Analisi dei dati
Esercizi

Aggiornati al 3 maggio 2018

1 Parte C

1. Sia dato un campione bernoulliano $Y_1, \cdots, Y_n$, $n \geq 2$, da una variabile casuale Y uniforme nell'intervallo $(0, \theta)$, $\theta > 0$.

   (a) Trovare lo stimatore di θ secondo il metodo dei momenti.

   (b) Sia T_1 lo stimatore calcolato al punto precedente, calcolare $EQM(T_1)$.

   (c) Mostrare che lo stimatore è consistente.

   (d) Trovare la costante a per cui lo stimatore

   $$T_2 = \frac{Y_1 + aY_n}{n}$$

   è uno stimatore non distorto.

   (e) Se T_2 è lo stimatore non distorto, mostrare quale stimatore tra T_1 e T_2 è più efficiente.

2. Considerate i dati sulla velocità della luce e supponete che questi costituiscano un campione bernoulliano da una v.c. continua Y.

   (a) Calcolate la stima per sostituzione della mediana ?

   (b) Calcolate l'intervallo di confidenza secondo il *bootstrap* basato sulla quantità pivot.

3. Si considerino i dati contenuti nel file `billings.txt` e si voglia costruire un intervallo di confidenza del 90% per la varianza.

   (a) Valutate la distorsione dei due stimatori

   $$T_1 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad T_2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

   (b) Costruite gli intervalli di confidenza per T_1 T_2 basati sui percentili

4. (tratto da Wasserman) Sia $n = 50$ e $T(F) = \int (y - \mu)^3 dF(y)/\sigma^3$ l'indice di asimmetria (con $\mu = \int y dF(y)$ e $\sigma^2 = \int y^2 dF(y) - \mu^2$) e $X_1, \ldots, X_n$ un campione bernoulliano da una v.c. $\mathcal{N}(0, 1)$ e si ponga $Y_i = e^{X_i}$. Costruite i tre tipi di intervalli di confidenza basati del bootstrap per $T(F)$.

5. '*Women who are union members earn* 2.50$ *per hour more than women who are not union members*". (The Wall Street Journal, July 26, 1994). Sembrerebbe quindi che per le donne statunitensi sia conveniente far parte di un sindacato. Per verificare l'affermazione del *Wall Street Journal* abbiamo scelto due campioni indipendenti di lavoratrici del settore industriale. Il primo campione è costituito da 15 lavoratrici iscritte ad un sindacato, mentre il secondo campione è costituito da 20 lavoratrici che non fanno parte di nessun sindacato. Su ciascuna unità statistica (la lavoratrice) abbiamo misurato il salario orario (in $). I dati sono contenuti nel dataset `workers.txt` (variabili `salary` e `union`).

(a) Considerate separatamente le due classi di lavoratrici e verificate se ci sono dei valori anomali.

(b) Verificate se potete supporre che i dati siano assimibilabili ad un campione da una v.c. normale

(c) Se $X \sim F_X$ rappresenta la variabile casuale salario delle donne iscritte e la distribuzione della e $Y \sim F_Y$ rappresenta la variabile casuale salario delle donne non iscritte, si proponga uno stimatore per sostituzione del parametro $\theta = \mathrm{E}_{F_X}(X) - \mathrm{E}_{F_Y}(Y)$.

(d) Sia $\widehat{\theta}_n$ lo stimatore ottenuto al punto precedente. Si scriva una prodedura bootstrap per il calcolo della varianza dello stimatore.

(e) Si calcoli un intervallo di confidenza per θ di livello 0.99.

(f) Si stimi con il metodo bootstrap la probabilità che $\mathrm{P}(\widehat{\theta}_n > 0)$.

6. Sia $y_1, \ldots, y_n$ un campione bernoulliano da una v.c. $\mathcal{U}(0, \theta)$.

   (a) Determinate una stima secondo il metodo dei momenti per θ.

   (b) Se il campione fosse estratto da $\mathcal{U}(\theta_1, \theta_2)$ con $\theta_1 < \theta_2$, cosa cambierebbe nella stima di $\theta = (\theta_1, \theta_2)'$?

7. Si consideri un'urna, contenente palline bianche e rosse. Sono ammesse tre possibili percentuali di palline rosse, precisamente $\Theta = \{0.2,\ 0.6,\ 0.8\}$. Data un'estrazione bernoulliana di tre palline, che restituisce una pallina bianca e due rosse, determinare la stima di massima verosimiglianza del parametro θ.

8. Sia $y_1, \ldots, y_n$ un campione bernoulliano da una v.c. Poisson(θ), $\theta > 0$.

   (a) Determinate una stima secondo il metodo dei momenti per θ.

   (b) Supponete che il parametro sia $\phi = \mathrm{P}(Y = 0)$, determinate una stima di ϕ, secondo il metodo dei momenti.

9. Si supponga che $y_1, \ldots, y_n$ sia un campione bernoulliano da una v.c. $Be(\theta)$

   (a) Si determini la stima di massima verosimiglianza per la probabilità di nascere femmina nel caso dei nati a Brisbane.

   (b) Si calcoli $J(\theta)$ e $I(\theta)$

10. Si supponga che $y_1, \ldots, y_n$ sia un campione bernoulliano da una v.c. $\mathcal{N}(\mu, \sigma^2)$

    (a) Si determini la stima di massima verosimiglianza per μ e σ^2 nel caso dei dati sulla velocità della luce.

    (b) Si supponga che μ sia noto e pari a 27. Si determini la stima di massima verosimiglianza per σ_2.

    (c) Si calcoli $J(\sigma^2)$ e $I(\sigma^2)$

11. Si assuma che $y_1, \ldots, y_n$ siano realizzazioni di variabili casuali continue indipendenti con densità comune

$$f(y; \beta) = \frac{1}{\beta^2}\,\frac{1}{y^3}\,\exp\left\{-\frac{1}{\beta y}\right\}\,, \qquad \beta > 0,\, y > 0\,.$$

    (a) Si indichi lo spazio parametrico si scriva la funzione di log-verosimiglianza per β.

    (b) Si ottenga lo stimatore di massima verosimiglianza, $\hat{\beta}$, per β e se ne calcoli il valore per i dati osservati $(1.6, 3.9, 1.5, 2.2, 2.8, 2.7, 1.3, 0.4)$.

    (c) Si ottengano la matrice di informazione osservata e la matrice di informazione attesa.

    (d) Si fornisca un'approssimazione per la distribuzione dello stimatore di massima verosimiglianza $\hat{\beta}$ utilizzando l'informazione osservata di Fisher.

    (e) Si sfrutti l'approssimazione ottenuta al punto precedente per ottenere un intervallo di confidenza con livello approssimato 0.95 per β.

12. Il *file* `KNMI_20160831.txt` contiene le precipitazioni giornaliere rilevate dal 1906 al 2016 nella stazione di De Bilt (Olanda)

    (a) Si estraggano tutte le rilevazioni giornaliere rispetto al mese di aprile

    (b) Si analizzino i dati per evidenziare se vi sono dati anomali.

    (c) Si adatti una distribuzione parametrica ai dati chiarendo le ipotesi sottostanti.

    (d) Si confronti la distribuzione parametrica stimata con la distribuzione empirica.

    (e) Si considerino ora le medie mensili e si ripeta l'analisi

    (f) Quale teorema della probabilità entra in gioco ?

13. Si vuole studiare l'affidabilità di un componente elettronico, cioè la probabilità che esso superi un tempo di missione prefissato. Allo scopo, vengono testati 420 componenti. Di questi 25 non superano il tempo di missione.

    (a) Stimate la probabilità che un componente superi il tempo di missione.

    (b) Costruite un intervallo di confidenza, ad un livello del 97%, della probabilità suddetta.

    (c) Costruite un intervallo di confidenza, ad un livello del 97%, della probabilità suddetta con il metodo *bootstrap*.

14. Il file `kevlar90.txt` contiene dei dati sulla resistenza del Kevlar 49/epoxy un materiale utilizzato nello Space Shuttle. Vengono riportati in tempi di resistenza (in ore) di 101 fili testati ad un livello di stress del 90%.

    (a) Si analizzino i dati per evidenziare se vi sono dati anomali.

    (b) Si adatti una distribuzione parametrica ai dati chiarendo le ipotesi sottostanti.

    (c) Si confronti la distribuzione parametrica stimata con la distribuzione empirica.

    (d) Si utilizzi la teoria asintotica per lo stimatore di massima verosimiglianza per ottenere un intervallo di confidenza per il parametro

    (e) Si utilizzi il *bootstrap* parametrico per ottenere un intervallo di confidenza per il parametro

15. I file `haliburton.txt` and `macdonalds.txt` contengono i rendimenti mensili sulle azioni di queste due compagnie dal 1975 al 1999.

    (a) Si analizzino i dati per evidenziare se vi sono dati anomali.

    (b) Si adatti una distribuzione parametrica ai dati chiarendo le ipotesi sottostanti.

    (c) Si confronti la distribuzione parametrica stimata con la distribuzione empirica.

    (d) In finanza si è interessati alla volatilità del titolo misurata attraverso la sua varianza e il suo scarto interquartile. Si calcoli una stima delle due quantità.

    (e) Quale delle due compagnie risulta più variabile ?

    (f) Si utilizzi il *bootstrap* non parametrico per ottenere un intervallo di confidenza per lo scarto interquartile.

    (g) Si utilizzi il *bootstrap* parametrico per ottenere un intervallo di confidenza per lo scarto interquartile.

16. Il file `Basilea.txt` contiene i dati sulle precipitationi (in mm) invernali nella stazione di Basilea (Svizzera) per il periodo 1902-2011 (fonte: MeteoSwiss, Zurigo, Svizzera)

    (a) Si analizzino i dati per evidenziare se vi sono dati anomali.

    (b) Si adatti una distribuzione normale ai dati chiarendo le ipotesi sottostanti.

    (c) Se la precipitazione segue una distribuzione normale con quale probabilità si supererà la soglia di 165 mm ?

(d) In alternativa si utilizzi la distribuzione gamma e si stimino i parametri con il metodo dei momenti e con il metodo della massima verosimiglianza.

(e) Cosa potete dire dell'adattamento rispetto al modello con distribuzione normale ?

(f) Si utilizzi il *bootstrap* parametrico per ottenere un intervallo di confidenza per la probabilità di superare la soglia di 165 mm.

(g) Si utilizzi il *bootstrap* non parametrico per ottenere un intervallo di confidenza per la probabilità di superare la soglia di 165 mm.

17. Si assuma che $y_1, \ldots, y_n$ sia un campione bernoulliano di una variabile casuale Y avente densità

$$f(y;\theta) = \theta y^{\theta-1}, \qquad 0 < y < 1,$$

con $\theta > 0$ parametro ignoto.

(a) Si dimostri che $\mathrm{E}(Y) = \theta/(1+\theta)$.

(b) In base al risultato precedente si stimi θ con il metodo dei momenti.

(c) Si scriva la funzione di verosimiglianza e log-verosimiglianza per θ.

(d) Si calcoli lo stimatore di massima verosimiglianza per θ, che sarà denotato con $\hat{\theta}$.

(e) Si calcoli l'informazione attesa di Fisher

(f) Si può dimostrare che $-\log(Y)$ si distribuisce come una v.c. esponenziale con valore atteso $1/\theta$.

i. Si provi che $1/\hat{\theta}$ è uno stimatore non distorto di $1/\theta$.

ii. Si dica se $\hat{\theta}$ è uno stimatore consistente.

18. Due ricercatori hanno esaminato il colore nelle piume della coda di 16 esemplari di una specie di picchio. Gli uccelli avevano una piuma "singolare" (`Odd`) che era di colore o lunghezza diversa dal resto del piumaggio della coda, probabilmente perché era ricresciuta dopo essere stata persa. Hanno misurato la gradazione di giallo di una piuma "singolare" prelevata da ogni uccello e l'hanno confrontata con la gradazione di giallo di una piuma "tipica" (`Typical`) dello stesso uccello. Nel *file* `piume.txt` sono contenuti i dati. Lo scopo dell'indagine era quello di vedere se vi era una differenza nella gradazione. In particolare si è interessati a stimare una differenza media

(a) Si cerchi di rispondere al quesito principale con un'analisi puramente descrittiva.

(b) Si descriva l'esperimento casuale e il numero di osservazioni.

(c) Si costruisca un intervallo di confidenza per la differenza media basato sul *bootstrap*.

(d) Si specifichi un modello statistico, le ipotesi sottostanti, lo spazio campionario e lo spazio parametrico.

(e) Si costruisca un intervallo di confidenza per la differenza media tenendo conto della distribuzione dei dati.

19. Si consideri il *file* `USStates.csv`

(a) Si utilizzi la *library* per disegnare un opportuno grafico della variabile `Smokers`, la percentuale di residenti che fuma. Si commenti la forma della distribuzione e il suo valore centrale.

(b) Si calcoli la proporzione media di fumatori

(c) Si scriva una funzione in R che estrae 10000 campioni casuali di 5 stati and ritorna come output la proporzione media di fumatori per ogni campione. Si salvi l'output in un vettore `estimates`

(d) Cosa rappresenta l'output di tale funzione ?

(e) Si calcoli la media aritmetica e lo scarto quadratico medio di `estimates` e utilizzate `ggplot2` per rappresentare la densità e la funzione di ripartizione empirica. Si commenti la forma della distribuzione e il suo valore centrale.

(f) Quale dovrebbe essere la relazione teorica tra lo scarto quadratico di `Smokers` e `estimates`. Vi sembra rispettata empiricamente ?

(g) Un campione di cinque stati è Arkansas, Florida, Pennsylvania, California, e Vermont. Si calcoli la proporzione media di questi stati.

(h) Si utilizzi il valore ricavato per costruire un intervallo di confidenza di livello 95% della proporzione media dei fumatori

(i) L'intervallo contiene la 'vera' proporzione media.

(j) Quale proporzione dei 10000 valori di `estimates` in realtà contiene il vero valore ?

(k) Quale proporzione ci si attende ? Si è vicini a questo valore