Proiect la Probabilități și Statistică

-partea I-

Important!

1) **Fiecare problemă este obligatorie și are asociată un punctaj de 5 puncte.**
2) **Rezolvările vor fi realizate în limbajul R**
3) **Respectați restricțiile din cerințe, dar acolo unde nu sunt precizări făcute aveți libertatea de a face orice alegere considerați potrivită**
4) **Trimiteți până la data de 31.05. 2024 ora 22:00 la adresa simona.cojocea@fmi.unibuc.ro o singura arhivă per echipă care conține codul sursa R împreună cu documentația asociată**
5) **Documentația va conține, în mod obligatoriu, numele membrilor echipei, prezentarea soluției, comentarii legate de rezultatele obținute sau de dificultățile întâmpinate, grafice și tabele și respectiv codul sursa comentat la fiecare cerință în parte.**

Cerințe

1. a) Construiți un *dataframe*(cu cel puțin 100 înregistrări) care să conțină toate concertele dintr-un an, organizate după prețul biletului, genul muzical, data, orașul în care au loc, artiștii(puși într-o structură de tip listă) și orice altă informație considerați relevantă.

   b) Construiți o funcție ***f1*** care, pentru un buget dat(introdus de utilizator) și un interval de timp(sau o reuniune de intervale) vă oferă combinații de evenimente la care puteți merge.

   c) Adaugați dataframe-ului inițial o coloană ***soldout***(cu opțiunile ***da/nu***). Construiți o funcție care, la rularea funcției ***f1***(pe dataframe-ul actualizat, care conține și coloana ***soldout***) să genereze automat cu o probabilitate de 1%, valoarea ***da*** pentru coloana soldout. Odată ce această poziție este ocupată cu valoarea ***da*** ea nu mai poate fi schimbată de o altă rulare a funcției(atenție la cum implementați aceasta restricție!).

   d) Pentru unul din concertele din dataframe-ul vostru se suplimentează bilete după ce, în prealabil, a fost declarat ***soldout***. Ce soluție tehnică găsiți de a schimba informația despre concertul respectiv *fără a dezactiva restricția de la c)?*

   e) Construiți o nouă funcție de filtrare care să ia în calcul și dacă respectivul concert e sau nu ***soldout***.

2. Despre ***linia 501 STB*** se cunosc următoarele:
   -*în medie* călătoresc **512 călători pe zi** (medie raportată la datele dintr-o luna pentru un singur tramvai)
   -*numărul minim* înregistrat în perioada studiată este de **210 călători pe zi** iar *numărul maxim* este **983 calatori pe zi**
   -*în medie*, **20%** din zilele dintr-o lună sunt clasificate ca fiind ***lejere***(mai puțin de **350 de calatori pe zi**), **50%** ca fiind ***normale***(între **351 și 670 calatori pe zi**), iar **30%** ca fiind ***aglomerate***(peste **671 calatori pe zi**).
   -*prețul* unui bilet este **3 lei** și, *în medie*, ***74%*** din pasagerii care *nu* au abonament platesc biletul la utilizarea tramvaiului
   -*prețul* unui abonament este **70 lei pe lună** și, *în medie*, **38%** din pasagerii care călătoresc cu tramvaiul îl achiziționează

   a) Generați, prin simulare, valori care să reprezinte numărul de călători dintr-o zi, pentru fiecare zi a lunii iunie 2024, respectând restricțiile de mai sus și stocați valorile obținute într-un vector. Construiți histograma acestor valori.
   b) Repetați procedul de la a) pentru fiecare luna a unul 2024 și centralizați rezultatele empirice într-un dataframe care să conțină, pentru fiecare lună valorile medii, minime și maxime de călători, precum și procentul de zile lejere, normale și respectiv aglomerate înregistrate.
   c) Completați dataframe-ul de la b) cu simularea nr de pasageri cu abonament, nr de pasageri care platesc bilet și respectiv număr de pasageri care nu plătesc bilet. Determinați pentru fiecare lună în parte veniturile provenite din bilete și abonamente și respectiv, veniturile nerealizate prin neplata biletului de unii dintre pasageri. Organizați informația într-o manieră ușor de vizualizat.
   d) Un tramvai de pe linia lui 501 face 11 trasee complete în timpul programului de lucru dintr-o zi. De două ori pe zi un controlor se urcă în unul din tramvaie și solicită prezentarea biletelor de călătorie unui număr de pasageri, aleși în mod aleator, după următorul algoritm:
      -dacă e o zi **lejeră**, verifică în mod aleator între 2 și 11 persoane, dar se oprește din verificare dacă a amendat deja 3 persoane.
      -dacă e zi **normală**, verifică în mod aleator un număr de persoane până reușește să amendeze 5 persoane(sau a verificat pe toata lumea prezentă în tramvai între 2 stații)
      -dacă e o zi **aglomerată**, verifică în mod aleator între 3 și 5 persoane și se oprește din verificare după prima amendă

      Știind că amenda este **50 lei**, determinați, în urma simulării, pentru fiecare zi a fiecarei luni din anul 2024, câți bani se strâng din aplicarea unor amenzi. Comparați această sumă cu pierderea realizată prin neplata biletelor și stabiliți în câte zile dintr-o lună(în medie) sumele obținute din amenzi depășesc pierderea prin neplata biletelor, considerând și faptul că pentru fiecare din cele 2 verificări zilnice există un cost asociat controlorului de **214 lei**.
   e) Studiați prin simulare, oportunitatea de a introduce un al treilea control pe zi(în condițiile menționate anterior, la care adaugam informația ca, *în medie*, 30% din amenzile colectate la fiecare din controale nu sunt raportate oficial ci sunt păstrate de controlor).

3. Construiți o funcție în R care să se comporte ca un *generator de numere aleatoare*(se generează n valori, unde n este dat de utilizator) având următoarele specificații:

   1) Pentru prima valoare(x_1) se citește timpul sistemului, se ia numărul format din minute și secunde(t_1) și se calculează modulo 23. (**ex**. Daca ora sistemului este 12:15:23 atunci t_1=1523)
   2) Dacă **t_1 mod 23=0** atunci x_1 se generează folosind funcția **rnorm** din R cu parametri dați de numărul minutelor și respectiv numărul secundelor.
   3) Dacă **t_1 mod 23=3** atunci x_1 se generează folosind funcția **rpois** din R cu parametru dat de număr reprezentând minutele și se adună la el un număr y_1 generat cu funcția **runif** din R(cu parametrii -1 și 1) .
   4) Dacă **t_1 mod 23=5** atunci x_1 se generează folosind funcția **rexp** din R cu parametru dat de numărul reprezentat de ora sistemului.
   5) Dacă **t_1 mod 23=7** atunci x_1 se generează folosind funcția **rbinom** din R cu parametrii dați de ora sistemului și 1/nr_minute și se adună la el un număr y_1 generat cu funcția **runif** din R(cu parametrii 0 și 5) .
   6) Dacă **t_1 mod 23=8** atunci x_1 se generează folosind funcția **runif** din R(de parametri -5 și 7).
   7) Dacă **t_1 mod 23=11** atunci x_1 se generează folosind funcția **rgamma** din R și se scade din el un număr y_1 generat cu funcția **rhyper** din R.
   8) În celelalte cazuri se reia procesul și se citește din nou ora sistemului. Dacă procesul a fost reluat de 2 ori și nu s-a intrat pe unul dintre cazurile de mai sus atunci x_1 se generează folosind funcția **rnorm** din R cu parametrii 0 și 1.

   Pentru valorile x_n cu n>1 se folosește următoarea formulă de recurență:
   $x_n=a* x_{n-1}+b$ unde a este o valoare generată cu funcția **rexp** din R de parametru 5, iar b este o valoare generată cu funcția rnorm din R de parametri 2 și 1

   Funcția returnează un vector cu valorile generate și realizează histograma lor.