Probabilităţi şi Statistică - Curs 10

mai, 2018

- Procesul de generare a valorilor aleatoare care au o anumită densitate se numeşte *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 şi observând fracţiunea acestor valori care au o anumită proprietate. Această metodă este utilă pentru a obţine soluţii numerice (aproximative) pentru probleme care sunt prea complicate pentru a fi rezolvate analitic. *mathworld.wolfram.com*
- O valoare a unei variable aleatoare (sau o valoare care urmează o densitate) este numită *quantilă* sau *valoare aleatoare*, *număr aleator* (în engleză *variate*).

- O *metodă Monte Carlo* poate genera foarte multe astfel de valori aleatoare (câteodată milioane) asociate unei distribuţii de probabilitate, iar acest proces se numeşte simulare a respectivei distribuţii.
- Simularea este utilizată pentru a determina media sau dispersia unei distribuţii sau un alt parametru asociat.
- Simularea depinde de "calitatea" valorilor aleatoare. Cele mai utilizate numere aleatoare sunt cele provenite din distribuţia uniformă continuă standard, $U(0,1)$, sau din distribuţia 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 utilizate 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 şir Monte Carlo de valori aleatoare care urmează distribuţia lui X (acestea pot fi privite şi ca variabile aleatoare independente şi identic distribuite cu X: $X_1, X_2, \ldots X_N$. Un estimator nedeplasat pentru μ este

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N},$$

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

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

- Exemplu. Un vânzător comercializează un produs perisabil şi în fiecare zi face o comanda de 100 de unităţi din acest produs. Fiecare unitate vândută aduce un profit de 55 cenţi, iar o unitate nevândută dă, la sfârşitul zilei, o pierdere de 40 cenţi. Cererea zilnică, X, urmează o distribuţie uniformă $U[80, 140]$. Estimaţi profitul mediu.
- Soluţie. Dacă P este profitul, atunci

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

- Generăm N valori pentru X şi calculăm $P_1, P_2, \ldots, P_N$, apoi determinăm media de selecţie.
- Pentru cinci eşantioane independente (cu $N = 10000$) obţinem

$$51.7796 \;\; 51.82632 \;\; 51.87036 \;\; 51.84095 \;\; 51.88509$$

- 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 independenţi. În fiecare zi, fiecare utilizator foloseşte serverul, independent, cu probabilitate 0.3. Numărul de job-uri lansate de fiecare utilizator pe server urmează o distribuţie Geometrică cu parametrul 0.15 şi fiecare job are nevoie de $\Gamma(10, 3)$ timp (în minute) pentru a fi executat. Job-urile sunt executate consecutiv. Estimaţi media timpului total de utilizare a serverului.
- Soluţie. Timpul total necesar $T = T_1 + \ldots + T_X$ constă din suma timpilor T_i ceruţi de cei X utilizatori activi. Numărul de utilizatori activi X este $Binomial(250, 0.3)$.

- Fiecare utilizator activ lansează Y_i job-uri, unde Y_i este $Geometric(0.15)$. Astfel $T_i = T_{i,1} + \ldots + T_{i,Y_i}$, unde $T_{i,j} : \Gamma(10, 3)$.
- Trei estimări independente oferă următoarele perioade de timp (în minute)

$$1494.901 \quad 1492.228 \quad 1489.696$$

- Aceste valori sunt puţin peste 24 de ore (1440 minute).
- Exemplu. Două servere web oferă (servesc) aceleaşi pagini posibililor clienţi (web). Timpul necesar procesării unei cereri HTTP urmează o distribuţie exponenţială cu $\lambda_1 = 0.03\text{ms}^{-1}$ pentru primul server şi $\lambda_2 = 0.04\text{ms}^{-1}$ pentru cel de-al doilea. Latenţa totală, care mai conţine timpul necesar cererii şi răspunsului de a parcurge distanţa între client şi server şi înapoi, are o distribuţie exponenţială cu $\lambda = 1\text{ms}^{-1}$.

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

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

where $U : U(0,1)$, $X : Exp(1)$, $Y : Exp(0.03)$, $Z : Exp(0.04)$. From $N = 10000$ obţinem o estimare a mediei de 29.48822 ms.

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

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

- Fie $X = \chi_A$ funcţia indicator (caracteristică) a mulţimii A şi $X_1, X_2, \ldots X_N$ un şir de variabile aleatoare identic distribuite cu X.

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

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N},$$

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

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

- Generăm un şir de valori uniforme şi independente pe $[a, b]$: $U_1, U_2, \ldots, U_N$ $(X_i = \chi_A(U_i))$. Lungimea lui A va fi cu aproximaţie $(b-a) \cdot \overline{X}$, adică proporţia valorilor U_i care se află în A înmulţită cu $(b-a)$.
- De obicei calculul unei lungimi nu pune probleme majore; metoda aceasta poate fi însă utilizată pentru estimarea ariilor şi a volumelor.

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

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

- Aria lui B este

$$P\left((U,V)\in B\right)=\iint_B 1\; du dv.$$

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

- Fie B o mulţime 2-dimensională care este inclusă în $[a, b] \times [a, b]$; două variabile uniforme pe $[a, b]$ independente au densitatea comună

$$f_{U,V}(u, v) = \begin{cases} 1/(b-a)^2, & (u, v) \in B \\ 0, & \text{altfel} \end{cases}.$$

- Aria lui B este

$$P((U, V) \in B) = \iint_B 1 \, du dv = (b-a)^2 \iint_B \frac{1}{(b-a)^2} \, du dv.$$

- Un algoritm pentru estimarea ariei unei mulţimi $B \subseteq [a, b]^2$:
  1. Generăm un număr par de valori uniforme pe $[a, b]$ independente: $U_1, \ldots, U_N, V_1, \ldots, V_N$;
  2. Fie N_B numărul de perechi (U_i, V_i) care aparţin lui B.
  3. Un estimator al ariei lui B este $(b-a)^2 \cdot N_B / N$.

- Exemplul 1. Fie B discul unitate din plan:

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

- Generăm $N = 10000$ valori uniforme pe $[-1,1]$ independente (în R folosim *runif(1, -1, 1)* de 10000 de ori sau runif(10000, -1, 1)).
- Obţinem o estimare de 3.1368 pentru aria acestui disc care în realitate este $\pi = 3.14159$.
- Exemplul 2. Fie B o elipsă ($a = 4$, $b = 3$):

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

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

- Fie C o mulţime 3-dimensională inclusă în $[a, b] \times [a, b] \times [a, b]$; trei variabile uniforme pe $[a, b]$ independente au densitatea comună

$$f_{U,V,W}(u, v, w) = \begin{cases} 1/(b-a)^3, & (u, v, w) \in C \\ 0, & \text{altfel} \end{cases}.$$

- Volumul lui C este

$$P((U, V, W) \in C) = \iiint_C 1 \, dudvdw = (b-a)^3 \iiint_C \frac{dudvdw}{(b-a)^3}$$

- Un algoritm pentru estimarea volumului lui $C \subseteq [a, b]^3$:
  1. Generăm un număr multiplu de 3 de valori uniforme pe $[a, b]$ independente : $U_1, \ldots, U_N, V_1, \ldots, V_N, W_1, \ldots, W_N$.
  2. Fie N_C numărul de triplete (U_i, V_i, W_i) care aparţin lui C.
  3. Estimăm volumul lui C prin $(b-a)^3 \cdot N_C / N$.

- Să estimăm volumul sferei (bilei) unitate[1]:

$$C = \left\{(u, v, w) \,:\, u^2 + v^2 + w^2 \leqslant 1\right\} \subseteq [-1, 1]^3.$$

- Mai întâi generăm $N = 10000$ triplete uniforme din $[-1, 1]$, independente şi apoi obţinem 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, obţinem o estimare de 4.18816 pentru volumul bilei unitate
- Pe măsură ce numărul de dimensiuni ale spaţiului în care lucrăm creşte, avem nevoie de tot mai multe valori aleatoare pentru a aproxima bine parametrul dorit.
- Acesta este *the curse of dimensionality* vizibil în spaţii cu multe dimensiuni.

[1] De obicei, prin sferă se înţelege doar frontiera mulţimii care urmează.

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

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

- Următorul tabel conţine conţine patru estimatori diferiţii 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
absolute error	0.321112	0.129112	0.075176	0.023128

- Pentru a aproxima arii sau volume cu metoda Monte Carlo nu este necesar să cunoaştem frontierele mulţimii în cauză.
- Pentru a aplica unul dintre algoritmii anteriori este suficient să ştim cum putem afla dacă un punct dat aparţine mulţimii (pentru care măsurăm aria, volumul etc).
- Astfel, nu este necesar ca mulţimea 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 conţine acea regiune până când obţinem un punct 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 locaţie dată.
- Soluţie. Un dreptunghi de 10×8 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 locaţii alese aleator dintre cele accesibile. Aria este estimată ca proporţia măsurătorilor peste nivelul admis înmulţită cu aria dreptunghiului.
- Să presupunem că radioactivitatea este măsurată în 50 de locaţii aleatoare şi că se găseşte un nivel peste cel normal în 18 locaţii. Aria expusă este estimată prin $\frac{18}{50} \cdot 80\ km^2 = 28.8\ km^2$.

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

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

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

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

- 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 distribuţiei uniforme pe $[a, b]$.

- Astfel

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

şi

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

- Folosind estimarea Monte Carlo pentru media de mai sus obţinem

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

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

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

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

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

- Cum deviaţia standard este o măsură a împrăştierii, ultima relaţie poate fi citită astfel: trebuie să mărim de patru ori dimensiunea eşantionului pentru a reduce la jumătate eroarea (deviaţia standard).

- Să estimăm următoarea integrală (improprie):

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

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

- Observăm mai întâi că $\lim\limits_{a\to\infty}\int_0^a e^{-u^2/2}\,du = \int_0^\infty e^{-u^2/2}\,du$, deci, pentru valori mari ale lui a avem $\int_0^\infty e^{-u^2/2}\,du \approx \int_0^a e^{-u^2/2}\,du$. Să alegem $a = 10$.

- Pentru diferite valori ale dimensiunii N am obţinut 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

- 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 distribuţie uniformă continuă pe $[a,b]$.

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

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

- Vom estima H alegând N valori aleatoare ale lui X ($X_1, \ldots, X_N$) şi 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 şi integrale improprii (convergente).
- Putem aproxima pe orice interval $(a, b) \subseteq \overline{\mathbb{R}}$ trebuie doar ca suportul lui f, i. e., $supp(f) = \{x \in R \, : \, f(x) \neq 0\}$ să fie inclus în (a, b).

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

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

folosind de data aceasta densitatea exponenţială $\lambda = 1$ (şi 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 unei probabilităţi este una dintre aplicaţiile tipice ale metodei Monte Carlo.
- Fie X o variabilă aleatoare reală şi $A \subseteq \mathbb{R}$; probabilitatea $p = P(X \in A)$ se estimează astfel

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

- Evident că numărul variabilelor $X_1, X_2, \ldots, X_N$ care aparţin lui A este o variabilă aleatoare discretă cu o distribuţie binomială ($B(N,p)$).
- Media şi dispersia lui $\widehat{p}_N$ sunt

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

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

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

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

- De unde

$$P(|\widehat{p} - p| > \epsilon) = P\left(\frac{|\widehat{p} - p|}{\sqrt{p(1-p)/N}} > \frac{\epsilon}{\sqrt{p(1-p)/N}}\right) \approx$$

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

unde $\Phi(\cdot)$ este funcţia de repartiţie a unei variabile normale standard (în R, $\Phi(z) = pnorm(z)$).

- Cum se proiectează un studiu Monte Carlo care să aibă o acurateţe anterior prescrisă?
- Adică, pentru un ϵ şi $0 < \alpha < 1$, cât de mare trebuie să fie N astfel ca
$$P(|\widehat{p} - p| > \epsilon) \leqslant \alpha\,?$$
- Principalul obstacol este acelă că în relaţia de mai sus valoarea lui p este necunoscută (altfel estimarea nu ar mai avea sens).
- Avem două posibilităţi 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
$$p(1-p) \leqslant 1/4, \forall p \in [0,1].$$

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

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

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

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

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

- Obţinem un minorant pentru N:

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

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

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

- Exemplu. Un server este utilizat de 250 clienţi independenţi. În fiecare zi, un client, în mod independen, foloseşte serverul cu probabilitate 0.3. Numărul de procese lansate în execuţie pe server de fiecare client activ urmează o distribuţie Geometrică cu parametrul 0.15, iar fiecare proces are nevoie pentru a fi executat de un timp care urmează o distribuţie $\Gamma(10, 3)$. Job-urile sunt procesate consecutiv. Care este probabilitatea ca timpul total necesar să fie mai puţin de 24 de ore? Estimaţi probabilitatea cu o eroare de cel mult ± 0.01 cu probabilite 0.99.
- Soluţie. Timpul total $T = T_1 + \ldots + T_X$ este format din suma timpilor necesari fiecărui clienţilor activi, T_i, care sunt în număr de X, variabilă distribuită $Binomial(250, 0.3)$.

- Fiecare client activ lansează în execuţie Y_i procese, Y_i : $Geometric(0.15)$. Astfel $T_i = T_{i,1} + \ldots + T_{i,Y_i}$, unde $T_{i,j} : \Gamma(10, 3)$.
- Nu avem o "aproximare" a probabilităţii în cauză, $P(T < 24)$. Pentru a obţine acurateţea cerută ($\alpha = 0.001$, $\epsilon = 0.001$) avem nevoie de

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

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 distribuţiei binomiale).

- Trei estimări independente dau următoarele probabilităţi

$$0.4262117 \ 0.4202435 \ 0.4259103$$

- Probabilitatea nu este chiar atât de mică; este posibil ca toate job-urile să fie terminate într-o singură zi.

Sfârşit

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.