

1

GLI INTERVALLI DI CONFIDENZA

Argomenti trattati:

- Stima per intervallo
- Analogie tra la stima puntuale e per intervallo
- Intervallo di confidenza per la media
- Intervallo di confidenza per la proporzione
- Intervallo di confidenza per la varianza
- Determinazione della numerosità campionaria

2 **PERCHE' LA STIMA PER INTERVALLO ?**

a) Abbiamo detto che l'inferenza statistica è il processo attraverso cui i risultati campionari vengono utilizzati per trarre conclusioni sulle caratteristiche di una Popolazione.

Lo stimatore puntuale è una singola statistica che viene usata per stimare il vero valore (incognito) di un parametro θ di una P (es. $\bar{X} \rightarrow \mu$ oppure $S^2 \rightarrow \sigma^2$).

Tuttavia, nonostante sia possibile individuare stimatori che godano di proprietà ottimali, sappiamo che nella pratica si seleziona un solo campione e che una statistica varia da campione a campione e perciò dipende dagli elementi che vengono selezionati; di ciò va necessariamente tenuto conto quando con un solo valore numerico cerchiamo di stimare una caratteristica incognita di P.



Pertanto è più opportuno per garantirsi contro eventuali deviazioni, positive e negative, dal vero valore, fornire un "intervallo di valori" per la stima del parametro incognito

b) Nel campionamento da una popolazione, a parità di condizioni, è ragionevole ritenere che ad una conoscenza più approfondita di P si giunga usando campioni più grandi. Ciò non è rispecchiato dalla stima puntuale (a meno che come abbiamo visto non venga accompagnata dall'errore standard come nel caso della media campionaria).

Ad esempio la stima puntuale della proporzione π di pezzi difettosi in un carico sarebbe la stessa sia osservando 1 pezzo in un campione di 10 pezzi, sia osservando 100 pezzi difettosi in un campione di 1000 pezzi.

La maggiore **precisione** delle nostre informazioni sul parametro della popolazione viene esplicitamente evidenziata nelle stime per intervallo. Infatti come vedremo, a parità delle altre condizioni, campioni più grandi determinano stime per intervallo più precise, riflettendo in tal modo la minore incertezza sull'effettivo valore del parametro incognito considerato.

 STIMATORE PER INTERVALLO

Uno **stimatore per intervallo** per un parametro di P è una funzione delle variabili campionarie (campione casuale): determina gli estremi di un intervallo di valori che "verosimilmente" contiene il parametro da stimare. La stima corrispondente viene chiamata **stima per intervallo**.

Ma che cos'è e come si perviene ad una stima per intervallo?

5 **Stima per intervallo**

Supponiamo di aver estratto un campione casuale da P e che sulla base delle informazioni campionarie sia possibile determinare due v. casuali:

$$L_1 = L_1(X_1, \dots, X_n) \quad \text{ed} \quad L_2 = L_2(X_1, \dots, X_n)$$

con $L_1 < L_2$.

In corrispondenza di un determinato campione siano l_1 ed l_2 i valori assunti da L_1 ed L_2 allora l'una delle due: l'intervallo da l_1 a l_2 o contiene o non contiene il parametro stimato.

Non avendo indagato l'intera P non possiamo tuttavia saperlo. Ipotizziamo allora di estrarre ripetutamente dei campioni casuali da P e di determinare gli intervalli corrispondenti a tutti i diversi L_1 e L_2 .

6 Stima per intervallo

Alla fine solo una certa percentuale di questi intervalli (ad es. il 95% o il 98%) conterrà il valore incognito.

In base alla definizione frequentista di probabilità si può dare la seguente interpretazione degli intervalli trovati:

*Se si estraggono successivamente più campioni indipendenti dalla stessa popolazione e si determinano i relativi intervalli (**di confidenza**) nel modo ora illustrato, alla fine il 95% (o un'altra percentuale qualsiasi) di intervalli conterrà il vero valore del parametro incognito.*

L'intervallo $[L_1, L_2]$ è definito uno stimatore per intervallo a **livello di confidenza** 95% per il parametro

Stima per intervallo

Sia X una v.c. che rappresenta un carattere osservato su una popolazione. Supponiamo che la v.c. sia definita da una funzione di probabilità $f(x; \theta)$ dipendente dal parametro incognito θ .

Sia $X_1, \dots, X_n$ un campione di dimensione n e $x_1, \dots, x_n$ il corrispondente campione osservato.

Obiettivo:

Determinare due statistiche campionarie:

$$L_1 = L_1(X_1, \dots, X_n)$$

$$L_2 = L_2(X_1, \dots, X_n)$$

tali che $L_1 \leq L_2$ per ogni possibile campione e che l'intervallo $[L_1, L_2]$ contenga il parametro θ con probabilità $1 - \alpha$

ossia tali che $P(L_1 \leq \theta \leq L_2) = 1 - \alpha$

8 Stima per intervallo

L'intervallo casuale $[L_1(X_1, \dots, X_n), L_2(X_1, \dots, X_n)]$ si definisce **intervallo di confidenza** di livello $1 - \alpha$ per il parametro θ se contiene con probabilità $1 - \alpha$ il parametro ignoto θ della popolazione, ossia:

$$Pr[L_1(X_1, \dots, X_n) \leq \theta \leq L_2(X_1, \dots, X_n)] = 1 - \alpha$$

In genere si fissano valori di $1 - \alpha$ pari a 0,99; 0,95; 0,90 e questo viene detto **livello di confidenza**. Una volta estratto il campione si ottiene **l'intervallo di confidenza stimato** $[l_1, l_2]$.

OSSERVAZIONE:

Non è possibile sapere se l'intervallo stimato contenga o meno il valore vero del parametro; d'altra parte se si estraesse dalla popolazione un numero sufficientemente elevato di campioni ripetutamente ed in modo indipendente, e calcolassimo i corrispondenti intervalli di confidenza, il vero valore del parametro θ sarà contenuto nel circa il $100(1-\alpha)\%$ di questi intervalli.

 **il livello di confidenza è quindi la frequenza di questi intervalli aleatori che contengono θ e perciò non si parla della probabilità che il nostro parametro sia contenuto nell'intervallo**

9 ***esempio***

Una nota azienda alimentare produce ogni giorno migliaia di scatole di cereali per la prima colazione. Se gli impianti di riempimento non funzionano adeguatamente le scatole possono risultare o troppo piene o troppo vuote. Poiché controllare ogni scatola risulterebbe lungo e dispendioso, il manager operativo deve elaborare una strategia che gli consenta di valutare a partire da un campione di scatole, la probabilità che il processo di riempimento sia sotto controllo:

una volta estratto un campione di scatole e pesata ognuna d'esse, si deve determinare la probabilità che tale campione, caratterizzato da una media $\bar{x}$, provenga da una popolazione la cui vera media μ , ha un certo valore, es. 368 gr.

Alla luce di tale risultato il manager prenderà una decisione circa il mantenere o meno lo stesso processo produttivo oppure modificarlo.

10

Supponiamo che il macchinario sia predisposto in maniera tale che la quantità in una scatola abbia distribuzione Normale di $\mu=368$ gr e che $\sigma=15$ gr. Se estraiamo casualmente un campione di 25 scatole ci potremmo ad es. chiedere qual è la probabilità che tale campione abbia una media inferiore a 365 gr.

Poiché $\bar{X} \square N(\mu, \sigma^2 / n)$ dobbiamo trovare l'area sottesa alla normale fino a 365; ossia in termini di $N(0,1)$:

$$Z = \frac{365 - 368}{15 / \sqrt{25}} = \frac{-3}{3} = -1$$

In corrispondenza di -1 sulle tavole troviamo che la probabilità cercata è 0.1587 $\implies$ il 15.87% di tutti i possibili campioni di ampiezza 25 ha una media campionaria al di sotto di 365 gr. Ciò non equivale alla probabilità che una singola scatola contenga meno di 365 gr di cereali; questa percentuale può essere calcolata come:

$$Z = \frac{x - \mu}{\sigma} = \frac{365 - 368}{15} = -0.20$$

In corrispondenza di -0.20 troviamo 0.4207 $\implies$ ci aspettiamo che il 42.07% delle singole scatole contenga meno di 365 gr. Tale % di singole scatole come si vede è superiore rispetto alla corrispondente % di medie campionarie: la probabilità che la media del campione sia lontana dalla media di P è inferiore alla probabilità che la singola osservazione lo sia.

Come cambia l'errore standard se n passa da 25 a 100 scatole?

Se $n=100$ $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$; quadruplicando l'ampiezza campionaria l'errore standard della media si riduce della metà: un aumento dell'ampiezza campionaria comporta minore variabilità delle possibili medie da campione a campione. La Probabilità di ottenere una media inferiore a 365 con un campione di 100 scatole sarà:

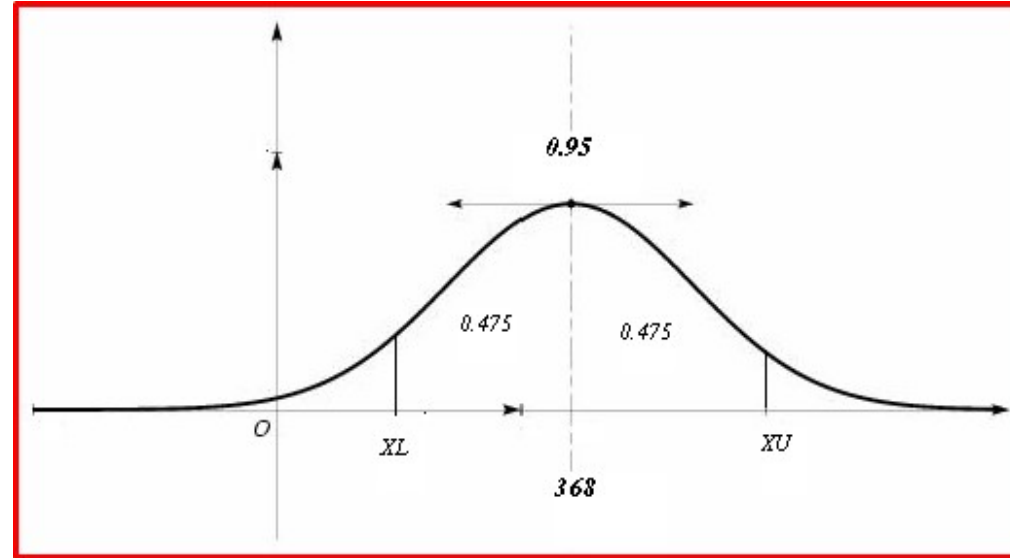
$$Z = \frac{365 - 368}{\frac{15}{\sqrt{100}}} = \frac{-3}{1.5} = -2$$

L'area a sinistra di -2 è 0.0228. Ci aspetteremo che il 2.28% dei campioni di $n=100$ ha una media inferiore a 365.

Determiniamo ora l'intervallo centrato sulla media della Popolazione di scatole, contenente il 95% delle medie campionarie calcolate a partire da campioni di 25 scatole.

Il 95% delle osservazioni si divide in due parti uguali, metà al di sotto della media e metà al di sopra





A sinistra il valore di Z cui corrisponde una prob. cum. di 0.025 è -1.96 ed il valore di Z cui corrisponde una prob. cum. pari a 0.975 è +1.96. Quindi:

$$\bar{X}_L = \mu - 1.96(\sigma / \sqrt{n}) = 368 - 1.96 \frac{15}{\sqrt{25}} = 362.12$$

$$\bar{X}_L = \mu + 1.96(\sigma / \sqrt{n}) = 368 + 1.96 \frac{15}{\sqrt{25}} = 373.88$$

Pertanto il 95% delle media di campioni di 25 scatole cadrà tra 362.12 e 373.88 gr.

La precedente conclusione si basa su un ragionamento deduttivo. Facciamo un tipo opposto di ragionamento di tipo induttivo: l'inferenza si basa infatti sui risultati di un solo campione per trarre conclusioni sulla intera Popolazione e non viceversa.

Consideriamo il caso in cui si voglia stimare la media di $P \mu$ (che supponiamo non conoscere) e ipotizziamo di conoscere $\sigma=15$ gr.

Pertanto invece di calcolare $\mu \pm 1.96(\sigma/\sqrt{n})$ per definire limite superiore ed inferiore dell'intervallo centrato in μ entro cui cade il 95% delle media campionarie, valutiamo che cosa succede quando a μ non nota sostituiamo $\bar{X}$ ed usiamo $\bar{X} \pm 1.96(\sigma/\sqrt{n})$ come intervallo per stimare μ .

Supponiamo inizialmente che un campione di 25 scatole abbia fornito come media 362.3 gr. L'intervallo costruito per stimare μ è $362 \pm 1.96(15/\sqrt{25})$ ossia 362.3 ± 5.88 . Lo stimatore per intervallo di μ sarà quindi:

$$356.42 \leq \mu \leq 368.18$$

Poiché la vera media di P è 368 essa è contenuta in tale intervallo; questo campione ha condotto ad una valutazione corretta!

¹⁴ Supponiamo che per un diverso campione di 25 scatole la media è 369.5. In questo caso l'intervallo sarebbe $369.5 \pm 1.96(15/\sqrt{25})$ ossia 369.5 ± 5.88 e lo stimatore per intervallo di μ :

$$363.62 \leq \mu \leq 375.38$$

Ed anche questo campione porterebbe a conclusioni corrette!

Tuttavia supponiamo di estrarre un terzo ipotetico campione di $n=25$ e di aver trovato che la media è 360 gr. In tal caso risulterebbe $360 \pm 1.96(15/\sqrt{25})$ ossia 360 ± 5.88 e lo stimatore per intervallo di μ sarà quindi:

$$354.12 \leq \mu \leq 365.88$$

Questo campione conduce ad una conclusione errata circa il vero valore di μ !

Quindi per alcuni campioni la stima (per intervallo) di μ è corretta mentre per altri non lo è.

Nella pratica estraiamo un solo campione e poiché non si conosce la media di P , non possiamo stabilire se le conclusioni a cui perveniamo siano corrette o meno.

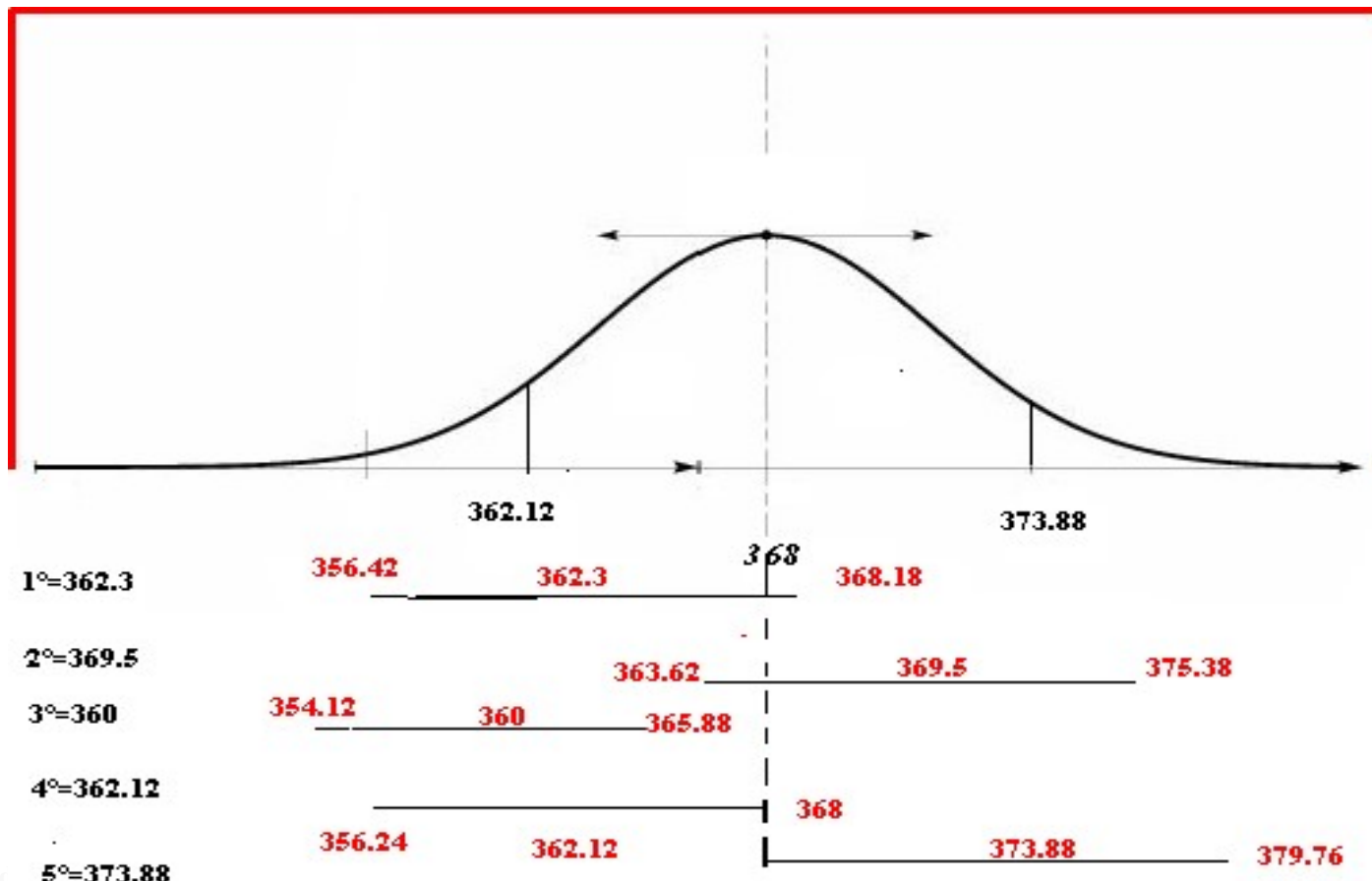


COME RISOLVIAMO QUESTO PROBLEMA?

15

Determiniamo la proporzione dei campioni che dà luogo ad affermazioni corrette sulla media μ .

Consideriamo altri due campioni di $n=25$ quello con media 362.12 e quello con media 373.88. Con gli analoghi calcoli fatti prima forniscono rispettivamente come stime per intervallo per μ : $356.24 \leq \mu \leq 368.00$ e $368.00 \leq \mu \leq 379.76$: la media di P coincide con i limiti inferiore e superiore dei due intervalli ed entrambi conducono a conclusioni corrette.



→ se la media campionaria calcolata per un campione $n=25$ è compresa tra 362.12 e 373.88, la media μ di P è inclusa nell'intervallo costruito sulla base di tale media.

Allo stesso tempo abbiamo visto che il 95% delle medie campionarie è compreso tra 362.12 e 373.88.

→ il 95% dei campioni di $n=25$ scatole ha una media campionaria che dà luogo ad intervalli contenenti la media di P .

*L'intervallo di estremi [362.12,373.88] è il nostro **intervallo di confidenza** al 95%*

Lo interpretiamo:

"se si considerano tutti i possibili campioni di ampiezza n , e per ciascuno la media campionaria e l'intervallo centrato su questa, il 95% degli intervalli così ottenuti contiene la media di P e solo il 5% di essi non la comprende"



non possiamo sapere se uno specifico intervallo include o meno μ ; tuttavia abbiamo una *confidenza (fiducia)* del 95% di aver selezionato un campione cui corrisponde un intervallo comprendente la media (incognita) di P

17 A volte è desiderabile un livello di fiducia maggiore (es. 99%), a volte possiamo accettare un grado minore di sicurezza (es 90%). In generale il livello di confidenza è indicato con $(1-\alpha)100\%$, dove " α è la massa che si trova nelle code della distribuzione fuori dall'intervallo di confidenza, che si bipartisce in maniera simmetrica in $\alpha/2$, nella coda sinistra e destra". In conclusione (supponendo nota la varianza di P!) abbiamo costruito l'intervallo per la media nel modo seguente:

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}} \quad \text{o equivalentemente} \quad \bar{X} - Z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}}$$

Dove Z (**valore critico** della distribuzione) è quel valore tale che l'area sottesa alla curva $N(0,1)$ tra $-Z$ e $+Z$ è pari ad $(1-\alpha)$; quindi Z si lascia a destra un'area pari ad $\alpha/2$ e l'area sottesa alla normale tra 0 e Z è pari ad $(1-\alpha)/2$.

A diversi livelli $(1-\alpha)$ corrispondono diversi valori critici:

$$(1-\alpha)=95\% \longrightarrow Z=1.96 ; (1-\alpha)=99\% \longrightarrow Z=2.58$$

La quantità $z_{\alpha/2} \sigma / \sqrt{n}$ viene detta anche margine di errore (o anche errore di campionamento) mentre $z_{\alpha/2}$ viene anche chiamato fattore di affidabilità .

In tabella diversi livelli di confidenza e corrispondenti soglie critiche nel caso di Normale

Livello di confidenza	90%	95%	98%	99%
α	0.10	0.05	0.02	0.01
$Z_{\alpha/2}$	1.645	1.96	2.33	2.58

Stima per intervallo – esempio (*fare su libro*)

Esempio:

Sia $X \sim N(\mu; \sigma^2 = 0,1)$

Si consideri un campione di dimensione $n=10$

La media campionaria è una v.c. che si distribuisce come

$$\bar{X} \sim N(\mu; \sigma^2/n = 0,01)$$

$Z = (\bar{X} - \mu)/\sqrt{0,01}$ è una v.c. Normale standardizzata

$$Pr(-1,96 \leq Z \leq +1,96) = 0,95$$

da cui possiamo ricavare che

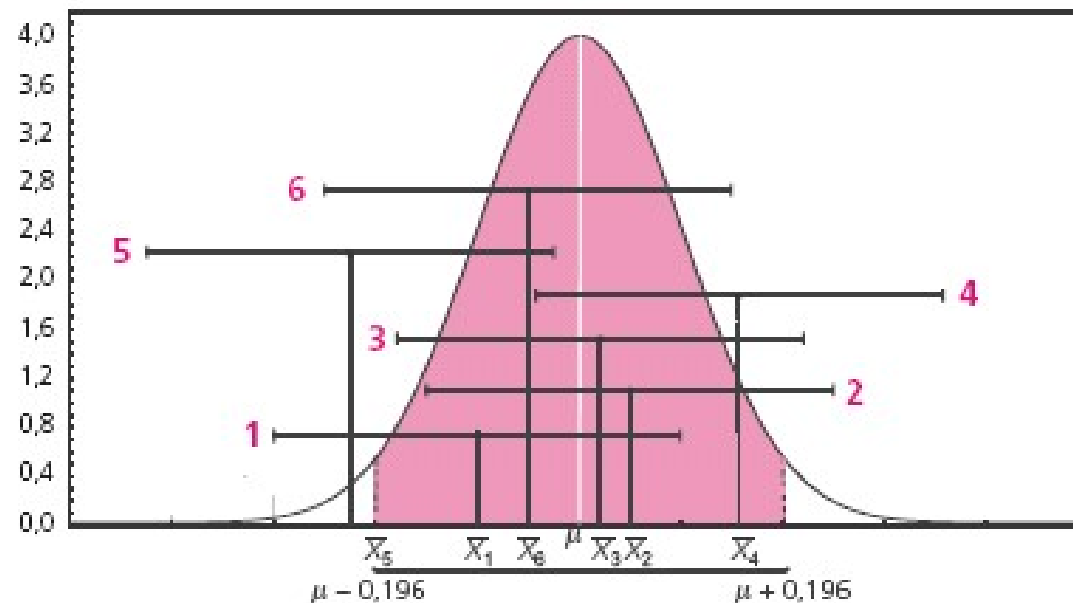
$$Pr(\bar{X} - 1,96 \cdot 0,1 \leq \mu \leq \bar{X} + 1,96 \cdot 0,1) = 0,95$$

Se dal campione estratto si osserva un valore della media pari a $\bar{X} = 10$, l'intervallo stimato risulta: $[9,804; 10,196]$

Stima per intervallo - esempio

Esempio (continua)

Nella seguente figura si mostrano, in corrispondenza di 6 campioni osservati, gl'intervalli di confidenza stimati per la media della popolazione a un livello di confidenza 0,95.



Osserviamo che dal campione 5 si ottiene un intervallo stimato che non contiene il vero parametro della popolazione.

Analogie con la stima puntuale

Nella seguente tabella sono riportate analogie e differenze tra la stima puntuale e la stima per intervallo.

	Stima puntuale	Stima intervallare
Campione casuale	$X_1, \dots, X_n$	$X_1, \dots, X_n$
Obiettivo	Stima puntuale del parametro θ	Stima per intervallo del parametro θ
Strumento	Stimatore puntuale: $T = T(X_1, \dots, X_n)$	Stimatore intervallo di confidenza: $[L_1, L_2] = [L_1(X_1, \dots, X_n), L_2(X_1, \dots, X_n)]$
Accuratezza	Errore quadratico medio $MSE(T) = E[(T - \theta)^2]$	Livello di confidenza $P(L_1 \leq \theta \leq L_2) = 1 - \alpha$
Dati campionari	$x_1, \dots, x_n$	$x_1, \dots, x_n$
Risultato	Stima puntuale: $t = T(x_1, \dots, x_n)$	Intervallo di confidenza stimato: $[l_1, l_2] = [L_1(x_1, \dots, x_n), L_2(x_1, \dots, x_n)]$

22 Intervallo di confidenza per la media (varianza nota e P Normale)

Sia X una v.c. che rappresenta un carattere osservato su una popolazione. Supponiamo che la v.c. sia distribuita come una **Normale con varianza nota**. Allora sappiamo che:

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \longrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$P(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1 - \alpha$$

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

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

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

23 Intervallo di confidenza per la media (con varianza nota e P distribuita come una Normale)

Dato un campione casuale estratto da una popolazione **Normale con media ignota e varianza nota**, l'intervallo di confidenza per la media della popolazione al livello di confidenza $1 - \alpha$ è:

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Esempio

Siano $n = 10$ $\sigma^2 = 9$ $1 - \alpha = 0,99$

Dalle tavole della Normale standardizzata si ottiene

$$z_{\alpha/2} = z_{0,005} = 2,576$$

Se $\bar{X} = 4,924$ si ottiene:

$$\left[4,924 \pm 2,576 \frac{\sqrt{9}}{\sqrt{10}} \right] \longrightarrow [2,4802, 7,3678]$$

24 Intervallo di confidenza per la media (varianza nota)

La **lunghezza** (ampiezza) dell'intervallo di confidenza si ricava dalla differenza tra estremo superiore e estremo inferiore:

$$\text{Lunghezza} = 2 z_{\alpha/2} (\sigma / \sqrt{n})$$

Dipende da:

- 1. la dimensione del campione**
- 2. il livello di confidenza**
- 3. la varianza della popolazione**

Intervenendo sulla dimensione del campione o sul livello di confidenza si può aumentare o diminuire la lunghezza dell'intervallo. Una volta fissati questi due elementi, al variare dei campioni estratti, la lunghezza degli intervalli corrispondenti rimane costante.

25 Intervallo di confidenza per la media (varianza nota)

La **lunghezza** dell'intervallo di confidenza si ricava dalla differenza tra estremo superiore e estremo inferiore:

$$\text{Lunghezza} = 2z_{\alpha/2}(\sigma/\sqrt{n})$$

Esempio:

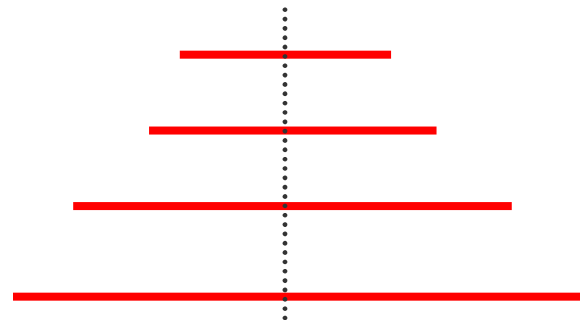
Fissato $1 - \alpha$

$n = 100$

$n = 70$

$n = 50$

$n = 10$



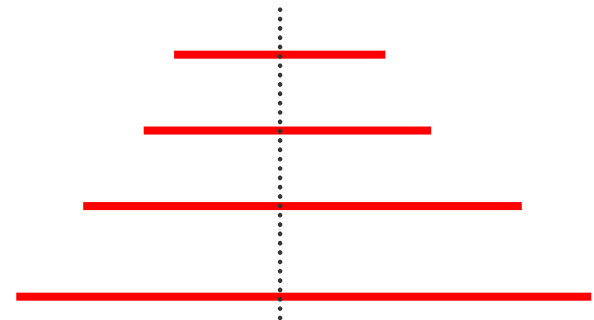
Fissato n

$1 - \alpha = 0,85$

$1 - \alpha = 0,90$

$1 - \alpha = 0,95$

$1 - \alpha = 0,99$



26 Esempio

Consideriamo un'azienda che produce fogli di carta per PC mediante un ciclo produttivo che opera ininterrottamente per ogni ciclo di produzione. I fogli di carta dovrebbero avere lunghezza media pari a 33 cm ed sqm pari a 0.06 cm. Ad intervalli regolari, vengono estratti dei campioni di fogli per stabilire se la lunghezza media è di 33 oppure se è il processo è andato fuori controllo. Supponiamo di estrarre un campione di 100 fogli e che risulta che la lunghezza media è 32.994 cm. Calcolare un intervallo di confidenza di livello 95% per la media della lunghezza dei fogli nella Popolazione.

$$n = 100 \qquad \bar{x} = 32.994 \qquad \sigma = 0.06$$

per il livello 95% risulta che $z=1.96$ ed avremo

$$\begin{aligned} \bar{X} \pm Z \frac{\sigma}{\sqrt{n}} &= 32.994 \pm (1.96) \frac{0.06}{\sqrt{100}} \\ &= 32.994 \pm 0.01176 \\ \Rightarrow \quad 32.98224 &\leq \mu \leq 33.00576 \end{aligned}$$

e non c'è motivo di ritenere che ci sia qualcosa che non va nel processo produttivo.

L'intervallo cambia se scegliamo un livello di confidenza diverso!

Supponiamo il livello desiderato sia del 99% e calcoliamo il corrispondente intervallo di confidenza:

$$n = 100 \qquad \bar{x} = 32.994 \qquad \sigma = 0.06$$

per il livello 99% risulta che $z=2.58$ ed avremo

$$\begin{aligned} \bar{X} \pm Z \frac{\sigma}{\sqrt{n}} &= 32.994 \pm (2.58) \frac{0.06}{\sqrt{100}} \\ &= 32.994 \pm 0.01548 \\ \Rightarrow \quad 32.97852 &\leq \mu \leq 33.00948 \end{aligned}$$

Un livello di confidenza più elevato ha condotto ad un intervallo più ampio.

Poichè 33 è incluso non si ha anche in tal caso motivo di ritenere che il processo non funzioni in modo corretto.

28

In genere σ di P al pari della media μ è non noto

➡ per ottenere un intervallo di confidenza della media dobbiamo basarci su entrambe le statistiche campionarie

$$\bar{X} \text{ ed } S^2$$

A questo punto si utilizza un risultato dovuto a W.S. Gosset, uno statistico impiegato presso la Guinness in Irlanda agli inizi del secolo che affrontò il problema della stima della media di una P Normale quando anche σ non è nota. Poiché agli impiegati della fabbrica non era concesso pubblicare lavori di ricerca utilizzò lo pseudonimo "Student". La distribuzione da lui determinata è da allora nota come ***t di Student***.

Se la v.c. $X \sim N(\mu, \sigma^2)$ allora la (statistica) v.c. $\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$

29 Sui gradi di libertà (g.d.l.)

Per determinare la varianza campionaria S^2 dobbiamo calcolare

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

Ossia bisogna conoscere la media; di conseguenza solo $n-1$ valori campionari sono liberi di variare: ci sono $n-1$ g.d.l.

Supponiamo di avere un campione di 5 valori che hanno media 20. Di quanti valori distinti si ha bisogno per ricostruire l'intero campione?

$$n = 5 \text{ e } \bar{X} = 20 \text{ ci dice anche che } \sum_{i=1}^n X_i = 100$$

$$(\text{poich\`e } \sum_{i=1}^n X_i / n = \bar{X})$$

Quindi se conosciamo solo 4 dei 5 valori, il quinto non sar\`a libero di variare, perch\`e la somma dei 5 elementi deve essere pari a 100. Per es. se i 4 valori sono pari a 18,24,19,16, il quinto potr\`a essere pari solo a 23 in modo da avere somma 100.

30 Intervallo di confidenza per la media (varianza ignota)

Sia X una v.c. che rappresenta un carattere osservato su una popolazione. Supponiamo che la v.c. sia distribuita come una **Normale con media e varianza ignota**.

Per stimare la varianza della popolazione si utilizza lo stimatore **varianza campionaria corretta**:

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

Pertanto la v.c. $T = (\bar{X} - \mu) / (S / \sqrt{n})$ si distribuisce come una v.c. ***t-Student*** con $n - 1$ gradi di libertà.

31 Intervallo di confidenza per la media (varianza ignota)

Dato un campione casuale di dimensione n estratto da una popolazione Normale con media e varianza entrambe ignote, l'intervallo di confidenza per la media a livello $1 - \alpha$ è dato da:

$$\left[\bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

La **lunghezza** dell'intervallo di confidenza è data in questo caso da:

$$\text{Lunghezza} = 2t_{\alpha/2} (S/\sqrt{n})$$

Nota

Al variare dei campioni estratti, la lunghezza degli intervalli corrispondenti non rimane costante poiché varia il valore di S .

32 Intervallo di confidenza per la media (popolazioni non Normali)

Quando **non è nota la popolazione** ma il campione ha una dimensione sufficientemente grande, possiamo considerare un'approssimazione dell'intervallo di confidenza per la media ottenuta attraverso il **teorema del limite centrale**.

Per n sufficientemente grande possiamo utilizzare il seguente intervallo di confidenza a livello $1 - \alpha$:

$$\left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

33 Esempio

Un manager di una società che vende combustibile per riscaldamento domestico vuole stimare il consumo medio annuo nelle case monofamiliari di una certa area geografica. Si estrae un campione di 35 case ed il loro consumo annuo risulta:

1150.25	1352.67	983.45	1365.11	942.71	1577.77	330.00
872.37	1126.57	1184.17	1046.35	1110.50	1050.86	851.60
1459.56	1252.01	373.91	1047.40	1064.46	1018.23	996.92
941.96	767.37	1598.57	1598.66	1343.29	1617.73	1300.76
1013.27	1402.59	1069.32	1108.94	1326.19	1074.86	975.86

Si vuole calcolare un intervallo di confidenza del 95% per la media della popolazione del consumo medio di combustibile in un anno.

I calcoli danno luogo a $\bar{x} = 1122.75$ ed $s^2 = 295.72$

Per ottenere l'intervallo di confidenza dobbiamo determinare il valore critico corrispondente ad un'area di 0.025 in ciascuna coda per 34 g.d.l. :

$$t_{(\alpha/2, n-1)} = t_{(0.025, 34)}$$

Dalla tavola si ricava $t = 2.0322$.

Abbiamo tutti gli "ingredienti" per determinare l'intervallo:

$$\bar{X} \pm t_{(\alpha/2, n-1)} \frac{S}{\sqrt{n}} = 1122.75 \pm (2.0322) \frac{295.72}{\sqrt{35}} = 1122.75 \pm 101.58$$

$$1021.17 \leq \mu \leq 1224.33$$

La validità della conclusione dipende dall'assunzione di Normalità; con un campione di ampiezza 35 l'uso della distribuzione t può ritenersi appropriato.

35 Intervallo di confidenza per una proporzione (campioni di dimensione elevata)

Quando la popolazione è riferita a un carattere che può assumere solo due modalità (popolazione Bernoulliana), siamo interessati all'intervallo di confidenza per una **proporzione** π , ad esempio, la proporzione di maschi nella popolazione. Come sappiamo un buon stimatore per π è la media campionaria $\bar{X}$.

Si ha: $E(\bar{X}) = \pi$ $V(\bar{X}) = \pi(1 - \pi)/n$

inoltre, dal **teorema del limite centrale** sappiamo che al crescere della dimensione campionaria la distribuzione della $\bar{X}$ tende alla Normale, pertanto

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

Intervallo di confidenza per una proporzione (campioni di dimensione elevata)

$$1 - \alpha \cong P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \pi}{\sqrt{\pi(1-\pi)/n}} \leq +z_{\alpha/2}\right) =$$

$$= P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq \bar{X} + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Tuttavia gli estremi dell'intervallo dipendono ancora dal parametro incognito e dunque devono essere sostituiti con degli stimatori, ottenendo il seguente intervallo di confidenza al livello $1 - \alpha$:

$$\left[\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right]$$

Una regola pratica: $n\bar{X} \geq 5$ e $n(1 - \bar{X}) \geq 5$

Esempio

Si vuole ottenere una stima intervallare della proporzione di fumatori presenti in una certa regione. A tal fine viene osservato un campione casuale di 120 persone, di cui 78 sono fumatori.

Quindi la stima puntuale della proporzione è data da:

$$\bar{X} = 78/120 = 0,65$$

quindi l'intervallo di confidenza al livello $1 - \alpha = 0,95$ è:

$$\left[0,65 - z_{0,025} \sqrt{\frac{0,65(0,35)}{120}}, 0,65 + z_{0,025} \sqrt{\frac{0,65(0,35)}{120}} \right] = [0,56, 0,74]$$

In questo caso: $n\bar{X} = 120 \cdot 0,65 = 78 \geq 5$

$$n(1 - \bar{X}) = 120 \cdot 0,35 = 42 \geq 5$$

Intervallo di confidenza per la varianza

Si consideri una popolazione Normale con media e varianza entrambe ignote.

Come stimatori puntuali dei due parametri si possono utilizzare: $\bar{X} \rightarrow \mu$ $S^2 \rightarrow \sigma^2$

Si può dimostrare che la v.c. $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ pertanto

$$\begin{aligned} 1 - \alpha &= P\left(\chi_{1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2}^2\right) = \\ &= P\left((n-1)S^2/\chi_{\alpha/2}^2 \leq \sigma^2 \leq (n-1)S^2/\chi_{1-\alpha/2}^2\right) \end{aligned}$$

E quindi l'intervallo per la varianza al livello $1 - \alpha$:

$$\left[(n-1)S^2/\chi_{\alpha/2}^2, (n-1)S^2/\chi_{1-\alpha/2}^2 \right]$$

Determinazione numerosità campionaria

Si consideri una popolazione Normale con media ignota e varianza nota. Ci si può chiedere quale debba essere la dimensione campionaria necessaria ad assicurare che la semi-lunghezza dell'intervallo non superi un certo valore δ . Dalla formula della lunghezza dell'intervallo di confidenza si ottiene:

$$n = \left(z_{\alpha/2} \frac{\sigma}{\delta} \right)^2$$

Quando la popolazione non è Normale o la varianza è ignota si utilizza:

$$n = \left(z_{\alpha/2} \frac{S}{\delta} \right)^2$$

tuttavia in questo caso è necessario che la numerosità risultante sia sufficientemente grande (>120)

Determinazione numerosità campionaria

Nel caso di popolazione Bernoulliana si ha:

$$n = z_{\alpha/2}^2 \frac{\hat{\pi}(1 - \hat{\pi})}{\delta^2}$$

dove $\hat{\pi}$ indica il valore della stima preliminare di π .

Se non si hanno informazioni a priori sul parametro incognito si usa fissare valore prudenziale pari a

$$\hat{\pi} = 0,5$$

Esempio

Si vuole stimare la numerosità necessaria per ottenere un intervallo di confidenza per π (ad es. la proporzione di persone propense a dare la preferenza a un certo candidato) in modo tale che la semi-lunghezza dell'intervallo di confidenza al livello 0,95 non sia superiore a 0,05.

$$n = 1,96^2 \frac{0,5(0,5)}{0,05^2} = 384,16 \approx 385$$