INSTITUT FOR MATEMATISKE FAG

AALBORG UNIVERSITET

FREDRIK BAJERS VEJ 7 G ▪ 9220 AALBORG ØST
URL: www.math.auc.dk
E-mail: jm@math.aau.dk

Tlf.: 96 35 88 63
Fax: 98 15 81 29

Simulation from specific distributions — especially the normal distribution

Special algorithms exist for the most important distributions. These algorithms are faster but more complicated than the algorithms considered so far. The algorithms are often based on clever forms of rejection sampling and/or transformation methods; examples of transformation methods are given in Exercises 1 and 2 below. R provides a comprehensive set of standard distributions, see pages 37–40 in "An Introduction to R". Another place to read is

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numercal Recipes in C*, 2nd edn, Cambridge University Press, Cambridge.

This book is also available at www.nr.com.

Exercise 1 (The normal distribution)

Recall the classical and extremely important

Theorem 1 — Central limit theorem (CLT) If $Y_1, Y_2, \ldots$ are iid random variables with mean μ and variance σ^2 where $\sigma > 0$, then the empirical mean $\bar{Y}_n = \frac{1}{n}(Y_1 + \cdots + Y_n)$ converges in distribution towards the normal distribution $\mathrm{N}(\mu, \sigma^2/n)$; more precisely, for $-\infty < x < \infty$,

$$P(\sqrt{n}(\bar{Y}_n - \mu)/\sigma \le x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \qquad \text{as } n \to \infty.$$

For short we write this as $\sqrt{n}(\bar{Y}_n - \mu)/\sigma \xrightarrow{\sim} \mathrm{N}(0,1)$.

1. Assuming you have an algorithm for simulating from N(0,1), how would you make simulations from $\mathrm{N}(\mu, \sigma^2)$?

2. *The CLT can be used for simulating from N(0,1)*: Let $U_1, \ldots, U_n$ be iid with distribution unif(0,1). Show that

$$W_n \equiv ((U_1 + \cdots + U_n) - n/2)/\sqrt{n/12} \xrightarrow{\sim} \mathrm{N}(0,1).$$

   Illustrate this result in R, setting $n = 5$, simulating 100 values of W_n, and making a histogram; superimpose the density of N(0,1); make a Q-Q plot using the R-function `qqnorm`. Repeat all this for $n = 10$ and $n = 100$. Discuss the results.

The so-called *Box-Muller method is a better way of simulating from N(0,1)*, since it is based on the following exact result.

Theorem 2 If U_1 and U_2 are iid with distribution unif(0,1) and

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

then X_1 and X_2 are iid with distribution N(0,1).

In fact Theorem 2 can be verified using the following

Theorem 3 Let $\mathbf{R}$ be an invertible mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$, that is, there exists a unique mapping $\mathbf{S}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ such that $\mathbf{y} = \mathbf{R}(\mathbf{x})$ if and only if $\mathbf{x} = \mathbf{S}(\mathbf{y})$. Let $\mathbf{J}(\mathbf{x})$ be the Jacobi matrix for $\mathbf{R}$, i.e.

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} R_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} R_1(\mathbf{x}) \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial x_1} R_n(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} R_n(\mathbf{x}) \end{bmatrix}.$$

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a stochastic vector with density function $f_{\mathbf{X}}(\mathbf{x}),\ \mathbf{x} \in T$, where $T \subset \mathbb{R}^n$ is an open set. Let $\mathbf{Y}$ be the stochastic vector given by $\mathbf{Y} = \mathbf{R}(\mathbf{X})$. Then $\mathbf{Y}$ has density function

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})/|\det(\mathbf{J}(\mathbf{x}))|, \qquad \mathbf{y} \in \mathbf{R}(T),$$

where $\mathbf{x} = S(\mathbf{y})$.

Exercise 2 (The multivariate normal distribution)

The multivariate normal distribution play an extremely important role in statistical applications. One way of defining this distribution is by the distribution of an n-dimensional random vector of the form

$$X = \mu + CU$$

where $U = (U_1, \ldots, U_n)^{\mathrm{T}}$ consists of n iid N(0,1) random variables (and $^{\mathrm{T}}$ denotes transposition), $\mu \in \mathbb{R}^n$, and C is an $n \times n$ matrix. Define the mean of X by $EX = (EX_1, \ldots, EX_n)^{\mathrm{T}}$ and the covariance matrix $VarX$ of X as the $n \times n$ matrix with entry $Cov(X_i, X_j)$ for row i and column j. It can be shown that $EX = \mu$ and $VarX = CC^{\mathrm{T}}$. Set $\Sigma = CC^{\mathrm{T}}$. The distribution of X is called the *n-dimensional normal distribution with mean μ and covariance matrix* Σ, and it is denoted $\mathrm{N}_n(\mu, \Sigma)$.

1. Show that Σ is a symmetric positive semi-definite matrix, i.e. $\Sigma = \Sigma^{\mathrm{T}}$ and $x^{\mathrm{T}}\Sigma x \geq 0$ for all $x \in \mathbb{R}^n$.

2. Suppose that C is invertible. Show that Σ is invertible and Σ is positive definite, i.e. $x^{\mathrm{T}}\Sigma x = 0$ implies $x = 0$.
   Hint: *A square matrix B is invertible if and only if the only solution to the matrix equation $Bx = 0$ is the trivial solution $x = 0$.*
   Furthermore, using Theorem 3 it can be shown that X has density

   $$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right).$$

   Check that this is in accordance with the well-known result for $n = 1$.

3. *Choleski decomposition* of a symmetric positive definite square matrix Σ produces an upper-triangular matrix R such that $R^{\mathrm{T}}R = \Sigma$; the corresponding R-function is `chol`. How would you use Choleski decomposition for simulation from $\mathrm{N}_n(\mu, \Sigma)$.

4. Simulate 1000 times from $\mathrm{N}_2(\mu, \Sigma)$ when $\mu = 0$ and

   $$\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}.$$

   Produce a scatter plot of this sample.