INFERENZA STATISTICA

A-D-E-G-T Anno Accademico 2008-2009

Verifica n. 2 - 31 maggio 2009

COMPITO A

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

Esercizio 1A. Sia $X_1, \dots X_n$ un campione casuale da una popolazione con funzione di massa di probabilità

$$f_X(x;\theta) = (2\theta-1)^x(2-2\theta)^{1-x} \qquad x = 0, 1, \qquad \theta \in \left(\frac{1}{2}, 1\right).$$

1. Verificare che
$$\mathbb{E}[X] = 2\theta - 1, \qquad \mathbb{V}[X] = (2\theta-1)(2-2\theta).$$

Soluzione – Osservando che il supporto della v.c. X_i è costituito dai soli valori 0 e 1, si deduce che siamo in presenza di una variabile casuale bernoulliana con probabilità di successo

$$p = 2\theta - 1$$

per la quale è immediato verificare che

$$E[X_i] = p = 2\theta - 1$$

e che

$$\text{Var}\,[X_i] = p(1-p) = (2\theta-1)(2-2\theta)$$

2. Determinare lo stimatore dei momenti di θ, $\hat{\theta}_M(\mathbf{X}_n)$, calcolarne distorsione, varianza ed errore quadratico medio e stabilire se si tratta di stimatore UMVUE.

Soluzione – Risolvendo in θ l'equazione che uguaglia il primo momento teorico (il valore atteso) di X_i con il primo momento empirico (la media campionaria) si ottiene

$$2\theta - 1 = \bar{X} \iff \theta = \frac{\bar{X}+1}{2}$$

e dunque la soluzione dell'equazione determina lo stimatore dei momenti

$$\hat{\theta}_M = \frac{\bar{X}+1}{2}$$

Dalla linearità dell'operatore valore atteso e dalla proprietà della media campionaria per cui $E[\bar{X}] = E[X_i] = 2\theta - 1$ si deduce che

$$\begin{aligned} E\left[\hat{\theta}_M\right] &= E\left[\frac{\bar{X}+1}{2}\right] = \frac{E\left[\bar{X}\right]+1}{2} \\ &= \frac{2\theta - 1 + 1}{2} = \theta \qquad \forall\theta \in \Theta \end{aligned}$$

e dunque siamo in presenza di uno stimatore non distorto per il parametro θ la cui distorsione vale

$$B_{\hat{\theta}_M}(\theta) = 0 \qquad \forall\theta \in \Theta.$$

Essendo lo stimatore dei momenti non distorto, si ha che la sua varianza coincide con l'errore quadratico medio e può essere calcolata come segue

$$\begin{aligned} MSE_{\hat{\theta}_M}(\theta) = \mathrm{Var}_{\hat{\theta}_M}(\theta) &= \mathrm{Var}\left[\frac{\bar{X}+1}{2}\right] \\ &= \frac{1}{4}\mathrm{Var}\left[\bar{X}\right] \\ &= \frac{1}{4}\frac{\mathrm{Var}\left[X_i\right]}{n} \\ &= \frac{(2\theta-1)(2-2\theta)}{4n} \end{aligned}$$

grazie alle ben note proprietà della varianza di trasformazioni lineari, nonché alla proprietà della varianza della media campionaria, valida per tutti i modelli statistici.

Ora per verificare *se* lo stimatore è anche UMVUE sono possibili due strade alternative:

- avendo già verificato che lo stimatore è non distorto ed avendo calcolato il suo errore quadratico medio potrebbe verificarsi che la sua varianza raggiunge il LICR (Limite Inferiore di Cramer Rao). Nel caso in cui ciò si verificasse avremmo potuto dedurre la relazione richiesta. Nel caso però in cui invece il LICR non fosse ragiunto non potremmo dedurre alcunché.
- avendo già verificato che lo stimatore è non distorto si potrebbe cercare di verificare se tale stimatore è funzione di una statistica sufficiente e completa per il parametro θ

La seconda strada sembra più agevole in quanto lo stimatore è funzione della statistica $S(X_1, ..., X_n) = \bar{X}$ che, in un modello bernoulliano, appartenente alla famiglia esponenziale, è noto essere una statistica sufficiente e completa per p. Quindi, dalle proprietà delle statistiche sufficienti e delle statistiche complete, se $S(X_1, ..., X_n)$ è sufficiente e completa per p lo è anche per una funzione biunivoca del parametro p nel nostro caso per

$$\theta = \frac{p+1}{2}.$$

3. Studiare la consistenza di $\hat{\theta}_M(\mathbf{X}_n)$ e determinare l'approssimazione normale della sua distribuzione campionaria.

   Soluzione – Dal momento che la media campionaria è consistente per $p = 2\theta - 1$

   $$\bar{X} \longrightarrow p$$

   sia in senso debole [convergenza in probabilità] che in senso forte [convergenza quasi certa], invocando il teorema di continuità per entrambe le forme di convergenza è immediato dedurre che una funzione continua della media aritmetica converge alla corrispondente funzione applicata nel limite $p = 2\theta - 1$ ovvero

   $$\hat{\theta}_M = \frac{\bar{X} + 1}{2} = g(\bar{X}) \longrightarrow g(p) = \frac{p+1}{2} = \frac{2\theta - 1 + 1}{2} = \theta \qquad \forall \theta \in \Theta.$$

   Inoltre, considerando l'espressione precedente derivata dell'errore quadratico medio si ottiene che

   $$MSE_{\hat{\theta}_M}(\theta) = \frac{(2\theta - 1)(2 - 2\theta)}{4n} \longrightarrow 0$$

   e dunque lo stimatore è consistente anche in media quadratica.

4. Verificare che lo stimatore di massima verosimiglianza di θ coincide con lo stimatore dei momenti.

   Soluzione – Avendo già osservato che il parametro usuale della distribuzione di Bernoulli $p \in (0, 1)$ è in relazione al parametro di interesse $\theta \in (\frac{1}{2}, 1)$ attraverso la relazione

   $$p = 2\theta + 1$$

   è vero anche il viceversa ovvero che il parametro di interesse θ è in relazione con p attraverso la relazione inversa

   $$\theta = g(p) = \frac{p+1}{2}$$

   e dunque è immediato argomentare che dalla nota proprietà di invarianza (equivarianza) dello stimatore di massima verosimiglianza e dal fatto che $\hat{p}_{MV} = \bar{X}$ si ha che

   $$\hat{\theta}_{MV} = g(\hat{p}_{MV}) = \frac{\hat{p}_{MV} + 1}{2} = \frac{\bar{X} + 1}{2}.$$

5. Calcolare un intervallo di confidenza approssimato per il parametro θ, assumendo di avere osservato un campione di numerosità $n = 50$ in cui la somma delle osservazioni campionarie è pari a 10.

   Soluzione – Per costruire un'opportuna quantità pivotale (approssimata) possiamo usare il teorema del limite centrale per cui

   $$\bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right) = N\left(2\theta - 1, \frac{(2\theta - 1)(2 - 2\theta)}{n}\right)$$

Analoga approssimazione è valida (teorema di Slutsky) se sostituiamo alla varianza teorica della distribuzione normale uno stimatore consistente, in questo caso facilmente ottenibile come $\frac{\bar{X}(1-\bar{X})}{n}$ e dunque

$$Pr_\theta \left\{ -z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - (2\theta - 1)}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \leq z_{1-\frac{\alpha}{2}} \right\} = 1 - \alpha \qquad \forall \theta \in \Theta$$

Con pochi passaggi algebrici è possibile invertire le relazioni di disuguaglianza in modo tale da avere un evento equivalente a quello nella parentesi graffa che vede il parametro θ all'intervallo di un intervallo aleatorio e precisamente

$$\begin{aligned}
-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} &\leq \bar{X} - (2\theta - 1) \leq & z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \\
-\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} &\leq -(2\theta - 1) \leq & -\bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \\
+\bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} &\geq 2\theta - 1 \geq & +\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \\
\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} &\leq 2\theta - 1 \leq & \bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \\
\bar{X} + 1 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} &\leq 2\theta \leq & \bar{X} + 1 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \\
\frac{\bar{X}+1}{2} - \frac{1}{2}z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} &\leq \theta \leq & \frac{\bar{X}+1}{2} + \frac{1}{2}z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}
\end{aligned}$$

Dalla precedente derivazione dell'intervallo aleatorio otteniamo che per $n = 50$ e $\sum_{i=1}^{50} X_i = 20$ si ha che

$$\bar{X} = \bar{x} = \frac{20}{50} = 0.4$$

ed inoltre per $1 - \alpha = 0.95$ si ha che $z_{1-\frac{\alpha}{2}} = 1.96$. Sostituendo tali numeri si ottiene

$$IC_{0.95}[\theta] = [0.632, 0.768]$$

Esercizio 2A. Sia $X_1, \ldots, X_n$ un campione casuale di dimensione n proveniente da una popolazione bernoulliana di parametro incognito θ, con distribuzione di probabilità:

$$f_X(x;\theta) = \theta^x \, (1-\theta)^{1-x}, \qquad x = 0, 1 \quad \theta \in (0,1).$$

Si consideri il sistema di ipotesi:

$$H_0 : \, \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \, \theta = \theta_1,$$

assumendo $\theta_0 < \theta_1$.

1. Verificare che la regione di accettazione del test basato sul rapporto delle verosimiglianze risulta essere
$$A = \{\mathbf{x}_n : \, S_n = \sum_{i=1}^{n} X_i \leq k\}. \tag{1}$$

   Soluzione – Ricordiamo che dal lemma di Neyman e Pearson si ottiene che la regione di accettazione di un test ottimo (ovvero con potenza maggiore o uguale a qualsiasi altro test con uguale probabilità di errore di prima specie) si ottiene considerando una regione di accettazione costituita dai campioni in corrispondenza dei quali il rapporto tra la verosimiglianza del parametro θ_0 sotto l'ipotesi nulla e quella del parametro θ_1 sotto l'ipotesi alternativa risulta elevato, in simboli

$$A \;\; = \;\; \left\{ \boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n : \frac{L_{\boldsymbol{x}}(\theta_0)}{L_{\boldsymbol{x}}(\theta_1)} \geq c \right\}.$$

   Notiamo che questo equivale a considerare la regione complementare alla regione di accettazione ovvero la regione di rifiuto costituita dai campioni in corrispondenza dei quali il rapporto tra la verosimiglianza del parametro θ_1 sotto l'ipotesi alternativa e quella del parametro θ_0 sotto l'ipotesi nulla risulta elevato, in simboli

$$\begin{aligned} R = A^c \;\; &= \;\; \left\{ \boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n : \frac{L_{\boldsymbol{x}}(\theta_0)}{L_{\boldsymbol{x}}(\theta_1)} < c \right\} \\ &= \;\; \left\{ \boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n : \frac{L_{\boldsymbol{x}}(\theta_1)}{L_{\boldsymbol{x}}(\theta_0)} > c \right\} \end{aligned}$$

   Il quesito posto richiedeva di verificare l'equivalenza delle disuguaglianze

$$\frac{L_{\boldsymbol{x}}(\theta_0)}{L_{\boldsymbol{x}}(\theta_1)} \geq c \Longleftrightarrow \sum_{i=1}^{n} x_i \leq k$$

   per opportuni valori delle soglie c e k. Infatti, dall'espressione della fuznione di verosimiglianza del modello bernoulliano

$$L_{\boldsymbol{x}}(\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i} = \frac{\theta^{\sum_{i=1}^n x_i}}{(1-\theta)^{\sum_{i=1}^n x_i}} (1-\theta)^n = \left(\frac{\theta}{1-\theta} \right)^{\sum_{i=1}^n x_i} (1-\theta)^n$$

si ha

$$
\begin{aligned}
& \frac{L_{\boldsymbol{x}}(\theta_0)}{L_{\boldsymbol{x}}(\theta_1)} \geq c \\
\iff \quad & \frac{\left(\frac{\theta_0}{1-\theta_0}\right)^{\sum_{i=1}^n x_i} (1-\theta_0)^n}{\left(\frac{\theta_1}{1-\theta_1}\right)^{\sum_{i=1}^n x_i} (1-\theta_1)^n} \geq c \\
\iff \quad & \left(\frac{\frac{\theta_0}{1-\theta_0}}{\frac{\theta_1}{1-\theta_1}}\right)^{\sum_{i=1}^n x_i} \geq c \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \qquad (*) \\
\iff \quad & \left(\sum_{i=1}^n x_i\right) \log\left(\frac{\frac{\theta_0}{1-\theta_0}}{\frac{\theta_1}{1-\theta_1}}\right) \geq \log c + n \log\left(\frac{1-\theta_1}{1-\theta_0}\right) \qquad (**) \\
\iff \quad & \sum_{i=1}^n x_i \leq \frac{\log c + n \log\left(\frac{1-\theta_1}{1-\theta_0}\right)}{\log\left(\frac{\frac{\theta_0}{1-\theta_0}}{\frac{\theta_1}{1-\theta_1}}\right)} \qquad (***) \\
\iff \quad & \sum_{i=1}^n x_i \leq k
\end{aligned}
$$

utilizzando l'opportuna soglia

$$
k = \frac{\log c + n \log\left(\frac{1-\theta_1}{1-\theta_0}\right)}{\log\left(\frac{\frac{\theta_0}{1-\theta_0}}{\frac{\theta_1}{1-\theta_1}}\right)}
$$

legata alla soglia di partenza c dalla relazione appena scritta. Osserviamo che dalla disuguaglianza in (**) passiamo a quella opposta in (***) dal momento che se $\theta_0 < \theta_1$ allora anche il rapporto degli *odds* mantiene lo stesso segno di disuguaglianza ovvero $\frac{\theta_0}{1-\theta_0} < \frac{\theta_1}{1-\theta_1}$ e dunque dalla disuguaglianza (**) abbiamo diviso entrambi i termini per un numero negativo, precisamente per

$$
\log\left(\frac{\frac{\theta_0}{1-\theta_0}}{\frac{\theta_1}{1-\theta_1}}\right) < 0.
$$

Era possibile evitare i passaggi (**) e (***) semplicemente osservando che l'espressione a sinistra della disuguaglianza

$$
\left(\frac{\frac{\theta_0}{1-\theta_0}}{\frac{\theta_1}{1-\theta_1}}\right)^{\sum_{i=1}^n x_i}
$$

è una funzione monotona decrescente della quantità ad esponente ovvero di

$$
\sum_{i=1}^n x_i
$$

e dunque la disuguaglianza (*) è soddisfatta se e soltanto se è soddisfatta l'ultima disuguaglianza per un'opportuna soglia k.

2. Assumendo $n = 2$, si ha che

$$A = \{\mathbf{x}_2 : \ x_1 + x_2 \leq k\}.$$

Calcolare la probabilità di errore di I specie, α, e la potenza del test, $1 - \beta$, che si ottengono ponendo $k = 1$ e assumendo $\theta_0 = \frac{1}{2}$, $\theta_1 = \frac{3}{4}$.

Soluzione – Ricordando che la somma S_n di n v.c. bernoulliane di parametro θ indipendenti e identicamente distribuite ha distribuzione binomiale di parametri (n, θ) per calcolare la probabilità di prima specie si farà riferimento alla seguente

$$\begin{aligned}
\alpha &= Pr\left\{(X_1, X_2) \in R = A^c; \theta = \theta_0 = \frac{1}{2}\right\} \\
&= Pr\left\{S_2 = X_1 + X_2 > 1; \theta_0 = \frac{1}{2}\right\} \\
&= Pr\left\{Bin\left(2, \frac{1}{2}\right) > 1\right\} \\
&= Pr\left\{Bin\left(2, \frac{1}{2}\right) = 2\right\} \\
&= \binom{2}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0 \\
&= \frac{1}{4} \\
&= 0.25
\end{aligned}$$

Per la potenza si dovrà invece utilizzare la distribuzione della statistica test S_n considerando come parametro il valore assunto sotto l'ipotesi alternativa

$$Bin(n, \theta_1) = Bin\left(2, \frac{3}{4}\right)$$

e dunque

$$\begin{aligned}
1 - \beta &= Pr\left\{(X_1, X_2) \in R = A^c; \theta = \theta_1 = \frac{3}{4}\right\} \\
&= Pr\left\{S_2 = X_1 + X_2 > 1; \theta_1 = \frac{3}{4}\right\} \\
&= Pr\left\{Bin\left(2, \frac{3}{4}\right) > 1\right\} \\
&= Pr\left\{Bin\left(2, \frac{3}{4}\right) = 2\right\} \\
&= \binom{2}{2}\left(\frac{3}{4}\right)^2\left(\frac{3}{4}\right)^0 \\
&= \left(\frac{3}{4}\right)^2 \\
&= 0.5625
\end{aligned}$$

3. Assumendo ora $n = 20$, determinare l'approssimazione normale per la statistica S_n sotto l'ipotesi alternativa ($\theta_1 = \frac{3}{4}$) e determinare il valore di k in (1) tale che il test abbia una potenza pari a 0.9.

Soluzione – Dal teorema del limite centrale possiamo dedurre che

$$\bar{X}_n = nS_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

dove

$$\begin{aligned} \mu &= E[X_i] = \theta \\ \sigma^2 &= \mathrm{Var}[X_i] = \theta(1-\theta) \end{aligned}$$

e quindi

$$S_n \sim N\left(n\theta, n\theta(1-\theta)\right).$$

Usando la suddetta approssimazione asintotica per la distribuzione della statistica S_n per $n = 20$ e sotto l'ipotesi alternativa in cui $\theta = \theta_1 = \frac{3}{4}$ otteniamo che la potenza approssimata del test con generica soglia k sarà data da

$$\begin{aligned} & Pr\left\{S_n > k; \theta_1 = \frac{3}{4}\right\} \\ & = Pr\left\{N\left(n\theta_1, n\theta_1(1-\theta_1)\right) > k\right\} \\ \\ & = Pr\left\{N\left(15, 3.75\right) > k\right\} \\ & = Pr\left\{N\left(0,1\right) > \frac{k-15}{\sqrt{3.75}}\right\} \\ & = 1 - F_{N(0,1)}\left(\frac{k-15}{\sqrt{3.75}}\right) \end{aligned}$$

e dunque imporre l'uguaglianza

$$0.9 = 1 - \beta = Pr\left\{S_n > k; \theta_1 = \frac{3}{4}\right\} = 1 - F_{N(0,1)}\left(\frac{k-15}{\sqrt{3.75}}\right)$$

equivale a risolvere l'equazione

$$0.1 = F_{N(0,1)}\left(\frac{k-15}{\sqrt{3.75}}\right).$$

Usando la notazione $z_{0.1}$ per il quantile della distribuzione normale standard $N(0,1)$ a livello di probabilità 0.1 si deve infine risolvere la seguente equazione in k

$$\frac{k-15}{\sqrt{3.75}} = z_{0.1} = -1.28$$

e dunque la soluzione al problema posto è data dal più piccolo intero che supera il valore

$$k^* = -1.28 \cdot \sqrt{3.75} + 15 = 12.52.$$

4. Supponendo di avere osservato un campione in cui S_{20} risulta pari a 15, determinare un intervallo di confidenza approssimato di livello $1-\alpha = 0.95$ per θ.

Soluzione – Dall' approssimazione

$$\bar{X}_n = nS_n \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

utilizzando al posto di $\theta(1-\theta)$ una sua stima consistente e precisamente $\bar{x}(1-\bar{x})$ si ottiene

$$1-\alpha \quad = \quad Pr\left\{-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\theta}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \leq z_{1-\frac{\alpha}{2}}\right\}$$

e quindi

$$1-\alpha \quad = \quad Pr\left\{\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq \theta \leq \bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right\}.$$

Poiché $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$, $n = 20$, $\bar{X}_n = S_n/n = 15/20 = 0.75$ e $\frac{\bar{X}_n(1-\bar{X}_n)}{n} = \frac{0.75 \cdot 0.25}{20} = 0.009375$ si ottiene dunque la determinazione dell'intervallo

$$IC_{0.95}[\theta] = [0.75 - 1.96 \cdot 0.009375, 0.75 + 1.96 \cdot 0.009375] = [0.5602, 0.9398]$$

N.B. Era possibile anche non sostituire la varianza teorica con una stima consistente ed ottenere un intervallo detto di Wilson (cfr. Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association, 22, 209-212 1927) ovvero del metodo basato sullo *score*) risolvendo la disequazione di secondo grado corrispondente alla seguente

$$\begin{aligned}
1-\alpha &= Pr\left\{-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq z_{1-\frac{\alpha}{2}}\right\} \\
&= Pr\left\{\frac{\left(\bar{X}-\theta\right)^2}{\frac{\theta(1-\theta)}{n}} \leq z^2_{1-\frac{\alpha}{2}}\right\} \\
&= Pr\left\{\left(\bar{X}-\theta\right)^2 \leq z^2_{1-\frac{\alpha}{2}}\frac{\theta(1-\theta)}{n}\right\} \\
&= Pr\left\{n\left(\bar{X}-\theta\right)^2 \leq z^2_{1-\frac{\alpha}{2}}\theta(1-\theta)\right\}
\end{aligned}$$