

Stime puntuali

Abbiamo visto che, data una v.a. X di cui non si conoscano valore atteso e varianza, tali numeri si possono stimare *puntualmente* nel seguente modo:

- si prende un campione casuale X_1, \dots, X_n di v.a. aventi la stessa legge di X ;
- si effettuano gli esperimenti che ci danno n numeri x_1, \dots, x_n (x_i è il valore osservato – o realizzazione – di X_i);
- per $\mathbb{E}(X)$ la stima è la realizzazione della media campionaria $\bar{x}_n = \frac{1}{n}(x_1 + \dots + x_n)$;
- per $\text{Var}(X)$ la stima è la realizzazione della varianza campionaria $s_n^2 = \frac{1}{n-1}(x_i - \bar{x}_n)^2$.

Notiamo che le lettere maiuscole indicano le v.a. (prima dell'esperimento), quelle minuscole indicano i numeri ottenuti (dopo l'esperimento, sono la realizzazione dell'esperimento).

Stime puntuali

Sia X_1, \dots, X_n un campione casuale di una legge che ha valore atteso μ .

Facciamo gli n esperimenti che realizzano il campione e calcoliamo la media campionaria.

Supponiamo che si abbia $\bar{x}_n = 5$. Questo significa che

possiamo dire con certezza che $\mu = 5$?

Certo che no! pensate solo ad una moneta equilibrata (cioè a una $\mathcal{B}(0.5)$) che in 20 lanci dia 12 teste e 8 croci. La media campionaria sarebbe quindi $\frac{12}{20} = 0.6$!

Stime puntuali

Oppure: possiamo forse dire che

$$\mathbb{P}(\bar{X}_n = \mu) \xrightarrow{n \rightarrow \infty} 1?$$

Questo ci direbbe che con grande probabilità la media campionaria coincide con μ .

Purtroppo anche **questa affermazione è falsa**. Anzi,

la probabilità che \bar{X}_n coincida con μ tende a zero!

Basta riscrivere la probabilità spostando μ e usare il teorema del limite centrale:

$$\begin{aligned}\mathbb{P}(\bar{X}_n - \mu = 0) &= \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = 0\right) \\ &\xrightarrow{n \rightarrow \infty} \mathbb{P}(\mathcal{N}(0, 1) = 0) = 0.\end{aligned}$$

Intervalli di confidenza

Allora che fare? Ci viene in soccorso la legge dei grandi numeri, che ci dice che, fissato $\varepsilon > 0$

la probabilità che $X_n \in (\mu - \varepsilon, \mu + \varepsilon)$ tende a 1 cioè
la probabilità che $\mu \in (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)$ tende a 1

L'idea è che invece di una stima puntuale

si può fornire un **intervallo di confidenza** cioè

- un intervallo (aleatorio)
- basato sul campione casuale
- che con grande probabilità* (a priori) contiene il parametro incognito.

* Chiederemo che tale probabilità sia uguale a un valore fissato che chiameremo *livello di confidenza*.

Definizione

Sia X_1, \dots, X_n un campione casuale estratto da una popolazione di legge $f(x; \theta)$ e si voglia stimare $g(\theta)$.

DEFINIZIONE DI INTERVALLO DI CONFIDENZA

Siano $T_1 = t_1(X_1, \dots, X_n)$ e $T_2 = t_2(X_1, \dots, X_n)$ due statistiche. Fissato $\alpha \in [0, 1]$, l'intervallo aleatorio (T_1, T_2) si dice **intervallo di confidenza per $g(\theta)$, al livello α** se

$$\mathbb{P}_\theta(T_1 < g(\theta) < T_2) = \alpha, \quad \text{per ogni } \theta \in \Theta.$$

A campionamento eseguito, l'intervallo numerico $[t_1(x_1, \dots, x_n), t_2(x_1, \dots, x_n)]$ si chiama intervallo di confidenza per $g(\theta)$, al livello α , **calcolato dal campione**.

Sul significato

(T_1, T_2) è un **intervallo aleatorio** (gli estremi dell'intervallo sono v.a.) che contiene il parametro $g(\theta)$ con probabilità α .

NON è invece **vero** che la probabilità che $t_1(x_1, \dots, x_n) < g(\theta) < t_2(x_1, \dots, x_n)$ è uguale ad α .

Ciò è dovuto al fatto che l'intervallo

$$\left(t_1(x_1, \dots, x_n), t_2(x_1, \dots, x_n) \right),$$

che si ottiene a esperimenti fatti, è numerico e non più aleatorio.

Per questo motivo si parla di *confidenza* e non di probabilità.

All'atto pratico

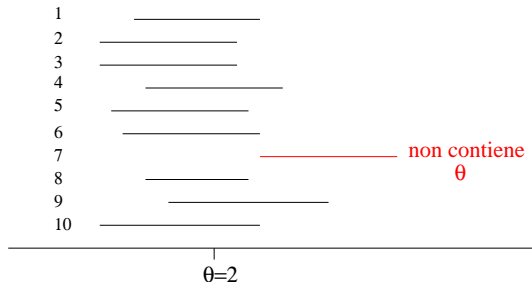
Tuttavia è vero che se

- decidiamo che il nostro esperimento è “ottenere un campione casuale di numerosità n ” e poi calcolare un intervallo di confidenza I , basato su quel campione, al livello α per $g(\theta)$;
- ripetiamo k volte l'esperimento ottenendo I_1, \dots, I_k intervalli di confidenza al livello α per $g(\theta)$;

in virtù della Legge dei Grandi Numeri, una proporzione del $100\alpha\%$ (circa) degli intervalli calcolati conterrà il valore di $g(\theta)$.

Esempio

Supponiamo che $\theta = 2$ e calcoliamo 10 intervalli di confidenza al livello 90%: *mediamente* un intervallo su 10 NON conterrà θ^* .



In genere nei problemi di stima non conosciamo il valore vero di θ e quindi non sappiamo quali intervalli contengono θ e quali no.

* Il numero di intervalli che non lo contengono è una $\mathcal{B}(10, 0.10)$.

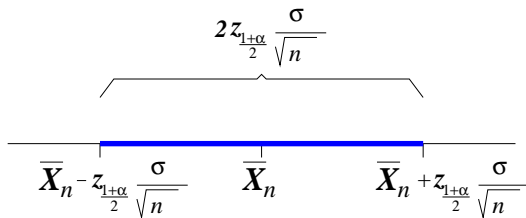
Intervallo per μ per popolazioni normali – σ^2 nota

Sia X_1, \dots, X_n un campione casuale estratto da una popolazione di legge $\mathcal{N}(\mu, \sigma^2)$ dove si conosca il valore di σ^2 e si voglia stimare μ .

INTERVALLO DI CONFIDENZA per μ , σ^2 nota

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

è un intervallo di confidenza di livello α per il valore atteso μ .



Perché vale la formula

Se la popolazione è $\mathcal{N}(\mu, \sigma^2)$

La media campionaria \bar{X}_n ha legge $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ e

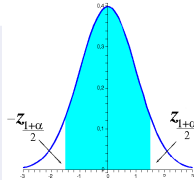
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

La chiave è mostrare che

$$\mathbb{P} \left(-z_{\frac{1+\alpha}{2}} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\frac{1+\alpha}{2}} \right) = \alpha.$$

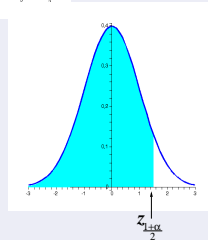
Perché vale la formula

Mostriamo cioè che



vale α .

Infatti l'area a sinistra di $z_{\frac{1+\alpha}{2}}$ è



e vale $\frac{1+\alpha}{2}$,

quella a destra vale $1 - \frac{1+\alpha}{2} = \frac{1-\alpha}{2}$.

Dunque l'area fra $-z_{\frac{1+\alpha}{2}}$ e $z_{\frac{1+\alpha}{2}}$ vale $\frac{1+\alpha}{2} - \frac{1-\alpha}{2} = \alpha$.

Perché vale la formula

Ne consegue che

$$\mathbb{P} \left(-z_{\frac{1+\alpha}{2}} < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\frac{1+\alpha}{2}} \right) = \alpha.$$

Ricavando μ da queste disuguaglianze si ricava la formula

$$\mathbb{P} \left(\bar{X}_n - z_{\frac{1+\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\frac{1+\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = \alpha.$$

Cosa fare se σ^2 è incognita

Se non si conosce σ^2 la formula dell'intervallo precedente non è utilizzabile, infatti gli estremi, dipendendo anche da σ , non sono più statistiche (perché? vi ricordate la definizione di statistica?).

Idea:

sostituire a σ lo stimatore $\sqrt{S_n^2}$

L'idea è quasi del tutto corretta, nel senso che diviene corretta se si fa anche la sostituzione

$t_{\frac{1+\alpha}{2}}(n-1)$ al posto di $z_{\frac{1+\alpha}{2}}$

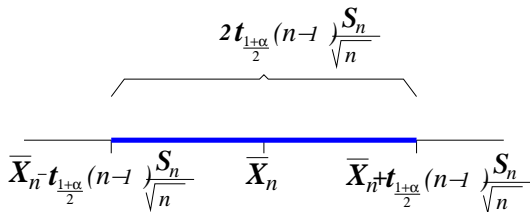
cioè al quantile della normale standard $z_{\frac{1+\alpha}{2}}$ sostituiamo il quantile della legge di Student a $n-1$ gradi di libertà $t_{\frac{1+\alpha}{2}}(n-1)$.

Intervallo per μ per popolazioni normali – σ^2 è incognita

INTERVALLO DI CONFIDENZA per μ , σ^2 incognita

$$\left(\bar{X}_n - t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}}, \quad \bar{X}_n + t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}} \right)$$

è un intervallo di confidenza di livello α per il valore atteso μ .



Perché vale la formula

Se la popolazione è $\mathcal{N}(\mu, \sigma^2)$

Che un intervallo di confidenza di livello α per il valore atteso μ sia

$$\left(\bar{X}_n - t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}}, \quad \bar{X}_n + t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}} \right)$$

è conseguenza del fatto che

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{S_n^2}{n}}} \sim t(n-1),$$

dove $t(n-1)$ indica la v.a. continua che ha legge di Student a $n-1$ gradi di libertà.

La t di Student – tavole

	0.9	0.95	0.975	0.99	0.995
1	3.07768	6.31375	12.70620	31.82052	63.65674
2	1.88562	2.91999	4.30265	6.96456	9.92484
3	1.63774	2.35336	3.18245	4.54070	5.84091
4	1.53321	2.13185	2.77645	3.74695	4.60409
5	1.47588	2.01505	2.57058	3.36493	4.03214
6	1.43976	1.94318	2.44691	3.14267	3.70743
7	1.41492	1.89458	2.36462	2.99795	3.49948
8	1.39682	1.85955	2.30600	2.89646	3.35539
9	1.38303	1.83311	2.26216	2.82144	3.24984
10	1.37218	1.81246	2.22814	2.76377	3.16927
11	1.36343	1.79588	2.20099	2.71808	3.10581
12	1.35622	1.78229	2.17881	2.68100	3.05454
13	1.35017	1.77093	2.16037	2.65031	3.01228
14	1.34503	1.76131	2.14479	2.62449	2.97684
15	1.34061	1.75305	2.13145	2.60248	2.94671
16	1.33676	1.74588	2.11991	2.58349	2.92078
17	1.33338	1.73961	2.10982	2.56693	2.89823
18	1.33039	1.73406	2.10092	2.55238	2.87844
19	1.32773	1.72913	2.09302	2.53948	2.86093
20	1.32534	1.72472	2.08596	2.52798	2.84534
21	1.32319	1.72074	2.07961	2.51765	2.83136
22	1.32124	1.71714	2.07387	2.50832	2.81876
23	1.31946	1.71387	2.06866	2.49987	2.80734
24	1.31784	1.71088	2.06390	2.49216	2.79694
25	1.31635	1.70814	2.05954	2.48511	2.78744
26	1.31497	1.70562	2.05553	2.47863	2.77871
27	1.31370	1.70329	2.05183	2.47266	2.77068
28	1.31253	1.70113	2.04841	2.46714	2.76326
29	1.31143	1.69913	2.04523	2.46202	2.75639
30	1.31042	1.69726	2.04227	2.45726	2.75000
40	1.30308	1.68385	2.02108	2.42326	2.70446
50	1.29871	1.67591	2.00856	2.40327	2.67779
60	1.29582	1.67065	2.00030	2.39012	2.66028
70	1.29376	1.66691	1.99444	2.38081	2.64790
80	1.29222	1.66412	1.99006	2.37387	2.63869
90	1.29103	1.66196	1.98667	2.36850	2.63157
100	1.29007	1.66023	1.98397	2.36422	2.62589
110	1.28930	1.65882	1.98177	2.36073	2.62126
120	1.28865	1.65765	1.97993	2.35782	2.61742

La colonna indica l'area α , la riga indica di quale legge t stiamo parlando. Infatti per ogni intero n c'è una diversa legge $t(n)$.

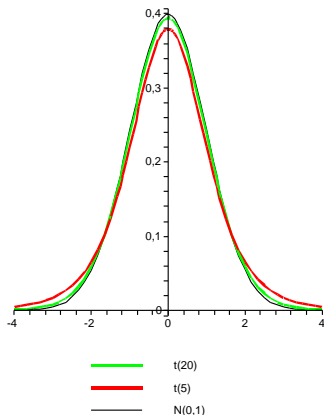
I quantili della t di Student

MOLTO IMPORTANTE

Nel simbolo del quantile $t_{\alpha}(n - 1)$, “ $(n - 1)$ ” sta ad indicare quale riga della tabella usare.

NON è un numero da moltiplicare per il quantile!!!

La t di Student – densità



La densità $t(n)$ ha le stesse proprietà di simmetria della $\mathcal{N}(0, 1)$ (quindi $\mathbb{E}(t(n)) = 0$).

Il picco è più basso, mentre le code sono più alte.

Al crescere di n , la densità $t(n)$ assomiglia sempre più alla $\mathcal{N}(0, 1)$.

Intervallo per μ per popolazioni qualsiasi

Se la popolazione non è normale, non è vero che

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Tuttavia per il Teorema del Limite Centrale, questa affermazione è **approssimativamente vera quando n è abbastanza grande**.

Allo stesso modo si potrebbe mostrare che anche

$$\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \sim t(n-1),$$

è **approssimativamente vera quando n è abbastanza grande**.

Conseguenza

Le formule viste per le popolazioni normali, valgono anche per **popolazioni qualsiasi** (purché il campione sia abbastanza numeroso da consentire l'applicazione del Teorema del Limite Centrale).

Intervallo di confidenza di livello α per il valore atteso, varianza nota

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

Intervallo di confidenza di livello α per il valore atteso, varianza incognita

$$\left(\bar{X}_n - t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}}, \quad \bar{X}_n + t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}} \right)$$

Intervallo per p di $\mathcal{B}(p)$

Se la popolazione è bernoulliana e si vuole stimare il parametro p , la stima puntuale è fornita da \bar{X}_n (p è anche il valore atteso).

Se n è abbastanza grande da consentire l'applicazione del Teorema del Limite Centrale*, allora sappiamo che

$$\bar{X}_n \approx \mathcal{N}(p, p(1 - p)/n).$$

* Ricordiamo che occorre $np \geq 5$ e $n(1 - p) \geq 5$.

Se p è incognito (come nei problemi di stima) si verifica che sia $n\bar{x}_n \geq 5$ e $n(1 - \bar{x}_n) \geq 5$.

Intervallo per p di $\mathcal{B}(p)$

Si considera quindi l'intervallo per le normali:

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

e al posto della varianza (della singola variabile) σ^2 si sostituisce $\bar{X}_n(1 - \bar{X}_n)$ (ricordiamo che la varianza di una $\mathcal{B}(p)$ è $p(1 - p)$).

Intervallo di confidenza di livello α per p

$$\left(\bar{X}_n - z_{\frac{1+\alpha}{2}} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}, \quad \bar{X}_n + z_{\frac{1+\alpha}{2}} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right).$$

Perché vale la formula

Ci si potrebbe chiedere il motivo per cui abbiamo utilizzato la formula dell'intervallo a varianza nota (quindi con i quantili z) e non quello a varianza incognita (quindi con i quantili t). In effetti sui testi si trovano anche formule con i quantili t , ma si possono osservare due fatti:

- per usare entrambe le formule occorre n grande e quindi i quantili $t(n - 1)$ e i quantili z non sono molto differenti;
- la varianza in un certo senso non è *del tutto incognita*, in quanto una volta stimato p con \bar{X}_n , la varianza risulta stimata da $\bar{X}_n(1 - \bar{X}_n)$ (dunque ha senso usare la formula per varianza nota).

Intervalli e livello



Ricordiamo che dare un intervallo di confidenza di livello α per un certo parametro è un po' come “sparare” a un bersaglio (il parametro) che non sappiamo dove sia.

Quello che vogliamo è colpirlo con probabilità α .

È chiaro che più grande è il “proiettile” (l'ampiezza dell'intervallo) e più probabile sarà colpire il bersaglio.

Intervalli e livello

Se consideriamo due intervalli di confidenza I_1 e I_2 per lo stesso parametro, basati sullo stesso campione, I_1 con livello α_1 e I_2 con livello α_2 , con $\alpha_1 < \alpha_2$, cosa possiamo dire di questi due intervalli?

Per definizione la probabilità che I_2 contenga il parametro è maggiore della probabilità che I_1 contenga il parametro.

È ragionevole aspettarsi che questo accada perché I_2 è più ampio di I_1 (**livello più alto \Rightarrow ampiezza maggiore**).

Vediamo che è in effetti così negli intervalli per il valore atteso della normale.

Gli intervalli per μ

Con varianza nota, un intervallo di livello α per il valore atteso di una $\mathcal{N}(\mu, \sigma^2)$ è:

$$\begin{array}{c}
 \underbrace{\hspace{10em}}_{2z_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}} \\
 \hline
 \bar{X}_n - z_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \quad \bar{X}_n \quad \quad \bar{X}_n + z_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}
 \end{array}$$

se la varianza è incognita invece è:

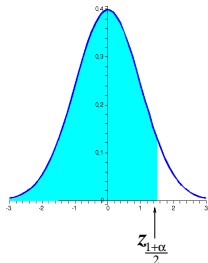
$$\begin{array}{c}
 \underbrace{\hspace{10em}}_{2t_{\frac{1+\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}}} \\
 \hline
 \bar{X}_n - t_{\frac{1+\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}} \quad \quad \bar{X}_n \quad \quad \bar{X}_n + t_{\frac{1+\alpha}{2}}(n-1) \frac{S_n}{\sqrt{n}}
 \end{array}$$

Ampiezze e livello

Le ampiezze dipendono da α attraverso i quantili $z_{\frac{1+\alpha}{2}}$ e $t_{\frac{1+\alpha}{2}}(n-1)$ rispettivamente.

Basta un'occhiata alle tavole per rendersi conto che i quantili crescono al crescere di α .

Oppure si può ragionare sul fatto che



l'area a sinistra di $z_{\frac{1+\alpha}{2}}$ vale $\frac{1+\alpha}{2}$ e se α (e dunque l'area azzurra) cresce, allora $z_{\frac{1+\alpha}{2}}$ deve spostarsi verso destra (e quindi crescere). Lo stesso ragionamento vale per i quantili $t_{\frac{1+\alpha}{2}}(n-1)$.

Intervalli e numerosità del campione

Chiaramente è desiderabile avere un intervallo di confidenza stretto perché dà una stima più precisa.

Da quanto appena visto sembrerebbe che per avere intervalli stretti occorra necessariamente scegliere α piccolo.

Ma d'altra parte α lo vogliamo vicino a 1!!! Come risolvere questo dilemma?

Intervalli e numerosità del campione

Ricordiamo le ampiezze degli intervalli per il valore atteso di una $\mathcal{N}(\mu, \sigma^2)$. Se σ è nota

$$2z_{\frac{1+\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

e se σ è incognita

$$2t_{\frac{1+\alpha}{2}}(n-1) \cdot \sqrt{\frac{S_n^2}{n}}.$$

La strada per restringere l'ampiezza

In entrambi i casi è aumentare n (il solito buon senso: più dati abbiamo meglio è!!!)