Laborator 4 - Simulare. Metode de tip Monte Carlo.

I. Estimarea ariilor şi a volumelor

RStudio. Nu uitaţi să va setaţi directorul de lucru: **Session → Set Working Directory → Choose Directory**.

Exerciţiu rezolvat. Aria discului unitate este π. Acoperim discul cu un pătrat de dimensiuni 2 pe 2 şi estimăm numărul π folosind 10000, 50000 şi 100000 valori uniforme aleatoare. Comparăm apoi rezultatele cu valoarea cunoscută a lui $\pi = 3.14159265358...$

Discul unitate este inlcus în $[-1, 1] \times [-1, 1]$. Următoarea funcţie estimează π utilizând N numere aleatoare.

```
disc_area = function(N) {
  N_C = 0;
  for(i in 1:N) {
    x = runif(1, -1, 1);
    y = runif(1, -1, 1);
    if(x*x + y*y <= 1)
      N_C = N_C + 1;
    }
  return(4*N_C/N); }
}
```

Dacă am estimat o valoare α_{actual} prin metoda Monte Carlo şi obţinem α_{MC}, putem măsura eroarea făcută (aceea de a folosi α_{MC} în loc de α_{actual}) în cel puţin două moduri:

- **Eroarea absolută**: $\epsilon_{abs} = |\alpha_{MC} - \alpha_{actual}|$.
- **Eroarea relativă**: $\epsilon_{rel} = \dfrac{|\alpha_{MC} - \alpha_{actual}|}{|\alpha_{actual}|}$. Această avaloare poate fi scrisă şi procentual, obţinând **eroarea procentuală**: $\epsilon_{per} = \epsilon_{rel} \cdot 100\%$.

Exercţii propuse.

I.1. Estimaţi volumul sferei unitate (care este $4\pi/3$) folosind eşantioane de numere aleatoare de dimensuni diferite şi apoi calculaţi erorile (absolută şi relativă) corespunzătoare.

I.2. Estimaţi aria următoarei elipse (care este 2π)

$$E = \left\{(x, y) \in \mathbb{R}^2 \ : \ x^2 + 4y^2 \leqslant 4\right\}.$$

Folosiţi eşantioane de dimensiuni diferite şi calculaţi erorile absolută şi procentuală.

I.3. Estimaţi aria dintre parabola de ecuaţie $y = -x^2 + 2x + 3$ şi axa Ox (abscisă) - folosind 10000 valori uniforme. Determinaţi aria exactă prin integrare şi calculaţi eroarea relativă.

Indicaţie: parabola intersectează axa Ox în punctele $(-1, 0)$ şi $(3, 0)$ şi are vârful în $(1, 4)$. Un domeniu rectangular din planul real care acoperă această arie poate fi $[-1, 3] \times [0, 4]$.

II. Integrarea Monte Carlo

Exerciţiu rezolvat. Estimaţi valoarea următoarei integrale folosind 20000 şi apoi 50000 de valori aleatoare uniforme (determinaţi 30 astfel de aproximări pentru fiecare din cele două dimensiuni şi calculaţi câte o medie şi câte o deviaţie standard).

$$\int_0^{10} e^{-u^2/2}\, du.$$

Următoarea funcţie oferă estimare pentru un eşantion de dimensiune N

```
MC_integration = function(N) {
   sum = 0;
   for(i in 1:N) {
      u = runif(1, 0, 10);
      sum = sum + exp(-u*u/2;
   }
   return(10*sum/N);
}
```

Putem calcula o medie pentru $k = 30$ astfel de aproximări şi şi deviaţia standard corespunzătoare folosind următoarea funcţie.

```
MC_integr_average= function(k, N) {
   for(i in 1:k)
      estimates[i] = MC_integration(N);
   print(mean(estimates));
   print(sd(estimates));
}
```

În urma execuţiei acestei funcţii obţinem

```
> MC_integr_average(30, 20000
[1] 1.249768
[1] 0.02327472
> MC_integr_average(30, 50000
[1] 1.253072
[1] 0.01373724
```

Exerciţiu rezolvat. Estimaţi valoarea următoarei integrale folosind 20000 şi apoi 50000 de valori aleatoare uniforme (determinaţi 30 astfel de aproximări pentru fiecare din cele două dimensiuni şi calculaţi câte o medie şi câte o deviaţie standard), utilizând metoda MC îmbunătăţită, anume cu distribuţia exponenţială ($\lambda = 1$)

$$\int_0^{+\infty} e^{-u^2}\, du.$$

(Valoarea exactă acestei integrale este $\sqrt{\pi}/2 \approx= 0.8862269$.)
Mai întâi, următoarea funcţie oferă o estimare pentru un eşantion de dimensiune N

```
MC_improved_integration = function(N) {
   sum = 0;
   for(i in 1:N) {
      u = rexp(1, 1);
      sum = sum + exp(-u*u)/exp(-u);
   }
   return(sum/N);
}
```

Putem calcula o medie pentru $k = 30$ astfel de aproximări şi şi deviaţia standard corespunzătoare folosind următoarea funcţie.

```
MC_imprvd_integr_average= function(k, N) {
   for(i in 1:k)
      estimates[i] = MC_improved_integration(N);
   print(mean(estimates));
   print(sd(estimates));
}
```

În urma execuţiei acestei funcţii obţinem

```
> MC_imprvd_integr_average(30, 20000
[1] 0.8858024
[1] 0.002743676
> MC_imprvd_integr_average(30, 50000
[1] 0.8861285
[1] 0.00213069
```

Exercţii propuse.

II.1. Estimaţi valoarile următoarelor integrale şi comparaţi rezultatul cu valorile exacte (dacă sunt date):

$$(a)\int_0^{\pi}\cos^2 x\,dx=\frac{\pi}{2},(c)\int_0^{3}e^x\,dx=19.08554$$

II.2 Estimaţi valoarile următoarelor integrale şi comparaţi rezultatul cu valorile exacte şi calculaţi erorile absolute şi relative corespunzătoare

$$(a)\int_0^{+\infty}\frac{dx}{x^2+1}=\frac{\pi}{2};(b)\int_2^{+\infty}\frac{dx}{x^2-1}=\ln 2/2.$$

II.3 Estimaţi valoarea următoarei integrale utilizând metoda MC îmbunătăţită, cu distribuţia exponenţială ($\lambda=1, N=40000$)

$$\int_0^{+\infty}e^{-u^2/2}\,du=\sqrt{\pi/2}.$$

Comparaţi rezultatul cu valoarea exactă, şi calculaţi erorile absolute şi relative corespunzătoare. Determinaţi apoi 30 astfel de aproximări şi calculaţi o medie şi o deviaţie standard. (Vezi exerciţiul rezolvat.)

III. Estimarea mediilor

Exerciţiu rezolvat. Modelul stochastic pentru numărul de erori (bug-uri)găsite într-un nou produs software se poate descrie după cum urmează. Zilnic cei care testează produsul software testers determină un număr aleator de erori care sunt apoi corectate. Numărul de erori găsite în ziua i urmează o distribuţie Poisson(λ_i) al cărei parametru este cel mai mic număr de erori din cele două zile anterioare:

$$\lambda_i=\min\{X_{i-2},X_{i-1}\}.$$

Care este numărul mediu de zile în care sunt detectate toate erorile? (Presupunem că în primele două zile sunt găsite 31 şi 27 erori, respectiv.) Folosiţi $N=10000$ de simulări ("runs") pentru estimatorul Monte Carlo.

Generăm un număr de erori pentru fiecare zi, până când acest număr este 0. Următoarea funcţie oferă numărul de zile până când nu mai apar erori (pentru o singură simulare - "run").

```
Nr_days = function() {
  nr_days = 1;
  last_errors = c(27, 31);
  nr_errors = 27;
  while(nr_errors > 0) {
    lambda = min(last_errors);
    nr_errors = rpois(1, lambda);
    last_errors = c(nr_errors, last_errors[1]) ;
    nr_days = nr_days + 1;
  }
  return(nr_days);
}
```

Executăm această funcţie de $N = 10000$ de ori şi returnăm media obţinută

```
MC_nr_days = function(N) {
  s = 0;
  for(i in 1:N)
    s = s + Nr_days();
  return(s/N);
}
```

Rezultatul este 28.0686, astfel, în 4 săptămâni toate erorile vor fi găsite.

Exercţii propuse.

III.1 Rezolvaţi din nou exerciţiul anterior considerând că λ_i este media erorilor din cele trei zile anterioare (În primele trei zile numărul de erori găsite este 18, 16 şi 11, respectiv).

III.2 Doi mecanici schimbă filtrele de ulei pentru autoturisme într-un service. Timpul de servire este exponenţial cu parametrul $\lambda = 20$ hrs^{-1} în cazul primului mecanic şi $\lambda = 5$ hrs^{-1} în cazul celui de al doilea. Deoarece primul mecanic este mai rapid, el serveşte de 4 ori mai mulţi clienţi decât partenerul său. Astfel când un client ajunge la rând, probabilitatea de a fi servit de primul mecanic este 4/5. Fie X timpul de servire pentru un mecanic. Explicaţi cum se generează valori pentru X şi estimaţi media lui X.

IV. Estimarea probabilităţilor

Exerciţiu rezolvat. Modelul stochastic pentru numărul de erori (bug-uri) ăsite într-un nou produs software se poate descrie după cum urmează. Zilnic cei care testează produsul software testers determină un număr aleator de erori care sunt apoi corectate. Numărul de erori găsite în ziua i urmează o distribuţie Poisson(λ_i) al cărei parametru este cel mai mic număr de erori din cele trei zile anterioare:

$$\lambda_i = \min\{X_{i-3}, X_{i-2}, X_{i-1}\}.$$

(a) Estimaţi probabilitatea de a mai avea încă erori după 21 de zile de teste folosind 500 de simulări MC. (În primele trei zile numărul de erori găsite este 28, 22 şi 18, respectiv.).

(b) Estimaţi această probabilitate, cu o eroare de cel mult ± 0.01 cu probabilitate 0.95.

(a) Utilizăm $N = 5000$ simulări ("runs") Monte Carlo; în fiecare simulare generăm un număr de erori găsite în fiecare zi până când acest număr devine 0. Următoarea funcţie returnează numărul de zile până când nu mai există erori (la o singură simulare).

```
Nr_days = function() {
  nr_days = 2;
  last_errors = c(18, 22, 28);
  nr_errors = 18;
  while(nr_errors > 0) {
    lambda = min(last_errors);
    nr_errors = rpois(1, lambda);
    last_errors = c(nr_errors, last_errors[1:2]) ;
    nr_days = nr_days + 1;
  }
  return(nr_days);
}
```

Aplicăm această funcţie de $N = 5000$ ori şi returnăm proporţia valorilor care depăşesc (strict) 21 de zile.

```
MC_nr_days_21 = function(N) {
  s = 0;
  for(i in 1:N) {
    if(Nr_days() > 21) ;
      s = s + 1;
  }
  return(s/N);
}
```

Proporţia calculată, 0.246, este o estimare a probabilităţii de a mai avea erori nedetectate şi după 21 de zile de testare.

(b) Vom estima probabilitatea în două moduri.

Mai întâi, folosim o valoare "prezumată", $p^* = 0.246$, şi $N \geqslant p^*(1-p^*)\left(\frac{z_{\frac{\alpha}{2}}}{\epsilon}\right)^2$:

```
> alfa = 1 - 0.95
> z = qnorm(alfa/2)
> epsilon = 0.01
> p = 0.246
> N_min = p(1 - p)*(z/epsilon)^ 2
> N_min
[1] 7125.291
> MC_nr_days_21(N_min + 1)
[1] 0.2547264
```

Obţinem $N \geqslant 7125.291$ şi putem aplica MC_nr_days(N_min + 1)

A doua metodă utilizează minorantul $N \geqslant \left(\frac{z_{\frac{\alpha}{2}}}{2\epsilon}\right)^2$:

```
> alfa = 1 - 0.95
> z = qnorm(alfa/2)
> epsilon = 0.01
> p = 0.246
> N_min = (1/4)(z/epsilon)^ 2
> N_min
[1] 9603.647
> MC_nr_days(N_min + 1)
[1] 0.2496968
```

Obţinem $N \geqslant 9603.647$ şi putem aplica MC_nr_days(N_min + 1). De obicei, cu a cea de-a doua metodă numărul de simulări ("runs") este mai mare.

Exercţii propuse.

IV.1 Estimaţi probabilitatea $P(X > Y)$, unde X şi Y sunt variabile Poisson independente cu parametrii 3 şi 5, respectiv. Apoi estimaţi aceeaşi probabilitate cu o eroare care să nu depăşească ± 0.005 cu probablitate 0.95. Care ar trebui să fie numărul de simulări ("runs")?

IV.2 Douăzeci de computere sunt conectate printr-un LAN. Unul dintre ele este infectat de un anumit virus. În fiecare zi, acest virus ajunge de la un computer infectat la unul "curat" cu probabilitate 0.1. De asemeni, în fiecare zi (începând din a doua zi), administratorul de sistem alege la întâmplare două computere infectate şi îndepărtează virusul.

(a) Estimaţi probabilitatea ca într-o anumită zi toate computerele să fie infectate.

(b) Estimaţi probabilitatea ca într-o anumită zi cel puţin 8 computere să fie infectate.

(c) Estimaţi această ultimă probabilitate cu o eroare de ± 0.01 cu probabilitea 0.95.

Temă pentru acasă.

14 puncte [2p: C1 sau C2] + [2p: C3] + [5p: C4 sau C5] + [5p: C6 sau C7]

C1. (2 puncte) Volumul unui elipsoid (de revoluţie)

$$E(a,b,c) = \left\{ (u,v,w) \ : \ \frac{u^2}{a^2} + \frac{v^2}{b^2} + \frac{w^3}{c^2} \leqslant 1 \right\} \subseteq [-a,a] \times [-b,b] \times [-c,c]$$

este $\frac{4\pi}{3}abc$. Estimaţi acest volum utilizând metoda Monte Carlo pentru $a = 2, b = 3, c = 4$ şi comparaţi rezultatul cu valoarea exactă. Folosiţi eşantioane de dimensiune 10000, 20000 şi 50000 trials şi calculaţi erorile relative.

C2. (2 puncte) Fie T un triunghi ale cărui laturi sunt incluse în următoarele trei drepte: $x - 2y = 0$, $2x - 3y = 0$ şi $3x + 2y = 6$, respectiv. Estimaţi aria (necunoscută) acoperită de T folosind metoda Monte Carlo 10000 de simulări.

Indicaţie: punctele interioare triunghiului sunt caracterizate de următoarele inegalităţi: $x - 2y \geqslant 0$, $2x - 3y \leqslant 0$, *şi* $3x + 2y \geqslant 6$. *Este necesar să găsiţi o zonă rectangulară care să acopere triunghiul în întregime.*

C3. (2 puncte) Estimaţi valoarele următoarelor integrale şi comparaţi rezultatul cu valorile exacte (dacă sunt date):

$$(a) \int_0^1 \left[\cos(50x) + \sin(20x)\right]^2 dx, (b) \int_0^3 \frac{x^3}{x^4+1}\, dx = 1.01168;$$

C4. (5 puncte) Trei servere web oferă (servesc) aceleaşi pagini posibililor clienţi (web). Timpul necesar procesării unei cereri (request) HTTP este distribuit $\Gamma(5,3)$ pe primul server, $\Gamma(7,5)$ pe cel de-al doilea şi $\Gamma(5,2)$ pe al treilea (în miliseconde). La această durată se adaugă latenţa dintre client şi servere pe Internet care are o distribuţie exponenţială cu $\lambda = 1$ (în miliseconde). Se ştie că un client este direcţionat către primul server cu probabilitatea 0.5 şi către al doilea server cu probabilitatea 0.3. Estimaţi timpul mediu necesar servirii unui client (de la lansarea cererii până la primirea răspunsului).

C5. (5 puncte) Două zeci de computere sunt conectate într-un LAN. Un virus infectează această reţea în felul următor: în fiecare zi, acest virus ajunge de la un computer infectat la unul

"curat" cu probabilitate 0.15. De asemeni, în fiecare zi (începând cu a doua zi), administratorul de sistem îndepărtează virusul cu probabilitate 0.2 de pe unul dintre calculatoarele infectate alese la întâmplare

(a) Estimaţi numărul mediu de zile necesare îndepărtării virusului din întreaga reţea.

(b) Estimaţi numărul mediu de computere infectate.

C6. (5 puncte) O pădure formată din 1000 de copaci are forma unui dreptunghi de dimensiuni 50×20. În colţul nord-vestic (stânga-sus) ia naştere un incendiu. Vântul bate dinspre vest şi, cu probabilitate 0.8, orice copac ia foc de la vecinul să vestic (dacă acesta deja arde). Probabilităţile ca un copac să ia foc de la vecinul din estic, nordic sau sudic (dacă aceştia ard) sunt toate egale cu 0.3. întreprindeţi un studiu Monte Carlo pentru a estima probabilitatea ca cel puţin 30% din pădure să ardă la un moment dat.

C7. (5 puncte) Două zeci de studenţi au conturi pe un acelaşi server de mail. Contul unul student este infectat de un malware. În fiecare zi, acest malware se răspândeşte şi ajunge de la un cont infectat la unul neinfectat cu probabilitate 0.25. De asemeni, în fiecare zi (începând cu a doua zi), administratorul de sistem curăţă la întâmplare 5 conturi (sau toate conturile dacă sunt infectate mai puţin de 5).

(a) Estimaţi probabilitatea ca fiecare cont să fie infectat cel puţin o dată.

(b) Estimaţi probabilitatea ca într-o anumită zi toate conturile să fie infectate.

(c) Cât de mare trebuie să fie numărul de simulări ("runs") - adică dimensiunea eşantionului - astfel ca prima probabilitate să aibă o eroare de cel mult ± 0.01 cu probabilitatea 0.99? Găsiţi o astfel de estimare.

Rezolvările acestor exerciţii (funcţiile R şi apelurile lor) vor fi redactate într-un script R.