Generating random variables II

Anatoli Iambartsev
IME-USP

Distribution of Random Variables and Simulation of Random Variables

• Random variable uniquely determined by the cumulative distribution function (cdf) $F(x)$

• There exists pseudo-random number generation for uniformly distributed on $[0, 1]$ r.v.

• Inverse method, if it is possible to calculate explicitly inverse or generalized inverse F^{-1}

• Accept-Reject method

Bernoulli Factory.

Let $X_1, X_2, \ldots$ be i.i.d. $X_i \sim B(p)$. A Bernoulli factory is an algorithm that takes (X_i) and auxiliary variables with known distributions, and simulates a Bernoulli r.v.s with success probability $f(p)$. Of course, the algorithm is not allowed to know the value p.

In [2] Assmussen raised the question of whether it was possible to construct a Bernoulli factory for $f(p) = Cp$, the application being perfect simulation for certain positive recurrent regenerative processes.

[1] Huber, M. (2016) Nearly Optimal Bernoulli Factories for Linear Functions. *Combin., Prob. and Computing*, 25, 577591.

[2] Assmussen, S, Glynn, P.W. and Thorisson, H. (1992) Stationarity detection in the initial transient problem. *ACM Trans. Model. Comput. Simul.* **2** 130157.

(Again) About uniform r.v.

Note that $U \sim U[0,1]$ can be viewed as an i.i.d. sequence of $B(1/2)$ r.v.s simply by reading off the bits in the number U. These bits can then be uses to build an i.i.d. sequence of uniform random numbers in $[0,1]$.

Lemma 1. Let $U \sim U[0,1]$, and let γ_i are uniforms on the set $S = \{0,1,\ldots,9\}$. Then

$$U = 0.\gamma_1\gamma_2\ldots.$$

Proof.

$\{\gamma_k = i\}$ iff $0.\gamma_1 \dots \gamma_{k-1} i \leq U < 0.\gamma_1 \dots \gamma_{k-1} i + 10^{-k}$

for any $\gamma_1 \dots \gamma_{k-1}$ fixed, then

$$\mathbb{P}(\gamma_k = i) = \sum_{\gamma_1,\dots,\gamma_{k-1}=0}^{9} 10^{-k} = 0.1$$

Let $1 \leq s < k$. Similarly we have

$$\mathbb{P}(\gamma_s = j, \gamma_k = i) = \sum_{\gamma_1,\dots,\gamma_{s-1},\gamma_s,\dots,\gamma_{k-1}=0}^{9} 10^{-k} = 0.01$$

Thus

$$\mathbb{P}(\gamma_{k_1} = i_1, \gamma_{k_2} = i_2, \dots, \gamma_{k_s} = i_s) = \mathbb{P}(\gamma_{k_1} = i_1) \dots \mathbb{P}(\gamma_{k_s} = i_s)$$

$\square$

(Again) About uniform r.v.

Lemma 2. Let a be some positive integer number, and let $U \sim U[0,1]$, then

$$\eta = \{aU\} \sim U[0,1],$$

where $\{x\}$ stands for fractional (noninteger) part of a positive real number $x \in \mathbb{R}_+$.

Proof. If $x \in (0,1)$ then

$$\begin{aligned}
\mathbb{P}(\eta < x) &= \sum_{k=0}^{a-1} \mathbb{P}(k \le aU < k + x) \\
&= \sum_{k=0}^{a-1} \mathbb{P}(ka^{-1} \le U < (k+x)a^{-1}) = \sum_{k=0}^{a-1} xa^{-1} = x
\end{aligned}$$

$\square$

Generation multi-dimensional r.v.s.

Let $Q = (X_1, \dots, X_n)$ be a vector with independent components, then

$$F_Q(x_1, \dots, x_n) = F_1(x_1) \dots F_n(x_n),$$

where F_i is cdf of component X_i. Here we generate vector Q simulating X_i independently.

Generation multi-dimensional r.v.s.

Example. Simulate (X, Y) uniform on the disc

$$\{(x, y) : \; x^2 + y^2 \leq 1\}.$$

Generation multi-dimensional r.v.s.

Example. Simulate (X, Y) uniform on the disc

$$\{(x, y) : \ x^2 + y^2 \leq 1\}.$$

Use polar coordinates (R, Θ). These are independent and s.t. $\Theta \sim U[0, 2\pi]$ and R has pdf $f_R(r) = 2r, r \in [0, 1]$. It gives cdf $F_R(r) = r^2, r \in [0, 1]$ and inverse $F^{-1}(r) = \sqrt{r}$. Thus, if $U, V \sim U[0, 1]$ are uniform, then

$$\Theta = 2\pi U, \ R = \sqrt{V}.$$

Generation multi-dimensional r.v.s.

$p_Q(x_1,\ldots,x_n) = p_1(x_1)p_2(x_2 \mid x_1)p_3(x_3 \mid x_1,x_2)\ldots p_n(x_n \mid x_1,\ldots,x_{n-1})$

$$
\begin{aligned}
p_1(x_1) &= \int\cdots\int p_Q dx_2\ldots dx_n \\
p_2(x_2 \mid x_1) &= [p_1(x_1)]^{-1}\int\cdots\int p_Q dx_3\ldots dx_n \\
p_3(x_3 \mid x_1,x_2) &= [p_1(x_1)p_2(x_2 \mid x_1)]^{-1}\int\cdots\int p_Q dx_4\ldots dx_n \\
\ldots &\quad \ldots \\
p_{n-1}(x_{n-1} \mid x_1,\ldots,x_{n-2}) &= [p_1(x_1)\ldots p_{n-2}(x_{n-2} \mid x_1,\ldots,x_{n-3})]^{-1}\int p_Q dx_n \\
p_n(x_n \mid x_1,\ldots,x_{n-1}) &= [p_1(x_1)\ldots p_{n-1}(x_{n-1} \mid x_1,\ldots,x_{n-2})]^{-1} p_Q
\end{aligned}
$$

Generation multi-dimensional r.v.s.

Let

$$F_i(x_i \mid x_1, \ldots, x_{i-1}) = \int_{-\infty}^{x_i} p_i(x \mid x_1, \ldots, x_{i-1}) dx$$

Lemma 3. Let $U_1, \ldots, U_n$ be i.i.d. $U[0,1]$. Then $Q = (X_1, \ldots, X_n)$ s.t.

$$X_1 = F_1^{-1}(U_1),\; X_2 = F_2^{-1}(U_2 \mid X_1), \ldots,$$
$$X_n = F_n^{-1}(U_n \mid X_1, \ldots, X_{n-1})$$

has the density $p_Q(x_1, \ldots, x_n)$.

Generation multi-dimensional r.v.s.

Example. Simulate (X, Y) uniform on the triangle T with corners in $(0,1),(0,0)$ and $(1,0)$, i.e.

$$T = \{(x,y): \ 0 \le x \le 1, \ 0 \le y \le 1-x\}.$$

The joint density

$$f(x,y) = \begin{cases} 2, & \text{if } (x,y) \in T, \\ 0, & \text{if } (x,y) \notin T. \end{cases}$$

gives us the marginal

$$f_X(x) = \int_0^{1-x} f(x,y)dy = 2\int_0^{1-x} dy = 2(1-x).$$

Generation multi-dimensional r.v.s.

Example. The marginal density

$$f_X(x) = \int_0^{1-x} f(x,y)dy = 2\int_0^{1-x} dy = 2(1-x).$$

gives us the marginal cdf

$$F_X(x) = \int_0^x 2(1-z)dz = 2x - x^2,\ x \in [0,1]$$

and inverse

$$F^{-1}(t) = 1 - \sqrt{1-t},\ t \in [0,1]$$

(note that when you solve the quadratic equation you choose only one correct solution with "$-$" within interval $[0,1]$)

Generation multi-dimensional r.v.s.

Example. Further, the conditional pdf of Y given $X = x$ is

$$f_Y(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{1-x},$$

i.e. $Y \mid X = x \sim U[0, 1-x]$.

The inverse transformation method thus gives that if U and V are independent uniform $[0,1]$, then

$$\begin{aligned} X &= 1 - \sqrt{1-U} \\ Y &= V(1-X) \end{aligned}$$

gives a pair (X,Y) which is uniform on the triangle T.

Change of variables method.

[S, p.58] Sometimes we can simplify formulas of modelling of multi-dimensional random variables by choosing new coordinate system.

The rule of density transformation under changing of variables: let $y_i = g_i(x_1, \ldots, x_n), i = 1, \ldots, n$, one-to-one differentiable transformation of an area B in the space $x_1, \ldots, x_n$ into an area B' em the space $y_1, \ldots, y_n$. Let $p_Q(x_1, \ldots, x_n)$ be the density of a random vector $Q = (\xi_1, \ldots, \xi_n)$ in the area B, then the density of the vector $Q' = (\eta_1, \ldots, \eta_2)$ in B', where $\eta_i = g_i(\xi_1, \ldots, \xi_n)$, is

$$p_{Q'}(y_1, \ldots, y_n) = p_Q(x_1, \ldots, x_n) \left| \frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)} \right|$$

in the right-hand side x_i should be expressed by $(y_i)_{i=1,\ldots,n}$.

Change of variables method. Box-Muller.

Example. (RC, Example 2.3) Generate standard normal r.v.s $X, Y \sim N(0,1)$. Consider a transformation to polar coordinates: $(x, y) \to (d, \theta)$

$$\begin{cases} d &= x^2 + y^2 \\ \theta &= \tan^{-1}\left(\frac{y}{x}\right). \end{cases}$$

To get the joint distribution of d and θ need Jacobian of the transformation

$$J = \begin{vmatrix} \frac{\partial d}{\partial x} & \frac{\partial d}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix} = \begin{vmatrix} 2x & 2y \\ \frac{1}{1+\frac{y^2}{x^2}}\left(-\frac{y}{x^2}\right) & \frac{1}{1+\frac{y^2}{x^2}}\left(\frac{1}{x}\right) \end{vmatrix} = 2$$

Change of variables method. Box-Muller.

Example. (RC, Example 2.3) Since $f(x,y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}$ then

$$f_{d,\theta}(d,\theta) = \frac{1}{2\pi}e^{-d/2} \cdot \frac{1}{2} = \frac{e^{-d/2}}{2} \cdot \frac{1}{2\pi}$$

for $0 < d < \infty$ and $0 < \theta < 2\pi$. It means that d and θ are independent. Furthermore

$$d \sim Exp(1/2),\ \ \theta \sim U[0,2\pi].$$

Thus, if $U, V \sim U[0,1]$ are independent, then the variables defined by

$$X = \sqrt{-2\ln(U)}\cos(2\pi V),\ \ Y = \sqrt{-2\ln(U)}\sin(2\pi V),$$

are independent and have standard normal distribution.

Change of variables method. Box-Muller.

[RC, p 47] " ... the Box-Muller algorithm is exact, producing two normal random variables from two uniform r.v.s, the only drawback (in speed) being the necessity of calculating transcendental functions s.t. ln, cos and sin."

Superposition method (Variant of mixture representation).

Suppose that the cdf of a r.v. that we are interested in can be represented as a composition

$$F(x) = \sum_{k=1}^{m} c_k F_k(x),$$

where all F_k's are cdf's, and $c_k > 0$. Obviously,

$$c_1 + \cdots + c_m = 1.$$

Let η be a discrete r.v. with probability distribution

$$\mathbb{P}(\eta = k) = c_k.$$

Superposition method (Compare with mixture representation).

Lemma 4. Let $U, V \sim U[0,1]$. If using V we generate a value $\eta = k$ of r.v. η, and using U, find ξ s.t. $F_k(\xi) = U$, then such generated ξ has cdf $F(x)$.

Proof.

$$\begin{aligned} \mathbb{P}(\xi \leq x) &= \sum_{k=1}^{m} \mathbb{P}(\xi \leq x \mid \eta = k)\mathbb{P}(\eta = k) \\ &= \sum_{k=1}^{m} F_k(x) c_k = F(x) \end{aligned}$$

□

Superposition method.

The generalization to infinite case is obvious.

Example. R.v. ξ is defined on the interval $[0,1]$ and has cdf

$$F(x) = \sum_{k=1}^{\infty} c_k x^k, \ c_k > 0.$$

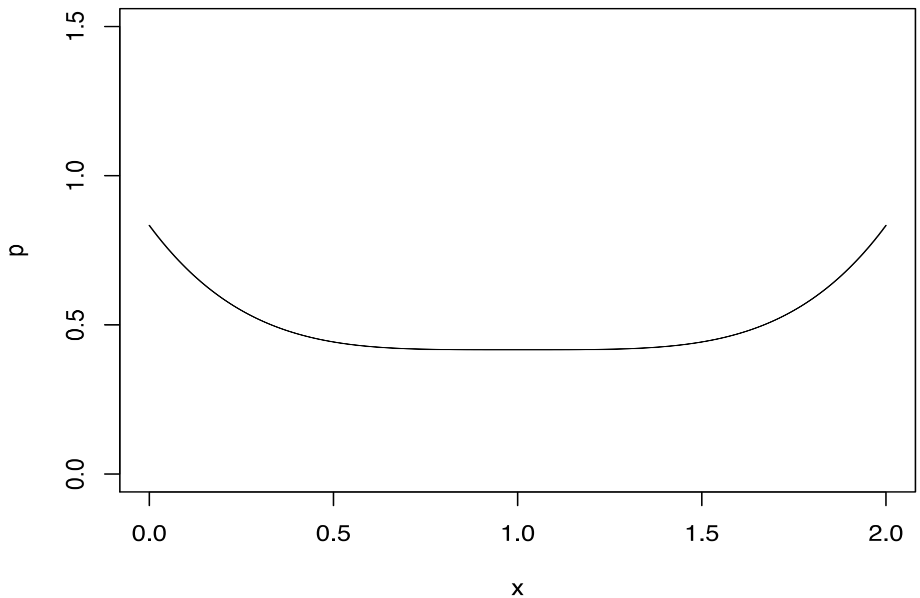
Here we can consider $F_k(x) = x^k, x \in [0,1]$, and by the method of superposition we obtain:

$$\text{if } \sum_{i=1}^{k-1} c_i \leq V < \sum_{i=1}^{k} c_i, \quad \text{then } \xi = (U)^{1/k}.$$

Superposition method.

Example. Let $\xi \in [0,2]$ with density

$$p(x) = \frac{5}{12}(1 + (x-1)^4), \; x \in [0,2].$$



Superposition method.

Example. Let $\xi \in [0,2]$ with density

$$p(x) = \frac{5}{12}(1 + (x-1)^4), \; x \in [0,2].$$

The inverse method gives the equation

$$(\xi - 1)^5 + 5\xi = 12u - 1,$$

and we should resolve the equation of 5th order.......
it is difficult

Superposition method.

Example. The density $p(x)$ can be represented

$$p(x) = \frac{5}{12}(1 + (x-1)^4) = \frac{5}{6}p_1(x) + \frac{1}{6}p_2(x), \; x \in [0,2],$$

where

$$p_1(x) = \frac{1}{2}, \; x \in [0,2], \quad p_2(x) = \frac{5}{2}(x-1)^4.$$

Then, with $U, V \sim U[0,1]$

$$\xi = \begin{cases} 2U, & \text{if } V < 5/6, \\ 1 + (2U-1)^{1/5}, & \text{if } V \geq 5/6. \end{cases}$$

Superposition method.

Lemma 4 utilizes two uniform random variables in order to simulate r.v. with distribution $F(x) = \sum_{i=1}^{m} c_k F_k(x)$. The next lemma shows that we can use only one uniform random variable.

Lemma 5.[*] Using $U \sim U[0,1]$ generate value $\eta = k$ of r.v. η, and after that define ξ from the equation $F(\xi) = \theta$, where

$$\theta = \frac{1}{c_k}\Big(U - \sum_{i=1}^{k-1} c_i\Big),$$

then the cdf of generated ξ is $F(x) = \sum_{i=1}^{m} c_k F_k(x)$.

Proof. It is enough to prove that θ is uniformly distributed on the interval $[0,1]$:

$$\mathbb{P}(\theta < y \mid \eta = k) = y).$$

□

[*]Mikhailov, G.A. *On the question of efficient algorithms for modeling of random variables.* (Russian) USSR Computational Mathematics and Mathematical Physics, 1966, **6**:6, 269273.

Superposition method.

Example. In the previous example with $U, V \sim U[0,1]$ the random variable

$$\xi = \begin{cases} 2U, & \text{if } V < 5/6, \\ 1 + (2U-1)^{1/5}, & \text{if } V \geq 5/6, \end{cases}$$

can be represented by the last lemma as

$$\xi = \begin{cases} \frac{12}{5}V, & \text{if } V < 5/6, \\ 1 + (12V-11)^{1/5}, & \text{if } V \geq 5/6. \end{cases}$$

References:

[RC] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R.* Series "Use R!". Springer

[S] Sobol, I.M. *Monte-Carlo numerical methods.* Nauka, Moscow, 1973. (In Russian)