Statistica Computazionale

Esempio di compito (4)

È richiesta una relazione contenente il codice R, debitamente commentato, necessario per risolvere i problemi posti. Le procedure usate devono essere giustificate a parole. Anche i grafici, se presenti, utilizzati nella discussione devono essere inseriti nell'elaborato.

Dove richiesto, devono essere usate 10000 replicazioni (Monte Carlo o bootstrap).

Non è possibile consultare alcun materiale, tranne l'"help" di R.

Il tempo a disposizione è di 135 minuti.

Esercizio 1

Si consideri una variabile aleatoria Y con funzione di densità[1]

$$f(y) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{y^2}{2}\right) , \qquad y > 0 .$$

a. Si scriva una funzione in R che genera valori di Y con l'algoritmo accetto/rifiuto, usando come proposte valori generati dalla distribuzione esponenziale di media 1, che ha funzione di densità[2]

$$g(y) = \exp(-y) , \qquad y > 0 .$$

Si calcoli anche il tasso di accettazione (la frazione di volte in cui l'algoritmo accetta).

b. Si verifichi graficamente la correttezza della funzione di tipo "r" scritta al punto precedente[3].

c. La variabile aleatoria X ha, condizionatamente a $Y = y$ distribuzione esponenziale di media y. Usando anche l'algoritmo sviluppato al punto **a** si approssimi, via Monte Carlo[4], la media *marginale* di $\sqrt{X}$. Si fornisca un intervallo di confidenza per l'approssimazione.

d. La radice quadrata di una distribuzione esponenziale ha distribuzione di Rayleigh. Da questo discende che

$$E(\sqrt{X}|y) = \frac{1}{2} \cdot \sqrt{\pi \cdot y} .$$

Si approssimi la media *marginale* di $\sqrt{X}$ usando Rao-Blackwell. Anche in questo caso, si fornisca un intervallo di confidenza per l'approssimazione. Si commentino i risultati.

[1] In R, il valore π si ottiene semplicemente con `pi`.

[2] Si noti che, in R, per ottenere n valori da tale densità è sufficiente usare `rexp(n)`.

[3] Può essere utile il comando `hist(y,prob=T,add=T,nclass=100)`, dove `y` contiene i valori generati.

[4] Si richiede la *soluzione diretta*. Per generare n valori da una esponenziale di media y si usi `rexp(n,1/y)`.

Esercizio 2

Il file `esercizio_2.dat` contiene osservazioni relative a due variabili casuali X e Y.

a. Si usi il test t di Student a due campioni, con correzione di Welch, per verificare l'ipotesi

$$H_0 \colon E(X) = E(Y) \qquad \text{contro} \qquad H_1 \colon E(X) \neq E(Y)$$

Si verifichino poi, con il test di Shapiro-Wilk, le assunzioni di normalità di X e Y.

b. Si usi un approccio bootstrap non parametrico per verificare la stessa ipotesi del punto precedente, fornendo il valore-p del test.

c. Si usi un approccio bootstrap non parametrico, fornendo il valore-p del test, per verificare l'ipotesi

$$H_0 \colon \text{Varianza}(X) = \text{Varianza}(Y) \qquad \text{contro} \qquad H_1 \colon \text{Varianza}(X) \neq \text{Varianza}(Y)$$

d. Si fornisca l'intervallo di confidenza bootstrap (non studentizzato), con grado di fiducia 0.95, per $\theta = E(X) - E(Y)$. Si usi poi questo intervallo per verificare l'ipotesi di cui al punto **a**.

Esercizio 3

Il file `esercizio_3.dat` contiene 70 coppie di osservazioni (x_i, y_i).

a. Si stimi, usando la funzione `spline.regression`[5], una regressione non parametrica basata sulle spline, usando un numero di parametri equivalenti pari a 5. Si commenti il risultato, dicendo se e perché 5 sembra essere un valore troppo grande o troppo piccolo per il numero di parametri equivalenti.

b. Si determini, con il metodo della convalida incrociata, il numero ottimale *df* di parametri equivalenti[6].

c. Un metodo alternativo, e meno computazionalmente oneroso, per determinare *df* è fornito dal criterio *GCV* (*generalized cross validation*), definito da

$$GCV = \frac{s^2(df)}{(1 - df/n)^2} \ ,$$

dove $s^2(\cdot)$ indica l'errore di adattamento. Si scriva una funzione che determina il valore ottimale di *df* usando il criterio *GCV*[7]. Si confronti il valore ottimale di *df* con quello ottenuto al punto **b**.

[5] La funzione `spline.regression` è contenuta nel file `spline.regression.R`, messo a disposizione. Il numero di parametri equivalenti è selezionato con l'argomento `df`.

[6] Si ricorda che la stima della curva non parametrica in corrispondenza del punto `z` è contenuta nell'oggetto `estimate` estratto con `spline.regression(x, y, df=df, display="none", eval.points=z)$estimate`.

[7] Si ricorda che le stime della curva non parametrica in corrispondenza del vettore `x` di valori osservati per la variabile esplicativa sono contenute nell'oggetto `estimate` estratto con `spline.regression(x, y, df=df, display="none", eval.points=x)$estimate`.