Simulation of random variables

Jean-Michel Marin

University of Montpellier
Faculty of Sciences

HAX918X / 2024-2025

Methods involving the uniform distribution on $[0, 1]$

Proposition Let X be a real random variable ($X(\Omega) \subseteq \mathbb{R}$), with cumulative distribution function $F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)d\mu(u)$

- If $F(x)$ is continuous, then $U = F(X)$ is distributed according to a uniform $[0, 1]$ distribution
- Even if $F(x)$ is not continuous, the inequality $\mathbb{P}(F(X) \leqslant t) \leqslant t$ is true for all $t \in [0, 1]$
- If $F^{[-1]}(y) = \inf\{x : F(x) \geqslant y\}$ ($0 < y < 1$) and if U is distributed from a uniform $[0, 1]$ distribution, then $F^{[-1]}(U)$ is distributed according to $F(x)$

Methods involving the uniform distribution on $[0, 1]$

To perform probabilistic simulations on a computer, a pseudo-random number generator is used

Such a generator returns a sequence $(x_n)_{n \in \mathbb{N}}$ of real numbers between 0 and 1

These numbers are calculated by a deterministic algorithm but imitate a realization of a sequence of iid uniform $[0, 1]$ random variables

The good behavior of the sequence is verified by means of statistical tests

Methods involving the uniform distribution on [0, 1]

A standard method to construct the sequence $(x_n)_{n\in\mathbb{N}}$ is the congruence $x_n = y_n/N$ where the y_n are integers such that

$$y_{n+1} = (ay_n + b) \mod (N)$$

The period of the congruence generator is always smallest than $N - 1$

The choice of a, b et N is done such that

- the period of the generator is the largest as possible
- the sequence $(x_n)_{n\in\mathbb{N}}$ is as close as possible to an iid uniform $[0, 1]$ sequence

Methods involving the uniform distribution on $[0, 1]$

Proposition If $U \sim \mathscr{U}([0, 1])$ then $X = a + (b - a)U \sim \mathscr{U}([a, b])$

Proposition If $U \sim \mathscr{U}([0, 1])$ then $X = \mathbb{I}_{U \leqslant p} \sim \mathscr{B}(1, p)$

Proposition If $U_1, \ldots, U_n$ are n iid uniform random variables on $[0, 1]$, then $X = \sum_{i=1}^{n} \mathbb{I}_{U_i \leqslant p} \sim \mathscr{B}(n, p)$

It is always possible to obtain a simulation following a random variable which takes the values $(x_i)_{i \in \mathbb{N}^*}$ with respective probabilities $(p_i)_{i \in \mathbb{N}^*}$ (with $p_i \geqslant 0$ such as $\sum_{i \in \mathbb{N}^*} p_i = 1$) using a single uniform variable on $[0, 1]$

Methods involving the uniform distribution on $[0, 1]$

Proposition If $U \sim \mathscr{U}([0,1])$, then

$$X = x_1 \mathbb{I}_{U \leqslant p_1} + x_2 \mathbb{I}_{p_1 < U \leqslant p_1 + p_2} + \ldots + x_i \mathbb{I}_{\sum_{j=1}^{i-1} p_j < U \leqslant \sum_{j=1}^{i} p_j} + \ldots$$

is distributes as a random variable that takes values $(x_i)_{i \in \mathbb{N}^*}$ with associates probabilities $(p_i)_{i \in \mathbb{N}^*}$

Requires coding a loop on i with stopping rule $\sum_{j=1}^{i-1} p_j < U \leqslant \sum_{j=1}^{i} p_j \implies$ it can take a while if the sequence $(p_i)_{i \in \mathbb{N}^*}$ converges slowly to 1.

Methods involving the uniform distribution on $[0, 1]$

Proposition If U_1 and U_2 are two $\mathscr{U}([0, 1])$ independent random variables, then

$$X_1 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$$

and

$$X_2 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$$

are two independent standard Gaussian random variables

Recall that if $X \sim \mathscr{N}(0, 1)$ then $\mu + \sigma X \sim \mathscr{N}(\mu, \sigma^2)$

The accept-reject algorithm

Target distribution with pdf p on $\mathbb{R}^d$

Instrumental distribution with pdf q on $\mathbb{R}^d$

There exists $k \geqslant 1$ such that

$$\forall x \in \mathbb{R}^d, \quad p(x) \leqslant kq(x)$$

The accept-reject algorithm

0) Set i=1
1) Generate Y_i from q
2) Calculate $M = \dfrac{p(Y_i)}{kq(Y_i)}$
3) Generate $U_i \sim \mathscr{U}([0,1])$
4) If $U_i > M$, then $i = i + 1$ and back 1)
   If $U_i \leqslant M$, then $X = Y_i$

The accept-reject algorithm

Note $N = \inf\{i \geqslant 1 : kq(Y_i)U_i \leqslant p(Y_i)\}$ (N is a random variable), we have $X = Y_N$

Proposition N is distributed according to a geometric distribution with parameter $1/k$, $\mathbb{E}(N) = k$

N is independent of $(Y_N, kq(Y_N)U_N)$ which is uniformly distributed on

$$\{(x, z) \in \mathbb{R}^d \times \mathbb{R} : 0 \leqslant z \leqslant p(x)\}$$

Typically, $X = Y_N$ is distributed from p