

Statistica Inferenziale

- PARAMETRICA -

NOTA: Correlazione non implica causalità

- I dati sono estratti da famiglie di distribuzioni note ma con parametri incogniti.

Ottieniamo informazioni sui parametri tramite stime risultanti dai dati

STIMATORE / STATISTICA

- È una funzione calcolabile in base al campione casuale

esempi:

- Media Campionaria $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Varianza Campionaria $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ → stimatore corretto della varianza della popolazione perché $E[s^2] = \sigma^2$
- Fornisce informazioni sui parametri reali con 2 teoremi $\begin{cases} \text{LGN} \\ \text{TLC} \end{cases}$

LEGGE DEI GRANDI NUMERI - LGN

TEOREMA: X_1, X_2, \dots, X_n v.s. indipendenti e identicamente distribuite IID

$$E(X_i) \text{ per } i=1, 2, \dots, n : \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow +\infty} E(X_i)$$

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

↓
 \bar{X} è uno stimatore corretto della popolazione perché il suo valore atteso è proprio la media della popolazione

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

TEOREMA: X_1, X_2, \dots, X_n v.s. indipendenti e identicamente distribuite

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \xrightarrow{n \rightarrow +\infty} \text{Var}(X_i)$$

TEOREMA DEL LIMITE CENTRALE

Mai più quanto velocemente gli stimatori convergono al valore del parametro

TEOREMA: X_1, X_2, \dots, X_n v.a. IID ; $E(X_i) < +\infty$, $Var(X_i) < +\infty$:

valore per
qualsiasi
distribuzione

$$P\left(\frac{\bar{X} - E(X_i)}{\frac{s(X_i)}{\sqrt{n}}} \leq x\right) \xrightarrow{n \rightarrow \infty} P(Z \leq x) \quad \text{con } Z \sim N(0,1)$$

Al crescere di n si possono ottenere i valori delle probabilità per questa variabile utilizzando la v.a. di Z ottenendo dei risultati con un'approssimazione tanto migliore quanto maggiore è n ($n \geq 30$ è già buona)

- Le probabilità al 1° membro si può approssimare con le probabilità al 2° membro sempre meglio al crescere di n

\bar{X} è circa la v.a. $N(E(x), \frac{stDev(X_i)}{\sqrt{n}})$

① $X \sim N(\mu, \sigma) \Rightarrow \sim N(E(x), \frac{stDev(X_i)}{\sqrt{n}})$

② $X \sim N(\mu, \sigma) \Rightarrow \frac{s^2(n-1)}{Var(X_i)} \sim \chi^2(n-1)$

simbolo "chi"
 gradi di libertà
 sta ad indicare che sono
 valori esatti, non approssimazioni
 (per la normale)

così particolari
 per la Normale

INTERVALLI DI CONFIDENZA

- anziché valori puntuali per le stime dei parametri si forniscono

$$\begin{aligned} [\underline{a}, \bar{b}] &\Rightarrow \text{intervallo di valori} \\ 1 - \alpha &\Rightarrow \text{livello di confidenza} \end{aligned}$$

L'intervallo di confidenza $[\underline{a}, \bar{b}]$ contiene il valore del parametro con confidenza 30% ($\alpha = 0.10$), 95% ($\alpha = 0.05$) ecc...
 $\mu = E$

LA CONFIDENZA NON È UNA PROBABILITÀ

Si parte da 10 aleatori che hanno come estremi v.a.

Per ottenere l'intervallo $[\underline{a}, \bar{b}]$:

- Trovo una quantità pivotale Q (v.a. funzione del campione e distribuzione nota)
 - Fisso il livello di confidenza estremi t_1, t_2 | $P(t_1 < Q < t_2) = 1 - \alpha$
- (Guarda qualche esempio nella risoluzione)

IC PER UNA MEDIA

IC PER UNA PROPORZIONE - prop test

p percentuale della popolazione $\Leftrightarrow \hat{p}$ percentuale campionaria

X v.a. dicotomica : $X \sim \text{Bernoulli}(p)$

$$P(X=s) = p \xrightarrow{\text{success}} ; P(X=f) = 1-p \xrightarrow{\text{fallimento}}$$

$(X_1, \dots, X_n) \quad X_i \sim \text{Bernoulli}(p) \quad \text{i.i.d.}$

$$\mathbb{E}(X_i) = \sum_{x=s,f} x_i P(X_i=x) = s \cdot P(X=s) + f \cdot P(X=f) = p = \mathbb{E}(X_i)$$

① Trovo le quantità pivotale

$$Q = \frac{1}{n} \sum_{i=1}^n X_i - p \xrightarrow{\mathbb{E}(X_i)} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \xrightarrow{\text{Var}(X_i)} \frac{\text{Var}(X)}{n}$$

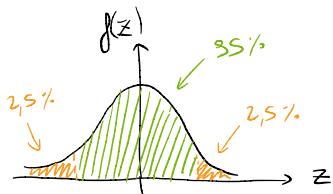
Funzione del t.p.s. $\frac{\bar{X} - \mathbb{E}(X_i)}{\sqrt{\frac{\text{Var}(X_i)}{n}}} \xrightarrow[n \rightarrow +\infty]{} N(0,1)$

Per $n > 30$ la distribuzione di Q è $N(0,1)$

- Se $n < 30$ non ha senso verificare la normalità perché il campione è di Bernoulli. L'intervallo di confidenza non sarebbe attendibile

② Trovo i quantili delle distribuzione di Q , z_1 e z_2 t.c.

$$P(z_1 < \frac{\bar{X} - p}{\sqrt{\frac{\text{Var}(X)}{n}}} < z_2) = 1 - \alpha = 0.95 \quad \text{per calcolare un per cento 95%}$$



$$z_1 = q_{\frac{\alpha}{2}} = q_{\frac{0.05}{2}} = q_{0.025}$$

$$z_2 = q_{1-\frac{\alpha}{2}} = q_{1-\frac{0.05}{2}} = q_{0.975}$$

③ Trovo la formula per a e b t.c.

$$P(A < p < B) = 0.95$$

\downarrow contiene \bar{X} \downarrow contiene \bar{X}

$$\text{IC}_{95\%}(p) = [a, b] = \left[\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

IC PER LA DIFFERENZA DI DUE MEDIE

t-test(x_1, x_2)

t-test(x_1, x_2 , paired=T)

oglie entrambi ≥ 30

- I due campioni possono essere dipendenti o indipendenti

METODO PER 2 CAMPIONI INDIPENDENTI

① Quantità piuttosto

$$\mu_1 - \mu_2 (\mathbb{E}(X_{1-}) - \mathbb{E}(X_{2-})) \text{ Parametro da stimare}$$

$$Q = \frac{X_{1-} - X_{2-} - (\mu_1 - \mu_2)}{\text{SE}(X_{1-} - X_{2-})}$$

standard error

$$\text{SE}(X_{1-} - X_{2-}) \begin{cases} \text{una forma nel caso di var uguali} \\ \text{una forma nel caso di var diverse} \end{cases}$$

1° caso: distribuzione t di Student con gradi di libertà $n_1 + n_2 - 2$

2° caso: i gradi di libertà possono non essere interi

② Non facciamo un test sulla differenza delle medie ma sulla media delle differenze

$$\begin{matrix} \checkmark (x_1 | y_1) & \rightarrow \text{minore dopo} \\ \text{minore prima} & (x_1 | y_2) \\ \dots & (x_n | y_n) \end{matrix}$$

$$\left. \begin{array}{l} d_1 = y_1 - x_1 \\ d_2 = y_2 - x_2 \\ \dots \\ d_n = y_n - x_n \end{array} \right\} \begin{array}{l} \text{campioni estratti} \\ \text{dalla variabile} \\ \text{differenze } D \\ D = Y - X \end{array}$$

Calcoliamo IC per il valore medio di D, $\mathbb{E}(D)$

Q ha distribuzione con $n-1$ gradi di libertà

IC PER LA DIFFERENZA DI PROPORZIONI - prop. test

- Utilizzo due V.A. Bernulliane

$$\begin{aligned} p_1 &: X_1 \rightarrow \hat{P}_1 \\ p_2 &: X_2 \rightarrow \hat{P}_2 \end{aligned}$$

$$\hat{P}_1 - \hat{P}_2$$

*Grande esempio
svolto in Esercizi*

QUANTITÀ PIVOTALE:

$$\frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\text{SE}(\hat{P}_1 - \hat{P}_2)}$$

per $n_1, n_2 \rightarrow +\infty$ è circa una Normale

TEST DI IPOTESI

- > Utensili metodi di controllare la verosimilità
- > Procedura statistiche che ci permettono di testare un'ipotesi
- > In statistica testiamo un'ipotesi NULLA contro una seconda che chiamiamo ALTERNATIVA

H_0 : IPOTESI NULLA
 H_1/H_A : IPOTESI ALTERNATIVA

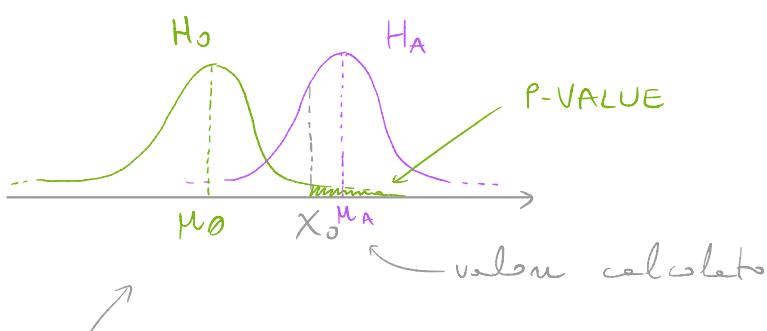
proposizioni
che riguardano
i parametri,
nel caso di
test parametrici

L'ipotesi nulla è quella ASSESSED

HP SEMPLICI : ipotesi del tipo $\mu = 10$
HP COMPOSTE : ipotesi del tipo $\mu > 10$
↳ ONE-SIDED (unilaterali) $<, >$
↳ TWO-SIDED (bilaterali) \neq

STATISTICA DEL TEST

- > Calcolo una quantità (ogni volta) sul campione (es. $\frac{1}{n} \sum_{i=1}^n x_i$)
- > Se H_0 vera il valore dovrebbe essere vicino a quello calcolato, se è vero posso anche approssimare $\frac{1}{n} \sum_{i=1}^n x_i \rightarrow N(\mu_0, \frac{\sigma^2}{n})$
- > Se H_0 falsa e fanno con H_1 abbondono H_0 in favore di H_1 .



La probabilità che H_0 sia vera è troppo estrema
Anche se H_0 viene scritta come semplice, in molti prende il range o $DX + SX$

TEST PER UNA MEDIA

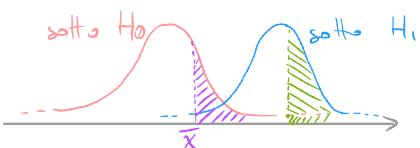
CASO HP SEMPLICE:

- > Se H_0 è vera, circa un'ipotesi che ottiene distribuzione normale e coinvolgono una statistica.
- > Se in base ai dati si ottiene qualcosa di attendibile secondo H_0 , allora l'ipotesi H_0 non è rifiutata

Indicatore: p-value

$$P\left(\frac{1}{n} \sum_{i=1}^n X_i > \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i\right) = p$$

serve il μ di H_0 statistica di test
probabilità di ottenere un valore almeno estremo quanto quello ottenuto dai dati



x : probabilità grande verso H_0

\bar{x} : probabilità piccola sotto H_0 ma grande per H_1