Media Campionaria

Intervallo di confidenza con σ nota

- Iniziamo assumendo, non realisticamente, di conoscere la media μ e la deviazione standard σ della popolazione
- Se questo fosse vero, la distribuzione campionaria della media dei campioni di dimensioni n sarebbe nota:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Illustrazione

- $\mu = €43236$,
- $\sigma = €15500$,
- e $n = 100$

ne segue che

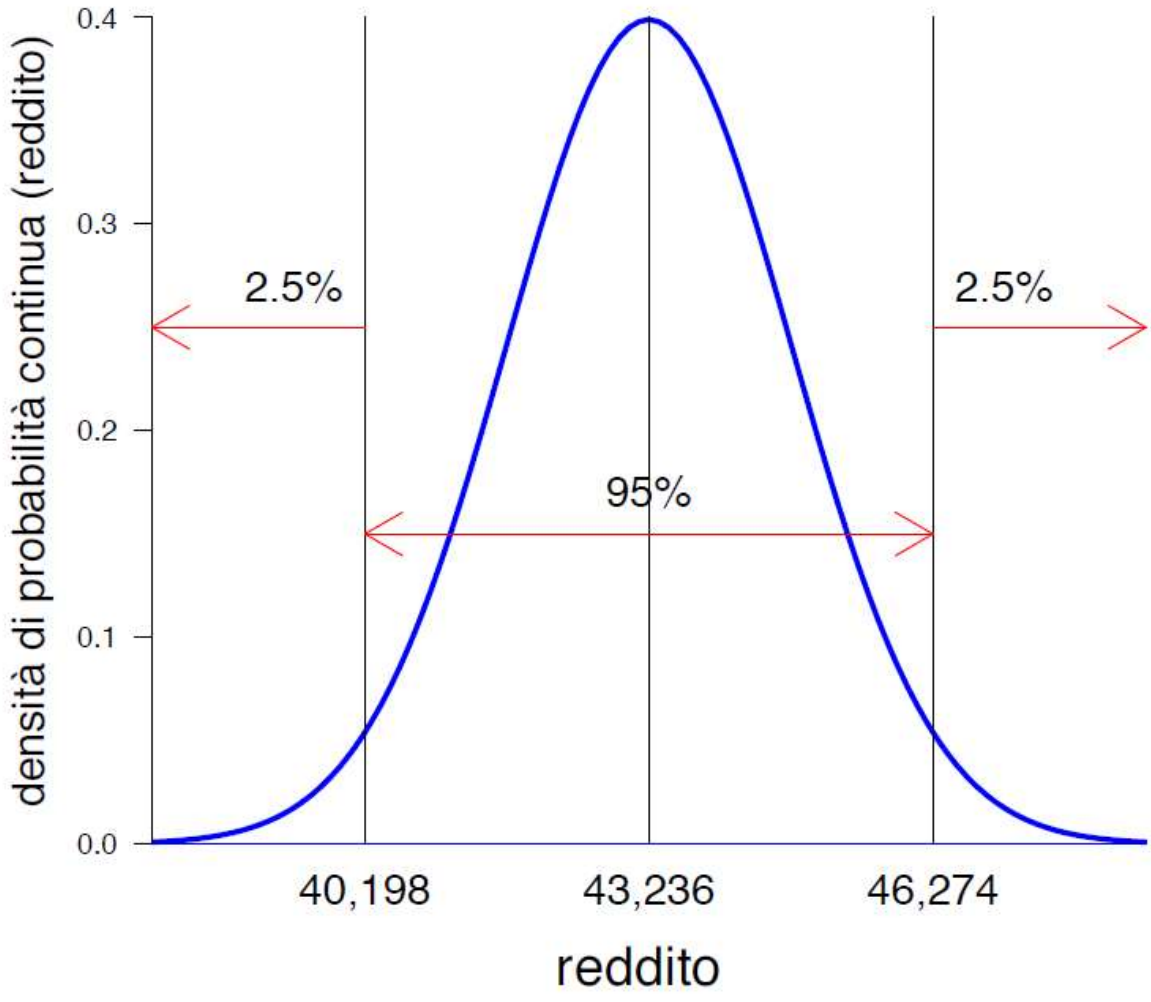
- $\bar{x} \sim N\left(43236, \frac{15500}{\sqrt{100}}\right)$

- Per le proprietà della distribuzione normale, sappiamo dunque che circa il 95% delle medie dei campioni di dimensioni $n = 100$ è contenuto nell'intervallo $43236 \pm 2 \times 1550$.

Se vogliamo essere più precisi, diciamo che il 95% delle medie dei campioni di dimensioni $n = 100$ è contenuto nell'intervallo
$43236 \pm \mathbf{1.96} \times 1550$.

Quindi:

- il 95% delle medie dei campioni è contenuto nell'intervallo
$\mu \pm 1.96 \times \sigma_{\bar{x}} = 43236 \pm 1.96 \times 1550 = 43236 \pm 3038$
- il 2.5% delle medie ha un valore minore di
$\mu - 1.96 \times \sigma_{\bar{x}} = 43236 - 3038 = 40198$
- il 2.5% delle medie ha un valore superiore a
$\mu + 1.96 \times \sigma_{\bar{x}} = 43236 + 3038 = 46274$



- In altri termini, se campioni di numerosità n venissero estratti ripetutamente dalla popolazione, nel 95% dei casi la statistica $\bar{x}$ sarebbe contenuta nell'intervallo $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$.
- L'affermazione precedente si può però rovesciare affermando che nel 95% dei casi (*campioni estratti*) la media μ della popolazione è contenuta nell'intervallo $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

$$P\left(-1.96 \leq \quad z_i \quad \leq 1.96\right) = 0.95$$

$$P\left(-1.96 \leq \frac{\bar{x}_i - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

$$P\left(-1.96 \times \frac{\sigma}{\sqrt{n}} \leq \bar{x}_i - \mu \leq 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-\bar{x}_i - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x}_i + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{x}_i + 1.96 \times \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x}_i - 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- Se dunque costruiamo un intervallo avente ampiezza

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

allora questo intervallo conterrà la media μ della popolazione nel 95% dei campioni casuali di ampiezza n estratti dalla popolazione.

Illustrazione. Supponiamo, per esempio, che il ricercatore disponga di un campione di $n = 100$ osservazioni avente media $\bar{x} = 41100$.

Supponiamo inoltre che il ricercatore non conosca la media μ della popolazione, ma sappia che la deviazione standard della popolazione è $\sigma = 15500$.

• Il ricercatore può dunque attribuire un grado di fiducia del 95% all'affermazione secondo cui la media della popolazione è contenuta nell'intervallo.

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm 1.96\frac{15500}{\sqrt{100}} = 41100 \pm 3038$$

Questo intervallo, che ha la forma

$$\boxed{\text{stima puntuale} \pm \text{margine d'errore}}$$

è chiamato intervallo di confidenza per la media sconosciuta della popolazione μ.

Si noti che, nel caso presente, la media della popolazione $\mu = 43236$ (che è conosciuta a noi ma non al ricercatore) è contenuta nell'intervallo di confidenza.

Questo esempio, però, è poco realistico: in qualsiasi applicazione concreta dell'inferenza statistica, se il parametro μ non è conosciuto, tantomeno lo è σ.

Intervallo di confidenza con σ ignota

- Nel caso in cui la deviazione standard della popolazione non sia conosciuta (in pratica, sempre!), si usa la deviazione standard del campione s quale stima di σ.
- E' importante ricordare che la deviazione standard del campione si calcola mediante la formula

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

dove la parte sotto radice è chiamata varianza campionaria corretta o stima corretta della varianza.

- Una stima della deviazione standard della media dei campioni di dimensioni n è calcolata mediante

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- L'intervallo di confidenza per la media μ della popolazione al grado di fiducia $1-\alpha$ diventa dunque

$$\bar{x} \pm z_{\alpha/2}\hat{\sigma}_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Illustrazione. Per l'esempio precedente, con $n = 100$, $\bar{x} = 41100$ e, poniamo, $s = 13950$, l'intervallo di confidenza al 95% per μ sarà:

$$\begin{aligned} \bar{x} \pm 1.96\hat{\sigma}_{\bar{x}} = & \quad 41100 \pm 1.96\frac{13950}{\sqrt{100}} \\ = & \quad 41100 \pm 2734.2 \end{aligned}$$

Illustrazione. Più correttamente, dovremmo usare il valore critico della distribuzione $t - Student$ con $n - 1 = 99$ gradi di libertà; quelli della stima della varianza della popolazione con s^2.

Nel caso presente, $t(\alpha/2 = .025;\ gdl = 99) = 1.984217$:

$$\bar{x} \pm t_{(\alpha/2,\, n-1)}\hat{\sigma}_{\bar{x}} = 41100 \pm 1.984217\frac{13950}{\sqrt{100}}$$

$$= 41100 \pm 2768$$

Quando n è grande $t \to N$.

Interpretazione dell'intervallo di confidenza

- Per attribuire l'interpretazione corretta all'intervallo di confidenza per μ dobbiamo supporre di estrarre dalla popolazione tutti i possibili campioni aventi numerosità n e di costruire tutti i possibili intervalli di confidenza (uno per ciascun campione).

  In queste circostanze,

  - una frazione uguale a $1 - \alpha$ degli intervalli di confidenza conterrà il valore μ, e
  - la rimanente frazione α non lo conterrà.

- Per questa ragione assegnamo un livello di fiducia pari a $1 - \alpha$ all'affermazione secondo cui l'intervallo di confidenza $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ contiene il vero valore μ della media della popolazione.

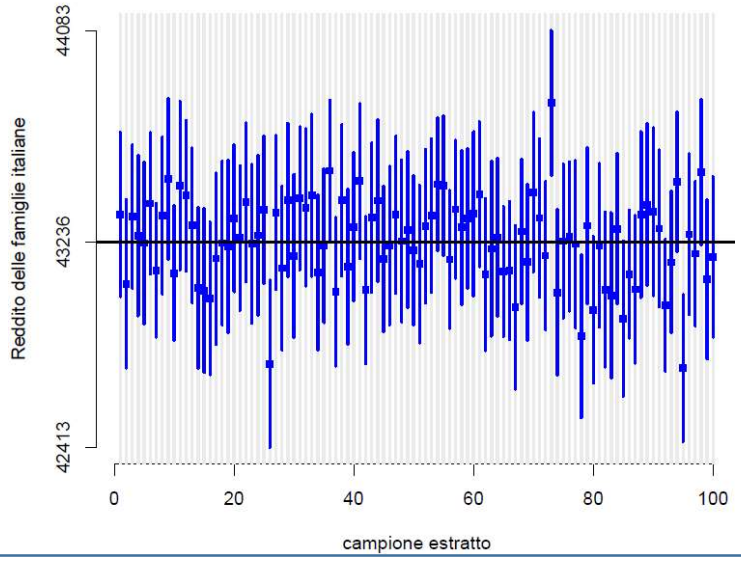
L'interpretazione corretta dell'intervallo di confidenza per μ al 95% potrebbe dunque essere formulata in questi termini:

> *il ricercatore attribuisce un grado di fiducia al* 95% *all'affermazione secondo cui la media* μ *della popolazione è contenuta nell'intervallo compreso tra* 38332 *e* 43868 *nel senso che ha usato una procedura la quale produce la risposta corretta nel* 95% *dei campioni casuali di numerosità* $n = 100$ *e la risposta sbagliata nel restante* 5% *dei campioni;*

> *il ricercatore però non può mai sapere se l'intervallo costruito utilizzando uno specifico campione contenga o meno il vero valore* μ *della media della popolazione.*

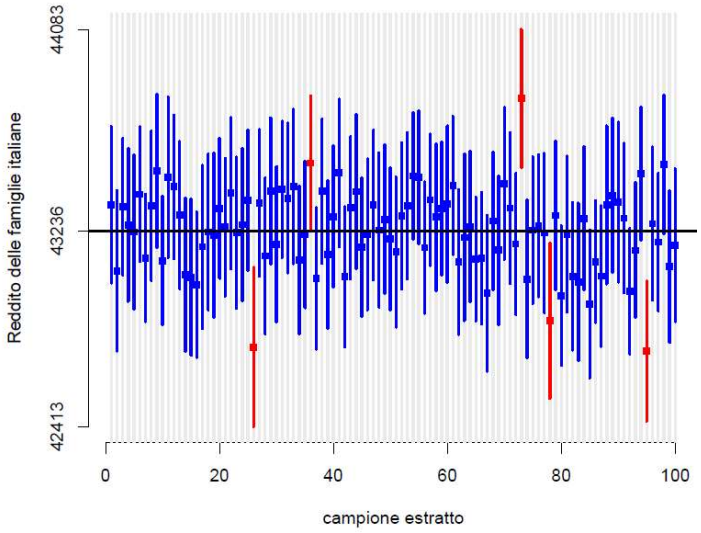
Nella figura seguente sono presentati i risultati di una simulazione in cui 100 campioni casuali di numerosità $n = 100$ vengono estratti da una popolazione $\approx N(43236,\ 15500)$

- Per ciascun campione $i - esimo$ vengono calcolate la media $\bar{x}_i$ e la deviazione standard s_i.
- Usando queste informazioni, vengono calcolati 100 intervalli di confidenza al 95% per μ.
- Gli intervalli di confidenza sono rappresentati nella figura con dei segmenti verticali. La linea orizzontale rappresenta il reddito medio $\mu = 43236$



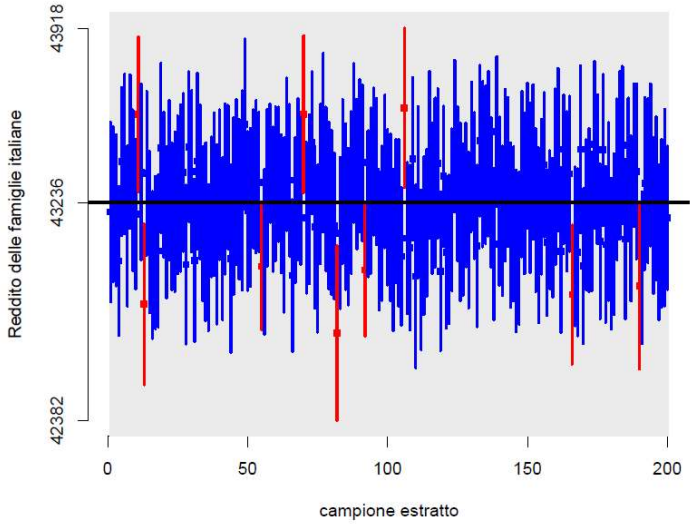
• I segmenti rossi rappresentano gli intervalli di confidenza al 95% che NON contengono la media della popolazione.

5 (5%) non contengono la media $\mu = 43236$



Simulazione su 200 campioni

9 ($\approx 5\%$) non contengono la media $\mu = 43236$



Probabilità d'errore

- La probabilità che un intervallo di confidenza NON contenga il parametro è chiamata probabilità d'errore.
- La probabilità d'errore è denotata da α.
- Il coefficiente di confidenza C è uguale a $(1 - \alpha)$

Per un intervallo di confidenza al 95%, per esempio, $C = 0.95$ e $\alpha = 0.05$.

- Questo è l'intervallo di confidenza più comunemente usato.
- Sono però comuni anche intervalli di confidenza al 90% e al 99%.

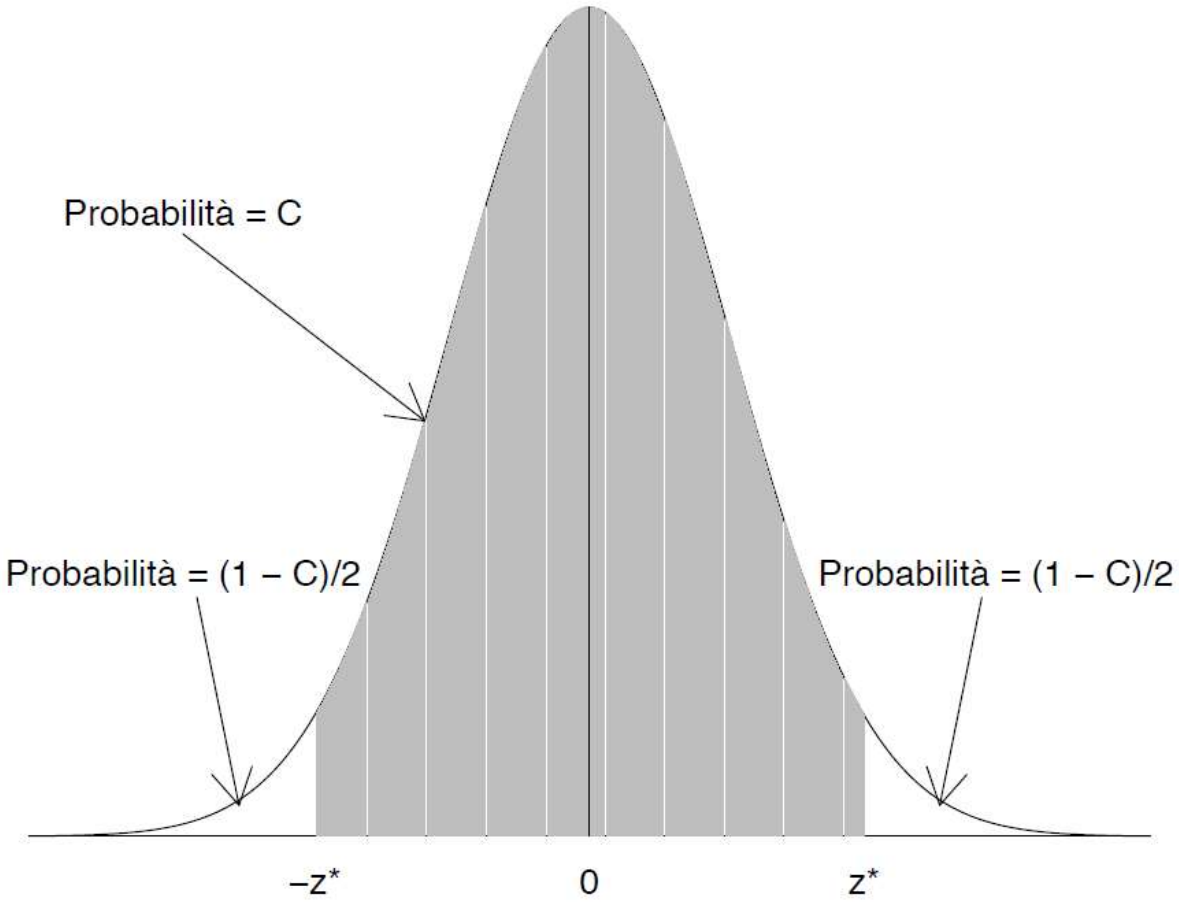
Livelli di fiducia

In generale, per costruire un intervallo di confidenza per μ al livello del $100(1 - \alpha)\%$, si usa la formula:

$$\bar{x} \pm z^* \hat{\sigma}_{\bar{x}} = \bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

laddove $1 - \alpha$ corrisponde all'area sottesa alla curva normale standardizzata nell'intervallo $[-z^*,\ z^*]$.

Si noti che l'area in ciascuna coda della distribuzione è $\alpha/2$.



Livelli di fiducia

- Di seguito sono riportati i valori critici z^* corrispondenti agli intervalli di confidenza più comunemente usati.

intervallo di confidenza al	coda	z^*
90%	0.050	1.645
95%	0.025	$1.960 \simeq 2$
99%	0.005	2.576

- Per il nostro esempio:

$$\begin{aligned}
&\text{intervallo di confidenza al } 90\% : \quad 41100 \pm 1.645 \times 1.395 \\
&\text{intervallo di confidenza al } 95\% : \quad 41100 \pm 1.960 \times 1.395 \\
&\text{intervallo di confidenza al } 99\% : \quad 41100 \pm 2.576 \times 1.395
\end{aligned}$$

- Si noti che:
- se si desidera un grado di confidenza maggiore, l'intervallo deve essere più ampio;
- è comunque desiderabile che l'intervallo di confidenza sia piccolo, ovvero che il margine d'errore

$$z^* s/\sqrt{n}$$

sia il minore possibile.

Margine d'errore

- Tre fattori influenzano il margine d'errore $z^* s/\sqrt{n}$.
1. Se vogliamo un grado di confidenza maggiore, dobbiamo scegliere un valore z^* più grande. Questo produce un margine d'errore più grande.
2. Il margine d'errore cresce all'aumentare della varianza di X nella popolazione (ovvero, al crescere di s). È più facile stimare precisamente μ nel caso di una popolazione omogenea che in una popolazione eterogenea – dato che s non è sotto il nostro controllo, non possiamo però ottenere una maggiore precisione agendo su s.
3. n si trova al denominatore del margine d'errore e dunque una precisione maggiore può essere ottenuta utilizzando un campione più grande – dato però che n è sotto radice, dobbiamo aumentare le dimensioni del campione di 4 volte se vogliamo ridurre a metà il margine d'errore.

Numerosità del campione

- Supponiamo di fissare il margine d'errore m, per un dato livello di confidenza $1-\alpha$. Supponiamo inoltre di conoscere la deviazione standard della popolazione σ.
- Quante osservazioni sono necessarie per ottenere il margine d'errore m?

Il margine d'errore è

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

risolvendo per n otteniamo

$$n = \left(z^* \frac{\sigma}{m} \right)^2$$

Illustrazione. Si costruisca un intervallo di confidenza al 95% per il reddito medio in una popolazione avente deviazione standard $\sigma = €15500$ (come nell'esempio precedente). Poniamo il margine d'errore uguale a $m = €500$.

- Le dimensioni n richieste sono

$$n = \left(z^* \frac{\sigma}{m} \right)^2 = \left(1.96^* \frac{15500}{500} \right)^2 = 3691.8$$

ovvero, quasi 3700 famiglie.

- Per $m = 250$ (un margine di errore dimezzato)

$$n = \left(z^* \frac{\sigma}{m} \right)^2 = \left(1.96^* \frac{15500}{250} \right)^2 = 14767.11$$

ovvero, quasi 14800 famiglie (quattro volte l'n precedente).

Conclusioni

- Una stima puntuale è una statistica (ovvero, un numero calcolato sui dati di un campione) che fornisce la valutazione del valore del parametro sconosciuto della popolazione.
- Una stima intervallare, chiamata intervallo di confidenza, fornisce un intervallo per il parametro al grado di fiducia $1 - \alpha$.
- Gli intervalli di confidenza hanno la forma

$$\text{stima puntuale } \pm z \times \widehat{\text{errore standard}}$$

- Gli intervalli di confidenza che abbiamo discusso richiedono campioni di grandi dimensioni dato che sono stati calcolati assumendo che la distribuzione campionaria della media sia normale.
- Per campioni di grandi dimensioni, il teorema del limite centrale garantisce la gaussianità della distribuzione campionaria della media anche se la popolazione non segue la distribuzione normale.

Parametro	Stima puntuale	Errore standard stimato	Intervallo di confidenza	n e margine d'errore m
Popolazione normale – grandi campioni				
μ	$\bar{x}$	$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$	$\hat{\bar{x}} \pm z\hat{\sigma}_{\bar{x}}$	$n = \left(\frac{zs}{m}\right)^2$