Tema di Statistica Matematica

Sessione estiva

12 luglio 2019

1) Siano $(X_1, X_2, \ldots, X_{n_1})$ e $(Y_1, Y_2, \ldots, Y_{n_2})$ due campioni casuali indipendenti da due distribuzioni di Poisson rispettivamente di parametri μ_1 e $\mu_2 = \mu_1 + \mu$.

a) Trovare lo stimatore di massima verosimiglianza del parametro μ.

b) Trovare la distribuzione asintotica dello stimatore di cui al punto a) e costruire un intervallo di confidenza per μ di livello $(1 - \alpha)$.

2) Sia $(X_1, X_2, \ldots, X_n, \ldots)$ un campione casuale proveniente da una distribuzione $F(x)$ che ammette varianza finita e sia $\{S_n^2\}_{n \in \mathbb{N}}$ la successione delle varianze campionarie. Dimostrare che la successione degli scarti quadratici medi $\{S_n\}_{n \in \mathbb{N}}$ converge in probabilità a σ, scarto quadratico medio della popolazione. Dal punto di vista inferenziale, che cosa possiamo dedurre da tutto ció?

3) Siano $(X_1, X_2, \ldots, X_n, \ldots)$ v.c. indipendenti con densità

$$f(x_i; \theta) = \begin{cases} \frac{1}{2i\theta} & -i\,(\theta - 1) < x_i < i\,(\theta + 1) \\ 0 & \text{altrimenti} \end{cases}$$

con $\theta > 0$. Trovare una statistica sufficiente per θ.

4) Le sintesi che seguono si riferiscono a un campione casuale semplice di 150 clienti di una banca che nel 2018 sono risultati insolventi non avendo restituito il prestito a essi accordato nei tempi stabiliti. Indicato con X il credito non recuperato (espresso in migliaia di euro) si è osservato:

$$\sum_{i=1}^{150} x_i = 6232.40 \quad \text{e} \quad \sum_{i=1}^{150} x_i^2 = 402709.08$$

a) Costruire un intervallo di confidenza di livello 99% per il valor medio del credito non recuperato.

b) Nel biennio 2015-2016 il credito mediamente non recuperato è stato di 45.25 migliaia di euro e nel 2017 la banca ha provveduto a modificare il sistema di valutazione del rischio di credito. Si può affermare, a un livello di significatività del 5%, che il nuovo sistema di valutazione del rischio di credito introdotto abbia ridotto l?ammontare medio del credito non recuperato? Commentare il procedimento di costruzione della regola di decisione e il risultato ottenuto.

5) Supponiamo che le v.c. $Y_1, Y_2, \ldots, Y_n$ soddisfino la relazione

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

con $x_1, x_2, \ldots, x_n$ costanti fissate e $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ v.c. i.i.d. secondo $N(0, \sigma^2)$ con σ^2 quantità non nota. Trovare

a) una statistica congiuntamente sufficiente per (β, σ^2);

b) lo stimatore a massima verosimiglianza di β e dimostrare che esso è non distorto per β;

c) la distribuzione dello stimatore di massima verosimiglianza di β

Inoltre,

d) dimostrare che

$$V_n = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}$$

è anch'esso uno stimatore non distorto per β;

e) calcolare la varianza esatta di V_n e confrontarla con quella dello stimatore di massima verosimiglianza di β poc'anzi trovata. Che conclusioni ne traete?

Es. 1)

$$X_1, X_2, \ldots, X_{n_1} \overset{IID}{\sim} \mathcal{P}(\mu_1)$$

$$Y_1, Y_2, \ldots, Y_{n_2} \overset{IID}{\sim} \mathcal{P}(\mu_2) \quad \text{con} \quad \mu_2 = \mu_1 + \mu$$

a) In virtù dell' INDIPENDENZA dei due campioni casuali,

$$L(\mu_1, \mu_2 \mid \underline{x}, \underline{y}) = \prod_{i=1}^{n_1} \frac{e^{-\mu_1} \mu_1^{x_i}}{x_i!} \cdot \prod_{i=1}^{n_2} \frac{e^{-\mu_2} \mu_2^{y_i}}{y_i!}$$

$$= \frac{e^{-n_1 \mu_1} \mu_1^{\sum_{i=1}^{n_1} x_i}}{\prod_{i=1}^{n_1} x_i!} \cdot \frac{e^{-n_2 \mu_2} \mu_2^{\sum_{i=1}^{n_2} y_i}}{\prod_{i=1}^{n_2} y_i!}$$

$$\propto e^{-n_1 \mu_1} \mu_1^{\sum_{i=1}^{n_1} x_i} \cdot e^{-n_2 \mu_2} \mu_2^{\sum_{i=1}^{n_2} y_i}$$

e

$$l(\mu_1, \mu_2 \mid \underline{x}, \underline{y}) = \ln L(\mu_1, \mu_2 \mid \underline{x}, \underline{y})$$

$$\propto -n_1 \mu_1 + \sum_{i=1}^{n_1} x_i \ln(\mu_1) - n_2 \mu_2 + \sum_{i=1}^{n_2} y_i \ln(\mu_2)$$

sicché

$$\begin{cases} \dfrac{d}{d\mu_1} l(\cdot) = -n_1 + \dfrac{1}{\mu_1} \displaystyle\sum_{i=1}^{n_1} x_i \\[2ex] \dfrac{d}{d\mu_2} l(\cdot) = -n_2 + \dfrac{1}{\mu_2} \displaystyle\sum_{i=1}^{n_2} y_i \end{cases}$$

e pertanto il SISTEMA di EQUAZIONI la cui SOLUZIONE restituisce gli STIMATORI di MV di μ_1 e μ_2 è dato da

$$\begin{cases} -n_1 + \dfrac{1}{\mu_1} \displaystyle\sum_{i=1}^{n_1} x_i = 0 \\ -n_2 + \dfrac{1}{\mu_2} \displaystyle\sum_{i=1}^{n_2} y_i = 0 \end{cases}$$

e dunque

$$(\hat{\mu}_1, \hat{\mu}_2) = (\bar{X}_{n_1}, \bar{Y}_{n_2})$$

Ora

$$\mu_1 - \mu_2 = \mu_1 - (\mu_1 + \mu) = -\mu$$

sicché

$$\mu = \mu_2 - \mu_1 \equiv g(\mu_1, \mu_2)$$

e per la PROPRIETÀ di INVARIANZA degli STIMATORI di MV si ha

$$\hat{\mu} = g(\hat{\mu}_1, \hat{\mu}_2) = \bar{Y}_{n_2} - \bar{X}_{n_1}$$

b) Ora

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}[\bar{Y}_{n_2} - \bar{X}_{n_1}] = \mu_2 - \mu_1 = \mu$$

$$Var(\hat{\mu}) = Var(\bar{Y}_{n_2} - \bar{X}_{n_1}) = Var(\bar{Y}_{n_2}) + Var(\bar{X}_{n_1}) = \frac{\mu_2}{n_2} - \frac{\mu_1}{n_1}$$

poiché $\bar{Y}_{n_2}$ e $\bar{Y}_{n_1}$ sono v.c. INDIPENDENTI in virtù del fatto che sono calcolate su CAMPIONI INDIPENDENTI sicché

$$Cov(\bar{Y}_{n_2} - \bar{X}_{n_1}) = 0$$

Essendo $\hat{\mu}$ stimatore di MV di μ, per le proprietà che caratterizzano questa classe di stimatori

$$\hat{\mu} \underset{a}{\sim} N\left(\mu, \frac{n_1 \mu_2 + n_2 \mu_1}{n_1 \cdot n_2}\right)$$

e stimata la varianza di μ con

$$\frac{n_1 \bar{y}_{n_2} + n_2 \bar{x}_{n_1}}{n_1 \cdot n_2}$$

possiamo costruire l'intervallo di confidenza per μ di livello fissato $(1-\alpha)$

$$IC_{\mu}(1-\alpha) = \left[(\bar{y}_{n_2} - \bar{x}_{n_1}) - z_{1-\alpha/2} \sqrt{\frac{n_1 \bar{y}_{n_2} + n_2 \bar{x}_{n_1}}{n_1 n_2}} \, , \, (\bar{y}_{n_2} - \bar{x}_{n_1}) + z_{1-\alpha/2} \sqrt{\frac{n_1 \bar{y}_{n_2} + n_2 \bar{x}_{n_1}}{n_1 \cdot n_2}} \right]$$

Es.2)

È sufficiente ricordare il TEOREMA del CONTINUOUS MAPPING

Thm. Sia $\{X_n\}_{n\in\mathbb{N}}$ una succ.ne di v.c. che converge in probabilità alla v.c. X e sia g una funzione continua. Allora

$$\{h(X_n)\}_{n\in\mathbb{N}} \xrightarrow{P} h(X)$$

Ora, la varianza campionaria S_n^2 è STIMATORE CONSISTENTE della VARIANZA della POPOLAZIONE (lo si può dimostrare ricordando che

$$Var(S_n^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$$

e tenuto conto della NON DISTORSIONE di S_n^2, concludere per la CONVERGENZA QUADRATICA di S_n^2 a σ^2 da cui la CONVERGENZA DEBOLE o in PROBABILITÀ)

Riconducendo al teorema poc'anzi enunciato posto

$$S_n = h(S_n^2) = \sqrt{S_n^2}$$

si ottiene il RISULTATO cercato.

EX. 3)

Tenuto conto dell'INDIPENDENZA di $X_1, X_2, \dots$ possiamo scrivere

$$L(\theta;\underline{x}) = \prod_{i=1}^{n} \frac{1}{2i\theta} \mathbb{1}_{(-i(\theta-1),\, i(\theta+1))}(x_i)$$

$$= \prod_{i=1}^{n} \frac{1}{2i\theta} \mathbb{1}_{(-(\theta-1),\,(\theta+1))}\left(\frac{x_i}{i}\right)$$

$$= \left(\frac{1}{2\theta}\right)^{n} \prod_{i=1}^{n} \frac{1}{i} \mathbb{1}_{(-(\theta-1),\,(\theta+1))}\left(\frac{x_i}{i}\right)$$

$$= \left(\frac{1}{2\theta}\right)^{n} \left(\prod_{i=1}^{n} \frac{1}{i}\right) \cdot \mathbb{1}_{(-(\theta-1),\,(\theta+1))}\left(\min\left(\frac{x_i}{i}\right)\right) \cdot$$

$$\cdot\, \mathbb{1}_{(-(\theta-1),\,(\theta+1))}\left(\max\left(\frac{x_i}{i}\right)\right)$$

sicché, per il TEOREMA di FATTORIZZAZIONE

$$\left(\min\left(\frac{X_i}{i}\right), \max\left(\frac{X_i}{i}\right)\right)$$

è STATISTICA SUFFICIENTE (minimale) per θ.

Ex. 4)

Dai dati a nostra disposizione possiamo ricavare

$n = 150$

$\bar{x}_n = 6232,40 / 150 = 41,55$

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \cdot \bar{x}_n^2 = \frac{1}{149} \cdot 402709,08 - \frac{150}{149} \cdot 41,55^2$$

$$= 964.81$$

$s_n = \sqrt{964,81} = 31,06$

a) Essendo $n > 30$, possiamo sfruttare i risultati asintotici per costruire un INTERVALLO di CONFIDENZA di livello (approssimato) 0.99 per il valor medio μ del CREDITO NON RECUPERATO. Sicché a partire dal fatto che

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

sicché

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

che, stimata σ con s_n, restituisce la STATISTICA PIVOT tramite la quale costruire l'INTERVALLO di CONFIDENZA per μ. Nel dettaglio,

$$IC_\mu(0.99) = \left[\bar{x}_n + z_{0.005} \frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{0.995} \frac{s_n}{\sqrt{n}}\right]$$

$$= [35,02 , 48,08]$$

con $z_{0.005} = -z_{0.995} = -2,58$

b) Il quesito inferenziale può essere espresso tramite il seguente SISTEMA di IPOTESI

$$\begin{cases} H_0: \mu \geq 45.25 \\ vs \\ H_1: \mu < 45.25 \end{cases}$$

dove l'ipotesi alternativa esprime quanto interessa verificare ossia il fatto che il nuovo sistema di valutazione abbia RIDOTTO l'ammontare medio del CREDITO NON RECUPERATO.

Da un punto di vista intuitivo è ragionevole adottare una REGOLA di DECISIONE che RIFIUTI H_0 per un valore di $\bar{x}_n$ significativamente più PICCOLO di $\mu_0 = 45.25$.

Sulla base del ragionamento sviluppato al punto a) che porta a

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \simeq \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \sim N(0,1)$$

tenuto conto del fatto che S_n è stimatore CONSISTENTE di σ; sicché sotto H_0

$$\frac{\bar{X}_n - \mu_0}{s_n/\sqrt{n}} \sim N(0,1)$$

che, fissato $\alpha = 0{,}05$, porta alla seguente REGIONE CRITICA

$$C_{0.05} = \left\{ \underline{x} \in \mathcal{X} : \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}} < -z_{0.05} \right\}$$

e dunque, essendo

$$\frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}} = \frac{41.55 - 45.25}{31{,}06 / \sqrt{150}} = -1{,}46 > -1{,}645 = z_{0.05}$$

sicché H_0 NON PUÒ ESSERE RIFIUTATA sulla base della INFORMAZIONE CONTENUTA nei dati a disposizione portando così a concludere che il NUOVO SISTEMA di VALUTAZIONE del RISCHIO di CREDITO NON HA PRODOTTO l'EFFETTO sperato.

IX.5

OMISSIS (argomento non trattato in questo anno accademico)