



## Table of contents

### 1 Simulare - metodele Monte Carlo (MC)

- Estimarea mediei cu metoda Monte Carlo
- Estimarea lungimilor, ariilor și a volumelor
  - Estimarea ariilor regiunilor cu frontiere necunoscute
- Integrarea Monte Carlo
- Estimarea probabilităților folosind metoda Monte Carlo

### 2 Bibliography

## Introducere

- Procesul de generare a valorilor aleatoare care au o anumită densitate se numește *simulare* (unii o numesc *simulare Monte Carlo*).  
*Statistics and Data with R by Y. Cohen, J. Y. Cohen*
- *Metodă Monte Carlo* este numită orice metodă care rezolvă o problemă prin generarea unor anumite valori aleatoare și observând fracțiunea acestor valori care au o anumită proprietate. Această metodă este utilă pentru a obține soluții numerice (aproximative) pentru probleme care sunt prea complicate pentru a fi rezolvate analitic. [mathworld.wolfram.com](http://mathworld.wolfram.com)
- O valoare a unei variabile aleatoare (sau o valoare care urmează o densitate) este numită *quantilă* sau *valoare aleatoare*, *număr aleator* (în engleză *variate*).

## Introducere

- O *metodă Monte Carlo* generează foarte multe astfel de valori aleatoare (câteodată milioane) asociate unei distribuții de probabilitate, iar acest proces se numește simulare a respectivei distribuții.
- Simularea este utilizată de exemplu pentru a determina media sau dispersia unei distribuții sau un alt parametru asociat.
- Simularea depinde de "calitatea" valorilor aleatoare. Cele mai folosite numere aleatoare sunt cele provenite din distribuția uniformă continuă standard,  $U(0, 1)$ , sau din distribuția uniformă discretă,  $U_n$ .
- Aproape orice limbaj de programare are un *generator de numere aleatoare*, dar aceste generatoare oferă doar numere pseudo-aleatoare sau quasi-aleatoare (valori uniforme).
- Unul dintre cele mai bune generatoare de numere pseudo-aleatoare (pseudorandom number generator - PRNG) este Mersenne-Twister (implicit în R).

## Estimarea mediei cu metoda Monte Carlo

- Fie  $X$  o variabilă aleatoare căreia dorim să îi estimăm media  $\mu = \mathbb{E}[X]$ .
- Generăm mai întâi un șir Monte Carlo de valori aleatoare care urmează distribuția lui  $X$  (acestea pot fi privite și ca variabile aleatoare independente și identic distribuite cu  $X$ :  $X_1, X_2, \dots, X_N$ ). Un estimator nedeplasat<sup>1</sup> pentru  $\mu$  este

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N},$$

deoarece  $\mathbb{E}[\bar{X}] = \mu$ . Dacă  $\text{Var}[X] = \sigma^2$ , atunci

$$\text{Var}[\bar{X}] = \frac{\sum_{i=1}^N \text{Var}[X_i]}{N^2} = \frac{\sigma^2}{N}.$$

<sup>1</sup>Un estimator a cărei medie este chiar parametrul estimat.

## Estimarea mediei cu metoda Monte Carlo - Exemple

- Exemplu.** Un vânzător comercializează un produs perisabil și în fiecare zi face o comanda de 100 de unități din acest produs. Fiecare unitate vândută aduce un profit de 55 cenți, iar o unitate nevândută dă, la sfârșitul zilei, o pierdere de 40 cenți. Cererea zilnică,  $X$ , urmează o distribuție uniformă  $U[80, 140]$ . Estimați profitul mediu.

- Soluție.** Dacă  $P$  este profitul, atunci

$$P = \begin{cases} 55, & \text{if } X \geq 100 \\ 0.55X - 0.4(100 - X), & \text{if } X < 100 \end{cases}$$

- Generăm  $N$  valori pentru  $X$  și calculăm  $P_1, P_2, \dots, P_N$ , apoi determinăm media de selecție.
- Pentru cinci eșantioane independente (cu  $N = 10000$ ) obținem

51.7796 51.82632 51.87036 51.84095 51.88509

## Estimarea mediei cu metoda Monte Carlo - Exemple

- Valoarea exactă a profitului mediu este

$$\int_{80}^{100} \frac{0.95x - 40}{60} dx + \int_{100}^{140} \frac{55}{60} dx = 51.83333$$

- **Exemplu.** Un server foarte performant este folosit de 250 utilizatori independenți. În fiecare zi, fiecare utilizator folosește serverul, independent, cu probabilitate 0.3. Numărul de job-uri lansate de fiecare utilizator pe server urmează o distribuție Geometrică cu parametrul 0.15 și fiecare job are nevoie de  $\Gamma(10, 3)$  timp (în minute) pentru a fi executat. Job-urile sunt executate consecutiv. Estimați media timpului total de utilizare a serverului.
- **Soluție.** Timpul total necesar  $T = T_1 + \dots + T_X$  constă din suma timpilor  $T_i$  ceruți de cei  $X$  utilizatori activi. Numărul de utilizatori activi  $X$  este  $Binomial(250, 0.3)$ .

## Estimarea mediei cu metoda Monte Carlo - Exemple

- Fiecare utilizator activ lansează  $Y_i$  job-uri, unde  $Y_i$  este  $G(0.15)$ . Astfel  $T_i = T_{i,1} + \dots + T_{i,Y_i}$ , unde  $T_{i,j} : \Gamma(10, 3)$ .
- Trei estimări independente oferă următoarele perioade de timp (în minute)  
1494.901 1492.228 1489.696
- Aceste valori sunt puțin peste 24 de ore (1440 minute).
- **Exemplu.** Două servere web oferă (servesc) aceleași pagini posibilelor clienți (web). Timpul necesar procesării unei cereri HTTP urmează o distribuție exponențială cu  $\lambda_1 = 0.03\text{ms}^{-1}$  pentru primul server și  $\lambda_2 = 0.04\text{ms}^{-1}$  pentru cel de-al doilea. Latența totală, care mai conține timpul necesar cererii și răspunsului de a parcurge distanța între client și server și înapoi, are o distribuție exponențială cu  $\lambda = 1\text{ms}^{-1}$ .

## Estimarea mediei cu metoda Monte Carlo - Exemple

- **Exemplu - continuare.** Un client oarecare este îndreptat către primul server cu probabilitate 0.4 și către al doilea cu probabilitate 0.6. Estimati timpul mediu de așteptare pe care un client îl petrece până la sosirea răspunsului la cererea sa.
- **Soluție.** O simulare (sau "run") pentru această problemă constă în generarea unei valori uniforme standard  $U$ , apoi în funcție de această valoare a unei valori care urmează o distribuție exponențială cu  $\lambda = 0.03$  sau  $0.04$ ; rezultatul este adăugat unei valori distribuite exponențial cu  $\lambda = 1$ :

$$T = X + \begin{cases} Y, & \text{if } U < 0.4 \\ Z, & \text{if } U \geq 0.4 \end{cases},$$

where  $U : U(0, 1)$ ,  $X : Exp(1)$ ,  $Y : Exp(0.03)$ ,  $Z : Exp(0.04)$ .

Pentru  $N = 10000$  obținem o estimare a mediei de 29.48822 ms.

## Estimarea lungimilor

- Fie  $U$  o variabilă uniformă standard;  $U$  aparține mulțimii  $A \subseteq [0, 1]$  cu probabilitatea

$$P(U \in A) = \int_A 1 \, du = \text{lungimea lui } A.$$

- Fie  $X = \chi_A(U)$  funcția indicator (caracteristică) a mulțimii  $A$  și  $X_1, X_2, \dots, X_N$  un șir de variabile aleatoare identic distribuite cu  $X$ .

$$X(u) = \chi_A(u) = \begin{cases} 1, & u \in A \\ 0, & \text{altfel} \end{cases}$$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N},$$

## Estimarea lungimilor

- Șirul  $(X_i)$  poate fi obținut prin generarea a  $n$  valori independente uniforme  $U_1, U_2, \dots, U_N$ , luând apoi  $X_i = \chi_A(U_i)$ .
- Lungimea lui  $A$  este aproximativ  $\bar{X}$  care este proporția valorilor  $U_i$  care se găsesc în  $A$ .
- Fie  $A \subseteq [a, b]$ ; dacă  $U$  este o variabilă uniformă definită pe  $[a, b]$ , atunci

$$P(U \in A) = \int_A \frac{du}{b-a} = \frac{1}{b-a} \int_A 1 \, du = \frac{\text{lungimea lui } A}{b-a}.$$

- Generăm un șir de valori uniforme și independente pe  $[a, b]$ :  $U_1, U_2, \dots$  ( $X_i = \chi_A(U_i)$ ). Lungimea lui  $A$  va fi cu aproximație  $(b-a) \cdot \bar{X}$ , adică proporția valorilor  $U_i$  care se află în  $A$  înmulțită cu  $(b-a)$ , deoarece  $P(U \in A) \approx \bar{X}$ .
- De obicei calculul unei lungimi nu pune probleme majore; metoda aceasta poate fi însă utilizată pentru estimarea ariilor și a volumelor.

## Estimarea ariilor

- Fie  $B$  o mulțime 2-dimensională care este inclusă în  $[0, 1] \times [0, 1]$ ; Două variabile uniforme standard independente au densitatea comună

$$f_{U,V}(u, v) = \begin{cases} 1, & (u, v) \in [0, 1] \times [0, 1] \\ 0, & \text{altfel} \end{cases}$$

- Aria lui  $B$  se determină astfel

$$P((U, V) \in B) = \iint_B 1 \, dudv.$$

- Un algoritm pentru estimarea ariei unei mulțimi  $B \subseteq [0, 1]^2$ :
  - Generăm un număr par de valori uniforme standard independente:  $U_1, \dots, U_N, V_1, \dots, V_N$ ;
  - Fie  $N_B$  numărul de perechi  $(U_i, V_i)$  care aparțin lui  $B$ .
  - Un estimator pentru aria lui  $B$  este  $N_B/N$ .

## Estimarea ariilor

- Fie  $B$  o mulțime 2-dimensională care este inclusă în  $[a_1, b_1] \times [a_2, b_2]$ ; două variabile uniforme una pe  $[a_1, b_1]$  și una pe  $[a_2, b_2]$ , independente, au densitatea comună

$$f_{U,V}(u, v) = \begin{cases} \frac{1}{(b_1 - a_1) \cdot (b_2 - a_2)}, & (u, v) \in [a_1, b_1] \times [a_2, b_2] \\ 0, & \text{altfel} \end{cases}$$

- Aria lui  $B$  se determină astfel

$$P((U, V) \in B) = \iint_B \frac{dudv}{(b_1 - a_1) \cdot (b_2 - a_2)} = \frac{1}{(b_1 - a_1) \cdot (b_2 - a_2)} \iint_B 1 \, dudv = \frac{\text{aria lui } B}{(b_1 - a_1) \cdot (b_2 - a_2)}$$

- Algoritm pt. estimarea ariei mulțimii  $B \subseteq [a_1, b_1] \times [a_2, b_2]$ :
  - Generăm  $N$  valori uniforme pe  $[a_1, b_1]$ ,  $U_1, \dots, U_N$  și  $N$  valori uniforme pe  $[a_2, b_2]$ ,  $V_1, \dots, V_N$ , toate independente;
  - Fie  $N_B$  numărul de perechi  $(U_i, V_i)$  care aparțin lui  $B$ .
  - Un estimator al ariei lui  $B$  este  $(b_1 - a_1) \cdot (b_2 - a_2) \cdot N_B / N$ .

## Estimarea ariilor - Exemple

- **Exemplul 1.** Fie  $B$  discul unitate din plan:

$$B = \{(u, v) : u^2 + v^2 \leq 1\} \subseteq [-1, 1]^2.$$

- Generăm  $N = 10000$  valori uniforme pe  $[-1, 1]$  independente (în R folosim  $\text{runif}(1, -1, 1)$  de 10000 de ori sau  $\text{runif}(10000, -1, 1)$ ).
- Obținem o estimare de 3.1368 pentru aria acestui disc care în realitate este  $\pi = 3.14159$ .

- **Exemplul 2.** Fie  $B$  o elipsă ( $a = 4, b = 3$ ):

$$B = \{(u, v) : u^2/a^2 + v^2/b^2 \leq 1\} \subseteq [-4, 4] \times [-3, 3] \subseteq [-4, 4]^2.$$

- Generăm  $N = 10000$  perechi de valori uniforme din  $[-4, 4]$  independente.
- Obținem o estimare de 37.4528 pentru aria acestei elipse care este  $\pi ab = 12\pi = 37.69911$ .

## Estimarea volumelor

- Fie  $C$  o mulțime 3-dimensională inclusă în  $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ ; trei variabile uniforme una pe  $[a_1, b_1]$ , una pe  $[a_2, b_2]$  și una pe  $[a_3, b_3]$ , independente, au densitatea comună

$$f_{U,V,W}(u, v, w) = \begin{cases} \frac{1}{(b_1 - a_1) \cdot (b_2 - a_2) \cdot (b_3 - a_3)}, & (u, v, w) \in [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] \\ 0, & \text{otherwise} \end{cases}$$

- Volumul lui  $C$  se determină astfel

$$P((U, V, W) \in C) = \frac{\iiint_C dudvdw}{(b_1 - a_1) \cdot (b_2 - a_2) \cdot (b_3 - a_3)} = \frac{\text{volumul lui } C}{(b_1 - a_1) \cdot (b_2 - a_2) \cdot (b_3 - a_3)}$$

- Algoritm pentru estimarea volumului,  $C \subseteq [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ :
  - Generăm  $N$  valori uniforme pe  $[a_1, b_1]$ ,  $(U_i)_{i=1, N}$ ,  $N$  valori uniforme pe  $[a_2, b_2]$ ,  $(V_i)_{i=1, N}$ , și  $n$  valori uniforme pe  $[a_3, b_3]$ ,  $(W_i)_{i=1, N}$ , toate independente;
  - Fie  $N_C$  numărul de triplete  $(U_i, V_i, W_i)$  care aparțin lui  $C$ .

## Estimarea volumelor - Exemplu

- 3. Estimăm volumul lui  $C$  prin  $(b_1 - a_1) \cdot (b_2 - a_2) \cdot (b_3 - a_3) \cdot N_C / N$ .
- Să estimăm volumul sferei (bilei) unitate<sup>2</sup>:  

$$C = \{(u, v, w) : u^2 + v^2 + w^2 \leq 1\} \subseteq [-1, 1]^3.$$
- Mai întâi generăm  $N = 10000$  triplete uniforme din  $[-1, 1]$ , independente și apoi obținem o estimare de 4.184 pentru volumul acestei bile care este  $4\pi/3 = 4.18879$ .
- Dacă generăm  $N = 50000$  triplete uniforme din  $[-1, 1]$ , independente, obținem o estimare de 4.18816 pentru volumul bilei unitate
- Pe măsură ce numărul de dimensiuni ale spațiului în care lucrăm crește, avem nevoie de tot mai multe valori aleatoare pentru a aproxima bine parametrul dorit.
- Acesta este *the curse of dimensionality* vizibil în spații cu multe dimensiuni.

<sup>2</sup>De obicei, prin sferă se înțelege doar frontiera mulțimii care urmează.

## Estimarea volumelor - Exemplu

- Să estimăm volumul bilei unitate 8-dimensionale (care are volumul egal cu  $\pi^4/24 = 4.058712$ ):

$$C = \left\{ (u_1, \dots, u_8) : \sum_{i=1}^8 u_i^2 \leq 1 \right\} \subseteq [-1, 1]^8.$$

- Următorul tabel conține patru estimatori diferiți pentru un număr diferit de simulări MC:

run	$N = 1000$	$N = 20000$	$N = 50000$	$N = 100000$
1.	2.816	3.3920	4.11136	3.99872
2.	4.096	4.1600	4.01408	3.98592
3.	3.584	4.3776	4.06528	4.04992
4.	3.328	4.0704	4.2496	4.13440
5.	4.864	3.6480	4.22912	4.00896
average	3.7376	3.9296	4.133888	4.035584

## Estimarea ariilor regiunilor cu frontiere necunoscute

- Pentru a aproxima arii sau volume cu metoda Monte Carlo nu este necesar să cunoaștem frontierele mulțimii în cauză.
- Pentru a aplica unul dintre algoritmi anteriori este suficient să știm cum putem afla dacă un punct dat aparține mulțimii (pentru care măsurăm aria, volumul etc).
- Astfel, nu este necesar ca mulțimea din care ne extragem punctele să aibă o formă rectangulară; cu scalări diferite ale axelor, putem genera puncte aleatoare dintr-o formă rectangulară sau dintr-o formă mai complexă.
- O metodă de a genera puncte aleatoare dintr-o regiune cu o formă arbitrară este de a genera puncte (de coordonate uniforme) într-o formă rectangulară care conține acea regiune numărând punctele din regiunea vizată.

## Estimarea ariilor regiunilor cu frontiere necunoscute - Exemplu

- **Exemplu.** O alertă este lansată la o centrală nucleară. Este necesar să se estimeze aria regiunii expuse la scurgeri radioactive. Frontierele acestei regiuni nu pot fi determinate, însă se poate măsura nivelul de radioactivitate în orice locație dată.
- **Soluție.** Un dreptunghi de  $10 \times 8$  km este ales în jurul ariei expuse. Se generează perechi de valori uniforme independente  $(U_i, V_i)$  în acest dreptunghi.
- Se măsoară radioactivitatea în câteva locații alese aleator dintre cele accesibile. Aria este estimată ca proporția măsurărilor peste nivelul admis înmulțită cu aria dreptunghiului.
- Să presupunem că radioactivitatea este măsurată în 50 de locații aleatoare și că se găsește un nivel peste cel normal în 18 locații. Aria expusă este estimată prin  $\frac{18}{50} \cdot 80 \text{ km}^2 = 28.8 \text{ km}^2$ .

## Integrarea Monte Carlo

- O lungime, o arie sau un volum pot fi văzute drept integrale definite ale unor anumite funcții.
- Metoda Monte Carlo se poate de altfel extinde la calculul integralelor definite. Să presupunem că avem de integrat o anumită funcție  $h$  între  $a$  și  $b$ :

$$H = \int_a^b h(u) du.$$

- Putem aproxima această integrală considerând media unor valori ale lui  $h$  în puncte aleatoare repartizate uniform pe  $[a, b]$ .
- Dacă  $U_1, U_2, \dots, U_N$  sunt valori uniforme pe  $[a, b]$  independente (pentru care densitatea este  $1/(b - a)$  pe acest interval și 0 altfel), estimatorul Monte Carlo pentru  $H$  este

$$F_N = \frac{b - a}{N} \sum_{i=1}^N h(U_i).$$

## Integrarea Monte Carlo

- Această aproximare are loc deoarece, pentru o variabilă uniformă,  $U$ , pe  $[a, b]$ , media lui  $h(U)$  este

$$\mathbb{E}[h(U)] = \int_a^b h(u)f(u) du,$$

unde  $f$  este densitatea distribuției uniforme pe  $[a, b]$ .

- Astfel

$$\mathbb{E}[h(U)] = \int_a^b h(u) \frac{1}{b-a} du,$$

și

$$H = \int_a^b h(u) du = (b-a)\mathbb{E}[h(U)].$$

- Folosind estimarea Monte Carlo pentru media de mai sus obținem

$$H \approx \frac{b-a}{N} \sum_{i=1}^N h(U_i) = F_N,$$

pentru variabilele uniforme pe  $[a, b]$  și independente  $(U_i)_{1 \leq i \leq N}$ .

## Integrarea Monte Carlo

- Din Legea (tare a) Numerelor Mari  $P\left(\lim_{N \rightarrow \infty} F_N = H\right) = 1$ ; dispersia acestui estimator este

$$\text{Var}[F_N] = \frac{(b-a)^2}{12N} = \mathcal{O}(1/N),$$

deoarece dispersia distribuției uniforme pe  $[a, b]$  este  $(b-a)^2/12$ .

- Cum deviația standard este o măsură a împrăstierii, ultima relație poate fi citită astfel: trebuie să mărim de patru ori dimensiunea eșantionului pentru a reduce la jumătate eroarea (deviația standard).

## Integrarea Monte Carlo - Exemplu

- Să estimăm următoarea integrală (improprie):

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

(se știe că  $\int_0^{\infty} e^{-u^2/2} du = \sqrt{\pi/2} = 1.253314$ ).

- Observăm mai întâi că  $\lim_{a \rightarrow \infty} \int_0^a e^{-u^2/2} du = \int_0^{\infty} e^{-u^2/2} du$ , deci, pentru valori mari ale lui  $a$  avem  $\int_0^{\infty} e^{-u^2/2} du \approx \int_0^a e^{-u^2/2} du$ ; alegem  $a = 10$ .
- Pentru diferite valori ale dimensiunii  $N$  am obținut următoarele medii pentru 30 de aproximări independente.

	$N = 1000$	$N = 10000$	$N = 20000$	$N = 50000$
media	1.247216	1.259898	1.250592	1.251562
dev. st.	0.08749	0.02256	0.01898	0.01045

## Integrarea Monte Carlo îmbunătățită

- Integrala definită de mai sus poate fi scrisă astfel:

$$H = \frac{1}{b-a} \int_a^b (b-a)h(u) du = \mathbb{E}[(b-a)h(U)],$$

unde  $U$  are o distribuție uniformă continuă pe  $[a, b]$ .

- Urmând următoarea procedură putem utiliza orice *distribuție continuă* în locul celei uniforme.
- Fie  $X$  o distribuție aleatoare continuă cu densitatea  $f$  astfel ca  $f(u) > 0$ , pentru orice  $u \in [a, b]$  și  $f(u) = 0$  pentru orice  $u \notin [a, b]$ .
- Putem scrie

$$H = \int_a^b h(x) dx = \int_a^b \frac{h(x)}{f(x)} f(x) dx = \mathbb{E} \left[ \frac{h(X)}{f(X)} \right].$$

## Integrarea Monte Carlo îmbunătățită

- Vom estima  $H$  alegând  $N$  valori aleatoare ale lui  $X$  ( $X_1, \dots, X_N$ ) și calculând următoarea medie:

$$H \approx \frac{1}{N} \sum_{i=1}^N \frac{h(X_i)}{f(X_i)}.$$

- Metoda de mai sus nu se limitează la intervale finite  $[a, b]$ . Putem aproxima în acest fel și integrale improprii (convergente).
- Putem aproxima pe orice interval  $(a, b) \subseteq \overline{\mathbb{R}}$  trebuie doar ca suportul lui  $f$ , i. e.,  $\text{supp}(f) = \{x \in \mathbb{R} : f(x) \neq 0\}$  să conțină  $(a, b)$ .

## Integrarea Monte Carlo îmbunătățită - Exemplu

- De exemplu, alegând  $f$  să fie densitatea normală standard putem aplica integrarea Monte Carlo de la  $-\infty$  până la  $\infty$  sau, dacă alegem  $f$  să fie densitatea exponențială putem aplica integrarea Monte Carlo de la 0 până la  $\infty$ .
- Să estimăm din nou

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

folosind de data aceasta densitatea exponențială  $\lambda = 1$  (și nu o aproximare a limitei de integrare).

	$N = 1000$	$N = 10000$	$N = 20000$	$N = 50000$
average	1.254416	1.254476	1.253978	1.253035
st. dev.	0.01454	0.00349	0.00313	0.00176

## Estimarea probabilităților folosind metoda Monte Carlo

- Estimarea unei probabilități este una dintre aplicațiile tipice ale metodei Monte Carlo.
- Fie  $X$  o variabilă aleatoare reală și  $A \subseteq \mathbb{R}$ ; probabilitatea  $p = P(X \in A)$  se estimează astfel

$$\hat{p}_N = \frac{\#\{X_i \in A\}}{N}.$$

- Evident că numărul variabilelor  $X_1, X_2, \dots, X_N$  care aparțin lui  $A$  este o variabilă aleatoare discretă cu o distribuție binomială ( $B(N, p)$ ).
- Media și dispersia lui  $\hat{p}_N$  sunt

$$\mathbb{E}[\hat{p}_N] = \frac{Np}{N} = p, \text{ respectiv}$$

$$\text{Var}[\hat{p}_N] = \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N}.$$

## Acuratețea estimării probabilităților cu metoda MC

- Cât de bună este această metodă de aproximare a lui  $p$  prin  $\hat{p}_N$  (care este un estimator nedeplasat)?
- Folosind aproximarea normală a distribuției binomiale,

$$\frac{N\hat{p} - Np}{\sqrt{Np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/N}} : N(0, 1).$$

- De unde

$$P(|\hat{p} - p| > \epsilon) = P\left(\frac{|\hat{p} - p|}{\sqrt{p(1-p)/N}} > \frac{\epsilon}{\sqrt{p(1-p)/N}}\right) \approx \\ \approx 2\Phi\left(-\frac{\epsilon}{\sqrt{p(1-p)/N}}\right) = 2 \cdot \text{pnorm}\left(-\frac{\epsilon}{\sqrt{p(1-p)/N}}\right),$$

unde  $\Phi(\cdot)$  este funcția de repartiție a unei variabile normale standard (în  $\mathbb{R}$ ,  $\Phi(z) = \text{pnorm}(z)$ ).

## Acuratețea estimării probabilităților cu metoda MC

- Cum se proiectează un studiu Monte Carlo care să aibă o acuratețe anterior prescrisă?
- Adică, pentru un  $\epsilon$  și  $0 < \alpha < 1$ , cât de mare trebuie să fie  $N$  astfel ca

$$P(|\hat{p} - p| > \epsilon) \leq \alpha ?$$

- Principalul obstacol este acela că în relația de mai sus valoarea lui  $p$  este necunoscută (altfel estimarea nu ar mai avea sens).
- Avem două posibilități pentru a estima cantitatea  $p(1 - p)$ :

- 1 Mai întâi, am putem utiliza o "aproximare" (o estimare preliminară) a lui  $p$ , dacă există.
- 2 În al doilea rând, putem utiliza un majorant din inegalitatea mediilor

$$p(1 - p) \leq 1/4, \forall p \in [0, 1].$$

## Acuratețea estimării probabilităților cu metoda MC

- În primul caz, dacă  $p^*$  este o "aproximare", trebuie să rezolvăm inegalitatea

$$2\Phi\left(\frac{\epsilon}{\sqrt{p^*(1-p^*)/N}}\right) \leq \alpha.$$

- Fie  $z_a = \Phi^{-1}(a) = \text{qnorm}(a)$ , unde  $a \in (0, 1)$ . Inegalitatea devine

$$-\frac{\epsilon}{\sqrt{p^*(1-p^*)/N}} \leq z_{\frac{\alpha}{2}} \text{ or } \sqrt{p^*(1-p^*)/N} \leq -\frac{\epsilon}{z_{\frac{\alpha}{2}}}.$$

(Să notăm că, pentru  $a < 1/2$ , avem  $z_a < 0$ .)

- Obținem un minorant pentru  $N$ :

$$N \geq p^*(1-p^*) \left(\frac{z_{\frac{\alpha}{2}}}{\epsilon}\right)^2.$$

## Acuratețea estimării probabilităților cu metoda MC

- În cel de-al doilea caz, dacă nu avem o "aproximare", atunci

$$N \geq \frac{1}{4} \left( \frac{z_{\frac{\alpha}{2}}}{\epsilon} \right)^2 = \left( \frac{z_{\frac{\alpha}{2}}}{2\epsilon} \right)^2$$

## Estimarea probabilităților folosind metoda MC - Exemplu

- **Exemplu.** Un server este utilizat de 250 clienți independenți. În fiecare zi, un client, în mod independent, folosește serverul cu probabilitate 0.3. Numărul de procese lansate în execuție pe server de fiecare client activ urmează o distribuție Geometrică cu parametrul 0.15, iar fiecare proces are nevoie pentru a fi executat de un timp care urmează o distribuție  $\Gamma(10, 3)$ . Job-urile sunt procesate consecutiv. Care este probabilitatea ca timpul total necesar să fie mai puțin de 24 de ore? Estimați probabilitatea cu o eroare de cel mult  $\pm 0.01$  cu probabilitate 0.99.
- **Soluție.** Timpul total  $T = T_1 + \dots + T_X$  este format din suma timpilor necesari fiecărui clienților activi,  $T_i$ , care sunt în număr de  $X$ , variabilă distribuită *Binomial*(250, 0.3).

## Estimarea probabilităților folosind metoda MC - Exemplu

- Fiecare client activ lansează în execuție  $Y_i$  procese,  $Y_i : Geometric(0.1)$ . Astfel  $T_i = T_{i,1} + \dots + T_{i,Y_i}$ , unde  $T_{i,j} : \Gamma(10, 3)$ .
- Nu avem o "aproximare" a probabilității în cauză,  $P(T < 24)$ . Pentru a obține acuratețea cerută ( $\alpha = 0.01$ ,  $\epsilon = 0.01$ ) avem nevoie de

$$N \geq \frac{1}{4} \left( \frac{z_{\frac{\alpha}{2}}}{\epsilon} \right)^2 = \frac{1}{4} \left( \frac{z_{0.005}}{0.01} \right)^2 = \frac{1}{4} \left( \frac{-2.57529}{0.01} \right)^2 = 16587.24,$$

as  $z_{0.005} = qnorm(0.005) = -2.57529$ .







- Astfel, vom avea nevoie de  $N = 16588$  simulări (valoare suficient de mare pentru a utiliza aproximarea normală a distribuției binomiale).

## Estimarea probabilităților folosind metoda MC - Exemplu

- Trei estimări independente dau următoarele probabilități  
0.4262117 0.4202435 0.4259103
- Probabilitatea nu este chiar atât de mică; este posibil ca toate joburile să fie terminate într-o singură zi.



## Bibliografie

-  Baron, M., *Probability and Statistics for Computer Scientist*, Chapman & Hall/CRC Press, 2013 or the electronic edition <https://ww2.ii.uj.edu.pl/~z1099839/naukowe/RP/rps-michael-byron.pdf>
-  Johnson, J. L., *Probability and Statistics for Computer Science*, Wiley Interscience, 2008.
-  Lipschutz, S., *Theory and Problems of Probability*, Schaum's Outline Series, McGraw-Hill, 1965.
-  Ross, S. M., *A First Course in Probability*, Prentice Hall, 5th edition, 1998.
-  Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
-  Stone, C. J., *A Course in Probability and Statistics*, Duxbury Press, 1996.