Analisi dei dati
Esercizi

Aggiornati al 30 marzo 2018

1 Parte B

1. Considerate il lancio di due dati e considerate la somma delle due facce. Sia Y la v.c. 'somma delle due facce'

   (a) Individuate lo spazio campionario.

   (b) Costruite un programma in `R` che estragga un campione bernoulliano di numerosità n dalla v.c. somma delle due facce.

   (c) Calcolate la distribuzione di probabilità di Y.

   (d) Calcolate la distribuzione di probabilità empirica di Y, con $n = 50$.

   (e) Confrontate le due distribuzioni.

   (f) Calcolate la funzione di ripartizione e la funzione di ripartizione empirica e sovrapponetele in un'unico grafico.

   (g) Commentate quanto accade se $n = 500$.

2. Sia $Y_1, \ldots, Y_n$ un campione bernoulliano da una v.c. $\mathcal{N}(\mu, \sigma^2)$ e si voglia stimare μ. Per questo vengono considerati due stimatori $\widehat{\mu}_{A,n} = \sum_{i=1}^{n-1} Y_i/n$, e $\widehat{\mu}_{B,n} = Y_1/n + (n-1)Y_n/n$. Nel seguito si suppone che σ^2 sia noto.

   - Calcolate la distorsione dei due stimatori.
   - Calcolate il loro errore standard.
   - Fornite una stima del loro errore standard.
   - Calcolate il loro errore quadratico medio.

3. Siano $(Y_1, \cdots, Y_4)$, quattro v. c. di Bernoulli indipendenti con probabilità di successo $\Pr\{Y_i = 1\} = \theta$, $0 < \theta < 1$. Si considerino i seguenti stimatori di θ:

$$T_1 = \frac{2}{3}Y_1 + \frac{1}{3}Y_4, \qquad T_2 = 3Y_2 - 2Y_3.$$

   (a) Dite quale dei due stimatori risulta preferibile.

   (b) Mostrate che lo stimatore

$$T_3 = (2 - a)T_1 + (a - 1)T_2$$

   è non distorto per qualsiasi valore di a.

   (c) Trovate il valore di a che rende minima la varianza di questo stimatore.

4. Riprendete l'esercizio precedente e fissate $n = 50$, $\mu = 4$ e $\sigma = 1$. Nel seguito utilizzate `R` .

   - Simulate $m = 1000$ campioni di numerosità n. Ogni campione sarà denotato con $(y_1^{(j)}, \ldots, y_n^{(j)})'$, $j = 1, \ldots, m$.

- Per ogni campione $(y_1^{(j)}, \ldots, y_n^{(j)})'$ calcolate le due stime $\widehat{\mu}_A^{(j)}$ e $\widehat{\mu}_B^{(j)}$, ottenendo i vettori

$$(\widehat{\mu}_A^{(1)}, \ldots, \widehat{\mu}_A^{(m)})' \qquad (\widehat{\mu}_B^{(1)}, \ldots, \widehat{\mu}_B^{(m)})'$$

- Calcolate le distorsioni ottenute in simulazione $\sum_{j=1}^m \widehat{\mu}_A^{(j)}/m - \mu$ e $\sum_{j=1}^m \widehat{\mu}_B^{(j)}/m - \mu$ e confrontatele con quelle teoriche si commentino i risultati.
- Si calcolino le varianze e lo scarto quadratico medio di $(\widehat{\mu}_A^{(1)}, \ldots, \widehat{\mu}_A^{(m)})'$ e $(\widehat{\mu}_B^{(1)}, \ldots, \widehat{\mu}_B^{(m)})'$ e si commentino i risultati.
- Si calcoli sulla base di $(\widehat{\mu}_A^{(1)}, \ldots, \widehat{\mu}_A^{(m)})'$ e $(\widehat{\mu}_B^{(1)}, \ldots, \widehat{\mu}_B^{(m)})'$ una misura dell'errore quadratico medio e si commentino i risultati.
- Riflettete sul ruolo di n e m. Cosa ottenete al variare dei due valori ?

5. Sia Y una *v.c.* di Poisson di parametro λ. Sia dato un campione bernoulliano $Y_1, \ldots, Y_n$, $n > 3$. Si considerino i seguenti stimatori di λ:

$$T_1 = \frac{\sum_{i=1}^{n-2} Y_i}{n-1} + \frac{Y_n}{n-1}, \quad T_2 = \frac{Y_1 + (n-2)Y_n}{n-1}.$$

   (a) Calcolare le distorsioni dei due stimatori proposti.

   (b) Calcolare le varianze dei due stimatori.

   (c) Calcolare gli errori quadratici medi.

   (d) Quale dei due stimatori risulta preferibile ?

6. Siano $Y_1, \cdots, Y_n$, n variabili casuali i.i.d. normali con valore atteso θ e varianza θ^2 e si considerino i seguenti stimatori di θ:

$$T_1 = \frac{Y_1 + (n-1)Y_n}{n}, \qquad T_2 = \frac{Y_1 + \ldots + Y_{n-1}}{n}$$

   (a) Dite quale dei due stimatori risulta preferibile secondo il criterio dell'errore quadratico medio.

7. Sia $y_1, \ldots, y_n$ un campione bernoulliano da una v.c. Y continua e si consideri la stima della funzione di densità

$$\widehat{f}_n(t) = \frac{1}{n\Delta} \sum_{i=1}^{n} K\left(\frac{t - y_i}{\Delta}\right).$$

   dove $K(t)$ è un prefissato nucleo. Si chiede di dimostrare che

   (a) $\widehat{f}_n(t)$ è effettivamente una funzione di densità ovvero $\widehat{f}_n(t) \geq 0$ e $\int_{-\infty}^{\infty} \widehat{f}_n(t)dt = 1$;

   (b) $\int_{-\infty}^{\infty} t\widehat{f}_n(t)dt = \overline{y}$.

8. Un amministratore ospedaliero che spera di migliorare i tempi di attesa decide di stimare il tempo medio di attesa del pronto soccorso nel suo ospedale. Raccoglie un campione casuale di 64 pazienti e determina il tempo (in minuti) tra il momento del *check-in* al pronto soccorso fino a quando i pazienti non sono stati visti per la prima volta da un medico. Un intervallo di confidenza del 95% basato su questo campione è (128 minuti, 147 minuti), che si basa sul modello normale per la media.

   Determina se le seguenti affermazioni sono vere o false e spiega il tuo ragionamento.

   (a) Questo intervallo di confidenza non è valido poiché non sappiamo se la distribuzione della popolazione dei tempi di attesa ER sia quasi normale.

   (b) Siamo condfidenti al 95% che il tempo medio di attesa di questi 64 pazienti del pronto soccorso sia tra i 128 ei 147 minuti.

   (c) Siamo confidenti al 95% che il tempo di attesa medio di tutti i pazienti nel pronto soccorso di questo ospedale sia tra i 128 e i 147 minuti.

   (d) il 95% dei campioni casuali ha una media campionaria compresa tra 128 e 147 minuti.

(e) Un intervallo di confidenza del 99% sarebbe più ristretto dell'intervallo di confidenza del 95% poiché dovremmo essere più sicuri della nostra stima.

(f) Il margine di errore è 9.5 e la media campionaria è 137.5.

(g) Al fine di ridurre il margine di errore di un intervallo di confidenza del 95% a metà di quello che è ora, dovremmo raddoppiare la dimensione del campione.

9. Sia $Y_1, \ldots, Y_n$ un campione bernoulliano da una v.c. Y e che μ sia un parametro della distribuzione (non necessariamente il valore atteso). Si supponga che $L_n = g(Y_1, \ldots, Y_n)$ e $U_n = h(Y_1, \ldots, Y_n)$ siano tali che $\mathrm{P}(L_N < \mu < U_n) = 0.95$ per ogni valore di μ.

   (a) Si supponga che il parametro sia $\theta = 3\mu + 7$, derivate un intervallo di confidenza di livello 0.95 per θ

   (b) Si supponga che $\theta = 1 - \mu$. Costruite un intervallo di confidenza per θ di livello 0.95.

   (c) Sia $\theta = \exp(\mu)$. Costruite, se possibile, un intervallo di confidenza per θ di livello 0.95.

   (d) Sia $\theta = \mu^2$. Costruite, se possibile, un intervallo di confidenza per θ di livello 0.95.

10. Sia $y_1, \ldots, y_n$ un campione bernoulliano da una v.c. $Y \sim \mathcal{N}(\mu, 3)$.

    (a) Derivate un intervallo di confidenza per μ di livello 0.99

    (b) Quante osservazioni sono necessarie affinché la lunghezza dell'intervallo di confidenza per μ di livello 0.99 sia pari a 0.57 ?

11. L'esercizio mira ad una verifica empirica del seguente risultato teorico:

$$\text{Se } Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2) \text{ allora } \textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2_{n-1}$$

    e prevede la realizzazione di un programma in `R` .

    (a) Considerate dei campione bernoulliano da una v.c. $\mathcal{N}(-2, 4)$ di numerosità $n = 20$.

    (b) Simulate un grande numero di questi, ad esempio 1000.

    (c) Calcolate per ognuno $t = \sum_{i=1}^n (y_i - \overline{y})^2$.

    (d) Verificate empircamente con l'aiuto di un istogramma e con la costruzione di un qq-plot che i valori di t/σ^2 si distribuiscono come χ^2_{n-1}

    (e) Ripetete la procedura per $n = 2$ e considerate la stima non parametrica della funzione di densità. Quali inconvenienti registrate ? Come potreste ovviare a questi ?

12. Considerate i dati sulla velocità della luce e supponete che questi costituiscano un campione bernoulliano da una v.c. $\mathcal{N}(\mu, \sigma^2)$

    (a) Derivate un intervallo di confidenza per σ^2 di livello 0.99.

    (b) In base al campione quanto vale l'intervallo di confidenza ?

13. La stagione di vendita al dettaglio delle vacanze del 2009, che ha segnato il 27 novembre 2009 (il giorno successivo al Giorno del Ringraziamento), è stata contrassegnata da una spesa per consumi auto-segnalata leggermente inferiore a quella osservata nel periodo comparabile del 2008. Per ottenere una stima della spesa dei consumatori, sono stati intervistati 436 adulti americani campionati a caso. E' stata esaminata la spesa giornaliera dei consumatori per il periodo di sei giorni successivo al Ringraziamento, che si è estesa al weekend del Black Friday e al Cyber Monday e i dati sono contenuti nel *file* `thanksgiving_spend.csv`.

    (a) Importate il file in R.

    (b) Disegnate un istogramma dei dati e confrontatelo con una stima della funzione di densità.

    (c) Proponete un intervallo di confidenza per la spesa media.

    (d) Quali potrebbero essere gli elementi di debolezza della vostra scelta ?

    (e) Trasformate ora i dati y_i secondo queste tre opzioni: a) $\log(y_i)$, b) $\sqrt{y_i}$, c) $y_i^{1/3}$.

    (f) A vostro parere quale di queste trasformazioni rende la distribuzione dei dati trasformati più simile a quella di una v.c. normale ?