# STATISTICA

venerdì 18 agosto 2023 09:39

## Statistica descrittiva

analisi dei dati $x_1, \dots, x_n$ e della loro distribuzione, senza l'interpretazione di un modello probabilistico

(non ci interessa la distribuzione di prob. associata all'esperimento)

## Indici statistici

media campionaria $\rightarrow \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

mediana campionaria $\rightarrow$ ordino in senso crescente, n dispari è il dato centrale, n pari è la media dei 2 dati centrali

} Indici di centralità

varianza campionaria $\rightarrow s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ indice di dispersione

coefficiente di correlazione campionario $\rightarrow r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{\frac{1}{2}}} \in [-1,1]$ indica quanto i dati sono allineati

$|r| \simeq 0 \Rightarrow$ poco allineati

$|r| \simeq 1 \Rightarrow$ dati molto vicini a una retta.

retta di regressione campionaria $\rightarrow y = \alpha^* x + \beta^*$ con $\alpha^* = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, $\beta^* = \bar{y} - \alpha\bar{x}$

è la retta che meglio approssima i dati

## Statistica inferenziale

A partire dai dati $x_1, \dots, x_n$ ottenuto da un esperimento ripetuto ricavare info sulla distrib. di prob.

## Statistica inferenziale parametrica

Supp che la distrib. dell'esp. sia nota a meno di uno o più parametri. (Siamo interessati a determinare uno o più parametri incognita)

## Modello Statistico parametrico

È una terna $\left(S, \mathcal{J}, (Q_\theta)_{\theta\in\Theta}\right)$

- $(S, \mathcal{J})$: spazio misurabile
- $\Theta$ = insieme dei parametri $\neq \emptyset$
- $Q_\theta$ prob su $(S, \mathcal{J})$ $\forall\, \theta \in \Theta$
  rappresenta la distribuzione di un carattere di un esperimento aleatorio

**esempio**

Lancio di moneta

$\theta = p$ = prob di testa
$\Theta = [0,1]$
$\} \quad \theta \in \Theta$

$(S, \mathcal{J}) = (\{0,1\}, \mathcal{P}(\{0,1\}))$

$Q_\theta = \mathcal{B}(\theta)$ Bernoulli di parametro $\theta$

**Campione**

Sia $(S, \mathcal{J}, (Q_\theta)_{\theta \in \Theta})$ modello statistico.

Si dice campione ogni successione $(X_n)_{n \geq 1}$ di v.a. reali definite su $S$

$\exists$ (fam. di Prob) $(P_\theta)_{\theta \in \Theta}$, tc. $\forall \theta \in \Theta$ le $X_i$ sono i.i.d sotto $P_\theta$ con legge $Q_\theta$

(con distribuzione che può dipendere da $\theta$)

$(X_1, X_2, \dots, X_n)$ è detta campione di taglia $n$ e legge $Q_\theta$.

Tipicamente

Se $Q_\theta$ è la distribuzione di un esperim. aleatorio
$X_1, \dots, X_n$ rappresentano gli esiti di $n$ ripetizioni

**es**

$n$ lanci di moneta

$X_i = \begin{cases} 1 & \text{se esce T all'i-esimo lancio} \\ 0 & \text{altrimenti} \end{cases}$

$\Omega = \{0,1\}^n \qquad \mathcal{J} = \mathcal{P}(\Omega) \qquad P_\theta = Q_\theta^{\otimes n}$ cioè $P_\theta(x_1, \dots, x_n) = q_\theta(x_1) \cdots q_\theta(x_n)$

Nella statistica inferenziale osservo i risultati $x_1, \dots, x_n$ di un campione $X_1, \dots, X_n$ di legge $Q_\theta$ e voglio studiare il parametro incognito $\theta$ a partire da questi risultati

Data una fam. di prob. $(Q_\theta)_{\theta \in \Theta}$ su $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ tutte discrete o tutte ass. cont.

$\exists$ un campione $(X_1, \dots, X_n)$ di legge $Q_\theta$:

$\Omega = \mathbb{R}^n, \quad \mathcal{J} = \mathcal{B}(\mathbb{R}^n)$

$X_i(\omega_1, \dots, \omega_n) = \omega_i$

$P_\theta$ caso discreto $P_\theta(x_1, \dots x_n) = q_\theta(x_1) \cdots q_\theta(x_n)$

caso ass. cont. $f_\theta(x_1 \dots x_n) = g_\theta(x_1) \cdots g_\theta(x_n)$

caso generale $Q_\theta = P_\theta^{\otimes n}$

## Statistica campionaria

Sia $(X_n)_{n\geq 1}$ un campione.

Una successione di v.a. $(Y_n)_{n\geq 1}$ della forma $Y_n = g_n(X_1, \dots, X_n)$

dove $g_n : \mathbb{R}^n \to \mathbb{R}$ è una f. misurabile si chiama statistica campionaria.

## Stimatore

è una statistica che viene chiamata stimatore per via del ruolo speciale che ricopre.

Infatti uno stimatore è una statistica che non dipende direttamente da $\theta$.

Lo scopo di uno stimatore è stimare $h(\theta)$ con $h : \Theta \to \mathbb{R}$ f. data

### Criteri per vedere la bontà di uno stimatore

### Stimatore corretto

Uno stimatore $U$ si dice stimatore corretto di $h(\theta)$ se

1. $\forall \theta \in \Theta$ $U$ è $P^\theta$-integrabile equiv. $U_n \in L^1\left(S, \mathcal{J}, (Q_\theta)_{\theta\in\Theta}\right)$
2. $E^\theta[U] = h(\theta)$ $\forall \theta \in \Theta$

↳ $E^\theta$ è il v. atteso rispetto alla prob. $Q_\theta$

slogan "la media di $U$ su tutti i campioni possibili è $h(\theta)$

### Stimatore asintoticamente corretto

Una successione di stimatori $U_n = g_n(X_1, \dots, X_n)$ di $h(\theta)$

è asintoticamente (corretto) non distorta se $\forall n \in \mathbb{N}^+$, $\forall \theta \in \Theta$,

1. $U_n$ è $P_\theta$-integrabile
2. $\lim_n E^\theta[U_n] = h(\theta)$ $\forall \theta \in \Theta$

Mi serve indebolire la def di correttezza [cioè non $\forall n$ ma solo nel $\lim_{n\to+\infty}$]

perché non è sempre possibile determinare stimatori con le proprietà desiderate

### Stimatore consistente

Una successione di stimatori $U_n = g_n(X_1, \dots, X_n)$ di $h(\theta)$ è

consistente se:

$\lim_n P^\theta\{|U_n - h(\theta)| > \varepsilon\} = 0$ $\forall \varepsilon > 0$ $\forall \theta \in \Theta$

cioè $(U_n)_n$ converge in $P^\theta$-prob a $h(\theta)$

rischio quadratico

Dato $U$ stimatore di $R(\theta)$

$R_\theta(U) = E^\theta[(U - R(\theta))^2]$ = rischio quadratico di $U$

preferibilità

Dati $U$ e $V$ stimatori di $R(\theta)$. $U$ è preferibile a $V$ se:

$R_\theta(U) \leq R_\theta(V) \quad \forall \theta \in \Theta$ cioè $U$ è meno incerto di $V$

Oss.

Se $U$ è corretto $R_\theta(U) = Var^\theta(U)$

Stimatori di massima verosimiglianza

notazione
$Q_\theta$ discreta con den. discreta $m_\theta = P_\theta$
$Q_\theta$ ass. con " " $m_\theta = f_\theta$

funzione di verosimiglianza

$L: \Theta \times \mathbb{R}^n \longrightarrow \mathbb{R}$

$L(\theta, x_1, \dots, x_n) = L_\theta(x_1, \dots, x_n) = m_\theta(x_1) \dots m_\theta(x_n)$

caso discreto

$L_\theta(x_1, \dots, x_n) = P_\theta(X_1 = x_1) \dots P_\theta(X_n = x_n) = P_\theta(X_1 = x_1, \dots, X_n = x_n)$

caso ass. cont

$L_\theta$ è la densità congiunta di un campione i.i.d $X_1, \dots, X_n$ di legge $Q_\theta$

stimatore di massima verosimiglianza

$U$ è stimatore di massima verosimiglianza (MLE) di $\theta$ se:

$L_U(X_1, \dots, X_n) = \sup_{\theta \in \Theta} L_\theta(X_1, \dots, X_n)$

poiché $U = g(X_1, \dots, X_n)$ trovare $U$ di max ver $\longrightarrow$ $g$ t.c. $L_{g(x_1, \dots, x_n)}(x_1, \dots, x_n) = \sup_{\theta \in \Theta} L_\theta(x_1, \dots, x_n)$

Attenzione si usa se ho interv. aperto di $\mathbb{R}$ / semiretta o $\mathbb{R}$ stesso.

Per poter calcolare lo stimatore MLE si ricorre alla funzione log-verosimiglianza ottenuta attraverso l'applicazione del logaritmo naturale, quindi risulta: $l(x,\theta) = \log L(x,\theta)$
Dato che la funzione logaritmica è una trasformazione monotona crescente, con il passaggio alla log-verosimiglianza non si perdono le caratteristiche della funzione $L(x,\theta)$ in termini di crescenza e decrescenza e soprattutto si ottiene una forma analitica più semplice da trattare.

Non si usa per $U(0,\theta)$

es. log-verosimiglianza

$Q_\theta = B(\theta) \quad \theta \in [0,1]$

$m_\theta(x) = \begin{cases} 1-\theta & \text{se } x=0 \\ \theta & \text{se } x=1 \end{cases} = \theta^x (1-\theta)^{1-x} \quad x \in \{0,1\}$

$Q_\theta = B(\theta) \quad \theta \in [0,1]$

$$m_\theta(x) = \begin{cases} 1-\theta & \text{se } x=0 \\ \theta & \text{se } x=1 \end{cases} = \theta^x (1-\theta)^{1-x} \qquad x \in \{0,1\}$$

$$L_\theta(x_1,\dots,x_n) = \prod_{i=1}^n m_\theta(x_i) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{\sum_{i=1}^n (1-x_i)} = \theta^{n\bar{x}} (1-\theta)^{n(1-\bar{x})}$$

con $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$

Cerco l'MLE con log-verosim.

$$\log L_\theta(x_1,\dots,x_n) = \log\left(\theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}\right) = \log\left(\theta^{n\bar{x}}\right) + \log\left((1-\theta)^{n(1-\bar{x})}\right) = n\bar{x}\log(\theta) + n(1-\bar{x})\log(1-\theta)$$

Cerco il max della f.

$$\frac{d}{d\theta}\log L_\theta(x_1,\dots,x_n) = \frac{n\bar{x}}{\theta} + \frac{n(1-\bar{x})}{1-\theta} \qquad n\left(\bar{x} - \bar{x}\theta + \theta\right)$$

$$\frac{d}{d\theta} = 0 \iff \frac{n\bar{x}(1-\theta) + n\theta(1-\bar{x})}{\theta(1-\theta)} = \frac{n\bar{x} - n\bar{x}\theta + n\theta - n\theta\bar{x}}{\theta(1-\theta)} = 0 \iff \bar{x} = \theta$$

Inoltre $\log L_\theta(x_1,\dots,x_n) \to -\infty$ per $\theta = 0,1 \Rightarrow \exists!$ max in $\bar{x} = \theta$

$\Rightarrow \bar{X}$ è l'MLE cercato

## Modello esponenziale

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_\theta)_{\theta\in\Theta})$ $\Theta \subseteq \mathbb{R}$ intervallo tc.

discreto $\exists\ T:\mathbb{N}\to\mathbb{R}$, $g:\mathbb{N}\to\mathbb{R}$ tc. $\forall \theta\in\Theta$ $Q_\theta$ ha den. discr. su $\mathbb{N}$

$$p_\theta(x) = c_\theta\, g(x) e^{\theta T(x)} \qquad x\in\mathbb{N} \quad c_\theta > 0 \text{ cost.}$$

ass. cont. $\exists\ T:\mathbb{R}\to\mathbb{R}$, $g:\mathbb{R}\to\mathbb{R}$ boreliana tc. $\forall \theta\in\Theta$ $Q_\theta$ ha dens.

$$f_\theta(x) = c_\theta\, g(x) e^{\theta T(x)} \qquad x\in\mathbb{R} \quad c_\theta > 0 \text{ cost.}$$

## Teorema consistenza dello stim. di MLE

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_\theta)_{\theta\in\Theta})$ modello statistico tc:

1. $\forall \theta_1 \neq \theta_2 \quad Q_{\theta_1} \neq Q_{\theta_2}$ (Olibro)
2. $\Theta \subseteq \mathbb{R}$ interv. aperto (A libro)
3. $(Q_\theta)_{\theta\in\Theta}$ è modello esp. ass. cont. $\to f_\theta(x) = c_\theta\, g(x) e^{\theta T(x)}$ (B+C) libro
4. $x \mapsto g(x)T^2(x)e^{\theta T(x)}$ è integrabile $\forall \theta\in\Theta$ (E libro) $\hookrightarrow E_\theta\left[(\log p(X_1,t))^2\right] < +\infty$ f $(x_i,t)$ caso cont.
5. Sia $(X_1,\dots,X_n)$ campione iid di legge $Q_\theta$, e supp che $\exists$ Un MLE a valori in $\Theta$

$\Rightarrow$ Allora Un è unico è consistente

## lemma

$Y_n$ v. a. reali $Y_n \xrightarrow{P} c$, $\varphi:\mathbb{R}\to\mathbb{R}$ cont. $\Rightarrow \varphi(Y_n) \xrightarrow{P} \varphi(c)$

## Intervalli di fiducia

$(S, \mathcal{J}, (Q_\theta)_{\theta\in\Theta})$ modello statistico
$(X_1,\dots,X_n)$ campione iid di legge $Q_\theta$

## regione di fiducia

Dato $\alpha\in(0,1)$, una regione di fiducia a livello $1-\alpha$ però è un insieme aleatorio $D(\omega)\subseteq\Theta$, $\omega\in\Omega$
(cioè è una mappa $\Omega \to \mathcal{P}(\Theta)$, $\omega \to D(\omega)$)

Dato $\alpha \in (0,1)$, una regione di fiducia a livello $1-\alpha$ però è un insieme aleatorio $D(\omega) \subseteq \Theta$, $\omega \in \Omega$

(cioè una mappa $\Omega \longrightarrow \mathcal{P}(\Theta)$, $\omega \longrightarrow D(\omega)$)

tale che $P_\theta(\theta \in D) \geq 1-\alpha \quad \forall \theta \in \Theta$

$= \{\omega \in \Omega \mid \theta \in D(\omega)\} \subseteq \Omega$

$\{\theta \in D\} \subseteq ?$

dato un campione, posso determ D t.c. con prob alta, $\theta \in D$

Oss: $\{\theta \in D\}$ (l'aleatorietà è in D non in $\theta$ che è deterministico

$P\{\theta \in D\} \geq 1-\alpha$ va inteso come in almeno una frazione $1-\alpha$ di tutti i campioni, $\theta$ cade in D

A questo intervallo di confidenza si associa quindi un valore di probabilità cumulativa che caratterizza, indirettamente in termini di probabilità, la sua ampiezza rispetto ai valori massimi assumibili dalla variabile aleatoria.
Cioè il valore di probabilità cumulativa indica la probabilità che l'evento casuale descritto dalla variabile aleatoria cada all'interno di suddetto intervallo di confidenza, graficamente pari all'area sottesa dalla curva di distribuzione di probabilità della variabile aleatoria nell'intervallo considerato.

È bene non confondere l'intervallo di confidenza con la probabilità.
Perciò l'espressione "vi è un livello di confidenza del 95% che $\mu$ sia nell'intervallo", non indica la probabilità che $\mu$ cada nell'intervallo, in quanto $\mu$ non è una variabile aleatoria nella logica frequentista ( $\mu$ è invece interpretata come una costante non nota); bensì, indica che nel 95% dei casi in cui questa tecnica viene adottata, questa produce un intervallo che contiene il valore vero di $\mu$

- La **stima puntuale** è lo specifico valore assunto da una statistica, calcolata in corrispondenza dei dati campionari e che viene utilizzata per stimare il vero valore non noto di un parametro di una popolazione
- Uno **stimatore per intervallo** è un intervallo costruito attorno allo stimatore puntuale, in modo tale che sia nota e fissata la probabilità che il parametro appartenga all'intervallo stesso
- Tale probabilità è detta **livello di confidenza** ed è in generale indicato con $(1-\alpha)\%$ dove $\alpha$ è la probabilità che il parametro si trovi al di fuori dell'intervallo di confidenza
- Quindi la confidenza è il grado di fiducia che l'intervallo possa contenere effettivamente il parametro di interesse

## Quantili

Dato $\beta \in (0,1)$, $\beta$-quantile di $P$ (o di $F$ di $R$ $F$) è il numero $r_\beta = \inf\{x \in \mathbb{R} \mid F(x) \geq \beta\}$

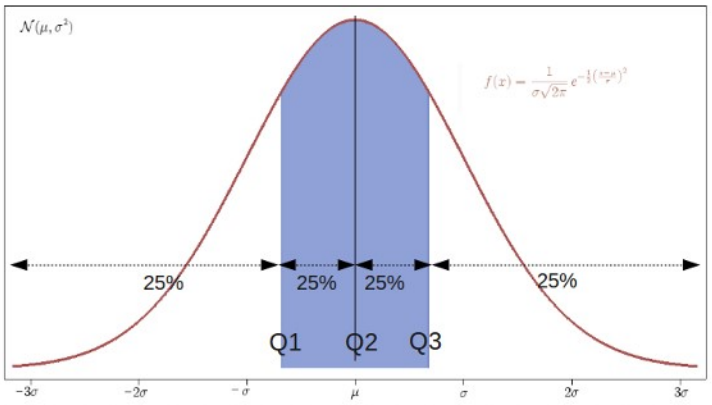
I quantili di $N(0,1)$ si denotano con $q_\beta$, $z_\beta$

Per simmetria 1 $q_{1-\beta} = -q_\beta$

2 $P\left(-q_{1-\frac{\alpha}{2}} \leq Z \leq q_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ per $Z \sim N(0,1)$, $0 < \alpha < 1$

Queste due proprietà valgono in generale se $Z$ ha densità $f$ pari, è più in generale se $Z \overset{(2)}{=} -Z$

In statistica il **quantile** di ordine $\alpha$ o *$\alpha$-quantile* (con $\alpha$ un numero reale nell'intervallo [0,1]) è un valore $q_\alpha$ che divide la popolazione in due parti, proporzionali ad $\alpha$ e *(1-$\alpha$)* e caratterizzate da valori rispettivamente minori e maggiori di $q_\alpha$. Per poter calcolare un quantile di ordine $\alpha$ è necessario che il carattere sia almeno ordinato, cioè sia possibile definire un ordinamento sui valori




Densità di probabilità di una distribuzione normale con quartili in evidenza. L'area sotto la curva rossa è la stessa negli intervalli $(-\infty,Q_1)$, $(Q_1,Q_2)$, $(Q_2,Q_3)$ e $(Q_3,+\infty)$

Nel caso di una densità di probabilità la funzione di ripartizione $F$ è continua e il quantile di ordine $\alpha$ è definito da $F(q_\alpha)=\alpha$. Questo quantile può non essere unico se la funzione di densità è nulla in un intervallo, ovvero se la funzione di ripartizione è costante ed assume il valore $\alpha$ per più di un valore $q\alpha$; ciononostante per ognuno di questi valori la distribuzione viene correttamente divisa in due parti proporzionali ad $\alpha$ e *(1-$\alpha$)*, in quanto un intervallo a densità nulla non contribuisce al calcolo della probabilità, quindi non fa differenza quale punto dell'intervallo si scelga come qα.

Nel caso di una densità discreta il quantile di ordine α è un valore $q_\alpha$ nel quale la somma delle probabilità discrete sia maggiore o uguale ad $\alpha$, ovvero tale che la somma delle probabilità *fino a* quel valore incluso sia almeno $\alpha$ e che la somma delle probabilità discrete da quel valore in poi (incluso) sia maggiore o uguale a *1-$\alpha$*. Nel caso discreto, oltre alla non unicità del quantile, si può avere una divisione della distribuzione non proporzionale ad $\alpha$ e *1-$\alpha$* (del resto una variabile discreta può essere divisa solo in un numero discreto di modi).

Il quantile di ordine α=0,1 (anche detto *primo decile*) è quel valore della distribuzione per cui la probabilità cumulata fino a qual valore, incluso, sia maggiore o uguale a 0,1, e la probabilità cumulata da quel valore, incluso, in poi sia maggiore o uguale a 0,9. Il quantile di ordine α=0,5 (la *mediana*) è quel valore della distribuzione per cui la probabilità cumulata fino a qual valore, incluso, sia maggiore o uguale a 0,5, e la probabilità cumulata da quel valore, incluso, in poi sia maggiore o uguale a 0,5.

## Modello della statistica pivotale

Data una statistica pivotale cioè $ST = g(\theta, X_1, \dots, X_n)$

1. Si è invertibile come funzione di $\theta$
2. Si ha legge che non dipende da $\theta$ cioè $P_\theta\{ST \in A\}$ non dipende da $\theta$

Allora ha una regione di fiducia di livello $1-\alpha$ per $\theta$ è data da

$D = g(\cdot, x_1, \dots, x_n)^{-1}(A)$ dove $A$ è tc. $P\{ST \in A\} = 1-\alpha$

infatti $P\{\theta \in D\} = P\{ST \in g(\cdot, X_1, \dots, X_n)(\theta)\} = 1-\alpha$

## regione di fiducia asintotica

$\alpha \in (0,1)$, una regione di fiducia asintotica di livello $1-\alpha \in (0,1)$

è una successione di insiemi aleatori $D_n(\omega) \subseteq \Theta$, $\omega \in \Omega$, $n \in \mathbb{N}$ tc.

$\liminf P_\theta\{\theta \in D_n\} \geq 1-\alpha$

## distribuzioni legate alle v.a. Gaussiane

$(\Omega, \mathcal{F}, P)$

### Proposizione

a) $Z \sim N(0,1) \Rightarrow Z^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) = \chi^2_1$

b) $Z_1, \dots, Z_n \sim N(0,1) \Rightarrow Z_1^2 + \dots + Z_n^2 \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right) = \chi^2_n$

c) $Z = (Z_1, \dots, Z_n)^T$ vettore di $N(0,1)$ indip, $M$ ortog. $\Rightarrow MZ$ vettore di $N(0,1)$ indip

d) $Z_1, \dots, Z_n \sim N(0,1)$ indip. $\Rightarrow \bar{Z}, S^2$ indip

$\bar{Z} \sim N\left(0, \frac{1}{n}\right)$, $(n-1)S_n^2 \sim \chi^2_{n-1}$

### distribuzione t - di - student

Date $Z \sim N(0,1)$, $Y \sim \chi^2_n$, $Z, Y$ indip

distribuzione t- di student a n gradi di libertà $(t_n)$ la legge di $T = \frac{\sqrt{n}\, Z}{\sqrt{Y}}$

$$f_{t_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

### corollario

$Z_1, \dots, Z_n \sim N(0,1)$ i.i.d $\Rightarrow \sqrt{n}\,\frac{\bar{Z}}{S} \sim t_{n-1}$

### Prop

$X_1, \dots, X_n$ i.i.d. $\sim N(m, \sigma^2)$ con $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$

a) $\bar{X}, S^2$ indip

b) $\bar{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$, $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

c) $\sqrt{n}\,\frac{\bar{X} - m}{S} \sim t_{n-1}$

## esempi notevoli di stimatori corretti

- La media campionaria $\bar{X} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ è un buon stimatore di $E(X_1)$

  infatti $E^*[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E^*[X_i] = E^*(X_1)$

- La varianza $S^2 = S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$ = media campionaria degli scarti quadratici

  è un buon stimatore di $Var(X_1)$

  infatti $E^*[\bar{X}^2] = \frac{1}{n^2}\sum_{j=1}^{n} E^*[X_i X_j] = \frac{1}{n^2}\sum_{i=1}^{n} E^*[X_i^2] + \frac{1}{n^2}\sum_{j\neq i}^{n} E^*[X_i]E^*[X_j] = \frac{1}{n}E^*(X^2) + \frac{n-1}{n}E^*(X_1^2)$

  $$E^*[S^2] = \frac{1}{n-1}\left(\sum_{i=1}^{n} E^*[X_i^2] - nE^*[\bar{X}^2]\right) = \frac{1}{n-1}\left(nE^*[X_i^2] - E^*[X_i^2] - (n-1)E^*[X_1]^2\right) = E^*[X_i^2] - E^*[X_1]^2 = Var^*(X_1)$$

- Il coeff. di correl. campionario $r^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)^{\frac{1}{2}}}$

  è un buon stimatore del coeff. di correl. tra le $X_i$ e $Y_i$ $\rho(X_i, X_i)$

  momento secondo $\alpha+\beta$ cioè $E^*(X_i^2)$ $\alpha+\beta$ b.stim per $E(X_1)$

  quarto $E^*(X_i^4)$ $\alpha+\beta$ " " $Var^*(X_1)$

## esempi di modelli esponenziali

leggi esponenziali $\rightarrow f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{(0,+\infty)}(x)$

$\theta = \lambda$
$T(x) = -x$
$g(x) = \mathbb{1}_{(0,+\infty)}(x)$
$c_\lambda = \lambda$

leggi di Poisson $\rightarrow p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

$\theta = \log(\lambda)$
$T(k) = k$
$g(k) = \frac{1}{k!}$
$c_\theta = e^{-\lambda}$

$\lambda^k = e^{k\log(\lambda)}$

Leggi Gaussiane $N(m, \sigma^2)$ $\sigma^2 = 1$ per semplicità $f_m(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-m)^2}$

$\theta = m$
$T(x) = 2x$
$g(x) = e^{-x^2}$
$c_m = \frac{1}{\sqrt{2\pi}} e^{-m^2}$

leggi geometriche $p_\lambda(k) = p(1-p)^{k-1} \mathbb{1}_{k\geq 1}$

$\theta = \log(1-p)$
$T(k) = k-1$
$g(k) = \mathbb{1}_{k\geq 1}$
$c_\theta = p$

## Stimatore di max verosomiglianza di media e var Gaussiana

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m,\sigma^2))_{m\in\mathbb{R}, \sigma^2\in(0,+\infty)}$ $\Theta = \mathbb{R}\times(0,+\infty)$

$L_{(m,\sigma^2)}(x_1,\dots,x_n) = f_{(m,\sigma^2)}(x_1)\cdot\dots\cdot f_{(m,\sigma^2)}(x_n) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n (x_i-m)^2\right)}$

passo al log.

$$\begin{cases} \frac{\partial}{\partial m} = 0 \\ \frac{\partial}{\partial \sigma} = 0 \end{cases}$$

e ottengo che $\left(\bar{X}, \frac{n-1}{n}\delta^2\right)$ è MLE per $(m,\sigma^2)$

se $\sigma^2$ è nota (cioè $\theta = m$) $\Rightarrow$ $\bar{X}$ è MLE per $m$
se $m$ è nota (cioè $\theta = \sigma^2$) $\Rightarrow$ $\frac{1}{n}\sum_i (X_i - m)^2$ è MLE per $\sigma^2$

# INTERVALLI DI FIDUCIA

media di una popolazione normale (varianza nota)

media di una proporzione Bernoulli (grandi campioni)

media di una popolazione, varianza non nota (grandi campioni)

media di una popolazione normale (varianza non nota)

varianza di una popolazione gaussiana

## Intervalli di fiducia per la media di una popolazione normale (varianza nota)

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathcal{N}(m,\sigma^2))_{m\in\mathbb{R}})$

Poiché $\bar{X} = \frac{1}{n}\sum X_i$ è uno stimatore di $m$ è ragionevole cercare $D$ nell'intervallo centrato in $\bar{X}$

$D = [\bar{X}-d, \bar{X}+d]$

Sappiamo che $\bar{X} \sim N(m, \frac{\sigma^2}{n})$ cioè $\frac{\sqrt{n}}{\sigma}(\bar{X}-m) \sim N(0,1)$

Calcoliamo $P_n\{ m\in[\bar{X}\pm d]\}$

$P_n(m\in[\bar{X}\pm d]) = P_n\{|\bar{X}-m|\le d\} = P_n\left(\frac{\sqrt{n}}{\sigma}|\bar{X}-m| \le \frac{d\sqrt{n}}{\sigma}\right) = P(|Z| \le \frac{\sqrt{n}}{\sigma}d) = \Phi(\frac{\sqrt{n}}{\sigma}d) - \Phi(-\frac{\sqrt{n}}{\sigma}d) = 2\Phi(\frac{\sqrt{n}}{\sigma}d) - 1$ con $Z\sim N(0,1)$

Imponiamo $P_n(m\in[\bar{X}\pm d]) = 1-\alpha$ per avere il più piccolo intervallo possibile

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

(quantile)

Quindi $\left[\bar{X} \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right]$ è int. di fiducia per $m$ di livello $1-\alpha$

Per TCL $\lim_n P\left\{-q_{1-\frac{\alpha}{2}} \le \underbrace{\frac{\sqrt{n}}{\sigma}(\bar{X}_n - m)}_{=ST} \le q_{1-\frac{\alpha}{2}}\right\} = 1-\alpha$, l'interv. di prima è asintotico

## Intervalli di fiducia per la media di una popolazione normale (varianza non nota)

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathcal{N}(m,\sigma^2))_{m\in\mathbb{R}})$

$St = \frac{\sqrt{n}}{S}(\bar{X}-m) \sim t_{n-1}$ (t-di-Student, $n-1$ gradi di libertà)

Chiamo $t_{\beta,n-1}$ il quantile di $t_{n-1}$ di ordine $\beta$ ($P\{\underbrace{T}_{t_{n-1}} \le t_{\beta,n-1}\} = \beta$)

$P\{St \in [-t_{1-\frac{\alpha}{2},n-1}, t_{1-\frac{\alpha}{2},n-1}]\} = 1-\alpha$

Ottengo $\left[\bar{X} \pm \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1}\right]$ è inter. di fiducia esatto di livello $1-\alpha$ per $m$

proprietà.

$t_{\beta,n-1} > q_\beta \quad \forall n, \forall \beta > \frac{1}{2}$ ⇒ incertezza nel caso var non nota > incertezza caso var nota

## Intervalli di fiducia asintotici per la media di una popolazione (varianza non nota), grandi campioni

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_m)_m)$, $m = E_m(X)$, $X\sim Q_m$

$P_n\{\frac{\sqrt{n}}{\cdot}(\bar{X}$ ...

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_m)_{m})$, $m = E_m(X)$, $X \sim Q_m$

Problema $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - m)$ è funzione anche di $\sigma$ non nota

Però vale il corollario

$E[X_i^2] < +\infty \Rightarrow \frac{\sqrt{n}}{S_n}(\bar{X}_n - m) \xrightarrow{\text{in legge}} Z \sim N(0,1)$

Uso $ST = \frac{\sqrt{n}}{S_n}(\bar{X}_n - m) \Rightarrow \left[\bar{X}_n \pm \frac{\sqrt{n}}{S_n} q_{1-\frac{\alpha}{2}}\right]$ è I. fiducia asintotico (di livello $\alpha$) per $m$.

## Intervalli di fiducia per la varianza di una popolazione normale

$\left(\mathbb{R}, \mathcal{B}(\mathbb{R}), (N(m, \sigma^2))_{\sigma^2 \in (0,+\infty)}\right)$

$m$ non nota

Uso $ST = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

Chiamo $\chi^2_{\beta, n-1}$ il quantile di ordine $\beta$ di $\chi^2_{n-1}$ $\left(P\left(Y \leq \chi^2_{\beta, n-1}\right) = \beta\right)$

$P\left\{ST \in \left[\chi^2_{\alpha, n-1}, +\infty\right)\right\} = 1 - \alpha$

$\left[0, \frac{(n-1)S^2}{\chi^2_{\alpha, n-1}}\right]$ è int. di fiducia unilatero per $\sigma^2$ di livello $1-\alpha$

## Intervallo di fiducia per una proporzione (media di una popol. di Bernoulli)

$\left(\mathbb{R}, \mathcal{B}(\mathbb{R}), (B(p))_{p \in [0,1]}\right)$

$p = E_p(X)$ $X \sim B(p)$

Uso $ST = \frac{n}{\bar{X}_n(1-\bar{X}_n)}(\bar{X}_n - p)$

$\left[\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} q_{1-\frac{\alpha}{2}}\right]$ è int. di fiducia

## Test Statistici

***Cosa sono i test statistici?***

Un test statistico è una procedura che consente di verificare con un elevato grado di fiducia una nostra ipotesi iniziale $H_0$, che si chiama anche *ipotesi di lavoro* o *ipotesi nulla*, inerente a un fenomeno che stiamo studiando. Il test essenzialmente è una procedura di calcolo che si basa sull'analisi dei dati numerici che abbiamo a disposizione, dati che sono interpretati come valori osservati di una certa variabile aleatoria. Tale procedura di calcolo si conclude fornendo un numero, un *consuntivo*, in inglese *test statistic*[1], che permette di confermare oppure di respingere l'ipotesi di lavoro, in favore di un'altra ipotesi iniziale $H_1$, detta *ipotesi alternativa*. Essendo la procedura aleatoria, possono comunque presentarsi quattro casi diversi:

| | $H_0$ è vera | $H_0$ è falsa |
|---|---|---|
| Accetto $H_0$ | $H_0$ è vera ed accetto $H_0$ | $H_0$ è falsa ma accetto $H_0$ |
| Rifiuto $H_0$ | $H_0$ è vera ma rifiuto $H_0$ | $H_0$ è falsa e rifiuto $H_0$ |

Le celle evidenziate con lo sfondo colorato rappresentano i casi in cui commettiamo degli errori: commettiamo un *errore di primo tipo* quando l'ipotesi $H_0$ è vera ma noi la rifiutiamo. È possibile andare a valutare la probabilità di commettere tale errore: tale probabilità di commetto si indica con $\alpha$, ed essa viene chiamata *livello di significatività*[2] del test. Esiste anche una probabilità $\beta$ di commettere un *errore di secondo tipo*, ossia l'errore di accettare l'ipotesi $H_0$ nonostante essa sia falsa. In linea di principio, uno sperimentatore desidera abbassare il livello $\alpha$ il più possibile verso zero; sfortunatamente, si può dimostrare che questo comporta il fatto che ciò aumenterà di conseguenza la probabilità $\beta$, e di riflesso questa diminuirà (il numero $1-\beta$, che viene definito con il termine di *potenza del test* (ossia, la probabilità di rifiutare l'ipotesi di lavoro giacché essa è falsa). Rifacendoci a quanto avevamo detto in [2] a proposito dei test diagnostici, la potenza di un test è legata alla sensibilità ed ai valori predittivi positivi, mentre la quantità $\beta$ si collega alla specificità ed ai valori predittivi negativi.
Per illustrare meglio la situazione, possiamo osservare questi due schemi (avendo in mente un processo indiziario, oppure un esame diagnostico):

| | innocente | colpevole |
|---|---|---|
| assoluzione | innocente assolto | colpevole assolto |
| condanna | innocente condannato | colpevole condannato |

| | sano | malato |
|---|---|---|
| negativo | vero negativo | falso negativo |
| positivo | falso positivo | vero positivo |

Quello che si vorrebbe realizzare è di non commettere l'errore di primo tipo ("condannare un innocente", "curare un sano"); inoltre si vuole anche abbassare la possibilità di commettere l'errore di secondo tipo ("assolvere il colpevole", "non curare un malato") allo scopo di massimizzare la potenza del test ("condannare il colpevole", "curare un malato"). Si noti però che c'è una sorta di "asimmetria" nei due errori. Tra i molti test statistici esistenti, il più celebre in ambito medico sicuramente è il *test t di Student*. Ci riproponiamo di trattare in una prossima dispensa anche un altro importante test, il test $\chi^2$ di indipendenza di Karl Pearson.

## Test statistici

Un test statistico è una procedura per verificare, sulla base dei dati del campione se una certa hp. (affermazione sul parametro) è plausibile o no

$(S, \mathcal{J}, (Q_\theta)_{\theta \in \Theta})$ modello statistico, $(X_1, \dots, X_n)$ iid di legge $Q_\theta$, def. su $(\Omega, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$

## Elt principali di un test (parametrico)

1 L'hp da verificare (di carattere generale): $\Theta = \Theta_0 \cup \Theta_1$ partizione di $\Theta$ $(\Theta_0 \cap \Theta_1 = \emptyset)$

- $H_0: \theta \in \Theta_0$ ipotesi nulla, di cui verificare se è plausibile o no
- $H_1: \theta \in \Theta_1$ ipotesi alternativa

Osservazione

- Se l'esito del test, sulla base dei dati, è il rifiuto di $H_0$, in favore di $H_1$, allora $H_0$ è poco plausibile, cioè c'è evidenza statistica per $H_1$
- Se l'esito del test è l'accettazione di $H_0 \Rightarrow H_0$ è plausibile, ma non c'è per forza evidenza per $H_0$

2 La regione critica o di rifiuto

Un evento $C \in \mathcal{J}$, solit. dipendente da $(X_1, \dots, X_n)$ tc:

se l'esito $\omega$ cade in $C$ rifiutiamo l'hp $H_0$ in favore di $H_1$

se l'esito $\omega$ non cade in $C$, accettiamo l'hp $H_0$

La scelta di $C$ deve essere tale che: $C$ sia poco prob. sotto $H_0$
se $\omega \in C$ allora questo è un'evidenza per $H_1$

### esempio

n lanci di moneta con $p$ = prob testa
Il banco dichiara che $p = \frac{1}{2}$
Sulla base degli esiti degli n lanci, vogliamo verificare se l'hp $p = \frac{1}{2}$ è plausibile o no

$H_0: p = \frac{1}{2}$
mi aspetto una regione critica della forma $C = \{|\bar{X} - \frac{1}{2}| > d\}$ per un $d$ opportuno, per cui $P_{\frac{1}{2}}(C)$ sia piccola

$H_0: p \leq \frac{1}{2}$

$C = \{\bar{X} - \frac{1}{2} > d\}$ per cui $P_p(C)$ sia piccola $\forall p \leq \frac{1}{2}$

Spesso per $C$ si usa una statistica $ST$ la cui legge non dipende da $\vartheta$

## Errori

errore di 1ª specie: $H_0$ vera ma l'esito del test è il rifiuto detto

errore di 2ª specie: $H_0$ falsa ma l'esito del test è l'accettazione di $H_0$

## Prob di commettere questi errori

$\alpha \in (0,1)$. Un test è di livello $\alpha$ se $\sup_{\theta \in \Theta_0} P_\theta(C) \leq \alpha$

cioè la prob. dell'errore di 1ª specie è al massimo $\alpha$
Più $\alpha$ è piccolo, maggiore è l'evidenza statistica formata dal test contro $H_0$ in caso di rifiuto

## potenza di un test

$\pi_C: \Theta_1 \longrightarrow [0,1]$ $\quad \pi_C(\theta) = P_\theta(C)$

$\pi_C : \Theta_1 \longrightarrow [0,1]$ $\quad \pi_C(\theta) = P_\theta(C)$

↓

prob. di rifiutare $H_0$ quando il v. del parametro è $\theta \in \Theta_1$

$\pi_C(\theta)$ rappresenta la capacità di accorgersi che l'ip $H_0$ è falsa

Più $\pi_C(\theta)$ è piccolo più commetti errore di 2ª specie
più $\alpha$ è piccolo più " " 1ª specie

Approccio classico.

Fisso $\alpha$ piccolo e cerco $C$ in modo da massimizzare la potenza

## Legame tra test e intervalli di fiducia

caso ip semplice $\Theta_0 = \{\theta_0\}$ e $\Theta_1 = \Theta \setminus \{\theta_0\}$

- dato $\alpha \in (0,1)$ se $D$ è una regione di fiducia di livello $1-\alpha$ per $\theta \Rightarrow$

$C = [\theta_0 \notin D] = \{\omega \in \Omega \mid \theta_0 \notin D(\omega)\}$ è una regione critica di livello $\alpha$

perché

$P_{\theta_0}(C) = 1 - P_{\theta_0}[\theta_0 \in D] \leq 1 - (1-\alpha) = \alpha$

Viceversa

- dato $\alpha \in (0,1)$ se $\forall \theta_0 \in \Theta$, $C_{\theta_0}$ è una regione critica di livello $\alpha$ per il Test $H_0 : \theta = \theta_0 \Rightarrow$

$D(\omega) = \{\theta \in \Theta \mid \omega \notin C_\theta\}$ è una regione di fiducia per $\theta$ di livello $1-\alpha$

perché

$P_{\theta_0}\{\theta_0 \in D\} = P_{\theta_0}\{C_{\theta_0}^c\} = 1 - P_{\theta_0}(C_{\theta_0}) \geq 1-\alpha$

## Metodo del p-value / soglia di accettazione

Abbiamo una famiglia di regioni critiche $(C_\alpha)_{\alpha \in (0,1)}$

$C_\alpha$ regione critica a livello $\alpha$ $\forall \alpha$

se $\alpha < \alpha' \Rightarrow C_\alpha \subset C_{\alpha'}$

$\bigcap_\alpha C_\alpha = \emptyset \quad \bigcup_\alpha C_\alpha = \Omega$

Dato $\omega \in \Omega$ esiste $\exists\, \bar{\alpha} = \bar{\alpha}(\omega) \in (0,1]$ tc. ← p-value di $\omega$

- se $\alpha < \bar{\alpha} \Rightarrow \omega \notin C_\alpha \Rightarrow$ accetto $H_0$
- se $\alpha \geq \bar{\alpha} \Rightarrow \omega \in C_\alpha \Rightarrow$ rifiuto $H_0$

$\bar{\alpha}(\omega)$ fornisce il grado di plausibilità di $H_0$ se si verifica $\omega$

più $\bar{\alpha}$ è basso meno plausibile è $H_0$, minore è il livello $\alpha$ per cui rifiuto

Quando si effettua un test d'ipotesi si fissa un'ipotesi nulla e un valore soglia α (per convenzione di solito 0,05) che indica il livello di significatività del test. Calcolato il *p-value* relativo ai dati osservati è possibile comportarsi come segue:

- se valore $p > \alpha$ l'evidenza empirica non è sufficientemente contraria all'ipotesi nulla che quindi non può essere rifiutata;
- se valore $p \leq \alpha$ l'evidenza empirica è fortemente contraria all'ipotesi nulla che quindi va rifiutata. In tal caso si dice che i dati osservati sono statisticamente significativi.

- se valore $p \leq \alpha$ l'evidenza empirica è fortemente contraria all'ipotesi nulla che quindi va rifiutata. In tal caso si dice che i dati osservati sono statisticamente significativi.

caso Hp composta

$\Theta_0 = (-\infty, \theta_0]$ $\quad H_0 : \theta \leq \theta_0$

$\Theta_1 = (\theta_0, +\infty)$ $\quad H_1 : \theta > \theta_0$

} o viceversa

def

Dati $(S, \mathcal{I}, (Q_\theta)_{\theta \in \Theta})$ modello statistico con $\Theta \subseteq \mathbb{R}$ intervallo

$(X_1, \dots, X_n)$ campione i.i.d. di legge $Q_\theta$ def su $(\Omega, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$

$T : \Omega \to \mathbb{R}$ v.a.

diciamo che il mo dello è a rapporto di verosomiglianza crescente rispetto a T

se $\forall \theta_1 < \theta_2 \ \exists \ f_{\theta_1, \theta_2} : \mathbb{R} \to (0, +\infty)$ strett. cresc. t.c.

$$\frac{L(\theta_2, x_1, \dots, x_n)}{L(\theta_1, x_1, \dots, x_n)} = f_{\theta_1, \theta_2}(T)$$

Teorema

Nelle Hp sopra Considero il Test $H_0 : \theta \leq \theta_0$ contro $H_1 : \theta > \theta_0$

sia $C = \{T \geq d\}$

Allora

a) $\sup_{\theta \leq \theta_0} P_\theta(C) = P_{\theta_0}(C)$

b) il Test di regione critica C è il più potente tra i test di livello $P_{\theta_0}(C)$

Lemma

$\Theta = \{\theta_0, \theta_1\}$ con $H_0 : \theta = \theta_0$ e $H_1 : \theta = \theta_1$,

$C = \{ L_{\theta_0}(x_1, \dots, x_n) \leq c \, L_{\theta_1}(x_1, \dots, x_n) \}$ $\quad c > 0$

Allora

a) il Test di regione critica C è il più potente tra i test di livello $P_{\theta_0}(C)$

b) $P_{\theta_0}(C) \leq P_{\theta_1}(C)$

TEST Z

Test Bilatero per la media di una popol. gaussiana (var nota)

Dati ↓

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m, \sigma^2)_{m \in \mathbb{R}}), \Theta = m \in \Theta = \mathbb{R}$, $\sigma^2$ nota $(X_1, \dots, X_n)$ campione iid. di $\mathcal{N}(m, \sigma^2)$

Test ↓

dato $m_0 \in \mathbb{R}$, considero il Test $H_0 : m = m_0$ vs. $H_1 : m \neq m_0$, a livello $\alpha$

Regione critica C ↓

Poiché $\bar{X}$ è uno stimatore di $m$ e $H_1 : m \neq m_0$, cerco C della forma

$C = (\bar{X} \dots$

TEST Z

Test unilatero per la media di una popol. gaussiana (var nota)

Dati ↓

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m, \sigma^2)_{m \in \mathbb{R}}), \Theta = m \in \Theta = \mathbb{R}$, $\sigma^2$ nota $(X_1, \dots, X_n)$ campione iid. di $\mathcal{N}(m, \sigma^2)$

Test ↓

dato $m_0 \in \mathbb{R}$, considero il Test $H_0 : m \leq m_0$ vs. $H_1 : m > m_0$, a livello $\alpha$

Regione critica C ↓

Per il Teorema $C = \{\bar{X} > d\}$ è ottimale & opportuno

Poiché $\bar{X}$ è uno stimatore di $m$ e $H_1: m \neq m_0$, cerco $C$ della forma

$C = \{ |\bar{X} - m| > d \}$ $d$ opportuno

**Procedura ↓**

Uso $ST = \frac{\sqrt{n}}{\sigma}(\bar{x} - m) \sim N(0,1)$ sotto $P_m$

Calcolo $P_{m_0}(C) = P_{m_0}\{ \frac{\sqrt{n}}{\sigma}|\bar{x} - m| > \frac{\sqrt{n}}{\sigma} d \} = P\{ |Z| > \frac{\sqrt{n}}{\sigma} d \}$

Impongo $P_{m_0}(C) = \alpha$ cioè $\frac{\sqrt{n}}{\sigma} d = q_{1-\frac{\alpha}{2}}$

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

↳ regione critica di livello $1-\alpha$

Per il Teorema $C = \{ \bar{X} > d \}$ è ottimale $d$ opportuno

**Procedura ↓**

Uso $ST = \frac{\sqrt{n}}{\sigma}(\bar{x} - m) \sim N(0,1)$ sotto $P_m$ uso statistica di Test.

Impongo $\sup_{m \leq m_0} P_m(\bar{X} > d) \overset{\text{Teorema}}{=} P_{m_0}(\bar{X} > d) = \alpha$

$\alpha = P_{m_0}(\bar{X} > d) = P_{m_0}\{ \frac{\sqrt{n}}{\sigma}(\bar{x} - m_0) > \frac{\sqrt{n}}{\sigma} d \} = P(z > \frac{\sqrt{n}}{\sigma} d)$

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

$C = \{ \frac{\sqrt{n}}{\sigma}(\bar{x} - m_0) > q_{1-\alpha} \}$

**p-value**

$\bar{\alpha} = P\{ z > \frac{\sqrt{n}}{\sigma}(\bar{x} - m) \}$

**potenza**

$P_m(C) = P_m\left[ \frac{\sqrt{n}}{\sigma}(\bar{x} - m) > q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}(m - m_0) \right] =$

$= P\{ z > q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}(m - m_0) \} = 1 - \Phi\left( q_{1-\alpha} - \frac{\sqrt{n}}{\sigma}(m - m_0) \right)$

## Test Bilatero per la media di una popol. Bernoulli, grandi camp

**Dati ↓**

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (B(p))_{p \in (0,1)})$, $\theta = p \in \Theta = (0,1)$, $(X_1, \dots, X_n)$ campione iid. di $B(p)$

**Test ↓**

dato $p_0 \in (0,1)$, considero il Test $H_0: p = p_0$ vs. $H_1: p \neq p_0$, a livello $\alpha$

**Regione critica $C$ ↓**

$C = \{ |\bar{X} - p_0| > d \}$ $d$ opportuno

**Procedura ↓**

Uso $ST = \sqrt{\frac{n}{p(1-p)}}(\bar{X} - p) \overset{TCL}{\sim} N(0,1)$ sotto $P_n$

Impongo $P_{n_0}(C) = \alpha$ cioè $\frac{\sqrt{n}}{\sigma} d = q_{1-\frac{\alpha}{2}}$

Ottengo $C = \left\{ \sqrt{\frac{n}{p_0(1-p_0)}} |\bar{X} - p_0| > q_{1-\frac{\alpha}{2}} \right\} = \left\{ \frac{n\bar{X} - np_0}{\sqrt{np_0(1-p_0)}} > q_{1-\frac{\alpha}{2}} \right\}$

↳ regione critica di livello $\alpha$

## Test unilatero per la media di una popol. gaussiana (var nota)

**Dati ↓**

$(\mathbb{R}, \mathcal{B}(\mathbb{R}), (B(p))_{p \in (0,1)})$, $\theta = p \in (0,1]$ $(X_1, \dots, X_n)$ campione iid.

**Test ↓**

dato $p_0 \in \mathbb{R}$, considero il Test $H_0: p \leq p_0$ vs. $H_1: p > p_0$, a livello $\alpha$

**Regione critica $C$ ↓**

Per il Teorema $C = \{ \bar{X} > d \}$ è ottimale $d$ opportuno

**Procedura ↓**

Uso $ST = \sqrt{\frac{n}{p(1-p)}}(\bar{x} - p) \overset{TCL}{\sim} N(0,1)$ uso statistica di Test.

ottengo come sopra

$C = \left\{ \sqrt{\frac{n}{p_0(1-p_0)}}(\bar{x} - p_0) > q_{1-\alpha} \right\}$ livello $\alpha$

**p-value**

$\bar{\alpha} = P\left\{ z > \sqrt{\frac{n}{p_0(1-p_0)}}(\bar{x} - p_0) \right\}$

In un test unilaterale la zona di rifiuto dell'ipotesi nulla è solamente in una coda della distribuzione; in un test bilaterale essa è equamente divisa nelle due code della distribuzione.