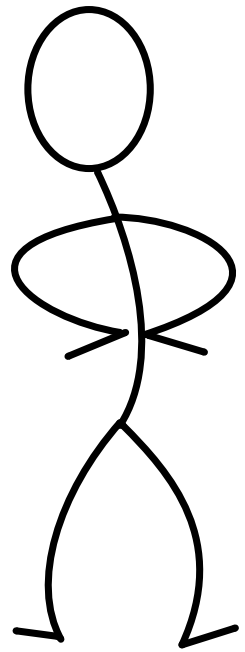
Intervallo di confidenza



Stima intervallare

Esempio: siamo intressati al peso dei maschi in una certa regione.
Vogliamo stimare il peso medio μ della popolazione.
Siano $x_1, \ldots, x_n$ i pesi di n individui scelti a caso.

La **media campionaria** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ è una **stima puntuale** di μ.

Il valore $\bar{x}$ ha poco significato: non sarà mai esattamente uguale a μ.

Il valore $\bar{x}$ è uno dei possibili della v.a. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Dove X_i è la v.a. che da il peso dell'i-esimo soggetto del campione.

Si tratta di v.a.i. equidistribuite.

Stima intervallare

Buona notizia: se le variabili X_i hanno media μ e varianza σ^2 la v.a. $\bar{X}$ ha media μ e varianza σ^2/n.

Cattiva notizia: nel caso di variabili continue, $P(\bar{X} = \bar{\mu}) = 0$.

Stima intervallare

Ma se sostituiamo il valore puntuale con un intervallo di raggio ε

$$P\Big(\bar{X} \in [\mu - \varepsilon, \mu + \varepsilon]\Big) = P\big(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon\big)$$

$$= P\big(\mu \in [\bar{X} - \varepsilon, \bar{X} + \varepsilon]\big) = 1 - \alpha.$$

$1 - \alpha$ si chiama **livello di confidenza** è la probabilità che la media campionaria sia a distanza $\leq \varepsilon$ da μ

α si chiama **livello di significatività**.

ε lo chiameremo **errore** (o raggio dell'intervallo di confidenza).

Tipicamente, $1 - \alpha \neq 0$, e cresce al cresere di ε.

Se osserviamo $\bar{x}$ allora μ è a distanza minore di ε con probabilità $1 - \alpha$.

Diremo che $\mu = \bar{x} \pm \varepsilon$ con un **livello di confidenza** $1 - \alpha$.

Stima intervallare

Consideriamo la distribuzione dei livelli di colesterolo in una data popolazione. Assumiamo che la distribuzione sia normale con media μ ignota la deviazione standard è $\sigma = 56\,\mathrm{mg/dL}$. Selezioniamo un campione di dimensione $n = 49$.

- Qual è la probabilità che μ sia a distanza inferiore a $20\,\mathrm{mg/dL}$ dalla media campionaria?

Ci interessa $P(\mu - 20 \leq \bar{X} \leq \mu + 20)$.

Con $\bar{X}$ normale con media μ e deviazione standard $\sigma/\sqrt{n} = 56/7 = 8$.

$$\begin{aligned} P(\mu - 20 \leq \bar{X} \leq \mu + 20) &= P\big(\bar{X} \leq \mu + 20\big) - P\big(\bar{X} \leq \mu - 20\big) \\ &= P\big(\bar{X} - \mu \leq 20\big) - P\big(\bar{X} - \mu \leq -20\big) \\ &= 1 - 2 \cdot P\big(\bar{X} - \mu \leq -20\big) \approx 99\% \end{aligned}$$

$$1 \;-\; 2 * \texttt{pnorm}(-20, 0, 8) \;=\; 0.9875807$$

Quindi il risultato non dipende da μ !

Stima intervallare

Consideriamo la distribuzione dei livelli di colesterolo in una data popolazione. Assumiamo che la distribuzione sia normale con media μ ignota la deviazione standard è $\sigma = 56\,\mathrm{mg/dL}$. Selezioniamo un campione di dimensione $n = 49$.

- Vogliamo un livello di confidenza almeno del 98%. Qual è il minimo errore con cui possiamo stimare μ?

Ci interessa ε tale che $P(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon) = 0.98$.

Con $\bar{X}$ normale con media μ e deviazione standard $\sigma/\sqrt{n} = 56/7 = 8$.

$$
\begin{aligned}
P(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon) &= P(-\varepsilon \leq \bar{X} - \mu \leq \varepsilon) \\
&= P\big(\bar{X} - \mu \leq \varepsilon\big) - P(\bar{X} - \mu \leq -\varepsilon) \\
\text{per simmetria} \qquad &= 1 - 2 \cdot P\big(\bar{X} - \mu \leq -\varepsilon\big) = 0.98
\end{aligned}
$$

Quindi vogliamo ε tale che $P\big(\bar{X} - \mu \leq -\varepsilon\big) = 0.01$.

Stima intervallare

La funzione **quantile** della v.a. X ha come input una probabilità p e come output un x tale che $p = P(X \leq x)$.

In R la funzione quantile di una v.a. normale si calcola con la funzione

$$\texttt{qnorm(p,}\mu\texttt{,}\sigma\texttt{)}$$

Se vogliamo ε tale che $P(\bar{X} - \mu \leq -\varepsilon) = 0.01$.

Poiché $\bar{X} - \mu$ ha media 0 e deviazione standard 8.

$$-\varepsilon \;=\; \texttt{qnorm}(0.01, 0, 8) \;=\; -18.61078$$

Stima intervallare

Consideriamo la distribuzione dei livelli di colesterolo in una data popolazione. Assumiamo che la distribuzione sia normale con media μ ignota la deviazione standard è $\sigma = 56\,\mathrm{mg/dL}$.

- Quanto grande dev'essere il campione per poter stimare μ con un errore di $\varepsilon = 20\,\mathrm{mg/dL}$ ed un livello di confidenza del 99%?

Ci interessa n tale che (N.B. $\bar{X} = \bar{X}(n)$.)

$$
\begin{aligned}
0.99 \;=\; P(\mu - \varepsilon \le \bar{X} \le \mu + \varepsilon) &= P(-\varepsilon \le \bar{X} - \mu \le \varepsilon) \\
&= P\left(-\frac{\varepsilon}{\sigma/\sqrt{n}} \;\le\; \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \;\le\; \frac{\varepsilon}{\sigma/\sqrt{n}} \right) \\
&= 1 - 2 \cdot P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \;\le\; -\frac{\varepsilon}{\sigma/\sqrt{n}} \right)
\end{aligned}
$$

La v.a. $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ è una normale standard.

Stima intervallare

Vogliamo n tale che $0.005 = P\left(Z \leq -\frac{\varepsilon}{\sigma/\sqrt{n}}\right)$.

$$-\frac{\varepsilon}{\sigma/\sqrt{n}} = \texttt{qnorm}(0.005) = -2.575829$$

$$n = \left(\frac{\sigma \cdot 2.575829}{\varepsilon}\right) = \left(\frac{56 \cdot 2.575829}{20}\right)^2 \approx 52$$

Stima intervallare con varianza nota (ripasso)

I livelli di colesterolo in una data popolazione hanno distribuzione $N(\mu, \sigma^2)$ con μ ignota e $\sigma = 49\,\mathrm{mg/dL}$. Da un campione di dimensione $n = 9$ otteniamo $\bar{x} = 220\,\mathrm{mg/dL}$.

- Vogliamo intervallo di confidenza per μ di livello $1 - \alpha = .98$.

Per $x \in [0, 1]$ sia q_x tale che $x = P(Z \leq q_x)$

q_x = `qnorm(x)`

$$
\begin{aligned}
1-\alpha &= P\left(q_{\frac{\alpha}{2}} \leq Z \leq q_{1-\frac{\alpha}{2}}\right) = P\left(q_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq q_{1-\frac{\alpha}{2}}\right) \\
&= P\left(q_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq q_{1-\frac{\alpha}{2}}\right) \\
&= P\left(\bar{X} - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - q_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}
$$

per simmetria $-q_x = q_{1-x}$

$$\mu = \bar{x} \pm q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 220 \pm \texttt{qnorm(0.99)*49/sqrt(9)} = \texttt{38.00}$$

qualcuno definisce $z_\alpha = q_{1-\alpha}$

Stima intervallare con varianza ingnota

I livelli di colesterolo in una data popolazione hanno distribuzione $N(\mu, \sigma^2)$ con μ e σ entrambe ignote. Da un campione di dimensione $n = 9$ otteniamo $\bar{x} = 220\,\mathrm{mg/dL}$ e $s = 49\,\mathrm{mg/dL}$.

- Vogliamo intervallo di confidenza per μ di livello $1 - \alpha = .98$.

Per $x \in [0, 1]$ sia q_x tale che $x = P(T \leq q_x)$ per $T \sim t(n-1)$

q_x= `qt(x)`

$$
\begin{aligned}
1-\alpha &= P\left(q_{\frac{\alpha}{2}} \leq T \leq q_{1-\frac{\alpha}{2}}\right) = P\left(q_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq q_{1-\frac{\alpha}{2}}\right) \\
&= P\left(q_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq q_{1-\frac{\alpha}{2}}\right) \\
&= P\left(\bar{X} - q_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} - q_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - q_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + q_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)
\end{aligned}
$$

per simmetria $-q_x = q_{1-x}$

$$\mu = \bar{x} \pm q_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} = 220 \pm \texttt{qt(0.99,8)*49/sqrt(9)} = 220 \pm 46.31$$

qualcuno definisce $t_\alpha = q_{1-\alpha}$

Stima intervallare (riassunto)

1. Deviazione standard σ nota: $\mu = \bar{x} \pm q_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

   dove $q_{1-\frac{\alpha}{2}}$ si riferisce alla distribuzione N(0,1).

2. Deviazione standard ignota: $\mu = \bar{x} \pm q_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

   dove $q_{1-\frac{\alpha}{2}}$ si riferisce alla distribuzione $T(n-1)$.

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