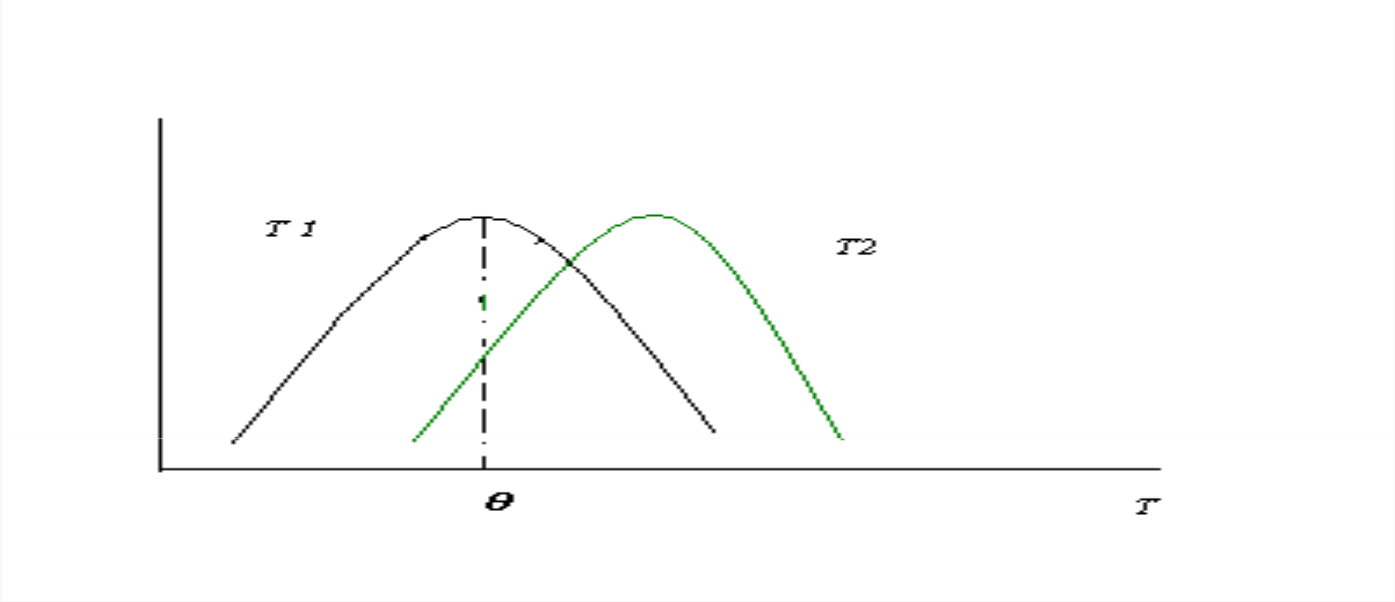
Il valore atteso di T può essere pensato come la media dei valori che T assumerebbe in corrispondenza di tutti i possibili campioni estraibili da P, o in alternativa alla *media* dei valori di T in un numero molto elevato di campioni ripetuti.

Quindi:

talvolta T sovrastimerà e altre volte sottostimerà il parametro ma dalla definizione di valore atteso segue che ripetendo la procedura di campionamento molte volte si otterrà che in media lo stimatore non distorto sarà uguale al parametro della P.

A parità di altre condizioni la proprietà di non distorsione è una proprietà ragionevolmente auspicabile per uno stimatore puntuale.

Funzioni di densità di probabilità di due stimatori: T_1 è non distorto per θ, T_2 non lo è



La media campionaria, la varianza campionaria e la proporzione campionaria sono stimatori non distorti dei corrispondenti parametri della popolazione:

$$E(\bar{X}) = \mu; \qquad E(S^2) = \sigma^2; \qquad E(\hat{P}) = p$$

Proprietà degli stimatori: efficienza

In molti problemi pratici si possono ottenere più stimatori non distorti per uno stesso parametro: *ad esempio se la P segue una distribuzione Normale sia la media campionaria sia la mediana campionaria sono stimatori non distorti per la media della popolazione.*

E' necessario un metodo per individuare il migliore:

è naturale preferire lo stimatore con la distribuzione il più possibile ***concentrata*** intorno al parametro da stimare; in tal caso infatti i valori dello stimatore si scosteranno dal parametro oggetto di stima con minor probabilità rispetto a quanto potrebbe succedere negli altri casi.

Per valutare la prossimità di T a θ possiamo usare l'**errore quadratico medio** (*mean square error*) dato dalla quantità:

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

Proprietà:

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

dove $Var(T) = E[T - E(T)]^2$

Diremo che T_1 è più efficiente di T_2 se

$$MSE(T_1) < MSE(T_2)$$

Per tutti i possibili valori di θ.

Se lo stimatore è corretto e quindi è nulla la distorsione si ha:

$$MSE(T) = Var(T)$$

per tutti i possibili valori di θ .

Dati due stimatori corretti T_1 e T_2 , si dirà che T_1 è più efficiente di T_2 se

$$Var(T_1) < Var(T_2)$$

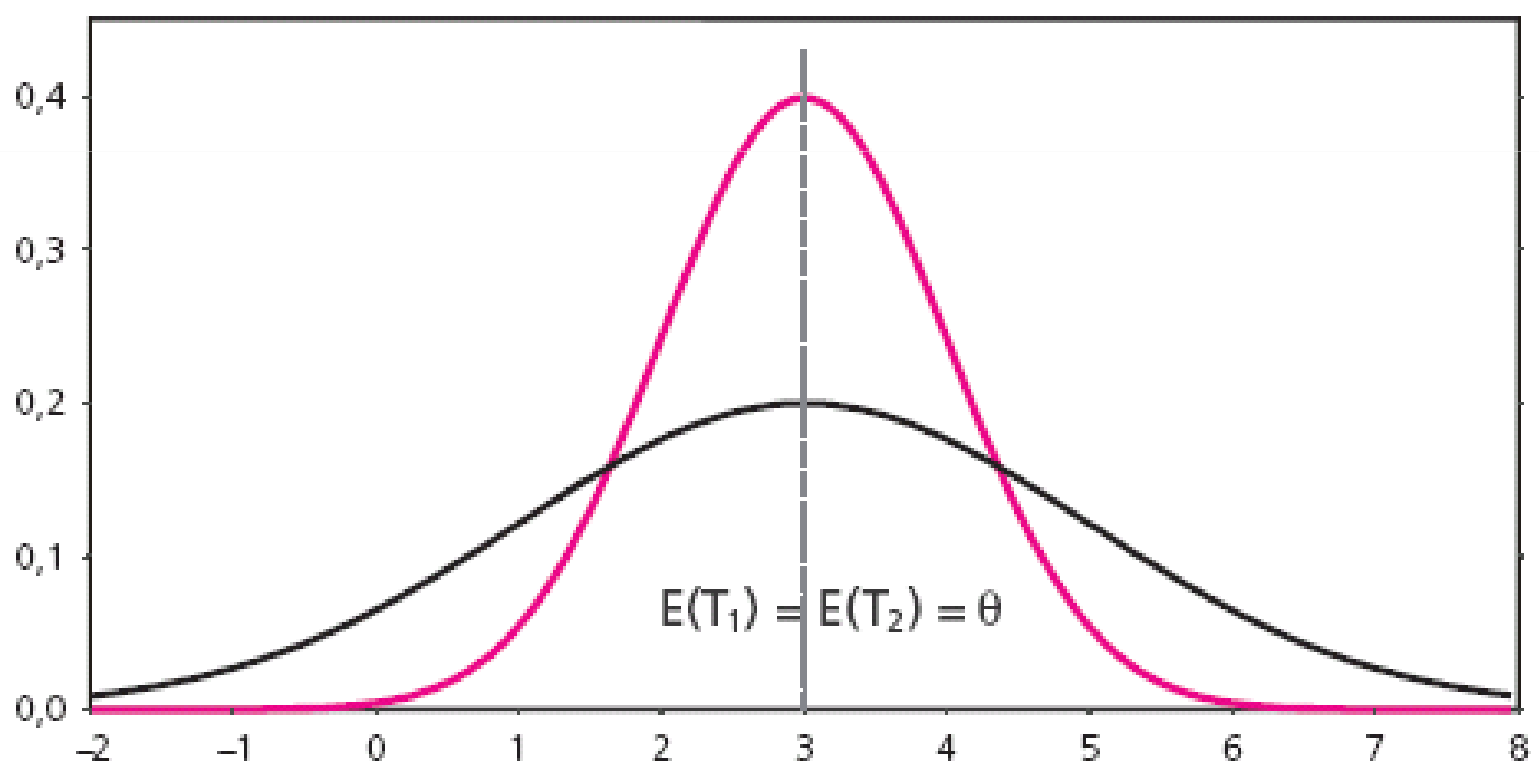
Per tutti i possibili valori di θ.

⟹ **se ci sono più stimatori non distorti per uno stesso parametro, è detto più efficiente quello che ha la varianza più piccola**.

La **efficienza relativa** di T_1 rispetto a T_2 è il rapporto tra le loro varianze:

$$efficienza\ relativa = \frac{\text{var}(T_2)}{\text{var}(T_1)}$$

Nella figura sono riportate le distribuzioni campionarie di due stimatori corretti. lo stimatore T_1 (linea rossa) possiede un errore quadratico medio (ossia una varianza) più piccolo di T_2 (linea nera).



Esempio 1

Sia $x_1, \ldots, x_n$ un campione casuale estratto da una Popolazione Normale con media μ e varianza σ^2. Quale stimatore preferire tra la Media e la Mediana Campionaria se si deve stimare la media della popolazione?

Sono ambedue stimatori non distorti:

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

Ma:

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n} \quad \text{mentre} \quad \text{var}(Me) = \frac{\pi}{2} \times \frac{\sigma^2}{n} = \frac{1.57\sigma^2}{n}$$

La media campionaria è più efficiente della mediana campionaria e l'efficienza relativa della prima rispetto alla seconda è pari a:

$$efficienza\ relativa = \frac{\text{var}(Me)}{\text{var}(\bar{X})} = 1.57$$

La var di Me supera del 57% quella della media. Afinchè la Me abbia una varianza pari a quella della Media dovrebbe essere basata sul 57% di osservazioni in più.

Esempio 2

Si consideri il campione $X_1, \ldots, X_4$ estratto da una popolazione con media μ e varianza σ^2, e i due stimatori:

$$T_1 = \bar{X} \qquad T_2 = \frac{X_1 + 2X_2 + 3X_3 + X_4}{4}$$

Allora si ha:

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

$$Var(T_1) = Var(\bar{X}) = \sigma^2/4$$

$$E(T_2) = (7/4)\mu$$

$$B(T_2) = (7/4)\mu - \mu = (3/4)\mu$$

$$Var(T_2) = (15/16)\sigma^2$$

$$MSE(T_1) = \sigma^2/4 \quad < \quad MSE(T_2) = (15/16)\sigma^2 + (9/16)\mu^2$$

Proprietà degli stimatori: Consistenza

Uno stimatore puntuale T è definito **stimatore consistente** di θ di P se la differenza tra il valore atteso dello stimatore e il parametro da stimare diminuisce al crescere dell'ampiezza del campione: ossia all'aumentare della ampiezza del campione, la distorsione diventa sempre più piccola.

Non tutti gli stimatori non distorti sono consistenti, così come non tutti gli stimatori consistenti sono non distorti.

Ad esempio se la varianza campionaria fosse calcolata come

$$\tilde{S}^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

Si tratterebbe di uno stimatore distorto per la varianza di P. Sarebbe però uno stimatore consistente, perché approssima lo stimatore non distorto $\tilde{S}^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$ al crescere della dimensione del campione.

⟹ **l'uso di uno stimatore consistente accompagnato da una quantità infinita di informazioni campionarie fornisce un risultato non distorto. Viceversa l'uso di uno stimatore non consistente fornisce un risultato distorto, anche qualora si disponga di una quantità infinita di informazioni campionarie**

Più in generale:

Consistenza

Lo stimatore T_n di un parametro θ, dove l'indice indica la dipendenza dello stimatore dalla numerosità campionaria, è uno stimatore **consistente in media quadratica** se

$$\lim_{n\to\infty} MSE(T_n) = \lim_{n\to\infty} E(T_n - \theta)^2 = 0$$

Quindi

$\lim_{n\to\infty} MSE(T_n) = 0$ se e solo se $\lim_{n\to\infty} Var(T_n) = 0$ $\lim_{n\to\infty} B(T_n) = 0$

Allora uno stimatore **corretto** è consistente se:

$$\lim_{n\to\infty} MSE(T_n) = \lim_{n\to\infty} Var(T_n) = 0$$

Proprietà degli stimatori

Correttezza asintotica

Uno stimatore T_n di un parametro θ è uno **asintoticamente corretto** se:

$$\lim_{n\to\infty} \mathrm{B}(T_n) = 0$$

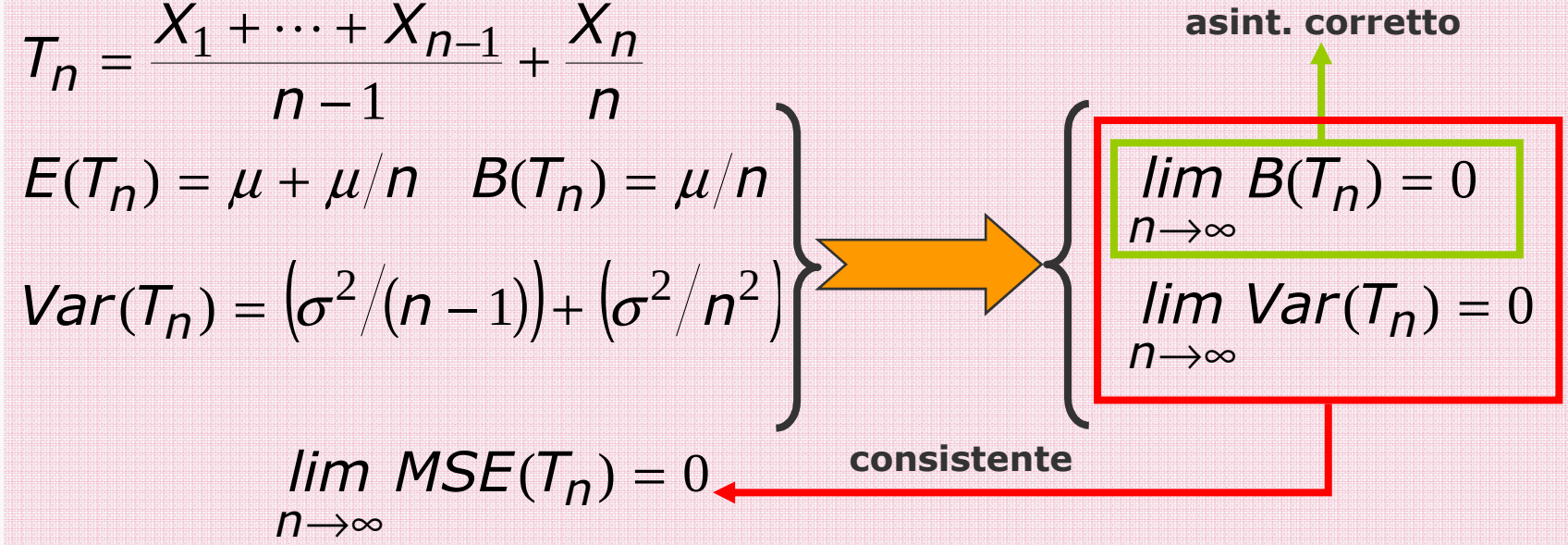
per ogni possibile valore di θ

Esempio

$$T_n = \frac{X_1 + \cdots + X_{n-1}}{n-1} + \frac{X_n}{n}$$

$$E(T_n) = \mu + \mu/n \quad B(T_n) = \mu/n$$

$$Var(T_n) = \left(\sigma^2/(n-1)\right) + \left(\sigma^2/n^2\right)$$



Stima puntuale della media della popolazione

Si consideri una popolazione X con media μ e varianza σ^2

- La media campionaria $\overline{X}$ è uno stimatore corretto per la media della popolazione, ossia $E(\overline{X}) = \mu$

- La varianza della media campionaria è $Var(\overline{X}) = \sigma^2 / n$ pertanto è uno stimatore consistente, poiché

$$\lim_{n\to\infty} MSE(\overline{X}_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

- Se la popolazione è distribuita come una Normale, $N(\mu, \sigma^2)$ allora anche la media campionaria si distribuisce come una Normale

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

Verifichiamo:

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

$$\text{è:} \quad E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

2.

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

$$\text{è:} \quad Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Oss. 1. la precedente varianza della somma uguale alla somma delle varianze è vera perché consideriamo una campione casuale

Oss.2. la varianza della distribuzione della media campionaria diminuisce all'aumentare della dimensione del campione: più grande è il campione meno dispersa è la distribuzione campionaria

⇒ campioni più grandi determinano una maggiore certezza nell'inferenza, in quanto si ottengono con essi maggiori informazioni su P e le sue caratteristiche.

Oss. 3 se la media campionaria segue una distribuzione normale allora possiamo standardizzarla e:

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

Qualora la distribuzione di partenza non dovesse essere normale il precedente risultato continua a valere per n sufficientemente grande in virtù del teorema del limite centrale

Stima puntuale della proporzione della popolazione

Si consideri una popolazione X distribuita come una Bernoulli con parametro π.

1) La media campionaria $\bar{X}$ è uno stimatore corretto della proporzione π della popolazione, ossia

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

2) La varianza della media campionaria è $Var(\bar{X}) = \pi(1-\pi)/n$ pertanto è uno stimatore consistente, poiché

$$\lim_{n\to\infty} MSE(\bar{X}_n) = \lim_{n\to\infty} \frac{\pi(1-\pi)}{n} = 0$$

Si noti che la media campionaria nel caso di v.c. di Bernoulli altri non è se non la frequenza relativa dei successi riscontrata nel campione $\hat{P}$

3) Se il campione ha ampiezza elevata risulta che:

$$Z = \frac{\hat{P} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \; \square \; N(0,1)$$

approssimativamente.

Normalmente l'approssimazione è buona se

$n\pi(1-\pi)>9$

Stima puntuale della varianza della popolazione

Si consideri una popolazione X con media μ e varianza σ^2 entrambe ignote.

Si definisce varianza campionaria corretta lo stimatore:

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

1) S^2 è uno stimatore corretto della varianza della popolazione ossia $E(S^2) = \sigma^2$

2) S_n^2 è uno stimatore consistente per σ^2 , ossia

$$\lim_{n\to\infty} MSE(S_n^2) = 0$$

3) inoltre nel caso in cui la popolazione è normale di varianza σ^2 si può dimostrare che

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

dove *n-1* sono i g.d.l (o valori indipendenti)

Poiché $E(\chi^2) = n-1 \quad \text{e} \quad \text{Var}(\chi^2) = 2(n-1)$

Segue che:

$$E\left(\frac{(n-1)S^2}{\sigma^2}\right) = (n-1)$$

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

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

Proprietà di alcuni stimatori puntuali

Proprietà degli stimatori puntuali più usati.

Parametro della Popolazione	Stimatore puntuale	Proprietà
Media, μ	$\overline{X}$	Non distorto, consistente, più efficiente (sotto ipotesi di normalità)
Media, μ	***Me***	Non distorto (sotto ipotesi di normalità), ma non il più efficiente
Proporzione, p	$\hat{P}$	Non distorto, consistente, più efficiente
Varianza, σ^2	S^2	Non distorto, consistente, più efficiente (sotto ipotesi di normalità)

Osservazioni ed esempio

Un aspetto importante della media campionaria $\overline{X}$ è che la sua varianza decresce all'aumentare della dimensione campionaria *n (n° di osservazioni)*; quindi al crescere di n essa tende ad assumere valori in un intorno del valore vero della media μ di ampiezza sempre più piccola.

Quando si ha a che fare con un campione si deve sempre fornire oltre al valore della stima anche lo *standard error* associato a questa che ci da una misura della **precisione** della stima stessa.

Esempio

Supponiamo che sulla base di un campione si ha che una percentuale stimata p=50.5% di preferenze accordate ad un partito A rispetto ad un partito B.

Possiamo concludere che A vincerà le elezioni?

No perché ci manca l'informazione sulla precisione del dato 50.5%.

Calcoliamo l'errore standard:

la varianza della proporzione è $\pi(1-\pi)/n$.

Ci mancano 2 informazioni: 1. la dimensione campionaria; 2. π (d'altronde se lo conoscessimo non avremmo bisogno di stimarlo!).

Possiamo utilizzare il valore stimato nel campione *p* come stima del valore incognito π.

Utilizziamo due valori per *n:* 100 e 1000

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.505 \cdot 0.495}{100}} = \sqrt{0.0025} = 0.05 = 5\%$$

mentre

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.505 \cdot 0.495}{1000}} = \sqrt{0.00025} = 0.0158 = 1.6\%$$

Ora se consideriamo l'intervallo: $p \pm \sqrt{\frac{p \cdot (1 - p)}{n}}$

abbiamo nei due casi:

$$(50.5\% - 5\%, 50.5\% + 5\%) = (45.5\%, 55.5\%)$$

e

$$(50.5\% - 1.6\%, 50.5\% + 1.6\%) = (48.9\%, 52.1\%)$$

e nel secondo caso il risultato che annuncia la vittoria di A è più attendibile

Per valutare la precisione della stima ci si è rifatti ad un risultato già visto nel caso di una v.c. Gaussiana ossia quando abbiamo mostrato che:

$$P(\mu - \sigma < X < \mu + \sigma) = 68\%$$

in un intorno della media di lunghezza σ si trovano il 68% di tutti i possibili risultati della v.c. Normale.

L'intervallo μ+/-σ è l'intervallo di riferimento che si usa per valutare la qualità delle nostre stime.

Per **precisione** si intende il reciproco dello scarto quadratico medio (o della varianza): più questa è elevata più i dati presentano variabilità e quindi ci si deve aspettare una perdita di precisione e viceversa.