Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 12 13

Piano della presentazione

Inferenza Statistica

Inferenza statistica

- In precedenza abbiamo ragionato in maniera deduttiva utilizzando le caratteristiche di una *popolazione conosciuta* per dedurre le caratteristiche dei campioni casuali che si possono estrarre da tale popolazione.

• **Illustrazione**. Supponiamo, per esempio, che il reddito italiano medio sia $\mu = €43236$ con deviazione standard $\sigma = €15500$.

• Avendo la popolazione tali caratteristiche, le medie dei campioni casuali di dimensioni $n = 100$ saranno distribuite in maniera approssimativamente normale con media $\mu_{\bar{x}} = €43236$ e deviazione standard $\sigma_{\bar{x}} = €15500/\sqrt{n} = 1550$;

• Diremo quindi:

$$\bar{x} \sim N(43236, 1550)$$

Inferenza statistica

- In qualsiasi applicazione concreta dell'inferenza statistica, però, quando un campione casuale viene estratto dalla popolazione, le caratteristiche della popolazione non sono conosciute.
- *Se le caratteristiche della popolazione fossero conosciute non avrebbe senso utilizzare le informazioni di un campione!*

Inferenza statistica

- Inoltre, il ricercatore ha a disposizione un unico campione di n osservazioni.
- *Se il ricercatore potesse misurare* 1000 *campioni contenenti* $n = 100$ *osservazioni ciascuno, per esempio, allora tratterebbe queste osservazioni come un unico campione di dimensioni* $n = 1000 \times 100 = 100000$.

Inferenza statistica

- L'approccio classico all'inferenza statistica si basa sulla concezione frequentista della probabilità.
- Nell'approccio classico, l'inferenza statistica è dunque fondata sul principio di ripetizione del campionamento, per cui i dati osservati vengono considerati come uno degli infiniti possibili campioni che teoricamente si otterrebbero se il processo di campionamento venisse ripetuto infinite volte nelle medesime condizioni.

Inferenza statistica

- L'inferenza statistica procede per mezzo di due approcci distinti:
  1. La stima (puntuale e intervallare) dei parametri.
  2. La verifica di ipotesi statistiche.

La stima dei parametri

- Un gran numero di problemi dell'inferenza statistica concerne la stima dei parametri ignoti della popolazione, avendo a disposizione delle statistiche, ossia le misure tratte da un campione.
- Supponiamo, per esempio, di porci il problema di stabilire, sulla base dei dati campionari, la "miglior stima possibile" del parametro μ, il reddito medio della popolazione italiana. Tale stima viene detta puntuale.

Intervallo di confidenza

- Le stime puntuali – ovvero le statistiche quali la media e la varianza del campione – variano da campione a campione.
- In generale, dunque, è preferibile tenere conto della variabilità campionaria costruendo la stima dell'intervallo di confidenza.
- Nel caso di distribuzioni campionarie simmetriche, un intervallo di confidenza prende la forma di una stima puntuale $\pm$ un margine d'errore.

Verifica di ipotesi

- I test statistici costituiscono il secondo metodo con cui può essere effettuata un'inferenza statistica.
- Un test statistico è una procedura che, sulla base dei dati osservati, consente di decidere se è ragionevole assegnare un certo valore ad un parametro.

Verifica di ipotesi

- Supponiamo, per esempio, di chiederci se il reddito medio degli uomini sia uguale a quello delle donne – ovvero, se la differenza tra il reddito medio di uomini e donne ($\mu_{uomini} - \mu_{donne}$) sia uguale a zero
- Il test di un'ipotesi statistica (la cosiddetta significatività di un test statistico) ci dice in che misura i dati osservati sono coerenti con l'asserzione che assegna un certo valore ad un parametro.
- Nel nostro esempio, ci chiederemo se il dato campionario

$$\bar{x}_{uomini} - \bar{x}_{donne} \neq 0$$

sia coerente con l'ipotesi

$$\mu_{uomini} - \mu_{donne} = 0$$

Stima puntuale

Stima puntuale

- Se da una popolazione con media reale μ si estrae a caso un campione di n soggetti, a causa dell'errore di campionamento la media campionaria $\bar{x}$ non avrà un valore identico alla media μ della popolazione.
- Inversamente, la media $\bar{x}$ di un campione non coincide esattamente con il valore reale della media μ della popolazione.

Stima puntuale

proprietà di uno stimatore puntuale

• Per arrivare alla stima migliore del parametro della popolazione partendo dal campione, sono state definite quattro proprietà, di cui uno stimatore puntuale dovrebbe godere:

1. **correttezza**,
2. **efficienza**,
3. **consistenza**,
4. **sufficienza**.

Correttezza

Correttezza. Una statistica si dice corretta (*unbiased*) se la media della sua distribuzione campionaria è uguale al parametro stimato.

- Una statistica corretta talvolta fornisce una stima troppo grande, talvolta troppo piccola. La media calcolata sui valori che tale statistica assume in campioni diversi, però, è uguale al valore del parametro.
- Uno stimatore non corretto è detto distorto (*biased*); la distorsione nella stima (*bias*) è la differenza tra la media generale che la statistica assume nei vari campioni e il valore (vero) del parametro della popolazione.

Efficienza

- Il valore medio di una statistica non è l'unica proprietà di una statistica dato che, in campioni diversi, una statistica varia attorno alla sua media.

Efficienza. Uno stimatore si dice efficiente quando le misure ottenute in campioni diversi sono vicine al valore reale della popolazione.

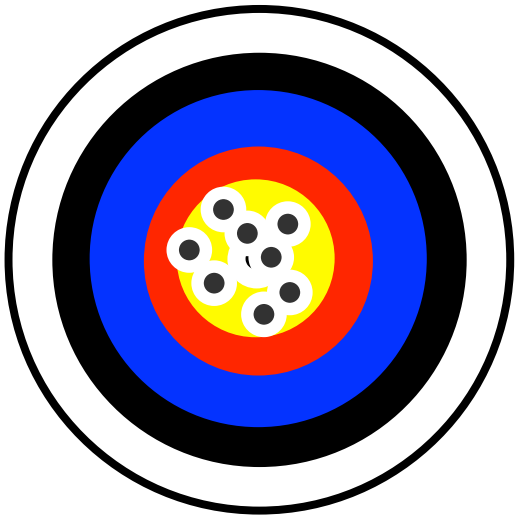
- Fra più stimatori *corretti*, il migliore è quello che ha la varianza minore.

Efficienza

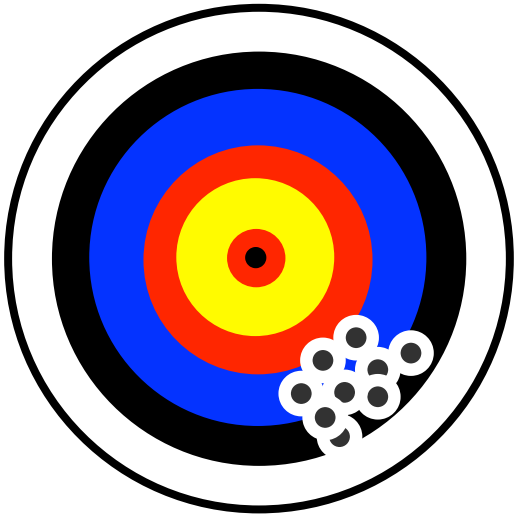
- La stima dei parametri ignoti della popolazione può essere paragonata al tiro al bersaglio, pensando al centro del bersaglio come al parametro e ai singoli tentativi come alle stime.
  - Una statistica corretta (non distorta) in media colpisce il centro del bersaglio.
  - Una statistica efficiente addensa i suoi colpi in un'area ristretta del bersaglio
  - L'errore (o scarto) quadratico medio (in inglese *mean-squared error*, MSE) di una statistica – ovvero, la media dei quadrati degli errori di stima – è uguale alla somma del quadrato della distorsione e della varianza:

$$\begin{aligned} SQM &= E\left[stat_i - (par + bias)\right]^2 \\ &= E\left[(stat_i - par) - bias)\right]^2 \\ &= E\left[(stat_i - par)^2 + bias^2 - 2(stat_i - par)bias)\right] \\ &= Var(stat_i) + bias^2 \end{aligned}$$

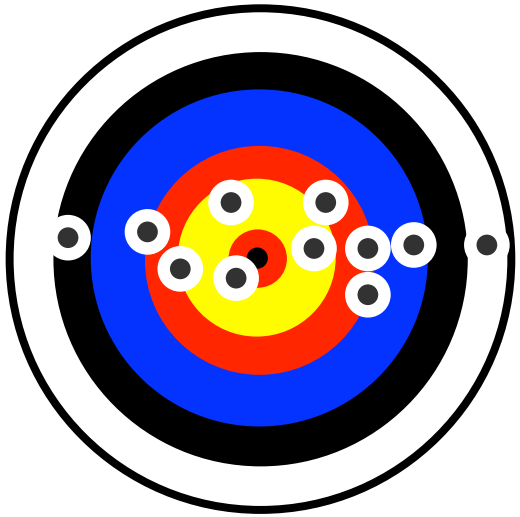
Correttezza e efficienza



Stimatore non distorto poco variabile



Stimatore distorto poco variabile



Stimatore non distorto molto variabile

Consistenza e sufficienza

Consistenza. Uno stimatore si dice consistente quando la differenza tra la stima ed il valore vero del parametro della popolazione diminuisce all'aumentare delle dimensioni del campione. Si tratta di una proprietà asintotica di uno stimatore: uno stimatore è consistente quando la sua distribuzione tende ad accentrarsi, al crescere di n, sempre più vicino al parametro (ignoto).

Sufficienza. Si ha sufficienza quando uno stimatore sintetizza tutte le informazioni presenti nel campione che sono importanti per la stima del parametro.

Intervallo di confidenza con σ nota

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:

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

Intervallo di confidenza con σ nota

Illustrazione. Per l'esempio precedente,

- $\mu = €43236$,
- $\sigma = €15500$,
- 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 $43236 \pm 2 \times 1550$.

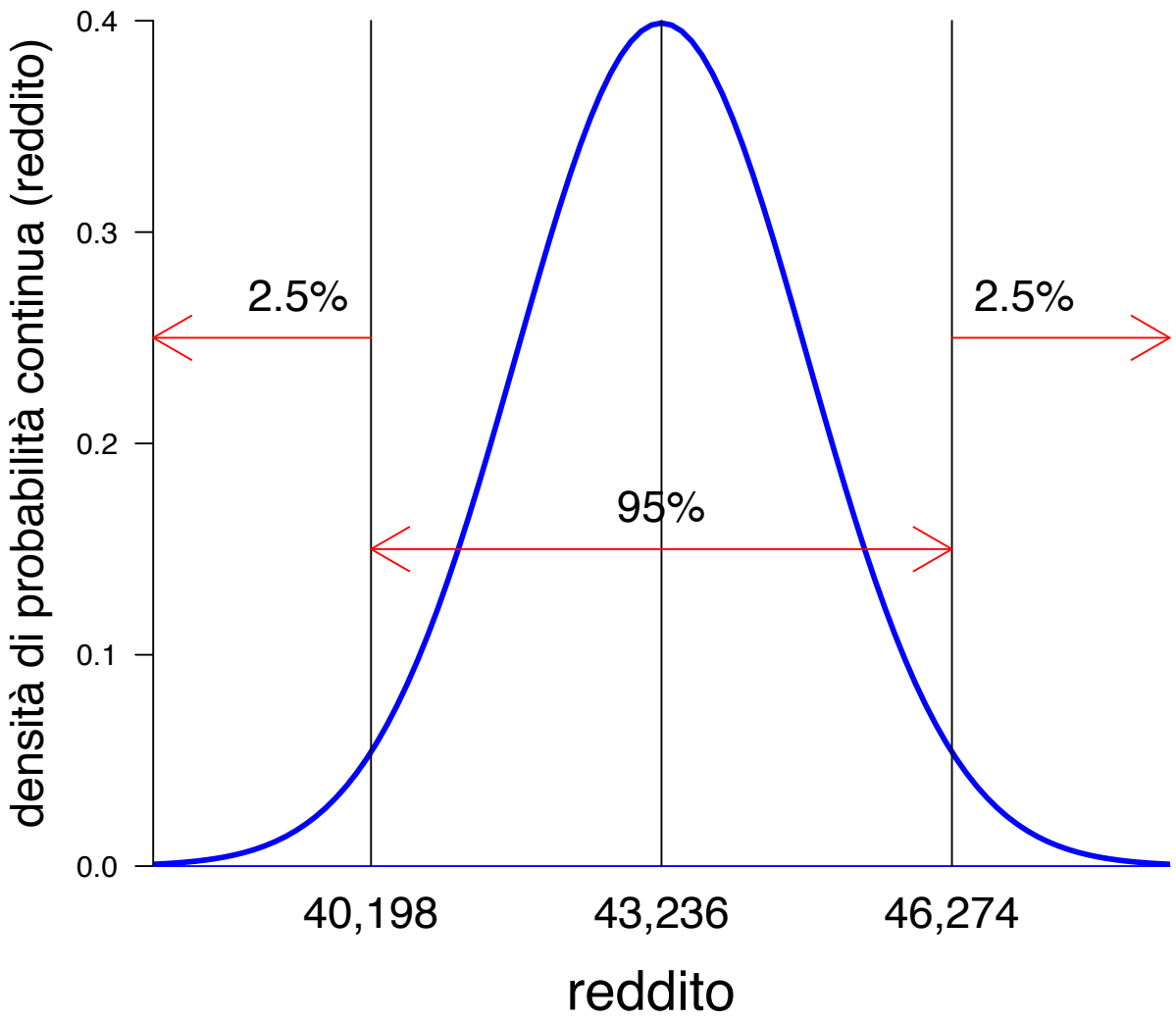
Intervallo di confidenza con σ nota

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
$\mu + 1.96 \times \sigma_{\bar{x}} = 43236 + 3038 = 46274$

Intervallo di confidenza con σ nota



Intervallo di confidenza con σ nota

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

Intervallo di confidenza con σ nota

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

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

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

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

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

Intervallo di confidenza con σ nota

- 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.

Intervallo di confidenza con σ nota

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.

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

Intervallo di confidenza con σ nota

Questo intervallo, che ha la forma

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

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.

Intervallo di confidenza con σ nota

- Questo esempio, però, è poco realistico: in qualsiasi applicazione concreta dell'inferenza statistica, se il parametro μ non è conosciuto, tantomeno lo è σ.
- Come procediamo quando la deviazione standard della popolazione non è conosciuta?

Intervallo di confidenza con σ ignota

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.

Stima corretta della varianza

$$
\begin{aligned}
E\left[s_j^2\right] &= E\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x}_j)^2}{n}\right] \\
&= E\left[\sum_{i=1}^{n}\frac{((x_i-\mu)-(\bar{x}_j-\mu))^2}{n}\right] \\
&= E\left[\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}+\frac{\not{n}(\bar{x}_j-\mu)^2}{\not{n}}-2(\bar{x}_j-\mu)\sum_{i=1}^{n}\frac{(x_i-\mu)}{n}\right] \\
&= E\left[\sum_{i=1}^{n}\frac{1}{n}(x_i-\mu)^2+(\bar{x}_j-\mu)^2-2(\bar{x}_j-\mu)\left(\sum_{i=1}^{n}\frac{x_i}{n}-\not{n}\frac{\mu}{\not{n}}\right)\right] \\
&= E\left[\sum_{i=1}^{n}\frac{1}{n}(x_i-\mu)^2+(\bar{x}_j-\mu)^2-2(\bar{x}_j-\mu)(\bar{x}_j-\mu)\right] \\
&= E\left[\sum_{i=1}^{n}\frac{1}{n}(x_i-\mu)^2-(\bar{x}_j-\mu)^2\right] \\
&= \not{n}\frac{1}{\not{n}}E\left[(x_i-\mu)^2\right]-E\left[(\bar{x}_j-\mu)^2\right]=\sigma^2-\frac{\sigma^2}{n}= \boxed{\frac{n-1}{n}\sigma^2<\sigma^2}
\end{aligned}
$$

Stima corretta della varianza

- la varianza campionaria $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$ è quindi uno stimatore distorto (*biased*) della varianza della popolazione σ^2.
- il suo valore atteso è sistematicamente inferiore al parametro σ^2 della popolazione.
- Possiamo quindi correggere lo stimatore campionario nel seguente modo:

$$E\left[s_j^2\right] \times \frac{n}{n-1} = \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2$$

$$E\left[\sum_{i=1}^{n} \frac{(x_i - \bar{x}_j)^2}{\cancel{n}} \times \frac{\cancel{n}}{n-1}\right] = \sigma^2$$

$$E\left[\sum_{i=1}^{n} \frac{(x_i - \bar{x}_j)^2}{n-1}\right] = \sigma^2$$

Stima corretta della varianza

- Quando nella formula della varianza viene utilizzata la media campionaria $\bar{x}$ come stima di μ, stiamo sottostimando la vera variabilità dei dati.
- La quantità al denominatore $n - 1$ è chiamata (numero di) gradi di libertà (gdl).
- L'operazione di aggiustamento del valore atteso di una statistica campionaria incorretta (biased) attraverso l'uso dei gdl, in luogo di n, è detta aggiustamento per i gradi di libertà.

Stima corretta della varianza

Inoltre,

- Si può dimostrare che

$$Var\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x}_j)^2}{n}\right] < Var\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x}_j)^2}{n-1}\right]$$

- La varianza campionaria corretta, che usa $n-1$ al denominatore, è quindi uno stimatore meno efficiente della varianza campionaria non corretta.
- La sua varianza campionaria è superiore a quella della statistica non corretta, che usa n al denominatore.
- Sarà comunque da preferire come *stima puntale non distorta* del parametro σ.

Intervallo di confidenza con $\hat{\sigma}_{\bar{x}} = s/\sqrt{n}$

- 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

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

Intervallo di confidenza con $\hat{\sigma}_{\bar{x}} = s/\sqrt{n}$

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

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

Intervallo di confidenza con $\hat{\sigma}_{\bar{x}} = s/\sqrt{n}$

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

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

$$= \quad 41100 \pm 2768$$

Quando n è grande $t \to N$.

Interpretazione dell'intervallo di confidenza

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.

Interpretazione dell'intervallo di confidenza

- 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.

Interpretazione dell'intervallo di confidenza

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.*

Interpretazione dell'intervallo di confidenza

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

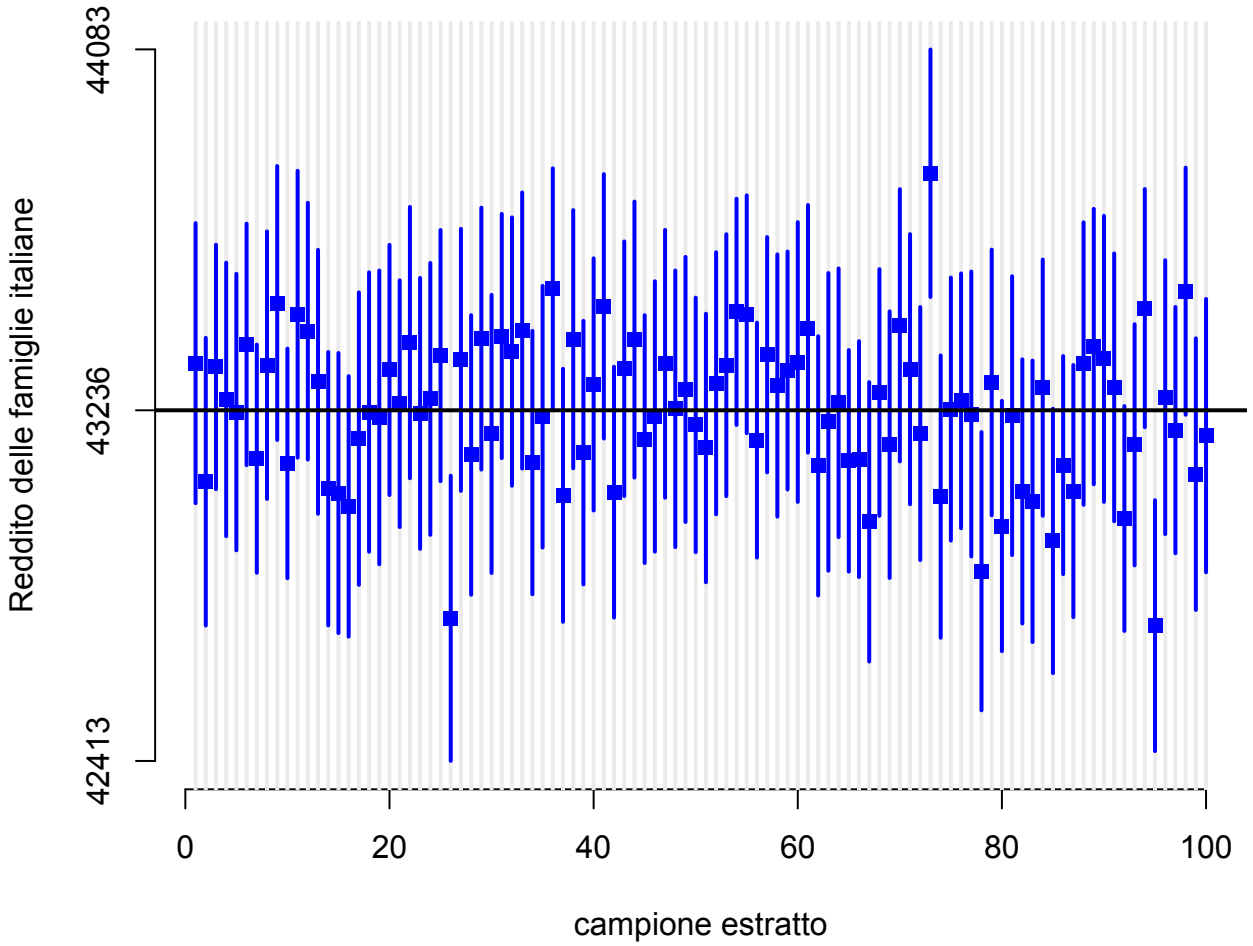
  • Per ciascun campione $i - esimo$ vengono calcolate la media $\bar{x}_i$ e la deviazione standard s_i.

  • 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 $\mu = 43236$

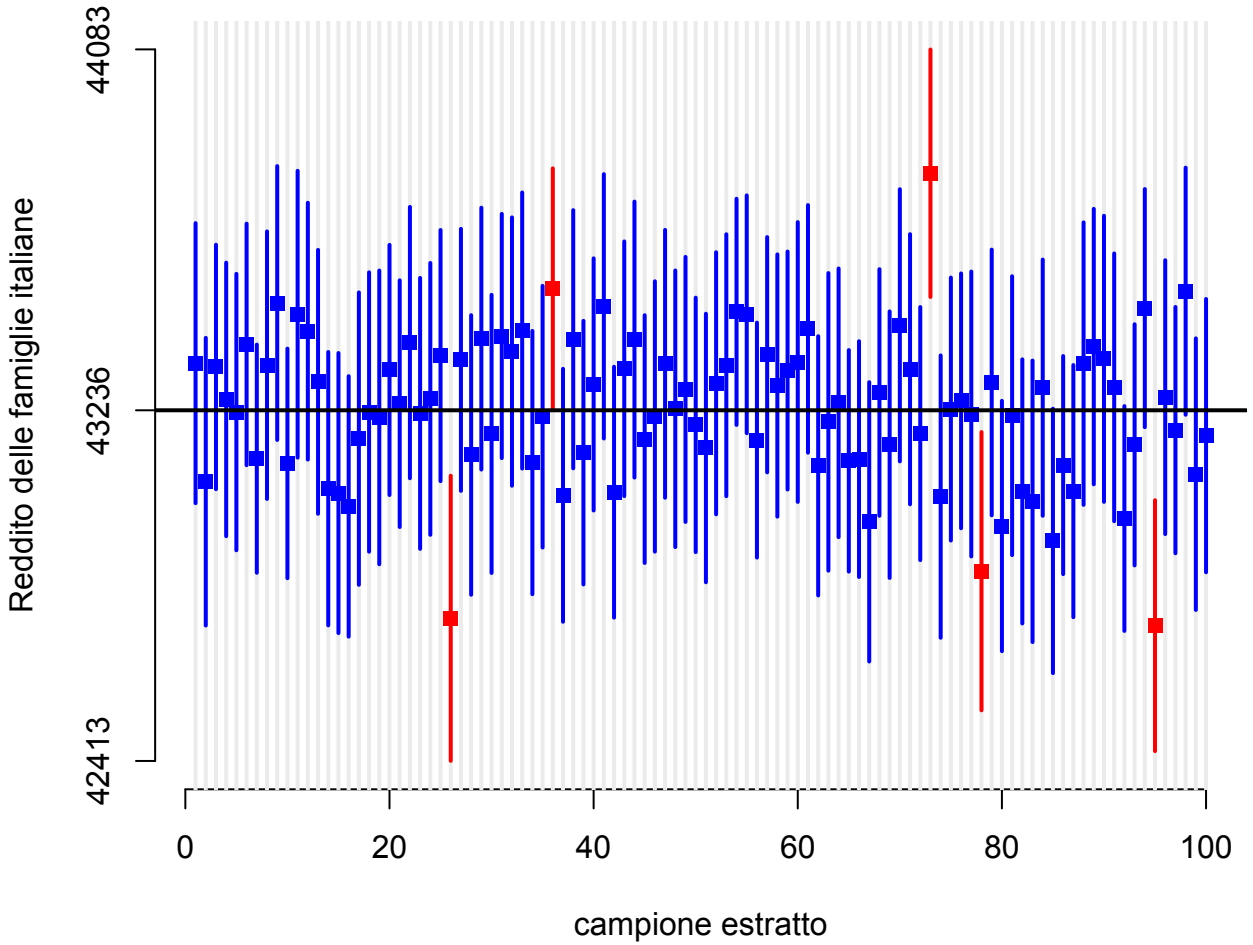
  • I segmenti rossi rappresentano gli intervalli di confidenza al 95% che NON contengono la media della popolazione.

Simulazione su 100 campioni

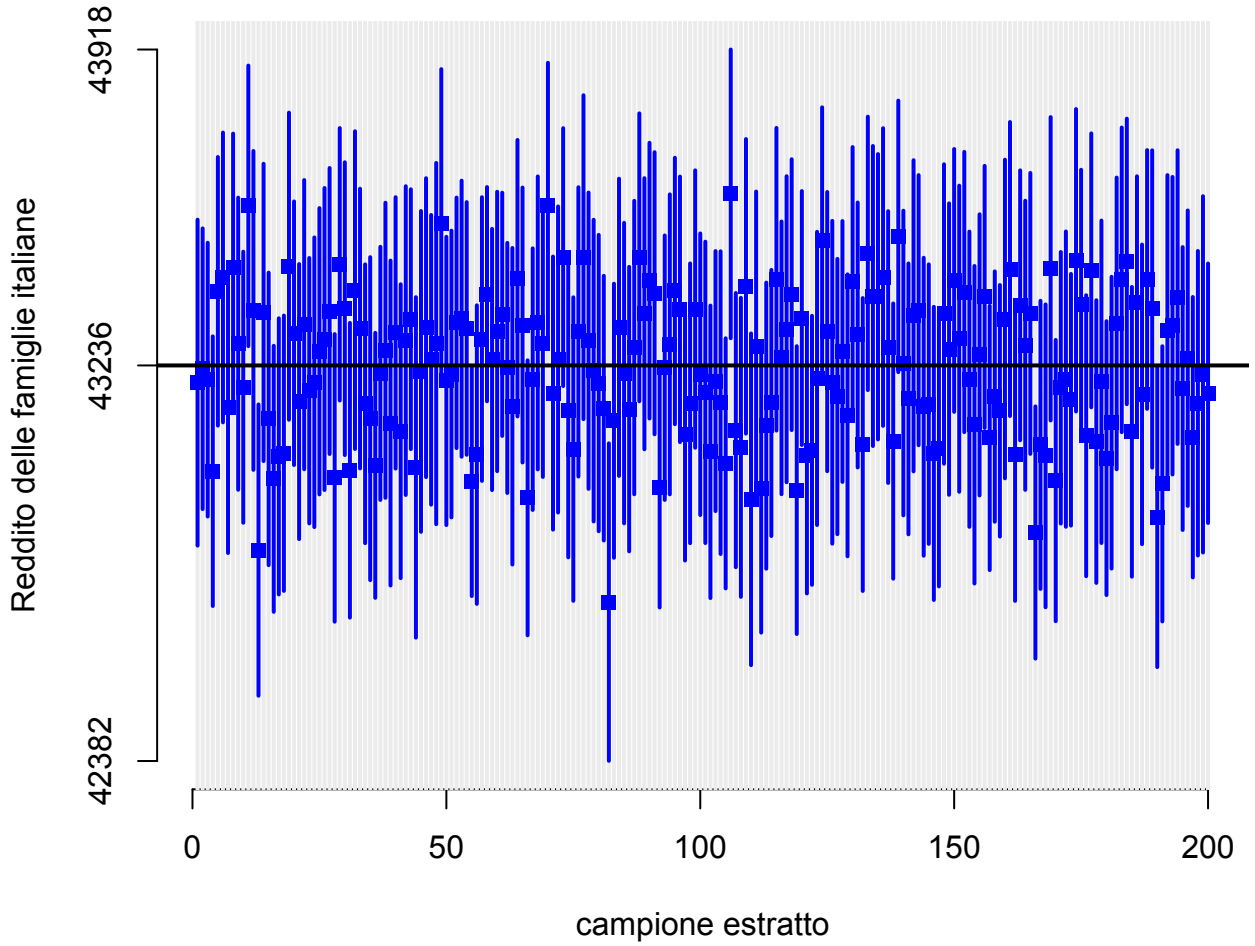


Simulazione su 100 campioni

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

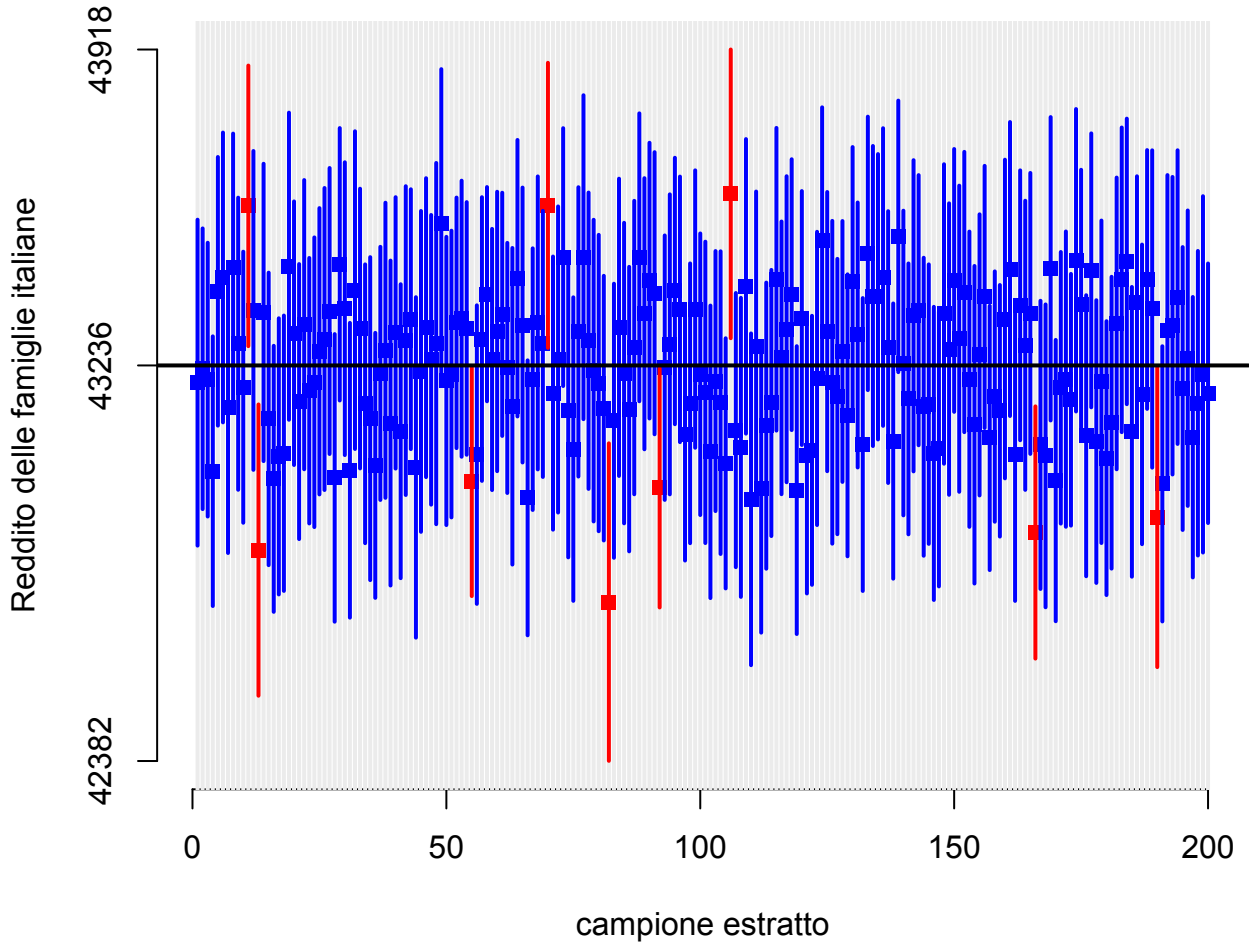


Simulazione su 200 campioni



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 $\alpha = 0.05$.

Interpretazione errata

Le seguenti interpretazioni dell'intervallo di confidenza NON sono corrette.

- La media μ della popolazione ha una probabilità 0.95 di essere contenuta nell'intervallo

$$\bar{x} \pm t_{(\alpha/2, n-1)} \hat{\sigma}_{\bar{x}} = 41100 \pm 2768$$

  La media della popolazione è contenuta nell'intervallo (nel qual caso la probabilità che sia compresa nell'intervallo è 1), oppure non è contenuta nell'intervallo (nel qual caso la probabilità che sia contenuta nell'intervallo è 0).

Interpretazione errata

- In questo esempio, noi sappiamo che la media della popolazione (€43236) è contenuta nell'intervallo di confidenza (dato che abbiamo costruito l'esempio in questo modo), ma in qualsiasi applicazione concreta il ricercatore non può saperlo.

Interpretazione errata

- Una famiglia scelta a caso dalla popolazione ha probabilità 0.95 di avere un reddito compreso nell'intervallo

$$\bar{x} \pm t_{(\alpha/2, n-1)} \hat{\sigma}_{\bar{x}} = 41100 \pm 2768$$

La distribuzione dei valori x_i nella popolazione ha media μ (non $\bar{x}$), deviazione standard σ (non $s/\sqrt{100}$) e non è necessariamente una distribuzione normale (tanto da usare $t \to N$).

Interpretazione errata

- La media del campione $\bar{x}$ ha probabilità 0.95 di essere un contenuta nell'intervallo

$$\bar{x} \pm t_{(\alpha/2, n-1)} \hat{\sigma}_{\bar{x}} = 41100 \pm 2768$$

Questa è un'affermazione priva di senso in quanto la media $\bar{x} = 41100$ di questo particolare campione è certamente contenuta all'interno dell'intervallo, dato che l'intervallo è stato costruito attorno a questo valore.

Interpretazione errata

- Se campioni di numerosità $n = 100$ venissero ripetutamente estratti dalla popolazione, il 95% delle medie campionarie $\bar{x}$ sarebbe contenuto nell'intervallo

$$\bar{x} \pm t_{(\alpha/2, n-1)} \hat{\sigma}_{\bar{x}} = 41100 \pm 2768$$

La distribuzione campionaria di $\bar{x}$ è centrata sulla media della popolazione μ e non sulla media del campione $\bar{x}$. Inoltre, la deviazione standard della distribuzione campionaria della media è $\frac{\sigma}{\sqrt{n}}$ e non $\frac{s}{\sqrt{n}}$. Quindi, il 95% delle medie dei campioni è contenuta nell'intervallo

$$\mu \pm 1.96 \times \sigma_{\bar{x}} = 43236 \pm 1.96 \frac{15500}{\sqrt{(100}} = 43236 \pm 3038$$

Livelli di fiducia

- Fino ad ora abbiamo costruito l'intervallo di confidenza per μ al 95%.
- Questo è l'intervallo di confidenza più comunemente usato.
- Sono però comuni anche intervalli di confidenza al 90% e al 99%.

Livelli di fiducia

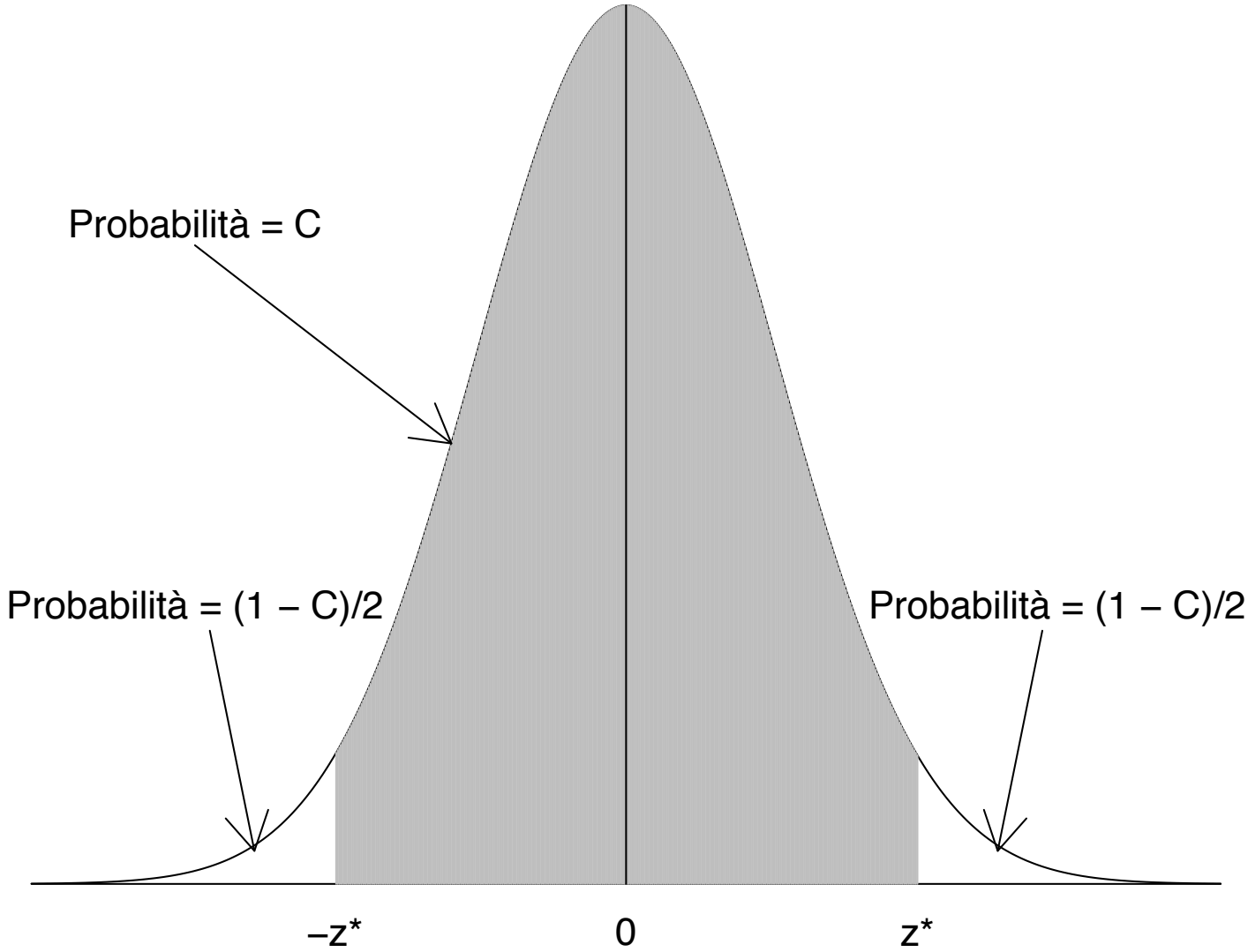
In generale, per costruire un intervallo di confidenza per μ al livello del $100(1-\alpha)\%$, si usa la formula:

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

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



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:

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

Livelli di fiducia

- 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

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

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

Numerosità del campione

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).

Numerosità del campione

- Affinché la formula $\left(z^* \frac{\sigma}{m}\right)^2$ sia accurata, è necessario che i dati provengano da un *campione casuale semplice estratto da una popolazione di grandi dimensioni*.
- La formula precedente non è corretta per campionamenti complessi, come il campionamento stratificato.
- Non c'è un metodo corretto per costruire gli intervalli di confidenza per dati provenienti da un campionamento non casuale.

Numerosità del campione

- Se la popolazione non è distribuita normalmente, allora la formula precedente potrebbe non essere accurata se n è molto piccolo.
- L'intervallo di confidenza, così come la media $\bar{x}$, può essere fortemente influenzato dai valori anomali o *outliers*.

Intervallo di confidenza della mediana

Intervallo di confidenza della mediana

- Quando la popolazione è asimmetrica, la media non è un utile indice di tendenza centrale. In queste condizioni, la mediana (M_e) è preferibile.
- Oppure, anche in caso di popolazione simmetrica, se nel campione esistono valori anomali, allora è possibile utilizzare la mediana come descrittore di tendenza centrale. Abbiamo visto come nelle distribuzioni simmetriche, Moda, Media e Mediana coincidano.
- Come la media, anche la mediana è dotata di una distribuzione campionaria.

Intervallo di confidenza della mediana

- Considereremo due casi:
- mediana calcolata su un campione di grandi dimensioni proveniente da una popolazione normale;
- mediana calcolata su un campione di grandi dimensioni proveniente da una popolazione non normale.

Intervallo di confidenza della mediana

- La distribuzione campionaria della mediana $\hat{M}_e$ ottenuta estraendo i campioni da una popolazione con funzione di densità $f(x)$ è asintoticamente $(n \to \infty)$ normale, con valore atteso pari alla mediana della popolazione,

$$E(\widehat{Me}_i) = M_e = \mu,$$

e varianza

$$Var(\widehat{Me}) = \frac{1}{4nf(Me)^2},$$

dove $f(M_e)$ è la funzione di densità della popolazione a livello della mediana (uguale a μ nelle distribuzioni simmetriche) e n è la grandezza del campione.

Intervallo di confidenza della mediana

- In pratica, la funzione di densità della popolazione a livello della mediana, $f(M_e)$, non è nota; tuttavia se si può assumere come normale, abbiamo

$$
\begin{aligned}
f(Me = \mu; \mu; \sigma) = & \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{ -\frac{1}{2} \frac{(\mu - \mu)^2}{\sigma^2} \right\}} \\
= & \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{ -\frac{1}{2} \frac{0}{\sigma^2} \right\}} \\
= & \frac{1}{\sqrt{2\pi\sigma^2}} = \frac{1}{\sqrt{2\pi}} \times \frac{1}{\sqrt{\sigma^2}} \\
= & 0.3989423 \times \frac{1}{\sqrt{\sigma^2}} \\
\approx & 0.4 \times \frac{1}{\sqrt{\sigma^2}}
\end{aligned}
$$

Intervallo di confidenza della mediana

- Il valore di ordinata massima (dove $f(x)$ raggiunge il valore massimo) delle funzioni di densità di probabilità normali è quindi

$$0.4 \times \frac{1}{\sqrt{\sigma^2}}$$

che si semplifica a 0.4 nel caso della normale standard, dove $\sigma^2 = 1$.

```
 dnorm(0,mean=0,sd=1)
## [1] 0.3989423
```

- Inoltre, nelle normali *non standard* viene determinato esclusivamente dal parametro $\sigma^2 \neq 1$

```
 dnorm(70,mean=70,sd=5);0.3989423 * 1/(5)
## [1] 0.07978846
## [1] 0.07978846
 dnorm(43236,mean=43236,sd=15500); 0.3989423 * 1/(15500)
## [1] 2.573821e-05
## [1] 2.573821e-05
```

Intervallo di confidenza della mediana

- La varianza della distribuzione campionaria della mediana, secondo l'assunzione di normalità nella popolazione, diventa

$$\widehat{Var(Me)} = \frac{1}{4n\left(0.4 \times 1/\sqrt{\sigma^2}\right)^2} = \frac{\sigma^2}{4n(0.4)^2} = \frac{1}{4(0.4)^2} \times \frac{\sigma^2}{n}$$

$$= 1.5625\frac{\sigma^2}{n}$$

$$= 1.25^2\sigma^2_{\bar{x}}$$

Intervallo di confidenza della mediana

- La formula precedente rivela che la mediana è meno efficiente della media quale operatore di tendenza centrale (ha una varianza campionaria 1.25^2 volte maggiore).
- Nelle condizioni sopra esposte (popolazione normale, campione di grandi dimensioni) l'intervallo di confidenza della mediana è dunque più grande del 25% dell'intervallo di confidenza della media e si calcola come:

$$Me \pm z^* 1.25 \frac{s}{\sqrt{n}}$$

laddove z^* dipende dal livello di confidenza richiesto.

Intervallo di confidenza della mediana

- Esaminiamo i risultati di una simulazione.
- Calcoliamo empiricamente la varianza della distribuzione campionaria di media e mediana di un campione casuale di $n = 200$ osservazioni, estratto (50000 volte) da una popolazione normale con media $\mu = 100$ e deviazione standard $\sigma = 36$.

Intervallo di confidenza della mediana

```
rep <- 50000
n <- 200
 DistrCampMedia <- rep(0, rep)
 DistrCampMediana <- rep(0, rep)
 for (i in 1:rep) {
 samp <- rnorm(n, mean=100, sd=36)
 DistrCampMedia[i]<-mean(samp)
 DistrCampMediana[i]<-median(samp)
 }

var.media <- var(DistrCampMedia)*(n-1)/n
var.media
## 6.499387
```

Intervallo di confidenza della mediana

I risultati confermano quello che ci aspettiamo:

```
var.media <- var(DistrCampMedia)*(n-1)/n
var.media
## 6.499387

36^2/200
## 6.48
```

la varianza della distribuzione campionaria della media è uguale al rapporto tra la varianza della popolazione e la numerosità del campione.

Intervallo di confidenza della mediana

Vediamo ora cosa è successo per la mediana calcolata sui medesimi campioni.

```
var.mediana <- var(DistrCampMediana)*(n-1)/n
var.mediana
## 10.06176

1/(4* dnorm(0,0,1)^2) * 36^2/200
## 10.17876
```

Intervallo di confidenza della mediana

- Di quanto è più grande la deviazione standard della distribuzione campionaria della mediana della deviazione standard della distribuzione campionaria della media?

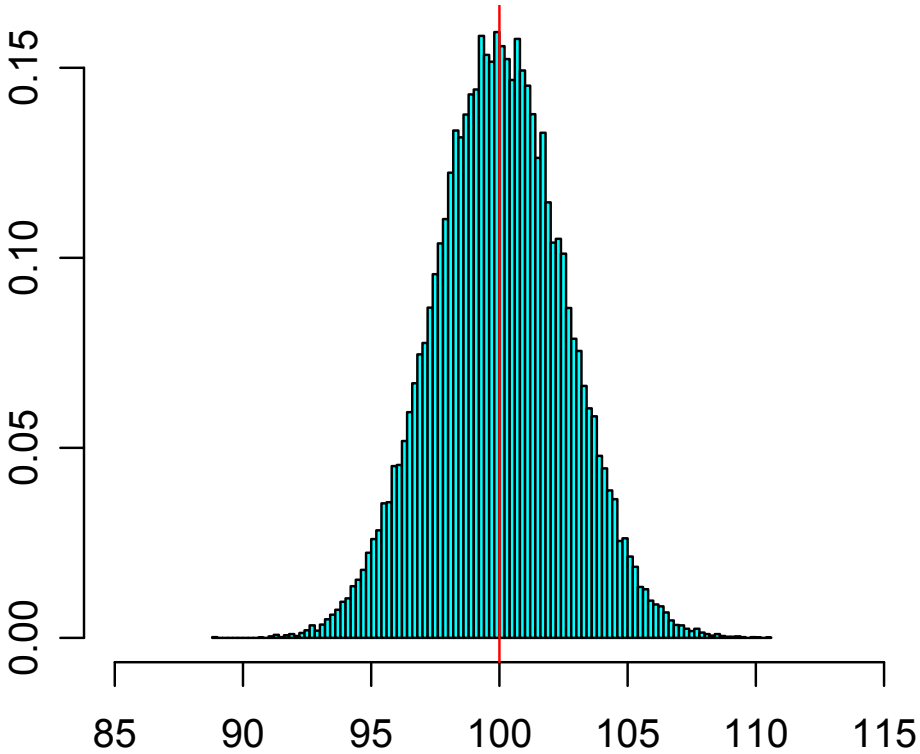
```
 sqrt(var.mediana)/sqrt(var.media)
## [1] 1.24423
```

```
 var.mediana/var.media
## [1] 1.548109
 sqrt(1.548109)
## [1] 1.24423
```

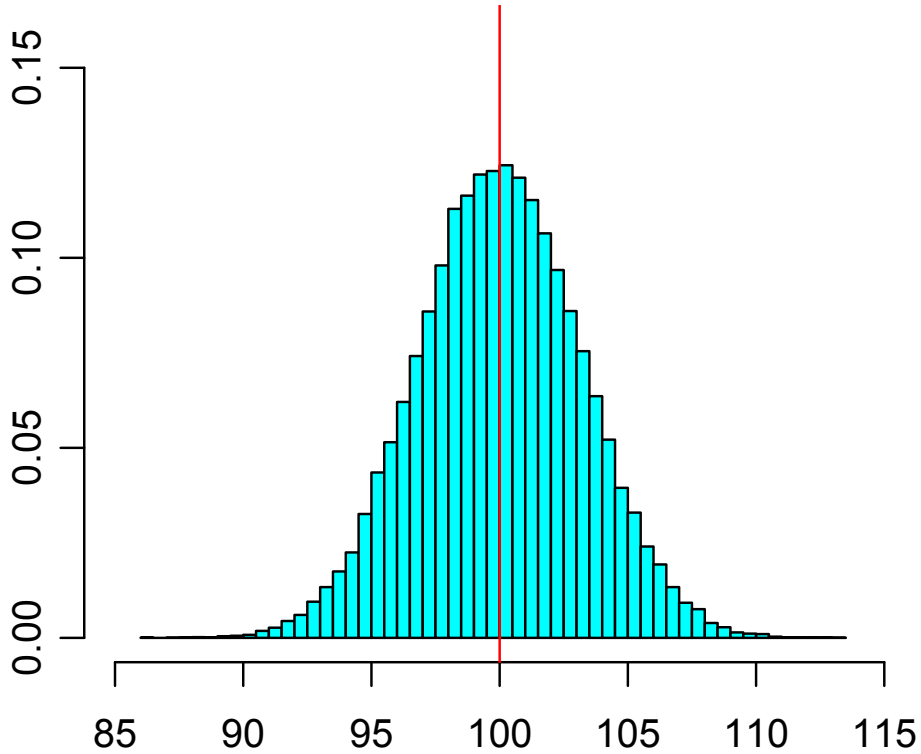
Come vi ho mostrato, quando n è grande e la popolazione si assume essere normale, la varianza della distribuzione campionaria della mediana è

$$V(Me_i) = 1.25^2 \frac{\sigma^2}{n} \approx 10.17876$$

Intervallo di confidenza della mediana



Distribuzione campionaria della Media



e della Mediana

Intervallo di confidenza della mediana

- La simulazione precedente rivela che la mediana è meno efficiente della media quale operatore di tendenza centrale (ha una varianza campionaria maggiore).
- Nelle condizioni sopra esposte (popolazione normale, campione di grandi dimensioni) l'intervallo di confidenza della mediana è dunque più grande del 25% dell'intervallo di confidenza della media e si calcola come:

$$Me \pm z^* 1.25 \frac{s}{\sqrt{n}}$$

Intervallo di confidenza della mediana

- Le considerazioni precedenti ci sono di poco aiuto, però, dato che, nel caso di una popolazione normale, la tendenza centrale della distribuzione verrà misurata utilizzando la media.
- Una eccezione potrebbe essere un campione che si assume essere distribuito normalmente, almeno secondo la popolazione di partenze, ma che invece presenta numerosi valori anomali, tanto da indurre il ricercatore ad utilizzare la mediana come stimatore di μ.
- Chiediamoci dunque come si calcola l'intervallo di confidenza della mediana nel caso di una popolazione asimmetrica (ovvero, quando non possiamo assumere che la popolazione sia distribuita normalmente).

Intervallo di confidenza della mediana

Agresti e Finaly: Tempo mediano di permanenza di un libro sullo scaffale di una biblioteca.

> *How long has it been since a typical book has been checked out?*
>
> *We suspected that the distribution of variables of this type may be heavily skewed to the right, so we used the median to describe central tendency.*

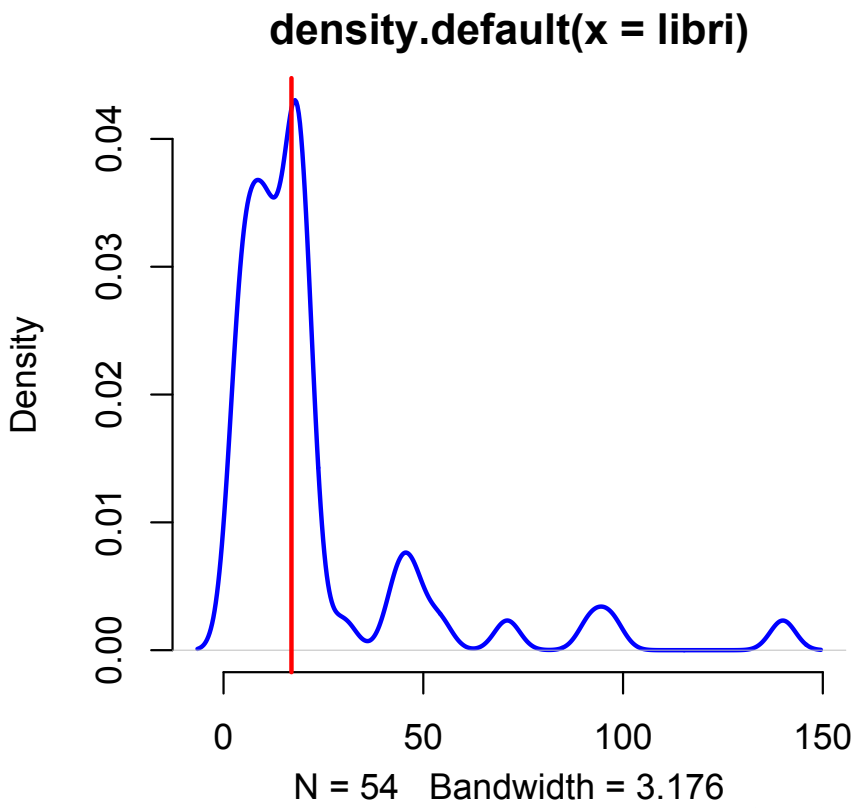
• ecco i dati di un campione casuale di 54 libri dalla collezione della biblioteca dell'Università della Florida:

```
libri <- c(3, 30, 19, 140, 5, 97, 4, 19, 13, 19, 92,
 9, 17, 5, 19, 22, 10, 11, 10, 71, 11, 44, 4, 7,
 47, 8, 11, 21, 20, 10, 19, 6, 5, 18, 12, 15, 10,
 19, 7, 14, 18, 17, 20, 54, 5, 13, 17, 18, 3, 19,
 43, 17, 48, 4)
```

Intervallo di confidenza della mediana

La mediana del tempo (anni) di permanenza sugli scaffali e la distribuzione (attesa) asimmetrica positiva:

```
median(libri)
## [1] 17
plot(density(libri))
```



Intervallo di confidenza della mediana

- Un intervallo di fiducia per la mediana, valido per grandi campioni, e che non richiede alcuna assunzione relativa alla popolazione (se non che sia una variabile aleatoria continua) è dato da:

$$\frac{n+1}{2} \pm n\left(z^* \frac{0.50}{\sqrt{n}}\right) = \frac{n+1}{2} \pm z^*(0.50)\sqrt{n}.$$

- Per il nostro esempio, un intervallo di fiducia al 95% della mediana corrisponde alle posizioni:

$$\frac{54+1}{2} \pm 1.96(0.50)\sqrt{54} = 27.5 \pm 7.02 = (20.3,\ 34.7)$$

che corrisponde ai valori della variabile `libri` $(11,\ 19)$

```
 sort(libri)[20]; sort(libri)[35]
## [1] 11
## [1] 19
```

Intervallo di confidenza della mediana

- Possiamo attribuire un livello di fiducia del 95% all'affermazione che la mediana del tempo di giacenza dei libri in biblioteca è di almeno 11 anni e non superiore ai 19 anni.
- Anche in questo caso, se volessimo ridurre il margine di errore $m = (z^*0.50/\sqrt{n})$ dovremmo raccogliere un campione n più grande.
- Un risultato simile, mediante un algoritmo diverso, si ottiene utilizzando la funzione `wilcox.test(x,conf.int=TRUE)`:

Intervallo di confidenza della mediana

```
 wilcox.test(libri,conf.int=TRUE)

Wilcoxon signed rank test with continuity correction

data:  libri
V = 1485, p-value = 1.639e-10
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 12.50007 19.50006
sample estimates:
(pseudo)median
      15.50004
```

Conclusioni

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

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

Conclusioni

- 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.

Conclusioni

Parametro	Stima puntuale	Errore standard stimato	Intervallo di confidenza	n e margine d'errore m
Popolazione normale – grandi campioni				
μ	$\bar{x}$	$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$	$\hat{\bar{x}} \pm z\hat{\sigma}_{\bar{x}}$	$n = \left(\frac{zs}{m}\right)^2$
Me	$\widehat{Me}$	$\hat{\sigma}_{Me} = 1.25\frac{s}{\sqrt{n}}$	$Me \pm z\hat{\sigma}_{Me}$	$n = \left(\frac{1.25 \times zs}{m}\right)^2$
Popolazione continua – grandi campioni				
Me	$\widehat{Me}$	$\hat{\sigma}_{50\%} = 0.50/\sqrt{n}$	$\frac{n+1}{2} \pm n\hat{z}\sigma_{Me}$	$n = \left(\frac{0.50 \times z}{m}\right)^2$