STATISTICA per ING INF TEL [2L] I. Epifani, A. Pievatolo 25.06.2011

Nello svolgere gli esercizi fornire passaggi e spiegazioni: non bastano i risultati finali.

Esercizio 1 Sia $X_1, \dots, X_n$ un campione casuale estratto dalla densità discreta

$$f(x,\theta) = \begin{cases} \frac{(\ln\theta)^x}{\theta x!} & x = 0,1,2,\dots \\ 0 & \text{altrove} \end{cases} \qquad \theta > 1$$

con θ parametro incognito e maggiore di uno.

1. Determinate uno stimatore $\widehat{\kappa}_{ML}$ della caratteristica $\kappa = \ln\theta$ e $\widehat{\theta}_{ML}$ del parametro θ usando il metodo di massima verosimiglianza.
2. Verificate che la varianza di $\widehat{\kappa}_{ML}$ raggiunge il confine inferiore di Frechét-Cramer-Rao per la varianza di uno stimatore (non distorto); quindi dimostrate che uno stimatore efficiente per θ non esiste. (*Giustificate rigrosamente la risposta*).
3. Determinate media, varianza e distribuzione asintotiche di $\widehat{\kappa}_{ML}$.
4. Costruite un intervallo di confidenza asintotico bilatero di livello approssimativamente 90% per κ, se $n = 169$ e $\sum_{j=1}^{169} x_j = 35.0$. Quindi, deducetene uno bilatero per θ sempre di livello approssimato 90%.

Soluzione

1. $f(x,\theta)$ è una densità di Poisson di parametro $\kappa = \ln\theta$. Riparametrizzando in κ otteniamo

$$L_\kappa(x_1,\dots,x_n) = \prod_{j=1}^n f(x_j,\kappa) = \frac{e^{-n\kappa}\kappa^{\sum_{j=1}^n x_j}}{\prod_{j=1}^n x_j!}$$

$$l_\kappa(x_1,\dots,x_n) = \ln L_\kappa(x_1,\dots,x_n) = -n\kappa + \sum_{j=1}^n x_j \ln\kappa - \ln\left(\prod_{j=1}^n x_j!\right)$$

$$\frac{\partial l_\kappa(x_1,\dots,x_n)}{\partial\kappa} = \frac{n}{\kappa}\left(\bar{x}_n - \kappa\right) \tag{1}$$

da cui deduciamo che $\widehat{\kappa}_{ML} = \overline{X}_n$ e $\widehat{\theta}_{ML} = e^{\widehat{\kappa}_{ML}} = e^{\overline{X}_n}$.

2. Osserviamo che $\widehat{\kappa}_{ML} = \overline{X}_n$ è stimatore non distorto della media della densità di Poisson, nel nostro caso data da κ. Inoltre, leggiamo nell'Equazione (1) che essenzialmente $\dfrac{\partial l_\theta(x_1,\dots,x_n)}{\partial\theta} = a(n,\theta)\,(\widehat{\kappa}_{ML} - \ln\theta)$, con $a(n,\theta) = n/(\theta\ln\theta)$. Ma l'ultima è condizione necessaria e sufficiente affinché la varianza di $\widehat{\kappa}_{ML}$ raggiunga il confine inferiore di Frechét-Cramer-Rao per la varianza di uno stimatore (non distorto).

Per quanto concerne θ, abbiamo che $\partial l_\theta/\partial\theta$ è funzione lineare di $\widehat{\kappa}_{ML}$, quindi non può esserlo di $\widehat{\theta}_{ML}$: non essendo soddisfatta una CNS per l'efficienza, allora $\widehat{\theta}_{ML}$ non è stimatore efficiente di θ. D'altro canto se uno stimatore efficiente per θ esiste, allora è necessariamente stimatore ML: rimane così stabilito che NON esiste nessun stimatore efficiente di θ.

3. Vale che $E(\widehat{\kappa}_{ML}) = E(\overline{X}) = E(X_1) = \kappa$ e $\mathrm{Var}(\widehat{\kappa}_{ML}) = \mathrm{Var}(\overline{X}) = \mathrm{Var}(X_1)/n = \kappa/n$. Inoltre, per il Teorema centrale del limite la f.d.r. asintotica di $\widehat{\kappa}_{ML}$ è $\mathcal{N}(\kappa, \kappa/n)$.

4. La f.d.r. asintotica di $(\widehat{\kappa}_{ML} - \kappa)/\sqrt{\kappa/n}$ è $\mathcal{N}(0,1)$. Segue che un IC asintotico bilatero per κ di livello approssimativamente 90% ha estremi $\widehat{\kappa}_{ML} \pm 1.645 \times \sqrt{\widehat{\kappa}_{ML}/n}$: cioè, siamo 90%-confidenti (approssimativamente) che $0.1495 < \kappa < 0.2647$. Sfruttando il fatto che θ è funzione strettamente crescente di κ: $\theta = e^\kappa$ siamo 90%-confidenti (approssimativamente) che $e^{0.1495} < \theta < e^{0.2647}$, cioè che $1.1316 < \theta < 1.3030$.

Soluzione alternativa: Dal fatto che approssimativamente $P_\theta\left(|\frac{\widehat{\kappa}_{ML}-\kappa)}{\sqrt{\kappa/n}}| < 1.645\right) = 0.90$ deduciamo che $P_\theta\left(|\frac{\widehat{\kappa}_{ML}}{\sqrt{\kappa}} - \sqrt{\kappa}| < \frac{1.645}{\sqrt{169}}\right) = 0.90$. Inoltre, $\widehat{\kappa}_{ML}/\sqrt{\kappa} - \sqrt{\kappa}$ è funzione strettamente decrescente di κ, nulla per $\kappa = \widehat{\kappa}$. Quindi $|\frac{\widehat{\kappa}_{ML}}{\sqrt{\kappa}} - \sqrt{\kappa}| < \frac{1.645}{\sqrt{169}}$ se e solo se $T_1 < \kappa < T_2$ con T_1 uguale al quadrato dell'unica soluzione positiva dell'equazione in κ: $\widehat{\kappa}_{ML}/\sqrt{\kappa} - \sqrt{\kappa} = 1.645/13$ e T_2 uguale al quadrato dell'unica soluzione positiva di $\widehat{\kappa}_{ML}/\sqrt{\kappa} - \sqrt{\kappa} = -1.645/13$. Scopriamo che $T_1 = [(-1.645/13 + \sqrt{1.645^2/169 + 4\widehat{\kappa}_{ML}})/2]^2 \simeq 0.15697$ e $T_2 = [(1.645/13 + \sqrt{1.645^2/169 + 4\widehat{\kappa}_{ML}})/2]^2 \simeq 0.27325$. Il corrispondente IC per θ è $(1.3142, 1.1700)$. ∎

Esercizio 2 Siano X una variabile gaussiana di media incognita μ_X e varianza $\sigma_X^2 = 1$ e Y una variabile gaussiana di media $\mu_Y = 0.5$ e varianza $\sigma_Y^2 = 3$; inoltre, siano X e Y indipendenti.

1. Determinate media, varianza e distribuzione della variabile $W = X - Y$.

Si supponga ora di avere osservato il seguente campione casuale di 5 osservazioni della popolazione W:

$$w_1 = 4.8;\ w_2 = 3.5;\ w_3 = 6.4;\ w_4 = 7.2;\ w_5 = 4.3 \tag{2}$$

2. Impostate un opportuno test che usi il campione osservato $w_1, ..., w_5$ per verificare se la media μ_X di X non supera il valore 3.51, tenendo conto del fatto che si è disposti a commettere un errore di primo tipo al più pari al 3% quando si accetta che μ sia strettamente maggiore di 3.51, ma in realtà è vero il contrario. Con i dati in (2), quale decisione prendete?

3. Determinate analiticamente e rappresentate graficamente la funzione di potenza del test costruito al punto 2.

4. Calcolate la probabilità di prendere una decisione SBAGLIATA quando il vero valore della media μ_X di X è 4.71.

Soluzione

1. In quanto differenza di due variabili aleatorie gaussiane indipendenti, anche W è gaussiana con media la differenza delle medie e varianza la somma delle varianze: $W \sim \mathcal{N}(\mu_X - 0.5, 4)$. Pertanto, $W_1, \dots, W_5$ è un campione casuale estratto da popolazione gaussiana di media incognita e varianza nota: $W_1, \dots, W_5$ i.i.d. $\sim \mathcal{N}(\mu, 4)$ ($\mu = \mu_X - 0.5$).

2. Ci viene richiesto di costruire un test per verificare $H_0 : \mu_X \leq 3.51$ *versus* $H_1 : \mu_X > 3.51$ che formulato in termini del campione $w_1, \dots, w_5$ e della sua media risulta essere:

$$H_0 : \mu \leq 3.01 \text{ } \textit{versus} \text{ } H_1 : \mu > 3.01 \ .$$

Avendo a disposizione un campione gaussiano di varianza nota e pari a 4, allora rifiutiamo H_0 con significatività 3% se $(\overline{W} - 3.01)/\sqrt{4/5} \geq z_{1-0.03}$, ossia se

$$\overline{W} \geq \sqrt{\frac{4}{5}} \cdot z_{1-0.03} + 3.01 = \sqrt{\frac{4}{5}} \cdot 1.881 + 3.01 \simeq 4.692418 \simeq 4.69$$

Con il nostro campione: $\overline{W} = 5.24$ e quindi rifiutiamo H_0 a livello 3%.

3. $$\pi(\mu_X) = P_{\mu_X}\left(\overline{W} \geq 4.69\right) = 1 - \Phi\left(\frac{4.69 - \mu_X + 0.5}{2/\sqrt{5}}\right) = 1 - \Phi\left(\frac{5.19 - \mu_X}{2/\sqrt{5}}\right), \quad \mu_X > 3.51 \ ;$$

La potenza è funzione crescente di μ_X, sempre maggiore di 0.03 con asintoto orizzontale in 1, per $\mu_X \to +\infty$.

4. Dobbiamo calcolare la probabilità di accettare H_0 ma H_0 è falsa perché il vero valore di $\mu_X = 4.71$ è maggiore di 3.51; in altre parole, dobbiamo calcolare la probabilità di errore di seconda specie in $\mu_X = 4.71$:

$$\beta(4.71) = 1 - \pi(4.71) = \Phi\left(\frac{5.19 - 4.71}{2/\sqrt{5}}\right) \simeq \Phi(0.54) \simeq 0.7054. \quad \blacksquare$$

Esercizio 3 Un campione casuale di 500 nuclei famigliari degli Stati Uniti è stato classificato per regione e reddito (in migliaia di dollari) ottenenndo i risultati che seguono:

Reddito	Sud	Nord
$0 - 10$	42	53
$10 - 20$	55	90
$20 - 30$	47	88
30 o più	36	89

1. [1] Verificate l'ipotesi che il reddito di una famiglia scelta a caso sia indipendente dalla regione di residenza.

Siamo ora interessati a verificare se la seguente densità di Pareto:

$$f(x, \theta) = \begin{cases} 8^\theta \theta x^{-\theta-1} & \text{se } x > 8 \\ 0 & \text{altrove}, \qquad \theta > 1 \end{cases} \tag{3}$$

che ha valore atteso

$$E_\theta(X) = \frac{8\theta}{(\theta - 1)}, \qquad \forall \theta > 1$$

si adatti ai dati forniti sui redditi delle famiglie americane. A tal fine:

2. fornite uno stimatore di θ usando i 500 dati raggruppati sui redditi delle famiglie americane;
3. determinate $P(a < X < b)$ quando X ha densità di Pareto $f(x, \theta)$ in (3);
4. valutate con un opportuno test la bontà di adattamento del modello di Pareto (3) ai dati sul reddito. (*Se non siete riusciti a risolvere il punto 2, scegliete $\theta = 2.00$ ed eseguite un opportuno test.*)

Soluzione

1. Impostiamo un test χ^2 di indipendenza per verificare H_0: *"reddito e regione di residenza sono indipendenti"* contro H_1: *"reddito e regione di residenza non sono indipendenti"*. La statistica test è

$$Q_{\text{ind}} = \sum_{i=1}^{4} \sum_{j=1}^{2} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{500}\right)^2}{\frac{N_{i.}N_{.j}}{500}} = 500 \sum_{i=1}^{4} \sum_{j=1}^{2} \frac{{N_{ij}}^2}{N_{i.}N_{.j}} - 500$$

e nel nostro caso troviamo il valore $q_{\text{ind}} \simeq 5.91$. Dato che la statistica test sotto l'ipotesi nulla ha distribuzione limite chiquadrato con $(4-1)(2-1) = 3$ gradi di libertà, la cui f.d.r. indichiamo con F_3, allora per il p-value, usando le tabelle, abbiamo $F_3(5.739) = 87.5\%$ e $F_3(6.251) = 90.0\%$. Il p-value è $1 - F_3(q_{\text{ind}}) \in (10\%, 12.5\%)$ (valore esatto del p-value con R = 0.1161182). Un p-value così alto indica mancanza di evidenza empirica contro H_0, cioè ai consueti livelli di significatività ($\leq 10\%$) concludiamo che il reddito di una famiglia scelta a caso è indipendente dalla regione di residenza.

2. Poiché $\mathrm{E}(X) = 8\theta/(\theta - 1)$ e

$$\overline{X}_{mid} = \frac{9775}{500} = 19.55,$$

allora, applicando il metodo dei momenti otteniamo

$$\widehat{\theta} = \frac{\overline{X}_{mid}}{\overline{X}_{mid} - 8} \simeq 1.69$$

3. Si trova facilmente che la fdr di X è $F(x) = 0$ quando $x \leq 8$ e

$$F(x) = \int_8^x \frac{\theta 8^\theta}{s^{\theta+1}}\, ds = 1 - \frac{8^\theta}{x^\theta}, \qquad \forall x > 8,$$

[1] Estratto da Sheldon Ross, Probabilità e statistica, Ed Apogeo, 2004

da cui otteniamo

$$P(a < X < b) = a \begin{cases} 0 & \text{se } a < b \leq 8 \\ 8^\theta \left(\frac{1}{a^\theta} - \frac{1}{b^\theta}\right) & \text{se } 8 < a < b \\ 1 - \frac{8^\theta}{b^\theta} & \text{se } a \leq 8 < b \end{cases}$$

4. Dobbiamo impostare un test chiquadrato di buon adattamento per dati raggruppati per l'ipotesi nulla composta H_0 :"X è paretiana" contro l'alternativa che X non sia paretiana.

Sostituendo $\widehat{\theta} = 1.69$ nelle probabilità calcolate al punto 3., otteniamo i valori "stimati" delle probabilità "teoriche": $p_1^{(0)}(\widehat{\theta}) = F_{H_0}(10) = 1 - (8/10)^1.69 \simeq 0.314$, $p_2^{(0)}(\widehat{\theta}) = 0.473, \ldots$; in sintesi

A_k	$0 - 10$	$10 - 20$	$20 - 30$	$(30, \infty)$
$p_i^{(0)}(\widehat{\theta})$	0.314	0.473	0.105	0.108
$np_i^{(0)}(\widehat{\theta})$	157.0	236.5	52.5	54
N_i	95	145	135	125

Dunque, la statistica test di Pearson $Q = \displaystyle\sum_{i=1}^{4} \frac{(N_i - np_i^{(0)}(\widehat{\theta}))^2}{np_i^{(0)}(\widehat{\theta})}$ vale $\simeq 282.9$ e sotto H_0 ha distribuzione $\mathcal{E}(2)$: rifiutiamo l'ipotesi di dati paretiani a qualunque livello del test, poiché $p - value = \mathrm{e}^{-282.9/2} \simeq 0$: fortissima evidenza empirica contro H_0. ▮