Tema di Statistica Matematica

Primo appello

23 gennaio 2019

1) Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale da una distribuzione Uniforme su $(0, \theta)$ e sia dato lo stimatore $\hat{\theta}_k = k \cdot \max(X_1, X_2, \ldots, X_n)$ per $k > 1$. Trovare

a) l'errore quadratico medio di $\hat{\theta}_k$;

b) il valore di k per il quale $\hat{\theta}_k$ é non distorto per θ;

c) il valore di k che minimizza l'errore quadratico medio e commentare il risultato ottenuto.

Dimostrare, infine, che il valore $k = 1$ restituisce lo stimatore di massima verosimiglianza $\hat{\theta}_n$ di θ e confrontare quest'ultimo stimatore con quello ottenuto al punto c).

2) Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale proveniente da una distribuzione la cui densitá sia data da

$$f_X(x; \theta) = \theta\, x^{\theta - 1}\, \mathbb{I}_{(0,1)}(x), \quad \theta > 0.$$

Esiste una funzione di θ per la quale esiste uno stimatore non distorto la cui varianza raggiunge il limite inferiore di Rao-Cramér? Se é cosí, individuarla; se no, dimostrare perché no.

3) Sia $(x_1, x_2, \ldots, x_n)$ la determinazione di un campione casuale proveniente da una distribuzione continua con densitá

$$f_X(x; \theta) = \theta\ (x + 1)^{-(\theta+1)}\ \mathbb{I}_{[0,\infty)}(x)$$

con $\theta > 0$ parametro ignoto. Indicata con F_X la funzione di distribuzione di X,

a) dimostrare che la variabile $V = -2\ln(1 - F_X(X))$ ha distribuzione chi-quadrato con $\nu = 2$ gradi di libertá;

b) usando il risultato ottenuto al precedente punto a), ricavare una statistica pivot per l'inferenza su θ e costruire un intervallo di confidenza per il parametro θ di livello (esatto) $(1 - \alpha) = 0.95$;

c) stabilire se lo stimatore di massima verosimiglianza di θ é o meno non distorto per θ;

(**<u>Sugg.</u>**: *disuguaglianza di Jensen*: $\mathbb{E}(g(X)) \geq g(E(X)$

d) ricavare la regione critica piú potente di livello $\alpha = 0.05$ per il problema della verifica del sistema di ipotesi $H_0 : \theta = \theta_0$ *vs.* $H_1 : \theta > \theta_0$.

EX. 1)

a) $$MSE_\theta(\hat{\theta}_k) = \left[B_\theta(\hat{\theta}_k)\right]^2 + Var_\theta(\hat{\theta}_k)$$

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

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

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

$$= \frac{\theta^2}{(n+1)^2(n+2)}\left[(n^2k^2 - 2(n+1)nk + (n+1)^2)((n+2) + nk^2\right]$$

$$= \frac{\theta^2}{(n+1)^2(n+2)}\left[(n+(n+2)\cdot n^2)k^2 - 2n(n+1)(n+2)k + (n+1)^2(n+2)\right]$$

b) $$E_\theta(\hat{\theta}_k) = \theta \Rightarrow k = \frac{n+1}{n}$$ dal momento che

$$E_\theta(X_{(n)}) = \frac{n}{n+1}\theta$$ dove

$$X_{(n)} = \max(X_1, \dots, X_n)$$

c) Come si può vedere al punto a) l'ERRORE QUADRATICO MEDIO di $\hat{\theta}_k$ visto come funzione di k è dato da

$$\theta^2\left[\frac{n+(n+2)n^2}{(n+1)^2(n+2)}k^2 - \frac{2n(n+1)(n+2)}{(n+1)^2(n+2)}k + \frac{(n+1)^2(n+2)}{(n+1)^2(n+2)}\right]$$

siccché

$$\frac{d}{dk} MSE_\theta(\hat{\theta}_k) = \theta^2 \left[\frac{2(n+(n+2)\cdot n^2)}{(n+1)^2(n+2)} k - \frac{2n(n+1)(n+2)}{(n+1)^2(n+2)} \right] = 0$$

e dunque, con qualche passaggio, si ottiene

$$k = \frac{n+2}{n+1}$$

Inoltre

$$\frac{d^2}{dk^2} MSE_\theta(\hat{\theta}_k) = \frac{2(n+(n+2)n^2)}{(n+1)^2(n+2)} > 0$$

e dunque $k = \frac{n+2}{n+1}$ è PUNTO di MINIMO.

Commento: dai calcoli appena fatti emerge che lo stimatore a MINIMO ERRORE QUADRATICO MEDIO in questa famiglia di distribuzioni è DISTORTO se confrontato con $k = \frac{n+1}{n}$ per lo STIMATORE (NON DISTORTO) di MASSIMA VEROSIMIGLIANZA al pto b) (opportunamente corretto)

d) $k=1$.

La funzione di VEROSIMIGLIANZA per la famiglia di distribuzioni in questione è data da

$$L(\theta; \underline{x}) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{(0,\theta)}(x_i)$$

Questa funzione risulta massima in corrispondenza del valore di θ PIÙ PICCOLO POSSIBILE ma NON PIÙ PICCOLO del massimo campionario altrimenti uno degli INDICATORI sarà ZERO.

Perciò

$$\hat{\theta}_n = X_{(n)} = \max(X_1, X_2, \dots, X_n)$$

Commento:

i) $\hat{\theta}_n$ NON È STIMATORE NON DISTORTO DI θ perché

$$\mathbb{E}_\theta(\hat{\theta}_n) \neq \theta$$

ii) $\hat{\theta}_n$ NON È LO STIMATORE A MINIMO ERRORE QUADRATICO MEDIO.

$\underline{\mathrm{X}}$.2)

In questo caso

$$L(\theta;\underline{x}) = \prod_{i=1}^{n} \theta\, x_i^{\theta-1} \mathbb{1}_{(0,1)}(x_i) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1} \cdot \prod_{i=1}^{n} \mathbb{1}_{(0,1)}(x_i)$$

sicché

$$\ell(\theta;\underline{x}) = n\ln(\theta) + (\theta-1)\sum_{i=1}^{n}\ln(x_i) + \ln\prod_{i=1}^{n}\mathbb{1}_{(0,1)}(x_i)$$

La FUNZIONE SCORE è data da

$$\frac{d}{d\theta}\,\ell(\theta;\underline{x}) = \frac{n}{\theta} + \sum_{i=1}^{n}\ln(x_i) \qquad \equiv S(\theta;\underline{x})$$

e, ricordando che $\mathbb{E}_\theta(S(\theta;\underline{X}))=0$ si ha

$$\mathbb{E}_\theta\left(\frac{n}{\theta} + \sum_{i=1}^{n}\ln(X_i)\right) = \frac{n}{\theta} + \mathbb{E}_\theta\left(\sum_{i=1}^{n}\ln(X_i)\right) = 0$$

sicché

$$\mathbb{E}_\theta\left(-\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)\right) = \frac{1}{\theta}$$

da cui segue che $T_n = -\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ è uno stimatore NON DISTORTO di $g(\theta) = \frac{1}{\theta}$.

$\underline{\text{Nota}}$: è immediato osservare che $T_n = -\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ è STIMATORE di MASSIMA VEROSIMIGLIANZA di $g(\theta)$ sicché

$$\hat{\theta}_n = -\frac{n}{\sum_{i=1}^{n}\ln(X_i)}$$

è STIMATORE di MASSIMA VEROSIMIGLIANZA di θ.

Ora essendo

$$T_n = -\frac{1}{n} \sum_{i=1}^{n} \ln(X_i)$$

stimatore NON DISTORTO di $g(\theta) = \frac{1}{\theta}$ e FUNZIONE di STATISTICA SUFFICIENTE MINIMALE di θ, è anche STIMATORE UMVU di $g(\theta) = \frac{1}{\theta}$ in virtù del COROLLARIO del TEOREMA di RAO-BLACKWELL.

Verifichiamo ora se T_n è anche STIMATORE EFFICIENTE per $g(\theta) = \frac{1}{\theta}$.

Per far ciò, calcoliamo la VARIANZA di T_n e confrontiamola con il LIMITE INFERIORE di RAO-CRAMER per la VARIANZA di un qualsiasi stimatore NON DISTORTO di $g(\theta)$.

Cominciamo con porre

$$Y_i = -\ln(X_i) \Rightarrow X_i = e^{-Y_i}, \left|\frac{d}{dY_i} X_i\right| = e^{-Y_i}$$

sicché

$$f_{Y_i}(y;\theta) = \theta\, e^{-y(\theta-1)}\, e^{-y}\, \mathbb{1}_{[0,+\infty)}(y)$$
$$= \theta\, e^{-\theta y}\, \mathbb{1}_{[0,+\infty)}(y)$$

ossia $Y_i \sim Exp(\theta)$ sicché $E_\theta(Y_i) = \frac{1}{\theta}$ e $Var_\theta(Y_i) = \frac{1}{\theta^2}$

Dunque per la PROPRIETÀ di RIPRODUCIBILITÀ

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

sicché $Var_\theta(T_n) = \frac{1}{n^2} \sum_{i=1}^{n} Var_\theta(Y_i) = \frac{1}{n^2} \frac{n}{\theta^2} = \frac{1}{n\theta^2}$

Ora, tenuto conto del fatto che il LIMITE INFERIORE per la VARIANZA di uno STIMATORE NON DISTORTO di una FUNZIONE $g(\theta)$ del parametro θ è dato da

$$Var_\theta(T_n) \geq \frac{[g'(\theta)]^2}{I_n(\theta)}$$

ed essendo

$$[g'(\theta)]^2 = \left[-\frac{1}{\theta^2}\right]^2 = \frac{1}{\theta^4}$$

$$I_n(\theta) = -E_\theta\left[\frac{d}{d\theta} S(\theta; \underline{X})\right] = -E\left[-\frac{n}{\theta^2}\right] = \frac{n}{\theta^2}$$

si ha

$$Var_\theta(T_n) \geq \frac{\theta^2}{n} \cdot \frac{1}{\theta^4} = \frac{1}{n\theta^2}$$

siccché

$$Var_\theta(T_n) = \frac{1}{n\theta^2}$$

con $T_n = -\frac{1}{n}\sum_{i=1}^{n} \ln(X_i)$, da cui la conclusione che T_n è STIMATORE EFFICIENTE di $g(\theta) = \frac{1}{\theta}$.

EX. 3)

a) Si ha

$$P(V \leq v) = P(-2\ln(1-F_X(X)) \leq v)$$
$$= P(F_X(X) \leq 1 - e^{-\frac{1}{2}v})$$
$$= 1 - e^{-\frac{1}{2}v}$$

essendo, come è noto, $F_X(X) \sim U(0,1)$. Quindi V ha una distribuzione ESPONENZIALE di parametro $\theta = 2$ ovvero

$$V \sim Exp(2) \equiv \mathcal{G}(1,2) \equiv \chi^2_2$$

ricordando che ESPONENZIALE e CHI-QUADRATO sono casi PARTICOLARI di GAMMA.

b) In base al risultato di cui al punto a) si ha che

$$-2\sum_{i=1}^{n} \ln(1-F_X(x_i)) \sim \mathcal{G}(n,2) \equiv \chi^2_{2n}$$

D'altra parte

$$F_X(x) = \int_0^x \theta(t+1)^{-(\theta+1)}\,dt = \int_1^{x+1} w^{-(\theta+1)}\,dw = 1-(x+1)^{-\theta}$$

Quindi

$$-2\sum_{i=1}^{n} \ln(1-F(x_i)) = -2\sum_{i=1}^{n} \ln(x_i+1)^{-\theta}$$
$$= 2\theta\sum_{i=1}^{n} \ln(x_i+1) \sim \chi^2_{2n}$$

è la STATISTICA PIVOT CERCATA per θ.

Portanto l' INTERVALLO di CONFIDENZA ESATTO per θ di livello 0.95 è dato da

$$IC_\theta(0.95) = \left[\frac{\chi^2_{2n;0.025}}{2\sum_{i=1}^{n} \ln(x_i+1)} , \frac{\chi^2_{2n;0.975}}{2\sum_{i=1}^{n} \ln(x_i+1)} \right]$$

dove $\chi^2_{2n;\alpha}$ indica il QUANTILE di ordine α della DISTRIBUZIONE CHI-QUADRATO con $2n$ gradi di libertà.

c) La funzione di VEROSIMIGLIANZA è data da

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

sicché

$$l(\theta;\underline{x}) = \ln L(\theta;\underline{x}) = n\ln(\theta) - (\theta+1)\sum_{i=1}^{n} \ln(x_i+1) + \ln\left(\prod_{i=1}^{n} \mathbb{1}_{\mathbb{R}^+}(x_i)\right),$$

e dunque

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} \ln(x_i+1)}$$

è lo STIMATORE di MASSIMA VEROSIMIGLIANZA di θ $\left[\frac{d^2}{d\theta} l(\theta;\underline{x}) < 0\right]$

Indicata con $S(\theta;\underline{x})$ la FUNZIONE SCORE ossia

$$S(\theta;\underline{x}) = \frac{d}{d\theta} l(\theta;\underline{x}) = \frac{n}{\theta} - \sum_{i=1}^{n} \ln(x_i+1)$$

e ricordando che

$$\mathbb{E}_\theta(S(\theta;\underline{X})) = 0$$

si ha che

$$\mathbb{E}_\theta\left[\sum_{i=1}^{n} \ln(X_i+1)\right] = \frac{n}{\theta}$$

Utilizzando la DISUGUAGLIANZA di JENSEN

$$\mathbb{E}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta\left[\frac{n}{\sum_{i=1}^{n} \ln(X_i+1)}\right] > \frac{n}{\mathbb{E}_\theta\left[\sum_{i=1}^{n} \ln(X_i+1)\right]} = \theta$$

e pertanto lo STIMATORE di MASSIMA VEROSIMIGLIANZA risulta essere DISTORTO per θ.

<u>Nota</u>: DISUGUAGLIANZA di JENSEN

$$\mathbb{E}_\theta(\varphi(X)) \geq \varphi(\mathbb{E}_\theta(X))$$

dove l'UGUAGLIANZA VALE se e solo se $\varphi(X)$ è FUNZIONE LINEARE di X.

d) È facile verificare che la famiglia parametrica considerata costituisce una famiglia ESPONENZIALE (regolare) a $k=1$ parametri con STATISTICA NATURALE (SUFFICIENTE, MINIMALE e COMPLETA)

$$T_n(X_1, \dots, X_n) = \sum_{i=1}^{n} \ln(X_i+1)$$

Pertanto, il test ottimo per il problema di verifica di ipotesi considerato ha regione critica C_α basata su $T_n(X_1, \dots, X_n)$ o opportuna trasformata.

Sappiamo che sotto H_0

$$2\theta_0 T_n(\underline{X}) \sim \chi^2_{2n}$$

e inoltre che

$$T_n(\underline{X}) = \frac{n}{\hat{\theta}_n}$$

e GRANDI VALORI della STIMA di MASSIMA VEROSIMIGLIANZA sono, evidentemente, CONTRARI all' IPOTESI NULLA.

Ne segue che la REGIONE CRITICA PIÙ POTENTE cercata è

$$C_\alpha = \left\{ \underline{x} \in \mathcal{X} : 2\theta_0 T_n(\underline{x}) < \chi^2_{2n;\alpha} \right\}$$

Con α LIVELLO di SIGNIFICATIVITÀ del TEST, e, nel nostro caso fissato in $\alpha = 0{,}05$.

Es. 4)

a) Si può osservare che

$$\tau = P_\theta(X \le 1) = P_\theta(X=0) + P_\theta(X=1) = \theta + \theta(1-\theta) = 2\theta - \theta^2$$

e perciò $\tau = g(\theta)$ con $g'(\theta) = 2 - 2\theta > 0$ per ogni valore di $\theta \in (0,1)$. Quindi τ è FUNZIONE MONOTONA CRESCENTE di θ.

D'altro canto, la funzione di VEROSIMIGLIANZA è

$$L(\theta; \underline{x}) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum_{i=1}^n x_i}$$

e quindi

$$\ell(\theta; \underline{x}) = \ln L(\theta; \underline{x}) = n \ln(\theta) + \ln(1-\theta) \cdot \sum_{i=1}^n x_i$$

Di conseguenza

$$S(\theta; \underline{x}) = \frac{d}{d\theta} \ell(\theta; \underline{x}) = \frac{n}{\theta} - \frac{1}{1-\theta} \sum_{i=1}^n x_i$$

da cui, uguagliando a zero la precedente e risolvendo rispetto θ si ha che lo STIMATORE di MV di θ è dato da

$$\hat{\theta}_n = \frac{n}{n + \sum_{i=1}^n X_i} = \frac{1}{1 + \bar{X}_n}$$

e inoltre

$$I_n(\theta) = -E_\theta\left[\frac{d^2}{d\theta^2} \ell(\theta; \underline{X})\right] = -E_\theta\left[-\frac{n}{\theta^2} - \frac{1}{(1-\theta)^2} \sum_{i=1}^n X_i\right]$$

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

dato che $E(X) = \frac{1-\theta}{\theta}$.

Ricordando che $\hat{\Theta}_n$ è lo STIMATORE di MV di θ, per le sue note proprietà si ha

$$(\hat{\Theta}_n - \theta) \underset{n}{\sim} N\left(0, \frac{1}{I_n(\theta)}\right) \equiv N\left(0, \frac{\theta^2(1-\theta)}{n}\right)$$

sicché

$$\left[\hat{\Theta}_n - 1.96\sqrt{\frac{\theta^2(1-\theta)}{n}}, \quad \hat{\Theta}_n + 1.96\sqrt{\frac{\theta^2(1-\theta)}{n}}\right]$$

costituisce un INTERVALLO di CONFIDENZA APPROSSIMATO per θ di livello 0,95.

Ora in virtù del METODO DELTA

$$g(\hat{\Theta}_n) \underset{n}{\sim} N\left(g(\theta), (g'(\theta))^2 I_n^{-1}(\theta)\right)$$

con $g(\theta) = 2\theta - \theta^2$ e $\underbrace{g'(\theta)\, I_n^{-1}(\theta)}_{\substack{\| \\ Var_\theta(g(\hat{\Theta}_n))}} = 4 \cdot \dfrac{\theta^2(1-\theta)^3}{n}$,

da cui

$$\frac{g(\hat{\Theta}_n) - g(\theta)}{\sqrt{Var_\theta(g(\hat{\Theta}_n))}} \underset{n}{\sim} N(0,1)$$

sicché

$$IC_{g(\theta)}(1-\alpha) = \left[g(\hat{\Theta}_n) - z_{1-\alpha/2}\sqrt{\hat{Var}_\theta(g(\hat{\Theta}_n))}, \quad g(\hat{\Theta}_n) + z_{1-\alpha/2}\sqrt{\hat{Var}_\theta(g(\hat{\Theta}_n))}\right]$$

dove

$$g(\hat{\Theta}_n) = 2\hat{\Theta}_n - \hat{\Theta}_n^2 \quad \text{e} \quad \hat{Var}_\theta(\hat{\Theta}_n) = 4 \cdot \frac{\hat{\Theta}_n^2(1-\hat{\Theta}_n)^3}{n}$$

b) Sotto il modello parametrico considerato, lo stimatore di massima verosimiglianza

$$\hat{\theta}_n = \frac{1}{1+\bar{X}_n}$$

è STIMATORE CONSISTENTE per $\theta = P_\theta(X=0)$.

Se la VERA legge di X fosse un membro della classe delle DISTRIBUZIONI di POISSON, avremo

$$P_\theta(X=0) = e^{-\theta}$$

D'altro canto, per la LEGGE dei GRANDI NUMERI,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}_\theta(X) = \theta$$

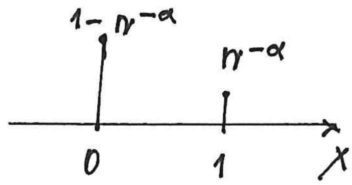
Allora, essendo $\hat{\theta}_n$ funzione continua di $\bar{X}_n$, al CRESCERE di n, si avrebbe

$$\hat{\theta}_n = \frac{1}{1+\bar{X}_n} \xrightarrow{P} \frac{1}{1+\theta} \neq e^{-\theta}$$

e quindi $\hat{\theta}_n$ NON RISULTEREBBE STIMATORE CONSISTENTE per $P_\theta(X=0)$ sotto modello di Poisson.

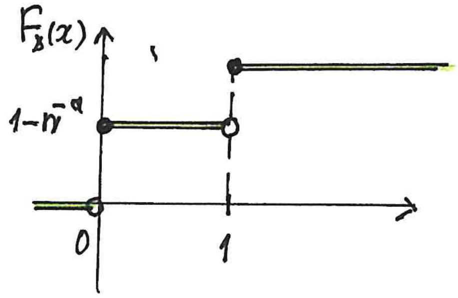
EX 5)

Sia $\{X_n\}_{n\in\mathbb{N}}$ una succ.ne di v.c. INDIPENDENTI $b(1, n^{-\alpha})$, con $\alpha > 0$.



Ora

$$F_{X_n}(x_n) = \begin{cases} 0 & x<0 \\ 1-\dfrac{1}{n^{\alpha}} & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$



sicché al crescere di n, qualunque sia $\alpha > 0$, X_n converge in distribuzione a ZERO.