Ordered Sample Generation

Xuebo Yu

November 20, 2010

1 Introduction

There are numerous distributional problems involving order statistics that can not be treated analytically and need to simulated through numerical way.Here we are focus on three algorithms of how to generate sample of order statistics. They are **Sort generation, Sequential generation and Simultaneous generation of order statistics.**

2 Sort Method

Sort method is a straightforward way of simulating order statistics. It is to generate a pseudo-random sample from the distribution F(x) and then sort the sample through an efficient algorithm like quick-sort. This is the general method, since once the F(x) is given, this algorithm will work. But it is the most expensive way to generate the sample to compare with other algorithms.
(note*: the best algorithm of sorting is $O(n.log(n))$

2.1 Theorem

(Probability integral transformation) *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_x(X)$. Then Y is uniformly distributed on $(0, 1)$, that is, $P(Y \leq y) = y, 0 < y < 1$.*

Proof: For $Y = F_X(X)$ we have, for $0 < y < 1$

$$\begin{aligned} P(Y \leq y) &= P(F_X(X) \leq y) \\ &= P(F_X^{-1}([F_X(X)]) \leq F_X^{-1}(y)) \\ &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned}$$

where $F_X^{-1}(u) = inf\{x : F_X \geq y\}$. Which show that Y has a uniform distribution.

2.2 Pseudo-Code

```
for(i=0; i<Samplesize; i++)
   Sample[i] = random(0,1);
QuickSort(Sample[]);
```

InversCDF(Sample[]);

Example:For standard exponential distribution
$F_Y(y) = 1 - exp(-y) \Rightarrow F_Y^{-1}(U) = -ln(1-U)$, is the function used in InversCDF.

3 Sequential generation of Random order statistics

The complexity of Sort algorithm is $O(n.ln(n))$. The sort function consumed most cpu time. So to avoid sort is the main idea to improve the algorithm above. Sequential generation of Random order statistics is one of them, which generate ordered sample of uniform without sort.

3.1 Theorem

The cdf **of uniform** distribution on interval $(0,1)$ is

$$U(x^*) = Pr(X^* \leq x^*) = \begin{cases} 0 & \text{if } x^* < 0, \\ x^* & \text{if } 0 \leq x^* \leq 1, \\ 1 & \text{if } x^* > 1. \end{cases}$$

Theorem 3.1: Given a distribution with cdf $F(x)$ then the cdf of first order statistics is: $F_{(1:n)}(x) = 1 - \{1 - F(x)\}^n$

Theorem 3.2: Let $X_1, X_2, ..., X_n$ be a random sample from an absolutely continuous population with cdf $F(X)$ and density function $f(x)$, and let $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ denote the order statistics obtained from this sample. Then the conditional distribution of $X_{j:n}$, given that $X_{i:n} = x_i$ for $i < j$, is the same as the distribution of the $(j-1)th$ order statistic obtained from a sample of size $n-i$ from a population whose distribution is simply $F(x)$ **truncated** on the left at x_i.
Proof: The conditional density function of $X_{j:n}$ given that $X_{i:n} = x_i$ is

$$\begin{aligned} f_{j:n}(x_j|X_{i:n} = x_i) &= f_{i,j:n}(x_i, x_j)/f_{i:n}(x_i) \\ &= \frac{(n-i)!}{(j-i-1)!(n-j)!}\left\{\frac{F(x_j)-F(x_i)}{1-F(x_i)}\right\}^{j-i-1}\left\{\frac{1-F(x_j)}{1-F(x_i)}\right\}^{n-j}\frac{f(x_j)}{1-F(x_i)} \end{aligned}$$

where $i < j \leq n$, $x_i \leq x_j < \infty$. By realizing that $F(x_j) - F(x_i)/1 - F(x_i)$ and $f(x_j)/1 - F(x_j)$ are the cdf and pdf of the population whose distribution is obtained by truncating the distribution $F(x)$ on the left at x_i. #

By theorem 3.1 we get the cdf of $u_{1:n}$ is given by

$$U_{(1)}(x) = 1 - \{1 - U(x)\}^n$$

From which we get its inverse function

$$U_{(1)}^{-1}(u) = 1 - \{1 - U(u)\}^{1/n}$$

which give us: $u_{(1)} = 1 - (1 - V_1)^{1/n} \quad \Leftrightarrow \quad u_{(1)} = 1 - V_1^{1/n} \qquad (1)$
Shows that $u_{(1)}$ could be generated by the help of the sample of uniform distribution

V. In our case, with $u_{(i)} = u_i$ is given, theorem 3.2 tell us u_{i+1} is the first order statistics of original truncated on the left at u_i. That is

$$u_{i+1:n} = u_{1:n-i} = 1 - (1 - u_i)V_{i+1}^{\frac{1}{n-i}} \qquad (2)$$

Equations (1) and (2) form the basis of an n-step loop subroutine which will generate the u_i sequentially.

3.2 Pseudo-Code

```
temp=0;
for(i=1;i<=n;i++){
   v=random(0,1)
   u=1-(1-temp)*v**(1/(n-i+1))
   temp = u;
   sample[i] = u;
}
InverseCDF(sample[]);
```

4 Simultaneous generation of order statistics for a multiplicity of sample sizes.

This method makes use of the fact that if $X_1, X_2, ...X_{n+1}$ are independent standard exponential random variables, then

$$\frac{X_1}{\sum_{i=1}^{n+1} X_i}, \frac{X_2}{\sum_{i=1}^{n+1} X_i}, ..., \frac{X_n}{\sum_{i=1}^{n+1} X_i}$$

are distributed as $U_1, U_2 - U_1, ..., U_n - U_{n-1}$. This is also known as the properties of uniform spacings. The following is the proof.

Definition: Let $U_1, ...U_n$ be iid uniform [0,1] random variables with order statistics $U_{(1)} \leq U_{(2)} \leq, ..., \leq U_{(n)}$. The statistics S_i defined by

$$S_i = U_{(i)} - U_{(i-1)} \qquad (1 \leq i \leq n+1)$$

where by convention $U_{(0)} = 0, U_{(n+1)} = 1$, are called the **uniform spacings** for this sample.

Theorem 3.1 $(S_1, ..., S_n)$ is uniformly distributed over the simplex

$$A_n = \{(x_1, ..., x_n) : x_i \geq 0, \sum_{i=1}^{n} x_i \leq 1\}.$$

Proof: We know that $U_{(1)}, ..., U_{(n)}$ is uniformly distributed over the simplex

$$B_n = \{(x_1, ..., x_n) : 0 \leq x_1 \leq ... \leq x_n \leq 1\}.$$

The transformation

$$\begin{array}{rclcrcl} s_1 &=& u_1 & & u_1 &=& s_1 \\ s_2 &=& u_2 - u_1 & & u_2 &=& s_1 + s_2 \\ \ldots & & & \Longrightarrow & \ldots & & \\ s_n &=& u_n - u_{n-1} & & u_n &=& s_1 + s_2 + \ldots + s_n \end{array}$$

and the determinant of the Jacobian of the transformation is 1. Shows that the density of $S_1, ..., S_2$ is uniformly distributed on the set A_n. #

Theorem 3.2 For any sequence of nonnegative numbers $x_1, ..., x_{n+1}$, we have

$$P(S_1 > x_1, ..., S_{n+1} > x_{n+1}) = (1 - \sum_{i=1}^{n+1} x_i)^n$$

Proof: By Theorem 3.1 the fact that $S_1, ..., S_n$ is uniformly distributed in A_n. Thus, the probability is equal to

$$P(S_1 > x_1, ..., S_n > x_n, 1 - \sum_{i=1}^{n} S_i > x_{n+1})$$

This is the probability of a set A_n^* which is a simplex just as A_n except that its top is not at $(0, 0, ..., 0)$ but rather at $(x_1, ..., x_n)$, and that its sides are not of length 1 but rather of length $1 - \sum_{i=1}^{n+1} x_i$. For uniform distributions, probabilities can be calculated as ratios of areas. In this case, we have

$$\frac{\int_{A_n^*} dx}{\int_{A_n} dx} = (1 - \sum_{i=1}^{n+1} x_i)^n \qquad \#$$

Theorem 3.3 $S_1, ..., S_{n+1}$ is distributed as

$$\frac{E_1}{\sum_{i=1}^{n+1} E_i}, ..., \frac{E_{n+1}}{\sum_{i=1}^{n+1} E_i}$$

where $E_1, ..., E_{n+1}$ is a sequence of iid exponential random variables.
Proof: Let $G = G_{n+1}$ be the random variable $\sum_{i=1}^{n+1} E_i$. We need to show that $E_1/G, ..., E_n/G$ is uniformly distributed in A_n. The last component E_{n+1}/G is taken care of by noting that it equals 1 minus the sum of the first n compnents. Let us use the symbols e_i, y, x_i for the running variabies corresponding to $E_i, G, E_i/G$. We first compute the joint density of $E_1, ..., E_n, G$:

$$f(e_1, ..., e_n, y) = \prod_{i=1}^{n} e^{-e_i} e^{-(y - e_i - ... - e_n)} = e^{-y}$$

valid when $e_i \geq 0$, and $y \geq \sum_{i=1}^{n} e_i$. Here we used the fact that the joint density is the product of the density of the first n variables and the density of G given $E_1 = e_1, ..., E_n = e_n$. Next, by a simple transformation of variables, it is easily seen that the joint density of $E_1/G, ..., E_n/G, G$ is

$$y^n f(x_1 y, ..., x_n y, y) = y^n e^{-y} \qquad (x_i y \geq 0, \sum_{i=1}^{n} x_i y \leq y)$$

This is easily obtained by the transformation $x_1 = \frac{e_1}{y}, ..., x_n = \frac{e_n}{y}, y = y$. Finally, the marginal density of $E_1/G, ..., E_n/G$ is obtained by integrating the last density with respect to dy, which gives us

$$\int_0^{\inf} y^n e^{-y} dy I_{A_n}(x_1, ..., x_n) = n! I_{A_n}(x_1, ..., x_n) \qquad \#$$

5 Reference

References

[1] D.LURIE and H.O. HARTLEY: *Machine-Generation of Order Statistics for Monte Carlo Computations*,

[2] BARRY C.ARNOLD: *A First Course in Order Statistics* ISBN:0-471-57416-3

[3] Luc Devroye *Non-Uniform Random Variate Generation* ISBN:3-540-96305-7