Chapter 2. Random Numbers: Generation and Tests

1 Introduction

In the first chapter, we just present an example of solving deterministic problem by formulating an equivalent problem involving a stochastic process. Namely, evaluate the nonstochastic integral $\int_0^1 g(x)dx$ by viewing it as $Eg(X)$ where X has a uniform distribution over the interval $(0,1)$. For this application, a mechanism for generating random processes is required. The techniques which can be implemented practically on a digital computer are deterministic; hence at best the process generated will be pseudostochastic. Therefore, many authors refer to this mechanism as a *pseudo random number generator.*

Analysis of the pseudo random number generator requires certain facts from number theory. We will just state these results without proof. The most basic random process we considered is to simulate a uniform $(0,1)$ distribution. If X is a continuous random variable with cumulative distribution function (cdf) $F(x)$, it is easy to show that the random variable Y given by $Y = F(X)$ has a uniform $(0,1)$ distribution. This simple fact provides us a general technique for generation of random samples from any other distribution by transformations. The technique is generate Y from $U(0,1)$ and take $X = F^{-1}(Y)$. As an example, we consider how to simulate sampling from the exponential distribution with density function $f(x) = \theta^{-1}\exp(-x/\theta)$ for $x \geq 0$ and $\theta > 0$. Letting

$$Y = \int_0^X \theta^{-1}\exp(-t/\theta)dt = 1 - \exp(-X/\theta)$$

we have $Y \sim U(0,1)$ and $X = -\theta\ln(1-Y)$ or equivalently $X = -\theta\ln Y$.

For the exponential distribution, the transformation from the uniform is simple and can be made exactly; in other cases the more complicated transformations must be approximated. This technique can also be used to generate deviates from discrete distributions using $U(0,1)$ deviates. Suppose the random variable X takes with nonzero probability only the values $x_1 < x_2 < x_3 < \cdots$, and $F(x)$ is the cdf of X. This distribution can be sampled by taking $X = x_i$, where $F(x_{i-1}) < Y < F(x_i)$ with the arbitrary value $x_0 < x_1$, and $Y \sim U(0,1)$. As an example, consider taking sample from the discrete uniform distribution defined by $Prob(X = k) = n^{-1}$ for $k = 1, 2, 3, \ldots, n$. This can be done by generating Y from $U(0,1)$ and taking X as $[nY] + 1$. ($[\cdot]$ denotes greatest integer function.)

Many univariate distributions can be generated easily in S-PLUS. They include beta, Cauchy, chisquared, exponential, F, gamma, log-normal, logistic, normal, normal range, stable, T, uniform, Weibull, binomial, geometric, hypergeometric, negative binomial, Poisson, and Wilcoxon. In fact, we can also generate 100 (say) samples from the contaminated normal distribution in which a sample is from $N(0,1)$ with probability 0.95 and otherwise from $N(0,9)$, by

$$contam < -rnorm(1, 0, (1 + 2 * rbinom(100, 1, 0.05))).$$

Now we present the simple fact that provides us a general technique for generation of random samples from any other distribution by transformations.

Theorem 1 *Let U have a uniform distribution on $[0,1]$ and let Y be any random variable with distribution function $F(y)$. Then there is a function ϕ such that Y and $\phi(U)$ have the same distribution.*

Proof. Suppose at first that F is continuous and strictly increasing. Then F is a one-to-one mapping of R onto $(0,1)$. Let $\phi(u) = \min\{y : F(y) \geq u\}$ which is the inverse mapping. Then ϕ is increasing and

$$P\{\phi(U) \leq x\} = P\{\min\{y : F(y) \geq u\} \leq x\} = P(F(x) \geq u) = F(x).$$

If F has discontinuities or is not strictly increasing, define

$$\phi(u) = \inf\{y : F(y) \geq u\}$$

for $0 < u < 1$. Since F is nondecreasing, $\{y : F(y) \geq u\}$ is an interval stretching to ∞; since F is right-continuous, this interval is closed on the left. For $0 < u < 1$, therefore, $\{y : F(y) \geq u\} = [\phi(u), \infty]$, and so $\phi(u) \leq x$ if and only if $u \leq F(x)$. The rest follows by the same reasoning.

Based on the above theorem, we now have a general technique for simulating variates from a given distribution if we can generate independent variates from $U(0,1)$. In this note, we will discuss the rejection method proposed by von Neumann (1951) in addition.

2 General Techniques for Generation of Nonuniform Random Deviates

2.1 Use of the Cumulative Distribution Function

Suppose X is a continuous random variable with cumulative distribution function (cdf) $F(x)$. We further assume that F can be inverted easily.

- the exponential distribution with parameter θ:

  Generate $u_1, \ldots, u_n$ from $U(0,1)$.

  Let $x_i = -\theta \ln u_i$.

- Weibull distribution with parameter α, β:

  Note that $P(X \geq x) = \exp[-(x/\beta)^\alpha]$ and $P(\beta Y^{1/\alpha} \geq x) = \exp[-(x/\beta)^\alpha]$ if Y is exponential distributed with parameter 1.

  Generate $u_1, \ldots, u_n$ from $U(0,1)$.

  Let $x_i = \beta(-\ln u_i)^{1/\alpha}$.

- The Gamma Distribution: $f(x) = x^{\alpha-1}\exp(-x)/\Gamma(\alpha)$

  Note that $G(1)$ is the exponential distribution with parameter 1 and $X+Y \sim G(\alpha+\beta)$ if $X \sim G(\alpha)$, $Y \sim G(\beta)$, and X and Y are independent. We then can generate $G(n)$ easily when n is an integer by the summation method. Namely,

  Generate $u_1, \ldots, u_n$ from $EXP(1)$.

  Let $x = \sum_{i=1}^{n} u_i$.

  How do we generate $G(n)$ if n is not an integer? Its discussion will be deferred to next section.

- The Chi-Square Distribution:

  Note that $\chi^2_J/2 = G(J/2)$ and $G(J+0.5) = G(J) + G(0.5)$.

  How do we generate $G(0.5)$?

Suppose the random variable X takes with nonzero probability only the values $x_1 < x_2 < x_3 < \cdots$, and $F(x)$ is the cdf of X. This distribution can be sampled by taking $X = x_i$, where $F(x_{i-1}) < Y < F(x_i)$ with the arbitrary value $x_0 < x_1$, and $Y \sim U(0,1)$. Here we consider three discrete distributions: Binomial, Multinomial, and Poisson distributions.

- The Binomial distribution with n trials and success probability p:

  Generate $u_1, \ldots, u_n$ from $U(0,1)$.

  Let y be the number of times $u_i \leq p$.

- The Multinomial distribution with n trials and J possible outcomes with probabilities $p_1, \ldots, p_J$:

  Generate $u_1, \ldots, u_n$ from $U(0,1)$.

  Count how many times falling in the intervals $(0, p_1], (p_1, p_1 + p_2], \ldots, (\sum_{i=1}^{J-1} p_j, 1]$.

- The Poisson Distribution with parameter λ is defined by the probability function

$$f(x) = \exp(-\lambda)\frac{\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots.$$

  Facts: Let $X_1, X_2, \ldots$ be independent exponential distributed random variables with parameter α. Set $S_n = X_1 + \cdots + X_n$ and $N_t = \max\{n : S_n \leq t\}$. (As a motivation, consider a stream or sequence of events, say arrivals of calls at an exchange. Let X_1 as the waiting time to the first event, let X_2 be the waiting time between the first and second events, and so on. Then N_t is the number of events that occur in the time interval $[0, t]$ and $S_1, S_2, \ldots$ are called arrival times.) Note that S_n has the distribution function specified by $\exp(-\alpha x)\sum_{i=n}^{\infty}(\alpha x)^i/i!$. Then $P(N_t \geq n) = \exp(-\alpha t)\sum_{i=n}^{\infty}(\alpha t)^i/i!$ and $P(N_t = n) = \exp(-\alpha t)(\alpha t)^n/n!$, $n = 0, 1, \ldots$. Thus N_t has the Poisson distribution with parameter αt.

  Use this fact, we have

$$P(m \text{ points in } [0,1]) = P(m+1 \text{ is the first time } X_1 + \cdots + X_{m+1} \geq 1) = \exp(-\lambda)\frac{\lambda^m}{m!}$$

  if X_i is exponentially distributed with mean $1/\lambda$. Now we describe the algorithm.

  1. Set $s = 0$.
  2. Generate u from $U(0,1)$ and set $s = s - \log u/\lambda$.
  3. Output $m - 1$ if $s \geq 1$; else go to step 2.

2.2 Rejection Methods

One of the most widely used techniques for generating random numbers from a given distribution was described by von Neumann (1951). It is variously referred to as the *rejection method* or the *acceptance sampling method.*

The method can be described as follows. Suppose it is desired to generate random numbers from the distribution having pdf $f(x)$. Let $h(x)$ and $k(x)$ be pdf's and let $\ell(x)$ be a given function (not necessarily a pdf).

1. Generate X from the distribution having pdf $h(x)$.
2. Generate independently Y from the distribution having pdf $k(y)$.
3. If $Y \leq \ell(X)$, deliver $Z = X$; else go to 1.

We now derive the cdf of Z. Recall that we generate sample from a sequence of independent pairs of random variables $(X_1, Y_1), (X_2, Y_2), \ldots$ having distributions $h(x)$ and $k(y)$, respectively. Set $Z = X_{K*}$ where $K* = \min\{K : Y_K \leq ell(X_K)$. Note that

$$\begin{aligned}
P(Z \leq z) &= P(X_{K*} \leq z) = \sum_{K=1}^{\infty} P(X_K \leq z, K* = K), \\
\{K* = K\} &= \{Z_K < 0, Z_{K-1} \geq 0, \cdots, Z_1 \geq 0\}, \\
P(X_K \leq z, K* = K) &= P(X_K \leq z, Z_K < 0)[P(Z_1 \geq 0)]^{K-1}, \\
P(Z \leq z) &= P(X_1 \leq z, Z_1 < 0) \sum_{K=1}^{\infty} [P(Z_1 \geq 0)]^{K-1} = P(X \leq z | Y \leq \ell(X)).
\end{aligned}$$

Hence,

$$\begin{aligned}
P(Z \leq z) &= P(X \leq z | Y \leq \ell(X)) \\
&= \frac{\int_{-\infty}^{z} \int_{-\infty}^{\ell(x)} h(x)k(y)dydx}{\int_{-\infty}^{\infty} \int_{-\infty}^{\ell(x)} h(x)k(y)dydx}.
\end{aligned}$$

Differentiation with respect to z of this cdf yields the pdf of Z as given.

$$f_Z(z) = \frac{h(z)G((\ell(z))}{\int_{-\infty}^{\infty} h(x)G(\ell(x))dx}.$$

It is of interest to know something about the efficiency of the rejection method. The time required to generate a variate by this method is a random variable. Let T be the number of executions of steps 1 through 3 before exit. Then T has the geometric distribution with

$$P(T = t) = \tau(1 - \tau)^{t-1} \quad t = 1, 2, \ldots,$$

where $\tau = \int_{-\infty}^{\infty} h(x)G(\ell(x))dx$. Hence $E(T) = \tau^{-1}$ and $Var(T) = (1 - \tau)/\tau^2$.

The foregoing description of the rejection method does not indicate the simplicity of its application, since the functions $h(x)$, $k(y)$ and $\ell(x)$ required may not be immediately apparent for a given distribution. We therefore will now give an alternate description that

begins with the pdf of the desired distribution and indicates the way in which other needed function can be obtained. In step 2, the function $k(y)$ is taken as the $U(0,1)$ pdf, that is, the constant 1. Pick a pdf $h(x)$ such that $f(x) \le ch(x)$ for some constant c. (In practice, we usually start with picking up a function $M(x)$ such that $M(x) \ge f(x)$. Set $c = \int_{-\infty}^{\infty} M(x)dx$. Then $h(x) = M(x)/c$.) Set $\ell(x) = f(x)/ch(x)$. By such a choice, we have

$$f_Z(z) = \frac{h(z)\ell(z)}{\int_{-\infty}^{\infty} h(x)\ell(x)dx} = f(z).$$

Now we consider the generation of deviates with a gamma distribution with shape parameter $0 < \alpha < 1$. Recall that $f(x) = x^{\alpha-1}\exp(-x)/\Gamma(\alpha)$. Note that

$$f(x) \le \left\{ \begin{array}{ll} \frac{x^{\alpha-1}}{\Gamma(\alpha)} & 0 \le x \le 1 \\ \frac{\exp(-x)}{\Gamma(\alpha)} & 1 \ge x \end{array} \right\} = ch(x),$$

where

$$c = \frac{1}{\Gamma(\alpha)}\int_0^1 x^{\alpha-1}dx + \frac{1}{\Gamma(\alpha)}\int_1^{\infty} e^{-x}dx = \frac{1}{\alpha\Gamma(\alpha)} + \frac{\exp(-1)}{\Gamma(\alpha)}.$$

Set $c\Gamma(\alpha) = \alpha^{-1} + e^{-1} = \beta$. Next we evaluate $H(x)$ and solve for $y = H^{-1}(u)$. Observe that

$$H(y) = \int_0^y h(x)dx = \left\{ \begin{array}{ll} \frac{y^{\alpha}}{\alpha\beta} & y \le 1 \\ 1 - \frac{1}{\beta}\exp(-y) & y \ge 1 \end{array} \right.$$

The algorithm is as follows:

1. Generate u_{2k-1}, u_{2k} from $U(0,1)$.
2. If $u_{2k-1} \le (\alpha\beta)^{-1}$ and $u_{2k} \ge \exp[-(\alpha\beta u_{2k-1})^{1/\alpha}]$, set $k = k+1$ and go to (1), else set $x = (\alpha\beta u_{2k-1})^{1/\alpha}$.
3. If $u_{2k-1} \ge (\alpha\beta)^{-1}$ and $u_{2k} \ge [-\log\beta(1-u_{2k-1})]^{\alpha-1}$, set $k = k+1$ and go to (1), else set $x = -\log\beta(1-u_{2k-1})]$.

2.3 Generation of Order Statistics

If we have a sample $x_1, \cdots, x_n$ from some distribution $F(x)$, order the sample in increasing order and relabel as $x_{(1)} \le \cdots \le x_{(n)}$. These are called the order statistics. Frequently in applications order statistics from a given distribution are needed. If the inverse cdf technique is appropriate for generating variates from the given distribution, the problem can be solved by first obtaining the order statistics from the uniform distribution. If other techniques are

used to transform the uniforms to variates from the given distribution, it may be necessary to obtain the full sample from the given distribution and then sort it to obtain the order statistics. We discuss below some methods for obtaining the order statistics from a uniform distribution.

Facts:

- The ith order statistic, $U_{(i)}$, in a sample of size n from $U(0,1)$ has a beta distribution with parameters i and $n-i+1$.
- If $X_1, X_2, \ldots, X_n$ are independently distributed as exponential (1), the ratio

$$V_i = \frac{X_1 + \cdots + X_j}{X_1 + \cdots + X_{n+1}}$$

  has the same distribution as $U_{(i)} - U_{(i-1)}$ where $U_{(0)} = 0$.

If we use the second fact, we can generate order statistics from uniform distribution easily as follows.

- Generate $n+1$ uniform variates $u_1, \ldots, u_{(n+1)}$.
- Put $\pi_j = u_1 \cdot u_1 \cdot \cdots \cdot u_j$.
- Let $V_j = \log \pi_j / \log \pi_{j+1}$.

For general distribution, we just do the inverse transformation $x_{(j)} = F^{-1}(u_{(j)})$.

2.4 Generation of Variates from Normal related Distributions

Probably the most often used nonuniform distributions are those of the normal family with mean μ and variance σ^2, which we denote as $N(\mu, \sigma^2)$. A random variable from the standard normal distribution, $N(0,1)$, can easily be transformed so as to have the $N(\mu, \sigma^2)$ distribution; hence, we only consider generation of variates from the standard normal.

We now introduce the method proposed in Box and Muller (1958). It is a direct transformation of two independent $U(0,1)$ variates, U_1 and U_2, to two independent $N(0,1)$ variates, X_1 and X_2:

$$\begin{aligned} X_1 &= [-2\ln(U_1)]^{1/2} \cos(2\pi U_2), \\ X_2 &= [-2\ln(U_1)]^{1/2} \sin(2\pi U_2). \end{aligned}$$

Suppose X_1 and X_2 are two independent $N(0,1)$ variates. Their joint density function is

$$f_{X_1,X_2}(x,y) = (2\pi)^{-1}\exp(-(x^2+y^2)/2).$$

Let $R = \sqrt{X_1^2 + X_2^2}$ and $\theta = arctan(X_1/X_2)$. Then θ is uniform on $[0, 2\pi]$ and R and θ are independent. Note that

$$\begin{aligned} P(R \le r) &= \int_{x^2+y^2 \le r^2} \frac{1}{2\pi}\exp(-\frac{x^2+y^2}{2})dxdy \\ &= \int_0^r \int_0^{2\pi} \frac{1}{2\pi}\exp(-\xi^2/2)\xi d\xi d\theta = 1 - \exp(-r^2/2). \end{aligned}$$

Consequently, let R and θ be independent where θ is uniform on $[0, 2\pi]$ and R and $P(R \le r) = 1 - \exp(-r^2/2)$. Then $R\sin\theta$ and $R\cos\theta$ are two independent $N(0,1)$ variables.

The p-variate normal distribution with mean μ and variance-covariance matrix V (non-singular) is defined by the pdf

$$f(\mathbf{x}) = (2\pi)^{-p/2}|V|^{-1/2}\exp\left\{\frac{1}{2}(\mathbf{x}-\mu)^t V^{-1}(\mathbf{x}-\mu)\right\}.$$

We denote this distribution $N_p(\mu, V)$. An obvious method to generate a p vector x with this distribution is to generate $z_1, z_2, \ldots, z_p$ as independent $N(0,1)$ variates, forming the p vector $\mathbf{z}$ having $N_p(\mathbf{0}, I)$ then form $\mathbf{x}$ as

$$\mathbf{x} = A\mathbf{z} + \mu$$

where A is such that $AA^t = V$, for example, A could be from a Cholesky decomposition of V. In Splus, the function *chol* returns the Choleski decomposition $B = U^tU$ of a nonnegative definite symmetric matrix. (Note that this is an unusual convention; more commonly the lower-triangular form $B = LL^t$ with $L = U^t$ is used.)

We group the F, T, and chi-square distributions together in this subsection because they are frequently associated in statistical applications and they are associated with normal variate. The chi-square distributions can be generated by the sum of n square $N(0,1)$ variates. A Snedecor F random deviate may be obtained as a transformation of two chi-squares or as a transformation of one beta. In particular, let Z have a Beta distribution with parameters $m/2$ and $n/2$. Then $(n/m)Z/(1-Z)$ has an $F_{m,n}$ distribution. Note that $t_n = \sqrt{n}Z/Y$ where Z and Y are independent, Z is a standard normal variate and Y is the square root of a chi-square variate. Hence, T distribution can be generated easily accordingly.

3 Pseudo Random Number Generator

There are at present two basic techniques for generating numbers from $U(0,1)$: the congruential methods and the feedback shift register methods. We first describe the generator adopted in S-PLUS. This generator is based on George Marsaglia's Super-Duper package from about 1973. It couples a multiplicative-congruential generator with a feedback shift register. This generator produces a 32-bit integer whose top 31 bits are divided by 2^{31} to produce a real number in $[0,1)$. The 32-bit integer is produced by a bitwise exclusive-or of two 32-bit integers produced by separate generators. The C code for the S random number generator is

unsigned long congrval, tausval

Static double xuni()

{

unsigned long n, lambda = 69069;

congrval = congrval*lambda;

tausval $=tausval >> 15$;

$tausval=tausval << 17$;

$n = tausval^congrval; return((((n >> 1)\&017777777777))/2147483648.);$

}

$The integer congrval follows the congruential generator X_{i+1} = 69069 X_i (mod\ 2^{32}) as unsigned integer overflow ar$

that is, the overflowing bits are discarded. As this is a multiplicative generator (no additive constant), its period is 2^{30} and the bottom bit must always be odd (including for the seed).

The integer tausval follows a 32-bit Tausworthe generator (of period $4,292,868,097 < 2^{32}-1$):

$$b_i = b_{i-p} \text{ xor } b_{i-(p-q)}, \quad p = 32, \quad q = 15.$$

This follows from a slight modification of Algorithm 2.1 of Ripley (1987, p. 28). (In fact, the period quoted is true only for the vast majority of starting values; for the remaining $2,099,198$ non-zero initial values there are shorter cycles.)

For most starting seeds the period of the generator is $2^{30} \times 4,292,868,097 \approx 6.6 \times 10^{14}$, which is quite sufficient for calculation which can be done in a reasonable time in S. The current values of congrval and tausval are envoded in the vector .Random.seed, a vector of

12 integers in the range $0, \ldots, 63$. If x represents .Random.seed, we have

$$congrval = \sum_{i=1}^{6} x_i 2^{6(i-1)} \quad \text{and} \quad tausval = \sum_{i=1}^{6} x_{i+6} 2^{6(i-1)}.$$

Additional materials will be presented next.

3.1 Congruential Methods

A very useful source of pseudorandom integers is a linear congruential sequence. It is defined by

$$X_{i+1} = (\alpha X_i + c)(mod\ m) \quad i = 1, 2, 3, \ldots$$

where X_i, α, c, and m are integers and $0 \leq X_i < m$. When $c = 0$, the method is referred to as multiplicative-congruential; otherwise, to emphasize the presence of an additive constant, it is referred to as mix-congruential.

The integers produced by X_i are in the interval $[0, m)$. They are transformed by X_i/m into $[0, 1)$ over which they approximate a $U(0, 1)$ process. For some j and k, it must be the case that $X_j = X_k$; and it is obvious from the relationship that the subsequence produced beginning with X_j must be the same as that beginning with X_k. For a given generator and any given X_0, called a seed, the length of the shortest subsequence which by repetition forms the entire sequence from some point on is called the *period* of the seed for that generator.

generator The bits are thought of as being numbered left to right: 0 through 31. The bit numbered 0 designates the sign of the number: 0 is positive, and 1 negative. In the case of positive integers, the remaining bits together form the binary representation of the integer.

3.2 Feedback Shift Register Methods

Thm: FSR has period $2^p - 1$ if $Q(x)$ is an irreducible polynomial mod(2). (i.e., We cannot factor $Q(x) = Q_1(x)Q_2(x)(mod\ 2)$ where coefficients Q_1, Q_2 are 0 or 1.

3.3 Randomizing Random Number Generators

4 Tests of Random Number Generators

There are two approaches: Structural and Empirical. Structural Analysis:

Graph successive pairs

- Test for uniformity:
- Extreme value test:
- Testing for independence: