STATISTICA INFERENZIALE

LEZIONE 21 - 15/05/25

Si ha un fenomeno regolato da un meccanismo, di cui si hanno delle parti del suo funzionamento a noi sconosciute. Osservando il fenomeno si ha un campione a disposizione, e analizzandolo si vogliono trovare le parti oscure del meccanismo, quelle che non si conoscono.
↳ Questo processo è detto INFERENZA

Si parte da una popolazione, un insieme di individui da cui si estrae un campione.
La popolazione è descritta da una var. al. X (Nella stat. inferenziale)

⊛ Il campione è definito, nella stat. inf. (DUE MODI) come: una serie di n var. al. $X_1, \dots, X_n$ i.i.d (INDIPEND. E IDENTICAMENTE DISTRIBUITE) come X
una serie di n specificazioni (VALORI OSSERVATI) $x_1, \dots, x_n$

La distribuzione di X può essere totalmente ignota: CAMPO DELLA STATISTICA INFERENZIALE NON PARAMETRICA

" " parzialmente ignota: SI CONOSCE LA FORMA GENERALE DELLA DISTRIBUZIONE MA CI MANCANO DEI DETTAGLI (I PARAMETRI)

[(moda: spec. dove la funz. di massa o densità assume maggior valore)
DISTRIBUZIONE UNIMODALE: È PRESENTE UN'UNICA MODA (MASSIMO NEL GRAFICO)]

STATISTICA INFERENZIALE PARAMETRICA

$X \sim F(\theta)$ X segue una qualche distribuzione (e sappiamo quale)

θ indica i parametri ignoti, ciò che non conosciamo

STATISTICA: o STIMATORE: l'algoritmo che a partire dal campione $(x \in D_X)$ stima un possibile valore per θ
Sono quasi sempre specifiche per il problema considerato

$t : (D_X)^n \to \mathbb{R}$ (#CAMPIONE) t non deve dipendere da parametri ignoti (θ)
↑ FUNZIONE STIMATORE

! $t(x_1, \dots, x_n)$ non è mai esattamente $= \theta$, dipende dal campione che si estrae.

⊛ CAMPIONE $\begin{cases} X_1, \dots, X_n \text{ i.i.d come } X \\ x_1, \dots, x_n \text{ (specificazioni)} \end{cases}$

gli individui vanno scelti tutti con la stessa probabilità, x avere un buon campione
⊛ (ogni individuo deve avere la stessa prob. di essere scelto) ≠ CAMPIONE DI CONVENIENZA

Se $X_1, \dots, X_n$ sono i.i.d allora questa richiesta è soddisfatta ⊛

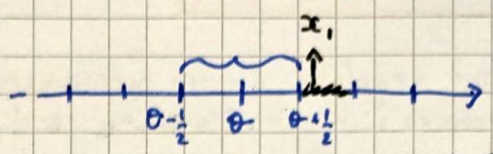
ma neanche da altre var. ignote. ← per questo t non può dipendere da θ ← $t(x_1, \dots, x_n) = \hat{\theta}$ stima di θ ← c'è una variabilità intrinseca nel valore restituito da t

Il valore restituito dalla statistica t è quindi $\hat{\theta} \approx \theta$

$\rightarrow t(x_1, \dots, x_n) = \hat{\theta} \approx \theta$

$\theta \in \mathbb{N}$ $\quad f_X(x) = I_{(\theta - \frac{1}{2}, \theta + \frac{1}{2})}(x)$

$x_1 = 5, 6$



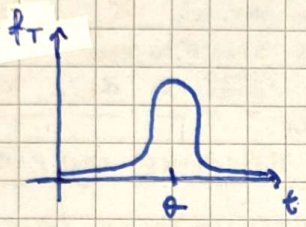
$\Rightarrow \theta = 6$

ESEMPIO SUL LIBRO

$t(X_1, \dots, X_n) = T$ var. aleatoria

modella l'incertezza che si ha della stima.
Distribuzione dei valori $\hat{\theta}$ che si possono ottenere

↳ Ci dice quali sono buoni e quali no.



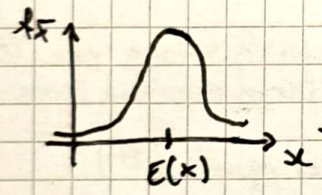
Esempio di statistica $\boxed{\text{Media campionaria}}$ Stima ragionevole del valore atteso

$t(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad t(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i := \overline{X}_n$

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

$X_i \sim X$ (stessa distribuzione)

Indipendentemente dalla distrib. che descrive la popolazione, il valore atteso della media camp. è = a quello della popolazione. (la stima bene, non è esattamente =)



→ è centrato sul valore atteso

Utile se si vuole calcolare il valore atteso della popolazione. (ossia è ciò che non si conosce)

Noi però non conosciamo i parametri di una distribuzione:

(A) $X \sim B(p) \quad E(X) = p$ nel modello di Bernoulli è utile, in quanto il parametro coincide con il valore atteso

$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \approx p = \hat{p}$

θ: PARAMETRO IGNOTO

$\tau(\theta)$: QUANTITÀ DA STIMARE CHE DIPENDE DA θ

(A) p PARAMETRO IGNOTO $\quad \tau(p) = E(X) = p$

$X \sim (n, p) \qquad \tau(p) = E(X) = np$

n → NOTO, p → IGNOTO

- Si vuole stimare $\tau(\theta)$
- si usa una statistica t
- si ha un campione $X_1, \dots, X_n$

$T = t(X_1, \dots, X_n)$ $\quad E(T) = E(t(X_1, \dots, X_n)) = \tau(\theta)$ (TAU)

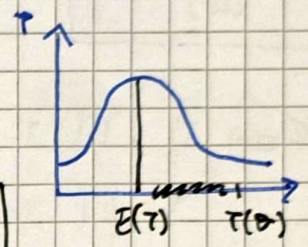
↳ se sì, t è NON DEVIATO per $\tau(\theta)$ quando la distribuz. della popolazione appartiene a quel particolare modello

NON DISTORTO

UNBIASED

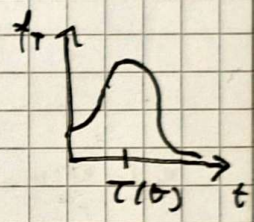
Si parla di BIAS di uno stimatore t rispetto a $\tau(\theta)$:

$$b_{\tau(\theta)}(T) = E(T) - \tau(\theta) = \tau(\theta) - E(T)$$



$\tau(\theta)$ $\quad t(X_1, \dots, X_n) = T$

NON DEVIATO RISPETTO A $\tau(\theta)$ $\Rightarrow$ DISTRIBUZIONE DI T CENTRATA RISPETTO A $\tau(\theta)$ $\Rightarrow$ NON C'È UN ERRORE SISTEMATICO



$E(T) = \tau(\theta)$ NON DISTORTO

$\Rightarrow$ LA MEDIA CAMPIONARIA NON È DISTORTA RISPETTO AL VALORE ATTESO DELLA POPOLAZIONE

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ è sempre non deviato rispetto a $E(X)$

Come si calcola la concentrazione attorno al valore atteso? VARIANZA

VARIANZA DELLA MEDIA CAMPIONARIA

$$\rightarrow Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \overset{\text{INDIPENDENZA DELLE } X_i}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \underset{X_i \sim X}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X) =$$

IDENTICAMENTE DISTRIBUITE

$$= \frac{1}{n^2}\, n\, Var(X) = \frac{Var(X)}{n}$$

Se la varianza dello stimatore è bassa, allora è buona. + n è grande, + il campione è grande e più la media campionaria è precisa.

[...]

tutto questo non può essere esteso facilmente a uno stimatore generalizzato.

MSE: Mean Squared Error

Lo scopo ultimo è stimare $\tau(\theta)$ e si usa $T = t(X_1, \dots, X_n)$

$\underbrace{(T - \tau(\theta))^2}_{\text{VAR. ALEATORIA}}$ non si può calcolare: 1) non si conosce $\tau(\theta)$
2) non si conosce T: ogni volta che la si osserva cambia

$$\boxed{E((T - \tau(\theta))^2) = MSE_{\tau(\theta)}(T)}$$

qual è mediamente l'MSE di T rispetto a $\tau(\theta)$

$$MSE_{\tau(\theta)}(T) = E((T - \tau(\theta))^2) = E(((T - \overbrace{E(T) + E(T)}^{=0} - \tau(\theta))^2) =$$

$$= E((T - E(T))^2 + 2(T - E(T))(\underbrace{E(T) - \tau(\theta)}_{\text{costanti}}) + (\underbrace{E(T) - \tau(\theta)}_{\text{costanti}})^2) =$$

$$= E((T - E(T))^2) + 2(E(T) - \tau(\theta))\underbrace{\boxed{E(T - E(T))}}_{\substack{= 0 \\ E(T) - E(E(T))}} + (E(T) - \tau(\theta))^2 =$$

$$= Var(T) + b_{\tau(\theta)}(T)^2$$

Se lo stimatore è non deviato, l'MSE è uguale alla sua varianza

$$MSE_{E(X)}(\bar{X}) = Var(\bar{X}) + \underbrace{\cancel{b_{E(X)}(\bar{X})^2}}_{=0} = \frac{Var(X)}{n}$$

→ La stima diventa accurata

Si introduce la CONSISTENZA IN MEDIA QUADRATICA: $MSE_{\tau(\theta)}(T) \xrightarrow[n \to +\infty]{} 0$

CAMPIONE

$X_1, X_2 \sim N(\mu, \sigma)$ $\qquad \theta = \sigma$

$\sim X$ (POPOLAZIONE) $\qquad \tau(\theta) = \sigma^2$

$X_1^2 - X_1 X_2 = T$ $\qquad$ T è non distorto per σ^2?

$t(X_1, X_2) = X_1^2 - X_1 X_2 = T$

$$E(T) = E(X_1^2 - X_1 X_2) = E(X_1^2) - E(X_1 X_2) \overset{\text{INDIPENDENZA}}{=} E(X_1^2) - E(X_1)E(X_2) \overset{X_1, X_2 \sim X}{=}$$

$$= E(X^2) - E(X)E(X) = Var(X) = \sigma^2$$ NON È DISTORTO

Popolazione X $\quad E(X) = \mu \quad Var(X) = \sigma^2$

$\uparrow$

$\theta = \mu, \ \tau(\mu) = \mu$

$T = \frac{1}{3}\sum_{i=1}^{n} X_i$ è distorto per μ?

$$E(T) = E\left(\frac{1}{3}\sum_{i=1}^{n} X_i\right) = \frac{1}{3}\sum_{i=1}^{n} \underbrace{E(X_i)}_{=\mu} = \frac{n}{3}\mu \overset{?}{=} \mu$$

distorto per $n \neq 3$, non distorto per $n = 3$ (è la media camp)

Si sono stimatori asintoticamente distorti ($n \to +\infty$) sono distorti:

$$MSE_\mu(T) = Var(T) + b_\mu(T)^2 = \qquad b_\mu(T) = \frac{n}{3}\mu - \mu = \frac{n-3}{3}\mu$$

$$= \frac{n}{3}\sigma^2 + \frac{(n-3)^2}{9}\mu^2 \qquad Var(T) = \frac{1}{9}\sum_{i=1}^{n} \underbrace{Var(X_i)}_{=\sigma^2} = \frac{n}{3}\sigma^2$$

$\xrightarrow{n \to +\infty} 0$ non è consistente in media quadratica

---

$X \sim B(p) \qquad t(X_1, X_2, X_3) = \frac{X_1 + 2X_2 + X_3}{5} = T \qquad$ Voglio stimare p

è distorto?

$$E(T) = \frac{1}{5}\left[\underbrace{E(X_1)}_{=p} + 2\underbrace{E(X_2)}_{=p} + \underbrace{E(X_3)}_{=p}\right] = \frac{4}{5}p \qquad b_p(T) = p - \frac{4}{5}p = +\frac{1}{5}p$$

$$MSE_p(T) = \frac{6}{25}p(1-p) + \frac{1}{25}p^2 = \qquad Var(T) = \frac{1}{25}\left[\underbrace{Var(X_1)}_{p(1-p)} + 4\underbrace{Var(X_2)}_{p(1-p)} + \underbrace{Var(X_3)}_{p(1-p)}\right]$$

$$= \frac{p}{25}\left(6(1-p) + p\right) = p(6-5p) \qquad = \frac{6}{25}p(1-p)$$

LO STIMATORE NON DISTORTO NON È UNICO:

$\tau(\theta) = E(X) \qquad t_1(X_1, \dots, X_n) = X_7 = T_1 \quad E(T_1) = E(X) = \tau(\theta)$

$\qquad\qquad\qquad\qquad t_2(X_1, \dots, X_n) = X_8 = T_2 \quad E(T_2) = E(X) = \tau(\theta)$

$\bar{X} \qquad E(\bar{X}) = E(X) \qquad Var(\bar{X}) = \frac{Var(X)}{n}$

$X \sim B(p) \qquad \forall i \ X_i \sim B(p) \qquad \sum_{i=1}^{n} X_i \sim B(n, p) \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Conosciamo la sua distribuzione, è una binomiale che invece di andare da 1 a n va da $\frac{1}{n}$ a 1.

$X \sim N(\mu, \sigma)$ $\quad \forall i \; \underline{X_i \sim N(\mu, \sigma)}$ $\quad \underbrace{\sum_{i=1}^{n} X_i}_{:= S} \sim \cancel{N(\mu, \sigma)}$ $N(n\mu, \sqrt{n}\,\sigma)$

$\left(Var(\Sigma X_i) = n\sigma^2 \right)$

$$\boxed{\frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)} = \overline{X}$$

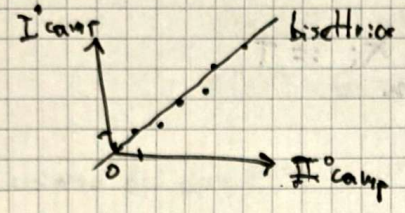
$$\left(\begin{array}{ll} E(S) = n\mu & Var(S) = n\sigma^2 \\ E\left(\frac{1}{n} S\right) = \frac{1}{n} E(S) = \mu & Var\left(\frac{1}{n} S\right) = \frac{1}{n^2} Var(S) = \frac{\sigma^2}{n} \end{array} \right)$$

SE LA POPOLAZIONE È NORMALE, LA MEDIA CAMPIONARIA È NORMALE

$\boxed{\text{LEZIONE 22 - 20/05/2025}}$ Foglio protocollo

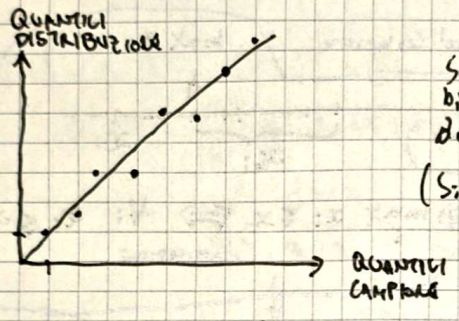
LEZIONE 22 - 20/05/2025

- Il quantile si può introdurre per una generica var. aleatoria.
- Il Q-Q plot si può estendere anche alle variabili aleatorie. Serviva per confrontare le distribuzioni di due campioni:



I Quantili individuano in modo univoco la popolazione.

- Ora si ha un solo campione, e la popolazione da cui è estratto ha una prefissata distribuzione:



Se i quant. si allineano sulla bisettrice, allora il campione è estratto da una popolazione con quella distribuzione.

(Si confrontano i quantili empirici con quelli teorici)

Sono usati nel contesto in cui le distribuzioni delle popolazioni sono normali, di cui però non si conoscono i parametri. Standardizzandole si ottiene una distribuzione normale standard, di cui conosciamo i parametri.

ASSENZA DI DEVIAZIONE → (DIREZIONE DELLA MIRA)

CONSISTENZA

} DUE PROPRIETÀ DESIDERATE PER GLI STIMATORI

↳ VALORI CONCENTRATI O MENO (BUONA MIRA)

TEOREMA CENTRALE DEL LIMITE

Se abbiamo n var. aleatorie $X_1, \dots, X_n$ i.i.d., allora $\sum_{i=1}^{n} X_i$ ha come val atteso $n\mu$ e varianza $n\sigma^2$

→ var. aleat. a sua volta

(con valore atteso $= \mu$ e varianza $= \sigma^2$ (e detto che esistono))

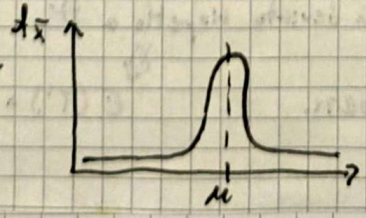
→ APPROSSIMA BENE

$$\sum_{i=1}^{n} X_i \dot{\sim} N(n\mu, \sqrt{n}\,\sigma)$$

più n aumenta è più questa approssimazione è precisa.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \dot{\sim} N\left(\underbrace{\mu}_{\frac{1}{n}\cdot n\mu}, \sigma/\sqrt{n}\right)$$

per un campione abbastanza grande, la media campionaria è uno stimatore di cui conosciamo approssimativamente la distribuzione.



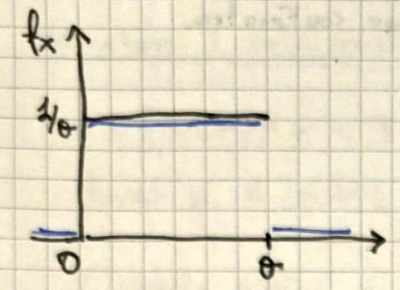
il picco è tanto + concentrato su μ per:

- σ piccolo
- n grande

La popolazione è distribuita secondo una legge uniforme continua:

$$X \sim U(0, \theta)$$

Sia $\tau(\theta) = \theta$



~~[illegible]~~

CAMPIONE $X_1, \dots, X_n$

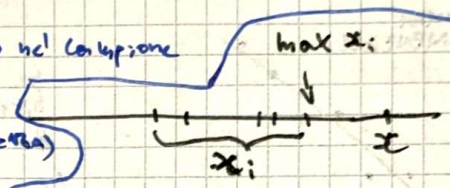
STATISTICA $t(X_1, \dots, X_n) = \max_{1 \le i \le n} X_i := T$

↳ statistica che utilizziamo

Questa statistica è non deviata? $E(T) = \theta$?

$$E(T) = \int_{-\infty}^{+\infty} x f_T(x)\, dx \quad (*)$$

X_i sono nel campione



$$F_T(x) = P(T \le x) = P(\max_i X_i \le x) =$$

(INDIPENDENZA)

$$\overset{(*)}{=} P(\forall i\ X_i \le x) = P\left(\bigcap_{i=1}^{n} \{X_i \le x\}\right) =$$

$$= \prod_{i=1}^{n} P(X_i \le x) = \prod_{i=1}^{n} F_{X_i}(x) \overset{i.i.d.}{=} \prod_{i=1}^{n} F_X(x) =$$

$$= F_X(x)^n$$

(*) $\max_i x_i \le x \iff \forall i\ x_i \le x$

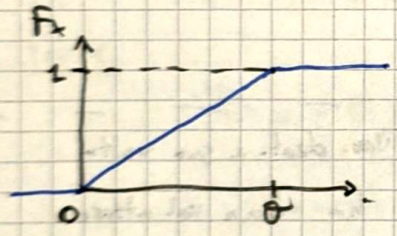
COINCIDENZA

Se al posto delle spec. ci mettiamo le relative var. aleat., queste si coimplicano

$\Rightarrow$ stessa probabilità

$\Rightarrow \{\forall i\ X_i \le x\} \overset{C}{\supseteq} \{\max_i X_i \le x\}$

In $X \sim U(0, \theta)$, $F_X(x) = \frac{x}{\theta} \underbrace{I_{[0,\theta)}(x) + I_{[\theta, +\infty)}(x)}_{\text{Due insiemi disgiunti}}$



Le nostre x stanno però da 0 a θ

$\Rightarrow F_X(x) = \frac{x}{\theta}$ (ignoriamo le funz. indicatrici)

$\Rightarrow F_T(x) = \left(\frac{x}{\theta}\right)^n \Rightarrow f_T(x) = \frac{d}{dx}\left(\frac{x}{\theta}\right)^n = n\left(\frac{x}{\theta}\right)^{n-1} \frac{1}{\theta} = n\frac{x^{n-1}}{\theta^n}$

$$(*)\ E(T) = \int_0^{\theta} n \frac{x^n}{\theta^n}\, dx = \frac{n}{\theta^n} \frac{x^{n+1}}{n+1}\Big|_0^{\theta} =$$

$$= \frac{n}{\theta^n} \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta \neq \theta$$

T SOTTOSTIMA

$\boxed{\hat{\theta} \le \theta}$ per $n \to +\infty$ diventa trascurabile

METODO PLUG-IN

TRASFORMAZIONE LINEARE

$\tau(\theta) = \theta \quad E(T) = \frac{n}{n+1}\theta \qquad \frac{n+1}{n} E(T) = \theta$

$\Rightarrow E\left(\frac{n+1}{n} T\right) = \theta$

$\boxed{T' = \frac{n+1}{n} T = \frac{n+1}{n} \max_i X_i}$ è non deviato rispetto a θ SI BILANCIA LA SOTTOSTIMA

$\Downarrow$

$E(T') = \theta$

SI APPLICA A STIMATORI RAGIONEVOLI PERÒ DISTORTI.

• $T = \max_i X_i \qquad E(T) = \frac{n}{n+1}\theta$

$T' = \frac{n+1}{n} \max_i X_i \qquad E(T') = \theta$

• Si consideri ora $U = \bar{X} \qquad E(U) = E(\bar{X}) = E(X) = \frac{\theta}{2}$

$U' = 2\bar{X} \qquad \cancel{E(U') = 2E(\bar{X})} \quad 2E(U) = \theta \Rightarrow E(2U) = \theta$

$E(U') = \theta \qquad U'$ non è distorto

• Quale usiamo dei due? T' o U'?

$MSE(T') = Var(T')$

$Var(T') = Var\left(\frac{n+1}{n}\max_i X_i\right) = \left(\frac{n+1}{n}\right)^2 Var\left(\overbrace{\max_i X_i}^{:=T}\right) = (*) \qquad f_T(x) = n\frac{x^{n-1}}{\theta^n}$

$Var(T) = E(T^2) - E(T)^2 = (*)$

$E(T^2) = \int_0^\theta x^2 f_T(x)\,dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx = \frac{n}{\theta^n}\left.\frac{x^{n+2}}{n+2}\right|_0^\theta =$

$= \frac{n}{\theta^n}\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2$

$(*)\ Var(T) = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\right)^2\theta^2 = n\theta^2\left(\frac{1}{n+2} - \frac{n}{(n+1)^2}\right) = n\theta^2\frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)^2} =$

$= n\theta^2\frac{\cancel{n^2}+\cancel{2n}+1-\cancel{n^2}-\cancel{2n}}{(n+2)(n+1)^2} = \frac{n}{(n+2)(n+1)^2}\theta^2$

$(*) = \frac{\cancel{(n+1)^2}}{n^{\cancel{2}}}\cdot\frac{\cancel{n}}{(n+2)\cancel{(n+1)^2}}\theta^2 = \underbrace{\boxed{\frac{1}{n(n+2)}\theta^2}}_{MSE(T')} \xrightarrow{n\to+\infty} 0$ CONSISTENTE IN MEDIA QUADRATICA

$X \sim U(0,\theta) \qquad Var(X) = \frac{\theta^2}{12}$

$MSE(U') = Var(U') = Var(2\bar{X}) = 4Var(\bar{X}) = 4\frac{Var(X)}{n} =$

$= 4\frac{\theta^2}{12n} = \boxed{\frac{\theta^2}{3n}} \xrightarrow{n\to+\infty} 0$ CONSISTENTE

• T' però tende a 0 quadraticamente, più velocemente rispetto a U'. Servono campioni più piccoli.

$MSE(T') \leq MSE(U') \qquad \frac{1}{\cancel{n}(n+2)}\cancel{\theta^2} \leq \frac{1}{3\cancel{n}}\cancel{\theta^2} \Rightarrow \frac{1}{n+2} \leq \frac{1}{3}$

$n+2 \geq 3 \qquad \boxed{n \geq 1}$ T' è da preferirsi a U' in $X \sim U(0,\theta)$

LEZIONE 23 - 22/05/2025

Legge dei grandi numeri

La versione forte implica quella debole.
Sia $\mu = E(X)$.

Debole $\underbrace{|\bar{X} - \mu|}$ $> \varepsilon$ $\quad P(|\bar{X} - \mu| > \varepsilon) \xrightarrow{n \to +\infty} 0 \quad \forall \varepsilon > 0$

Questo valore tende a essere basso
Di conseguenza più si aumenta n, più $\bar{X}$ è preciso e più $|\bar{X} - \mu|$ è basso. La probabilità che quindi sia $>$ di un qualsiasi ε diminuisce sempre di più fino a essere 0.

Forte $\bar{X} = \mu \quad P(\lim_{n \to +\infty} \bar{X} = \mu) = 1$

La media campionaria diventa una var. aleatoria degenere, la costante μ.

Metodo Plug-in

$E(T) \neq \tau(\theta)$

$E(T) = \alpha \neq \tau(\theta)$

$X \sim E(\lambda)$ voglio stimare λ $\quad (\tau(\theta) = \theta = \lambda)$

Propongo $\bar{X}$ come stimatore. È distorto?

$E(\bar{X}) = E(X) = \frac{1}{\lambda} \Rightarrow$ sì è distorto

Non si riesce a usare il metodo Plug-in

$\frac{1}{E(\bar{X})} = \lambda$ va al di là delle operazioni lineari del valore atteso

$\not\Rightarrow$ non è molto formale, non è un'implicita

$\bar{X} \approx \frac{1}{\lambda} \rightsquigarrow \lambda \approx \frac{1}{\bar{X}}$

PROPONGO $\frac{1}{\bar{X}}$ COME STIMATORE DI λ. NON POSSIAMO PERÒ PIÙ ANDARE AVANTI.

$E\left(\frac{1}{\bar{X}}\right)$ non è facile da calcolare, non ci si riesce (con le nostre conoscenze)

Non possiamo dire se godo di particolari proprietà in quanto è stato utilizzato il metodo plug-in con trasformazioni non lineari.

$X \sim G(p)$ voglio stimare p

$$E(\bar{X}) = E(X) = \frac{1-p}{p}$$

$$\bar{X} \approx \frac{1-p}{p} \rightsquigarrow p\bar{X} \approx 1-p \rightsquigarrow p(1+\bar{X}) \approx 1 \rightsquigarrow p \approx \frac{1}{1+\bar{X}}$$

PROPONGO $\frac{1}{1+\bar{X}}$ COME STIMATORE. CI SI FERMA QUA, NON POSSIAMO DIRE SE SIA DISTORTO O MENO.

Varianza campionaria | Statistica

Popolazione X, campione $X_1, \dots, X_n$

MEDIA CAMPIONARIA

Varianza campionaria $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

TEOREMA: Sia $\sigma^2 = Var(X)$. S^2 è uno stimatore non distorto per σ^2. $S^2 = \sigma^2$

Dim:

(i) $$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\underbrace{\boxed{\sum_{i=1}^{n} x_i}}_{= n\bar{x}} + n\bar{x}^2 =$$

$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Questo risultato non cambia se applicato sulle var. aleatorie

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

(iii) $E(X^2) = Var(X) + E(X)^2$

(•) $$\underbrace{E((n-1)S^2)}_{\circledast} = E\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right) = E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \underbrace{\sum_{i=1}^{n} \overbrace{E(X_i^2)}^{\text{i.i.d.} = E(X^2)}}_{n E(X^2)} - n E(\bar{X}^2) =$$

$$= n\left[E(X^2) - E(\bar{X}^2)\right] = n\left[Var(X) + \cancel{E(X)^2} - Var(\bar{X}) - \cancel{E(\bar{X})^2}\right] =$$

$$= n\left[\sigma^2 - \frac{\sigma^2}{n}\right] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

$$\circledast = (n-1)E(S^2) = (n-1)\sigma^2 \Rightarrow E(S^2) = \sigma^2$$

La varianza campionaria è sempre non distorta per la varianza della popolazione

Si divide per $n-1$ perché se si dividesse per n la var. camp. sarebbe distorta

Taglia del campione per il valore atteso di una popolazione usando la media campionaria

$X \sim N(d, 2)$

d → distanza del corpo celeste dalla terra (in anni luce)

2 → dev. standard σ

Visto che le misure presentano errori, si fanno tante misure e si prende la media:

- RIPETO n VOLTE L'ESPERIMENTO: $X_1, \dots, X_n$
  ↳ Le X_i sono tra loro indipendenti e distribuite come X.
- CALCOLO $\bar{X}$, LA MEDIA ARITMETICA

$$\boxed{\begin{aligned} &E(\bar{X}) = E(X) = d \text{ in questo caso} \\ &Var(\bar{X}) = \frac{Var(X)}{n} = 4/n \\ &\sigma_{\bar{X}} = \frac{2}{\sqrt{n}} \end{aligned}}$$

$|\bar{X} - d|$ errore che si compie in una misurazione. È una var. aleat.

$|\bar{X} - d| \leq 0.5$ a volte è vera, a volte falsa. È un evento

$P(|\bar{X} - d| \leq 0.5) \geq 0.95$ (95% delle volte è vero)

$$P(|\bar{X} - d| \leq 0.5) = P(-0.5 \leq \bar{X} - d \leq 0.5)$$

Si applica l'operazione di standardizzazione:

$$P\left(-\frac{1}{2} \cdot \frac{\sqrt{n}}{2} \leq \boxed{\frac{\bar{X} - d}{2/\sqrt{n}}} \leq \frac{1}{2} \cdot \frac{\sqrt{n}}{2}\right)$$

$:= Z \sim N(0,1)$ è approssimativamente normale.

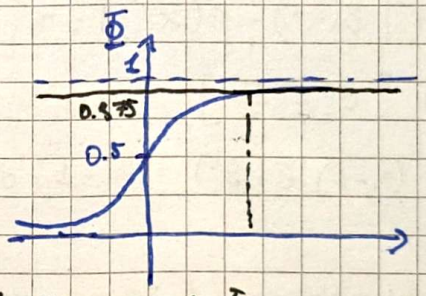
Si usa il teorema centrale del limite: (approssimazione)

$$\approx P\left(-\frac{\sqrt{n}}{4} \leq Z \leq \frac{\sqrt{n}}{4}\right) = \Phi\left(\frac{\sqrt{n}}{4}\right) - \Phi\left(-\frac{\sqrt{n}}{4}\right) =$$

$\boxed{\Phi(-x) = 1 - \Phi(x)}$

↳ VA BENE CHE NON SIA < POICHÉ CONTINUA

$$= \Phi\left(\frac{\sqrt{n}}{4}\right) - 1 + \Phi\left(\frac{\sqrt{n}}{4}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1$$



$$P(|\bar{X} - d| \leq 0.5) \approx 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1 \geq 0.95$$

$$\Rightarrow \Phi\left(\frac{\sqrt{n}}{4}\right) \geq \frac{1.95}{2} \approx 0.975$$

Per via della monotonicità crescente:

$$\Rightarrow \frac{\sqrt{n}}{4} \geq \Phi^{-1}(0.975)$$

↳ 1) esiste l'inversa di Φ
2) essendo monotona crescente non cambia il verso della disequaz.

$$n \geq \lceil (4\Phi^{-1}(0.975))^2 \rceil$$

$$n \geq \lceil 61.4 \rceil = 62$$

Con 62 esperimenti si ha una prob. del 95% di commettere un errore ≤ 0.5

- Problema: approssimazione dovuta al TCL. Per $n = 62$ però questa appross. è buona. Fosse stato + basso si sarebbe preso un n + grande che garantisse una buona appross. del T.C.L.

TUTTO L'ES PRIMA È GENERALIZZABILE:

Popolazione $\boxed{X \sim N(\mu, \sigma)}$ $E(X) = \mu$ $Var(X) = \sigma^2$ $Var(\bar{X}) = \frac{\sigma^2}{n}$ $\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$

$|\bar{X} - \mu|$ Media campionaria per stimare μ

$|\bar{X} - \mu| \leq \underset{\text{SOGLIA}}{\varepsilon}$ è un evento con prob. $\underbrace{P(|\bar{X} - \mu| \leq \varepsilon) \geq 1 - \delta}_{= \text{MOLTO PROBABILE}}$ ⊛ per $\varepsilon, \delta > 0$ e piccoli: $1 - \delta \in [0, 1]$

$$⊛ = P(-\varepsilon \leq \bar{X} - \mu \leq \varepsilon) = P\left(-\varepsilon \frac{\sqrt{n}}{\sigma} \leq \boxed{\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}} \leq \varepsilon \frac{\sqrt{n}}{\sigma}\right) \underset{\text{(Per via del T.C.L.) approssimazione}}{\approx} P\left(-\varepsilon \frac{\sqrt{n}}{\sigma} \leq Z \leq \varepsilon \frac{\sqrt{n}}{\sigma}\right) =$$

$:= Z \sim N(0,1)$ (TCL)

$$= \Phi\left(\varepsilon \frac{\sqrt{n}}{\sigma}\right) - \overbrace{\Phi\left(-\varepsilon \frac{\sqrt{n}}{\sigma}\right)}^{= 1 - \Phi\left(\varepsilon \frac{\sqrt{n}}{\sigma}\right)} = 2\Phi\left(\varepsilon \underbrace{\frac{\sqrt{n}}{\sigma}}_{1/\sigma_{\bar{X}}}\right) - 1$$

$$⊛ \quad 2\Phi\left(\varepsilon \frac{\sqrt{n}}{\sigma}\right) - 1 \geq 1 - \delta \qquad \Phi\left(\varepsilon \frac{\sqrt{n}}{\sigma}\right) \geq 1 - \frac{\delta}{2} \qquad \varepsilon \frac{\sqrt{n}}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\delta}{2}\right)$$

$$\boxed{n \geq \left(\frac{\sigma}{\varepsilon} \Phi^{-1}\left(1 - \frac{\delta}{2}\right)\right)^2}$$

- Precisione maggiore = errore minore = ε + piccolo = n + grande
- Più δ è piccolo, e più $\Phi^{-1}(\dots)$ aumenta
- Più σ è grande, e più n è grande

$\boxed{\varepsilon \geq \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\delta}{2}\right)}$ con n fissato bisogna tollerare almeno quell'errore là

$\frac{\delta}{2} \geq 1 - \Phi\left(\varepsilon \frac{\sqrt{n}}{\sigma}\right)$ $\boxed{\delta \geq 2\left(1 - \Phi\left(\varepsilon \frac{\sqrt{n}}{\sigma}\right)\right)}$

$$P(|\bar{X} - \mu| \leq \varepsilon) = 1 - P(|\bar{X} - \mu| \geq \varepsilon) \underset{\text{CHEBYSHEV}}{\geq} 1 - \frac{Var(\bar{X})}{\varepsilon^2} = 1 - \frac{\sigma^2}{n\varepsilon^2} \geq 1 - \delta$$

$$\Rightarrow \frac{\sigma^2}{n\varepsilon^2} \leq \delta \qquad \Rightarrow n \geq \frac{\sigma^2}{\delta \varepsilon^2}$$

Per via però di CHEBYSHEV questa disuguaglianza è molto permissiva. n può risultare tt grande di quello realmente necessario:

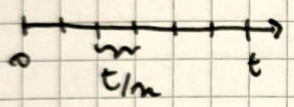
$$\frac{\cancel{\sigma^2}}{\cancel{\varepsilon^2}} \Phi^{-1}\left(1 - \frac{\delta}{2}\right)^2 \leq \frac{\cancel{\sigma^2}}{\delta \cancel{\varepsilon^2}} \Rightarrow \Phi^{-1}\left(1 - \frac{\delta}{2}\right) \leq \frac{1}{\delta} \quad [\dots]$$

T tempo

$N(t)$ numero di ingressi fino a t istanti dopo l'istante 0. Se gli ingressi sono casuali, allora è una var. alea.

<u>PROCESSO STOCASTICO di Poisson</u> Ipotesi:

1. $N(0)=0$ tempo in cui si osserva il fenomeno
2. Intervalli temporali disgiunti $\Rightarrow$ indipendenza tra le relative var. aleatorie
3. distribuzione # eventi dipende solo dalla lunghezza dell'intervallo (e non da dove iniziano)
4. $\dfrac{P(N(h)=1)}{h} \xrightarrow{h\to 0} \lambda$ Se h è piccolo la $P(N(h)=1)=\lambda h$ dove h è un intervallo
5. $\dfrac{P(N(h)\geq 2)}{h} \xrightarrow{h\to 0} 0$ $P(N(h)\geq 2)=0$ per h piccolo



t suddiviso in n intervalli equiampi

$N(t)=k \iff \underbrace{\left\{\begin{matrix}\text{ogni avvenimento in}\\ \text{un diverso sottoinsieme}\end{matrix}\right\}}_{A} \cup \underbrace{\left\{\begin{matrix}\text{tutti gli altri}\\ \text{casi possibili}\end{matrix}\right\}}_{B}$ A e B sono disgiunti: $A\cap B=\emptyset$

a garantire da un certo n, i k eventi stanno in intervalli separati

$$P(N(t)=k)=P(A)+\underbrace{P(B)}_{\substack{=0 \text{ per la 5 e dato}\\ \text{che per } n \text{ grande}\\ \text{gli intervalli sono piccoli}}}=\left(\lambda\frac{t}{n}\right)^k\left(1-\lambda\frac{t}{n}\right)^{n-k}\binom{n}{k}$$

$N(t)\sim B\left(n,\lambda\frac{t}{n}\right)$

$N(t)\sim P(\lambda t)$ (per n suff. grande)

X_1 = tempo primo avvenimento X_2 = tempo tra primo avv. e secondo

$$P(X_1>t)=P(N(t)=0)=e^{-\lambda t}\frac{(\lambda t)^0}{0!}=e^{-\lambda t}$$

$$F_{X_1}(t)=1-P(X_1>t)=1-e^{-\lambda t} \Rightarrow X_1\sim E(\lambda)$$

$$P(X_2>t \mid X_1=s)=P(\text{nessun evento in } (s,s+t] \text{ dato che } X_1=s)\overset{\text{Per la 2}}{=}$$

$$=P(\text{nessun evento in } (s,s+t])=P(N(t)=0) \Rightarrow X_2\sim E(\lambda) \text{ e così via}$$

$\boxed{X_i \text{ sono indipendenti}}$ dimostrabile

METODO MAX. VERO SIMIGLIANZA

~~X~~ popolazione $X \sim D(\theta)$

↳ modello individuato da θ, che non si conosce

$X_1, \dots, X_n$ campione

Per ind. delle X_i

FUNZ. DENSITÀ CONGIUNTA

Funzione di verosimiglianza: $f_{X_1,\dots,X_n}(x_1,\dots,x_n;\theta) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta) =$ ident. distribuite

(Servono per stimare il parametro)

↳ DIPENDE DAL PARAMETRO

$$= \prod_{i=1}^{n} f_X(x_i;\theta) \quad (*) \quad := \ell(\theta)$$

↳ solo dopo aver osservato il campione

Se osserviamo le x_i (ossia il campione), in $(*)$ rimane solo θ come incognita.

Si usa il θ che massimizza quel valore.

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta) \quad \underline{\text{OPPURE}} \quad \arg\max_{\theta} \ln \ell(\theta)$$

$X \sim B(p)$

→ MASSA DI PROB (ABUSO NOTAZIONE)

$$\ell(p) = \prod_{i=1}^{n} f_X(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \underbrace{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}}_{(*)}$$

Trovare p che massimizza $(*)$

$$\ln \ell(p) = \ln p^{\sum_i x_i} + \ln (1-p)^{n-\sum_i x_i} = \left(\sum_i x_i\right)\ln p + \left(n - \sum_i x_i\right)\ln(1-p)$$

~~[illegible]~~

$$\frac{d}{dp}\ln \ell(p) = \frac{\sum_i x_i}{p} - \frac{n-\sum_i x_i}{1-p}$$

$$\frac{d}{dp}\ln \ell(\hat{p}) = 0 \quad \text{QUANDO} \quad \frac{\sum_i x_i}{\hat{p}} = \frac{n-\sum_i x_i}{1-\hat{p}}$$

(unica poiché $\ln$ è convessa)

$$(1-\hat{p})\sum_i x_i = n\hat{p} - \hat{p}\sum_i x_i$$

$$\sum_i x_i - \cancel{\hat{p}\sum_i x_i} = n\hat{p} - \cancel{\hat{p}\sum_i x_i}$$

$$\boxed{\hat{p} = \frac{1}{n}\sum_i x_i = \bar{x}}$$

Per Bernoulli, lo stimatore di max. verosimiglianza coincide con la media campionaria

Vale anche per ~~B~~ la poisson.

~~Esp~~ ESPONENZIALE: $\ell(\lambda) = \prod_{i=1}^{n} f_X(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \prod_i e^{-\lambda x_i} =$

$$= \lambda^n e^{-\lambda \sum_i x_i}$$

$$\ln \ell(\lambda) = n \ln \lambda - \lambda \sum_i x_i \qquad \frac{d}{d\lambda} \ln \ell(\lambda) = \frac{n}{\lambda} - \sum_i x_i$$

$$\frac{d}{d\lambda} \ln \ell(\hat{\lambda}) = 0 \qquad \frac{n}{\hat{\lambda}} = \sum_i x_i \qquad \hat{\lambda} = \frac{n}{\sum_i x_i} = \frac{1}{\bar{x}}$$

(visto a inizio lezione con il metodo PLUG IN)

→ FUNZ. DI MASSIMA DI PROB.

POISSON: $X \sim P(\lambda) \qquad f_X(x;\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$

$$\ell(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \lambda^{\sum_i x_i} \prod_{i=1}^{n} \frac{1}{x_i!}$$

$$\ln \ell(\lambda) = -n\lambda + \sum_i x_i \ln \lambda + \sum_{i=1}^{n} \ln \frac{1}{x_i!}$$

$$\frac{d}{d\lambda} \ln \ell(\lambda) = -n + \frac{\sum_i x_i}{\lambda}$$

$$\frac{d}{d\lambda} \ln \ell(\hat{\lambda}) = 0 \qquad n = \frac{\sum_i x_i}{\hat{\lambda}} \qquad \hat{\lambda} = \frac{\sum_i x_i}{n} = \bar{x}$$

Si può fare per la normale senza conoscere μ e σ. Servono le derivate parziali.

FINE TEORIA