INTERVALLI di FIDUCIA (CONFIDENCE INTERVAL)

Consideriamo un campione $X_1, \dots, X_n$ (variabili aleatorie i.i.d.),
con legge dipendente da un parametro $\theta \in \Theta \subseteq \mathbb{R}$,
scriveremo $\mathbb{P}_\theta$ per indicare la legge e di conseguenza
$F_\theta(t) = \mathbb{P}_\theta(X_i \le t)$, f_θ o p_θ se le v.a. hanno densità o sono discrete

Cerchiamo, a partire da dati $x_1, \dots, x_n$, un intervallo numerico in cui
ci aspettiamo di trovare θ

Definizione: un intervallo di fiducia per θ DI LIVELLO $1-\alpha$ ($\alpha \in (0,1)$, tipicamente piccolo)
é un intervallo aleatorio della forma $I = [a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$
tale che $\mathbb{P}_\theta(\theta \in I) = \mathbb{P}(a(X_1, \dots, X_n) \le \theta \le b(X_1, \dots, X_n)) \ge 1-\alpha$

Osservazione: nella ricerca di un intervallo di fiducia abbiamo due richieste in contrapposizione:

- Vogliamo che I sia piccolo (stima precisa di θ)
- Vogliamo che α sia piccolo (probabilità bassa che I non contenga affatto θ)

INTERVALLI di FIDUCIA per GAUSSIANE

Sia $X_1, \dots, X_n$ campione di legge $N(m, \sigma^2)$ che dipende solo dai parametri (m, σ).

- σ NOTO: vogliamo trovare un intervallo di fiducia per $m \in \mathbb{R}$

sia $\bar{X}_n = \dfrac{X_1 + \dots + X_n}{n}$ uno stimatore consistente per m

é ragionevole cercare I nella forma $[\bar{X}_n - d, \bar{X}_n + d]$ con d da determinare

Imponiamo dalla definizione:

$$1-\alpha \le \mathbb{P}_m(m \in I) = \mathbb{P}_m(\bar{X}_n - d \le m \le \bar{X}_n + d)$$
$$= \mathbb{P}_m(|\bar{X} - m| \le d) = \mathbb{P}_m\left(\left|\sqrt{n}\,\frac{\bar{X}_n - m}{\sigma}\right| \le \frac{\sqrt{n}\,d}{\sigma}\right) =$$
$$= \mathbb{P}\left(|Z| \le \frac{\sqrt{n}\,d}{\sigma}\right) = 2\,\Phi\left(\frac{\sqrt{n}\,d}{\sigma}\right) - 1$$

Nota: se $X_1, \dots, X_n$ hanno legge $N(m, \sigma^2)$ allora $\bar{X}_n$ ha legge $N(m\,n, n\sigma^2)$ e $Z = \frac{\bar{X}_n - m}{\sigma}\sqrt{n}$ é $N(0,1)$

Nota: il β-quantile delle distribuzione $N(0,1)$ é il numero q_β tale che $\Phi(q_\beta)=\beta$

Abbiamo quindi $1-\alpha \leq 2\,\Phi\left(\frac{\sqrt{n}\,d}{\sigma}\right)-1$

e prendiamo l'uguaglianza $\Phi\left(\frac{\sqrt{n}\,d}{\sigma}\right)=\frac{2-\alpha}{2}=1-\frac{\alpha}{2}$

$$\frac{\sqrt{n}\,d}{\sigma}=q_{1-\frac{\alpha}{2}} \Rightarrow d=\frac{\sigma}{\sqrt{n}}\cdot q_{1-\frac{\alpha}{2}}$$

Dunque

$$I=\left[\overline{X}_n-\frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}},\ \overline{X}_n+\frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}\right]$$

é un intervallo di fiducia per m a livello $1-\alpha$

Esempio: descriviamo l'altezza di un italiano scelto a caso con una variabile aleatoria $N(m,\sigma^2)$ di cui supponiamo di conoscere $\sigma=5$ cm. Date misurazioni $x_1,\dots,x_n$ con $\overline{x}_n=170$ cm a livello 95% ($\alpha=0{,}05$).

Ho un intervallo di fiducia per m del tipo:

$$\left[170-\frac{5}{\sqrt{n}}q_{0.975},\ 170+\frac{5}{\sqrt{n}}q_{0.975}\right]$$

Osservazione: d é anche detta "precisione della stima":

(*)
- all'aumentare di n, d decresce
- all'aumentare di σ, d cresce
- al diminuire di α, $q_{1-\frac{\alpha}{2}}$ aumenta e così d

- σ NON NOTO: stimiamo σ con $S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$ che é uno stimatore CORRETTO e CONSISTENTE. Ci serve:

Fatto: se X e C_n sono variabili aleatorie indipendenti con densità rispettivamente $N(0,1)$ e $\chi^2(n)$

allora $T_n=\sqrt{n}\,\frac{X}{\sqrt{C_n}}$ ha densità $f_{T_n}(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n}{2}-\frac{1}{2}}$

che chiamiamo densità di Student con n gradi di libertà

Nota: per $n\to\infty$ la variabile di Student con n gradi di libertà converge in distribuzione alla Gaussiana $N(0,1)$

Fatto: se $X_1,\dots,X_n$ sono un campione $N(m,\sigma^2)$

1) $\overline{X}_n$ e S_n^2 sono INDIPENDENTI
2) La densità $N\left(m,\frac{\sigma^2}{n}\right)$ e $\frac{n-1}{\sigma^2}S_n^2$ ha densità $\chi^2(n-1)$
3) $T_n=\sqrt{n}\,\frac{\overline{X}_n-m}{S_n}$ é di Student con $n-1$ gradi di libertà

Cerchiamo l'intervallo nella forma $[\bar{X}_n - d, \bar{X}_n + d]$

$$1-\alpha \leq \mathbb{P}_{m,\sigma^2}(m \in I) = \mathbb{P}_{m,\sigma^2}(|\bar{X}_n - m| \leq d) =$$
$$= \mathbb{P}_{m,\sigma^2}\left(\left|\sqrt{n}\frac{\bar{X}_n - m}{S_n}\right| \leq \frac{\sqrt{n}\,d}{S_n}\right)$$

Indichiamo con $\tau_{\beta,n}$ il β-quantile della variabile di Student T_n, ovvero $F_{T_n}(\tau_{\beta,n-1}) = \beta$.

Siccome la densità f_{T_n} è simmetrica vale $F_{T_n}(t) + F_{T_n}(-t) = 1$

$$= F_{T_n}\left(\frac{\sqrt{n}\,d}{S_n}\right) - F_{T_n}\left(-\frac{\sqrt{n}\,d}{S_n}\right) =$$
$$= 2F_{T_n}\left(\frac{\sqrt{n}\,d}{S_n}\right) - 1$$
$$\Downarrow$$
$$F_{T_n}\left(\frac{\sqrt{n}\,d}{S_n}\right) = 1 - \frac{\alpha}{2} \qquad \frac{\sqrt{n}\,d}{S_n} = \tau_{1-\frac{\alpha}{2},n-1}$$

Nota: quando n è generale $\tau_{\beta,n}$ è indistinguibile da q_β

$$d = \frac{S_n}{\sqrt{n}}\tau_{1-\frac{\alpha}{2},n-1}$$

Dunque $I = \left[\bar{X}_n - \frac{S_n}{\sqrt{n}}\tau_{1-\frac{\alpha}{2},n-1}, \bar{X}_n + \frac{S_n}{\sqrt{n}}\tau_{1-\frac{\alpha}{2},n-1}\right]$

Esempio: analogamente al precedente se $\bar{X}_n = 170$ cm e $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2 = (5\text{ cm})^2$

allora
$\left[170 \pm \frac{5}{\sqrt{n}}\tau_{0.975,n-1}\right]$ è di fiducia con $\alpha = 0{,}05$

Valgono le osservazioni di (*)

Intervallo di Fiducia per variabili aleatorie Bernoulli (p)

Consideriamo un campione $X_1, \dots, X_n$ con legge Bernoulli (p).

Sappiamo che $\overline{X}_n$ é uno stimatore corretto e costante di p, quindi cerchiamo

$$I = [\overline{X}_n \pm d]$$

$$1-\alpha = \mathbb{P}_p(p \in I) = \mathbb{P}_p(|\overline{X}_n - p| \leq d)$$

Sappiamo per TCL che $\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$ é approssimativamente $N(0,1)$

ma non possiamo dividere per $\sqrt{p(1-p)}$ poiché p non é noto. Si puó peró mostrare che:

$$\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} = \sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to\infty]{} N(0,1)$$

$$\mathbb{P}_p\left(\left|\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}\right| \leq \sqrt{n}\,\frac{d}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}\right)$$

$$\approx$$

$$\mathbb{P}_p\left(|Z| = \frac{\sqrt{n}\,d}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}\right)$$

Abbiamo quindi la condizione approssimata:

$$1-\alpha \approx 2\,\Phi\left(\frac{\sqrt{n}\,d}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}\right) - 1 \qquad d \approx \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}}$$

dunque

$$\left[\overline{X}_n \pm \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}}\right]$$

Esempio: intervistiamo n persone che hanno votato $X_1, \dots, X_n$ $(=0,1)$ con $\overline{X}_n = 0{,}4$

allora $\left[0.4 \pm \frac{\sqrt{0.4 \cdot 0.6}}{\sqrt{n}}\,q_{0.975}\right]$ é di fiducia per p e rappresenta la percentuale di SÍ su tutta la popolazione a livello 95%