STATISTICA

Inferenza: Stima & Intervalli di confidenza, 1

Inferenza per la media

Siano $X_1, X_2, \cdots, X_n$ variabili casuali ***i.i.d***

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

media campionaria: v.c. che *predice* il valore della media aritmetica dei dati nel campione

E' uno *stimatore* della media $\boldsymbol{E(X_i)}$

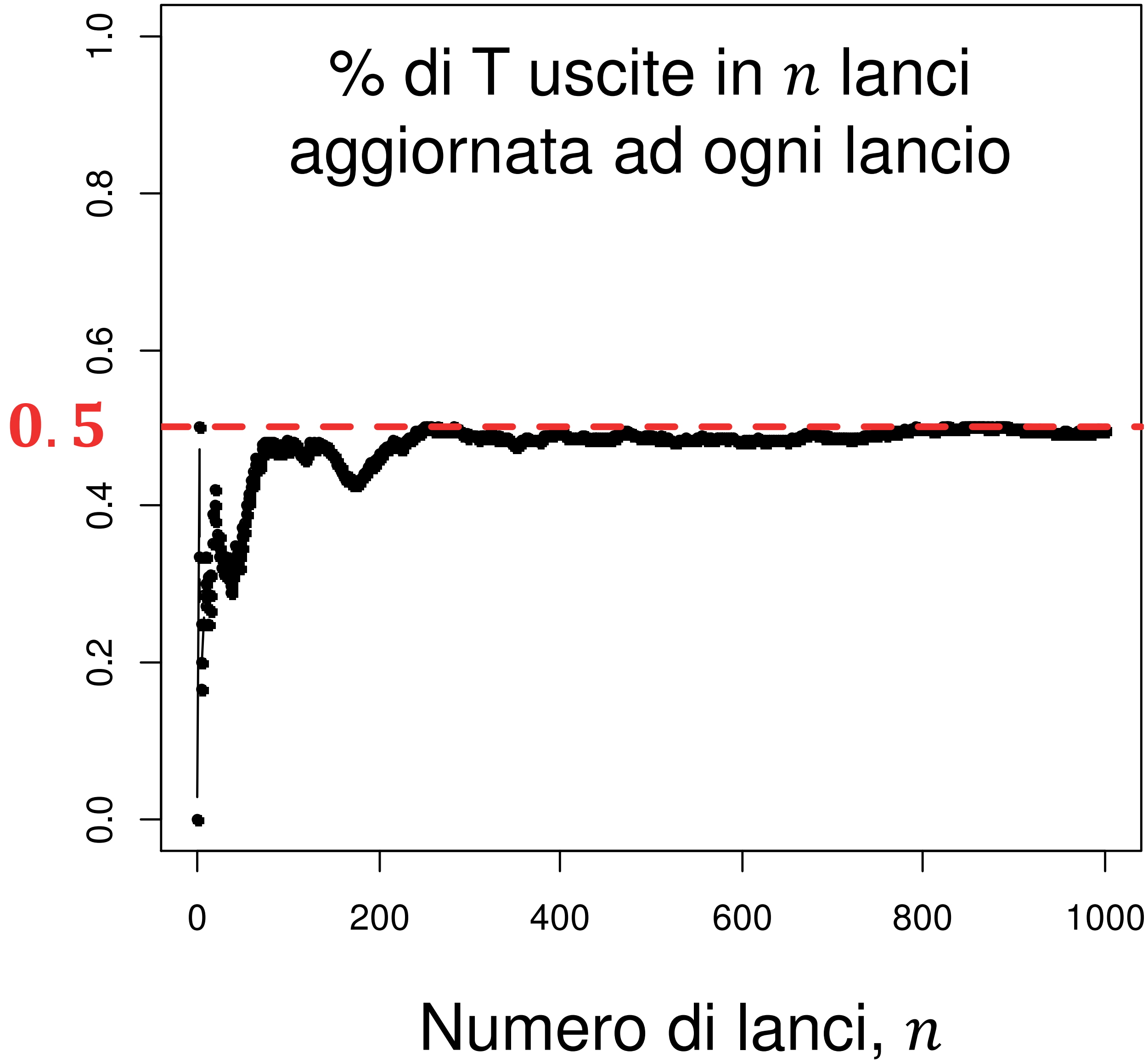
$X_1, \dots, X_n, \dots$ v.c. **i.i.d** con media $E(X_i) = \mu$:

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n} \xrightarrow[n \to \infty]{} \mu$$

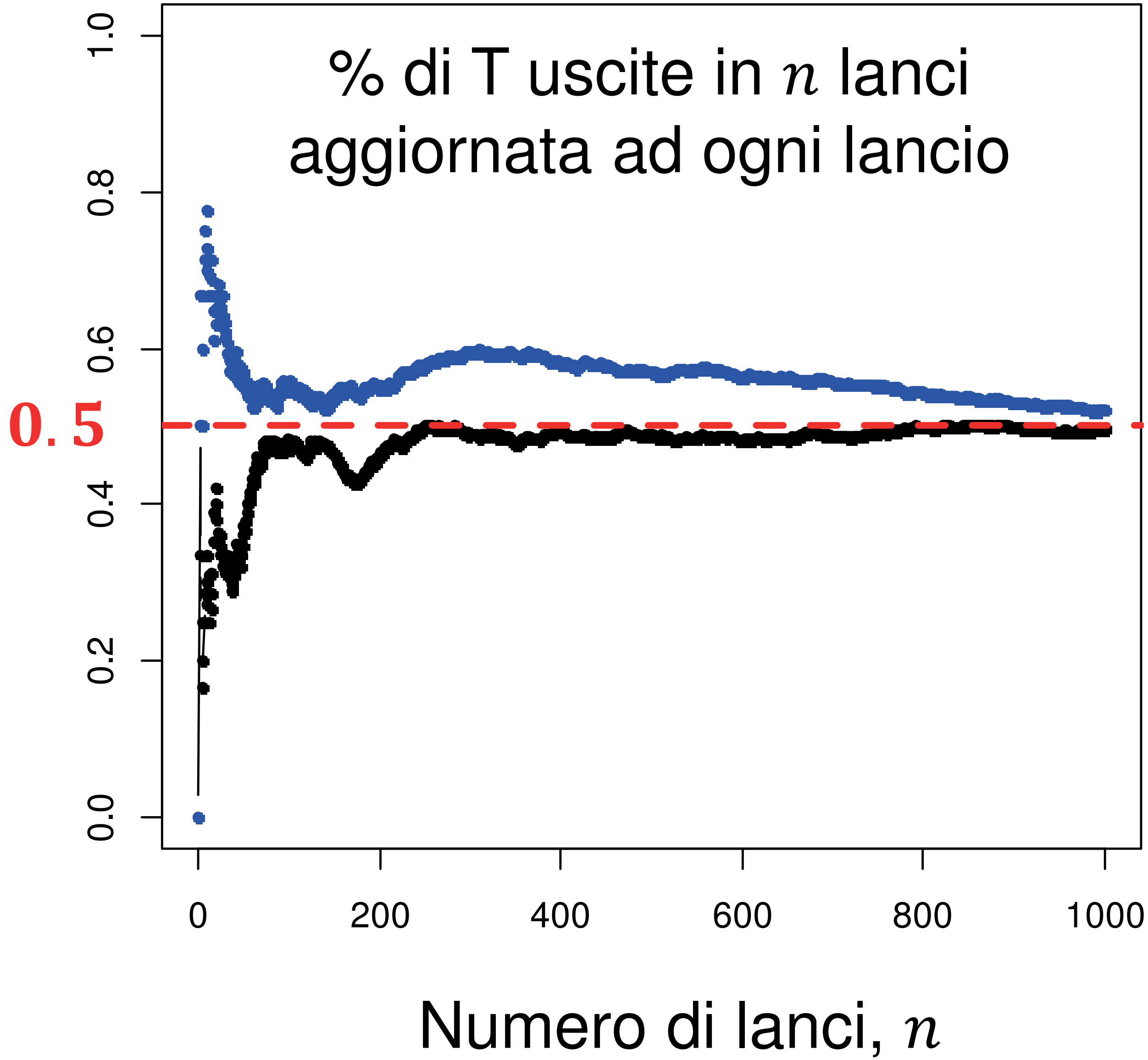
(quasi certamente)

Legge dei grandi numeri

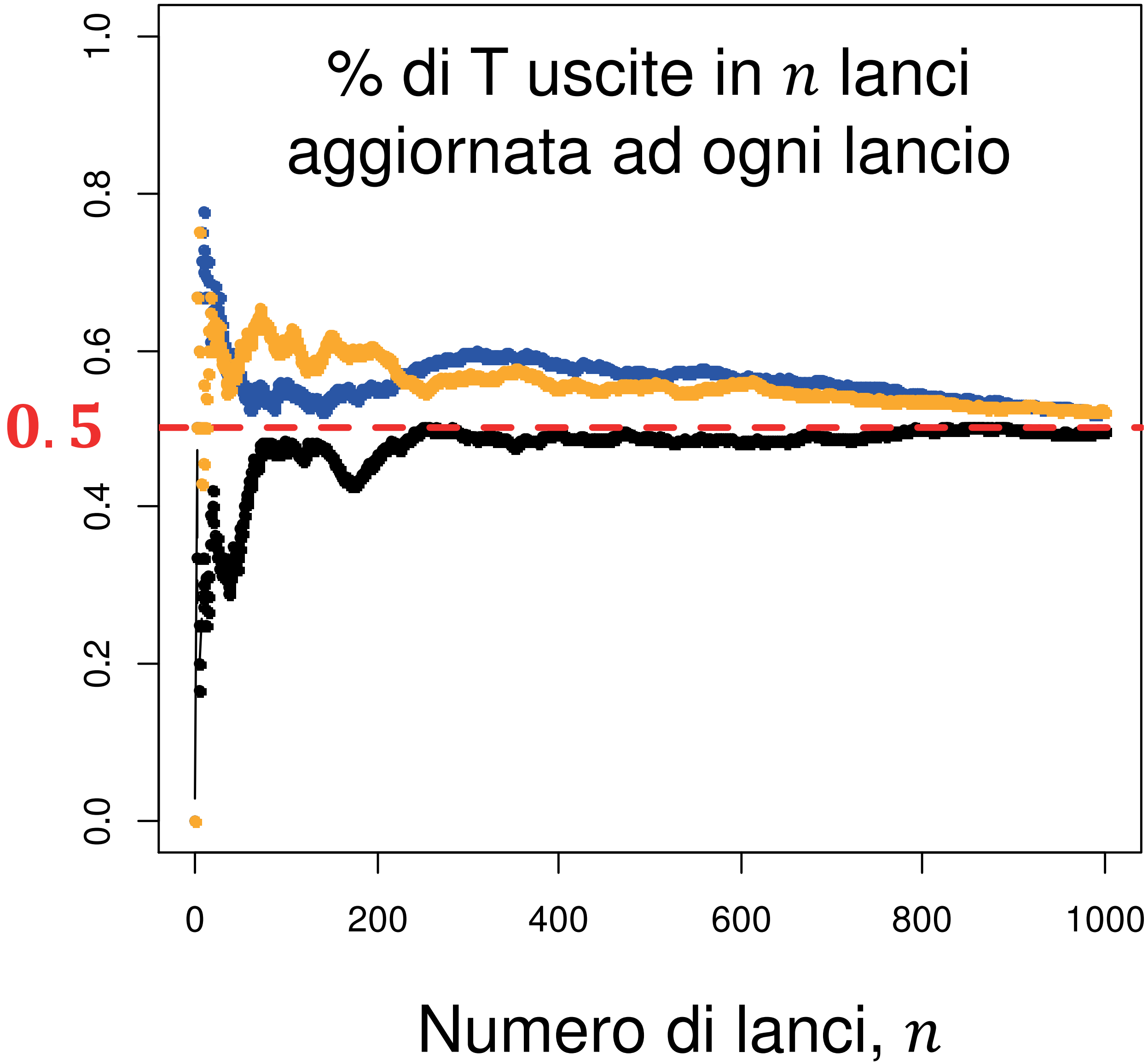
Esempio


% di T uscite in n lanci
aggiornata ad ogni lancio
0.5
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Numero di lanci, n

Esempio

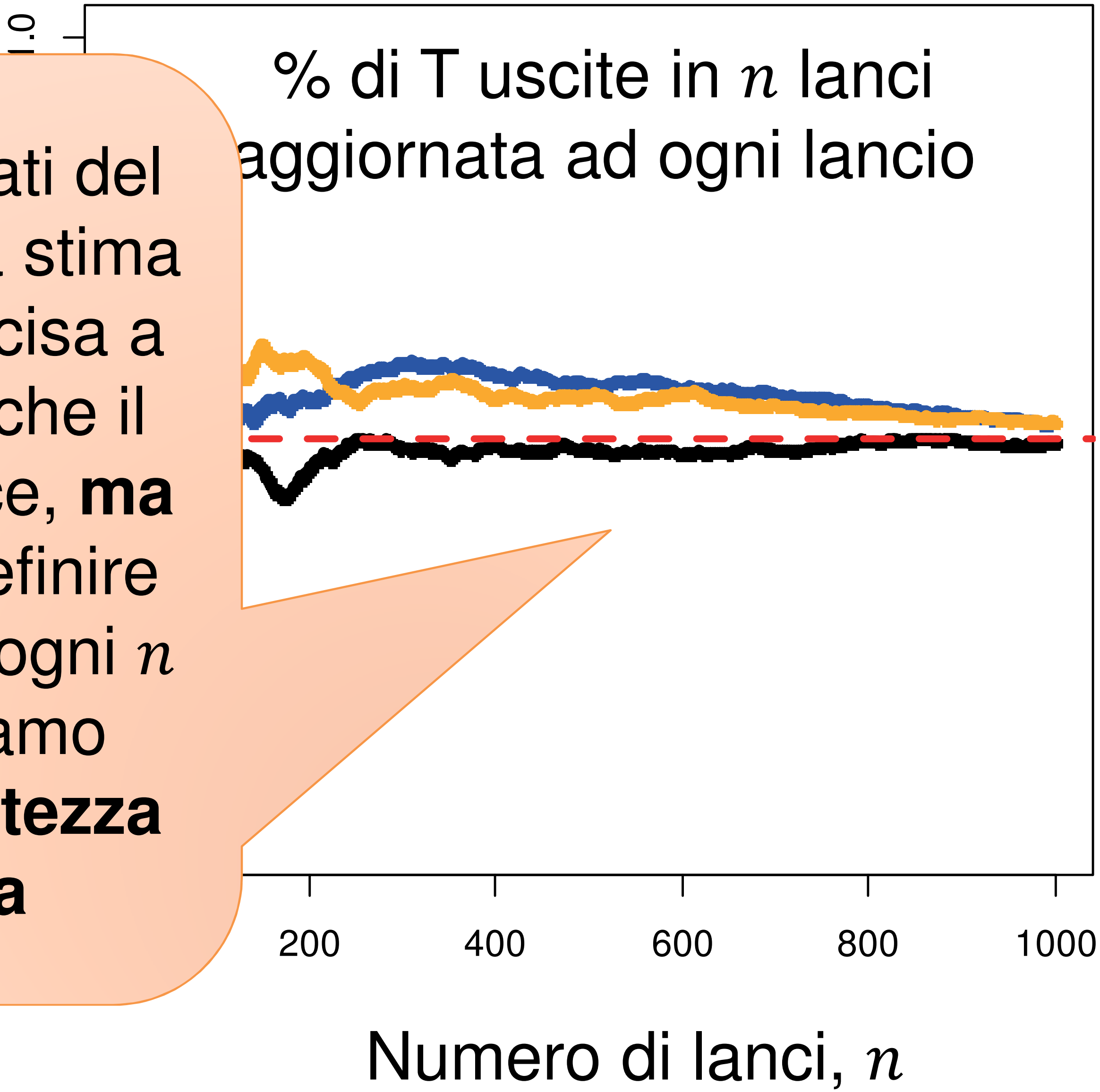

% di T uscite in n lanci
aggiornata ad ogni lancio
1.0
0.8
0.6
0.5
0.4
0.2
0.0
0
200
400
600
800
1000
Numero di lanci, n

Esempio



% di T uscite in n lanci
aggiornata ad ogni lancio
1.0
0.8
0.6
0.5
0.4
0.2
0.0
0
200
400
600
800
1000
Numero di lanci, n

Esempio



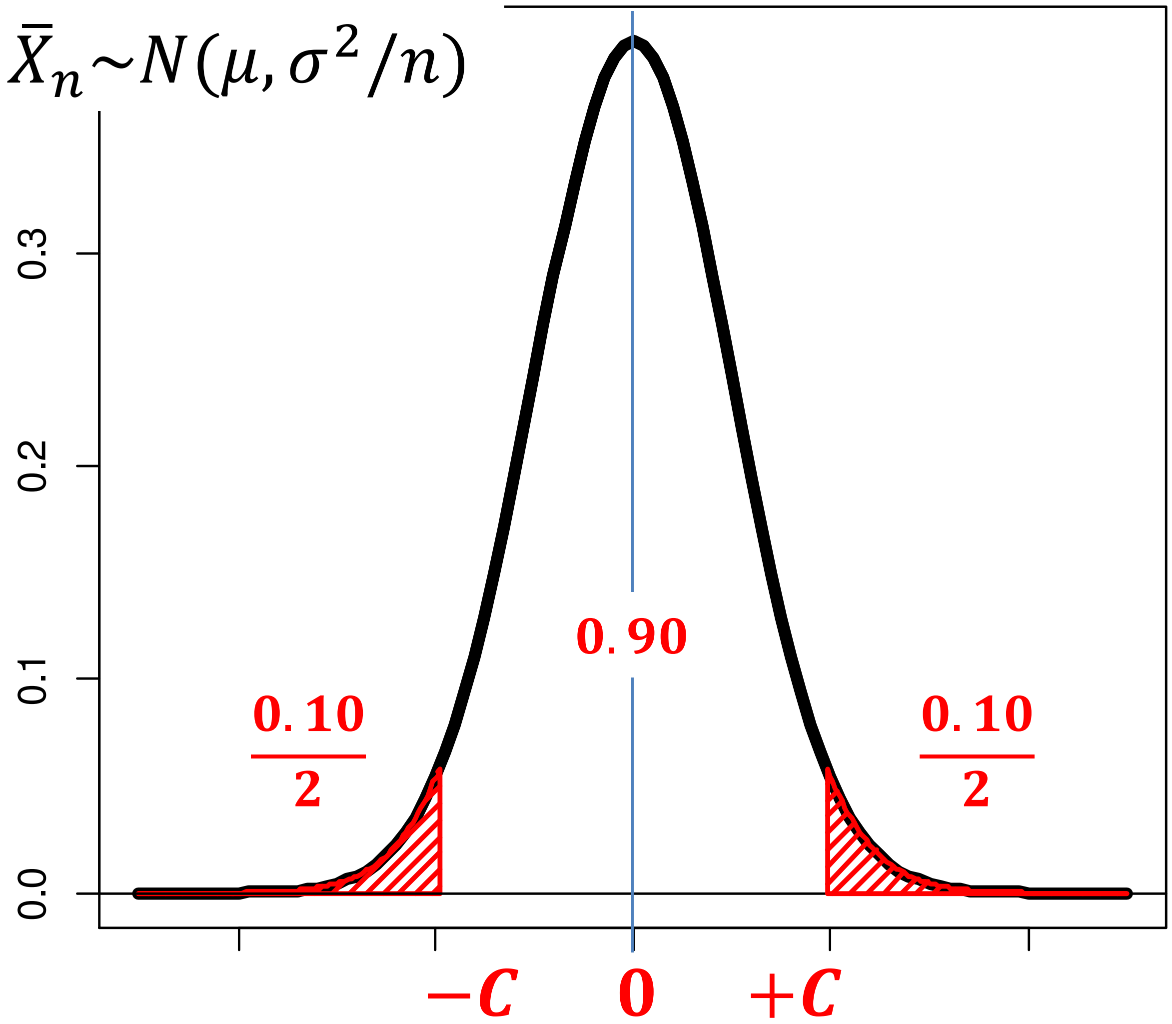
La media dei dati del campione è una stima sempre più precisa a mano a mano che il campione cresce, **ma** non potendo definire una soglia, per ogni n fissato dobbiamo valutare **l'incertezza della stima**

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

$$P\left(\mu - C \times \sqrt{\frac{\sigma^2}{n}} \leq \bar{X}_n \leq \mu + C \times \sqrt{\frac{\sigma^2}{n}}\right) = 0.90$$

$$P\left(\frac{\mu - C \times \sqrt{\frac{\sigma^2}{n}} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq \frac{\mu + C \times \sqrt{\frac{\sigma^2}{n}} - \mu}{\sqrt{\frac{\sigma^2}{n}}}\right) = 0.90$$

$$P(-C \leq Z \leq C) = 0.90$$



$$P(-C \leq Z \leq C) = 0.90$$

$$C = z_{1-\frac{0.10}{2}}$$

qualunque sia μ !

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

$$P\left(\mu - C \times \sqrt{\frac{\sigma^2}{n}} \leq \bar{X}_n \leq \mu + C \times \sqrt{\frac{\sigma^2}{n}}\right) = 0.90$$

$$P\left(\frac{\mu - C \times \sqrt{\frac{\sigma^2}{n}} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq \frac{\mu + C \times \sqrt{\frac{\sigma^2}{n}} - \mu}{\sqrt{\frac{\sigma^2}{n}}}\right) = 0.90$$

$$P(-C \leq Z \leq C) = 0.90 \Leftrightarrow \quad \boldsymbol{C = z_{1-\frac{0.10}{2}}}$$

$$\boldsymbol{P\left(\bar{X}_n - z_{1-\frac{0.10}{2}}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{0.10}{2}}\sqrt{\frac{\sigma^2}{n}}\right) = 0.90}$$

$\bar{X}_n \sim N(\mu, \sigma^2/n)$

abbiamo il 90% di probabilità che, quando prenderemo il nostro campione casuale (di numeri!), la media $\bar{\boldsymbol{x}}_{\boldsymbol{n}}$ del campione sarà vicina al valore da stimare, μ, nel senso che

con il 90% di probabilita' l'intervallo conterrà μ

$$\left(\bar{\boldsymbol{x}}_{\boldsymbol{n}} - \boldsymbol{z}_{\mathbf{1}-\frac{\mathbf{0.10}}{\mathbf{2}}}\sqrt{\frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\boldsymbol{n}}}\ , \bar{\boldsymbol{x}}_{\boldsymbol{n}} + \boldsymbol{z}_{\mathbf{1}-\frac{\mathbf{0.10}}{\mathbf{2}}}\sqrt{\frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\boldsymbol{n}}}\right)$$

$$\leq Z \leq C) = 0.90 \Leftrightarrow \quad C = \boldsymbol{z}_{\mathbf{1}-\frac{\mathbf{0.10}}{\mathbf{2}}}$$

$$\boldsymbol{P}\left(\bar{\boldsymbol{X}}_{\boldsymbol{n}} - \boldsymbol{z}_{\mathbf{1}-\frac{\mathbf{0.10}}{\mathbf{2}}}\sqrt{\frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\boldsymbol{n}}} \leq \boldsymbol{\mu} \leq \bar{\boldsymbol{X}}_{\boldsymbol{n}} + \boldsymbol{z}_{\mathbf{1}-\frac{\mathbf{0.10}}{\mathbf{2}}}\sqrt{\frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\boldsymbol{n}}}\right) = \mathbf{0.90}$$

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da

$$\left(\bar{x}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\ , \bar{x}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right)$$

$$\bar{x}_n = 9.75$$

$$\alpha = 0.05$$

$$z_{0.975} = 1.96$$

$$n = 30$$

Esempio

La quantità di arsenico nell’acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall’acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L’intervallo di confidenza del 95% per μ, basato sul campione, è dato da

$$\bar{x}_n = 9.75$$
$$\alpha = 0.05$$
$$z_{0.975} = 1.96$$
$$n = 30$$

$$\left(\bar{x}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\,,\ \bar{x}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right)$$

$$\left(9.75 - 1.96\sqrt{\frac{3.5}{30}}\,, 9.75 + 1.96\sqrt{\frac{3.5}{30}}\right)$$

$$(9.08, 10.42)\ \mu g/l$$

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da $\mathbf{(9.08, 10.42)}\ \mu g/l$. Se il limite imposto dalla legge è di 10 $\mu g/l$, possiamo stare tranquilli?

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da $\mathbf{(9.08, 10.42)}\ \mu g/l$. Se il limite imposto dalla legge è di 10 $\mu g/l$, possiamo stare tranquilli?

Mica tanto! Con una confidenza del 95%, il vero valore di μ **nella popolazione di riferimento** (ogni possibile campione dell'acqua di Milano) può anche essere > 10 (fino a 10.42).

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da $\mathbf{(9.08, 10.42)}\ \mu g/l$. Se il limite imposto dalla legge è di 10 $\mu g/l$, possiamo stare tranquilli? E se aumentiamo il livello di confidenza da 95% a 99%?

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da $\mathbf{(9.08, 10.42)}\ \mu g/l$. Se il limite imposto dalla legge è di 10 $\mu g/l$, possiamo stare tranquilli? E se aumentiamo il livello di confidenza da 95% a 99%?

$$\left(\bar{x}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\ , \bar{x}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right)$$

$\alpha = 0.05$ $\qquad$ $\alpha = 0.01$

$z_{0.975} = 1.96$ $\qquad$ $z_{0.995} = ???$

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da $\mathbf{(9.08, 10.42)}\ \mu g/l$. Se il limite imposto dalla legge è di 10 $\mu g/l$, possiamo stare tranquilli? E se il campione fosse stato più grande, $n = 50$?

$$\left(\bar{\boldsymbol{x}}_n - \boldsymbol{z}_{1-\frac{\alpha}{2}} \sqrt{\frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}} \;, \bar{\boldsymbol{x}}_n + \boldsymbol{z}_{1-\frac{\alpha}{2}} \sqrt{\frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}} \right)$$

Esempio

La quantità di arsenico nell'acqua potabile ha distribuzione $N(\mu, 3.5)$. In un campione di 30 prelievi dall'acquedotto di Milano abbiamo ottenuto un contenuto medio di arsenico pari a $\bar{x}_n = 9.75\ \mu g/l$. L'intervallo di confidenza del 95% per μ, basato sul campione, è dato da $\mathbf{(9.08, 10.42)}\ \mu g/l$. Se il limite imposto dalla legge è di 10 $\mu g/l$, possiamo stare tranquilli? E se il campione fosse stato più grande, $n = 50$?

$$\left(\bar{\boldsymbol{x}}_{\boldsymbol{n}} - \boldsymbol{z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}}\ ,\ \bar{\boldsymbol{x}}_{\boldsymbol{n}} + \boldsymbol{z}_{1-\frac{\alpha}{2}}\sqrt{\frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}}\right)$$

$n = 30 \qquad (\mathbf{9.08}, \mathbf{10.42})\ \mu g/l$

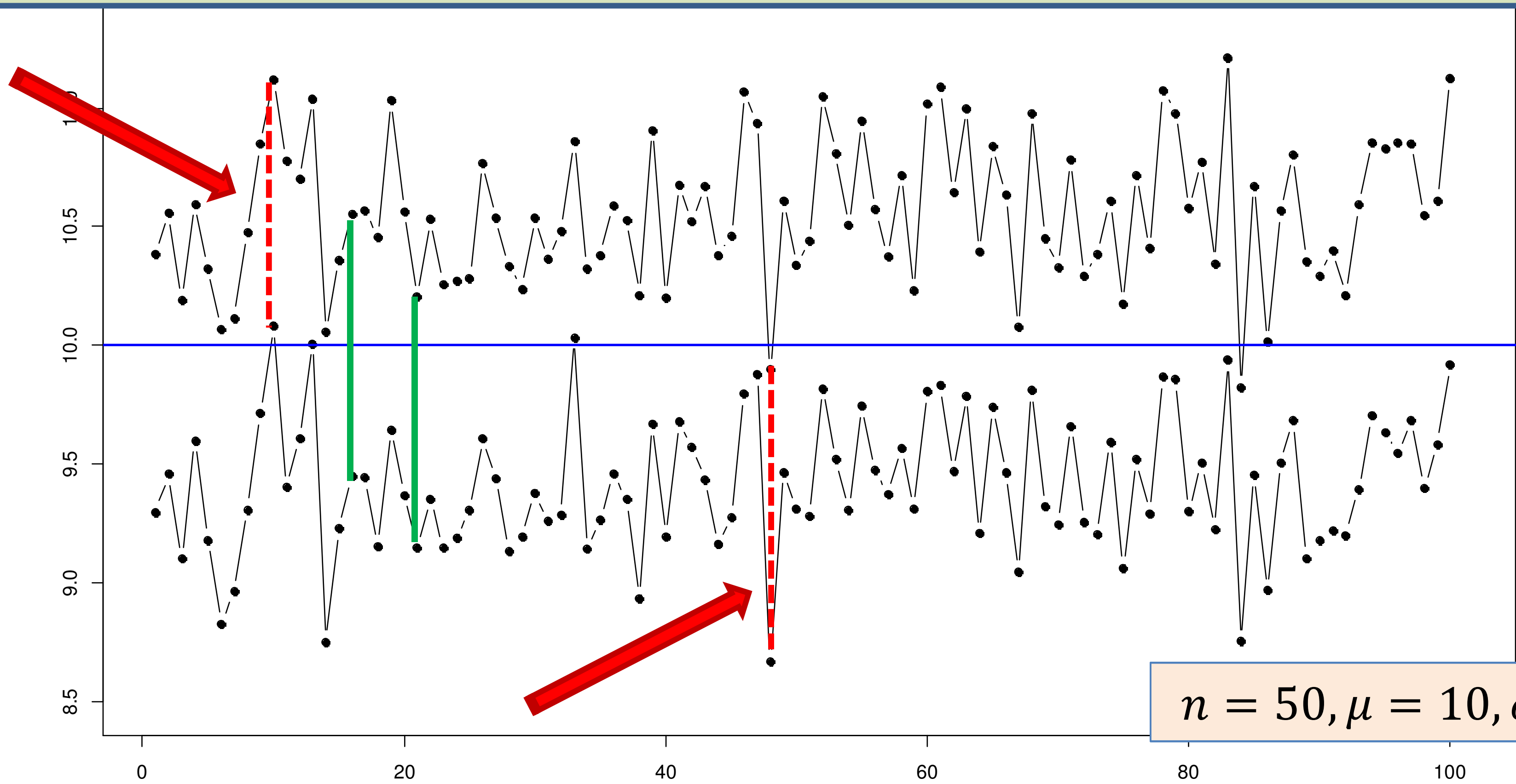
$\boldsymbol{n} = \mathbf{50} \qquad (\mathbf{9.23}, \mathbf{10.27})\ \mu g/l$

Interpretazione

$X_1, X_2, \dots, X_n$ i.i.d, $\sim N(\mu, \sigma^2)$ **con σ^2 nota**

$$\Rightarrow \left(\bar{x}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \, , \, \bar{x}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right)$$

intervallo di confidenza di livello $1-\alpha$ per μ



$n = 50, \mu = 10, \sigma = 2$