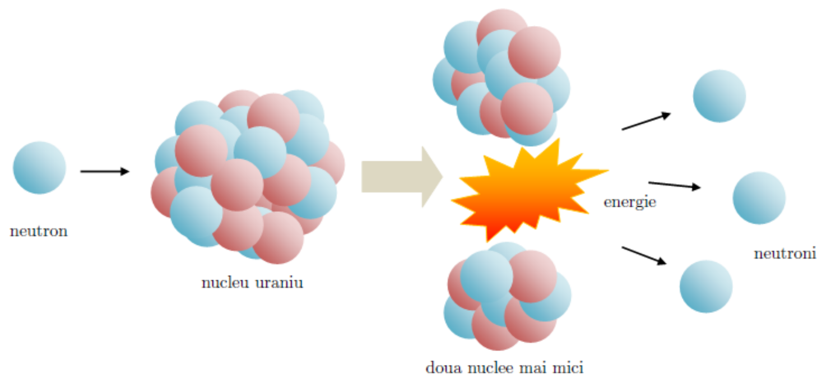
"A human is a random variable, with a hope to find a function to let his life have some value."

Jinesh Parakh

10
Simularea variabilelor aleatoare

Simulări Monte Carlo

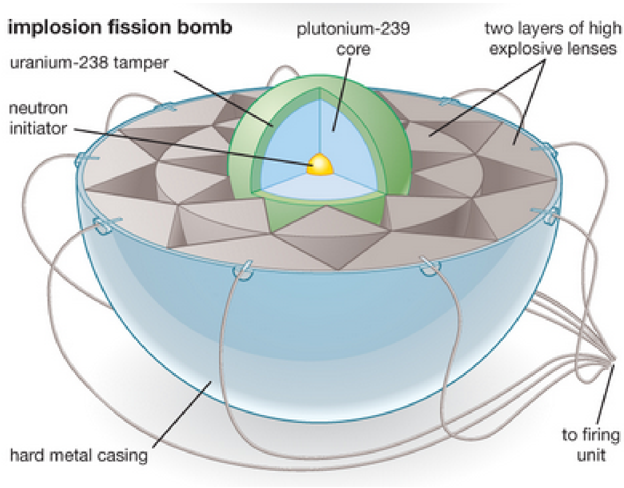


Anul 1945, doua evenimente socante au loc: explozia primei bombe nucleare si aparitia primului computer electronic. Primul eveniment a condus la doua regretabile erori, o cursa a inarmarilor, dar si la crearea unei discutabile surse de energie. Al doilea a trecut aproape neobservat, insa cateva minti luminate ca ale lui John von Neumann sau Stanislaw Ulam i-au inteles imediat impactul, cu toate ca nu aveau cum sa intuiasca pe deplin modul in care computerele ne vor schimba realitatea.

ENIAC (Electronic Numerical Integrator And Computer) a fost proiectat si construit pentru a calcula tabele balistice, pentru armata americana. John von Neumann, deja consultant in proiectul Manhattan, era preocupat de probleme termonucleare si de constructia primei bombe cu hidrogen. Prin urmare i-a contactat pe Nicholas Metropolis si Stan Frankel pentru a pregati un model computational al unei reactii termonucleare, pentru a testa ENIAC. Neumann considera ca acesta trebuia sa fie adevaratul test al calculatorului. La Los Alamos lucrau in acea vreme si matematicienii Stanislaw Ulam si Robert Richtmyer.

Impreuna au pornit de la cateva probleme cunoscute in studiul difuziei neutronilor in materialele fisionabile si au dezvoltat ceea ce astazi numim simulari sau metode Monte Carlo. Numele se pare ca a fost ales de N. Metropolis care si-a amintit ca un unchi al lui S. Ulam obisnuia sa imprumute bani de la rude pentru a pleca la Monte Carlo unde juca la cazinouri.

Una dintre probleme, discutata intr-o scrisoare adresata de Neumann lui Richtmyer, presupunea considerarea unui nucleu sferic al unui material fisionabil, invelit intr-un material numit tamper, care poate reflecta inapoi neutronii evadati provocand noi coliziuni intre acestia. Se presupunea o anumita distributie a neutronilor in spatiu si viteza dar nu se luau in considerare efectele radiative sau hidrodinamice. Ideea era sa poata fi urmarita dezvoltarea unui numar mare de lanturi neutronice, ca efect al imprastierii, absortiei, fisiunii sau evadarii.



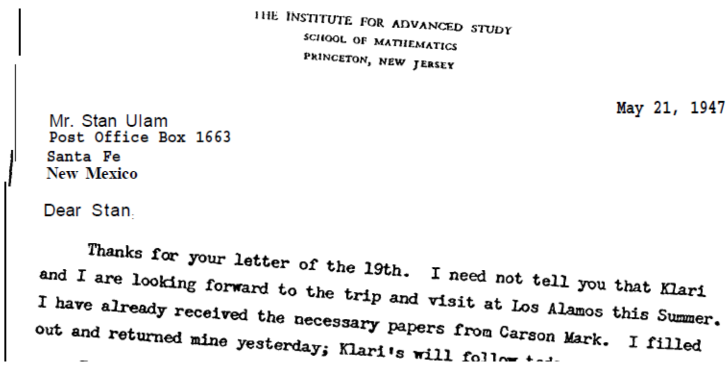
In fiecare stadiu trebuiau luate o serie de decizii bazate pe probabilitati statistice, adecvate factorilor fizici si geometrici. Printre acestea enumeram: pozitia in spatiu a neutronilor, viteza, pozitia primei coliziuni, natura acestei coliziuni. Daca se decidea ca o fisiune a avut loc trebuia decis numarul de neutroni rezultati. Daca coliziunea rezulta intr-o imprastiere trebuia decis noul impuls al neutronilor. Cand un neutron traversa un material se luau in considerare caracteristicile noului mediu, rezultand in final o istorie genealogica a fiecarui neutron. Procesul era repetat pentru alti neutroni, pana ce o imagine statistica valida era obtinuta.

Initial problema transportului neutronilor fusese abordata analitic, dar rezultasera ecuatii greu de manevrat. Stanislaw Ulam povestea ca in 1946, in timp ce juca Solitaire, ii venise in gand ideea sa estimeze sansele de a castiga la acest joc, cu un pachet de 52 de carti bine amestecate.

> *"Dupa ce am pierdut o multime de timp incercand sa le estimez, prin calcule combinatoriale, m-am intrebat daca nu cumva o metoda mai practica decat "gandirea abstracta" ar insemna sa asez cartile, sa zicem de 100 de ori, si sa calculez numarul de jocuri incununate de succes. Asta era deja posibil de imaginat, odata cu inceputul noii ere a computerelor rapide, si m-am gandit imediat la problemele difuziei*

neutronilor si mai general la cum ar putea fi schimbate procese descrise de ecuatii diferentiale intr-o forma echivalenta ca sa poata fi interpretata ca fiind o succesiune de operatii aleatoare."

In cele din urma S. Ulam i-a prezentat ideea lui von Neumann, care a fost fascinat de posibilitatea de a face esantionari folosind noile metode electronice de calcul. O astfel de abordarea parea potrivita pentru explorarea reactiilor in lant ale neutronilor in dispozitivele de fisiune. In formularea lui von Neumann, a problemei difuziei neutronilor, istoria fiecarui neutron era analoaga cu un singur joc de Solitaire iar folosirea numerelor aleatorii, pentru a face alegeri de-a lungul traiectoriei acestora, era analoaga cu intoarcerea unei noi carti in timpul jocului. Astfel, pentru a realiza o simulare Monte Carlo era nevoie de un generator de numere aleatoare si de simulari ale unor variabile aleatoare, specifice fiecarui tip de decizie. Se stia din fizica ca un model bun pentru traiectoriile neutronilor intre coliziuni presupunea o distributie exponentiala. Pentru a simula o astfel de distributie von Neumann si Ulam au pornit de la o distributie uniforma pe intervalul $(0, 1)$ si au construit, intr-o serie de corespondenţe, ceea ce azi numim metodele inversării funcţiei de repartiţie si metoda respingerii.



THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan,

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow

Ulterior, deoarece utilizarea simularilor Monte Carlo necesita un numar mare de numere aleatorii, s-a ajuns la dezvoltarea generatoarelor de numere pseudo-aleatorii. Este suprinzator cum natura poate genera cu usurinta evenimente aleatoare dar tot ceea ce omul creaza sfarseste prin a fi pseudo-aleator. Spre exemplu, dezintegrarea radioactiva survine aleator. S-ar putea, deci, folosi un contor de particule Geiger pentru a capta radiatii emise în urma dezintegrarii unui element radioactiv si utiliza valorile contorului pentru a genera numere aleatoare.

Proiectele de tip Big Science (cercetare stiintifica la scara larga, care implica colaborarea a mii de oameni de stiinta), ca de exemplu proiectul Manhattan, proiectul genomului uman sau multe proiecte ale NASA, sunt proiecte extrem de vulnerabile din punct de vedere al finantarii, fiind dependente de factorul politic. Impactul asupra stiintei este vizibil si dupa o lunga perioada de la finalizarea acestora. In ziua de azi, metodele Monte Carlo se folosesc pe scara larga, in studiul fenomenelor care implica o anumita incertitudine in desfasurarea lor, in domenii ca si grafica computerizata, inteligenta artificiala aplicata in studierea jocurilor, in operatiuni de cautare si salvare , in estimarea riscurilor (finante), programarea robotilor autonomi (localizarea Monte Carlo), procesarea semnalelor sau chiar in astrofizica (modelarea evolutiei unei galaxii).

▶ *Metode de simulare a variabilelor aleatoare*

• majoritatea limbajelor de programare au functii predefinite de generare aleatoare (pseudo-aleatoare) a unor numere care de obicei apartin intervalelor $[0,1)$ sau $(0,1)$, pe care le vom nota cu **rand(0,1)** intr-un context general

• aceste functii au la baza de cele mai multe ori algoritmul Mersenne Twister (MT19937) sau batranul algoritm congruential liniar

$\Longrightarrow$ de amintit faptul ca J. von Neumann isi crease propriul generator care nu era insa suficient de performant

• plecand de la aceste functii vom construi algoritmi de simulare a valorilor unor variabile aleatoare clasice, in conformitate cu densitatile de probabilitate f sau functiile de repartitii F date

Metoda inversarii functiei de repartitie

Ideea generala: putem genera o variabila aleatoare X cu o functie de repartitie dorita F daca stim inversa sa F^{-1}, pornind de la o variabila uniform distribuita $U \sim \mathcal{U}(0,1)$, prin formula

$$X = F^{-1}(U)$$

deoarece pentru o variabila uniform distribuita avem $P(U \leq u) = u$ si astfel

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

asadar X va avea functia de repartitie F.

• in cazul variabilelor aleatoare discrete

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

functia de repartitie $F : \mathbb{R} \to [0,1]$, $F(x) = P(X \leq x)$, are urmatoarea forma

$$F(x) = \begin{cases} 0, & x < x_1 \\ p_1, & x_1 \leq x < x_2 \\ p_1 + p_2, & x_2 \leq x < x_3 \\ \quad\vdots & \\ p_1 + p_2 + \ldots + p_k & x_k \leq x < x_{k+1} \\ \quad\vdots & \\ \quad 1 & x_n \leq x \end{cases}$$

• definim o "functie inversa" $F^{-1} : [0,1] \to \mathbb{R}$ prin formula

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$$

• daca generam $U \sim \mathcal{U}(0,1)$ pentru a obtine o variabila aleatoare cu functia de repartitie F, prin $X = F^{-1}(U)$, trebuie sa tinem cont de relatiile

$$X = \begin{cases} x_1, & u < p_1 \\ x_2, & p_1 \leq u < p_1 + p_2 \\ \vdots & \\ x_k, & p_1 + p_2 + \ldots + p_{k-1} \leq u < p_1 + p_2 + \ldots + p_{k-1} + p_k \\ \vdots & \\ x_n, & p_1 + p_2 + \ldots + p_{n-1} \leq u < 1 \end{cases}$$

• cu alte cuvinte, stim sa generam automat un numar aleator $u \in [0, 1)$ dar apoi trebuie sa il conectam la o valoare x corespunzatoare variabilei aleatoare X si vom realiza asta plecand de la faptul ca functia de repartitie F are domeniul de valori intervalul $[0, 1]$

• pentru a determina valoarea x, a lui X, trebuie sa identificam intervalul $[F(x_i), F(x_{i+1}))$ caruia ii apartine valoarea generata u

Algoritmul general

Input: functia de repartitie F
Output: o variabila X cu functia de repartitie F

Pas 1: $u := rand(0, 1)$ // genereaza o valoare uniform distribuita in $[0, 1)$
Pas 2: initializeaza $i = 1$ si $p = F(x_1)$
Pas 3: daca $u < p$ atunci seteaza $X := x_i$ si stop
Pas 4: incrementeaza $i \leftarrow i + 1$ si $p := F(x_i)$
Pas 5: repeta Pasul 3

• pentru o variabila continua X algoritmul este mai simplu, dar presupune cunoasterea expresiei lui F^{-1}

Algoritmul general

Input: inversa F^{-1} a functiei de repartitie
Output: valori ale unei variabile aleatoare X cu distributia data de F

Pas 1: $u := rand(0, 1)$
Pas 2: $x := F^{-1}(u)$
Pas 3: returneaza x

Metoda transformarilor

• foarte multe variabile aleatoare pot fi obtinute in functie de alte variabile aleatoare prin diferite transformari, de exemplu o variabila aleatoare normal distribuita $X \sim \mathcal{N}(m, \sigma^2)$ poate fi obtinuta dintr-una normal standard distribuita $Z \sim \mathcal{N}(0, 1)$

$$X = m + \sigma Z$$

• din acest motiv vom prezenta mai tarziu doar un procedeu de simulare a unei variabile Z normal standard distribuita (Box-Muller), cazul general fiind apoi obtinut prin adaugarea unei linii $x := m + \sigma * z$ in algoritm

• analog, putem obtine o variabila $X \sim \mathcal{U}(a,b)$ uniform distribuita pe $[a,b]$ prin transformarea

$$X = a + (b-a)Y$$

unde $Y \sim \mathcal{U}(0,1)$.

Metoda combinarii

• uneori o functie de repartitie F poate fi descompusa ca o suma de functii de repartitie

$$F(x) = \sum_{i=1}^{n} p_i \cdot G_i(x)$$

unde $p_i > 0$ si $\sum_{i=0}^{n} p_i = 1$

• o metoda de a simula o astfel de variabila presupune simularea unei variabile aleatoare Y cu proprietatea $p_i = P(Y = i)$ iar la pasul urmator, in functie de valoarea i obtinuta, simulam X considerand F_i ca fiind functia sa de repartitie

Algoritmul general

Input: functiile de repartitie G_i si un algoritm pentru simularea varaibilei Y
Output: valori ale unei variabile aleatoare X cu distributia data de F

Pas 1: genereaza o valoare i corespunzatoare variabilei Y
Pas 2: returneaza valoarea generata prin algoritmul de simulare al variabilei cu functia de repartitie G_i

Metoda convolutiei

• unele variabile aleatoare pot fi obtinute prin simpla insumare a unora clasice, de exemplu

$$X_1, X_2, \ldots X_n \sim Ber(p) \ \text{ independente } \implies Y = \sum_{i=1}^{n} X_i \sim Bin(n,p)$$

$$X_1, X_2, \ldots X_n \sim Exp(\lambda) \ \text{ independente } \implies Y = \sum_{i=1}^{n} X_i \sim Erlang(n,\lambda)$$

sau

$$X_1, X_2, \ldots X_n \sim Geo(p) \ \text{ independente } \implies Y = \sum_{i=1}^{n} X_i \sim NB(n,p)$$

• spunem ca distributia lui Y este convolutia distributiilor lui X_i, $i = \overline{1,n}$

Algoritmul general

Input: metoda de simulare a variabilei X, unde $X_i \sim X$, $i = \overline{1,n}$
Output: valori ale variabilei aleatoare Y

Pas 1: simuleaza valori ale variabilelor $X_1, X_2, \ldots X_n$
Pas 2: insumeaza valorile si returneaza y

__Metoda Box-Muller pentru simularea variabilelor normale__

• cea mai cunoscuta metoda, numita metoda Box-Muller, presupune generarea unor variabile U_1, U_2 uniform distribuite cu valori in $(0,1)$ si apoi construirea unei bijectii

$$g : (0,1) \times (0,1) \to \mathbb{R}^2$$

astfel incat $(Y_1, Y_2) = g(U_1, U_2)$ sa fie normal distribuite

• din motive de diferentiabilitate Box si Muller au ales functia

$$g(u_1, u_2) = \left(\sqrt{-2\ln u_1}\cos(2\pi u_2), \sqrt{-2\ln u_1}\sin(2\pi u_2)\right)$$

• aceasta functie nu poate fi utilizata direct deoarece sin si cos creaza probleme de eficienţă algoritmului, prin urmare se prefera metoda polara Box-Muller

Algoritmul general

Input: doua variabile uniform distribuite U_1, U_2
Output: doua variabile normal standard distribuite Z_1, Z_2

Pas 1: simuleaza $U_1, U_2 \sim \mathcal{U}(0,1)$ generand u_1 si u_2
Pas 2: seteaza $v_1 := 2u_1 - 1$, $v_2 := 2u_2 - 1$ si $r := v_1^2 + v_2^2$
Pas 3: daca $r > 1$ repeta Pas 1
Pas 4: returneaza valorile z_1, z_2, corespunzatoarea unor variabile independente, normal standard distribuite

$$z_1 = \sqrt{\frac{-2\ln r}{r}}v_1 \quad \text{si} \quad z_2 = \sqrt{\frac{-2\ln r}{r}}v_2$$

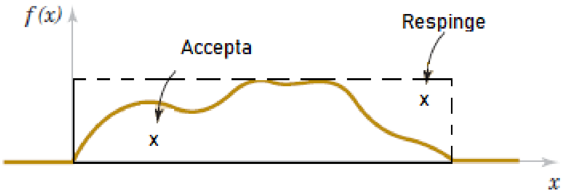
__Metoda respingerii__

• aflarea unei formule pentru inversa F_X^{-1} nu este intotdeauna posibila sau algoritmul propus nu este suficient de eficient, prin urmare propunem o metoda alternativa, construita de catre J. von Neumann

Ideea generala: daca stim densitatea de probabilitate $f(x)$ a unei variabile aleatoare X si aceasta densitate este nenula doar in intervalul $[a,b]$ iar valoarea maxima a lui f este c, atunci generam uniform puncte (x_i, y_i) din dreptunghiul $[a,b] \times [c,d]$ dupa care le respingem pe cele care nu convin dupa urmatorul algoritm

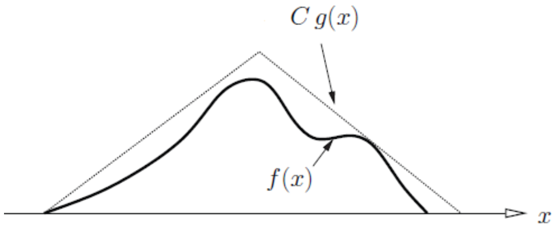
Algoritmul general

Pas 1: seteaza $x := a + (b-a) * rand(0,1)$ si $y := c * rand(0,1)$
// genereaza $x \in [a,b]$ si $y \in [0,c]$ uniform distribuite cu metoda
// transformarilor
Pas 2: daca $y \leq f(x)$ se accepta valoarea x
daca $y > f(x)$ se respinge valoarea x
Pas 3: incrementeaza contorul de numarare al incercarilor (pentru a putea controla procesul generarii) si revino la primul pas



• pentru cazul in care intervalul $[a, b]$ este nemarginit este nevoie de o metoda mai generala si anume presupunem ca f este majorata de o densitate de probabilitate g, care corespunde unei variabile aleatoare Y, pentru care avem deja un algoritm de simulare

$$f(x) \leq Cg(x), \qquad C \geq 1.$$



• un exemplu des intalnit este reprezentat de variabila aleatoare $X = |Z|$, unde $Z \sim \mathcal{N}(0, 1)$, si prin urmare are densitatea $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, care poate fi majorata pe intervalul $[0, \infty)$ printr-o densitate de probabilitate exponentiala $g(x) = e^{-x}$, $x \geq 0$

• in cazul exponential putem folosi usor metoda inversarii functiei de repartitie caci $G(x) = 1 - e^{-x}$, pentru $x \geq 0$, si prin urmare

$$Y = G^{-1}(U) = -\ln(1 - U), \qquad U \in \mathcal{U}(0, 1)$$

• putem sa folosim aceeasi abordare pentru a simula o variabila normal standard distribuita $Z \sim \mathcal{N}(0, 1)$, prin intermediul metodei combinarii, datorita relatiei

$$F_Z(z) = \frac{1}{2} F_{-|Z|}(z) + \frac{1}{2} F_{|Z|}(z)$$

Algoritmul general

Input: densitatea de probabilitate g si o constanta C
Output: o variabila aleatoare X cu densitatea de probabilitate f

Pas 1: found← false
Pas 2: while not found do
    genereaza x cu distributia Y
    alege $u := C * g(x) * rand(0, 1)$
// genereaza u cu distributia uniforma pe $(0, Cg(x))$
    daca $u \leq f(x)$ atunci found← true
Pas 3: returneaza x

• in cazul discret poate fi aplicata o strategie similara, acolo avand $p_i \leq Cq_i$, pentru doua variabile aleatoare X, Y, cu $p_i = P(X = i)$ si $q_i = P(Y = i)$

Simularea variabilelor in C++

• vom particulariza, pentru limbajul C++, cateva aspecte legate de cei doi piloni ai simularii variabilelor aleatoare: simularile variabilelor uniform distribuite si ale celor normal distribuite

• versiunea standard a C++ contine functia *rand* inclusa in libraria *cstdlib*, un apel al functiei *rand* returneaza o valoare intreaga intre 0 si o constanta numita $RAND\_MAX$, care de obicei este $2^{31} - 1$

• cel mai bine este sa definim propriile functii care sa genereze numere pseudo aleatoare si in principiu doua astfel de functii sunt utile: una care sa genereze numere naturale cuprinse 1 si un numar natural dat si una care sa genereze numere reale dintr-un interval fixat $[a, b]$

```
#ifndef   UNIFORM_H
#define   UNIFORM_H
double unif();
   /** genereaza un numar aleator intre 0 si 1
     *  returneaza o valoarea uniform distribuita din intervalul [0, 1]
     */
double unif(double a, double b);
   /** genereaza un numar real intr-un interval dat [a, b]
     */
long unif(long n);
   /** genereaza un numar natural intre 1 si o valoare data n
     * parametrul n reprezinta cea mai mare valoare pe care
     * o poate produce functia
     */
void seed();
   /** e nevoie sa resetam generatorul de numere pseudo-aleatoare
     */
  # endif
```

• procedura *rand* si celelalte care se bazeaza pe *rand* de mai sus returneaza un sir imprevizibil de valori, dar de fiecare data acelasi sir, din cauza ca se doreste a avea caracterul de reproductibilitate

• mai mult, valoarea de start, *seed*, este fixata

• in cazul nostru acest aspect este nedorit si putem evita elegant acest fenomen folosind *srand* si functia *time* (adaugam headerul *ctime*) care va returna o valoare *time(0)*, insemnand numarul de secunde scurse de la un moment de timp fixat (care depinde de computer)

• uneori compilatorul cere modificarea valorii timp returnata, intr-una fara semn, si atunci se poate inlocui ultima linie cu *srand(unsigned(time(0)));*

• in cele ce urmeaza este prezentat modul de definire al functiilor unif, putandu-se observa particularizarea pseudocodului de pe o pagina anterioara

```
#include "uniform.h"
#include <cstdlib>
#include <ctime>
#include <cmath>
using namespace std;

double unif() {
return rand() / double(RAND_MAX);
}

double unif(double a, double b) {
return (b-a)*unif() + a;
}

long unif(long a) {
if (a < 0) a = -a;
if (a==0) return 0;
return long(unif()*a) + 1;
}

void seed() {
srand(time(0));
 }
```

• in cele de urmeaza ne vom ocupa de implementarea metodei Box-Muller, un program C++ bazat pe metoda polara Box-Muller este prezentat mai jos

```
#include <cmath>
#include "uniform.h"
using namespace std;
double randn() {
double x,y,r;
do {
x = unif(-1.,1.);
y = unif(-1.,1.);
r = x*x + y*y;
} while (r >= 1.);
double mu = sqrt(-2.0 * log(r) / r);
return mu*x;
}
```

• header-ul *randn.h* arata in felul urmator

```
#ifndef RANDN_H
#define RANDN_H
#include "uniform.h"
double randn();
#endif
```

• o versiune completa poate arata in felul urmator

```
#include "randn.h"
#include <cmath>
using namespace std;

double randn() {
static bool has_saved = false;
static double saved;

if (has_saved) {
has_saved = false;
return saved;
}

double x,y,r;
do {
x = unif(-1.,1.);
y = unif(-1.,1.);
r = x*x + y*y;
} while (r >= 1.);

double mu = sqrt(-2.0 * log(r) / r);

saved = mu*y;
has_saved = true;

return mu*x;
}
```

• prin modificarea facuta, progamul este de doua ori mai rapid decat cel anterior care avea o problema de eficienta, deoarece genera doua variabile aleatoare normal distribuite Z_1 si Z_2

• noua versiune contine doua variabile statice: una booleana *has_saved* si una reala *saved*

• variabila *has_saved* este presetata ca fiind falsa, pentru a arata ca procedura nu contine in acel moment o valoare normal distribuita salvata, apoi se verifica daca exista una care nu a fost folosita

• apoi se genereaza doua variabile normal distribuite prin metoda Box-Muller polara si vom salva z_2 in *saved* si setam *has_saved* ca fiind *true* si returnam z_1

• prin salvarea lui z_2 partea inceata a algoritmului ruleaza doar o singura data, dubland viteza

▶ *Vectori aleatori*

• pentru a analiza anumite experimente aleatorii este nevoie de doua variabile aleatoare X si Y, vezi primele doua probleme rezolvate, si vom numi perechea (X, Y) vector aleator

• in cazul discret, functia de repartitie comuna perechii X,Y este definita prin

$$F_{X,Y}(x, y) = P(X \leq x,\ Y \leq y)$$

• probabilitatea unui eveniment $(X, Y) = (x, y)$ se calculeaza prin

$$P_{X,Y}(x, y) = P(X = x,\ Y = y)$$

folosind un tabel de repartitie comuna, la fel ca in fisa seminarului 10

• insa in cazul unor variabile X, Y independente avem formula

$$P_{X,Y}(x, y) = P(X = x,\ Y = y) = P(X = x) \cdot P(Y = y)$$

• de retinut ca $\{X = x,\ Y = y\}$ reprezinta acum un eveniment al unui experiment iar multimea starilor devine

$$S_{X,Y} = \{(x, y) : \quad P_{X,Y}(x, y) > 0\}$$

• probabilitatea unui eveniment E va fi

$$P(E) = \sum_{(x,y)\in E} P_{X,Y}(x, y)$$

iar functiile de probabilitate marginale sunt

$$P_X(x) = \sum_{y\in S_Y} P_{X,Y}(x, y), \qquad P_Y(y) = \sum_{x\in S_X} P_{X,Y}(x, y)$$

aflandu-se prin adunarea pe linie sau coloana a valorilor din tabelul de repartitie comun, vezi fisa seminarului despre variabile aleatoare discrete.

• in cazul unui vector aleator (X, Y) continuu este nevoie sa manevram integrale duble, vezi fisa de seminar daca e necesar

• densitatea comuna de probabilitate $f_{X,Y}(x, y)$ va avea proprietatile caracteristice

$$f_{X,Y}(x, y) \geq 0 \ \text{ si } \ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\ dxdy = 1$$

iar functia de repartitie comuna este definita prin

$$F_{X,Y}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)dudv$$

in consecinta

$$P(a \leq X \leq b, \quad c \leq Y \leq d) = F_{X.Y}(b, d) - F_{X.Y}(b, c) - F_{X.Y}(a, d) + F_{X.Y}(a, c)$$

• pentru variabile aleatoare independente avem relatia

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

• probabilitatea unui eveniment E se calculeaza prin

$$P(E) = \iint_E f_{X,Y}(x, y)\ dxdy$$

• densitatile de probabilitate marginale sunt obtinute din

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$$

Probleme rezolvate

Problema 1

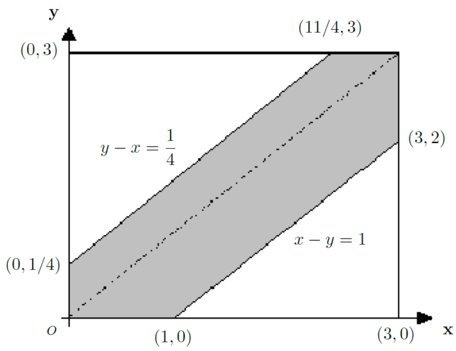
Problema intalnirii revizitata
Un barbat si o femeie decid sa se intalneasca intr-un restaurant dupa ora 21. Restaurantul se inchide la ora 24. Din cauza programului incarcat al fiecaruia ei decid ca in cazul in care unul dintre ei va intarzia fiecare sa astepte dupa celalalt un anumit timp. Barbatul este dispus sa astepte o ora iar femeia doar 15 minute! Care este probabilitatea ca cei doi sa se intalneasca la acel restaurant?

Solutie: Notam cu X variabila aleatoare continua care masoara timpul la care soseste femeia (fara vreo aluzie cromozomiala) si cu Y variabila aleatoare care masoara timpul la care soseste barbatul. Fiecare poate ajunge la restaurant in orice moment de timp intre ora 21 si 24 cu aceeasi probabilitate, deci X si Y au distributia uniform continua. Mai mult, cele doua variabile sunt independente si putem sa consideram ca ambele iau valori in intervalul $[0, 3]$ unde 0 inseamna ora 21 si 3 inseamna ora 24.

Avem asadar un vector aleator (X, Y) care modeleaza matematic problema si masoara timpii la care sosesc cei doi. Notam cu E evenimentul: cei doi se intalnesc in restaurant. Probabilitatea lui E se calculeaza conform formulei

$$P(E) = \iint_E f_{X,Y}(x, y)\, dxdy$$

unde E este regiunea gri din desenul de mai jos.



Deoarece X si Y sunt independente densitatile de probabilitate sunt in relatia

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

unde

$$f_X(x) = \begin{cases} \frac{1}{3-0}, & x \in [0, 3] \\ 0, & \text{altfel} \end{cases} \quad \text{si} \quad f_Y(y) = \begin{cases} \frac{1}{3-0}, & y \in [0, 3] \\ 0, & \text{altfel} \end{cases}$$

sunt densitatile de probabilitate ale celor doua variabile aleatoare continue uniform distribuite. Asadar

$$P(E) = \iint_E \frac{1}{9}\, dxdy.$$

Pentru a calcula aceasta integrala dubla notam cu Δ_1 si Δ_2 cele doua trinughiuri albe din desenul anterior. Se observa ca

$$\Delta_1 \cup A \cup \Delta_2 = [0,3] \times [0,3]$$

Prin urmare

$$\iint_{[0,3]\times[0,3]} \frac{1}{9}\, dxdy = \iint_{\Delta_1} \frac{1}{9}\, dxdy + \iint_A \frac{1}{9}\, dxdy + \iint_{\Delta_2} \frac{1}{9}\, dxdy.$$

Aplicam teorema lui Fubini si scriem cele doua triunghiuri ca domenii simple in raport cu axa Oy (numite regiuni de tip I in acest fisier)

$$\int_0^3 \int_0^3 \frac{1}{9}\, dxdy = \int_0^{\frac{11}{4}} \int_{x+\frac{1}{4}}^3 \frac{1}{9}\, dydx + \iint_A \frac{1}{9}\, dxdy + \int_1^3 \int_0^{x-1} \frac{1}{9}\, dydx$$

$$\iff \quad 1 = \frac{121}{288} + \iint_A \frac{1}{9}\, dxdy + \frac{4}{18}$$

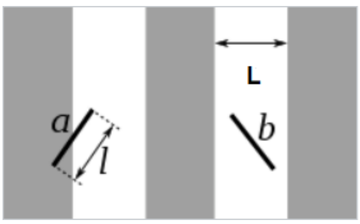
$$\implies P(E) = \iint_A \frac{1}{9}\, dxdy = 0.35 = 35\%$$

In final, ar fi indicata o comparare cu abordarea prezentata in introducerea fisei seminarului despre probabilitati. Pe parcursul solutiei, in evaluarea unor integrale am preferat o abordare tipica integralelor duble (pentru antrenament), evitand investigarea mult mai naturala a ariilor.

Problema 2

Problema acului lui Buffon
O podea are linii paralele situate la distanta L una de cealalta. Un ac de lungime ℓ este aruncat aleator pe podeaua. Gasiti probabilitatea ca acul sa intersecteze una dintre linii.

Solutie: Problema are o istorie aparte, fiind propusa de contele de Buffon in secolul al XVIII-lea. Poate fi interpretata ca descriind o metoda Monte Carlo (aruncari repetate ale acului pe podea) pentru aproximarea numarului π, deoarece solutia problemei este $p = \frac{2}{\pi}\frac{\ell}{L}$.



Pentru a simplifica lucrurile vom trata doar cazul in care lungimea acului este mai mica decat distanta dintre liniile podelei $\ell < L$.

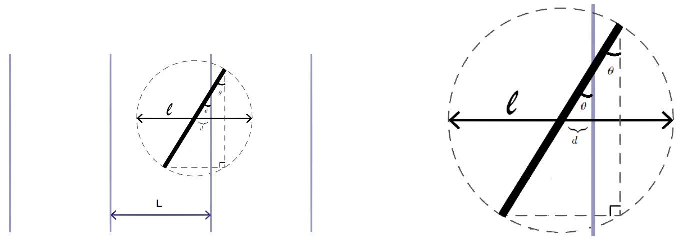
Inainte de toate trebuie remarcat faptul ca unghiul ascutit θ, pe care acul il formeaza cu directia descrisa de liniile podelei, influenteaza sansa ca acul sa atinga o linie. Un ac care cade paralel cu liniile nu va atinge prea des liniile (decat daca ajunge exact pe una dintre linii), indiferent de lungimea sa ℓ.

Se poate observa ca nu este suficient sa cunoastem unghiul θ format de ac si linii pentru a preciza pozitia acului si implicit a modela matemcatic problema, dupa cum decurge si din desen. Pozitia finala a acului depinde si de distanta dintre ac si oricare dintre liniile podelei. Vom considera pozitia fata de cea mai apropiata linie (dupa ce cade pe podea). Putem estima aceasta distanta in multe moduri, dar cea mai eleganta cale pare a fi sa masuram distanta de la mijlocul acului la aceasta cea mai apropiata linie. Notam cu d aceasta distanta este evident ca d poate avea valori intre 0 (cand mijlocul acului se afla pe linie) si $\frac{L}{2}$ (cand mijlocul acului se afla exact la mijlocul distantei dintre doua linii). Orice valoare din intervalul $[0, \frac{L}{2}]$ poate fi atinsa de catre d cu aceasi probabilitate, prin urmare d va fi o valoare a unei variabile aleatoare D cu distributie uniforma pe $[0, \frac{L}{2}]$. Si in cazul unghiului θ se poate observa ca valoarea minima este 0 iar valoarea maxima este $\frac{\pi}{2}$ si din nou sansa ca acul sa formeze oricare unghi din acest interval este aceeasi, deci θ este o valoare a unei variabile aleatoare Θ uniform distribuite pe $[0, \frac{\pi}{2}]$.

Asadar folosim vectorul aleator (D, Θ) pentru a modela matematic caderea unui ac pe podea. Prin definitie densitatile de probabilitate ale celor doua variabile uniform distribuite sunt

$$f_D(d) = \begin{cases} \frac{1}{\frac{L}{2}} & 0 \le d \le \frac{L}{2} \\ 0 & \text{in rest} \end{cases} \qquad \text{si} \qquad f_\Theta(\theta) = \begin{cases} \frac{1}{\frac{\pi}{2}} & 0 \le \theta \le \frac{\pi}{2} \\ 0 & \text{in rest} \end{cases}$$

Definim evenimentul E: acul intersecteaza una dintre liniile podelei. Este recomandat sa desenam aceasta situatie pentru a observa ca evenimentul are loc doar daca $d \le \frac{\ell}{2}\sin\theta$



In desenul de mai sus este descrisa situatia in care acul intersecteaza cea mai apropiata linie. Se formeaza doua triunghiuri dreptunghice cu un unghi θ si cateta opusa d, respectiv $\frac{\ell}{2}\sin\theta$. Pentru a intersecta linia respectiva trebuie sa avem relatia $d \le \frac{\ell}{2}\sin\theta$ altfel linia albastra se afla in afara zonei cu care acul poate avea puncte comune. Prin urmare putem defini evenimentul E ca fiind

$$E = \left\{ (d, \theta) \in \mathbb{R}^2 \ : \ d \le \frac{\ell}{2}\sin\theta \right\}$$

si conform teoriei prezentate in fisa

$$P(E) = \iint_E f_{D,\Theta}(d, \theta) \ ddd\theta.$$

Din cauza notatiei alese apare neplacuta asociere dd, insemnand integrare in raport cu d :)). Deoarece variabilele D si Θ sunt independente

$$f_{D,\Theta}(d,\theta) = f_D(d)\cdot f_\Theta(\theta) = \begin{cases} \frac{4}{L\pi}, & 0 \le d \le \frac{L}{2},\quad 0 \le \theta \le \frac{\pi}{2} \\ 0, & \text{in rest} \end{cases}$$

Pentru a evalua integrala dubla de mai sus vom scrie multimea E ca un domeniu simplu in raport cu axele

$$E = \left\{ (d,\theta) \in \mathbb{R}^2 \ : 0 \le \theta \le \frac{\pi}{2},\quad 0 \le \ d \le \frac{\ell}{2}\sin\theta \right\}$$

prin urmare

$$P(E) = \iint_E f_{D,\Theta}(d,\theta)\ dd d\theta = \int_0^{\frac{\pi}{2}} \left(\int_0^{\frac{\ell}{2}\sin\theta} \frac{4}{L\pi}\ dd \right) d\theta$$
$$= \int_0^{\frac{\pi}{2}} \frac{4}{L\pi}\frac{\ell}{2}\sin\theta\ d\theta = -\frac{4}{L\pi}\frac{\ell}{2}\cos\theta \Big|_0^{\frac{\pi}{2}} = \frac{2\ell}{\pi L}$$

Problema 3

Scrieti un algoritm (pseudocod) pentru a simula o variabila aleatoare binomiala $X \sim Bin(n,p)$.

Solutie: Functia de probabilitate a unei variabile aleatoare $X \sim Bin(n,p)$ este $f(k) = C_n^k p^k (1-p)^{n-k}$, $k = \overline{1,n}$. Este important sa remarcam recurenta $f(k+1) = \frac{n-k}{k+1}\frac{p}{1-p} f(k)$. Vom crea un algoritm bazat pe metoda inversarii functiei de repartitie:

Algoritm

Pas 1: citeste n si p si seteaza $c := \frac{p}{1-p}$, $i := 0$, $prob := (1-p)^n$ si $F := prob$
Pas 2: genereaza $u := rand(0,1)$
Pas 3: daca $u < F$ seteaza $x := i$ si stop
Pas 4: incrementeaza prob←prob*c*(n-i)/(i+1) apoi F← F+prob
si $i \leftarrow i+1$
Pas 5: repeta Pas 3

Problema 4

Scrieti un algoritm (pseudocod) pentru a simula o variabila aleatoare cu o distributie beta $X \sim \beta(\mu,\nu)$, pentru $\mu, \nu < 1$. Simulati o variabila aleatoare beta $X \sim \beta(2,2)$ prin metoda respingerii.

Solutie: Reamintim ca $f(x) = \frac{x^\mu (1-x)^\nu}{\beta(\mu,\nu)}$, pentru $x \in [0,1]$, este densitatea de probabilitate corespunzatoare lui $X \sim \beta(\mu,\nu)$. In general algoritmii de simulare a valorilor lui X se bazeaza pe diverse estimari ale lui f si metoda respingerii. In cazul $\mu,\nu < 1$ putem incerca urmatorul algoritm

Algoritm

Pas 1: genereaza $u := rand(0,1)$ si $v := rand(0,1)$

Pas 2: seteaza $x := u^{1/\mu}$ si $y := v^{1/b}$

Pas 3: daca $x + y > 1$ repeta Pas 1, altfel returneaza $\frac{x}{x+y}$ ca valoare a lui X

Pentru o variabila X cu distributie $\beta(2,2)$, densitatea de probabilitate este $f(x) = 6x(1-x)$, pentru $0 \leq x \leq 1$ si $f(x) = 0$ in rest. Deoarece are loc inegalitatea $f(x) \leq 6 \cdot 1$, pentru $x \in (0,1)$, putem alege $g(x) = 1$ care este densitatea de probabilitate a unei variabile uniform distribuite $Y \sim \mathcal{U}(0,1)$, si constanta $C = 6$. Putem crea acum usor un algoritm bazat pe metoda respingerii.

Problema 5

Doua variabile aleatoare discrete X si Y au functia de probabilitate comuna

$$P_{X,Y}(x,y) = \begin{cases} cxy, & x = 1,2,4 \quad y = 1,3 \\ 0, & \text{altfel} \end{cases}$$

i) Cat este $P(Y < X)$?

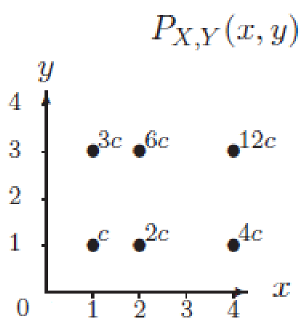
ii) Estimati probabilitatea $P(Y = 3)$.

iii) Aflati functiile de probabilitate marginale ale lui X si Y

iv) Aflati valorile asteptate $M(X)$ si $M(Y)$

v) Calculati deviatia standard σ_X

Solutie: i) Putem aborda aceasta problema fie completand un tabel de repartitie comun la fel ca in problema rezolvata 8 din fisa seminarului despre variabile discrete, fie realizand o reprezentare grafica a acelor valori (x, y), a vectorului aleator (X, Y), care au probabilitatea nenula, la fel ca mai jos



Multimea starilor $S_{X,Y}$, corespunzatoare vectorului aleator (X,Y) este

$$S_{X,Y} = \{(x,y) : \; P_{X,Y}(x,y) > 0\} = \{(1,1),(2,1),(4,1),(1,3),(2,3),(4,3)\}$$

Probabilitatea caracteristica a functiei de probabilitate comuna este

$$\sum_{x \in S_X} \sum_{y \in S_Y} P_{X,Y}(x,y) = 1$$

prin urmare

$$\sum_{x=1,2,4} \sum_{y=1,3} cxy = 28c = 1 \quad \Longrightarrow \quad c = \frac{1}{28}$$

Evenimentul $Y < X$ poate fi descris prin

$$A = \{(x, y) \in S_{X,Y} : y < x\} = \{(2, 1), (4, 1), (4, 3)\}$$

si probabilitatea de realizare va fi

$$P(A) = \sum_{(x,y)\in A} P_{X,Y}(x, y) = \frac{2 \cdot 1 + 4 \cdot 1 + 4 \cdot 3}{28} = \frac{18}{28}$$

ii) Pentru a afla probabilitatea $P(Y = 3)$ fie definim evenimentul $\{Y = 3\}$ care poate fi descris prin

$$B = \{(x, y) \in S_{X,Y} : y = 3\} = \{(1, 3), (2, 3), (4, 3)\}$$

sau putem calcula functia de probabilitate marginala a lui Y conform formulei

$$P_Y(y) = \sum_{x\in S_X} P_{X,Y}(x, y)$$

In particular pentru $y = 3$ otbinem

$$P_Y(3) = \sum_{x\in S_X} P_{X,Y}(x, 3) = \sum_{x=1,2,4} \frac{1}{28} x \cdot 3 = \frac{21}{28}$$

iii) Expresia completa a functiei de probabilitate corespunzatoare lui Y este

$$P_Y(y) = \sum_{x\in S_X} P_{X,Y}(x, y) = \sum_{x=1,2,4} P_{X,Y}(x, y) = \begin{cases} \frac{7}{28}, & y = 1 \\ \frac{21}{28}, & y = 3 \\ 0, & \text{in rest} \end{cases}$$

iar a functiei de probabilitate marginala corespunzatoare lui X

$$P_X(x) = \sum_{y\in S_Y} P_{X,Y}(x, y) = \sum_{y=1,3} P_{X,Y}(x, y) = \begin{cases} \frac{4}{28}, & x = 1 \\ \frac{8}{28}, & x = 2 \\ \frac{16}{28}, & x = 4 \\ 0, & \text{in rest} \end{cases}$$

iv) Valoarea asteptata a lui X poate fi acum calculata folosind functia de probabilitate a lui X

$$M(X) = \sum_{x\in S_X} x \cdot P_X(x) = 1 \cdot \frac{4}{28} + 2 \cdot \frac{8}{28} + 4 \cdot \frac{16}{28} = 3$$

v) Pentru deviatia standard $\sigma_X = \sqrt{D^2(X)}$ trebuie sa calculam dispersia

$$D^2(X) = M(X^2) - M(X)^2 = \sum_{x\in S_X} x^2 \cdot P_X(x) - \left(\sum_{x\in S_X} x \cdot P_X(x) \right)^2 = 10/7$$

In final se obtine $\sigma_X = \sqrt{\frac{10}{7}}$

Problema 6

Fie X variabila aleatoare care masoara timpul scurs pana cand un server se conecteaza la computerul tau (in milisecunde) si Y variabila care masoara timpul scurs pana cand serverul te autorizeaza ca user valid (in milisecunde). Se considera ca X, Y masoara asteptarea dintr-un moment initial comun al timpului si atunci evident avem $X < Y$. Se poate argumenta ca urmatoarea densitate de probabilitate comuna

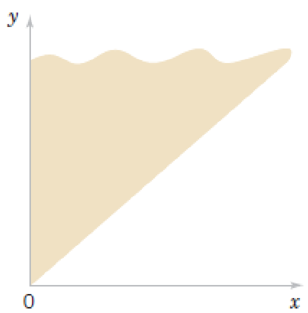
$$f_{X,Y}(x,y) = \begin{cases} 6 \cdot 10^{-6} \cdot e^{-0.001x - 0.002y} & , \quad \text{pentru} \quad 0 \le x < y \\ 0 & , \quad \text{in rest} \end{cases}$$

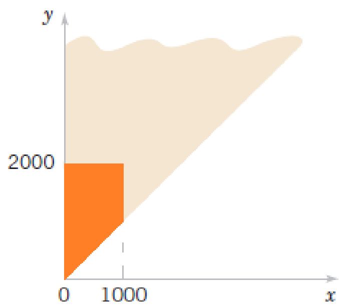
este una potrivita. Aratati ca $f_{X,Y}$ este o densitate de probabilitate si apoi aflati probabilitatea $P(X \le 1000, Y \le 2000)$.

Solutie: Pentru a argumenta ca $f_{X,Y}$ este o densitate de probabilitate, intrucat este o functie pozitiva, va fi suficient sa aratam ca

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dxdy = 1$$

Regiunea din plan unde $f_{X,Y}$ este nenula este cea reprezentata in imaginea din stanga, fiind cuprinsa intre axa Oy si prima bisectoare.





Daca notam cu A respectiva multime

$$A = \{(x,y) \in \mathbb{R}^2 :\ 0 \le x < y\ \}$$

se obtine imediat

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dxdy = \iint_A 6 \cdot 10^{-6} \cdot e^{-0.001x - 0.002y}\ dxdy$$

Pentru ca calcula integrala dubla vom exprima multimea A ca regiune de tip I

$$A = \{(x,y) \in \mathbb{R}^2 : 0 \le x < \infty,\ x < y < \infty\}$$

Prin urmare

$$
\begin{aligned}
\iint_A 6\cdot 10^{-6}\cdot e^{-0.001x-0.002y}\;dxdy &= \int_0^{\infty}\left(\int_x^{\infty} 6\cdot 10^{-6}\cdot e^{-0.001x-0.002y}\;dy\right)\;dx\\
&= \int_0^{\infty} 6\cdot 10^{-6}\cdot e^{-0.001x}\left(\int_x^{\infty} e^{-0.002y}\;dy\right)\;dx\\
&= \int_0^{\infty} 6\cdot 10^{-6}\cdot e^{-0.001x}\frac{e^{-0.002x}}{0.002}\;dx\\
&= 0.003\cdot\int_0^{\infty} e^{-0.003x}\;dx\\
&= 0.003\cdot\frac{1}{0.003}=1
\end{aligned}
$$

Pentru a estima probabilitatea ceruta va trebui sa tinem cont de faptul ca

$$P(X\le 1000,\;Y\le 2000)=F(1000,2000)=\int_{-\infty}^{1000}\int_{-\infty}^{2000} f_{X,Y}(x,y)\;dydx$$

unde F este functia de repartitie comuna. Integrala dubla rezultata poate fi interpretata ca fiind o integrala dubla pe dreptunghiul infinit

$$D=(-\infty,1000]\times(-\infty,2000]$$

Deoarece $f_{X,Y}$ este nenula doar pe multimea A descrisa anterior, se obtine

$$\int_{-\infty}^{1000}\int_{-\infty}^{2000} f_{X,Y}(x,y)\;dydx=\iint_{D\cap A} 6\cdot 10^{-6}\cdot e^{-0.001x-0.002y}\;dxdy$$

Multimea $D\cap A$ este cea reprezentata in imaginea din dreapta, pe pagina anterioara. Pentru a putea calcula integrala reprezentam aceasta multime ca o regiune de tip I

$$D\cap A=\{(x,y)\in\mathbb{R}^2:\;0\le x\le 1000,\;x<y\le 2000\}$$

In consecinta

$$
\begin{aligned}
\iint_{D\cap A} 6\cdot 10^{-6}\cdot e^{-0.001x-0.002y}\;dxdy &= \int_0^{1000}\left(\int_x^{2000} 6\cdot 10^{-6}\cdot e^{-0.001x-0.002y}\;dy\right)\;dx\\
&= \int_0^{1000} 6\cdot 10^{-6}\cdot e^{-0.001x}\left(\int_x^{2000} e^{-0.002y}\;dy\right)\;dx\\
&= 6\cdot 10^{-6}\cdot\int_0^{1000} e^{-0.001x}\frac{e^{-0.002x}-e^{-4}}{0.002}\;dx\\
&\approx 0.915=91\%
\end{aligned}
$$

Problema 7

Doua variabile aleatoare X si Y au densitatea de probabilitate comuna

$$f_{X,Y}(x,y)=\begin{cases}6e^{-2x-3y}, & x\ge 0,\;y\ge 0\\ 0, & \text{altfel}\end{cases}$$

Estimati probabilitatile $P(X\ge Y)$ si $P(X+Y\le 1)$.

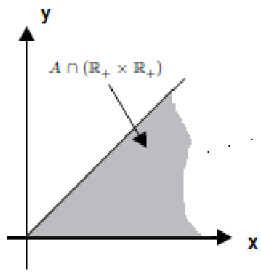
Solutie: Sa consideram evenimentul

$$A = \{X \geq Y\} = \{(x, y) \in \mathbb{R}^2 : x \geq y\}.$$

Deoarece $f_{X,Y}$ este nenula doar pentru $x \geq 0$ si $y \geq 0$, ne putem restrictiona la primul cadran $\mathbb{R}_+ \times \mathbb{R}_+$ cand vom estima probabilitatile

$$P(A) = \iint_A f_{X,Y}(x, y)\ dxdy = \iint_{A \cap (\mathbb{R}_+ \times \mathbb{R}_+)} f_{X,Y}(x, y)\ dxdy$$

De fiecare data cand avem de-a face cu integrale duble este foarte important sa desenam domeniul pe care integram. In cazul de fata avem de integrat pe $A \cap (\mathbb{R}_+ \times \mathbb{R}_+)$ care este regiunea cuprinsa intre dreapta $y = x$ si axa Ox



Multimea $A \cap (\mathbb{R}_+ \times \mathbb{R}_+)$ poate fi rescrisa ca o regiune de tip I

$$A \cap (\mathbb{R}_+ \times \mathbb{R}_+) = \{(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+ :\ 0 \leq x < \infty,\ 0 \leq y \leq x\}$$

Urmariti cu atentie modul de manipulare a integralelor duble in fisierul din linkul de mai sus. Ca o consecinta

$$\begin{aligned} P(A) &= \iint_A f_{X,Y}(x, y)\ dxdy = \iint_{A \cap (\mathbb{R}_+ \times \mathbb{R}_+)} 6e^{-2x-3y}\ dxdy \\ &= \int_0^\infty \left(\int_0^x 6e^{-2x-3y}\ tcdy \right) dx = \int_0^\infty 2e^{-2x} \left(-e^{-3y} \Big|_0^x \right)\ dx \\ &= \int_0^\infty (2e^{-2x} - 2e^{-5x})\ dx = \frac{3}{5} \end{aligned}$$

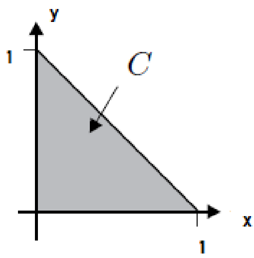
La ultimul pas am sarit peste calcule, dar avem spre exemplu

$$\int_0^\infty 2e^{-2x}\ dx = \lim_{b \to \infty} \int_0^b 2e^{-2x}\ dx = \lim_{b \to \infty} -e^{-2x} \Big|_0^b = \lim_{b \to \infty} (-e^{-2b} + 1) = 1$$

Pentru a doua parte a problemei definim evenimentul

$$B = \{(x, y) \in \mathbb{R}^2 : x + y \leq 1\}$$

si din nou, din moment ce f este nenula doar pentru valori pozitive ale lui x, y, este suficient sa calculam probabilitatea $C = B \cap (\mathbb{R}_+ \times \mathbb{R}_+)$



Acum, C poate fi scrisa ca regiune de tip I sau de tip II. Ca regiune de tip II va fi

$$C = \{(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+ : \ 0 \leq y \leq 1, \ 0 \leq x \leq 1 - y\}$$

$$P(C) = \iint_C f_{X,Y}(x, y) \ dxdy = \int_0^1 \left(\int_0^{1-y} 6e^{-2x-3y} dx \right) \ dy = 1 - 2e^{-3} - 3e^{-2}$$

Remarca

Intotdeauna verificati daca probabilitatile obtinute reprezinta numere reale pozitive subunitare (insemnand ca probabilitatea este cuprinsa intre 0% si 100%). Se verifica usor ca $0 < 1 - 2e^{-3} - 3e^{-2} < 1$.

Limbajul matematic si modelarea matematica

"Mathematics is not a science from our point of view, in the sense that it is not a natural science. The test of its validity is not experiment."

Richard P. Feynman, The Feynman lectures on Physics, vol 1

Ideea e ca in momentul in care se cauta o definitie a stiintelor naturale, pe care urmau sa o satisfaca chimia, biologia, fizica, s-a observat ca matematica, pe care toti o credeau din start o stiinta, *nu satisface conditiile definitiei.*

Toate definitiile stiintei se refera la observatii ale lumii exterioare si experimentari, toate legile stiintelor sunt aproximari ale realitatii si in diverse situatii se dovedesc false, de ex: legea atractiei universale. Legile matematice nu sunt niciodata false, intrucat se bazeaza pe un sistem axiomatic. Construirea cunoasterii matematicii fara o baza axiomatica a condus la paradoxuri, vezi paradoxul Russell. Aceasta abordare initiala, numita uneori abordare naiva, a fost schimbata cu o constructie axiomatica si atunci *matematica a inceput sa se indeparteze definitiv de definitia clasica a stiintei.*

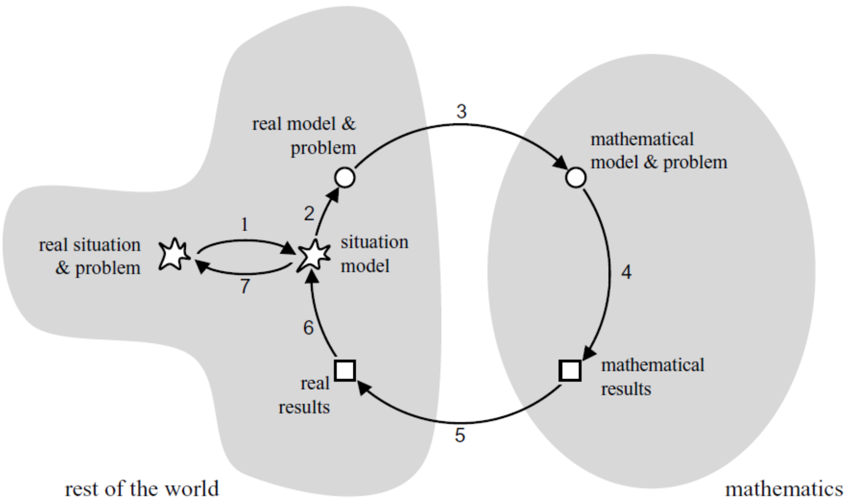
In alta ordine a ideilor, stiinta reprezinta intelegerea lumii exterioare imaginatiei umane iar matematica este produsul imaginatiei umane si nu ceva ce poate fi observat direct. Ideile stiintifice sunt validate prin experiment si observatie, ideile matematice sunt fie presupuse a fi axiome fie deduse dintr-un sistem axiomatic. Concordanta cu realitatea este masura succesului pentru o idee stiintifica. Ideile matematice sunt judecate dupa alte criterii, cum ar fi de exemplu consistenta interna: demonstratia teoremei lui Pitagora nu necesita nicio observatie stiintifica, decurge din definitia unui triunghi dreptunghic intr-un spatiu euclidian

Aceeasi discutie poate fi abordata din perspectiva entitatilor de baza din stiinta sau matematica. Numarul 2 este o entitate diferita de, spre exemplu, luna; luna poate fi perceputa direct, putem folosi matematica pentru a-i masura proprietatile, ca marime sau forma. Prin contrast, nu putem percepe direct numarul 2, putem percepe 2 banane, 2 carti dar nu numarul 2. De fiecare data cand contorizam doua unitati dintr-o entitate, pare ca folosim un concept pre-existent de ”2-ime”. Numarul 2 este un concept dupa care masuratorile sunt realizate, nu reprezinta o masura (text preluat dintr-o discutie de pe Quora).

Matematica nu este o stiinta naturala. Matematica este o unealta sau un limbaj, folosit de catre celelalte stiinte naturale pentru argumentare, vezi si articolul medaliatului Nobel, Eugene Wigner

Prin urmare, limbajul matematic (substituent pentru cuvantul matematica) este utilizat, prin intermediul modelarii matematice, pentru a trage concluzii valide despre fenomenele studiate. Provocarea cea mare a modelarii matematice nu e reprezentata de producerea celui mai comprehensiv model descriptiv ci de producerea celui mai *simplu model care incorporeaza trasaturile majore ale fenomenului studiat.*

Vom discuta in continuare despre etapele modelarii matematice. O reprezentare schematica a acestor etape este oferita in figura de mai jos.



Semnificatia cifrelor (activitatilor) din figura este data mai jos

1. Intelegerea sarcinii
2. Structurare/Simplificare
3. Matematizare
4. Munca matematica
5. Interpretare
6. Validare
7. Expunere

Ne vom folosi din nou de celebra problema a intalnirii pentru a exemplifica acesti pasi. Vom utiliza simulari ale unor variabile aleatoare pentru a valida estimarile obtinute pana acum si a completa *ciclul de modelare* inceput in fisa unui seminar trecut.

Un barbat si o femeie decid sa se intalneasca intr-un restaurant dupa ora 21. Restaurantul se inchide la ora 24. Din cauza programului incarcat, al fiecaruia, ei decid ca in cazul in care unul dintre ei va intarzia fiecare sa astepte dupa celalalt un anumit timp. Barbatul este dispus sa astepte o ora iar femeia doar 15 minute. Care este sansa ca cei doi sa se intalneasca in acel restaurant?

▶ **Intelegerea sarcinii**

Primul pas consta in descifrarea situatiei descrise, cand se incearca construirea unui *model situational* (situation model)

▷ in situatia de fata avem un restaurant deschis intre 21 si 24 si doua persoane care sosesc in acel restaurant

▷ apoi apar doua actiuni: "a astepta" si "a se intalni"

▶ **Structurare/ Simplificare**

La acest pas este nevoie uneori sa simplificam situatia descrisa/sa o idealizam pentru a o putea modela matematic. Aici probabil una dintre *simplificarile necesare* consta in faptul ca nu luam in considerare situatia in care are loc un ambuteiaj, sau intervine ceva neprevazut, si una dintre cele doua persoane nu sosesc in restaurant in intervalul $[21, 24]$. Vom *structura situatia descrisa* mai sus introducand semnificatia celor doua actiuni. Scopul este ca cream un model real al problemei (nu matematic).

▷ daca barbatul asteapta o ora, va fi prezent in restaurant cel mult o ora de la momentul in care soseste, femeia cel mult 15 minute (luam in considerare, spre exemplu, faptul ca femeia poate sosi la 23.50 in restaurant)

▷ avem deja un conflict intre modul in care masuram timpul, va trebui convertit in aceeasi unitate de masura

▷ putem sa remarcam usor ca cei doi se intalnesc daca perioadele petrecute in restaurant se intersecteaza

▶ **Matematizare**

Al treilea pas consta in transformarea modelului real intr-unul matematic, introducand ecuatii si variabile. Se remarca aici si aparitia normalizarii discutate la problem-solving, intrucat dorim sa cautam reprezentarea cat mai eleganta a situatiei descrise in problema. La acest pas va fi probabil nevoie de ghidajul profesorului, intrucat sunt cateva modele matematice posibile. Este evidenta necesitatea de a nota timpii de sosire in restaurant, insa nu e clar daca vom reprezenta doua puncte x, y pe o dreapta sau (x, y) intr-un plan

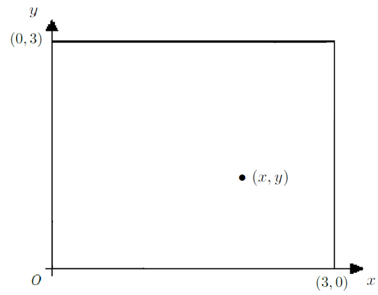
▷ probabil problema admite mai multe reprezentari vizuale, cel descris mai jos pare cel mai elegant

▷ notam cu x timpul la care soseste femeia la restaurant si cu y timpul la care soseste barbatul

▷ un prim obstacol: putem descrie un eveniment posibil folosind perechea (x, y) unde $x, y \in [21, 24]$!

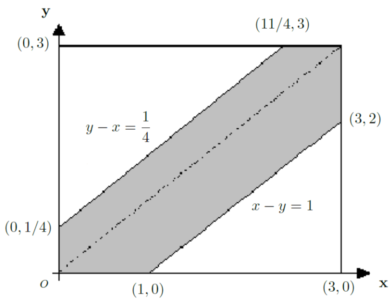
▷ in acest fel obtinem un dreptunghi in $\mathbb{R}^2$, intre dreptele $x = 21$, $x = 24$ si $y = 21$, $y = 24$, si fiecare punct din interiorul sau de pe laturile dreptunghiului reprezinta un eveniment posibil

Aici intervine simplificarea (normalizarea) reprezentarii vizuale. Putem sa consideram ora 21 ca fiind timpul 0 si atunci 24 va fi reprezentat de numarul 3. Asadar $x, y \in [0, 3]$. Toate situatiile posibile sunt reprezentate de punctele (x, y) din interiorul patratului $[0, 3] \times [0, 3]$ de mai jos.



Pentru a **matematiza asteptarea si intalnirea** trebuie sa remarcam ca avem nevoie de inecuatii. In cazul in care barbatul soseste primul, adica $y \leq x$, atunci cei doi se vor intalni daca $x - y \leq 1$ (timpul la care soseste femeia este cu cel mult o ora peste cel al sosirii barbatului). Toti timpii de sosire care satisfac aceste restrictii sunt continuti in regiunea gri, din interiorul patratului, mai precis partea din regiune cuprinsa intre prima bisectoare $y = x$ si dreapta $x - y = 1$ (vezi figura urmatoare)

In cazul in care femeia soseste prima, adica $x \leq y$, atunci cei doi se intalnesc doar daca $y - x \leq \frac{1}{4}$. Toti timpii de sosire care satisfac aceste restrictii sunt continuti in partea superioara a regiunii gri, din interiorul patratului, si anume partea cuprinsa intre prima bisectoare $y = x$ si dreapta $y - x = \frac{1}{4}$



▷ probabilitatea ca cei doi sa se intalneasca este definita in mod clasic

$$P = \frac{\text{numar cazuri favorabile}}{\text{numar cazuri posibile}}.$$

• un alt obstacol: *sunt o infinitate de cazuri favorabile si o infinitate de cazuri posibile.*

▶ **Munca matematica**

La acest pas trebuie sa obtinem rezultate matematice, sa facem estimari, calcule. Pentru inceput trebuie sa depasim obstacolul gasit anterior.

▷ va trebui sa contorizam intr-un alt mod punctele (x, y) care corespund celor doua multimi; in loc sa numaram puncte, vom ”masura” multimi

▷ estimam probabilitatea utilizand ariile regiunilor care descriu geometric multimea cazurilor favorabile, respectiv multimea cazurilor posibile.

▷ probabilitatea ca cei doi sa se intalneasca

$$P = \frac{\text{aria regiunii gri}}{\text{aria patratului}} = \frac{\frac{103}{32}}{3^2} \approx 35.7\%.$$

▶ **Interpretare**

La acest pas trebuie sa interpretam rezultatele obtinute ca rezultate in lumea reala si eventual sa formulam recomandari/ extragem concluzii posibile cu ajutorul modelului imperfect construit.

▶ **Validare**

O invalidare in lumea reala a rezultatelor obtinute poate conduce la o reparcurgere a ciclului de modelare. Va trebui sa construim un alt model, care eventual sa tina cont si de alte variabile. Prin urmare e posibil la acest pas sa constatam ca modelul simplist construit nu contine toate trasaturile esentiale ale fenomenului studiat.

▷ in functie de factorii noi introdusi in ciclu, recomandarile finale pot fi foarte diferite

▷ o validare partiala a rezultatului obtinut se poate realiza prin simulari, folosind spre exemplu Matlab vom simula doua variabile aleatoare uniform distribuite (timpii de sosire a celor doi)

```
function p=meeting_sim(n)
X=unifrnd(0,3,1,n); Y=unifrnd(0,3,1,n);
count=0;
for j=1:n
if (X(j)<Y(j)&&Y(j)<=X(j)+1/4)||(Y(j)<=X(j)&&X(j)<=Y(j)+1)
count=count+1;
end
end
p=count/n;    % probabilitatea estimata
end
```

```
>> meeting_sim(5000)
ans=0.3576
```

▷ cele 5000 de simulari realizate de algoritmul de mai sus sugereaza ca probabilitatea ca cei doi sa se intalneasca in acel restaurant este aproximativ de $35,76\%$

▷ validarea in practica a rezultatului depinde insa de alegerile facute in construirea modelului/ simplificarea lui

De remarcat ca rezultatul gasit prin simularea timpilor de sosire in restaurant este identic cu cel obtinut folosind limbajul matematic. Unii matematicieni considera ca problemele referitoare la estimarea sansei trebuie tratate doar in acest mod, argumentand prin intermediul unor simulari.

▶ **Expunere**

Pasul final presupune o prezentare a rezultatelor finale.

Observatii finale

Putem vedea intr-un context bine definit cum este utilizat limbajul matematic pentru argumentare, in studiul unor fenomene. Am studiat un fenomen care implica sansa. Intr-un top al celor mai importante calitati, pe care un inginer de succes trebuie sa le aiba, se afla in mod constant abilitatea de a utiliza limbajul matematic. Exista printre lingvisti crezul ca invatarea unui limbaj nou iti poate schimba modul in care gandesti sau in care percepi realitatea. Invatarea limbajului matematic *trebuie sa lase urme* in modul in care gandim sau percepem realitatea. Matematica nu inseamna totul, matematica este pretioasa intrucat reprezinta o unealta utilizata in argumentare dar si un mediu in care poti antrena abilitati de problem-solving.

Probleme propuse

B. Tehnica de calcul

Problema B.1. *Functia de probabilitate comuna a variabilelor aleatoare X si Y este data in tabel*

(a) Aflati functiile probabiliste marginale ale lui X, respectiv Y.
(b) Aflati $P(1 \leq X < 3, Y \geq 1)$.
(c) Determinati daca X si Y sunt independente

X \ Y	0	1	2
0	1/18	1/9	1/6
1	1/9	1/18	1/9
2	1/6	1/6	1/18

Problema B.2. *Densitatea de probabilitate comuna a doua variabile aleatoare X si Y este*

$$f(x, y) = \begin{cases} cxy, & 0 < x < 4, 1 < y < 5 \\ 0, & \textit{in rest} \end{cases}$$

a) Aflati valoarea lui c
b) Aflati $P(1 < X < 2, 2 < Y < 3)$
c) Aflati densitatea de probabilitate marginala a lui X, respectiv Y

Problema B.3. *Scrieti un algoritm de simulare (pseudocod) a unei variabile aleatoare Poisson prin metoda inversarii functiei de repartitie.*

Problema B.4. *Folositi metoda convolutiei pentru a scrie un algoritm de simulare a unei variabile aleatoare $X \sim Erlang(n, \lambda)$. Faceti acelasi lucru apoi pentru o variabila aleatoare cu distributie negativ binomiala $Y \sim NB(n, p)$*

Problema B.5. *Variabilele continue X, Y au densitatea de probabilitate comuna*

$$f_{X,Y}(x, y) = \begin{cases} ce^{-2x-3y}, & x + y \leq 1, \ x \geq 0, \ y \geq 0 \\ 0, & \textit{in rest} \end{cases}$$

i) Aflati valoarea lui c

ii) Evaluati probabilitatea $P(X \leq Y)$

iii) Calculati probabilitatea $P(X+Y \geq \frac{1}{2})$

Problema B.6. *Aruncam o moneda pana in momentul in care se obtine de doua ori pajura. Notam cu* X_1 *numarul de aruncari pana la prima aparitie a pajurei si cu* X_2 *numarul de aruncari aditionale necesare pana la a doua aparitie a pajurei. Consideram variabila aleatoare* $Y = X_1 - X_2$. *Aflati* $M(Y)$ *si* $D^2(Y)$.

Problema B.7. *Scrieti un algoritm de generare a unor puncte uniform distribuite in discul eliptic*

$$D = \left\{ (x,y) \in \mathbb{R}^2 : \; \frac{x^2}{4} + \frac{y^2}{9} \leq 1 \right\}.$$

Problema B.8. *Sa presupunem ca* X_1, X_2 *si* X_3 *reprezinta grosimea exprimata in micrometri a unui substrat, respectiv a unui strat activ si al unui strat de acoperire, corespunzatoare unui produs chimic. Grosimile* X_1, X_2 *si* X_3 *sunt considerate independente si normal distribuite cu* $m_1 = 10000$, $m_2 = 1000$, $m_3 = 80$ *si respectiv* $\sigma_1 = 250, \sigma_2 = 20$, $\sigma_3 = 4$. *Specificatiile tehnice pentru grosimea celor trei straturi sunt* $9200 < x_1 < 10.800$, $950 < x_2 < 1050$, *si respectiv* $75 < x_3 < 85$. *Ce proportie a produselor chimice satisface aceste specificatii ? Care dintre specificatii sunt cel mai greu de realizat ?*

Problema B.9. *Variabilele aleatoare continue* X, Y *au densitatea de probabilitate comuna data de*

$$f_{X,Y}(x,y) = \begin{cases} 2, & 0 \leq y \leq x \leq 1 \\ 0, & \textit{in rest} \end{cases}$$

Aflati densitatile de probabilitate marginale ale lui X, *respectiv* Y. *Aflati functia de repartitie comuna* $F_{X,Y}$. *Estimati probabilitatea evenimentului* $Y - X \geq \frac{1}{2}$.

Problema B.10. *Variabilele aleatoare discrete* X, Y *au tabelul de repartitie comun data de*

$P_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$
$x=0$	0.06	0.18	0.24	0.12
$x=1$	0.04	0.12	0.16	0.08

Aflati functiile de probabilitate marginale ale lui X, *respectiv* Y. *Aflati functia de repartitie comuna* $F_{X,Y}$. *Estimati probabilitatea evenimentului* $Y - X \geq \frac{1}{2}$.

Bibliografie

[1] R. Eckhardt. *Stam Ulam, John von Neumann, and the Monte Carlo Method*, Los Alamos Science, Special Issue, 1987.

[2] N. Metropolis. *The Beginning of the Monte Carlo Method*, Los Alamos Science, Special Issue, 1987.

[3] R. Yates and D. Goodman. *Probability and Stochastic processes*, Wiley&Sons, 2005.

[4] R. Negrea. *Note de curs MS*, 2021.

[5] D. Montgomery and G. Runger. *Applied Statistics and Probability for Engineers*, Wiley, 2014.

[6] S. Ross. *Simulation*, Academic Press, 2013.

[7] R. Rubinstein and D. Kroese. *Simulation and the Monte Carlo Method*, Wiley&Sons, 2017.

[8] E. Scheinerman. *C++ for Mathematicians*, Chapmann & Hall/CRC, 2006.