20
domenica 30 aprile 2023 15:18

Statistica descrittiva

Statistica descrittiva: analisi dei dati $x_1, \dots x_n$ (generalmente ottenuti da ripetizioni di un esperimento) e della loro distribuzione, senza l'interpretazione di un modello probabilistico (cioè non ci interessa la distribuzione di prob. associata all'esperimento)

Ad es: lanciamo 10 volte una moneta, studiamo la distrib del n° di teste tra i 10 lanci
considerimo 20 individui, studiamo la distribuzione dell'altezza tra questi 20 individui

Supponiamo qui che i dati siano quantitativi, cioè $x_i \in \mathbb{R}$ (o $\mathbb{R}^d$)

Alcuni indici statistici (funzioni di $x_1 \dots x_n$) rilevanti

- media campionaria o empirica
  $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  indice di centralità (indica il centro della distrib)
- mediana campionaria o empirica:
  ordinati i dati in senso crescente, il dato centrale se n è dispari, la media dei due dati centrali se n è pari.
  indice di centralità
- varianza campionaria
  $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
  indice di dispersione (indica la dispersione dei dati)

Oss: · la media campionaria è il valore atteso di una v.a. con distrib. uniforme sugli x_i
· $\frac{n-1}{n} s^2$ è la varianza di una v.a con distrib uniforme sugli x_i

Per coppie di dati $(x_1, y_1), \dots (x_n, y_n)$, ci interessa la distribuzione congiunta degli (x_i, y_i)

Ad es: dati 20 individui, x_i = altezza, y_i = peso dell' i-simo individuo

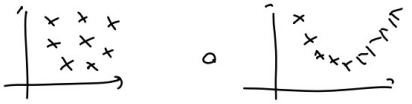
Alcuni indici statistici rilevanti:

- coefficiente di correlazione campionaria

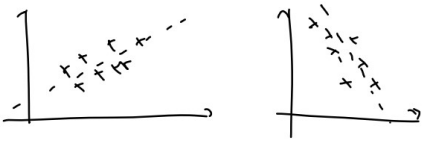
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2} \left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2}} \in [-1, 1]$$

indice quanto i dati sono allineati

- $|r| \approx 0$: i dati sono poco allineati



- $|r| \approx 1$: i dati sono molto vicini ad una retta



- retta di regressione campionaria

$y = \alpha^* x + \beta^*$ con

$$\alpha^* = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \beta^* = \bar{y} - \alpha \bar{x}$$

è la retta che meglio approssima i dati

21

domenica 30 aprile 2023 17:05

Teoria degli stimatori

Statistica inferenziale: a partire dai dati $x_1, \dots x_n$ ottenuti da un esperimento ripetuto, ricavare informazioni sulla distribuzione di probabilità associata a quell'esperimento

Esempio: lanciamo una moneta n volte, ottenendo risultati $x_1, \dots x_n$, $\left(x_i = \begin{cases} 1 & \text{se esce testa all'i-simo lancio} \\ 0 & \text{altrimenti} \end{cases}\right)$

vogliamo avere info sulla prob p di testa (non nota):

1) stima per p? → stimatore ($\overline{x}$ = freq relativa di "teste")

2) intervallo "ragionevole" per p? → intervalli di fiducia

3) dato $p_0 \in (0,1)$, l'ipotesi $p = p_0$ è "ragionevole" (sulla base dei dati)? → test di ipotesi

Altri esempi:

- date le altezze di n persone (adulte di sesso maschile in Italia), avere info sulla distribuzione delle altezze su tutta la popolazione (persone adulte di sesso maschile in Italia)
- dati i risultati di un sondaggio sul gradimento del governo tra n persone, avere info sul gradimento del governo tra tutta la popol.
- dati gli esiti di un certo farmaco su n persone, avere info sulla distrib. degli esiti del farmaco su un individuo generico

Qui ci occupiamo di statistica (inferenziale) parametrica: supponiamo che la distrib. dell'esperimento sia nota a meno di una o più parametri (o comunque siamo interessati a determinare uno o più parametri incogniti).

(parametrica)

Def: Chiamiamo modello statistico una terna $(S, \mathcal{S}, (Q_\theta)_{\theta \in \Theta})$ con

- $(S, \mathcal{S})$ spazio misurabile
- Θ insieme dei parametri ($\neq \emptyset$)
- Q_θ prob su $(S, \mathcal{S})$, $\forall \theta \in \Theta$

Q_θ rappresenta la distribuzione di un (carattere di) un esperimento aleatorio.

Esempi:

- lancio di moneta: θ = p.prob. di testa $\in \Theta = [0,1]$

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

  $Q_\theta = B(\theta)$ Bernoulli di parametro θ

- carica elettrica di un corpo, supponiamo di distrib. gaussiana di media e varianza non note

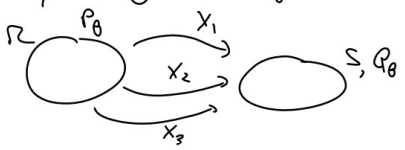
  $\theta = (m, \sigma^2) \in \mathbb{R} \times (0, +\infty) = \Theta$

  $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$

  $Q_{(m,\sigma^2)} = N(m, \sigma^2)$

Def: Un campione (i.i.d.) di taglia n e legge Q_θ è una famiglia $X_1 \dots X_n$ di v.a. i.i.d., a valori in S, di legge Q_θ, "al variare di $\theta \in \Theta$".

Più precisamente, un campione è una famiglia $X_1 \dots X_n$ con $X_i : \Omega \to S$ v.a., $(\Omega, \mathcal{F})$ spazio mis., tali che esista una famiglia di prob. $(P_\theta)_{\theta \in \Theta}$ su $(\Omega, \mathcal{F})$ tali che, $\forall \theta \in \Theta$: sotto P_θ $X_1 \dots X_n$ siano i.i.d. di legge Q_θ



Tipicamente, se Q_θ è la distribuzione di un esperimento aleatorio, $X_1 \dots X_n$ rappresentano gli esiti di n ripetizioni dell'esperimento.

Esempi:

- n lanci di moneta
  $X_i = \begin{cases} 1 & \text{se esce testa all'i-simo lancio} \\ 0 & \text{altrimenti} \end{cases}$
  $\Omega = \{0,1\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$, $P_\theta = Q_\theta^{\otimes n}$, cioè $p_\theta(x_1 \dots x_n) = q_\theta(x_1) \cdot \dots \cdot q_\theta(x_n)$ (con p_θ, q_θ densità discrete di P_θ, Q_θ risp.)

- n misurazioni della carica elettrica di un corpo
  X_i = esito della i-sima misurazione $X_i(\omega) = \omega_i$
  $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, P_θ misura con densità $f_\theta(x_1 \dots x_n) = g_\theta(x_1) \cdot \dots \cdot g_\theta(x_n)$
  (con $g_\theta = g_{(m,\sigma^2)}$ densità $\mathcal{N}(m, \sigma^2)$)

- dato un carattere di una popolazione, con distrib. Q_θ (ad es., n° di figli di un italiano)
  X_i = carattere per l'i-simo individuo estratto della popolazione (n° di figli dell'i-simo individuo)

[spesso si identificano una popolazione e il suo carattere, quando ci interessa solo tale carattere]

Oss: Data una famiglia di prob $(Q_\theta)_{\theta \in \Theta}$ su $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, tutte discrete o tutte assolutamente continue, esiste un campione $(X_1 \dots X_n)$ di legge Q_θ:

- $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$
- $X_i(\omega_1 \dots \omega_n) = \omega_i$ $\quad i = 1 \dots n$
- P_θ: nel caso discreto: avente densità discreta $p_\theta(x_1 \dots x_n) = q_\theta(x_1) \cdot \dots \cdot q_\theta(x_n)$
  con q_θ densità discrete di Q_θ
  nel caso assol. cont: avente densità $f_\theta(x_1 \dots x_n) = g_\theta(x_1) \cdot \dots \cdot g_\theta(x_n)$
  con g_θ densità di Q_θ

Nel caso generale, $Q_\theta = P_\theta^{\otimes n}$ è la prob prodotto

Nella statistica inferenziale, osserviamo i risultati $x_1 \dots x_n$ di un campione $X_1 \dots X_n$ di legge Q_θ e vogliamo stimare il parametro incognito θ a partire da questi risultati.

Def: Chiamiamo statistica una v.a. funzione del campione $(X_1 .. X_n)$: $g(X_1, .. X_n)$

Def: Chiamiamo stimatore una statistica che non dipende direttamente dal parametro θ.
Scopo di uno stimatore è stimare $h(\theta)$, con $h: \Theta \to \mathbb{R}$ funzione data.

Esempi notevoli:

- media campionaria ($X_1, .. X_n$ campione i.i.d. di v.a reali)
$\bar{X} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
come vedremo, è un "buon" stimatore del valore atteso $E[X_1]$

- varianta campionaria ($X_1 . X_n$ campione i.i.d. di v.a. reali)
$S^2 = S^2_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n \bar{X}^2 \right)$ (≈media campionaria degli scarti quadratici da $\bar{X}$)
come vedremo, è un "buon" stimatore della varianta $Var(X_1)$

- dato un campione bivariato $(X_1, Y_1), .. (X_n, Y_n)$, ($X_i, Y_i$ v.a. reali, (X_i, Y_i) i.i.d. sotto P_θ)
coefficiente di correlazione campionario
$$r^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \cdot \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right)^{1/2}}$$
è un "buon" stimatore del coeff. di correlazione tra X_i e Y_i $\rho(X_1, Y_1)$

Oss: In una sequenza di Bernoulli di n ripetizioni di esperimento successo/insuccesso, con
p prob (incognita) di successo ($X_i = \begin{cases} 1 & \text{se successo all'i-sima prova} \\ 0 & \text{altrimenti} \end{cases}$)
$\bar{X}$ = freq. relativa del successo sul campione
$E[X_1] = p$ = prob del successo

Criteri per valutare la bontà di uno o più stimatori:

Def: Uno stimatore U è uno stimatore corretto, o non distorto (unbiased), di $h(\theta)$: se, $\forall \theta \in \Theta$,
U è P^θ-integrabile e vale
$$E^\theta[U] = h(\theta) \quad \forall \theta \in \Theta$$
"la media di U su tutti i campioni possibili è il valore $h(\theta)$"

Esempi notevoli:

- Se $X_1 .. X_n$ è campione i.i.d. di v.a. reali con $E^\theta[|X_1|] < \infty \ \forall \theta \in \Theta$, $(h(\theta) = E^\theta[X_1])$
$\bar{X}$ è uno stimatore corretto di $E^\theta[X_1]$: infatti
$$E^\theta[\bar{X}] = \frac{1}{n} \sum_{i=1}^{n} E^\theta[X_i] = E^\theta[X_1]$$

- Se $X_1 .. X_n$ è campione i.i.d. di v.a. reali con $E^\theta[X_1^2] < \infty \ \forall \theta \in \Theta$, $(h(\theta) = Var^\theta(X_1))$,
S^2 è uno stimatore corretto di $Var^\theta(X_1)$: infatti

$$E^\theta[\bar X^2] = \frac{1}{n^2}\sum_{i,j=1}^n E^\theta[X_iX_j] = \frac{1}{n^2}\sum_{i=1}^n E^\theta[X_i^2] + \frac{1}{n^2}\sum_{i\neq j} E^\theta[X_i]E^\theta[X_j]$$
$$= \frac{1}{n}E^\theta[X_1^2] + \frac{n-1}{n}E^\theta[X_1]^2$$

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

Sia $(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ un modello e, $\forall n\in\mathbb{N}$, sia $X_1,..X_n$ un campione i.i.d. di taglia n e di legge Q_θ (è possibile def. tutte le X_i sullo stesso modello $(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$)

Def: Una successione di stimatori $U_n = g_n(X_1,..X_n)$, $n\in\mathbb{N}^+$, di $h(\theta)$ è asintoticamente non distorta se, $\forall n\in\mathbb{N}^+$, $\forall\theta\in\Theta$, U_n è P_θ-integrabile e

$$\lim_n E^\theta[U_n] = h(\theta) \quad \forall\theta\in\Theta$$

Def: Una successione di stimatori $U_n = g_n(X_1,..X_n)$, $n\in\mathbb{N}^+$, di $h(\theta)$ è consistente: se

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

cioè $(U_n)_n$ converge in P^θ-prob. a $h(\theta)$, $\forall\theta\in\Theta$

"per n grande, U_n è vicino con alta prob. a $h(\theta)$"

Esempi notevoli:

- Se $X_1,..X_n,...$ è un campione i.i.d. di v.a. reali con $E^\theta[X_1^2]<\infty$ $\forall\theta\in\Theta$, $(\bar X_n)_{n\in\mathbb{N}}$ è una successione di stimatori consistenti di $E^\theta[X_1]$:
  infatti per LGN, $\bar X_n \xrightarrow{P_\theta} E^\theta[X_1]$ $\forall\theta\in\Theta$
- Se $X_1,..X_n,...$ è un campione i.i.d. di v.a. reali con $E^\theta[X_1^4]<\infty$ $\forall\theta\in\Theta$, $(S_{n-1}^2)_{n\in\mathbb{N}}$ è una successione di stimatori consistenti di $\mathrm{Var}^\theta(X_1)$
  infatti per LGN,
  $$S_{n-1}^2 = \underbrace{\frac{n}{n-1}}_{\to 1}\left(\underbrace{\frac{1}{n}\sum_{i=1}^n X_i^2}_{\xrightarrow{P_\theta} E[X_1^2]} - \underbrace{\bar X_n^2}_{\xrightarrow{P_\theta} E[X_1]}\right) \xrightarrow{P_\theta} E^\theta[X_1^2] - E^\theta[X_1]^2 = \mathrm{Var}(X_1)$$
  (dove abbiamo usato che la convergenza in prob. è stabile per somma e prodotto)

Def: Dato U stimatore di $h(\theta)$, rischio quadratico di U

$$R_\theta(U) = E^\theta[(U - h(\theta))^2]$$

Dati U e V stimatori di $h(\theta)$, diciamo che U è preferibile a V: se $R_\theta(U)\le R_\theta(V)$ $\forall\theta\in\Theta$

"U è meno incerto di V"

Oss: Se U è corretto, $R_\theta(U) = \mathrm{Var}^\theta(U)$

Oss: $R_\theta(\bar X_n) = \mathrm{Var}^\theta(\bar X_n) = \frac{1}{n}\mathrm{Var}^\theta(X_1) \searrow 0$ per $n\to\infty$.

22
mercoledì 3 maggio 2023 19:12

Stimatori di massima verosimiglianza

Sia $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_\theta)_{\theta\in\Theta})$ un modello statistico
- discreto: Q_θ discreta con densità discreta $m_\theta = p_\theta$, $\forall \theta\in\Theta$
- oppure assolutamente continuo: Q_θ assol continua con densità $m_\theta = f_\theta$, $\forall\theta\in\Theta$

Def: Funzione di verosimiglianza: $L: \Theta\times\mathbb{R}^n \to \mathbb{R}$ data da
$$L(\theta, x_1 \dots x_n) = L_\theta(x_1 \dots x_n) = m_\theta(x_1)\cdot\dots\cdot m_\theta(x_n)$$

Oss: • Nel caso discreto, L_θ è la densità discreta congiunta di un campione $\overset{i.i.d.}{\forall} X_1 \dots X_n$ di legge Q_θ
$$L_\theta(x_1 \dots x_n) = P_\theta\{X_1 = x_1\}\cdot\dots\cdot P_\theta\{X_n = x_n\} = P_\theta\{X_1 = x_1, \dots X_n = x_n\}$$
• Nel caso assol. cont., L_θ è la densità congiunta di un campione $\overset{i.i.d.}{\forall} X_1 \dots X_n$ di legge Q_θ

Sia $(X_1 \dots X_n)$ un campione i.i.d. di legge Q_θ — maximum likelihood estimator

Def: Uno stimatore U è detto stimatore di massima verosimiglianza di θ (MLE in inglese): se
$$L_U(X_1 \dots X_n) = \sup_{\theta\in\Theta} L_\theta(X_1 \dots X_n)$$
(cioè $L_{U(\omega)}(X_1(\omega), \dots X_n(\omega)) = \sup_{\theta\in\Theta} L_\theta(X_1(\omega), \dots X_n(\omega))$ $\forall\omega\in\Omega$, si assume implicitamente $U(\omega)\in\Theta$)

Poiché $U = g(X_1 \dots X_n)$, trovare U di max. ver. equivale a trovare g t.c.
$$L_{g(x_1\dots x_n)}(x_1 \dots x_n) = \sup_{\theta\in\Theta} L_\theta(x_1 \dots x_n) \quad \forall (x_1 \dots x_n)\in\mathbb{R}^n$$

Significato: • Caso discreto: se i dati del campione sono $(x_1 \dots x_n)$, scegliamo la stima $g(x_1 \dots x_n)$ che massimizza $L_\theta(x_1 \dots x_n)$ = prob. P_θ di ottenere $X_1 = x_1 \dots X_n = x_n$
• Caso assol cont: scegliamo la stima $g(x_1 \dots x_n)$ che massimizza "$L_\theta(x_1 \dots x_n)$ ≈ prob di ottenere dati $X_1 \approx x_1 \dots X_n \approx x_n \cdot \delta x_1 \cdot\dots\cdot \delta x_n$"

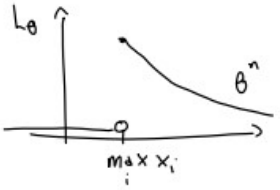
Esempio 1.: $Q_\theta = B(\theta)$ (Bernoulli di par. θ), $\theta\in[0,1]$
$$m_\theta(x) = \begin{cases} 1-\theta & \text{se } x=0 \\ \theta & \text{se } x=1 \end{cases} = \theta^x(1-\theta)^{1-x} \qquad x\in\{0,1\}$$
$$L_\theta(x_1 \dots x_n) = \prod_{i=1}^n m_\theta(x_i) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{x_i}$$
$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)} = \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}$$
con $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$
$$\log L_\theta(x_1 \dots x_n) = n\bar{x}\log\theta + n(1-\bar{x})\log(1-\theta)$$
$$\frac{d}{d\theta}\log L_\theta(x_1 \dots x_n) = \frac{n\bar{x}}{\theta} + \frac{n(1-\bar{x})}{1-\theta}$$
$\frac{d}{d\theta}\log L_\theta(x_1 \dots x_n) = 0 \iff \theta = \bar{x}$ e $\log L_\theta(x_1 \dots x_n) = -\infty$ per $\theta = 0, 1$
quindi $L_\theta(x_1 \dots x_n)$ ha un unico max in $\theta = \bar{x}$
quindi $\bar{X}$ è lo stimatore di max verosimiglianza di θ

Esempio 2: $Q_\theta = \mathcal{U}([0,\theta])$, $\theta > 0$

$m_\theta(x) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x)$

$L_\theta(x_1, \dots x_n) = \prod_{i=1}^n m_\theta(x_i) = \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \mathbb{1}_{0 \le \min_i x_i \le \max_i x_i \le \theta}$

$= \mathbb{1}_{[0,+\infty)}(\min_i x_i) \frac{1}{\theta^n} \mathbb{1}_{[\max_i x_i, +\infty)}(\theta)$



Le ha max in $\theta = \max_i x_i$

Quindi $\max\{X_1, \dots X_n\}$ è lo stimatore di max ver. di θ

Oss: Usando $\theta = 2E[X_i]$, potremmo stimare θ anche con $2\overline{X}$.

Quale tra $\max_i X_i$ e $2\overline{X}$ è "meglio" come stimatore di θ? si dimostra

- $\max_i X_i$ è distorto, ma asintoticamente non distorto
- $\max_i X_i$ è consistente
- $R(\max_i X_i) \le R(2\overline{X})$ quindi $\max_i X_i$ è preferibile a $\overline{X}$

Def: Modello esponenziale: modello statistico $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_\theta)_{\theta \in \Theta})$, $\Theta \subseteq \mathbb{R}$ intervallo, t.c.

- caso discreto: $\exists\, T: \mathbb{N} \to \mathbb{R}$, $g: \mathbb{N} \to \mathbb{R}$ t.c, $\forall \theta \in \Theta$, Q_θ ha densità discreta su $\mathbb{N}$

  $p_\theta(k) = c_\theta g(k) e^{\theta T(k)}$ $k \in \mathbb{N}$ $(c_\theta > 0$ costante$)$
- caso assolutamente continuo: $\exists\, T: \mathbb{R} \to \mathbb{R}$, $g: \mathbb{R} \to \mathbb{R}$ boreliane t.c, $\forall \theta \in \Theta$, Q_θ ha densità

  $f_\theta(x) = c_\theta g(x) e^{\theta T(x)}$ $x \in \mathbb{R}$ $(c_\theta > 0$ costante$)$

Esempi:

1. leggi esponenziali: $f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{(0,+\infty)}(x)$ $\theta = \lambda$, $T(x) = -x$, $g(x) = \mathbb{1}_{(0,+\infty)}(x)$, $c_\lambda = \lambda$
2. leggi Poisson: $p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ $\theta = \log\lambda$, $T(k) = k$, $g(k) = \frac{1}{k!}$, $c_\theta = e^{-\lambda}$ $(\lambda^k = e^{k \log\lambda})$
3. leggi gaussiane $\mathcal{N}(m, \sigma^2)$ $(\sigma^2 = 1$ per semplicità$)$: $f_m(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-m)^2}$ $(e^{-(x-m)^2} = e^{-x^2} e^{2xm} e^{-m^2})$

   $\theta = m$, $T(x) = 2x$, $g(x) = e^{-x^2}$, $c_m = \frac{1}{\sqrt{2\pi}} e^{-m^2}$
4. leggi geometriche $p_\lambda(k) = p(1-p)^{k-1} \mathbb{1}_{k \ge 1}$ $\theta = \log(1-p)$ $T(k) = k-1$, $g(k) = \mathbb{1}_{k \ge 1}$, $c_\theta = p$
5. leggi unif. su $(0, \theta)$, $f_\theta(x) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x)$; non è modello esponenziale

Teor: Sia $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_\theta)_{\theta \in \Theta})$ modello statistico t.c.

- $\forall \theta_1 \neq \theta_2$, $Q_{\theta_1} \neq Q_{\theta_2}$
- $\Theta \subseteq \mathbb{R}$ è intervallo aperto
- $(Q_\theta)_{\theta \in \Theta}$ è modello esponenziale assol. cont.: $f_\theta(x) = c_\theta g(x) e^{\theta T(x)}$
- $x \mapsto g(x) T(x)^2 e^{\theta T(x)+}$ è integrabile $\forall \theta \in \Theta$

Sia $(X_1, \dots X_n)$ un campione i.i.d. di legge Q_θ. Supponiamo inoltre che

- esiste uno stimatore U_n di max verosimiglianza (a valori in Θ)

Allora lo stimatore U_n di max verosimiglianza (è unico ed) è consistente.

La dim si basa su

- legame tra U_n e "funzione di partizione" $\theta \mapsto c_\theta$
- LGN

Dim:

$\psi(\theta) = -\log c_\theta = \log \int_{\mathbb{R}} g(x) e^{\theta T(x)} dx$ $\left(\int_{\mathbb{R}} c_\theta g(x) e^{\theta T(x)} dx = 1 \text{ da cui } \frac{1}{c_\theta} = \int_{\mathbb{R}} g(x) e^{\theta T(x)} dx \right)$

$\psi'(\theta) = \frac{1}{\int g(x) e^{\theta T(x)} dx} \int g(x) T(x) e^{\theta T(x)} dx = c_\theta \int T(x) g(x) e^{\theta T(x)} dx = \int T(x) f_\theta(x) dx = E_\theta[T(X_1)]$

$\psi''(\theta) = -\frac{1}{\left(\int g(x) e^{\theta T(x)} dx\right)^2} \left(\int g(x) T(x) e^{\theta T(x)} dx \right)^2 + \frac{1}{\int g(x) e^{\theta T(x)} dx} \cdot \int g(x) T(x)^2 e^{\theta T(x)} dx$

$= -\left(\int T(x) f_\theta(x) dx \right)^2 + \int T(x)^2 f_\theta(x) dx = -E_\theta[T(X_1)]^2 + E_\theta[T(X_1)^2] = \mathrm{Var}_\theta(T(X_1)) \geq 0$

quindi ψ è convessa.

Mostriamo che $\psi'' > 0$. Per assurdo: $\exists \theta_0 \in \Theta$, $\psi''(\theta_0) = 0$. Quindi $T(X_1) = \text{costante } P_{\theta_0}\text{-q.c.}$, cioè $T = \text{costante } Q_{\theta_0}\text{-q.c}$ (cioè $\exists t \in \mathbb{R}$, $Q_{\theta_0}\{T = t\} = P_{\theta_0}\{T(X_1) = t\} = 1$)

Poiché Q_θ ha densità $c_\theta g e^{\theta T}$, deve essere $T = \text{costante Lebesgue-q.o su } \{g > 0\}$ (cioè $\exists t \in \mathbb{R}$ t.c. $\{T \neq t, g > 0\}$ ha misura di Lebesgue nulla).

Ma allora ogni Q_θ ha densità $c_\theta g(x) e^{\theta t} = c\, g(x)$ indipendente (q.o.) da θ e quindi tutte le Q_θ coincidono $\lightning$.

Quindi ψ' è invertibile su Θ

$L_\theta(x_1, \dots x_n) = f_\theta(x_1) \cdot \dots \cdot f_\theta(x_n) = c_\theta^n\, g(x_1) \cdot \dots \cdot g(x_n)\, e^{\theta(T(x_1) + \dots + T(x_n))}$

$\log L_\theta(x_1, \dots x_n) = n\left[-\psi(\theta) + \theta \frac{1}{n} \sum_{i=1}^n T(x_i)\right] + \log(g(x_1) \cdot \dots \cdot g(x_n))$

$\frac{d}{d\theta} \log L_\theta(x_1, \dots x_n) = n\left[-\psi'(\theta) + \frac{1}{n} \sum_{i=1}^n T(x_i)\right]$

Poiché Θ è intervallo aperto, se U_n stimatore di max ver esiste (a valori in θ), U_n soddisfa

$\frac{d}{d\theta} \log L_{U_n}(X_1, \dots X_n) = 0 \quad (\forall \omega \in \Omega)$

quindi $\psi'(U_n) = \frac{1}{n} \sum_{i=1}^n T(X_i)$ da cui, per invertibilità di ψ' in Θ

$U_n = (\psi')^{-1}\left(\frac{1}{n} \sum_{i=1}^n T(X_i)\right)$.

Per LGN, $\frac{1}{n} \sum_{i=1}^n T(X_i) \xrightarrow{P_\theta} E_\theta[T(X_1)] = \psi'(\theta) \quad \forall \theta$

quindi, per continuità di $(\psi')^{-1}$, $U_n = (\psi')^{-1}\left(\frac{1}{n} \sum_{i=1}^n T(X_i)\right) \xrightarrow{P_\theta} (\psi')^{-1}(\psi'(\theta)) = \theta$ (vedi lemma sotto)

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

Dim:

$\forall \varepsilon > 0$, $\exists \delta > 0$, $|x - \ell| \leq \delta \Rightarrow |\varphi(x) - \varphi(\ell)| \leq \varepsilon$, quindi $\{|\varphi(X_n) - \varphi(\ell)| > \varepsilon\} \subseteq \{|X_n - \ell| > \delta\}$

$P\{|\varphi(X_n) - \varphi(\ell)| > \varepsilon\} \leq P\{|X_n - \ell| > \delta\} \to 0$ per $n \to \infty$.

Oss: Abbiamo usato il teor di deriv. sotto il segno di integrale:

data $F(t)=\int f(t,x)\,dx$, $t_0\in\mathbb{R}$ t.c.
- $x\mapsto f(t,x)$ è integrabile $\forall t\in$ intorno $\mathcal{I}(t_0)$ di t_0
- $t\mapsto f(t,x)$ è differenziabile in $\mathcal{I}(t_0)$ per q.o. x
- $\left|\frac{\partial}{\partial t}f(t,x)\right|\le g(x)$ con g integrabile, $\forall t\in\mathcal{I}(t_0)$, per q.o. x

allora $\frac{d}{dt}F(t_0)=\int\frac{\partial f}{\partial t}(t_0,x)\,dx$

Oss: Il teor. vale anche nel caso discreto, con enunciata del tutto analoga
(l'ipotesi di integrabilità diventa $\sum_{k=0}^{\infty} g(k)T(k)^2 e^{\theta T(k)_+}<\infty \quad \forall\theta\in\Theta$)

Oss: Il teor vale anche per $\Theta\subseteq\mathbb{R}^d$ aperto convesso, $T:\mathbb{R}\to\mathbb{R}^d$ boreliana:
$f_\theta(x)=c_\theta\, g(x)\exp(\theta\cdot T(x))$

Stimatori di massima verosimiglianza di media e varianza di gaussiana

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

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

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

$$\left.\begin{aligned}\frac{\partial}{\partial m}\log L_{(m,\sigma^2)}(x_1,\dots x_n)&=\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-m)\\ \frac{\partial}{\partial\sigma}\log L_{(m,\sigma^2)}(x_1,\dots x_n)&=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n(x_i-m)^2\end{aligned}\right\}=0\iff\begin{cases}m=\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i\\ \sigma^2=\frac{1}{n}\sum_{i=1}^n(x_i-\bar{x})^2=\frac{n-1}{n}s^2\end{cases}$$

e si verifica che $(\bar{x},\frac{n-1}{n}s^2)$ è un pt di max

Quindi $\left(\bar{X},\frac{n-1}{n}S^2\right)$ è lo stimatore di max verosimiglianza per (m,σ^2)

Oss: Con gli stessi conti si verifica, per $Q_\theta=\mathcal{N}(m,\sigma^2)$:
- se σ^2 è nota (cioè $\theta=m\in\Theta=\mathbb{R}$), $\bar{X}$ è lo stimatore di max verosimiglianza per m
- se m è nota (cioè $\theta=\sigma^2\in\Theta=(0,+\infty)$), $\frac{1}{n}\sum_{i=1}^n(X_i-m)^2$ è lo stimatore di max verosimiglianza per σ^2

23
lunedì 8 maggio 2023 20:49

Intervalli di fiducia

Esempio/motivazione:
n lanci di moneta, con p=prob di testa,
$\bar{X}$ = freq relativa campionaria di teste $(= \frac{\#\text{teste}}{n})$ è uno stimatore per p
vogliamo misurare l'incertezza della stima
intuitivamente, più n è grande minore dovrebbe essere l'incertezza

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

Def: Dato $\alpha\in(0,1)$, una regione di fiducia a livello $1-\alpha$ per il parametro θ è un insieme aleatorio $D(\omega)\subseteq\Theta$, $\omega\in\Omega$ (cioè una mappa $\Omega\to\mathcal{P}(\Theta)$, $\omega\mapsto D(\omega)$) t.c.

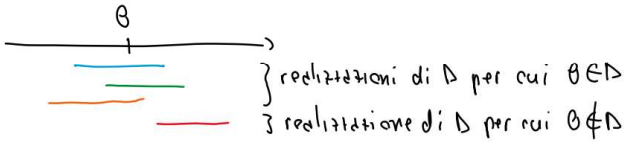
$$P_\theta\{\theta\in D\} \geq 1-\alpha \quad \forall\theta\in\Theta$$

(dove $\{\theta\in D\} = \{\omega\in\Omega \mid \theta\in D(\omega)\}\subseteq\Omega$ e si sottointende che $\{\theta\in D\}\in\mathcal{F}$ $\forall\theta\in\Theta$)

Oss: · α è piccolo, tipicamente $\alpha=0.05,\ 0.01$
· D solitamente è una funzione del campione $X_1 \dots X_n$

Significato: dato un campione $X_1 \dots X_n$, possiamo determinare D t.c., con prob. alta, $\theta\in D$

Oss: nell'evento $\{\theta\in D\}$, l'aleatorietà è in D (che dipende da $X_1 \dots X_n$), non in θ, che è non nota ma deterministico
$P\{\theta\in D\}\geq 1-\alpha$ va inteso intuitivamente come: "in almeno una frazione $1-\alpha$ di tutti i campioni, θ cade in D(campione)"



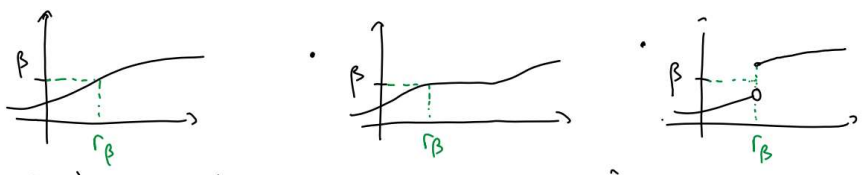
Quantili di una prob. P su $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$; con F d.R F

Def: Dato $\beta\in(0,1)$, quantile di ordine β (o β-quantile) di P (o di F): il numero

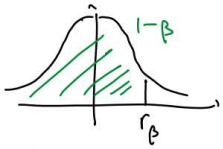
$$r_\beta = \inf\{x\in\mathbb{R} \mid F(x)\geq\beta\}$$

Oss: se F, ristretta ad un intervallo (a,b), è invertibile da (a,b) a $(0,1)$, allora $r_\beta = F^{-1}(\beta)$.

Esempi
-(F invertibile)

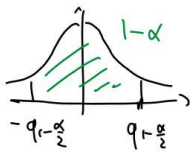


I quantili di $N(0,1)$ si denotano con q_β o z_β



Per simmetria, · $q_{1-\beta} = -q_\beta$
· $P\{-q_{1-\frac{\alpha}{2}} \leq Z \leq q_{1-\frac{\alpha}{2}}\} = 1-\alpha$ per $Z\sim N(0,1)$, $0<\alpha<1$

queste due proprietà valgono in generale se Z ha densità f pari, e più in generale se $Z \overset{(d)}{=} -Z$



Intervalli di fiducia per la media di una popol. normale (varianza nota)

- $(S, \mathcal{I}, (Q_\theta)_{\theta\in\Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), (N(m,\sigma^2))_{m\in\mathbb{R}})$, $(\theta = m \in \Theta \in \mathbb{R})$, σ^2 nota (sia $X \sim N(m,\sigma^2)$)

$(X_1, \dots X_n)$ campione i.i.d. di $N(m,\sigma^2)$

Poiché $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ è uno stimatore di m, è ragionevole cercare D intervallo centrato in $\bar{X}$,

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

Notiamo che, per riproducibilità e invarianza delle gaussiane per trasformazioni lineari,

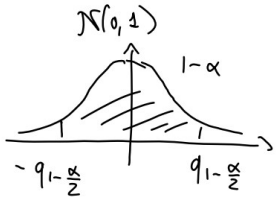
$\bar{X}$, combinazione lineare di gaussiane indip., è gaussiana, con $E[\bar{X}] = m$, $Var(\bar{X}) = \frac{\sigma^2}{n}$, cioè

$\bar{X} \sim N(m, \frac{\sigma^2}{n})$ (sotto P_m)

cioè $\frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim N(0,1)$ (sotto P_m)

Usiamo questo fatto per calcolare $P_m\{m \in [\bar{X} \pm d]\}$

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



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

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

cioè $\frac{\sqrt{n}}{\sigma} d = q_{1-\frac{\alpha}{2}}$ (con q_β quantile di $N(0,1)$ di ordine β)

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

Oss: L'ampiezza dell'intervallo, $2\frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}$,

- è crescente in $1-\alpha$
- è crescente in σ
- è decrescente in n

Esempio: Un metodo per misurare la carica elettrica di un corpo produce misure con distrib. gaussiana di media il valore vero della carica del corpo e dev. standard 0.1 (in Coulomb). Vengono effettuate 16 misurazioni, ottenendo una media campionaria di 5.2. Cerchiamo un intervallo di fiducia di livello 95% per la carica del corpo.

$(S, \mathcal{I}, (Q_\theta)) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), N(m,\sigma^2))_{m}$, $\sigma = 0.1$, $(X$ = carica elettrica rilevata $\sim N(m,\sigma^2))$ $m \in \mathbb{R}$

X_i = risultato i-esima rilevazione $i = 1, \dots n = 16$ $X_1, \dots X_n$ campione i.i.d. di $N(m,\sigma^2)$

livello $1-\alpha = 0.95$, $\alpha = 0.05$, $q_{1-\frac{\alpha}{2}} = q_{0.975} \approx 1.96$

int. di fiducia $= \left[\bar{X} \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \approx \left[\bar{X} \pm \frac{0.1}{4} \cdot 1.96\right] = [\bar{X} \pm 0.049]$

Dopo le misurazioni, con $\bar{x} = 5.2$, l'int. di fiducia risulta $[5.2 \pm 0.049] = [5.151, 5.249]$

Metodo della statistica pivotale: data una statistica pivotale, cioè una statistica $ST = g(\theta, X_1, \dots X_n)$

- ST è invertibile come funzione di θ, dato il campione:
  $\forall (x_1, \dots x_n)$, $\theta \mapsto g(\theta, x_1, \dots x_n)$ è invertibile $\quad \forall A \in \mathcal{B}(\mathbb{R})$
- ST ha legge che non dipende da θ: $ST_{\#} P_\theta$ è indipendente da θ (cioè $P_\theta\{ST \in A\}$ non dip. da θ)

allora una regione di fiducia di livello $1-\alpha$ per θ è data da $D = g(\cdot, X_1, \dots X_n)^{-1}(A)$, dove A è t.c. (deterministica)
$P\{ST \in A\} = 1-\alpha$: infatti

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

Nell'esempio precedente (media di popol. normale), $ST = \frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0,1)$ sotto P_m

Def: Dato $\alpha \in (0,1)$, una regione di fiducia asintotica (o approssimata) di livello $1-\alpha \in (0,1)$ per θ è una successione di insiemi aleatori $D_n(\omega) \subseteq \Theta$, $\omega \in \Omega$, $n \in \mathbb{N}$, t.c.

$$\liminf_n P_\theta\{\theta \in D_n\} \geq 1-\alpha \qquad \forall \theta \in \Theta$$

Intervalli di fiducia asintotici per la media di una popolazione (var. nota), grandi campioni

- $(S, \mathcal{I}, (Q_\theta)_{\theta \in \Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_m)_{m \in \mathbb{R}})$, $m = E_m[X]$, $X \sim Q_m$, $\sigma^2 = \mathrm{Var}(X)$ nota
  $X_1, \dots X_n$ campione i.i.d. di Q_m
  Per il TCL, se n è grande, $ST = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - m)$ è approx $\sim \mathcal{N}(0,1)$, in particolare

$$\lim_n P\{-q_{1-\frac{\alpha}{2}} \leq ST \leq q_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

  Ripetendo i passaggi del caso popol. gaussiana (ST come statistica pivotale), si ottiene che
  $\left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right]$ è un intervallo di fiducia asintotico di livello $1-\alpha$ per m

Intervalli di fiducia asintotici per una proporzione (media di una popol. Bernoulli), grandi campioni

- $(S, \mathcal{I}, Q_\theta)_{\theta \in \Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), (B(p))_{p \in [0,1]})$ (o $(\{0,1\}, \mathcal{P}(\{0,1\}), (B(p))_{p \in [0,1]})$)
  $p = E_p[X]$ con $X \sim B(p)$ (p = prob del successo "= proporzione")
  $X_1, \dots X_n$ campione i.i.d. di $B(p)$ ($\bar{X}_n$ = freq. relativa campionaria del successo)
  Per il TCL, se n è grande, $\sqrt{\frac{n}{p(1-p)}}(\bar{X}_n - p)$ è approx $\sim \mathcal{N}(0,1)$
  Problema: $p \mapsto \sqrt{\frac{n}{p(1-p)}}(\bar{X}_n - p)$ non è invertibile
  Cor: Date $X_1, \dots X_n$ i.i.d. $\sim B(p)$,

$$\sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}(\bar{X}_n - p) = \frac{n\bar{X}_n - np}{\sqrt{n\bar{X}_n(1-\bar{X}_n)}} \xrightarrow{\text{in legge}} Z \sim \mathcal{N}(0,1) \quad \text{per } n \to \infty$$

  Questo risultato si dimostra come Corollario del TCL e di $\bar{X}_n \xrightarrow{\text{in prob.}} p$ (usando però fatti più avanzati)
  Quindi usando $ST = \sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}(\bar{X}_n - p)$ come statistica pivotale (asintotica), si ottiene che

$$\left[\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\, q_{1-\frac{\alpha}{2}}\right] \text{ è un intervallo di fiducia asintotica di livello } 1-\alpha \text{ per } p$$

Intervalli di fiducia asintotici per la media di una popolazione, var. non nota, grandi campioni

- $(S, \mathcal{I}, (Q_\theta)_{\theta\in\Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), (Q_m)_m)$ $m = E_m[X]$, $X \sim Q_m$, $X_1 \dots X_n$ campione i.i.d. di Q_m

  questa volta però supponiamo $\sigma^2 = \mathrm{Var}(X)$ non nota

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

  Cor: Date $X_1 \dots X_n$ i.i.d. con $E[X_i^2] < \infty$,

  $$\frac{\sqrt{n}}{S_n}(\bar{X}_n - m) \xrightarrow{\text{in legge}} Z \sim N(0,1) \quad \text{per } n \to \infty$$

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

  Questo risultato si dimostra come corollario del TCL e di $S_{n-1}^2 \to \sigma^2$ (usando fatti più avanzati)

  Usando $ST = \frac{\sqrt{n}}{S_n}(\bar{X}_n - m)$ come statistica pivotale (asintotica), si ottiene che

  $$\left[\bar{X}_n \pm \frac{\sqrt{n}}{S_n} q_{1-\frac{\alpha}{2}}\right] \text{ è un intervallo di fiducia asintotico di livello } 1-\alpha \text{ per } m$$

Esempio: Viene rilevata l'opinione di 100 persone (estratte a caso) su un certo partito politico. Di queste 100 persone, 25 mostrano gradimento per il partito. Fornire un intervallo di fiducia (approssimato) per la percentuale di persone che apprezzano il partito, di livello 95%.

$(S, \mathcal{I}, (Q_\theta)_\theta) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), B(p)_p)$ $\left(X = \begin{cases} 1 & \text{se la persona apprezza il partito} \\ 0 & \text{altrimenti} \end{cases} \sim B(p)\right)$

p = prob gradimento

$X_i = \begin{cases} 1 & \text{se l'}i\text{-sima persona estratta apprezza il partito} \\ 0 & \text{altrimenti} \end{cases}$, $i = 1, \dots n = 100$ (abbastanza grande), $X_1 \dots X_n$ campione i.i.d. di $B(p)$

$1-\alpha = 0.95$, $q_{1-\frac{\alpha}{2}} = q_{0.975} \simeq 1.96$

int. di fiducia approx per p $\left[\bar{X} \pm \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} q_{1-\frac{\alpha}{2}}\right] \approx \left[\bar{X} \pm \frac{\sqrt{\bar{X}(1-\bar{X})}}{10} \cdot 1.96\right] = \left[\bar{X} \pm 0.196 \cdot \sqrt{\bar{X}(1-\bar{X})}\right]$

Dopo le rilevazioni, con $\bar{x} = \frac{25}{100} = 0.25$, l'int. di fiducia diventa $[0.25 \pm 0.196 \cdot \sqrt{0.25 \cdot 0.75}] \simeq [0.165, 0.335]$

Intervallo di fiducia per la media di una popolazione gaussiana, varianza non nota

- $(S, \mathcal{I}, (Q_\theta)_{\theta\in\Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), N(m, \sigma^2)_{m\in\mathbb{R}})$, $X_1 \dots X_n$ campione i.i.d. di $N(m, \sigma^2)$

  supponiamo σ^2 non nota (e n non necessariamente grande)

  Usiamo:

  $$ST = \frac{\sqrt{n}}{S}(\bar{X} - m) \sim t_{n-1} \quad (t\text{-di-Student a } n-1 \text{ gradi di libertà})$$

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

  In particolare, chiamando $t_{\beta, n-1}$ il quantile di t_{n-1} di ordine β ($P\{T \le t_{\beta, n-1}\} = \beta$),

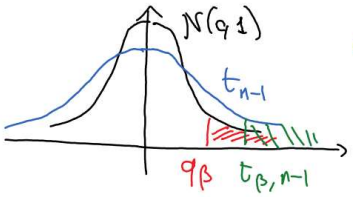
  $$P\left\{ST \in \left[-t_{1-\frac{\alpha}{2}, n-1},\ t_{1-\frac{\alpha}{2}, n-1}\right]\right\} = 1-\alpha$$

  Quindi usando ST come statistica pivotale si ottiene che

  $$\left[\bar{X} \pm \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1}\right] \text{ è un intervallo di fiducia (esatto) di livello } 1-\alpha \text{ per } m$$

Oss: Poiché la densità di t_{n-1} è polinomiale, essa ha code più pesanti della densità $N(0,1)$

e quindi $t_{\beta,n-1} > q_\beta \quad \forall n, \ \forall \beta > \frac{1}{2}$



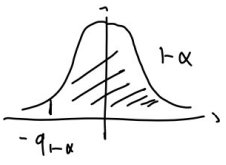
area rossa = area verde = $1-\beta$

Dunque l'incertezza nel caso varianza non nota è maggiore rispetto al caso varianza nota

Oss: Si può dimostrare che, per $n \to \infty$, $t_{n-1} \xrightarrow{\text{in legge}} N(0,1)$, in particolare $t_{\beta,n-1} \to q_\beta$, $\forall \beta \in (0,1)$.

Oss: È possibile considerare anche semirette come regioni di fiducia (intervalli unilateri)
Ad es., data una popolazione $N(m,\sigma^2)$ di varianza σ^2 nota, possiamo cercare una regione di fiducia per la media m della forma $(-\infty, \bar{X}+d]$. Imponiamo $P_m\{m \in (-\infty, \bar{X}+d]\} = 1-\alpha$ e usiamo la statistica pivotale $ST = \frac{\sqrt{n}}{\sigma}(\bar{X}-m)$

$$P_m\{m \in (-\infty, \bar{X}+d]\} = P_m\{\bar{X}+d \geq m\} = P_m\left\{\frac{\sqrt{n}}{\sigma}(\bar{X}-m) \geq -\frac{\sqrt{n}}{\sigma}d\right\}$$
$$= \Phi\left(\frac{\sqrt{n}}{\sigma}d\right)$$
$$= 1-\alpha \iff \frac{\sqrt{n}}{\sigma}d = q_{1-\alpha}$$

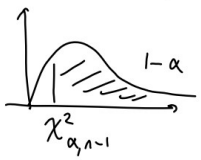


e troviamo $(-\infty, \bar{X} + \frac{\sigma}{\sqrt{n}}q_{1-\alpha}]$ come una regione di fiducia di livello $1-\alpha$ per m

Intervalli di fiducia per la varianza di una popol. gaussiana,

- $(S, \mathcal{Y}, (P_\theta)_{\theta\in\Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), (N(m,\sigma^2))_{\sigma^2\in(0,+\infty)})$, $\sigma^2 = Var_{\sigma^2}(X)$ con $X \sim N(m,\sigma^2)$

  $X_1, \dots X_n$ campione i.i.d.
  supponiamo m non nota
  Usiamo
  $$ST = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{(chi-quadrato a } n-1 \text{ gradi di libertà)}$$
  In particolare, chiamando $\chi^2_{\beta,n-1}$ il quantile di ordine β di χ^2_{n-1} ($P\{Y \leq \chi^2_{\beta,n-1}\} = \beta$)
  $$P\{ST \in [\chi^2_{\alpha,n-1}, +\infty)\} = 1-\alpha$$
  Quindi, usando ST come statistica pivotale ($ST \in [\chi^2_{\alpha,n-1}, +\infty) \iff \sigma^2 \in (0, \frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}]$)
  si ottiene che
  $$\left(0, \frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}\right]$$
  è un intervallo di fiducia (unilatero) per σ^2 di livello $1-\alpha$



Oss: Se m è nota, si usa la statistica pivotale
$$ST = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 \sim \chi^2_n$$

24
mercoledì 10 maggio 2023 00:20

Distribuzioni legate alle v.a. gaussiane

$(\Omega, \mathcal{F}, P)$ sp di prob.

Prop: a) $Z \sim \mathcal{N}(0,1) \Rightarrow Z^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) =: \chi^2_1$ (distribuzione chi-quadrato a un grado di libertà)

b) $Z_1, \dots Z_n \sim \mathcal{N}(0,1)$ i.i.d. $\Rightarrow Z_1^2 + \dots + Z_n^2 \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right) =: \chi^2_n$ (distrib chi-quadrato a n gradi di lib.)

Dim:

a) F_{Z^2} FdR di Z^2, dobbiamo mostrare $F_{Z^2}(u) = \int_{-\infty}^{u} f(v)\,dv$ con f densità $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$,

$$f(u) = \frac{1}{\sqrt{2}\,\Gamma\left(\frac{1}{2}\right)} r^{-1/2} e^{-\frac{1}{2}u} \mathbb{1}_{(0,+\infty)}(u)$$

$F_{Z^2}(u) = 0 \quad \forall u < 0$. Per $u \geq 0$,

$$F_{Z^2}(u) = P\{Z^2 \leq u\} = P\{-\sqrt{u} \leq Z \leq \sqrt{u}\} = 2P\{0 \leq Z \leq \sqrt{u}\} = 2\int_0^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$$

$$\begin{matrix} x^2 = v \\ 2x\,dx = dv \end{matrix} \quad = \int_0^u \frac{1}{\sqrt{2\pi}} e^{-u/2} \frac{1}{\sqrt{v}}\,dv = \int_0^u f(v)\,dv$$

b) Z_i^2 sono indipendenti e $\sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$. Quindi per riproducibilità di Γ, $Z_1^2 + \dots + Z_n^2 \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$.

Prop: $Z = (Z_1, \dots Z_n)^T$ vettore di $\mathcal{N}(0,1)$ indip., $M \in \mathbb{R}^{n \times n}$ ortogonale $\Rightarrow MZ$ vettore di $\mathcal{N}(0,1)$ indip.

Dim:

Notiamo che $Z_1, \dots Z_n$ sono i.i.d. $\sim \mathcal{N}(0,1) \Leftrightarrow Z = (Z_1, \dots Z_n)$ ha densità

$$f_Z(x_1, \dots x_n) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \cdot \dots \cdot \frac{1}{\sqrt{2\pi}} e^{-x_n^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2} \quad (\text{dove } |x|^2 = \sum_{i=1}^n x_i^2)$$

Se M è ortogonale (quindi $|Mx| = |x|$, $|\det M| = 1$), MZ ha densità

$$f_{MZ}(y_1, \dots y_n) = f_Z(M^{-1}y)\,|\det M^{-1}| = \frac{1}{(2\pi)^{n/2}} e^{-|M^{-1}y|^2/2} \frac{1}{|\det M|} = f_Z(y_1, \dots y_n)$$

quindi $MZ \overset{(d)}{=} Z$ è vettore di $\mathcal{N}(0,1)$ indip.

Prop: $Z_1, \dots Z_n \sim \mathcal{N}(0,1)$ indip., $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2 \Rightarrow$

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

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

Dim:

a) Prendiamo $M \in \mathbb{R}^{n \times n}$ definita da ($e_1, \dots e_n$ base canonica di $\mathbb{R}^n$)

$$M^T e_1 = \frac{1}{\sqrt{n}}(1, \dots 1)^T = \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i$$

$M^T e_2, \dots M^T e_n$ base ortonormale di $(M^T e_1)^\perp$

M^T e quindi M sono ortogonali.

Detta $Z = (Z_1, \dots Z_n)^T = \sum_{i=1}^n Z_i e_i$, MZ è vettore di $\mathcal{N}(0,1)$ indip, in particolare

$(MZ)_1$ e $\sum_{i=1}^n (MZ)_i^2$ sono indipendenti

$(Hz)_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i = \sqrt{n}\,\bar{z}$

$\sum_{i=2}^n (Hz)_i^2 = |Hz|^2 - (Hz)_1^2 = |z|^2 - n\bar{z}^2 = \left(\sum_{i=1}^n z_i^2\right) - n\bar{z}^2 = \sum_{i=1}^n (z_i - \bar{z})^2 = (n-1)S_n^2$

Quindi $\bar{z}$ e S_{n-1}^2 sono indip.

b) $\bar{z}$ è gaussiano per riproducibilità e $E[\bar{z}] = 0$, $\mathrm{Var}(\bar{z}) = \frac{1}{n}$

$(n-1)S_{n-1}^2 = \sum_{i=2}^n (Hz)_i^2$ è somma di $(n-1)$ v.a. $N(0,1)$ indip, quindi $(n-1)S_{n-1}^2 \sim \chi^2_{n-1}$.

Def: Date $z \sim N(0,1)$, $Y \sim \chi^2_n$, con z e Y indipendenti, chiamiamo distribuzione t-di-Student a n gradi di libertà (t_n) la legge di

$$T = \frac{\sqrt{n}\,z}{\sqrt{Y}}$$

Cor: $z_1, \dots z_n \sim N(0,1)$ i.i.d. $\Rightarrow \sqrt{n}\,\frac{\bar{z}}{S} \sim t_{n-1}$

Dim:

$\sqrt{n}\,\frac{\bar{z}}{S} = \frac{\sqrt{n-1}\cdot\sqrt{n}\,\bar{z}}{\sqrt{(n-1)S^2}}$ e $\sqrt{n}\,\bar{z} \sim N(0,1)$, $(n-1)S^2 \sim \chi^2_{n-1}$ e sono indipendenti

Prop: t_n ha densità data da

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

In particolare, f_{t_n} è pari.

Dim:

Dobbiamo verificare $F_{t_n}(x) = \int_{-\infty}^x f_{t_n}(x')\,dx'$

$$F_{t_n}(x) = \iint \mathbb{1}_{\sqrt{n}\,z/\sqrt{y} \le x}\, f_{N(0,1)}(z)\, f_{\chi^2_n}(y)\,dy\,dz$$

$$= \iint \mathbb{1}_{\sqrt{n}z/\sqrt{y} \le x}\, c_n\, e^{-z^2/2} y^{-n/2} e^{-y/2}\,dy\,dz \qquad (\text{con } c_n > 0 \text{ costante})$$

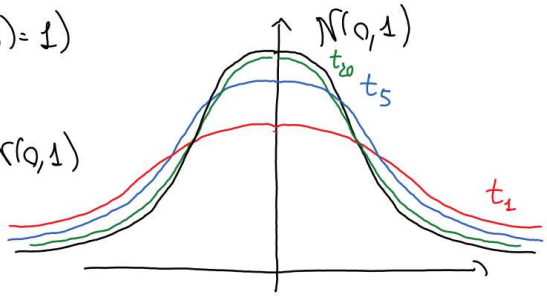
$\frac{\sqrt{n}\,z}{\sqrt{y}} = x'$, $\frac{\sqrt{n}}{\sqrt{y}}\,dz = dx'$

$$= \int_{-\infty}^x c_n' \int e^{-(x')^2 y/2n} e^{-y/2} y^{-(n-1)/2}\,dy\,dx' \qquad (\text{con } c_n' > 0 \text{ costante})$$

$\left(1 + \frac{(x')^2}{n}\right)y = \tilde{y}$, $\left(1 + \frac{(x')^2}{n}\right)dy = d\tilde{y}$

$$= \int_{-\infty}^x c_n' \left(1 + \frac{(x')^2}{n}\right)^{-\frac{n+1}{2}} \underbrace{\int e^{-\tilde{y}/2}\, \tilde{y}^{-\frac{n-1}{2}}\,d\tilde{y}}_{\text{cost.}}\,dx' = \int_{-\infty}^x c_n''\, f_{t_n}(x')\,dx'$$

L e necessariamente $c_n'' = 1$ (cosicché $F(+\infty) = 1$)



Oss: • t_n ha code polinomiali, più "pesanti" delle code $N(0,1)$

• per $n \to \infty$, "$t_n \to N(0,1)$"

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

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

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

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

Dim: dai risultati precedenti e da standardizzazione.

25
giovedì 11 maggio 2023 01:02

Test statistici 1

Esempio/motivazione:

n lanci di moneta, con p = prob di testa

il banco dichiara che la moneta è equilibrata, cioè $p = \frac{1}{2}$

sulla base degli esiti degli n lanci, vogliamo verificare se l'ipotesi $p = \frac{1}{2}$ è plausibile o no:

- se su 1000 dati escono $\begin{cases} 501 \text{ teste, l'ipotesi } p=\frac{1}{2} \text{ è} & \text{plausibile} \\ 900 \text{ " , " " " } & \text{non plausibile} \\ 540 \text{ " , " " " } & ? \end{cases}$

Un test statistico è una procedura per verificare, sulla base dei dati $(x_1 \dots x_n)$ del campione, se una certa ipotesi (=affermazione sul parametro) è plausibile o no.

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

Useremo spesso $(x_1, \dots x_n)$ per indicare i dati ottenuti dal campione $(X_1, \dots X_n)$

Elementi principali di un test (parametrico)

1) L'ipotesi da verificare: $\Theta = \Theta_0 \cup \Theta_1$ partizione di Θ $(\Theta_0 \cap \Theta_1 = \emptyset)$

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

$H_1 : \theta \in \Theta_1$ " alternativa

Nell'esempio precedente: $H_0 : \theta = \frac{1}{2}$, $\Theta_0 = \{\frac{1}{2}\}$, $\Theta_1 = [0,1] \setminus \{\frac{1}{2}\}$

Oss: asimmetria tra H_0 e H_1:

- 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 evidente statistica per H_1
- se invece l'esito del test è l'accettazione di H_0, allora H_0 è plausibile, ma non c'è per forza evidente per H_0.

2) La regione critica o di rifiuto: un evento $C \in \mathcal{F}$, solitamente dipendente dal campione $X_1 \dots X_n$, t.c., se l'esito ω cade in C rifiutiamo l'ipotesi H_0 in favore di H_1
   - se l'esito ω non cade in C, accettiamo l'ipotesi H_0.

   La scelta di C deve essere t.c.: • C sia poco probabile sotto H_0 (poco plausibile)
   • se $\omega \in C$, allora questa è un'evidente per H_1

   Nell'esempio precedente ($H_0 : p = \frac{1}{2}$), ci aspettiamo una regione critica C della forma
   $C = \{ |\bar{X} - \frac{1}{2}| > d \}$ per un $d > 0$ opportuno, per cui $P_{\frac{1}{2}}(C)$ sia piccola

   Se invece consideriamo $H_0 : p \le \frac{1}{2}$, ci aspettiamo una regione critica C della forma
   $C = \{ \bar{X} - \frac{1}{2} > d \}$ per un $d > 0$ opportuno, per cui $P_p(C)$ sia piccola $\forall p \le \frac{1}{2}$

Oss: Nella scelta di C spesso si usa una statistica (statistica di test) ST, la cui legge non dipenda da θ $(C = \{ ST \in A \}$ per un opportuno $A)$

L'esito del test (che è aleatorio poiché dipende dai dati del campione) può essere errato, in due modi:
- errore di 1ª specie: H_0 è vera ma l'esito del test è il rifiuto di H_0
- " " 2ª specie: H_0 è falsa ma l'esito del test è l'accettazione di H_0

Def: Dato $\alpha \in (0,1)$ (piccolo, ad es. $\alpha = 0.05, 0.01$), un test è di livello α se

$$\sup_{\theta \in \Theta_0} P_\theta(C) \leq \alpha$$

cioè la prob. dell'errore di 1ª specie (che dipende da $\theta \in \Theta_0$) è al massimo α.

Più α è piccolo, maggiore è l'evidenza statistica fornita dal test contro H_0 in caso di rifiuto.

Def: Potenza di un test: funzione

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

cioè $\pi_C(\theta)$ è la prob di rifiutare H_0 quando il valore del parametro è $\theta \in \Theta_1$,
quindi $\pi_C(\theta)$ rappresenta la capacità di accorgersi che l'ipotesi H_0 è falsa

Nell'approccio "classico" ai test, si fissa un livello α piccolo e si cerca C in modo da massimizzare la potenza

Caso ipotesi semplice: $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \Theta \setminus \{\theta_0\}$: Legame tra test e intervalli di fiducia

- Dato $\alpha \in (0,1)$, se D è una regione di fiducia di livello $1-\alpha$ per θ, allora
$$C = \{\theta_0 \notin D\} = \{\omega \in \Omega \mid \theta_0 \notin D(\omega)\}$$
è una regione critica di livello α: infatti
$$P_{\theta_0}(C) = 1 - P_{\theta_0}\{\theta_0 \in D\} \leq 1 - (1-\alpha) = \alpha$$
- Viceversa, dato $\alpha \in (0,1)$, se, $\forall \theta_0 \in \Theta$, C_{θ_0} è una regione critica di livello α per il test $H_0 : \theta = \theta_0$, allora
$$D(\omega) = \{\theta \in \Theta \mid \omega \notin C_\theta\}$$
è una regione di fiducia per θ di livello $1-\alpha$: infatti
$$P_{\theta_0}\{\theta_0 \in D\} = P_{\theta_0}(C_{\theta_0}^c) = 1 - P_{\theta_0}(C_{\theta_0}) \geq 1-\alpha \quad \forall \theta_0 \in \Theta$$

Test bilatero per la media di una popolazione gaussiana, varianza nota (test z)

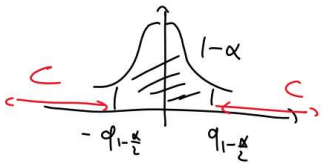
- $(S, \mathcal{I}, (Q_\theta)_{\theta \in \Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m, \sigma^2)_{m \in \mathbb{R}})$, $\theta = m \in \Theta = \mathbb{R}$, σ^2 nota
$X_1, \dots X_n$ campione i.i.d. di $\mathcal{N}(m, \sigma^2)$
Dato $m_0 \in \mathbb{R}$, consideriamo il test $H_0 : m = m_0$ contro $H_1 : m \neq m_0$, a livello α
Regione critica C: poiché $\bar{X}$ è uno stimatore di m e $H_1 : m \neq m_0$, è ragionevole cercare C della forma $C = \{|\bar{X} - m_0| > d\}$ per d opportuno
Usiamo la statistica di test
$$ST = \frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0,1) \text{ sotto } P_m$$

e calcoliamo $P_{m_0}(C) = P_{m_0}\left\{\frac{\sqrt{n}}{\sigma}|\bar{X}-m_0| > \frac{\sqrt{n}}{\sigma}d\right\} = P\{|Z| > \frac{\sqrt{n}}{\sigma}d\}$ con $Z \sim N(0,1)$

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

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

è una regione critica di livello α



Oss: $\omega \in C \Leftrightarrow |\bar{X}(\omega)-m_0| > \frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}} \Leftrightarrow m_0 \in \left[\bar{X}(\omega) \pm \frac{\sigma}{\sqrt{n}}q_{1-\alpha}\right]$ (Int. di fiducia per m_0 di livello $1-\alpha$)

Test bilatero per una proporzione (media di popol. Bernoulli), grandi campioni (test z approssimato ✓)

- $(S, \mathcal{J}, (Q_\theta)_{\theta\in\Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), B(p)_{p\in[0,1]})$ $\quad \theta = p \in \Theta = [0,1]$

  $X_1, \dots X_n$ campione i.i.d. di $B(p)$, assumiamo n grande (approssimativamente ✓)

  Dato $p_0 \in (0,1)$, consideriamo il test $H_0: p = p_0$ contro $H_1: p \neq p_0$ di livello α ✓

  Regione critica C: è ragionevole cercare C della forma $C = \{|\bar{X}-p_0| > d\}$

  Usiamo la statistica di test

  $$ST = \sqrt{\frac{n}{p(1-p)}}(\bar{X}-p), \text{ per il TLC} \overset{\text{approx}}{\sim} N(0,1)$$

  Otteniamo che

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

  è una regione critica di livello approx. α

  Oss: qui non abbiamo bisogno di considerare $\bar{X}(1-\bar{X})$ al posto di $p(1-p)$, poiché conosciamo, sotto H_0, $p = p_0$

Analogamente agli intervalli di fiducia, e con le stesse statistiche pivotali, si possono formulare i test bilateri per media di popol. gaussiana con varianza non nota (test t), media nel caso grandi campioni (test z approssimato), varianza di popol. gaussiana (test χ^2)

Metodo del p-value (soglia di accettazione):

Spesso si ha una famiglia di regioni critiche $(C_\alpha)_{\alpha\in(0,1)}$ con

- C_α regione critica a livello α, $\forall \alpha$
- se $\alpha \leq \alpha'$, allora $C_\alpha \subseteq C_{\alpha'}$ (ragionevole: minore è $\alpha \geq \sup_{\theta\in\Theta_0} P_\theta(C)$, più piccola è C_α)
- $\bigcap_\alpha C_\alpha = \emptyset$, $\bigcup_\alpha C_\alpha = \Omega$ (si può prendere $\Omega = \{(x_1 \dots x_n) | x_i \in S\} = S^n$) ✓

Quindi, $\forall \omega \in \Omega$ (esito del campione), $\exists\, \bar{\alpha} = \bar{\alpha}(\omega) \in (0,1)$

- se $\alpha < \bar{\alpha}$, allora $\omega \notin C_\alpha$ e quindi accettiamo H_0 se l'esito del campione è ω
- se $\alpha > \bar{\alpha}$, allora $\omega \in C_\alpha$ e quindi rifiutiamo H_0 " " " " " ω

$\bar{\alpha} = \bar{\alpha}(\omega)$ è detto p-value di ω (ed è funzione dell'esito del campione, solitamente $\bar{\alpha} = \bar{\alpha}(X_1 \dots X_n)$) e fornisce il "grado di plausibilità" di H_0 se si verifica ω (più $\bar{\alpha}$ è basso, minore è il livello

α per cui rifiuto, quindi meno plausibile è H_0).

Informalmente, $\bar\alpha(\omega)$ è "la prob., sotto H_0, di avere dati più estremi (rispetto a H_0) di ω"

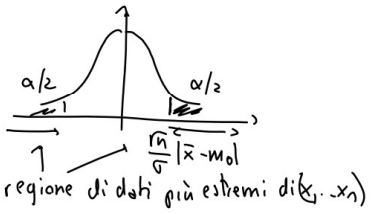
Esempi:

- media di popol. normale $\mathcal{N}(m,\sigma^2)$, varianza nota, $H_0: m = m_0$, $H_1: m \neq m_0$

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

  $\bar\alpha(x_1 \dots x_n)$ è t.c $\frac{\sqrt{n}}{\sigma}|\bar x - m_0| = q_{1-\frac{\alpha}{2}}$, cioè $\frac{\alpha}{2} = P\left\{ Z > \frac{\sqrt{n}}{\sigma}|\bar x - m_0| \right\}$ con $Z \sim \mathcal{N}(0,1)$

  $= P\left\{ |Z| > \frac{\sqrt{n}}{\sigma}|\bar x - m_0| \right\}$



- media di popol. $B(p)$, n grande, $H_0: p = p_0$, $H_1: p \neq p_0$

  $C_\alpha = \left\{ \sqrt{\frac{n}{p_0(1-p_0)}} |\bar X - p_0| > q_{1-\frac{\alpha}{2}} \right\}$ (approx.)

  $\bar\alpha(x_1 \dots x_n)$ è t.c $\frac{\alpha}{2} = P\left\{ Z > \sqrt{\frac{n}{p_0(1-p_0)}} |\bar x - p_0| \right\}$ (approx.)

  con $Z \sim \mathcal{N}(0,1)$

Esempi:

1. Un certa modello teorico afferma che la carica elettrica di un dato corpo (in Coulomb) è 5. Dai risultati di 16 misurazioni della carica sul corpo, risulta una media campionaria di 5.2. Supponiamo che le misurazioni siano gaussiane con media il valore reale della carica e dev. standard pari a 0.1. C'è evidenza che il modello teorico proposto sia errato? Effettuare un test di livello 0.01 e calcolare il p-value per i dati forniti.

  $(S, \mathcal{I}, (Q_\theta)_{\theta\in\Theta}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m,\sigma^2)_{m\in\mathbb{R}})$, $\sigma = 0.1$ (X = carica rilevata in una misur. $\sim \mathcal{N}(m,\sigma^2)$)

  X_i = i-sima misurazione, $i = 1, \dots n = 16$, campione i.i.d. di $\mathcal{N}(m,\sigma^2)$

  $H_0: m = 5 (= m_0)$, $H_1: m \neq 5$

  Statistica di test $Z = \frac{\sqrt{n}}{\sigma}(\bar X - m) \sim \mathcal{N}(0,1)$

  Regione critica di livello $\alpha = 0.01$ ($q_{1-\frac{\alpha}{2}} = q_{0.995} \approx 2.576$)

  $C = \left\{ \frac{\sqrt{n}}{\sigma}|\bar X - m_0| > q_{1-\frac{\alpha}{2}} \right\} = \{ 8\,|\bar X - 5| > 2.576 \}$

  Per $\bar x = 5.2$, $8\cdot|\bar x - 5| = 1.6 \leq 2.576$ $\Rightarrow (x_1 \dots x_n) \notin C$, accettiamo H_0: non c'è evidenza statistica dei dati, a livello 0.01, contro il lavoro teorico

  p-value di $\bar x = 5.2$: $\bar\alpha(\bar x = 5.2) = 2P\{ Z > 8|\bar x - 5| \} = 2(1 - \Phi(1.6)) \approx 2\cdot(1 - 0.945) = 0.11$

  accettiamo H_0 per $\alpha < 0.11$, rifiutiamo H_0 per $\alpha > 0.11$

- Il banco dichiara che una certa moneta è equilibrata. Lanciamo 1000 volte la moneta, ottenendo 540 teste. C'è evidenza statistica contro l'affermazione del banco? Effettuare un test di livello 0.05 e calcolare il p-value sulla base dei dati.

$(S, \mathcal{I}, (Q_\theta)_{\theta\in\Theta}) = (\mathbb{N}, \mathcal{B}(\mathbb{N}), B(p)_{p\in[0,1]})$ $\quad p$ = prob di testa

$X_i = \begin{cases} 1 & \text{se testa al lancio } i\text{-simo} \\ 0 & \text{altrimenti} \end{cases}$ $\quad i = 1, \dots n = 1000$ (grande), campione i.i.d. di $B(p)$

$H_0: p = \frac{1}{2} (= p_0)$ (affermazione del banco) $\quad H_1: p \neq \frac{1}{2}$

Statistica di test: $ST = \sqrt{\frac{n}{p(1-p)}}(\bar{X} - p) \overset{approx}{\sim} N(0,1)$

Regione critica di livello approx. $\alpha = 0.05$ $\quad (q_{0.975} \approx 1.96)$

$$C = \left\{ \sqrt{\frac{n}{p_0(1-p_0)}} |\bar{X} - p_0| > q_{1-\frac{\alpha}{2}} \right\} = \left\{ 63.25 \cdot |\bar{X} - 0.5| > 1.96 \right\}$$

Per $\bar{x} = \frac{540}{1000} = 0.54$, $\quad 63.25 \cdot |\bar{x} - 0.5| = 2.53 > 1.96 : (x_1 \dots x_n) \in C$, rifiutiamo H_0: dai dati c'è evidenza, a livello $\alpha = 0.05$, contro $p = \frac{1}{2}$

p-value (approx.) di $\bar{x} = 0.54$: $\quad \alpha(\bar{x} = 0.54) = 2P\left\{ Z > \sqrt{\frac{n}{p_0(1-p_0)}} |\bar{x} - p_0| \right\}$

$$= 2P\{Z > 2.53\} = 2(1 - \Phi(2.53)) \approx 2(1 - 0.9943) = 0.0114$$

martedì 16 maggio 2023 22:29

Test statistici 2

Caso ipotesi composte, test unilateri: $\Theta_0 = (-\infty, \theta_0]$ ($H_0: \theta \le \theta_0$), $\Theta_1 = (\theta_0, +\infty)$ ($H_1: \theta > \theta_0$)
(analogamente per $\Theta_0 = [\theta_0, +\infty)$, $\Theta_1 = (-\infty, \theta_0)$)

Esempio: per una moneta con prob. di testa p, ($\theta = p \in \Theta = [0,1]$, $Q_\theta = B(p)$), $H_0: p \le \frac{1}{2}$, $H_1: p > \frac{1}{2}$

È ragionevole prendere una regione critica della forma $\{\bar X > d\}$

Dato α livello del test, per determinare d dobbiamo imporre $\sup_{\theta \le \theta_0} P_\theta\{\bar X > d\} \le \alpha$,
intuitivamente ci aspettiamo che il sup è realizzato per $\theta = \theta_0$

Domande: a) è vero che il sup è realizzato per $\theta = \theta_0$?
b) la regione $\{\bar X > d\}$ è effettivamente la scelta "migliore"? precisamente, fissato α (e quindi d), essa massimizza la potenza del test?

D'ora in poi, considereremo regioni critiche della forma $C = \{(X_1 \dots X_n) \in \tilde C\}$ (con $\tilde C \in \mathcal{J}^{\otimes n}$)

Def: Dati $(S, \mathcal{J}, (Q_\theta)_{\theta \in \Theta})$ modello statistico con $\Theta \subseteq \mathbb{R}$ intervallo, $(X_1 \dots X_n)$ campione i.i.d. di legge $(Q_\theta)_{\theta \in \Theta}$, definita su $(\Omega, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$, $T: \Omega \to \mathbb{R}$ v.a.,
diciamo che il modello è a rapporto di verosimiglianza crescente rispetto a T: se
$\forall \theta_1 < \theta_2$, esiste una funzione $f_{\theta_1,\theta_2}: \mathbb{R} \to (0, +\infty)$ strettamente crescente t.c.

$$\frac{L(\theta_2, X_1 \dots X_n)}{L(\theta_1, X_1 \dots X_n)} = f_{\theta_1,\theta_2}(T)$$

(cioè $L(\theta_2, X_1 \dots X_n)/L(\theta_1, X_1 \dots X_n)$ è funz. strett. cresc. di T)
(Si sottointende che, se $L(\theta_1, x_1 \dots x_n) = 0$, allora $L(\theta_2, x_1 \dots x_n) = 0$

Esempi:

- Modello gaussiano $(S, \mathcal{J}, (Q_\theta)_\Theta) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m, \sigma^2)_m)$

$$L_m(x_1 \dots x_n) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - m)^2\right) = e^{-\frac{1}{2\sigma^2}(m_2 - m_1)(2n\bar x + (m_2 - m_1))}$$

$$\frac{L_{m_2}(x_1 \dots x_n)}{L_{m_1}(x_1 \dots x_n)} = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \left((x_i - m_2)^2 - (x_i - m_1)^2\right)\right) = \exp\left(-\frac{1}{2\sigma^2}(m_1 - m_2)\sum_{i=1}^n (2x_i - m_1 - m_2)\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}(m_2^2 - m_1^2)\right) \exp\left(\frac{n(m_2 - m_1)}{\sigma^2}\bar x\right)$$

per $m_2 > m_1$, il rapporto è crescente in $\bar x$

- Modello Bernoulli $(S, \mathcal{J}, (Q_\theta)_\Theta) = (\mathbb{N}, \mathcal{B}(\mathbb{N}), B(p)_p)$

$$L_p(x_1 \dots x_n) = p^{n\bar x}(1-p)^{n(1-\bar x)}$$

$$\frac{L_{p_2}(x_1 \dots x_n)}{L_{p_1}(x_1 \dots x_n)} = \left(\frac{p_2}{p_1}\right)^{n\bar x}\left(\frac{1-p_2}{1-p_1}\right)^{n(1-\bar x)}$$

per $p_2 > p_1$, il rapporto è crescente in $\bar x$

Teor: Siano $(S, \mathcal{T}, (P_\theta)_{\theta\in\Theta})$ modello statistico con $\Theta \subseteq \mathbb{R}$ intervallo, $(X_1 .. X_n)$ campione i.i.d. di legge Q_θ, supponiamo che il modello sia a rapporto di ver. crescente rispetto a T.

Consideriamo il test $H_0: \theta \le \theta_0$ contro $H_1: \theta > \theta_0$, sia

$C = \{T \ge d\}$

Allora:

(i) $\sup_{\theta\le\theta_0} P_\theta(C) = P_{\theta_0}(C)$

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

Lemma (Neyman-Pearson): Siano $(S, \mathcal{T}, (P_\theta)_{\theta\in\Theta})$ modello statistico con $\Theta = \{\theta_0, \theta_1\}$,

consideriamo il test $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$, per $c > 0$ sia

$C = \{L_{\theta_0}(X_1 .. X_n) \le c L_{\theta_1}(X_1 .. X_n)\}$

Allora:

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

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

Dim (del lemma di N.-P.): Per semplicità, consideriamo il caso assol. cont.

a) Notiamo che, se il test ha regione critica D, la sua potenza è $P_{\theta_1}(D)$, vogliamo quindi dim. $P_{\theta_1}(D) \le P_{\theta_1}(C)$.

Scriviamo $C = \{(X_1 .. X_n) \in \tilde{C} = \{L_{\theta_0} \le c L_{\theta_1}\}\}$, $D = \{(X_1 .. X_n) \in \tilde{D} \subseteq \mathbb{R}^n\}$

Notiamo $P_\theta(D) = P_\theta\{(X_1 .. X_n) \in \tilde{D}\} = \int_{\tilde{D}} L_\theta(x_1 .. x_n)\, dx_1 .. dx_n$

Per def di $\tilde{C} = \{L_{\theta_0} \le c L_{\theta_1}\}$

$$(\mathbb{1}_{\tilde{C}} - \mathbb{1}_{\tilde{D}})(L_{\theta_0} - c L_{\theta_1}) \le \underbrace{\mathbb{1}_{\tilde{C}}(L_{\theta_0} - c L_{\theta_1})}_{\le 0} - \underbrace{\mathbb{1}_{\tilde{D}\setminus\tilde{C}}(L_{\theta_0} - c L_{\theta_1})}_{\ge 0} \le 0$$

Integriamo su $\mathbb{R}^n$ (risp. alla misura di Lebesgue):

$$\int (\mathbb{1}_{\tilde{C}} - \mathbb{1}_{\tilde{D}}) L_{\theta_0}(x_1 .. x_n)\, dx_1 .. dx_n \le c \int (\mathbb{1}_C - \mathbb{1}_D) L_{\theta_1}(x_1 .. x_n)\, dx_1 .. dx_n$$

$$P_{\theta_0}(C) - P_{\theta_0}(D) \qquad\qquad c(P_{\theta_1}(C) - P_{\theta_1}(D))$$

Poiché D ha livello $P_{\theta_0}(C)$, $P_{\theta_0}(D) \le P_{\theta_0}(C)$, quindi $P_{\theta_1}(C) - P_{\theta_1}(D) \ge 0$

b) $$(\mathbb{1}_{\tilde{C}} - P_{\theta_0}(C))(L_{\theta_0} - c L_{\theta_1}) = \underbrace{(1 - P_{\theta_0}(C))}_{\ge 0} \underbrace{\mathbb{1}_{\tilde{C}}(L_{\theta_0} - c L_{\theta_1})}_{\le 0} - P_{\theta_0}(C) \underbrace{\mathbb{1}_{\tilde{C}^c}(L_{\theta_0} - c L_{\theta_1})}_{\ge 0} \le 0$$

Integrando come prima,

$$c(P_{\theta_1}(C) - P_{\theta_0}(C)) \ge 0$$

Dim del teor:

i) Basta dim. che $\theta_1 \le \theta_2 \le \theta_0 \Rightarrow P_{\theta_1}(C) \le P_{\theta_2}(C)$

$$C = \{T \ge d\} = \{f_{\theta_1,\theta_2}(T) \ge \underbrace{f_{\theta_1,\theta_2}(d)}_{=: 1/c}\} = \{L_{\theta_1}(X_1 .. X_n) \le c L_{\theta_2}(X_1 .. X_n)\}$$

T rapporto a ver crescente; $f_{\theta_1,\theta_2}(T) = \frac{L_{\theta_2}(X_1 .. X_n)}{L_{\theta_1}(X_1 .. X_n)}$

quindi per Lemma di N-P. (b) (con $\Theta=\{\theta_1,\theta_2\}$) $P_{\theta_1}(C) \le P_{\theta_2}(C)$

ii) Dobbiamo dim. che, se D è un'altra regione critica di livello $P_{\theta_0}(C)$, $\forall \theta>\theta_0$, $P_\theta(D) \le P_\theta(C)$

Poiché $P_{\theta_0}(D) \le P_{\theta_0}(C)$, per lemma di N-P. (a) (con $\Theta=\{\theta_0,\theta\}$), $P_\theta(D) \le P_\theta(C)$.

Test unilatero per la media di una popol. gaussiana, varianza nota

- $(S, \mathcal{I}_r(Q_\theta)) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(m,\sigma^2)_{m\in\mathbb{R}})$ $\quad \theta = m \in \mathbb{R}$, σ^2 nota

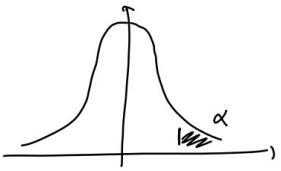
$X_1 \dots X_n$ campione i.i.d.

Consideriamo il test $H_0: m \le m_0$ contro $H_1: m > m_0$, a livello α

Regione critica C: $C = \{\bar X > d\}$, ottimale per il teor.

Usiamo la statistica di test

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



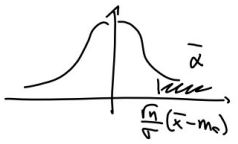
Imponiamo $\sup_{m\le m_0} P_m\{\bar X > d\} \overset{\text{teor}}{=} P_{m_0}\{\bar X > d\} = \alpha$

$$\alpha = P_{m_0}\{\bar X > d\} = P_{m_0}\left\{\frac{\sqrt n}{\sigma}(\bar X - m_0) > \frac{\sqrt n}{\sigma}(d - m_0)\right\} = P\{Z > \tfrac{\sqrt n}{\sigma}(d-m_0)\} \quad \text{con } Z\sim\mathcal{N}(0,1)$$

$$\iff \frac{\sqrt n}{\sigma}(d-m_0) = q_{1-\alpha}$$

Quindi $C = \left\{\frac{\sqrt n}{\sigma}(\bar X - m_0) > q_{1-\alpha}\right\}$

Il p-value di $(x_1 \dots x_n)$ è $\bar\alpha = P\{Z > \frac{\sqrt n}{\sigma}(\bar x - m_0)\}$ con $Z\sim\mathcal{N}(0,1)$

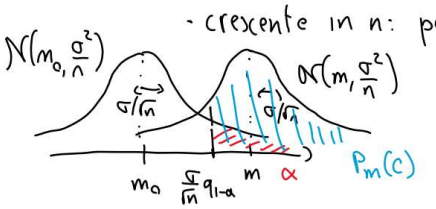


Oss: La potenza del test in $m > m_0$ è

$$\beta_m(C) = P_m\left\{\frac{\sqrt n}{\sigma}(\bar X - m) > q_{1-\alpha} - \frac{\sqrt n}{\sigma}(m-m_0)\right\} = P\{Z > q_{1-\alpha} - \tfrac{\sqrt n}{\sigma}(m-m_0)\} \quad (Z\sim\mathcal{N}(0,1))$$

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

$P_m(C)$ è:
- crescente in m (tenendo fissi gli altri parametri)
- crescente in α: se si vuole abbassare α, purtroppo anche la potenza si abbassa
- crescente in n: per abbassare il livello α e la potenza, è possibile aumentare n



Test unilatero per una proporzione (media di popol. Bernoulli), grandi campioni

- $(S, \mathcal{I}_r(Q_\theta)_{\theta\in\Theta}) = (\mathbb{N}, \mathcal{B}(\mathbb{N}), \mathcal{B}(p)_{p\in[0,1]})$, $\quad \theta = p \in [0,1]$

$X_1 \dots X_n$ campione i.i.d., assumiamo n grande

Consideriamo il test $H_0: p \le p_0$ contro $H_1: p > p_0$, a livello approssimativamente α

Regione critica C: $C = \{\bar X > d\}$ ottimale per il teor.

Usiamo la statistica di test

$$ST = \sqrt{\frac{n}{p(1-p)}}(\bar X - p), \text{ per il TLC } \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

Otteniamo come sopra

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

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

Il p-value di $(x_1 \dots x_n)$ è (approssimativamente) $\bar{\alpha}(x_1 \dots x_n) = P\left\{ Z > \sqrt{\frac{n}{p_0(1-p_0)}} (\bar{x} - p_0) \right\}$

Analogamente si possono formulare i test unilateri per media di popol. gaussiana con varianza non nota (t test), media nel caso grandi campioni (z test approssimato), varianza di popol. gaussiana (χ^2 test)

Esempio:

- Un'azienda dichiara che la percentuale di pezzi difettosi, sul totale dei pezzi prodotti, non supera il 10%. Su 1000 prodotti testati, 120 risultano difettosi. C'è evidenza contro l'affermazione dell'azienda? Effettuare un test a livello $\alpha = 0.05$

  $(S, \mathcal{I}, (Q_\theta)_\theta) = (\mathbb{N}, \mathcal{B}(\mathbb{N}), B(n)_{p \in (0,1]})$. p = prob pezzo difettoso

  $X_i = \begin{cases} 1 & \text{se } i\text{-simo pezzo difettoso} \\ 0 & \text{altrimenti} \end{cases}$, $i = 1, \dots n = 1000$ (grande) campione i.i.d. di $B(p)$

  $H_0: p \le 0.1 (= p_0)$ $\quad H_1: p > 0.1$

  Statistica di test $ST = \sqrt{\frac{n}{p(1-p)}} (\bar{X} - p) \overset{\text{approx}}{\sim} N(0,1)$

  Regione critica di livello $\alpha = 0.05$ (approx.)

  $$C = \left\{ \sqrt{\frac{n}{p_0(1-p_0)}} (\bar{X} - p_0) > q_{1-\alpha} \right\} = \{ 63.25 \cdot (\bar{X} - 0.1) > 1.96 \}$$

  Per $\bar{x} = \frac{120}{1000} = 0.12$, $63.25 \cdot (\bar{x} - 0.1) = 1.265 \le 1.96$: $(x_1 \dots x_n) \notin C$, non c'è evidenza contro dichiarazione azienda

  p-value di $\bar{x}$ (approx.): $\bar{\alpha}(\bar{x} = 0.12) = P\left\{ Z > \sqrt{\frac{n}{p_0(1-p_0)}} (\bar{x} - p_0) \right\}$

  $= P\{ Z > 1.265 \} = 1 - \phi(1.265) \approx 1 - 0.898 = 0.102$

  si accetta per $\alpha < \bar{\alpha}$, si rifiuta per $\alpha > \bar{\alpha}$