TEORIA E LABORATORIO DI INFERENZA STATISTICA - STATISTICA II
CORSI DI LAUREA D E A
ANNO ACCADEMICO 2005-2006

Verifica n. 1 - 21 aprile 2006

COMPITO A

COGNOME E NOME ——————————————————————————-

Esercizio 1. Da un'indagine demografica condotta dal *Census Bureau* statunitense, risulta che il 9.96 % dei cittadini americani di età superiore a 18 anni è di origine ispanica. Si consideri un campione casuale di $n = 1200$ cittadini americani e determinare

a) il numero medio di cittadini ispanici in un campione di tale ampiezza;

b) la probabilità che di un campione di tale ampiezza facciano parte meno di 100 cittadini ispanici.

Svolgimento.

$X_i =$ etnia cittadino i-esimo campionario (1=ispanico, 0=non ispanico);

$Y_n = \sum_{i=1}^{n} X_i =$ numero di cittadini ispanici nel campione;

$\theta = 0.0996$

$X_i|\theta \sim \text{Ber}(\theta), \qquad Y_n|\theta \sim \text{Binom}(n, \theta).$

a) $E_\theta[Y_n] = n\theta = (1200) * (0.096) = 119.52.$

b) Utilizzando l'approssimazione normale della distribuzione binomiale e ricordando che $V_\theta(Y_n) = n\theta(1 - \theta)$ si ottiene

$$\Pr(Y_n < 100) \approx \Pr(Z < (100 - 119.52)/\sqrt{107.64}) = \Phi(-1.88) = 0.03,$$

dove $\Phi(\cdot)$ indica la funzione di ripartizione della v.a. normale standardizzata (Z).

Esercizio 2. Si consideri un campione casuale di n osservazioni da una popolazione X con distribuzione di probabilità

$$f_X(x;\theta) = \frac{\theta}{(1+x)^{1+\theta}} \qquad x > 0, \qquad \theta > 0.$$

a) Scrivere il modello statistico-probabilistico associato al campione casuale.

b) Determinare l'espressione della funzione di verosimiglianza del parametro associata a un generico campione osservato, $\mathbf{x}_n$; individuare il nucleo della funzione di verosimiglianza e, se esiste, una statistica sufficiente unidimensionale.

c) Determinare la stima di massima verosimiglianza di θ e calcolarne il valore per un campione di dimensione $n = 10$ in cui si ha che

$$\sum_{i=1}^{10} \ln(1 + x_i) = 5.$$

d) Verificare se la famiglia di densità $\mathcal{F} = \{f_X(\cdot;\theta);\ \theta \in \Theta\}$ è una famiglia esponenziale.

Svolgimento.

a) Il modello statistico è:

$$\left((\mathbb{R}^+)^n,\ \frac{\theta^n}{[\prod_{i=1}^n (1+x_i)]^{(1+\theta)}} \prod_{i=1}^n I_{(0,+\infty)}(x_i), \mathbb{R}^+ \right).$$

b)

$$L(\theta;\mathbf{x}_n) = \prod_{i=1}^n f_X(x_i;\theta) = \frac{\theta^n}{[\prod_{i=1}^n (1+x_i)]^{(1+\theta)}} \prod_{i=1}^n I_{(0,+\infty)}(x_i) = \frac{\prod_{i=1}^n I_{(0,+\infty)}(x_i)}{\prod_{i=1}^n (1+x_i)} \frac{\theta^n}{\prod_{i=1}^n (1+x_i)^\theta}.$$

da cui si evince che L si fattorizza nel prodotto $h(\mathbf{x}_n)g(T(\mathbf{x}_n),\theta)$, dove il nucleo della funzione di verosim. è

$$g(T(\mathbf{x}_n),\theta) = \frac{\theta^n}{\prod_{i=1}^n (1+x_i)^\theta}$$

mentre $h(\mathbf{x}_n) = \frac{\prod_{i=1}^n I_{(0,+\infty)}(x_i)}{\prod_{i=1}^n (1+x_i)}$. Pertanto, per il teorema di fattorizzazione si ha che $T(\mathbf{x}_n) = \prod_{i=1}^n (1+x_i)$ è stat. suff. unidimensionale per il modello considerato.

c) Consideriamo la f.ne di log-verosimiglianza:

$$\ell(\theta;\mathbf{x}_n) = n \ln\theta - \theta \sum_{i=1}^n \ln(1+x_i).$$

Per trovare la stima di massima verosimiglianza consideriamo l'equazione di log-verosimiglianza:

$$\frac{d}{d\theta}\ell(\theta;\mathbf{x}_n) = 0 \qquad \text{ovvero} \qquad \frac{n}{\theta} - \sum_{i=1}^n \ln(1+x_i) = 0,$$

la cui unica radice è $\frac{n}{\sum_{i=1}^n \ln(1+x_i)}$. Poichè la derivata seconda di $\ell(\theta;\mathbf{x}_n)$ è $-n/\theta^2$, ovvero sempre negativa, la radice trovata è punto di massimo per ℓ (e per L) e quindi stima di massima verosimiglianza, $\hat{\theta}_{mv}(\mathbf{x}_n)$. Per il campione considerato si ha $\hat{\theta}_{mv}(\mathbf{x}_n) = 10/5 = 2$.

c) Si tratta di famiglia esponenziale. Infatti, osservando che

$$f_X(x;\theta) = \frac{1}{1+x} \exp\left(-\theta \ln(1+x) + \ln\theta\right),$$

le funzioni di densità per il modello considerato possono essere scritte nella forma generale

$$f_X(x;\theta) = h(x) \exp\left(\eta(\theta) T(x) - c(\theta)\right),$$

con

$$h(x) = \frac{1}{1+x}, \quad \eta(\theta) = -\theta, \quad T(x) = \ln(1+x), \quad c(\theta) = -\ln\theta.$$

Esercizio 3. Si suppone che la durata di funzionamento (X, in decine ore) di una popolazione di macchinari prodotti da una fabbrica sia una v.a. di Weibull con funzione di densità

$$f_X(x;\theta) = 2\,\theta\,x\,e^{-\theta x^2}, \qquad x \geq 0, \qquad \theta > 0.$$

In un esperimento si è riscontrato che, per $n = 100$ pezzi esaminati, la somma del quadrato dei tempi di durata è pari a 25.5 (decine di ore). Sulla base del campione osservato,

a) Determinare la stima di massima verosimiglianza di θ.

b) Determinare una stima per intervallo per θ, utilizzando l'insieme di verosimiglianza approssimato di livello $q = 0.147$.

Svolgimento.

a)

$$f_n(\mathbf{x}_n;\theta) = 2^n\theta^n\left(\prod_{i=1}^n x_i\right)e^{-\theta\sum_{i=1}^n x_i^2}; \qquad \ell(\theta;\mathbf{x}_n) = \ln L(\theta;\mathbf{x}_n) \propto n\ln\theta - \theta\sum_{i=1}^n x_i^2.$$

Pertanto

$$\frac{d}{d\theta}\ell(\theta;\mathbf{x}_n) = 0 \quad\Leftrightarrow\quad \frac{n}{\theta} - \sum_{i=1}^n x_i^2 = 0 \quad\Leftrightarrow\quad \theta = \frac{n}{\sum_{i=1}^n x_i^2}.$$

Poichè la derivata seconda di $\ell(\theta;\mathbf{x}_n)$ è $-n/\theta^2$, ovvero sempre negativa, la radice trovata è punto di massimo per ℓ (e per L) e quindi stima di massima verosimiglianza, $\hat{\theta}_{mv}(\mathbf{x}_n)$. Per il campione osservato si ha: $\hat{\theta}_{mv}(\mathbf{x}_n) = 100/25.5 = 3.92$.

b) Intervallo di verosimiglianza approssimato di livello q è:

$$\hat{\theta}_{mv}(\mathbf{x}_n) \pm K \times \sqrt{1/I_n^{oss}},$$

dove

$$k = \sqrt{-2\ln q} = 1.96, \qquad I_n^{oss} = -\frac{d}{d\theta}\ln L(\theta;\mathbf{x}_n)|_{\theta=\hat{\theta}_{mv}(\mathbf{x}_n)} = \frac{n}{[\hat{\theta}_{mv}(\mathbf{x}_n)]^2} = \frac{(\sum_{i=1}^n x_i^2)^2}{n} = 6.05.$$

Si ha quindi che l'intervallo richiesto è: $(3.16, 4.68)$.