Politecnico di Milano – Facoltà di Ingegneria Informatica

Anno Accademico 2009/2010

Corso di Statistica (2L) per INF e TEL

Docente: Antonio Pievatolo; esercitazioni: Raffaele Argiento

Esercitazione del 30/04/08

Esercizio 1

Avete raccolto $n = 100$ misurazioni ripetute di una grandezza incognita μ e avete ottenuto:

$$\sum_{i=1}^{n} x_i = 2371.2 \qquad \sum_{i=1}^{n} x_i^2 = 57682.1$$

1. Siete ingrado di fornire una stima intervallare di μ?
2. Con quale confidenza pensate che μ sia maggiore di 22.960?

Soluzione

1. Utilizzando il teorema del limite centrale ed alcuni risultati di convergenza (Teorema di Slutsky) si ricava che:
$$Z := \frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{approx}{\sim} N(0,1) \tag{1}$$
La variavile Z è quindi una quantità pivotale approssimata per la grandezza μ in esame. Se ne ricava che
$$\mathbb{P}(-z(\frac{1+\gamma}{2}) < Z < z(\frac{1+\gamma}{2}) = \gamma \Rightarrow \mathbb{P}\left(\bar{X} - \frac{S}{\sqrt{n}}z(\frac{1+\gamma}{2}) < \mu < \bar{X} + \frac{S}{\sqrt{n}}z(\frac{1+\gamma}{2})\right) = \gamma$$
Quindi $\left(\bar{x} - \frac{s}{\sqrt{n}}z(\frac{1+\gamma}{2}), \bar{s} + \frac{s}{\sqrt{n}}z(\frac{1+\gamma}{2})\right)$ è un intervallo di confidenza di livello γ per μ. Utilizzando i dati dell'esercizio si ottiene l'intervallo $(20.829, 26, 595)$.
2. La domanda dell'esercizio equivale a chiedere il livello γ dell'intervallo di confidenza per μ ad una coda superiore $(22.960, +\infty)$. Utilizzando la quantità pivotale approssimata utilizzata al punto precedente si ricava che tale intevallo ha la forma $\left(\bar{x} - \frac{s}{\sqrt{n}}z(\gamma), +\infty\right)$. Per rispondere al quessito risolviamo l'equazione in γ
$$\bar{x} - \frac{s}{\sqrt{n}}z(\gamma) = 22.960 \quad \Rightarrow z(\gamma) = 1.96 \quad \Rightarrow \quad \gamma = 0.975$$

∎

Esercizio 2

Tema d'esame del 19/09/05
Abbiamo estratto un campione casuale $X_1, ..., X_n$ dalla densità di parametro $\theta > 0$:

$$f(x, \theta) = \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right) \mathbf{1}_{(0,+\infty)} \tag{2}$$

1. Determinate lo stimatore della caratteristica $\kappa(\theta) = \theta^2$ usando il metodo di massima verosimiglianza.

2. Determinate la densità di $Y = X^2$.

3. Chiamiamo $\hat{\kappa}_n$ lo stimatore di massima verosimiglianza di $\kappa(\theta) = \theta^2$: discutete qualche proprietà di $\hat{\kappa}_n$.

4. Determinate la distribuzione della variabile aleatoria $Q_n = \frac{2n\hat{\kappa}_n}{\theta^2}$.

5. Fornite un intervallo di confidenza a due code per $\kappa(\theta)$ di livello $\gamma = 90\%$ per $n = 10$ e $\hat{\kappa}_n = 0.0387$. Quindi deducete un intervallo di confidenza per θ, sempre bilatero e di livello 90%.

Soluzione

1. La distribuzione in esame appartiene alla famiglia esponenziale, infatti possiamo scrivela come:

$$f(x,\theta) = x \exp\left\{-\frac{x^2}{2\theta^2} - 2\log(\theta)\right\}$$

Con la solita notazione per le famiglie esponenziali si ha che: $C(x) = x$, $A(\theta) = -\frac{1}{\theta^2}$, $g(x) = \frac{x^2}{2}$, $B(\theta) = 2\log(\theta)$. Tutte e sole le caratteristiche stimabili in modo efficiente sono funzioni lineari di

$$k(\theta) = -\frac{B'(\theta)}{A'(\theta)} = \theta^2$$

Si ricava che

$$\hat{k}_n = \frac{1}{2n}\sum_{i=1}^n x_i^2$$

è lo stimatore efficiente di $k(\theta) = \theta^2$, esso è dunque ML per θ^2. Per la proprietà di invarianza degli stimatori ML $\hat{\theta}_n = \sqrt{\frac{1}{2n}\sum_{i=1}^n x_i^2}$ è uno stimatore ML per θ.

1.*Alternativa* Sia $\underline{x} = (x_1, \dots, x_n)$ una realizzazione campionaria, scriviamo la funzione di verosimiglianza basata su $\underline{x}$

$$L(\theta, \underline{x}) = \prod_{i=1}^n f(x_i, \theta) = \frac{\prod_{i=1}^n x_i}{\theta^{2n}} \exp\left\{-\frac{\sum_{i=1}^n x_i^2}{2\theta^2}\right\} \mathbf{1}_{(0,\infty)}(x_{(1)}).$$

Passiamo dunque al logaritmo

$$l(\theta, \underline{x}) = \log(L(\theta, \underline{x})) = -n\log(\theta^2) + \sum_{i=1}^n \log(x_i) - \frac{\sum_{i=1}^n x_i^2}{2\theta^2}.$$

Differenziando rispetto a θ^2 otteniamo

$$\frac{\partial}{\partial\theta^2} l(\theta, \underline{x}) = -\frac{n}{\theta^2} + \frac{\sum_{i=1}^n x_i^2}{2(\theta^2)^2}.$$

Studiando il segno di quest'ultima, rispetto a θ^2 otteniamo

$$\hat{k}_n = \frac{\sum_{i=1}^n x_i^2}{2n}.$$

2. Sia $Y = X^2$ e $y > 0$, allora

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}(X^2 \le y) = \mathbb{P}(X \le \sqrt{y}) = \int_0^{\sqrt{y}} \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right) \\ &= \left[-\exp\left\{-\frac{x^2}{2\theta^2}\right\}\right]_0^{\sqrt{y}} = 1 - \exp\left\{-\frac{y}{2\theta^2}\right\};\end{aligned}$$

Concludiamo che $Y \sim Exp(2\theta^2) \overset{\mathrm{d}}{=} \Gamma(1, 2\theta^2)$

3. Sia $X_1, \ldots, X_n$ un campione dalla densità (2).

   (a) Dal punto 1. dell'esercizio si ha che $\hat{k}_n$ è efficiente, quindi esso è non distorto e consistente in media quadratica. Se si è risolto il punto 1 studiando la funzione di verosimiglianza (come descritto in 1. *Alternativa*) si può osservare che la derivata rispetto a θ^2 della log-verosimiglianza è
   $$\frac{\partial}{\partial \theta^2} l(\theta, \underline{x}) = -\frac{n}{\theta^2} + \frac{\sum_{x_i}^{2n}}{2(\theta^2)^2} = \frac{n}{(\theta^2)^2}\left(\hat{k}_n - \theta^2\right).$$
   Quindi $\hat{k}_n$ è lo stimatore efficiente per $k(\theta) = \theta^2$.

   (b) Se non si è riconosciuto che $\hat{k}_n$ è efficiente si osservi che dal punto 2. dell'esercizio si deduce che $\sum_{i=1}^n X_i^2 \sim \Gamma(n, 2\theta^2)$, di conseguenza $\hat{k}_n = \frac{\sum_{i=1}^n X_i^2}{2n} \sim \Gamma(n, \frac{\theta^2}{n})$ quindi $\mathbb{E}(\hat{k}_n) = \theta^2$: $\hat{k}_n$ è NON distorto!

   (c) Dato che il modello (2) è regolare, $\hat{k}_n$ è asintoticamente normale con media θ^2 e varianza $(\theta^2)^2/n$.

4. Dal punto 2. dell'esercizio si deduce che
   $$Q_n = \frac{2n}{\theta^2}\hat{k}_n \sim \Gamma(n, 2) \stackrel{\mathrm{d}}{=} \Gamma\left(\frac{2n}{2}, 2\right) \stackrel{\mathrm{d}}{=} \chi^2_{2n}$$

5. La variabile Q_n è una quantità pivotale in quanto la sua distribuzione non dipende da θ. Si fissino i due percentili $q_1 = \chi^2_{2n}\left(\frac{1-\gamma}{2}\right)$ e $q_2 = \chi^2_{2n}\left(\frac{1+\gamma}{2}\right)$, dove γ è compreso fra 0 e 1. Allora;
   $$\mathbb{P}(q_1 \le Q_n \le q_2) = \gamma \Rightarrow \mathbb{P}\left(q_1 \le \frac{2n}{\theta^2}\hat{k}_n \le_2\right) = \gamma \Rightarrow \mathbb{P}\left(\frac{2n}{q_2}\hat{k}_n \le \theta^2 \le \frac{2n}{q_1}\hat{k}_n\right) = \gamma.$$
   Concludiamo che
   $$\left[\frac{2n}{q_2}\hat{k}_n, \frac{2n}{q_1}\hat{k}_n\right] \tag{3}$$
   è un intervallo di confidenza di livello γ per θ^2. Inoltre dato che $k(\theta) = \theta^2$ per $\theta > 0$ è una funzione monotona crescente, l'intervallo
   $$\left[\sqrt{\frac{2n}{q_2}\hat{k}_n}, \sqrt{\frac{2n}{q_1}\hat{k}_n}\right] \tag{4}$$
   è di confidenza al livello γ per θ.

   Sia ora $\gamma = 0.9$, $n = 10$ e $\hat{\kappa}_n = 0.0387$, allora $q_1 = \chi^2_{20}(0.05)$ e $q_2 = \chi^2_{20}(0.95) = 31.4$. Sostituendo in tali valori in (3) si ottiene la realizzazione $(0.0246, 0.07118)$, mentre $(0.1570, 0.2665)$ é la realizzazione dell'intervallo in (4).

∎

Esercizio 3

Una moneta viene lanciata $n = 1000$ volte ottenendo 432 volte testa.

1. Determinare lo stimatore di massima verosimiglianza (ML) per θ, la probabilità di ottenere testa in un lancio.
2. Discutere qualche proprietà esatta e asintotica dello stmimatore $\hat{\theta}_n$ ricavato al punto 1.
3. Determinate l'informazione di Fisher del modello $I(\theta)$ ed un suo stimatore ML.
4. Costruire un intervallo di confidenza bilatero asintotico per θ di livello $\gamma = 0.95$.
5. Quanto deve essere grande n affinché l'intervallo calcolato al punto 4. abbia ampiezza al più pari a 0.02.

Soluzione

Per $i = 1, \ldots, n$ si definiscano le variabili

$$X_i = \begin{cases} 1 & \text{se si ottiene testa all'i-esimo lancio} \\ 0 & \text{altrimenti} \end{cases}.$$

Costruendo così il campione $X_1, \ldots, X_n$ da una popolazione Bernoulli(θ). La cui densità discreta è

$$p_X(x;\theta) = \theta^x (1-\theta)^{1-x}; \quad x \in \{0,1\}, \quad \theta \in [0,1].$$

1. Il modello bernulliano per $\theta \neq 0, 1$ appartiene alla famiglia esponenziale infatti possiamo scrivere:

$$p_X(x;\theta) = \mathbb{1}_{\{0,1\}}(x) \exp\left\{ x \log(\frac{\theta}{1-\theta}) + \log(1-\theta) \right\}$$

   Con la solita notazione abbamo $C(x) = \mathbb{1}_{\{0,1\}}(x)$, $A(\theta) = \log(\frac{\theta}{1-\theta})$, $B(\theta) = \log(1-\theta)$ e $g(x) = x$. Tutte e sole le caratteristiche stimabili in modo efficiente sono quelle nella forma:

$$k(\theta) = -\frac{B'(\theta)}{A'(\theta)} = \theta.$$

   Quindi lo stimatore efficiente per θ è

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

1. *Alternativa* Data la realizzazione $\underline{x} = (x_1, \ldots, x_n)$ la verosimiglianza è

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

   La log-verosimiglianza

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

   Derivando rispetto a θ si ottiene

$$\frac{\partial}{\partial\theta} l(\theta;\underline{x}) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta} = \frac{1}{\theta(1-\theta)} \left(\frac{\sum_{i=1}^n x_i}{n} - \theta \right).$$

   Dato che $\frac{\partial}{\partial\theta} l(\theta;\underline{x}) > 0 \;\Rightarrow\; \theta < \bar{x}_n$,, Si ottiene che

$$\hat{\theta}_n = \bar{X}_n$$

   dove $\bar{X}_n = \sum_{i=1}^n X_i$ è la statistica media campionaria

2. $\hat{\theta}_n$ è efficiente, quindi esso è non distorto e consistente in media quadratica. Oppure si osservi che dato che $X_i \sim$Bernoulli(θ), si ha che $\hat{\theta} = \bar{X}_n \sim \frac{1}{n}\text{Bin}(n,\theta)$ é uno stimatore non distorto, inoltre, esso ha varianza $\frac{\theta(1-\theta)}{n}$, ed è quindi consistente in media quadratica.

3. Dato che $\hat{\theta}$ è efficiente $I(\theta) = \frac{1}{n\text{Var}(\hat{\theta})} = \frac{1}{\theta(1-\theta)}$. Per la proprietà di invarianza dello stimatore di massima verosimiglianza $\hat{I} = I(\hat{\theta})$.

4. Dato che il modello è regolare, lo stimatore di massima verosimiglianza $\hat{\theta}_n$ è asintoticamente gaussiano con media $\mathbb{E}(\hat{\theta}_n) = \theta$, e $\text{Var}\hat{\theta}_n = \frac{1}{nI(\theta)} = \frac{\theta(1-\theta)}{n}$ (Si osservi come tali considerazioni si possono dedurre anche dal teorema de limite centrale). In conclusione, la quantità $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta)$

ha distribuzione asintotica approssimativamente normale standard. Essa è dunque una quantità (asintoticamente) pivotale, dalla quale ricaviamo l'intervallo di confidenza asintotico:

$$\left[\hat{\theta} - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}; \hat{\theta} + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right],$$

la cui realizzazione sulla base del campione fornito nel testo è $[0.4013; 04627]$.

5. La lunhezza dell'intervallo definito al punto 4. è

$$\mathcal{L}_\gamma(n) = 2z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

Questa è una quantità aleatoria, tuttavia osserviamo che se θ appartiene all'intervallo $[0,1]$, allora

$$0 < \theta(1-\theta) < \frac{1}{4}.$$

Quindi con probabilità 1 si ha che

$$\mathcal{L}_\gamma(n) < z_{\frac{1+\gamma}{2}}\frac{1}{\sqrt{n}}.$$

Affinche la lunghezza dell'intervallo sia minore di 0.02 si scegliere n che soddisfa la disequazione

$$z_{\frac{1+\gamma}{2}}\frac{1}{\sqrt{n}} < 0.02$$

ovvero $n > 9506.25$.

∎

Esercizio 4

Tema d'esame del 12/09/06
Il tempo di funzionamento in mesi di un componente prima di guastarsi può essere modellato come una variabile aleatoria T assolutamente continua con densità

$$f(t,\theta) = \frac{1}{6^{\frac{1}{\theta}}} \cdot \frac{1}{2\theta} t^{\frac{1}{2\theta}-1} \mathbb{1}_{(0,36)}(t), \qquad \theta > 0$$

Il parametro θ è incognito.

1. Determinate in funzione di θ la caratteristica κ data dalla probabilità che un componente non si guasti nei primi 6 mesi di vita.

In un esperimento vengono messi in prova simultaneamente n componenti e $T_1, \ldots, T_n$ sono i tempi di funzionamento rilevati.

2. Determinate gli stimatori di massima verosimiglianza di θ e di κ.
3. Sia $\hat{\theta}$ lo stimatore di massima verosimiglianza individuato al punto precedente. Determinate media, varianza e distribuzione asintotica di $\hat{\theta}$. (Per comodità vi forniamo i seguenti risultati che è necessario usare: $\mathbb{E}_\theta[\log(\sqrt{T})] = \log 6 - \theta$ e $\mathbb{E}_\theta[(\log(\sqrt{T}))^2] = (\log 6 - \theta)^2 + \theta^2$.

Soluzione ...

∎