Intervalli di fiducia

Intervalli di fiducia per la media di un campione gaussiano

Varianza nota

$X_1, \dots, X_n$ campione gaussiano con σ^2 nota e m da determinare. Visto che $E[\overline{X}] = m$, cioè $\overline{X}$ è una stima per la media, prendiamo l'intervallo

$$I = \left[\overline{X} - d, \overline{X} + d\right]$$

e un d, quanto più piccolo possibile, tale che:

$$P_m(m \in I) = \underbrace{P_m\left(\left|\overline{X} - m\right| \le d\right)}_{\text{usiamo questa}} = \underbrace{P_m\left(\overline{X} \in [m-d, m+d]\right)}_{\text{ha più senso rispetto a } P} \ge 1 - \alpha,$$

cioè cerchiamo d tale che:

$$P_m\left(\left|\overline{X} - m\right| \le d\right) = 1 - \alpha.$$

Visto che $X_i \sim N(m, \sigma^2)$,

$$\overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right) \qquad \Longrightarrow \qquad \frac{\sqrt{n}}{\sigma}\left(\overline{X} - m\right) \sim N(0,1),$$

quindi:

$$P_m\left(\left|\overline{X} - m\right| \le d\right) = P\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X} - m\right| \le d\frac{\sqrt{n}}{\sigma}\right) = 1 - \alpha$$

$$\Phi\left(\frac{\sqrt{n}}{\sigma}d\right) - \Phi\left(-\frac{\sqrt{n}}{\sigma}d\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}d\right) - \left(1 - \Phi\left(\frac{\sqrt{n}}{\sigma}d\right)\right) = 2\Phi\left(\frac{\sqrt{n}}{\sigma}d\right) - 1 = 1 - \alpha$$

$$\Phi\left(\frac{\sqrt{n}}{\sigma}d\right) = 1 - \frac{\alpha}{2} \qquad \frac{\sqrt{n}}{\sigma}d = q_{1-\frac{\alpha}{2}}$$

Quindi *precisione della stima* e *precisione relativa* sono:

$$d = \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}} \qquad\qquad \frac{\frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}}{\overline{X}}$$

e l'intervallo di fiducia al livello $1 - \alpha$ è:

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}.$$

Più sono i dati a disposizione, migliore è la precisione.

Varianza sconosciuta

Si sostitiusce σ^2 con la varianza campionaria $S^2 = \frac{\sum(X_i^2 - \overline{X})^2}{n-1}$. Allora

$$\sqrt{n}\frac{\overline{X} - m}{S} \sim T(n-1),$$

quindi l'intervallo è:

$$\overline{X} \pm \frac{S}{\sqrt{n}}\tau_{1-\frac{\alpha}{2}, n-1}.$$

Per n grande (≥ 60) si approssima il quantile a quello della gaussiana standard.

Unilatero sinistro

$I = \left(-\infty, \overline{X} + d\right]$,

$$P_m(m \in I) = P_m\left(\overline{X} - m \ge -d\right) = 1 - \alpha,$$

$$-d\frac{\sqrt{n}}{\sigma} = q_\alpha \qquad d = \frac{\sigma}{\sqrt{n}}q_{1-\alpha}.$$

Per l'intervallo destro il quantile diventa q_α. Gli stessi risultati si adattano al caso con varianza sconosciuta.

Intervalli di fiducia per la varianza di un campione gaussiano

$X_1, \dots, X_n$ campione gaussiano. Usiamo:

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

$$P_{\sigma^2}\left(d_1 < (n-1)\frac{S^2}{\sigma^2} < d_2\right) = P_{\sigma^2}\left((n-1)\frac{S^2}{d_2} < \sigma^2 < (n-1)\frac{S^2}{d_1}\right),$$

quindi cerchiamo intervalli del tipo:

$$I = \left[\frac{\sum(X_i - \overline{X})^2}{d_2}, \frac{\sum(X_i - \overline{X})^2}{d_1}\right].$$

Nella pratica servono solo gli intervalli unilateri.

Sinistro

$$I = \left(0, \frac{\sum(X_i - \overline{X})^2}{a}\right]$$

$$\{\sigma^2 \in I\} = \left\{\sigma^2 \le \frac{\sum(X_i - \overline{X})^2}{a}\right\} = \left\{a \le \frac{\sum(X_i - \overline{X})^2}{\sigma^2}\right\}$$

Poniamo:

$$P_{\sigma^2}\left(a \le \underbrace{\frac{\sum(X_i - \overline{X})^2}{\sigma^2}}_{\chi^2(n-1)}\right) = 1 - \alpha,$$

ovvero:

$$1 - G_{n-1}(a) = 1 - \alpha \quad \Longrightarrow \quad a = \chi^2_{\alpha,n-1},$$

che non dipende da σ^2. Quindi:

$$I = \left(0, \frac{\sum(X_i - \overline{X})^2}{\chi^2_{\alpha,n-1}}\right]$$

Destro

Analogamente si ottiene:

$$I = \left[\frac{\sum(X_i - \overline{X})^2}{\chi^2_{1-\alpha,n-1}}, +\infty\right)$$

Intervalli di fiducia per la media di un campione di Bernoulli

$X_1, \ldots, X_n \sim B(1,p)$. Per il TLC, per n grande:

$$\sqrt{n}\frac{\overline{X}-p}{\sqrt{p(1-p)}} \approx N(0,1).$$

p è anche la media, quindi possiamo usare $\overline{X}$ (che concide con la stima di massima verosimiglianza $\hat{p}$) come stima per p. Da questo si trova l'intervallo:

$$\hat{p} \pm \frac{\sqrt{p(1-p)}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}.$$

Non conosciamo p, quindi approssimiamo ulteriormente con $p \simeq \hat{p}$:

$$\hat{p} \pm \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}.$$

Studio della precisione della stima

Per esempio, vogliamo determinare la dimensione del campione per avere una determinata precisione a un dato livello di fiducia. È facile nel caso del campione gaussiano con varianza nota, ma negli altri la precisione della stima dipende dalle misurazioni (e.g. varianza campionaria), quindi:

- calcoliamo $S^2(\omega)$, $\hat{p}$, ... su un campione piccolo, oppure
- nel caso $B(1,p)$, consideriamo il caso pessimo – $\max p(1-p) = \frac{1}{4}$.

Verifica delle ipotesi

Test sulla media di un campione gaussiano con varianza nota (Z-test)

Campione $X_1, \dots, X_n \sim N(m, \sigma^2)$, con σ^2 nota. Test bilatero con ipotesi $\mathcal{H}_0 : m = m_0$, $\mathcal{H}_1 : m \neq m_0$.

Poiché $\overline{X}$ è la stima di m, rifiutiamo l'ipotesi se $\overline{X}$ si scosta troppo da m_0, ovvero

$$C = \{\left|\overline{X} - m_0\right| > d\}$$

con d da determinare in base al livello α:

$$P_{m_0}\left(\left|\overline{X} - m_0\right| > d\right) \leq \alpha$$

vogliamo aumentare la potenza, quindi C più grande possibile, dunque:

$$P_{m_0}\left(\left|\overline{X} - m_0\right| > d\right) = \alpha$$

Nel calcolo di C supponiamo che l'ipotesi sia vera, quindi:

$$\frac{\sqrt{n}}{\sigma}(\overline{X} - m_0) = Z \sim N(0, 1),$$

allora:

$$P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X} - m_0\right| > \frac{\sqrt{n}}{\sigma}d\right) = P\left(|Z| > \frac{\sqrt{n}}{\sigma}d\right) = \alpha$$

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

Quindi la regione critica è:

$$C = \left\{\left|\overline{X} - m_0\right| > \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}\right\}$$

Legame tra intervalli di fiducia e test l'intervallo di fiducia per la media al livello $1 - \alpha$ è:

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}},$$

quindi l'ipotesi $\mathcal{H}_0$ è accettata al livello α se m_0 appartiene all'intervallo di fiducia al livello $1 - \alpha$. Per questo motivo usiamo $1 - \alpha$ anziché β.

Utilizzo dei dati e p-value

Date le misurazioni $x_1, \dots, x_n$ e $\overline{x} = \frac{x_1 + \dots + x_n}{n}$, rifiutiamo l'ipotesi se $\omega \in C$:

$$|\overline{x} - m_0| > \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}},$$

la accettiamo se $\leq$. Il p-value quindi sarà $\overline{\alpha}$ tale che:

$$|\overline{x} - m_0| = \frac{\sigma}{\sqrt{n}}q_{1-\frac{\overline{\alpha}}{2}},$$

quindi:

$$\begin{aligned}
q_{1-\frac{\overline{\alpha}}{2}} &= \frac{\sqrt{n}}{\sigma}|\overline{x} - m_0| \\
1 - \frac{\overline{\alpha}}{2} &= \Phi\left(\frac{\sqrt{n}}{\sigma}|\overline{x} - m_0|\right) \\
\overline{\alpha} &= 2\left(1 - \Phi\left(\frac{\sqrt{n}}{\sigma}|\overline{x} - m_0|\right)\right).
\end{aligned}$$

Definizione alternativa del p-value

Consideriamo:

$$P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X}-m_0\right| \geq \frac{\sqrt{n}}{\sigma}\left|\overline{x}-m_0\right|\right)$$

- caso $\alpha < \overline{\alpha}$: accettiamo l'ipotesi, quindi $|\overline{x}-m_0| \leq \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}$. Allora:

$$P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X}-m_0\right| \geq \frac{\sqrt{n}}{\sigma}\left|\overline{x}-m_0\right|\right) \leq P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X}-m_0\right| \geq q_{1-\frac{\alpha}{2}}\right) = P_{m_0}(C) = \alpha;$$

- caso $\alpha > \overline{\alpha}$: rifiutiamo l'ipotesi, quindi $|\overline{x}-m_0| > \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}$. Allora:

$$P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X}-m_0\right| \geq \frac{\sqrt{n}}{\sigma}\left|\overline{x}-m_0\right|\right) \geq P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X}-m_0\right| \geq q_{1-\frac{\alpha}{2}}\right) = P_{m_0}(C) = \alpha.$$

Cioè:

$$\overline{\alpha} = P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left|\overline{X}-m_0\right| \geq \frac{\sqrt{n}}{\sigma}\left|\overline{x}-m_0\right|\right) = P\left(|Z| \geq \frac{\sqrt{n}}{\sigma}\left|\overline{x}-m_0\right|\right) = \underbrace{2\left(1-\Phi\left(\frac{\sqrt{n}}{\sigma}\left|\overline{x}-m_0\right|\right)\right)}_{\text{già visto}}$$

è la probabilità che un'altra indagine statistica ω' restituisca una media dei dati $\overline{x}'$ distante da m_0 almeno quanto $\overline{x}$ se l'ipotesi è vera.

Se il p-value è basso (< 0.1), significa che se l'ipotesi fosse stata vera difficilmente avremmo ottenuto il risultato osservato, per cui l'ipotesi è da scartare. Un p-value alto indica che l'ipotesi è plausibile, cioè che non possiamo scartarla, ma non è la probabilità che l'ipotesi sia corretta – per calcolarlo supponiamo che sia vera, quindi non ci dà informazioni nel caso in cui sia falsa.

In generale, se t è il valore osservato per una distribuzione T,

- $\overline{\alpha} = P(T \geq t \mid \mathcal{H}_0)$ per un test unilatero destro;
- $\overline{\alpha} = 2\min\left\{P(T \geq t \mid \mathcal{H}_0), P(T \leq -t \mid \mathcal{H}_0)\right\}$ per un test bilatero;
- $\overline{\alpha} = P(|T| \geq |t| \mid \mathcal{H}_0)$ per un test bilatero con T simmetrica rispetto all'origine (togliendo m_0 sopra viene simmetrico).

Curva operativa

$$\begin{aligned}
\beta(m) = P_m(A) &= P_m\left(\overbrace{\left|\overline{X}-m_0\right| \leq \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}}^{A}\right) \\
&= P_m\left(m_0 - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \leq \overline{X} \leq m_0 + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right) \qquad \overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right) \\
&= P_m\left(\frac{m_0-m}{\sigma}\sqrt{n} - q_{1-\frac{\alpha}{2}} \leq \underbrace{\frac{\sqrt{n}}{\sigma}(\overline{X}-m)}_{N(0,1)} \leq \frac{m_0-m}{\sigma}\sqrt{n} + q_{1-\frac{\alpha}{2}}\right) \\
&= \Phi\left(\frac{m_0-m}{\sigma}\sqrt{n} + q_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{m_0-m}{\sigma}\sqrt{n} - q_{1-\frac{\alpha}{2}}\right)
\end{aligned}$$

Z-test unilatero

Campione $X_1, \ldots, X_n \sim N(m, \sigma^2)$, con σ^2 nota. Test unilatero sinistro con ipotesi $\mathcal{H}_0 : m \leq m_0$, $\mathcal{H}_1 : m > m_0$.

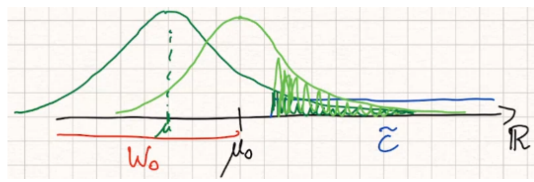
La regione critica è della forma:

$$C = \{\overline{X} - m_0 > d\},$$

con d da determinare in base al livello α:

$$\forall m \leq m_0 \, . \, P_m(\overline{X} - m_0 > d) \leq \alpha$$

$P_m(\overline{X} - m_0 > d)$ cresce all'aumentare di m:



quindi basta soddisfare la disuguaglianza per m_0. Quindi:

$$\alpha = P_{m_0}(\overline{X} - m_0 > d) = P_{m_0}\left(\frac{\sqrt{n}}{\sigma}\left(\overline{X} - m_0\right) > \frac{\sqrt{n}}{\sigma}d\right) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}d\right)$$

$$d = \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$$

e

$$C = \left\{\frac{\sqrt{n}}{\sigma}(\overline{X} - m_0) > q_{1-\alpha}\right\}$$

Utilizzo dei dati e p-value

Rifiutiamo l'ipotesi se e solo se:

$$\overline{x} - m_0 > \frac{\sigma}{\sqrt{n}} q_{1-\alpha},$$

quindi $\overline{x} - m_0 = \frac{\sigma}{\sqrt{n}} q_{1-\overline{\alpha}}$, da cui si ricava:

$$\overline{\alpha} = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(\overline{x} - m_0)\right).$$

Si trova lo stesso risultato con $\overline{\alpha} = P_{m_0}\left(\frac{\sqrt{n}}{\sigma}(\overline{X} - m_0) > \frac{\sqrt{n}}{\sigma}(\overline{x} - m_0)\right)$.

Curva operativa

$$\beta(m) = P_m\Bigg(\underbrace{\frac{\sqrt{n}}{\sigma}\left(\overline{X} - m_0\right) \leq q_{1-\alpha}}_{A}\Bigg) = P_m\Bigg(\underbrace{\frac{\sqrt{n}}{\sigma}(\overline{X} - m)}_{N(0,1)} \leq \frac{m_0 - m}{\sigma}\sqrt{n} + q_{1-\alpha}\Bigg) = \Phi\left(\frac{m_0 - m}{\sigma}\sqrt{n} + q_{1-\alpha}\right)$$

Destro

$\mathcal{H}_0 : m \geq m_0$, $\mathcal{H}_1 : m < m_0$.

$$C = \left\{\frac{\sqrt{n}}{\sigma}(\overline{X} - m_0) < q_\alpha\right\} \quad \beta(m) = 1 - \Phi\left(\frac{m_0 - m}{\sigma}\sqrt{n} + q_\alpha\right) \quad \overline{\alpha} = \Phi\left(\frac{\sqrt{n}}{\sigma}(\overline{x} - m_0)\right)$$

e l'ipotesi è rifiutata se

$$\overline{x} - m_0 < \frac{\sigma}{\sqrt{n}} q_\alpha.$$

Test sulla media di un campione gaussiano con varianza sconosciuta (T-test)

σ^2 qualsiasi, $\mathcal{H}_0 : m = m_0$, $\mathcal{H}_1 : m \neq m_0$. Come lo Z-test, ma usando $\frac{\sqrt{n}}{S}(\overline{X} - m) \sim T(n-1)$.

- regione critica al livello α:
$$C = \left\{ \left|\overline{X} - m_0\right| > \frac{S}{\sqrt{n}} \tau_{1-\frac{\alpha}{2}, n-1} \right\};$$

- p-value (importante sapere la formula):
$$P_{m_0}\left(\underbrace{\frac{\sqrt{n}}{S}\left|\overline{X} - m_0\right|}_{T \text{ del modello}} > \underbrace{\frac{\sqrt{n}}{s}\left|\overline{x} - m_0\right|}_{t \text{ osservata}} \right) = 2\left(1 - F_{n-1}\left(\frac{\sqrt{n}}{s}\left|\overline{x} - m_0\right|\right)\right)$$
con F_{n-1} c.d.f. di $T(n-1)$, che per $n > 70$ è approssimativamente Φ;

- la curva operativa non ha senso: se $m \neq m_0$ $\frac{\sqrt{n}}{S}(\overline{X} - m_0) \not\sim T(n-1)$, e se ci riconduciamo a $\frac{\sqrt{n}}{S}(\overline{X} - m)$ nell'espressione della curva compare S, che non conosciamo a priori.

Test approssimato su un campione di Bernoulli

$X_1, \ldots, X_n$ Bernoulli di parametro p. $\mathcal{H}_0 : p = p_0$, $\mathcal{H}_1 : p \neq p_0$.

Se $p = p_0$,
$$\frac{\sqrt{n}(\overline{X} - p_0)}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

- regione critica al livello α:
$$C = \left\{ \frac{\sqrt{n}\left|\overline{X} - p_0\right|}{\sqrt{p_0(1-p_0)}} > q_{1-\frac{\alpha}{2}} \right\};$$

- p-value
$$P_{p_0}\left(\frac{\sqrt{n}\left|\overline{X} - p_0\right|}{\sqrt{p_0(1-p_0)}} > \frac{\sqrt{n}\left|\hat{p} - p_0\right|}{\sqrt{p_0(1-p_0)}} \right) \simeq 2\left(1 - \Phi\left(\frac{\sqrt{n}\left|\hat{p} - p_0\right|}{\sqrt{p - (1-p_0)}}\right)\right)$$
dove $\hat{p} = \overline{x} = \frac{x_1 + \cdots + x_n}{n}$;

- la curva operativa si può calcolare ma è difficile.

Test unilatero sulla varianza di un campione gaussiano

$X_1, \dots, X_n \sim N(m, \sigma^2)$ indipendenti, $\mathcal{H}_0 : \sigma^2 \le \sigma_0^2$, $\mathcal{H}_0 : \sigma^2 > \sigma_0^2$.
Non cambia se si conosce la media.

Sappiamo che

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

quindi prendiamo come regione critica:

$$C = \left\{ \frac{\sum (X_i - \overline{X})^2}{\sigma_0^2} > d \right\} = \left\{ S^2 > \frac{d\sigma_0^2}{n-1} \right\}$$

Troviamo d al livello α:

$$\forall \sigma \le \sigma_0 \quad . \quad P_\sigma \Bigg(\underbrace{\frac{\sum (X_i - \overline{X})^2}{\sigma_0^2}}_{\text{non } \chi^2(n-1)} > d \Bigg) \le \alpha.$$

È sufficiente verificare per $\sigma = \sigma_0$, infatti $\forall \sigma \le \sigma_0 \; . \; P_\sigma(C) \le P_{\sigma_0}(C)$:

$$P_\sigma \left(\frac{\sum (X_i - \overline{X})^2}{\sigma_0^2} > d \right) = P_\sigma \left(\frac{\sum (X_i - \overline{X})^2}{\sigma^2} > \frac{d\sigma_0^2}{\sigma^2} \right) = P \left(C_{n-1} > \frac{d\sigma_0^2}{\sigma^2} \right) \le P \left(C_{n-1} > d \right) = P_{\sigma_0}(C).$$

Quindi:

$$P_{\sigma_0} \Bigg(\underbrace{\frac{\sum (X_i - \overline{X})^2}{\sigma_0^2}}_{\chi^2(n-1)} > d \Bigg) = \alpha \qquad d = \chi^2_{1-\alpha, n-1}.$$

- regione critica:

$$C = \left\{ \frac{\sum (X_i - \overline{X})^2}{\sigma_0^2} > \chi^2_{1-\alpha, n-1} \right\}$$

- p-value:

$$P_{\sigma_0} \left(\frac{\sum (X_i - \overline{X})^2}{\sigma_0^2} > \frac{\sum (x_i - \overline{x})^2}{\sigma_0^2} \right) = 1 - F_{n-1} \left(\frac{\sum (x_i - \overline{x})^2}{\sigma_0^2} \right) = 1 - F_{n-1} \left(\frac{(n-1)\,\mathrm{var}(x)}{\sigma_0^2} \right)$$

- curva operativa:

$$\beta(\sigma) = F_{n-1} \left(\frac{\sigma_0^2}{\sigma^2} \chi^2_{1-\alpha, n-1} \right)$$