Assegnamento Finale - Corso di Livellamento R

Creare uno script di R e salvarlo come "Cognome_Nome.R". La prima riga di tale script deve essere un commento del tipo:

```
# Cognome Nome Matricola
```

Risolvere i seguenti esercizi e riportarne le soluzioni ed eventuali commenti richiesti sullo script, opportunamente commentato ed inviarlo a roberto.ascari@unimib.it entro **e non oltre** le ore 12 del 18 ottobre 2021.

Esercizio 1

Si importi il dataset `animal.csv`, contenente informazioni su 38 specie di animali. In particolare si hanno le seguenti variabili:

1. `specie_tr`: specie dell'animale
2. `peso_corpo`: peso del corpo, in Kg
3. `peso_cerv`: peso del cervello, in grammi
4. `leggero`: numero di ore di sonno leggero (ore al giorno)
5. `profondo`: numero di ore di sonno profondo (ore al giorno)
6. `anni_vita`: speranza di vita alla nascita, in anni
7. `gestazione`: tempo di gestazione, in giorni
8. `preda`: indice qualitativo di predazione (0 = "raramente è una preda", 1 = "è molto spesso una preda")
9. `esposto`: indice qualitativo di esposizione durante il sonno (0 = "poco esposto", 1 = "molto esposto")
10. `pericolo`: indice qualitativo di pericolo (0 = "gli altri animali non rappresentano un pericolo per lui", 1 = "gli altri animali rappresentano un grave pericolo per lui")

N.B.: Prima di procedere con l'esercizio, assicurarsi che le variabili siano state importate in un formato idoneo.

- Si calcoli una nuova variabile "sonno", rappresentante il numero totale di ore di sonno giornaliere.
- Quale animale dorme complessivamente di più? E quale dorme di meno?
- Si scelga una variabile qualitativa (Q) e due quantitative tra quelle disponibili. Tra le due variabili quantitative scelte, si distingua tra una variabile risposta/dipendente (Y) ed una esplicativa/indipendente (X).
  - Si proponga un opportuno grafico per studiare la relazione tra Y ed X. Commentare il risultato.
  - Si proponga un opportuno grafico per studiare la relazione tra Y e Q. Commentare il risultato.
  - Si ritiene possibile proporre un grafico per studiare la relazione tra Y ed X, tenendo conto anche della presenza della variabile Q? Se si, si riporti tale grafico e lo si commenti. Altrimenti si motivi il perché non lo si ritiene possibile.

Esercizio 2

Si considerino due variabili casuali indipendenti $Y_1 \sim \text{Gamma}(\alpha_1, \beta)$ e $Y_2 \sim \text{Gamma}(\alpha_2, \beta)$, la cui generica funzione di densità è data da:

$$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \qquad y > 0, \alpha > 0, \beta > 0.$$

Dalla letteratura è noto che la variabile casuale

$$X = \frac{Y_1}{Y_1 + Y_2} \in (0, 1) \tag{1}$$

è distribuita come una $\text{Beta}(\alpha_1, \alpha_2)$, indipendentemente dal valore assunto dal parametro β, purché comune alle due Gamma. Per semplicità, di seguito si consideri $\beta = 1$. La funzione di densità di una variabile casuale $\text{Beta}(\alpha_1, \alpha_2)$ è la seguente:

$$f(x; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}, \qquad x \in (0, 1), \alpha_1 > 0, \alpha_2 > 0.$$

- Implementare una **funzione** che riceva come argomenti un numero intero $n \geq 1$ ed i numeri reali positivi α_1 e α_2. Tale funzione dovrà generare n valori secondo la relazione (1) (estrarre quindi prima valori da Y_1 ed Y_2 ed utilizzarli per determinare X). Inserire dei controlli sui sugli argomenti della funzione in modo tale che la funzione porti ad un messaggio di errore qualora l'utente scegliesse un argomento non idoneo.

- Impostare come seme per il generatore di numeri casuali la propria matricola. Si utilizzi la funzione definita al punto precedente per estrarre $n = 2000$ osservazioni pseudo-casuali da una variabile $\text{Beta}(5, 10)$. Si rappresenti la distribuzione del campione estratto mediante un istogramma e si sovraimponga la funzione di densità della corrispondente Beta (utilizzare una funzione già definita in R oppure scrivere una propria funzione). Si commenti il grafico.

- Impostare una finestra grafica con due righe e due colonne. Rappresentare la funzione di densità della variabile casuale Beta, scegliendo i valori di α_1 e α_2 secondo i seguenti scenari:

  (i) $\alpha_1 = \alpha_2$;

  (ii) $\alpha_1 > 1$ e $\alpha_2 < 1$;

  (iii) $\alpha_1 < 1$ e $\alpha_2 < 1$ $(\alpha_1 \neq \alpha_2)$;

  (iv) $\alpha_1 > 1$ e $\alpha_2 > 1$ $(\alpha_1 \neq \alpha_2)$.

  In che modo i due parametri α_1 e α_2 sembrano influenzare la distribuzione?
  (N.B.: **non** viene chiesto di rappresentare anche un istogramma relativo a dati simulati.)

Esercizio 3

Si supponga che in una stanza ci siano $k \in \mathbb{N}$ persone. Qual è la probabilità che due di esse condividano il compleanno? Questo esercizio si ispira al 'paradosso dei compleanni' (`https://en.wikipedia.org/wiki/Birthday_problem`). Per rispondere a questo quesito vi viene chiesto di impostare una simulazione.

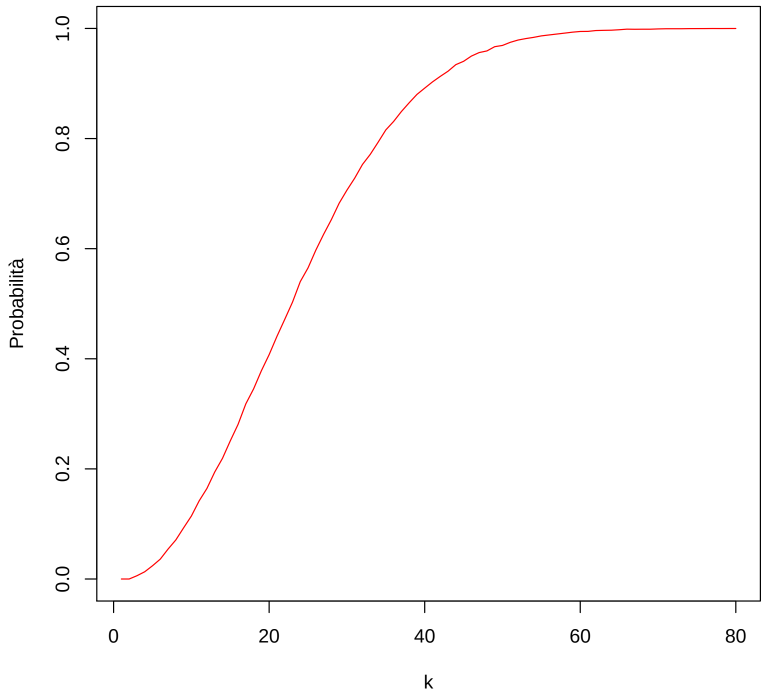
- L'utente deve poter impostare il numero k di persone nella stanza ed il numero di iterazioni.

- Per generare i compleanni casualmente, utilizzare la funzione `sample(x = 1:365, size = k, replace = T)` (per dettagli: `help(sample)`) (per semplicità si considerano unicamente anni composti da 365 giorni).

- Per stimare la probabilità, dividere il numero di scenari nei quali almeno una coppia di persone condivide il compleanno per il numero di iterazioni effettuato:

$$\hat{\text{Prob}}_k = \frac{\text{Numero scenari in cui almeno due persone su } k \text{ condividono il compleanno}}{\text{Numero scenari considerati}}$$

- Utilizzare la simulazione appena implementata per costruire un grafico che permetta di studiare come il numero k di persone nella stanza influenza la probabilità stimata che almeno due tra le k persone condividano il compleanno. Per stimare queste probabilità utilizzare almeno 20000 iterazioni. Impostare il grafico in modo tale che si considerino soltanto valori di k minori di 80. Un possibile grafico è quello riportato di seguito, ma altre rappresentazioni grafiche sono ammesse. Commentare il grafico ottenuto. Perché si parla di "paradosso dei compleanni"?



Esercizio 4

In una foresta incantata, alla fine del giorno 0 vengono piantati 6 semi magici. Dal giorno 1 in avanti, alla fine di ogni giornata, ogni seme ha una probabilità del 50% di trasformarsi in un albero. Una volta trasformato, resterà un albero per sempre. Si implementi una simulazione per approssimare il numero **atteso** (medio) di giorni necessari affinché tutti i semi diventino degli alberi.