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

Basic methods for simulation of random variables: 1. Inversion

Simulation of non-uniform random variables are often done by transforming (pseudo-random) uniform random variables. Here we consider the simplest method called *inversion*.

In the simplest case of inversion, we have a continuous random variable X with a strictly increasing distribution function F. Then F has an inverse F^{-1} defined on the open interval (0,1): for $0 < u < 1$, $F^{-1}(u)$ is the unique real number x such that $F(x) = u$. In other words,

$$F(F^{-1}(u)) = u, \qquad F^{-1}(F(x)) = x.$$

Let $U \sim \mathtt{unif}(0, 1)$ denote a uniform random variable on (0,1). Then

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

so $F^{-1}(U)$ has distribution function F. Hence we can simulate a realisation of X by simulating a realisation of $F^{-1}(U)$.

To extend this result to a general distribution function F (which is not necessarily strictly increasing), we need to introduce the *generalised inverse of* F: this is the function defined by

$$F^{-}(u) = \inf\{x : F(x) \geq u\} \qquad 0 < u < 1,$$

i.e. $F^{-}(u)$ is the u-quantile (the smallest real number x such that $F(x) \geq u$). We leave it as an exercise to show that

$$F(F^{-}(u)) \geq u, \qquad F^{-}(F(x)) \leq x, \tag{1}$$

and hence that

$$F^{-}(u) \leq x \iff F(x) \geq u. \tag{2}$$

Therefore,

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

and so we have verified the following useful result:

Theorem 1 If X is a random variable with distribution function F and $U \sim \texttt{unif}(0,1)$, then we can simulate a realisation of X by simulating a realisation of $F^{-}(U)$.

Exercise 1

Verify (1) and (2).
Hint: *Make a drawing of various distribution functions with and without jumps.*

Exercise 2

Show that if $U \sim \texttt{unif}(0,1)$ and $\lambda > 0$ is a constant, then

$$-\frac{1}{\lambda}\log(U)$$

is exponentially distributed with parameter $\lambda > 0$.

Exercise 3

Pareto distribution: If $\alpha > 0$ is a constant and X is a random variable with density

$$f(x) = \alpha x^{-(\alpha+1)}, \qquad x \geq 1,$$

then X is said to be *Pareto* distributed with parameter α (among other things, this distribution has been used for describing file transport on the Internet when using the so-called TCP protocol).

1. Show that the corresponding distribution function is given by

$$F(x) = 1 - x^{-\alpha}, \qquad x \geq 1.$$

2. Consider two distrbution functions G_1 and G_2 such that $G_1(x) = G_2(x) = 0$ for all $x < 0$. We say that G_2 has a *thicker tail* than G_1 if $G_1(x) \geq G_2(x)$ for all sufficiently large x. Discuss what this means (e.g. consider two random variables with distrbution functions G_1 and G_2).

3. Argue why F (as given above) has a thicker tail compared to any exponential distribution.

4. Show that for any real number k, $E(X^k) = \alpha/(\alpha - k)$ if $k < \alpha$, while $E(X^k) = \infty$ if $k \geq \alpha$.

5. Show that $F^{-}(u) = (1-u)^{-1/\alpha}$ and make an R-function `rpareto(n,a)` which generates n Pareto distributed random variables with parameter a.

6. Produce 10000 Pareto distributed random variables with parameter 0.5, plot their histogram (using `hist` and `boxplot`), and calculate the empirical mean (`mean`). Repeat this a number of times and study the variability of the empirical means (compare with question 4).

7. Calculate emperical means based on simulations when $n = 10^2, 10^3, 10^4, 10^5, 10^6$ and discuss the result.

Exercise 4

1. Consider a discrete random variable X with a distribution concentrated on $0, 1, \ldots$. For $j = 0, 1, \ldots$, let $p_j = P(X = j)$ and $q_j = p_0 + \cdots + p_j = P(X \leq j)$. Consider the following

   Algorithm (Inversion for discrete distributions)

   ```
   A  $j := -1$
   B  Generate $U \sim$ unif$(0,1)$
   C  Repeat $j := j + 1$ until $U < q_j$
   D  Return $j$
   ```

   Show that the output has the same distribution as X.

2. Let $X_1, X_2, \ldots$ be iid random variables with $P(X_i = 0) = 1 - p$ and $P(X_i = 1) = p$ where $0 < p < 1$ is a parameter. If r is a given positive integer, then

   $$X = \inf\{n : X_1 + \cdots + X_n = r\}$$

   can be interpreted as follows: Think of each X_i as an experiment where the event $X_i = 1$ is a success and $X_i = 0$ is a failure. Then X is the number of experiments needed in order to obtain r successes. It can be shown (you are welcome to try, but is not asked to do so) that

$$P(X = j) = \left(\begin{array}{c} j-1 \\ r-1 \end{array} \right) p^r (1-p)^{(j-r)}, \qquad j = r, r+1, \ldots$$

Here

$$\left(\begin{array}{c} n \\ m \end{array} \right) = \frac{n!}{m!(n-m)!}$$

is the binomial coefficient defined for non-negative integers m, n. Implement the inversion algorithm in R so that simulations from X are produced. Estimate the mean and variance of X when $p = 0.1, 0.5, 0.9$ and $r = 1, 10$; discuss the result.
Hint: `choose(n,m)` *can be used for calculating binomial coefficients. Read also Section 9.2.2 in "An Introduction to R".*

3. Read the help page for the R-function `sample` and discuss in which cases this can be used for simulating a discrete random variable.

Exercise 5

As a specific example of a factor that may influence the sex ratio, we consider the maternal condition *placenta previa*, an unusual condition of pregnancy in which the placenta is implemented very low in the uterus, obstructing the fetus from a normal vaginal delivery. An early study concerning the sex of placenta previa births in Germany found that of a total of 980 births, 473 were female. How much evidence does this provide for the claim that the proportion of female births in the population of placenta previa births is less than 0.485, the proportion of female births in the general population?

1. Let Θ be the probability of a female birth in the population of placenta previa births (we assume that Θ is the same for all births). Since we don't know this probability, we consider Θ as a random variable with a distribution centered around 0.485 but is flat far away from this value to admit the possibility that the truth is far away. Specificially, we let the density of Θ be

$$f_\Theta(\theta) = \begin{cases} 0.5 & \text{if } 0 < \theta \leq 0.385 \\ -20.675 + 55\theta & \text{if } 0.385 < \theta \leq 0.485 \\ 32.675 - 55\theta & \text{if } 0.485 < \theta \leq 0.585 \\ 0.5 & \text{if } 0.585 < \theta < 1. \end{cases}$$

Make a drawing of this density. Argue that 40% of the probability mass is outside the interval [0.385.0.585].

2. Conditional on $\Theta = \theta$ (where θ is a real number on the interval (0,1)) we naturally assume the following model: the sex of the $n = 980$ placenta previa births are realisations of iid random variables $X_1, \ldots, X_n$ with $P(X_i = 1|\Theta = \theta) = \theta$ where $X_i = 1$ is interpreted as "female" and $X_i = 0$ as "male". Let $X = X_1 + \cdots X_n$ denote the number of female births. We have observed $X = x$ where $x = 473$, while we are not told what the values of $X_1, \ldots, X_n$ are. As we shall see this is not a problem.

   Show that
   $$P(X_1 = x_1, \ldots, X_n = x_n|\Theta = \theta) = \theta^x(1 - \theta)^{n-x}$$
   where $x_1, \ldots, x_n$ denote the unobserved values of $X_1, \ldots, X_n$ (we only know that $x_1 + \ldots + x_n = x$).

   Consequently,
   $$P(X_1 = x_1, \ldots, X_n = x_n|\Theta = \theta) = \theta^{473}(1 - \theta)^{507}$$
   depends only on $x_1, \ldots, x_n$ through $x = 473$ and $n - x = 507$. In other words, it is sufficient to report that "of a total of 980 births, $X = 473$ were female".

3. Show that the conditional density of Θ given $X_1 = x_1, \ldots, X_n = x_n$ is specified by
   $$f(\theta|X_1 = x_1, \ldots, X_n = x_n) = \frac{1}{c(x)} f_\Theta(\theta) P(X_1 = x_1, \ldots, X_n = x_n|\Theta = \theta)$$
   where $c(x) = P(X_1 = x_1, \ldots, X_n = x_n)$ is a constant in the sense that it only depends on the data $x = 473$ and not on θ.

   Thus we can write
   $$f(\theta|X = 473) = f(\theta|X = x) = \frac{1}{c} f_\Theta(\theta)\theta^{473}(1 - \theta)^{507} \tag{3}$$
   where $c = c(473)$.

4. The density f_Θ is called the *prior density* and the conditional density (3) is called the *posterior density of* Θ *given* $X = 473$ (we return to this terminology later in the course). Calculate the posterior density in R when we approximate this density by a discrete density on the grid of θ values 0.000, 0.001, ..., 1.000.

5. Show a plot of this approximative posterior density together with the prior density.

6. Use inversion for simulating 1000 realisations from the (discrete approximation of the) posterior density (3). Show a histogram of the result and compare with the results above. Find also the posterior median (i.e. the 50% quantile in the posterior distribution), the posterior mean, and the 95% central posterior interval (this is the interval given by the 2.5% and 97.5% quantiles in the posterior distribution).
   Hint: *Here* `sample` *can be useful.*