Livellamento al Software R - Esame

Andrea Gilardi

2022-10-19

Informazioni generali

Le soluzioni agli esercizi proposti in questo documento devono essere riportate in un file con estensione `.R` ed inviate entro le 23:59 del 23 Ottobre 2022 al seguente indirizzo email tramite l'account di ateneo: *andrea.gilardi@unimib.it*. Vi chiedo cortesemente di rinominare il file con le soluzioni seguendo il pattern `nome-cognome-matricola.R` e specificare `Soluzioni Esame Livellamento R` come oggetto della mail.

Nel caso in cui utilizziate funzioni definite in pacchetti `R` non presentati a lezioni, dovrete illustrarne il funzionamento in fase di discussione orale.

La prova orale verrà fissata indicativamente 5/10 giorni dopo la consegna degli esercizi.

Per qualsiasi dubbio sull'interpretazione dei quesiti o domanda non esitate a contattarmi.

Esercizio 1

Dato il seguente vettore

```
y <- c(0.849, 0.723, 0.696, 0.430, 0.702, 0.651, 0.844, 0.120, 0.781, 0.417, 0.885, 0.543)
```

1. si calcoli la media aritmetica, la varianza, il minimo ed il massimo di `y`;
2. dopo aver convertito i primi 9 elementi di `y` in una matrice chiamata `M` avente 3 righe e 3 colonne, se ne calcoli il determinante, il rango (SUGGERIMENTO: provate a leggere la pagina di help delle funzioni `rank()` e `qr()`) e la traccia;
3. quanto valgono gli autovalori di `M'M`?
4. si calcoli la somma degli elementi di `M` posizionati fuori dalla diagonale maggiore;
5. si definisca una funzione `R` che permetta il calcolo dell'indice di Curtosi. Si calcoli quindi tale indicatore per la variabile `y`. Cosa possiamo concludere riguardo la sua distribuzione?

Supponiamo ora che `y` rappresenti un campione casuale estratto da una variabile casuale `Y` avente supporto in $[0, 1]$ e funzione di densità pari a:

$$f_Y(y) = 2y$$

6. si implementino in R tre funzioni denominate `d_myf`, `p_myf` e `q_myf` che permettano di ricavare la funzione di densità, la CDF, ed i quantili di Y. NB: Le tre funzioni devono restituire un messaggio di errore informativo nel caso in cui l'input non sia di tipo `numeric`. Non è necessario testare che l'input giaccia nel supporto di `Y`.
7. quanto vale $P(Y \leq 0.5)$? E qual è il quantile di ordine 0.6 di Y?
8. si calcoli il valore della funzione di densità $f_Y(y)$ per ogni elemento del vettore `y`.
9. si rappresenti la ECDF del vettore `y` e si sovrapponga al grafico ottenuto la CDF teorica tematizzando opportunamente le due curve.

10. si dimostri che la funzione di densità da voi implementata (i.e. `d_myf`) integra ad 1 se valutata su tutto il supporto di Y. (SUGGERIMENTO: Provate a controllare la pagina di help e gli esempi della funzione `integrate()`).

Esercizio 2

Sia $X_1, \ldots, X_n$ un campione casuale estratto da $X \sim f_X(x; \theta)$, variabile casuale unidimensionale avente media μ e varianza σ^2 entrambe finite. In maniera un pochino informale, potremmo dire che, grazie al Teorema Centrale del Limite, lo stimatore *media campionaria* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ è approssimabile con una variabile casuale normale avente media μ e varianza σ^2/n per $n >> 0$.

Si sviluppi uno script R per provare a verificare empiricamente e graficamente questo teorema nei seguenti casi:

- $X_1, \ldots, X_n$ i.i.d. ad $X \sim \text{Poisson}(\lambda)$ dove $n \in \{10, 100, 500\}$.
- $X_1, \ldots, X_n$ i.i.d. ad $X \sim \text{Beta}(\alpha, 1)$ dove $n \in \{100, 500\}$.

SUGGERIMENTO: Oltre agli strumenti visti a lezione, provate a consultare la pagina di help relativa alla funzione `qqplot`.

Esercizio 3

E' il 14 Marzo 2023. Mentre vi apprestate a trascorrere un tranquillo pomeriggio primaverile ad oziare sul vostro divano guardando Netflix, ricevete una visita da parte della Morte. Ella vi comunica che, purtroppo, è giunta la vostra ora... Dopo tante suppliche e lunghe contrattazioni, riuscite a convincerla a posticipare la vostra dipartita siglando il seguente accordo.

A partire da quel momento, la Morte passerà a trovarvi il 14 Marzo di ogni anno per l'eternità. Giunta alla vostra porta, lancerà una coppia di dadi e, nel caso in cui esca "doppio 1", allora verrete trascinati via da Lei; altrimenti potrete continuare a vivere fino all'anno successivo.

Tuttavia, i due dadi hanno una struttura speciale. Il primo anno hanno solamente due facce (così che la probabilità di morire è pari a 1/4). Il secondo anno hanno 4 facce, il terzo anno ne hanno 6 e così via per sempre. Di conseguenza, nonostante la Morte continuerà a venire alla vostra porta per l'eternità, le vostre probabilità di sopravvivere aumenteranno di anno in anno.

L'accordo stipulato con la Morte è sorprendentemente vantaggioso! L'obiettivo di questo esercizio è sviluppare una simulazione in `R` per calcolare la probabilità di sopravvivere in eterno dopo aver stipulato l'accordo.

SUGGERIMENTO: La funzione `sample()` può essere usata per generare estrazioni casuali da un vettore di elementi. Ad esempio:

```
set.seed(2)
dado <- c(1, 2)
estrazione <- sample(x = dado, size = 2, replace = TRUE)
estrazione # RIP
```

```
## [1] 1 1
```

La probabilità di morire dopo 1 anno può essere approssimata nel seguente modo:

```
n_simulazioni <- 1e5
risultati <- logical(length = n_simulazioni)
for (i in seq_along(risultati)) {
  estrazione <- sample(dado, size = 2, replace = TRUE)
  risultati[i] <- all(estrazione == c(1, 1)) # Perchè dobbiamo usare all()?
}
mean(risultati)
```

```
## [1] 0.24967
```

SPOILER: E' possibile dimostrare che la probabilità di sopravvivere in eterno è esattamente pari a $\frac{2}{\pi}$. In particolare, questo problema è stato preso da https://twitter.com/3blue1brown/status/1503423352207147010 e vi rimando al thread per maggiori dettagli sulla dimostrazione matematica.

Esercizio 4

Dopo aver caricato il pacchetto `MASS` ed il dataset `Cars93` tramite il seguente comando

```
library(MASS)
data(Cars93)
```

si risponda ai seguenti quesiti.

1. Quante righe ha il dataset? Quante colonne?
2. Dopo aver selezionato unicamente le vetture prodotte negli USA, si stampi a schermo e si commenti la "struttura" del dataset risultante.
3. Si rappresenti l'istogramma della variabile `MPG.city` aggiungendo anche la stima di densità non-parametrica opportunamente tematizzata.
4. Si commenti e si produca il grafico a dispersione delle variabili `Price` e `MPG.city`. Ci sono valori anomali? Sapreste individuarli e descriverli?
5. Dopo aver selezionato unicamente le auto aventi 4, 6, o 8 cilindri, si produca una rappresentazione grafica per descrivere la relazione tra le variabili `Cylnders` e `MPG.city`. SUGGERIMENTO: Potrebbe essere utile consultare l'help delle funzioni `droplevels()` e `boxplot()`.
6. Si calcoli il valore medio della variabile `Price` per ogni livello di `Manufacturer`. SUGGERIMENTO: Potrebbe essere utile consultare l'help della funzione `tapply()` ed i relativi esempi.