Pramod Parajuli
Simulation and Modeling, CS-331

Chapter 7

- Uniform Random Generators
- Testing of Uniform Random Generators
- Methods of Generating Non-Uniform Variables
- Inversion, Rejection, Composition

Introduction

Random numbers

R1, R2, R3, ...

Must have two essential properties;

- Uniformity
- Independence

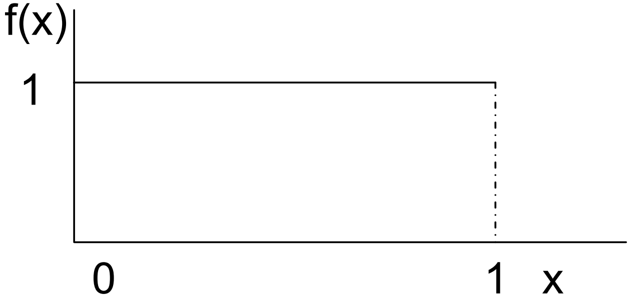
Continuous random variable – Random variable X is continuous if its sample space is a range or collection of ranges of real values

Continuous Uniformly Distributed RNs

Each random number R_i is an independent sample from a continuous uniform distribution between 0 and 1

Therefore, PDF = $f(x) = \begin{cases} 1 \;\; ,0 \le x \le 1 \\ 0, otherwise \end{cases}$

Expected value, E(R) $= \int_0^1 x dx$



$$= \left| \frac{x^2}{2} \right|_0^1$$

$$= \frac{1}{2}$$

Variance

$$V(R) = \int_0^1 x^2 dx - [E(R)]^2 = 1/12$$

Generation of RNs – RNGs (RN Generators)

-Should be fast

-Should be portable

-Should have sufficiently long cycle

-Should be replicable

-Should hold ideal statistical properties (uniformity and independence)

-Linear Congruent Generator

Linear Congruent Generator (LCG)

-Produces integers between 0 and m-1

$$x_{i+1} = (a.x_i + c) \bmod m$$

$$i = 0, 1, 2, . .$$

Initial value i.e. x_0 is called seed

a = constant multiplier

c = constant increment

Example;

$m = 100, x_0 = 27, a = 17, c = 43$

For random number between 0 and 1,

$= R_i/m$

Multiplicative Congruent Generator (MCG)

-If c = 0

-Example;

A = 13, m = 2^6 = 64, x_0 = 1, 2, 3, and 4

Look for the period/cycle length

Maximum period is 64/4 = 16

Here, 4 is the gap between the elements

Guidelines for selecting the values

Choose ‘m’ be one more than largest integer that can be stored in one word of the computer being used. Problem, how to represent such a number? Therefore, choose ($2^{maxword-1}$ – 1) i.e. in 32 bit number system, choose (2^{31}-1)

The seed x_0 must be relatively prime to m

The constant multiplier ‘a’ must be selected properly (prime). Normally, it is selected as;

$$a = k.8 + 3$$

or $a = k.8 + 5$

or $a = 2^{max\ wordsize/2} + 3$

or $a = 65,539$

Testing Uniform Random Generators

The general idea is to eliminate undesirable characteristics

- Empirical tests – statistical tests, generates uniform random numbers for tests
- Theoretical tests – uses numerical parameters but do not generate numbers

- Frequency test (uniformity test)
- Runs test
- Autocorrelation test
- Gap test
- Poker test

Frequency Tests (uniformity test)

- Counts how often numbers in a given range occur in the sequence to ensure that the numbers are uniformly distributed

Example;

If there are 5,000 3-digit numbers from 000 to 999, then we expect that there are about 500 numbers in the range of 00 to 99. If there are exact 500 numbers in each 10 segments, then the system is not a random one

Therefore, how much deviation should we allow from expected value of 500 and still accept the sequence as uniformly distributed?

Use chi-square (goodness of fit) test

Chi-square χ^2 test

- Statistical test
- Checks how well certain observed data fit the theoretically expected data
- First divide the observed data (here, the random numbers) into k non-overlapping classes, k must be >= 3
- Then count O_i, the number of times the observed data falls in each class i, i = 1, 2, 3, 4, . .
- Then measure how far the observed frequency deviates from the expected

$$chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Chi-square χ^2 test

$$chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- O_i is the observed number of sequences of value x_i and E_i is the expected number of occurrences of x_i
- This is showing the square of the differences the observed values and the expected values (divided by the expected values to normalize it)
- Now consider two hypotheses:
  - NULL Hypothesis, H0 : Y matches distribution of X
  - Hypothesis H1: Y does not match distribution of X
- If χ^2 is too large, we reject the null hypothesis (i.e. the distributions do not match)
- If χ^2 is small(ish) we do not reject the null hypothesis (i.e. the distributions may match)

Chi-square test

- If the observed data depart more from the expected value, the value of chi^2 will be larger
- In our example, E_i = 500 for i=1, 2, 3, . . . , 10

i	Range	Oi no. of observed occurrences	No. of expected occurrences	$(Oi\text{-}Ei)^2$	$(Oi\text{-}Ei)^2/Ei$
1	000-099	468	500	1024	2.048
2	100-199	519	500	361	0.722
3	200-299	480	500	400	0.8
4	300-399	495	500	25	0.05
5	400-499	508	500	64	0.128
6	500-599	426	500	5476	10.952
7	600-699	497	500	9	0.018
8	700-799	515	500	225	0.45
9	800-899	463	500	1369	2.738
10	900-999	529	500	841	1.682

Chi-square test

- Chi² = 19.588
- For large values, the hypothesis is rejected if

$$X^2 > X^2_{k-1,1-\alpha}$$

where $X^2_{k-1,1-\alpha}$ is upper critical point

$$X^2_{k-1,1-\alpha} \approx (k-1)\left\{1-\frac{2}{9(k-1)}+z_{1-\alpha}\sqrt{\frac{2}{9(k-1)}}\right\}^3$$

Where $Z_{1-\alpha}$ is the upper $1-\alpha$ critical point of the N(0,1) distribution

Degree of freedom v = k – 1, where k is no. of sets

α = Pr(reject H_0 | H_0 true), literally – level of significance

Chi-square test

- Let’s consider, probability of neglecting H0 is 0.05
- Then,

$$\alpha = 0.05$$
$$1-\alpha = 0.95$$
$$k = 10$$
$$k-1 = 9$$

??

$$\alpha = 0.1$$
$$1-\alpha = 0.9$$
$$k = 10$$
$$k-1 = 9$$

??

Are we going to reject the hypothesis or not?

Kolmogorov-Smirnov Test

- Another kind of frequency test
- Compares continuous cdf, F(x), of the uniform distribution to the empirical cdf, $S_N(x)$, of the sample of the N observations

$$F(x) = x, \qquad 0 \le x \le 1$$

- If sample from the random-number generator is R_1, R_2, . . . , R_N, then the empirical cdf, $S_N(x)$ is defined by

$$S_N(x) = \frac{\text{number of R1, R2, . . . , RN which are} \le x}{N}$$

- As N becomes larger, $S_N(x)$ should become a better approximation to F(x), provided that the null hypothesis is true

Kolmogorov-Smirnov Test

- This test is based on the largest absolute deviation between F(x) and $S_N(x)$ over the range of the random variable

Steps

1. Rank the data from smallest to largest. E.g. $R_{(1)} <= R_{(2)} <= R_{(3)} <= \ldots\ldots <= R_{(N)}$
2. Compute

$$D^+ = \max_{1 \le i \le N} \left\{ \frac{i}{N} - R_{(i)} \right\}$$

$$D^- = \max_{1 \le i \le N} \left\{ R_{(i)} - \frac{i-1}{N} \right\}$$

Kolmogorov-Smirnov Test

3. Compute $D = \max(D^+, D^-)$
4. Determine the critical value $D\alpha$ from Kolmogorov-Smirnov Critical values table
5. If $D > D\alpha$, the null hypothesis is rejected, otherwise accepted

Example 7.6, page 267

Runs test – runs up and runs down

- Direct test of independence (only) assumption
- Chi-square test might show some numbers are uniformly distributed but the ordering matters in case of independence
- Look at example on page 270 of Discrete Event System Simulation, Banks, Nicol, Nelson
- The general idea is to look at the patterns of runs up and runs down
- E.g.

  0.08 0.18 0.23 0.36 0.42 0.55 0.63 0.72 0.89 0.91

  One run – run up

  0.08 0.93 0.15 0.96 0.26 0.84 0.28 0.79 0.36 0.57

  Nine run – five up, four down

Runs test – runs up and runs down

- The run pattern is likely to be somewhere between two extremes
- If N is the number of numbers in the sequence, the maximum number of runs is N-1 and minimum number of runs is one
- Now, if ***a*** is total number of runs in truly random sequence, the mean and variance of ***a*** are given by

$$\mu_a = \frac{2N-1}{3}$$

$$\sigma_a^2 = \frac{16N-29}{90}$$

- For N>20, the distribution of a is approximated by normal distribution

$$N(\mu_a, \sigma_a^2)$$

Runs test – runs up and runs down

- Therefore, standardized normal test parameter

$$Z_0 = \frac{a - \mu_a}{\sigma_a}$$

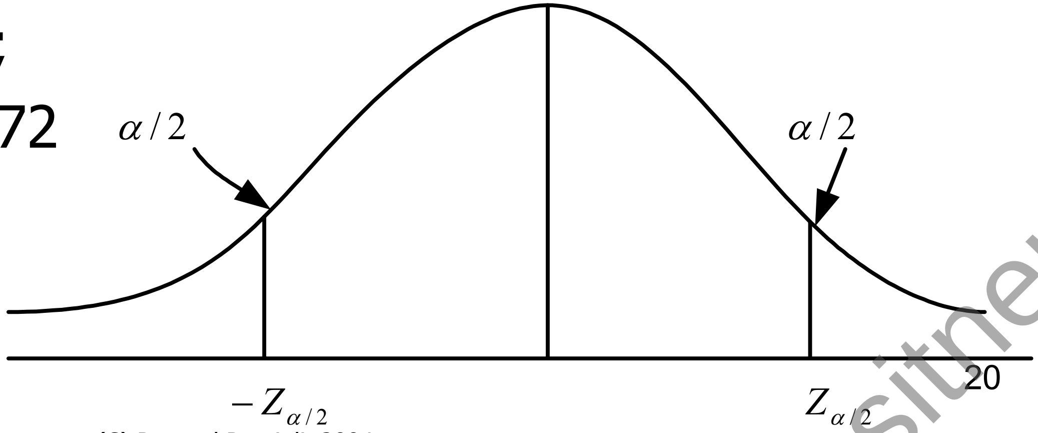
$$Z_0 = \frac{a - [(2N-1)/3]}{\sqrt{(16N-29)/90}}$$

- Where, Z_0 is approximately equal to N(0,1)
- Therefore, the hypothesis can not be rejected if

$$-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2}$$

Class-demonstration;

Example 7.8, page 272



Runs test – runs above and below the mean

- Runs up and down don't indicate any absolute organization of the data (or lack thereof)
- We may also want to test the number of runs (i.e consecutive values) above and below the mean
  - Too many or too few is not a good sign
- Calculation for mean and variance is more complex, but we handle this test in more or less the same way as the runs up and runs down test
  - Don't want the our number to be too far from the mean
  - Use the standard normal distribution to test

Runs test – runs above and below the mean

If n1 and n2 is the number of individual observation above and below the mean and b is total number of runs,

mean $$\mu_b = \frac{2n_1n_2}{N} + \frac{1}{2}$$

variance of b $$\sigma_b^2 = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}$$

If n1 or n2 is greater than 20, b is approximately normally distributed

Runs test – runs above and below the mean

Let's define standardized variable Z_0 for testing

$$Z_0 = \frac{b-(2n_1n_2/N)-1/2}{\left[\dfrac{2n_1n_2(2n_1n_2-N)}{N^2(N-1)}\right]^{1/2}}$$

Failure to reject the hypothesis of independence

occurs when $-Z_{\alpha/2} \le Z_0 \le Z_{\alpha/2}$

where α is the level of significance

Class-demonstration

Example 7.9, page 274

Runs test – length of runs

Given a sequence of random numbers, we can expect a certain number of runs of each length 1, 2, 3, …

If the run lengths in our empirical tests differ too much from the expected distribution, we can reject the generator

Idea is to determine the empirical distribution of runs of various lengths, then compare it with the expected distribution given random data

The distributions are compared using the Chi-Square test

Sample from book, page 274

In the example, we saw that run with length two is occurring. In reality, it must be random

Runs test – length of runs

As in chi-square test for the samples, the runs themselves should also poses uniformity

Determined by;
$$\chi_0^2 = \sum_{i=1}^{L} \frac{\left[O_i - E(Y_i)\right]^2}{E(Y_i)}$$

where,

Yi = number of runs of length in a sequence of N numbers

L = N – 1 for runs up and down

L = N for runs above and below the mean

Derivations – on blackboard

Runs test – length of runs

If null hypothesis of independence is true, then $\chi_0 2$ is approximately chi-square distributed with L-1 degrees of freedom

Example 7.10 – page 275

Autocorrelation test

- Tests relationship of items m (lag) locations apart
- Autocorrelation is determined for

$$\mathbf{R_{i}, R_{i+m}, R_{i+2m}, \ldots, R_{i+(M+1)m}}$$

- The value of M is the largest integer that satisfies;

$$i+(M+1)m <= N$$

where N is total number of values in the sequence

- A nonzero autocorrelation implies a lack of independence. Let

$$H_0 : \rho_{im} = 0$$

$$H_1 : \rho_{im} \neq 0$$

Autocorrelation test

- For large value of M, the distribution of the estimator of ρ_{im} ($\hat{\rho}_{im}$) is approximately normal if the values **R_i, R_{i+m}, R_{i+2m}, ..., $R_{i+(M+1)m}$** are uncorrelated
- Therefore;

$$Z_0 = \frac{\hat{\rho}_{im}}{\sigma\hat{\rho}_{im}}$$

where

$$\hat{\rho}_{im} = \frac{1}{M+1}\left[\sum_{k=0}^{M} R_{i+km} R_{i+(k+1)m}\right] - 0.25$$

$$\sigma\hat{\rho}_{im} = \frac{\sqrt{13M+7}}{12(M+1)}$$

Autocorrelation test

- Do not reject the null hypothesis of independence if $-Z_{\alpha/2} \le Z_0 \le Z_{\alpha/2}$
- If $\rho_{im} > 0$, the subsequence is said to exhibit positive autocorrelation. High random numbers in the subsequence followed by high, and low followed by low)
- If $\rho_{im} < 0$, the subsequence is said to exhibit negative autocorrelation. Low random numbers tend to be followed by high ones, and vice versa
- For desired property of independence, which implies zero autocorrelation

Gap test

- The gap test is used to determine the significance of the interval between the recurrences of the same digit
- The gap of length ***x*** occurs between the recurrences of some specified digit
- E.g.

4, 1, 3, 5, 1, 7, 2, 8, 2, 0, 7, 9, 1, 3, 5, 2, 7, 9,
4, 1, 6, 3, 3, 9, 6, 3, 4, 8, 2, 3, 1, ..

For digit 3, the first gap is 10 and second gap is 7.

$$P(\text{gap of } 10) = \overbrace{P(\text{no } 3), \ldots (P \text{ no } 3)}\, P(3)$$

Gap test

- The probability of the digits not being 3 is 0.9

  Therefore, P(gap of 10)= $(0.9)^{10}(0.1)$

- The theoretical frequency distribution for randomly ordered digits is given by

  P(gap <= x) = F(x) = $0.1\sum_{n=0}^{x}(0.9)^{n} = 1 - 0.9^{x+1}$

Steps

1. Specify the cdf for the theoretical frequency distribution given by P(gap<=x) based on the selected class interval width
2. Arrange the observed sample of gaps in a cumulative distribution with these same classes

Gap test

Steps

3. Find D, the maximum deviation between F(x) and $S_N(x)$ as in $D = | F(x) - S_N(x) |$
4. Determine the critical value, Dα , from Kolmogorov-Smirnov Critical Values table for the specified value of α and the sample size N
5. If the calculated value of D is greater than the tabulated value of Dα, the null hypothesis of independence is rejected

Example 7.13, page 282

Poker test

Test for interdependence based on the frequency with which certain digits are repeated in a series of numbers

e.g.

0.255, 0.577, 0.331, 0.414, 0.828, 0.909, 0.303, 0.001..

Here, there are three possibilities

1. The individual numbers can all be different
2. The individual numbers can all be the same
3. There can be one pair of like digits

The probability associated with each of these possibilities is given as;

P(three different digits) = P(second different from the first) * P(third different from the first and second)

= (0.9)*(0.8)

= 0.72

Poker test

P(three like digits) = P(second digit same as the first) * P(third digit same as the first)

= (0.1) * (0.1) = 0.01

P(exactly one pair) = 1 - 0.72 – 0.01 = 0.27

Alternatively, the last result can be obtained as follows;

P(exactly one pair) = $\binom{3}{2}(0.1)(0.9) = 0.27$

Normally used with chi-square test.

Example 7.14, page 283

Code snippet

```
/* Returns values between 0 and 1*/
/* Depends on 32 bit representation for integers */
double randu( seed )
  long int  *seed;
{
  long int  a = 16807,
            mod = 2147483647;    /* 2^31 - 1 */
  double    dmod = 2147483647.0; /* 2^31 - 1 */

  *seed = *seed * a;
  if ( *seed < 0 )
    *seed += mod;
  return( ( double ) *seed / dmod );
}
```

Discrete distributions

In real life, the probability of occurring one kind of event/data has different probability than other kind of event/data

Normally, there is a need for discrete but non-uniform distribution

Therefore, while generating a uniform random number 'U', the value is compared with the values of y. If the value falls in a interval $y_i < U \leq y_{i+1}$ (I = 0, 1, . . . 4) the corresponding value of x_{i+1} is taken as the desired output

No. of items (x)	Probability p(x)	Cumulative prob. (y)
0	0	0
1	0.10	0.10
2	0.51	0.61
3	0.19	0.80
4	0.15	0.95
5	0.05	1.00

Discrete distributions

When the number U is being tested, only 10% of the numbers are validated by using the first entry where as only 61% of numbers are validated by using second entry

For such cases, the table can be rearranged as the ordering doesn’t matter for testing

Cumulative prob. (y)	No. of items (x)
0.51	2
0.70	3
0.85	4
0.95	1
1.00	5

Now, the first entry can satisfy 51% of data

Non-uniform continuously distributed RNs

- Random numbers drawn from a distribution that is continuous & non-uniform
- These methods are based on use of sequences of uniformly distributed random numbers
- The algorithm must satisfy
  - Exactness – the random variates must follow the desired distribution
  - Efficient – in terms of time (setup time, marginal execution time) and space
  - Less Complexity
- Inverse transformation method – most widely used
- Rejection, Composition, Convolution etc.

Non-uniform continuously distributed RNs

Inverse transformation method

If U_i (i = 1, 2, 3, 4, …) are independent variables, uniformly distributed over the interval 0 to 1, and

$F^{-1}(x)$ is the inverse of the cumulative distribution function for random variable x, then the random variables defined by $x_i=F^{-1}(U_i)$ are the random sample of x.

If desired PDF is

$$f(x) = \begin{cases} a(1-x)\,(0 \le x \le 1) \\ 0 \quad otherwise \end{cases}$$

Cumulative;

$$F(x) = a(x - \frac{x^2}{2})$$

Non-uniform continuously distributed RNs

Let $U = F(x)$

$$x = 1 - (1 - U)^{1/2}$$

U	X
0.1009	69.23
0.3754	88.48
0.0842	66.65
0.9901	142.33

Non-uniform continuously distributed RNs

For exponential distribution

CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

now, find U(0,1)

then,

F(x) = R

or

$$1 - e^{-\lambda x} = R$$

$$e^{-\lambda x} = 1 - R$$

$$-\lambda x = \ln(1 - R)$$

$$x = \left(-\frac{1}{\lambda} \ln(R) \right)$$

$\beta = \frac{1}{\lambda}$ mean value

R is cumulative

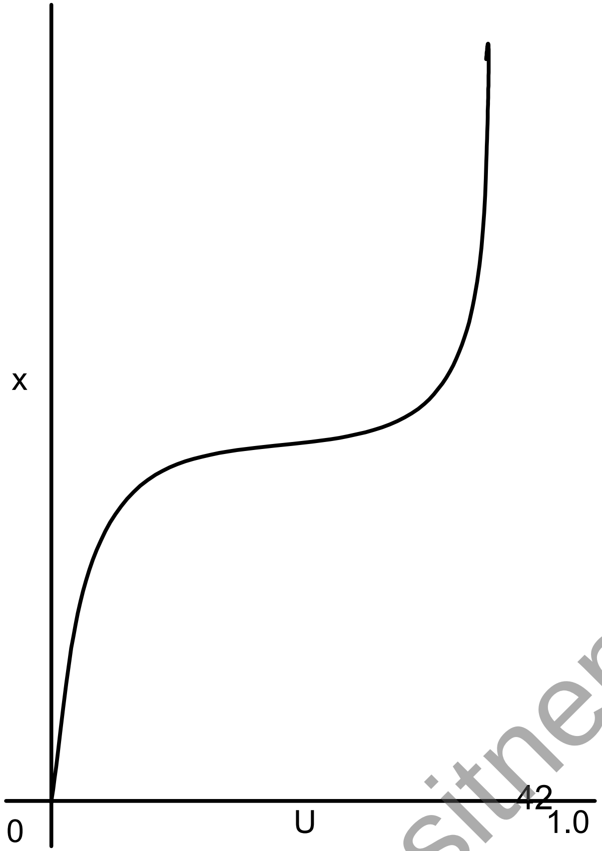
Non-uniform continuously distributed RNs

In case of discrete data, if the result falls between two values, then upper value is taken

But in case of continuous system, any value might occur

In such case, interpolation is applied

The plot of uniform random numbers and the input data (x) is;



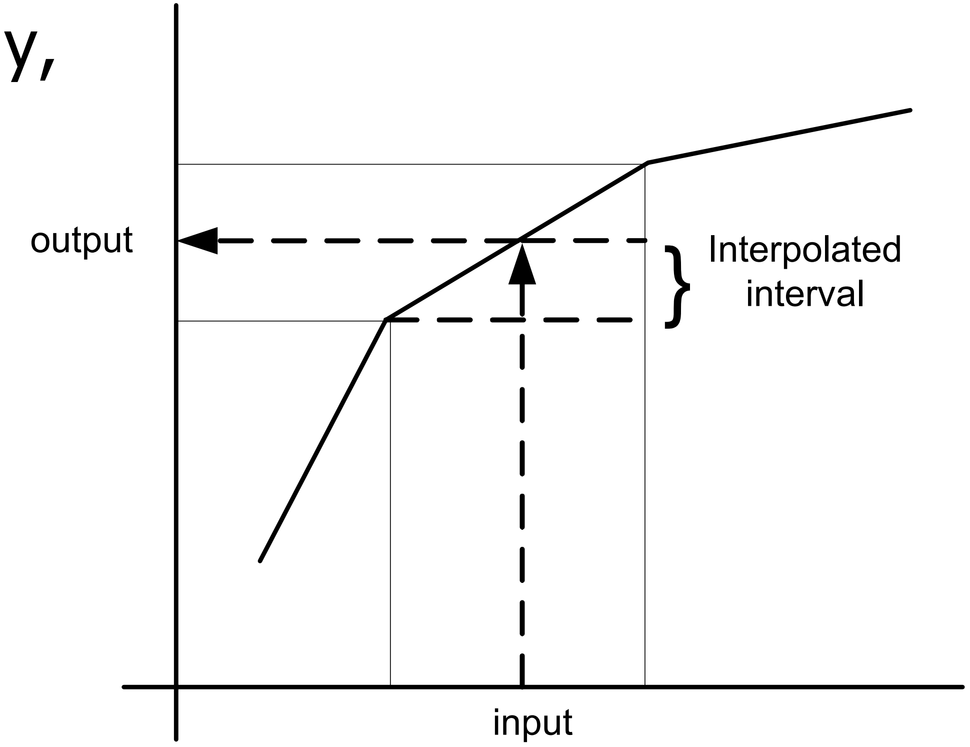
(C) Pramod Parajuli, 2004

Source: www.csitnepal.com

Non-uniform continuously distributed RNs

The general idea is, if a value falls in between two values, the result is calculated by using lower value plus portion of the output that is divided by the input

Graphically,



Non-uniform continuously distributed RNs

The standard curve and its points do not change

Let's define the slope of the segment as;

$$a_i = \frac{x_{i+1} - x_i}{y_{i+1} - y_i} \qquad (i = 0, 1, 2, 3, . . . N-1)$$

now,

if $Y_i < U < Y_{i+1}$ (i = 0, 1, 2, 3, . . . N-1)

then,

$$x = x_i + a_i(U-Y_i)$$

U	X
0.1009	69.29
0.3754	68.48
0.6842	66.65 . .

Non-uniform continuously distributed RNs

Disadvantage of inverse-transform

1. For normal and gamma distribution, there is no mathematical expression of F^{-1}
2. For given distribution, it may not be the fastest way

Advantages

1. Even if the distribution is truncated, can generate the output
2. Facilitates variance reduction that rely on correlation

Non-uniform continuously distributed RNs

Composition

Applies when the distribution function F from which we wish to generate can be expressed as a convex combination of other distributions F_1, F_2, . . .

For all x,

$$F(x) = \sum_{j=1}^{\infty} p_j F_j(x)$$

$$p_j \geq 0$$

$$\sum_{j=1}^{\infty} p_j = 1$$

F_j is distribution function

Non-uniform continuously distributed RNs

If, $p_k > 0$, $p_j = 0$, $j>k$

then, it will be finite

pdf = f(x) = $\sum_{j=1}^{\infty} p_j . f_j(x)$

Algorithm

1. Generate positive random number J such that

   $P(J - j) = pj$ for j = 1, 2, 3, 4, . .

2. Return X with distribution function F_J

Non-uniform continuously distributed RNs

Convolution

Probability distribution of a sum of two or more independent random variables is a convolution of the distributions of the original variables

Algorithm

Let F be the distribution function of X and G be the distribution function of Y_j

1. Generate Y_1, Y_2, Y_3, . . . Y_m, independent and identically distributed random variables each with distribution function G
2. Return, $X = Y_1 + Y_2 + Y_3 + . . . + Y_m$

Non-uniform continuously distributed RNs

Rejection method

Applied when the PDF f(x) has a lower and upper limit to its range a and b and upper bound c.

However, the upper bound ‘c’ is not very much relevant

Steps

1. Compute the values of two independent uniformly distributed variables U_1 and U_2
2. Compute, $X_0 = a + U(b – a)$
3. Compute, $Y_0 = c.U_2$
4. If $Y_0 <= f(X_0)$ accept X_0 otherwise repeat with two new uniform variates

Non-uniform continuously distributed RNs

Recall the Monte Carlo method

The probability density function is enclosed in a rectangle i.e. if a point is accepted, ‘n’ is incremented by 1

The curve f(x) represents probability density function. Therefore, the area under the curve must be 1. But,

$$c.(a-b) = 1$$

In such case, the graph must be chosen so that c.(a-b) and f(x) both are 1

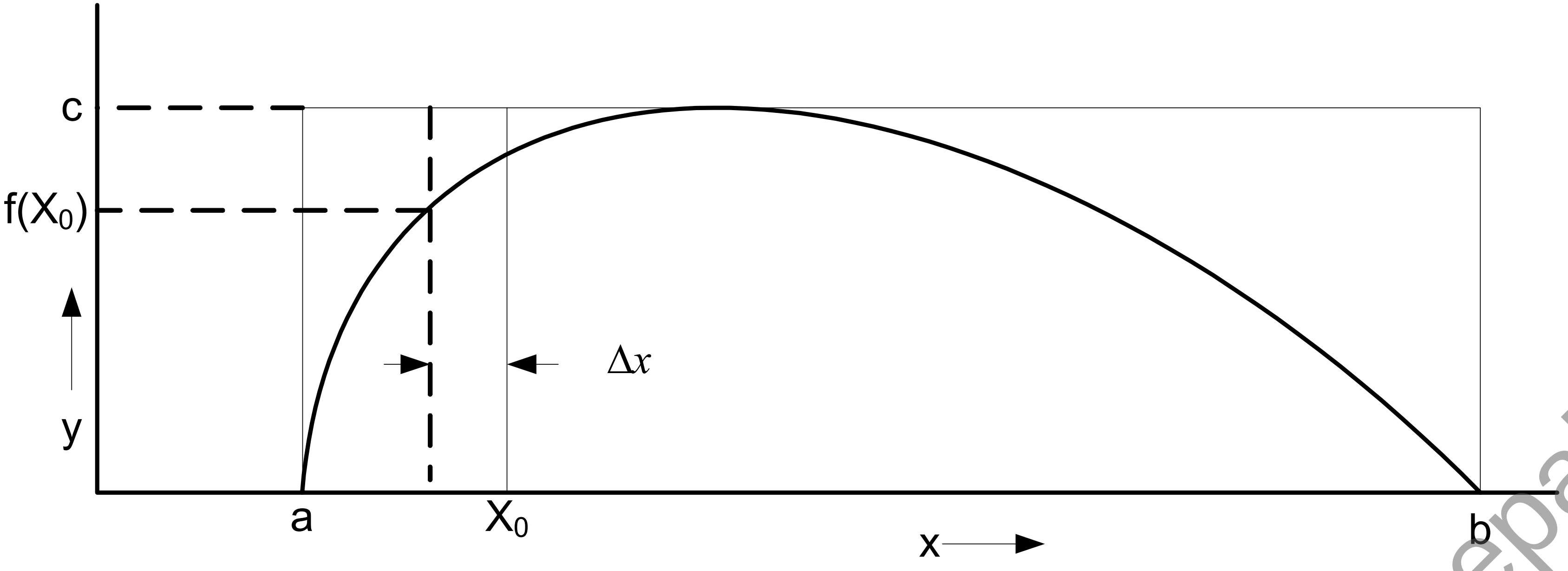
Therefore, is this a valid method?

Non-uniform continuously distributed RNs

Given, X_0,

X_0 is accepted if,

$Y_0 <= f(X_0)$



Non-uniform continuously distributed RNs

Let Δx be small area. If Y_0 still is inside the small rectangle bounded by Δx, under the curve, now output range will be,

$X_0 - \Delta x$ to X_0

In such case,

P(X <= X_0) = P (Y_0 bounded by left of X_0)

$$F(X_0) = \frac{\int_a^{X_0} f(x)dx}{c(x_0 - a)} \cdot \frac{(x_0 - a)}{(b-a)}$$

Therefore,

$$F(X_0) = \int_a^{X_0} f(x)dx$$

Non-uniform continuously distributed RNs

Disadvantage

- Two uniform variates must be calculated
- The situation will be even more complex if more iterations need to be performed

Therefore, it is usable only if,

The probability density function is given by a mathematical function

The PDF must be limited i.e. F(x) = 0 for < a and > b. If not, then use inverse transformation method

Sample function

```
/* C implementation Park & Miller’s random function     */
double random ( seed )
 long int   *seed;
{
   long int  a = 16807,            /* 7^5 */
            m = 2147483647,       /* 2^31 - 1 */
            q = 127773,           /* m / a (int divide) */
            r = 2836,             /* m % a */
            lo, hi, test;
   double dm = 2147483647;

   hi = *seed / q;
   lo = *seed % q;
   test = a * lo - r * hi;
   *seed = test > 0 ? test : test + m;
   return( (double ) *seed / dm );
}
```