

Inferenza Statistica

L'inferenza statistica cerca di risalire al modello del fenomeno sulla base delle osservazioni.

Non conosciamo il modello del fenomeno cioè la v.c. X . A volte la conoscenza può essere parziale (conosciamo la forma ma non i parametri)

Sulla base di n osservazioni x_1, x_2, \dots, x_n vogliamo risalire al modello, cioè alla conoscenza di X .

Ipotesi fondamentale: le osservazioni x_1, x_2, \dots, x_n sono le realizzazioni di n v.c. X_1, X_2, \dots, X_n indipendenti e identicamente distribuite come X (i.i.d.)

Esempio: supponiamo che la resistenza alla compressione di tipo di calcestruzzo sia una v.c. X gaussiana con media incognita μ e varianza $\sigma^2 = 4$. Scriveremo $X \sim N(\mu, \sigma^2)$. Supponiamo di osservare la resistenza alla compressione su n campioni di quel tipo di calcestruzzo. Indichiamo con x_1, x_2, \dots, x_n i valori osservati. Sulla base di questi valori vogliamo risalire al valore incognito di μ .

Abbiamo varie possibilità:

1. Stima puntuale
2. Stima per intervalli
3. Verifiche di ipotesi

Stima Puntuale

I parametri che ci interessa stimare (sono incogniti) sono media e varianza. Presentiamo gli stimatori della media incognita e della varianza incognita ottenuti come momenti campionari.

Lo **stimatore** è una funzione che ci fornisce il valore da stimare in funzione del campione osservato.

Nella parte di statistica inferenziale avremo a che fare con v.c. ottenute come funzioni di v.c. note.

Per gli stimatori della media e della varianza incogniti calcoleremo i valori di sintesi più importanti: media e varianza!

Variabile casuale media campionaria

Abbiamo un modello X di cui non conosciamo la media $E(X) = \mu$. μ è l'incognita da stimare. Siano x_1, x_2, \dots, x_n le osservazioni di X .

Denotiamo con X_1, X_2, \dots, X_n un campione i.i.d. di cui x_1, x_2, \dots, x_n sono la realizzazione. La media campionaria è definita come

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Stimeremo quindi il valore incognito μ con

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

Proprietà della media campionaria

Se $E(X_i) = \mu$ e $\text{Var}(X_i) = \sigma^2$ allora

$$E(\bar{X}_n) = \mu \quad \text{e} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

Infatti

$$E(\bar{X}_n) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu$$

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Esempio: consideriamo la variabile casuale X avente distribuzione

x_i	1	2	3
p_i	1/3	1/3	1/3

Abbiamo: $E(X) = \mu = 2$ e $\text{Var}(X) = \sigma^2 = 2/3$. Estraiamo un campione di numerosità 2 con riposizione. I campioni possibili sono

$(1, 1); (1, 2); (1, 3); (2, 1); (2, 2); (2, 3); (3, 1); (3, 2); (3, 3)$

Consideriamo la v.c. \bar{X}_n . La sua distribuzione è

\bar{x}_i	1	1.5	2	2.5	3
p_i	1/9	2/9	3/9	2/9	1/9

$$E(\bar{X}_n) = 2 = \mu, \text{ e } \text{Var}(\bar{X}_n) = 1/3 = \sigma^2/2$$

Esempio: un laboratorio produce molle per orologi con due macchine diverse denotate con A e B . La seguente tabella riporta i dati delle resistenze delle molle prodotte dalle due macchine.

A	22	25	24	25	25	27	26	26	22	26	23	25
B	23	24	23	26	24	22	25	24	24			

Stimare la media incognita delle resistenza delle molle prodotte dalle due macchine

Denotiamo con X_A e X_B rispettivamente i modelli per la resistenza delle molle prodotte rispettivamente dalla macchina A e dalla macchina B.

La media $E(X_A) = \mu_A$ e $E(X_B) = \mu_B$ sono incognite. Le stimiamo rispettivamente con la media campionaria calcolata per i due gruppi:

$$\hat{\mu}_A = \frac{1}{n} \sum_{i=1}^{12} x_i^A = \frac{296}{12} = 24.67$$

$$\hat{\mu}_B = \frac{1}{n} \sum_{i=1}^9 x_i^B = 23.89$$

Queste sono le stime puntuali per la media incognita delle popolazioni considerate.

Si faccia attenzione tra:

1. la media incognita μ
2. la media campionaria stimatore $\frac{1}{n} \sum X_i$
3. la media stimata $\frac{1}{n} \sum x_i$
4. la media dello stimatore $E(\frac{1}{n} \sum X_i)$

Variabile casuale varianza campionaria

Supponiamo ora che del modello X sia incognita la varianza σ^2 . Se, come prima, X_1, X_2, \dots, X_n è un campione i.i.d. allora la varianza campionaria è definita come

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Stimeremo il valore incognito σ^2 con

$$\hat{\sigma}^2 = s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Si verifica che

$$E(S_n^2) = \sigma^2, \quad \text{Var}(S_n^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)$$

Riprendiamo l'esempio e calcoliamo la distribuzione di S^2

campione	\bar{x}_i	s_i^2
1 1	1	0
1 2	1.5	0.5
1 3	2	2
2 1	1.5	0.5
2 2	2	0
2 3	2.5	0.5
3 1	2	2
3 2	2.5	0.5
3 3	3	0

La distribuzione è

s_i^2	0	0.5	2
p_i	3/9	4/9	2/9

Da cui $E(S_n^2) = 2/3 = \sigma^2$

Esempio: (continua) stimare la varianza incognita per la varianza della resistenza delle molle prodotte dalle due macchine A e B.

$$\hat{\sigma}_A^2 = \frac{1}{n-1}(\sum x_{Ai}^2 - n(\bar{x}_A)^2) = \frac{1}{11}(7330 - 12(24.67)^2) = 2.61$$

$$\hat{\sigma}_B^2 = \frac{1}{n-1}(\sum x_{Bi}^2 - n(\bar{x}_B)^2) = \frac{1}{8}(5147 - 9(23.89)^2) = 1.36$$

Si noti che le due stime ottenute sono diverse dalla varianza che era stata calcolata precedentemente

$$\sigma_A^2 = \frac{1}{n} \sum x_i^2 \cdot n_i - (\bar{x}_A)^2 = \frac{1}{12} 7330 - (24.67)^2 = 2.39$$

Facendo i conti analoghi per il gruppo B si trova $\sigma_B^2 = 1.21$.

Le stime ottenute si chiamano **non distorte** in quanto il valore atteso degli stimatori è uguale al parametro da stimare.

Casi particolari importanti

Fino ad ora non abbiamo fatto nessuna ipotesi sulla distribuzione del modello X . Abbiamo solo supposto che la media e la varianza fossero incognite.

Supponiamo ora che X sia un modello gaussiano

Se $X \sim N(\mu, \sigma^2)$ si ha che anche la media campionaria è gaussiana

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Supponiamo ora che X sia un modello bernoulliano

Se $X \sim Ber(p)$ si ha che

$$E(\bar{X}_n) = E(\hat{p}_n) = p \quad \text{e} \quad \text{Var}(\hat{p}_n) = \frac{p(1-p)}{n}$$

La distribuzione di \hat{p}_n è la seguente

$$P\left(\hat{p}_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n$$

Lo stimatore \hat{p}_n è detto *proporzione campionaria* ed è ottenuto come media di variabili che possono solo assumere valore 0 o 1.

Osservazione

Quando diciamo che la media stimata è un certo valore è meglio dare anche l'errore standard, in modo da avere un'idea della variabilità di tale stima

Esempio: sulla base di un campione, la percentuale di preferenze al partito A è $\hat{p}_n = 50.5\%$. Possiamo concludere che A ha vinto le elezioni?

Quello che ci manca è la precisione del dato 50.5%. Sappiamo che $\text{Var}(\hat{p}_n) = p(1 - p)/n$. Possiamo ottenere una stima della varianza di \hat{p}_n sostituendo al valore incognito p il valore \hat{p}_n .

Vediamo cosa accade per n pari a 100 o a 1000.

$$\sqrt{\frac{\hat{p}_n(1 - \hat{p}_n)}{n}} = \sqrt{\frac{0.505 \cdot 0.495}{100}} = 0.05 = 5\%$$

$$\sqrt{\frac{\hat{p}_n(1 - \hat{p}_n)}{n}} = \sqrt{\frac{0.505 \cdot 0.495}{1000}} = 0.0158 = 1.6\%$$

Consideriamo l'intervallo $\hat{p}_n \pm \sqrt{\frac{\hat{p}_n(1 - \hat{p}_n)}{n}}$

$$(50.5\% - 5\%, 50.5\% + 5\%) = (45.5\%, 55.5\%)$$

e

$$(50.5\% + 1.6\%, 50.5\% - 1.6\%) = (48.9\%, 52.1\%)$$

Gli intervalli di confidenza

Gli *intervalli di confidenza per la media* forniscono un campo di variazione (centrato sulla media campionaria) all'interno del quale ci si aspetta di trovare il parametro incognito μ

Ad ogni intervallo di confidenza viene associato un *livello di confidenza* $(1 - \alpha)$ che rappresenta il grado di attendibilità del nostro intervallo

Il nostro scopo è quello di determinare un intervallo di valori (a, b) che contenga il valore incognito μ . Vorremmo poter scrivere

$$P(a < \mu < b) = 1 - \alpha$$

Ma questa scrittura è priva di senso!!!!

Infatti μ è un numero.

Chiariamo come uscire dall'inghippo con un esempio fondamentale

Se X_1, X_2, \dots, X_n è un campione i.i.d. di variabili casuali Gaussiane di media incognita μ e varianza σ^2 , sappiamo che la media campionaria \bar{X}_n è una variabile aleatoria Gaussianiana di media μ e varianza σ^2/n .

Possiamo quindi scrivere la relazione sopra come

$$P\left(a < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < b\right) = 1 - \alpha$$

che corrisponde a scrivere

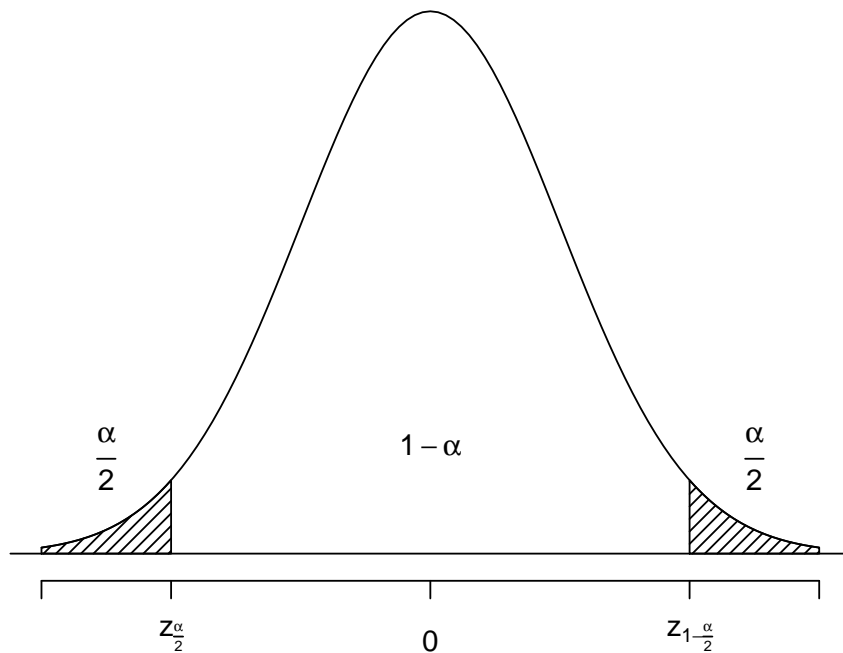
$$P(a < Z < b) = 1 - \alpha$$

con $Z \sim N(0, 1)$

Possiamo scegliere a e b come

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

Basta infatti osservare il disegno



Ricordando che $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, possiamo riscrivere l'espressione

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

Svolgiamo ora i calcoli necessari per arrivare ad un intervallo in termini della media incognita μ

Abbiamo

$$\begin{aligned}
 & P \left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \right) \\
 &= P \left(-z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \\
 &= P \left(-z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \bar{X}_n < -\mu < z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \bar{X}_n \right) \\
 &= P \left(z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \bar{X}_n > \mu > -z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \bar{X}_n \right) \\
 &= P \left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)
 \end{aligned}$$

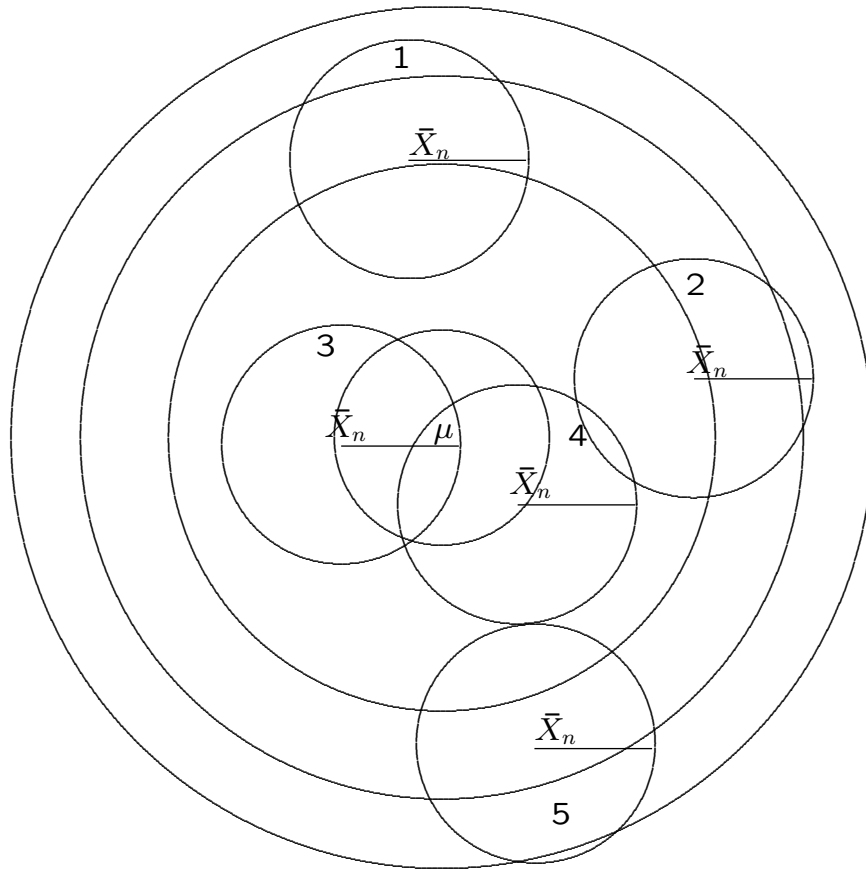
Osservazione: la probabilità non è riferita a μ , che è una costante, ma a \bar{X}_n che è una v.c.

In sostanza potremmo scrivere che

$$\mu \in \left(\bar{X}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

e siamo fiduciosi che questo accada nell' $(1 - \alpha)\%$ dei casi, cioè nell' $(1 - \alpha)\%$ dei campioni estratti

L'intervallo di confidenza è un intervallo i cui estremi sono aleatori



Il livello di confidenza può essere visto come la frequenza di questi intervalli aleatori che contengono il valore incognito μ

Alcuni intervalli (cerchi) non contengono il valore μ (gli intervalli 1, 2 e 5) altri invece lo contengono (gli intervalli 3 e 4). Si può interpretare il livello di confidenza $1 - \alpha$ come la frequenza degli intervalli che contengono il valore incognito μ .

Ecco perché è scorretto parlare del livello di confidenza come della probabilità che il nostro parametro sia contenuto nell'intervallo.

Esempio: consideriamo un campione di ampiezza $n = 27$ proveniente da una popolazione Normale con media μ incognita e varianza nota $\sigma^2 = 44$. Determinare l'intervallo di confidenza a livello $1 - \alpha = 0.95$.

Sappiamo che

$$Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_n - \mu}{\sqrt{44/27}} = N(0, 1)$$

Inoltre dalle tavole della Z ricaviamo

$$P(Z < -1.96) = P(Z > 1.96) = \frac{\alpha}{2} = 0.025$$

Abbiamo

$$\begin{aligned} P\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(-1.96 < \frac{\bar{X}_n - \mu}{\sqrt{\frac{44}{27}}} < 1.96\right) \\ &= P\left(\bar{X}_n - 1.96\sqrt{\frac{44}{27}} < \mu < \bar{X}_n + 1.96\sqrt{\frac{44}{27}}\right) \\ &= P(\bar{X}_n - 2.5 < \mu < \bar{X}_n + 2.5) = 0.95 \end{aligned}$$

Possiamo dire che con un grado di fiducia pari a 0.95 (95%) riteniamo che l'intervallo $(\bar{X}_n - 2.5, \bar{X}_n + 2.5)$ contenga la suo interno il valore incognito μ

A questo punto entrano in gioco le osservazioni!! Se abbiamo rilevato $\bar{x}_n = 174.5$, l'intervallo di confidenza risulta $(172, 177)$

È ragionevole pensare che la media incognita μ sia compresa nell'intervallo $(172, 177)$ nel 95% dei casi

È sbagliato affermare che la probabilità che μ sia nell'intervallo $(172, 177)$ è 0.95, perché μ non è una variabile casuale

Riassumendo: Intervallo di confidenza per la media (σ^2 nota) nel caso di popolazione Gaussiana

Sia X una v.c Gaussiana di media μ e varianza σ^2 . Se X_1, X_2, \dots, X_n è un campione i.i.d. estratto da X allora l'intervallo di confidenza per μ di livello $1 - \alpha$ si scrive nella seguente forma

$$\mu \in \left(\bar{X}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Lo stesso risultato vale se X è qualunque purché l'ampiezza del campione sia sufficientemente elevata