R per l'analisi Statistica Multivariata

Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Docente: Tommaso Rigon

9 Settembre 2021

Informazioni

Il tempo a disposizione del candidato è di **2 ore**. Si ricordi di firmare tutti i fogli che si è intenzionati a consegnare (se presenti) con il proprio nome, cognome e numero di matricola.

Prova d'esame

1. **Problema**
   Si supponga che X sia una variabile casuale Beta (classe di funzioni `*beta`) di parametri di forma (`shape1, shape2`) tali che $a = 0.5, b = 0.5$. Inoltre, sia Y una variabile casuale tale che Y condizionata a $X = x$ si distribuisce come una Binomiale di parametri $n = 10$ e probabilità di successo $\pi = x$.

   (a) (5pt) Calcolare via simulazione il valore atteso $E(Y^2)$.

   Grazie alle proprietà della distribuzione Binomiale, si ottiene che

   $$E(Y^2 \mid X = x) = n^2x^2 + nx(1 - x)$$

   (a) (5pt) Si sfrutti questo risultato per ottenere una stima alternativa (ma equivalente) del valore atteso $E(Y^2)$.

   Nota. Si consegni il file .R che produce le risposte alle domande richieste. Si risponda inoltre alle domande aperte direttamente in tale file, avendo cura di commentare con un cancelletto (`#`) quanto scritto.

2. **Problema**
   Si ponga $X_1 = 1$. Si consideri una collezione di variabile casuali **binarie ed indipendenti** $X_1, \ldots, X_n$, tali per per cui

   $$\mathrm{P}(X_i = 1) = \frac{1}{1 + (i - 1)^{1-\sigma}}, \qquad i = 2, \ldots, n,$$

   dove $0 < \sigma < 1$ è un parametro positivo. Inoltre, si definisca

   $$S = \sum_{i=1}^{n} X_i = 1 + \sum_{i=2}^{n} X_i.$$

   Nota. La variabile aleatoria S **non** segue una distribuzione binomale, perchè le probabilità di successo sono diverse tra loro.

(a) (4pt) Si scriva in **R** la funzione `rS(R, n, sigma)` che simula un `R` valori pseudo-casuali distribuiti come la variable S.

(b) (2pt) Utilizzando $n = 100$ e $\sigma = 1/2$, si ottenga una stima Monte Carlo del valore atteso $\mathrm{E}(S)$, utilizzando un numero di repliche `R` appropriato.

(c) (2pt) Utilizzando $n = 500$ e $\sigma = 1/2$, si ottenga una stima Monte Carlo dell'evento $\mathrm{P}(30 \leq S \leq 40)$, utilizzando un numero di repliche `R` appropriato.

(d) (2pt) Utilizzando $n = 500$ e $\sigma = 1/2$, si ottenga una stima Monte Carlo della distribuzione di S e se ne faccia un grafico, utilizzando un numero di repliche `R` appropriato.

Nota. Si consegni il file .R che produce le risposte alle domande richieste. Si risponda inoltre alle domande aperte direttamente in tale file, avendo cura di commentare con un cancelletto (`#`) quanto scritto.

3. **Problema**
La varianza campionaria dei dati $\mathbf{x} = (x_1, \ldots, x_n)$ è definita come

$$\mathrm{var}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

dove $\bar{x}$ è la media campionaria. Si noti che $\mathrm{var}(\mathbf{x})$ ammette la rappresentazione alternativa

$$\mathrm{var}(\mathbf{x}) = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2.$$

(a) (2pt) Si scriva una funzione `var2(x)` che calcola la varianza di $\mathbf{x}$ utilizzando la prima definizione.

(b) (4pt) Si scriva quindi una funzione `var3(x)` che calcola la varianza di $\mathbf{x}$ utilizzando la formula basata sulle distanze tra coppie di elementi.

(c) (1pt) Si utilizzino i dati `x <- c(5, 10, 8, 8, 22)`. Si verifichi che le due funzioni `var2(x)` e `var3(x)` forniscono lo stesso risultato.

(d) (1pt) Si supponga ora che `x <- 1:1500`. Si notano differenze rispetto al punto precedente?

(e) (2pt) Si confrontino le funzioni `var2` e `var3` con la funzione `var` implementata in **R**, utilizzando i dati `x <- c(5, 10, 8, 8, 22)`. Come mai i risultati differiscono, anche se di poco?

Nota. Si consegni il file .R che produce le risposte alle domande richieste. Si risponda inoltre alle domande aperte direttamente in tale file, avendo cura di commentare con un cancelletto (`#`) quanto scritto.