Principi di Inferenza

Inferenza di verosimiglianza

Laura D'Angelo
Università Milano Bicocca

1 Verosimiglianza e stimatore di massima verosimiglianza

Un metodo generale per costruire stimatori

Finora gli stimatori $(\bar{X},\, S^2)$ sono nati da un'*intuizione*: "sostituiamo il valore atteso con la media campionaria". Per modelli più complessi questa strada non è sempre chiara.

Idea: assumiamo un modello parametrico, cioè che i dati abbiano densità (o funzione di probabilità) $f(x;\theta)$ con θ parametro ignoto. Scegliamo come stima il valore di θ che rende i dati osservati i più plausibili.

Esempio: un nuovo trattamento viene somministrato a $n = 10$ pazienti; 7 rispondono positivamente al trattamento. Sia p la probabilità di risposta positiva. Un modello ragionevole è:

$$X_i \sim \text{Bernoulli}(p), \quad i = 1, \dots, 10 \text{ indipendenti}; \qquad \sum_{i=1}^{10} x_i = 7$$

$$f(x_1, \dots, x_n; p) = \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{10} x_i}(1-p)^{n-\sum_{i=1}^{10} x_i}$$

- Se $p = 0.1$, osservare 7 risposte su 10 è molto poco probabile.
- Se $p = 0.7$, è molto più probabile.
- Qual è il valore di p **più plausibile** alla luce dei dati?

La funzione di verosimiglianza

Definizione

Dati osservati $x_1, \ldots, x_n$, assumiamo sia un campione da $(X_1, \ldots, X_n)$ i.i.d. con densità $f(x\,;\theta)$. Definiamo

$$\text{Funzione di verosimiglianza:} \quad L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

$$\text{Funzione di log-verosimiglianza:} \quad \ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

- La verosimiglianza è la densità congiunta dei dati, **vista come funzione di θ con i dati $x_1, \ldots, x_n$ fissati ai valori osservati**.
- **Non** è una distribuzione di probabilità per θ (non integra a 1): conta solo il confronto tra valori di L.
- Si usa $\ell(\theta)$ perché i prodotti diventano somme, il massimo è lo stesso (il logaritmo è crescente), e spesso i conti risultano semplificati.

Esempio Bernoulli (7 successi su 10): $L(p) = p^7(1-p)^3, \quad \ell(p) = 7\log p + 3\log(1-p).$

Stima di massima verosimiglianza

Definizione

La stima di massima verosimiglianza $\hat{\theta}$ è il valore che massimizza $L(\theta)$ (o, equivalentemente, $\ell(\theta)$):

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta)$$

Come si trova? Procedura:

1. Scrivere la funzione di verosimiglianza $L(\theta)$ e log-verosimiglianza $\ell(\theta)$.
2. Calcolare la funzione score $\ell_*(\theta) = \ell'(\theta)$, ovvero la derivata prima (o vettore di derivate parziali, se θ ha dimensione >1)
3. Risolvere l'equazione di verosimiglianza $\ell_*(\theta) = 0$. La soluzione dell'equazione di verosimiglianza (se esiste ed è unica) è $\hat{\theta}$, valore candidato a stima di massima verosigmiglianza.
4. Verificare che $\hat{\theta}$ sia un massimo (es. $\ell''(\hat{\theta}) < 0$).

Esempio: modello Bernoulli

- $n = 10$ pazienti; $\sum_{i=1}^{10} x_i = 7$ risposte positive.
- Modello: $X_i \sim \text{Bernoulli}(p)$, $i = 1, \ldots, 10$ indipendenti, con p probabilità di risposta positiva.
- La distribuzione di probabilità di $(X_1, \ldots, X_n)$ è

$$f(x_1, \ldots, x_n; p) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

- Funzione di verosimiglianza:

$$L(p; x_1, \ldots, x_n) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

- Funzione di log-verosimiglianza:

$$\ell(p) = \left(\sum_{i=1}^n x_i \right) \log p + \left(n - \sum_{i=1}^n x_i \right) \log(1-p)$$

- Funzione score e equazione di verosimiglianza:

$$\ell_*(p) = \frac{d}{dp} \ell(p) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0 \implies \hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

- Stima di massima verosimiglianza: è la media campionaria $\hat{p} = \bar{x}$. Con i nostri dati, $\hat{p} = 7/10 = 0.7$.
- Stimatore di massima verosimiglianza: è la variabile aleatoria $\hat{P} = \bar{X}$

Esempio: modello normale

Generico campione $(X_1, \ldots, X_n)$ da $X_i \sim \mathrm{N}(\mu, \sigma^2)$ i.i.d., entrambi i parametri ignoti:

- Densità congiunta:

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

- Verosimiglianza

$$L(\mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

- Log-verosimiglianza

$$\ell(\mu, \sigma^2) \propto -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

- Funzione score (derivate parziali)

$$\ell_*(\mu, \sigma^2) = \left[\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma^2} \right]^\top$$

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu), \qquad \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Esempio: modello normale

- Equazioni di verosimiglianza: poniamo le derivate parziali uguali a zero e risolviamo le equazioni

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Le stime di massima verosimiglianza sono $\hat{\mu} = \bar{x}$ (media campionaria) e $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ (varianza campionaria).
- Gli stimatori di massima verosimiglianza sono $\hat{M} = \bar{X}$ e $S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ (variabili aleatorie).
- Lo stimatore di massima verosimiglianza di σ^2 è distorto: $\mathbb{E}[S^2] = \frac{n-1}{n} \sigma^2$.

  Gli stimatori di massima verosimiglianza possono essere distorti, ma sono **asintoticamente non distorti**.
- **Invarianza**: se $\hat{\theta}$ è la SMV di θ, allora $g(\hat{\theta})$ è la SMV di $g(\theta)$. Ad esempio, la SMV di σ è $\sqrt{S^2}$.

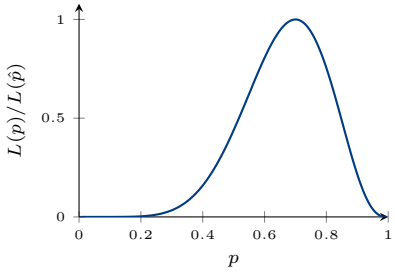
Verosimiglianza relativa

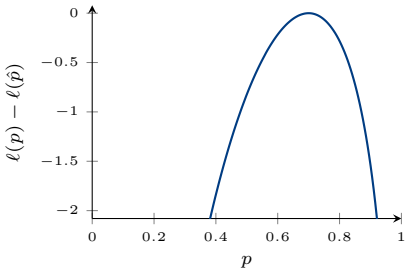
I valori della verosimiglianza ci interessano solo in termini relativi: quanto varia $L(\theta)$ se considero due valori θ' e θ''? Inoltre, non ci interessa *quanto vale* $L(\theta)$ nel massimo, ma solo per quale valore di θ è massimizzata.

Definizione

La verosimiglianza relativa (o normalizzata) è $L(\theta)/L(\hat{\theta})$. La verosimiglianza normalizzata nel massimo vale 1.

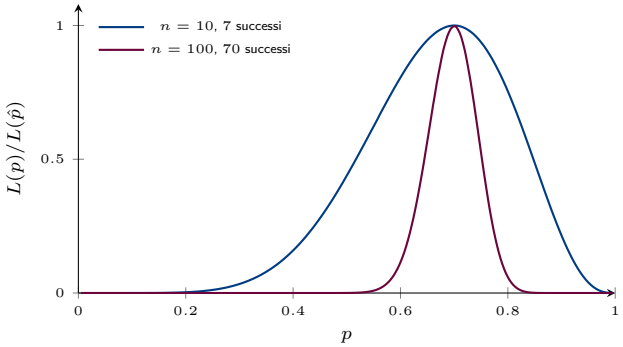
La log-verosimiglianza relativa è $\ell(\theta) - \ell(\hat{\theta})$. La log-verosimiglianza normalizzata nel massimo vale 0.





Verosimiglianza e numerosità campionaria

Esempio: studiamo due campioni Bernoulli, entrambi con proporzione di successi pari a 0.7, ma numerosità campionaria diversa: $n = 10$ e $n = 100$.



- Il massimo è in $p = 0.7$ in entrambi i casi, ma con più dati la funzione è molto più concentrata: i dati discriminano meglio tra valori di p.
- La **curvatura** intorno al massimo misura quanta informazione contengono i dati.

2 Precisione della stima e distribuzione asintotica

Quanto è precisa la stima? La curvatura

Nella slide precedente: più dati $\Rightarrow$ log-verosimiglianza più "appuntita" $\Rightarrow$ stima più precisa.

La curvatura si misura con la derivata seconda: $-\ell''(\hat{\theta})$ è grande quando la curva è stretta.

Informazione osservata

$$J(\hat{\theta}) = -\ell''(\hat{\theta})$$

Più è grande, più i dati sono informativi su θ e più la stima è precisa.

Esempio Bernoulli

7 successi su 10:

$$\ell''(p) = -\frac{7}{p^2} - \frac{3}{(1-p)^2} \quad \Longrightarrow \quad J(\hat{p}) = \frac{7}{0.7^2} + \frac{3}{0.3^2} \approx 14.3 + 33.3 = 47.6$$

70 successi su 100:

$$\ell''(p) = -\frac{70}{p^2} - \frac{30}{(1-p)^2} \quad \Longrightarrow \quad J(\hat{p}) = \frac{70}{0.7^2} + \frac{30}{0.3^2} \approx 143 + 333 = 476$$

Vedremo che la varianza dello stimatore di massima verosimiglianza $\hat{\Theta}$ è circa $1/J(\hat{\theta})$: curvatura grande $\Rightarrow$ errore standard piccolo.

La versione "teorica" dell'informazione osservata è il suo valore atteso, che dipende solo dal modello.

Informazione di Fisher (per una osservazione)

$$I(\theta) = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

Per n osservazioni i.i.d. l'informazione è $n\,I(\theta)$: cresce con la numerosità campionaria.

Esempio Bernoulli: $\dfrac{\partial^2}{\partial p^2}\log f(x;p) = -\dfrac{x}{p^2} - \dfrac{1-x}{(1-p)^2}$. Poiché $\mathbb{E}[X] = p$,

$$I(p) = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

Controllo con i dati: $n\,I(\hat{p}) = \dfrac{10}{0.7 \times 0.3} \approx 47.6$, lo stesso valore di $J(\hat{p})$.

Normalità asintotica (sotto condizioni di regolarità)

Per n grande, lo stimatore di massima verosimiglianza $\hat{\Theta}$ ha distribuzione approssimata data da

$$\hat{\Theta} \;\dot{\sim}\; \mathrm{N}\Big(\theta,\; \frac{1}{n\, I(\theta)}\Big)$$

- È un risultato analogo al TLC, ma vale per qualsiasi parametro che stimiamo con la massima verosimiglianza, non solo per la media campionaria.
- Lo SMV è (approssimatamente) non distorto, con varianza $1/(nI(\theta))$: "più informazione, meno variabilità".
- **Varianza stimata**:

$$\hat{\mathrm{var}}(\hat{\Theta}) = \frac{1}{n\, I(\hat{\theta})} \qquad \text{oppure} \qquad \hat{\mathrm{var}}(\hat{\Theta}) = \frac{1}{J(\hat{\theta})}$$

Intervallo di Wald

Dalla normalità asintotica, per la SMV otteniamo un intervallo di confidenza a livello $1-\alpha$ come :

$$\mathrm{IC}_{1-\alpha}(\theta) = \hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\mathrm{var}}(\hat{\Theta})} \qquad \text{(intervallo di Wald)}$$

Esempio: trattamento, 7 risposte su $n = 10$.

- $\hat{p} = 0.7, \quad \widehat{\mathrm{SE}} = \sqrt{\dfrac{0.7 \times 0.3}{10}} \approx 0.145$
- IC al 95%: $0.7 \pm 1.96 \times 0.145 \;\Rightarrow\; [0.416,\ 0.984]$
- IC al 99%: $0.7 \pm 2.57 \times 0.145 \;\Rightarrow\; [0.327; 1.073]$

Con n piccolo l'approssimazione normale è grossolana: l'intervallo è simmetrico intorno a $\hat{p}$ e al livello 0.99 supera 1, mentre $p \in [0, 1]$.

Quando l'approssimazione non è adatta (per esempio, n è troppo piccolo), esiste qualcosa di meglio? A volte sì: l'**inferenza esatta**. In alcuni modelli siamo in grado di calcoalre la distribuzione esatta dello stimatore, e quindi di fare inferenza esatta.

3 Inferenza esatta nel modello normale

Distribuzione degli stimatori

Siano $X_1, \ldots, X_n$ i.i.d. $\mathrm{N}(\mu, \sigma^2)$, abbiamo visto che gli stimatori di massima verosimiglianza sono

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{e} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Inoltre, sappiamo che lo stimatore della varianza è distorto, ma possiamo definire la versione corretta

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \left(\text{cioè: } S_c^2 = \tfrac{n}{n-1}S^2\right)$$

Risultato 1. Distribuzione dello stimatore della media: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ è combinazione lineare di normali indipendenti, quindi è normale, con

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n} \quad \Longrightarrow \quad \bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$$

Risultato 2. Distribuzione dello stimatore della varianza: si può dimostrare che

$$\frac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$$

Risultato 3. $\bar{X}$ e S_c^2 sono **indipendenti**

Risultato preliminare: distribuzione t di Student

Siano $Z \sim \mathrm{N}(0,1)$ e $V \sim \chi^2_k$ indipendenti, allora

$$T = \frac{Z}{\sqrt{V/k}}$$

ha distribuzione t di Student con k gradi di libertà, $T \sim t_k$.

Per calcolare un intervallo di confidenza dobbiamo prima trovare una quantità pivotale. Proviamo a standardizzare:

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0,1)$$

Tuttavia, σ^2 non è nota. Quindi questa **non** è una quantità pivotale. Consideriamo la trasformazione in cui a σ^2 sostituiamo il suo stimatore S_c^2:

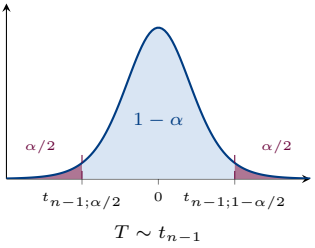
$$T = \frac{\bar{X}-\mu}{\sqrt{S_c^2/n}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{S_c^2/\cancel{n}}{\sigma^2/\cancel{n}}}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{S_c^2}{\sigma^2}}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{S_c^2(n-1)}{\sigma^2(n-1)}}} = \frac{Z}{\sqrt{V/(n-1)}}$$

con $Z = \sqrt{n}(\bar{X}-\mu)/\sigma \sim \mathrm{N}(0,1)$ (risultato 1) e $V = (n-1)S_c^2/\sigma^2 \sim \chi^2_{n-1}$ (risultato 2), indipendenti (risultato 3). Per il risultato sopra, $T \sim t_{n-1}$.

Intervallo di confidenza per μ

T dipende dai dati e da μ soltanto, e la sua distribuzione è completamente nota (non dipende da σ^2, ignota). Per questo T è una quantità pivotale.
Intervallo di confidenza:

$$\mathrm{P}\left(t_{n-1;\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{S_c^2/n}} \leq t_{n-1;1-\alpha/2}\right) = 1-\alpha$$



Isolando μ nella disuguaglianza otteniamo l'intervallo di confidenza di *livello esatto* $1-\alpha$:

$$\mathrm{P}\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{S_c}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{S_c}{\sqrt{n}}\right) = 1-\alpha$$

(per simmetria, $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$).

Coi dati, otteniamo l'intervallo

$$\mathrm{IC}_{1-\alpha}(\mu) = \bar{x} \pm t_{n-1;1-\alpha/2}\,\frac{s_c}{\sqrt{n}}$$

Intervallo di confidenza per μ - esempio

Esempio: $n = 10$ misure, $\bar{x} = 5.2$, $s_c = 1.8$. Quantile: $t_{9,\,0.975} = 2.262$.

- $\dfrac{s_c}{\sqrt{n}} = \dfrac{1.8}{\sqrt{10}} \approx 0.569$
- IC esatto al 95%: $\;5.2 \pm 2.262 \times 0.569 \;\Rightarrow\; [3.91,\ 6.49]$
- Intervallo basato sulla distribuzione asintotica normale al 95% (usando quindi il quantile $z_{0.975} = 1.96$): $5.2 \pm 1.96 \times 0.569 \Rightarrow [4.08,\ 6.32]$

L'intervallo con i quantili normali è troppo stretto: per $n = 10$ il suo livello di copertura reale è inferiore al 95%. L'intervallo basato su t ha copertura esattamente 95%.

n	5	10	30	100	1000
$t_{n-1,\,0.975}$	2.776	2.262	2.045	1.984	1.962

Intervallo di confidenza per σ^2

Abbiamo già visto che la distribuzione dello stimatore è

$$\frac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}$$

questa trasformazione è già una quantità pivotale, che può essere usata per costruire un intervallo di confidenza.

$$\mathrm{P}\left(\chi^2_{n-1;\alpha/2} \leq \frac{(n-1)S_c^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) = 1-\alpha$$

$$\mathrm{P}\left(\frac{1}{\chi^2_{n-1;1-\alpha/2}} \leq \frac{\sigma^2}{(n-1)S_c^2} \leq \frac{1}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha$$

$$\mathrm{P}\left(\frac{(n-1)S_c^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_c^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha$$

Con i dati, otteniamo l'intervallo

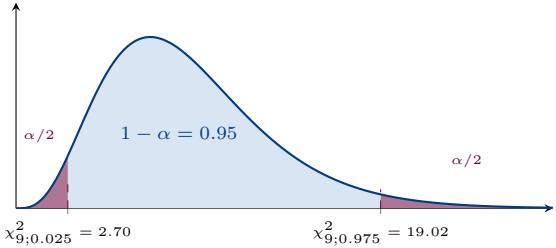
$$\mathrm{IC}_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)s_c^2}{\chi^2_{n-1,\,1-\alpha/2}},\ \frac{(n-1)s_c^2}{\chi^2_{n-1,\,\alpha/2}}\right]$$

Intervallo di confidenza per σ^2 - esempio

Con i dati di prima, $n = 10$, $s_c = 1.8$: $(n-1)s_c^2 = 29.16$.
I quantili sono: $\chi^2_{9,0.975} = 19.02$, $\chi^2_{9,0.025} = 2.70$. Otteniamo quindi l'intervallo

$$\mathrm{IC}_{95\%}(\sigma^2) \approx [1.53,\ 10.80]$$

L'intervallo è asimmetrico (non è della forma "stima $\pm$ margine") e abbastanza ampio: 10 osservazioni non sono sufficienti per stimare la varianza in modo preciso.



Test sulla media

Vogliamo testare il seguente sistema di ipotesi:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Sotto H_0, lo stimatore della media ha distribuzione

$$\bar{X} \overset{H_0}{\sim} \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \implies \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \overset{H_0}{\sim} \mathrm{N}(0,1)$$

Tuttavia, anche in questo caso σ^2 non è nota. Consideriamo la stessa trasformazione che abbiamo usato per costruire l'intervallo di confidenza, in cui a σ^2 sostituiamo il suo stimatore S_c^2:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S_c^2/n}} = \frac{\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{S_c^2(n-1)}{\sigma^2(n-1)}}} = \frac{Z}{\sqrt{V/(n-1)}} \overset{H_0}{\sim} t_{n-1}$$

Con i dati, calcoliamo il valore osservato del test,

$$t^{oss} = \frac{\bar{x} - \mu_0}{\sqrt{s_c^2/n}}$$

Test sulla media - livello α fissato

Test con **livello di significatività α fissato**

- Valori grandi (in valore assoluto) di t^{oss} sono in contrasto con H_0 (assumendo H_0 vera). La regione di rifiuto sarà quindi per valori grandi di t^{oss}, ovvero sulle code

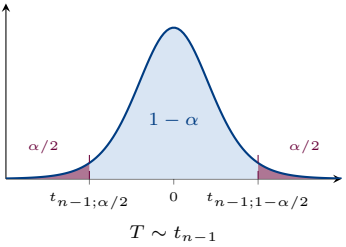
$$\mathcal{R} = (-\infty, a) \cup (b, +\infty)$$

Vogliamo inoltre che

$$\mathrm{P}_{H_0}(T \in \mathcal{R}) = \alpha$$

ovvero

$$\mathrm{P}_{H_0}((T < a) \cup (T > b)) = \alpha$$



I valori a e b che stiamo cercando sono quindi i quantili $t_{n-1;\alpha/2}$ e $t_{n-1;1-\alpha/2}$. Per simmetria, $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$

$$\mathcal{R} = (-\infty, -t_{n-1;1-\alpha/2}) \cup (t_{n-1;1-\alpha/2}, +\infty)$$

- Conclusione del test: se $t^{oss} \in \mathcal{R}$ rifiutiamo H_0 a livello α; altrimenti non rifiutiamo H_0.

Test sulla media - α osservato

- Se vogliamo invece calcolare il **p-value**, ci chiediamo
  "qual è la probabilità di osservare valori più estremi (in entrambe le direzioni) di t^{oss}, se H_0 è vera?".

$$\begin{aligned}\alpha^{oss} &= \mathrm{P}_{H_0}(|T| > |t^{oss}|) \\ &= \mathrm{P}_{H_0}((T < -|t^{oss}|) \cup (T > |t^{oss}|))\end{aligned}$$

