SAMPLING METHODS

3/17/2013

1. INTRODUCTION

The Monte Carlo simulation of different phenomena requires the sampling of the distributions that represent their behavior. Mastering the process of sampling these distributions allows for the simulation of any process one can think of.

The concepts of discrete and continuous random variables, probability density function (pdf), and cumulative distribution function (cdf) will be introduced. Next some relevant probabilistic quantities such as the mean value and the variance will be introduced. Then methods of sampling discrete and continuous probability density functions will be exposed. When analytical forms of the distributions are not readily available, the rejection method is shown to provide an approach for sampling difficult distributions.

It is important to thoroughly analyze the sampling process, since when variance reduction is attempted; this would involve modifying the sampling process with the objective of effectively simulating complex processes and phenomena.

2. THE CONCEPT OF A RANDOM VARIABLE

In practice, the term "random variable" or "stochastic variable" is used to emphasize the fact that one does not exactly know what specific value this kind of variable will take at a given occurrence. However the probabilities of different occurrences are usually known. Such a variable is shrouded by the kind of uncertainty designated as "randomness" that probability theory tries to estimate mathematically in an *exact*, surely not random way. Other forms of uncertainty exist in nature such as "fuzziness" that possibility theory also tries to estimate in an *exact* mathematical, surely not fuzzy, way.

Random variables are encountered in diverse fields of science and engineering that Monte Carlo simulation can handle. Examples of random variables, among many others, are:

1. The number of cosmic ray particles falling on an area of the Earth's surface within a certain time interval.
2. The number of phone calls arriving at a telephone exchange in a certain amount of time.
3. The deviation of the point of impact of a shell around the center of its target.
4. The velocity of a gas molecule in a gas, particle in a plasma, or neutron in a shield or a nuclear reactor.
5. Electrons energies in a conductor or a superconductor.
6. Number of photons of light, x-rays, or gamma rays falling on a surface.
7. Velocity of fluid particles in a turbulent flow in a pipe.
8. Temperature distributions in a car, plane or ship engine.

Mathematically speaking, the last examples share several similarities:

1. In each case we have to deal with a quantity of interest describing the phenomenon under study,
2. Under the effect of random circumstances, each quantity can take a variety of values,
3. One cannot state a priori what value the quantity of interest will assume, since it varies in a random fashion from one instance to the other.

To describe a random variable we must:

First: Know the values that it can assume.

Second: Know how often, or with what probability, it assumes these values, under a sufficiently large number of trials. This is possible by means of a "probability density function," abbreviated as pdf, of the random variable.

3. PROBABILITY DENSITY FUNCTIONS

Two situations will present themselves to us: discrete and continuous random variables.

If a random variable ξ assumes the discrete values: $x_1, x_2, x_3, \ldots, x_n$; it can be specified by the probability distribution function:

$$\xi : \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix} \tag{1}$$

where: $p_1, p_2, p_3, \ldots, p_n$ are the probabilities of occurrence of the values $x_1, x_2, x_3, \ldots, x_n$.

This probability density function p_i can be formally expressed by the expression:

$$P(x) \equiv P\{\xi = x_i\} = p_i \tag{2}$$

The outcome values $x_1, x_2, x_3, \ldots, x_n$ are arbitrary, but the probabilities p_i must satisfy the two conditions:

$$p_i \geq 0 \tag{3}$$

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \ldots + p_n = 1 \tag{4}$$

As an example one can consider the random variable representing the flipping of a coin, with the possible outcomes heads (H) and tails (T) and their associated probabilities p(H) = 0.5, p(T) = 0.5:

$$\xi_{coiu\ \ flip} : \begin{pmatrix} H & T \\ p(H) & p(T) \end{pmatrix} = \begin{pmatrix} H & T \\ 0.5 & 0.5 \end{pmatrix} \tag{5}$$

Another example is the random variable representing the points obtained in throwing a single die:

$$\xi_{die\ \ throw} : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} \tag{6}$$

Such a discrete probability density function is shown in Fig. 1.

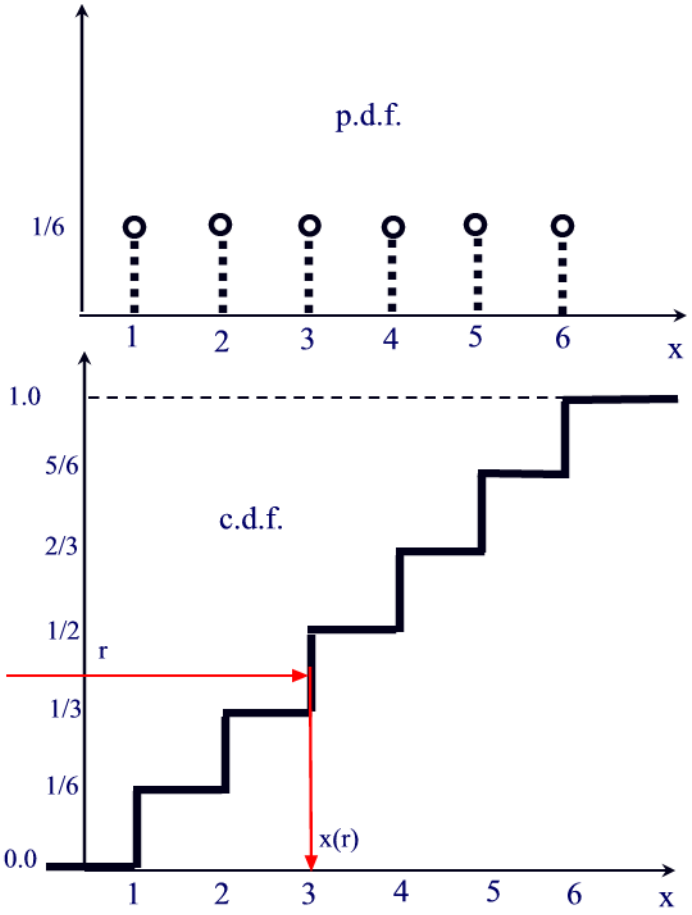

Figure 1. Discrete probability density function (pdf) and the associated cumulative distribution function (cdf) for the single die throws random variable.

Another example that is analogous to a five-faced die is the random variable that represents the expected number of neutrons outgoing a collision with a nucleus. One can derive the probability density function for this random variable as follows.

The total macroscopic cross section for a neutron colliding with a nucleus can be written as:

$$\Sigma_t = \Sigma_s + \Sigma_{n,2n} + \Sigma_{n,3n} + \Sigma_c + \Sigma_f \quad [cm^{-1}] \tag{7}$$

where: $\sum_t$ is the total macroscopic cross section,
$\sum_s$ is the macroscopic scattering cross section,
$\sum_{n,2n}$ is the macroscopic (n,2n) reaction cross section,
$\sum_{n,3n}$ is the macroscopic (n,3n) reaction cross section,\
$\sum_c$ is the macroscopic capture cross section,
$\sum_f$ is the macroscopic fission cross section.

Dividing into $\sum_t$ one can derive the probability density function for the different reactions as:

$$\frac{\Sigma_s}{\Sigma_t} + \frac{\Sigma_{n,2n}}{\Sigma_t} + \frac{\Sigma_{n,3n}}{\Sigma_t} + \frac{\Sigma_c}{\Sigma_t} + \frac{\Sigma_f}{\Sigma_t} = 1 \tag{8}$$

Equation 8 satisfies the conditions on the probabilities imposed by Eqns. 3 and 4.

Using the pdf of Eqn. 8, one can now express the random variable representing a neutron collision as:

$$\xi_{neutron\ \ collision} : \begin{pmatrix} 1 & 2 & 3 & 0 & \nu_f \\ \frac{\Sigma_s}{\Sigma_t} & \frac{\Sigma_{n,2n}}{\Sigma_t} & \frac{\Sigma_{n,3n}}{\Sigma_t} & \frac{\Sigma_c}{\Sigma_t} & \frac{\Sigma_f}{\Sigma_t} \end{pmatrix} \tag{9}$$

where: ν_f is the average number of neutrons outgoing a fission reaction.

4. CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The probability density function of Eqn. 2 is associated with a cumulative distribution function (cdf) defined as:

$$C(x) = \sum_{x_i \le x} p_i \tag{10}$$

As a result of the condition in Eqn. 4:

$$C(x_n) \equiv 1.0 \tag{11}$$

In the case of a continuous random variable, we must define the interval [a,b] of its variation, and its probability density function p(x) can be defined by the equation:

$$P\{a \le \xi \le b\} = \int_a^b p(x)dx \tag{12}$$

The probability density function p(x) must satisfy the two conditions:

$$\begin{aligned} &1.\ p(x) \ge 0 \\ &2.\ \int_a^b p(x)dx = 1 \end{aligned} \tag{13}$$

A cumulative distribution function is similarly defined over a continuous random variable as:

$$C(x) = \int_{-\infty}^{x} p(x')dx' \tag{14}$$

As an example, the pdf for a distance to the next collision, or the transport kernel of a particle diffusing in an infinite medium with a total macroscopic cross section Σ_t is:

$$p(x) = \Sigma_t e^{-\Sigma_t x}, \quad \in [0, \infty] \tag{15}$$

and its cdf is:

$$C(x) = \int_0^x p(x')dx' = \int_0^x \Sigma_t e^{-\Sigma_t x'} dx' = 1 - e^{-\Sigma_t x} \tag{16}$$

The pdf and cdf are shown for this case in Fig. 2.

5. MATHEMATICAL EXPECTATION OF A RANDOM VARIABLE

The objective of a Monte Carlo simulation is normally the determination of the mathematical expectation, mean value, or expected value of a given random variable. In the discrete random variable case, the mathematical expectation or first moment is given by:

$$\mu(\xi) = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} = \sum_{i=1}^{n} x_i p_i \tag{17}$$

by virtue of the normalization condition of Eqn. 4.

For the continuous random variable case, the summation is replaced by an integral, yielding:

$$\mu(\xi) = \frac{\int_a^b xp(x)dx}{\int_a^b p(x)dx} = \int_a^b xp(x)dx \qquad (18)$$

by use of the normalization condition of Eqn. 13.

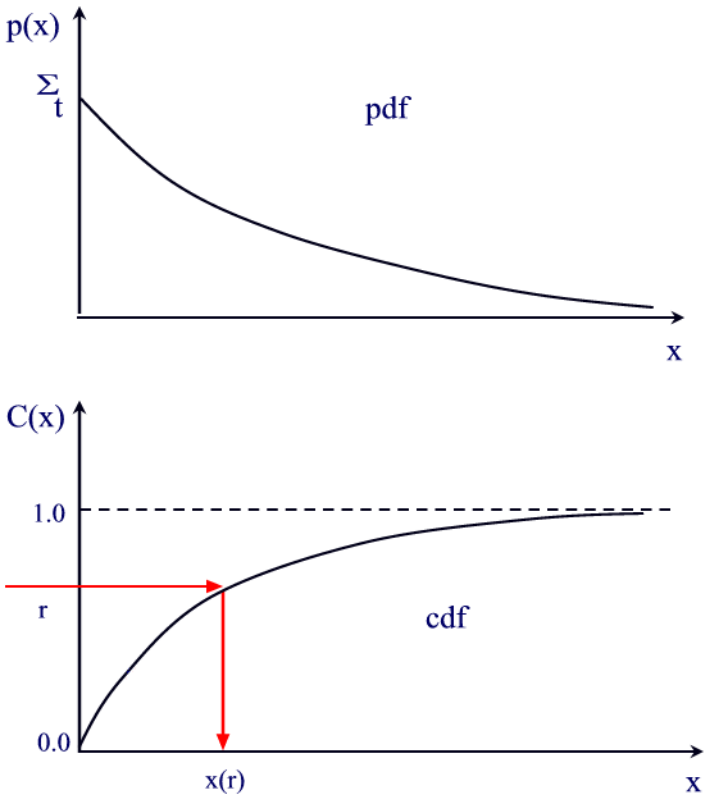

Figure 2. Probability density function (pdf) and cumulative distribution function (cdf) for the distance to the next collision of a particle diffusing in an infinite medium of total macroscopic cross-section Σ_t .

EXAMPLE 1

As an example, considering the discrete die throwing random variable, we can estimate its mean value, mathematical expectation or first moment as:

$$\mu(\xi)=\frac{1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}}{\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}}=\frac{21}{6}=3.5$$

EXAMPLE 2

As another example, considering the continuous particle diffusion random variable, we can estimate its mean value, mathematical expectation or first moment as:

.
$$\mu=\frac{\int_0^\infty x\Sigma_t e^{-\Sigma_t x}dx}{\int_0^\infty \Sigma_t e^{-\Sigma_t x}dx}=\frac{1}{\Sigma_t}=\lambda_t$$

which as expected suggests that the average distance traveled by the particle is equal to the inverse of the macroscopic total cross section, which is in turn is equal to the mean free path λ_t of the particle in the medium with macroscopic total cross section Σ_t.

6. THE VARIANCE OF A RANDOM VARIABLE

The variance of a random variable ξ is defined as:

$$V(\xi)=\mu\{[\xi-\mu(\xi)]^2\} \tag{19}$$

which is the expected value of the square of the deviation of the random variable from its mean value.

The last equation can be transformed in the following way:

$$V(\xi)=\mu\{\xi^2-2\xi\mu(\xi)+\mu^2(\xi)\} \tag{20}$$

Since:

$$\mu(c\xi)=c\mu(\xi), \tag{21}$$

we can write:

$$\begin{aligned}V(\xi)&=\mu(\xi^2)-2\mu(\xi)\mu(\xi)+\mu^2(\xi)\\&=\mu(\xi^2)-\mu^2(\xi)\end{aligned} \tag{22}$$

This second expression of the variance is equivalent to the one in Eqn. 19, and both equations can be used for the estimation of the variance. It defines the variance using a tongue twister as: "The mean of the squares, minus the square of the mean."

EXAMPLE 3

For the die throwing random variable, the variance can be estimated according to Eqn. 22 as:

$$V(\xi) = \frac{1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6}}{\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}} - (3.5)^2$$

$$= \frac{91}{6} - 12.25 = 2.917$$

7. SAMPLING DISCRETE PROBABILITY DENSITY FUNCTIONS

Two situations arise in Monte Carlo simulations where one needs to sample discrete probability density functions:

1. When the density functions themselves are discrete. Examples of this situation are the single die throwing random variable, the number of neutrons outgoing a certain reaction, discrete scattering direction cosines of particles, or multi-group cross sections data in particle transport calculations.
2. When the probability density functions are too complex, and one then needs to simplify their treatment by discretizing them and storing the discrete values in table form for sampling.

The following steps are usually followed:

1. The sampling of a discrete probability density function requires the construction of a cumulative probability density function (cdf) as shown in Fig. 1, and storing its values in table form.
2. A pseudo random number:

$$\rho \in [0,1]$$

is sampled over the unit interval.

3. The cumulative distribution function is inverted by searching its values for the position of the sampled random variable.

If:

$$p_1 + p_2 + ... + p_{i-1} < \rho \leq p_1 + p_2 + ... + p_i$$

then i is chosen and the corresponding outcome x_i is taken as a sample.

For instance in Fig. 1:

$$p_1 + p_2 < \rho \leq p_1 + p_2 + p_3$$

this results in i = 3, accordingly the inversion process yields $x_i = 3$ as the sampled value of x.

A procedure displaying the sampling of the coin flipping random variable is shown in Fig. 3. From the probability density function a cumulative distribution function is constructed, and is the sampled using a sequence of pseudo random numbers generated over the unit interval. The generated sample is stored to a file and reconstructed from the sample data to check the validity of the procedure and is shown in Fig. 4. It can be noticed that the correct probability density function is generated as we increase the number of trials to fully sample the sample space.

```
!       coin_flipping.for
!       Simulating the process of flipping a coin
!       M. Ragheb
        program coin_flipping
        dimension x(2), pdf(2), cdf(2), freq(2)
        integer :: trials = 10000000
        real :: c=0.5
        real pdf,cdf,freq,xtrials
        xtrials=trials
!       Initialize frequency distribution
        do i=1,
                freq(i)=0.0
        end do
!       Initialize possible outcomes of random variable
!       heads, i=1
!       tails, i=2
        do i=1,2
                x(i)=i
        end do
!       Initialize probability density function
        do i=1,2
                pdf(i)=c
        end do
!       write(*,*) pdf
!       Generate cumulative distribution function
        cdf(1)=pdf(1)
        cdf(2)=pdf(1)+pdf(2)
!       write(*,*) cdf
!       Open output file
        open(44, file = 'random_out')
!       Sample random variable and accumulate scores in frequency distribution
        do i= 1, trials
                call random(rr)
                if(rr.LE.cdf(1))then
                        freq(1)=freq(1)+1.0
                end if
                if((rr.GT.cdf(1)).AND.(rr.LE.cdf(2)))then
                        freq(2)=freq(2)+1.0
                end if
        end do
!       Normalize frequency distribution
```

```
        do i=1,2
                freq(i)=freq(i)/xtrials
        end do
!       Write results to output file
        do i=1,2
                write (44,100) x(i), freq(i)
                write(*,*) x(i),freq(i)
        end do
100     format (2e14.8)
        end
```

Figure 3. Procedure for the Monte Carlo simulation of coin flipping.

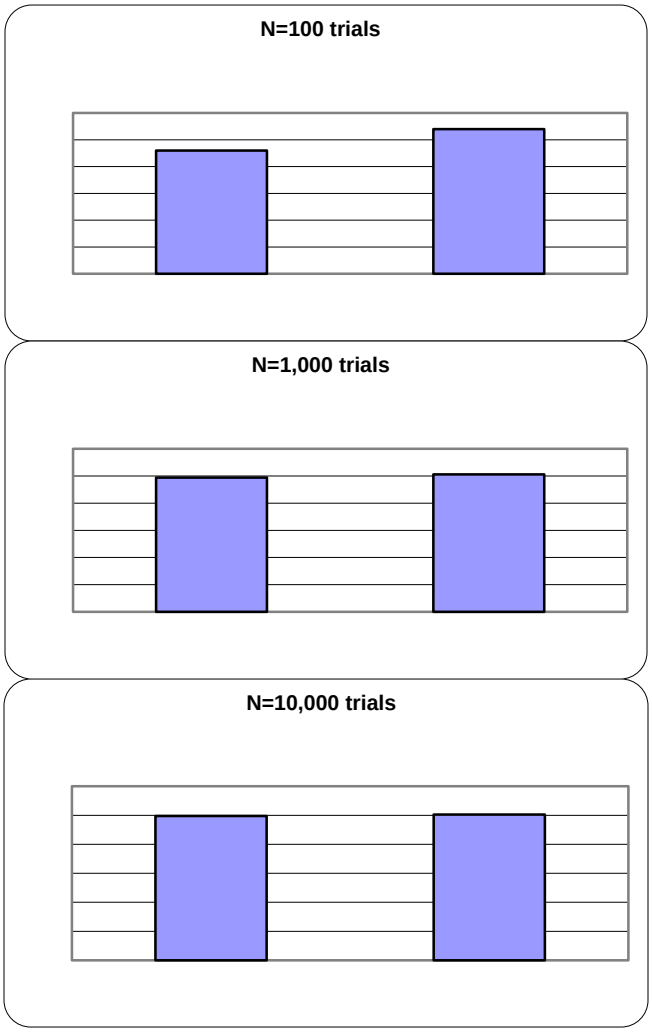

Figure 4. Results of increased number of trials for the Monte Carlo Simulation of coin flipping. Heads occurrences are represented by 1, and tails are represented by 2.

We also show a procedure for the sampling of the random variable representing the throw of a single die in Fig. 5. Notice that the construction of the cumulative distribution function follows directly from the procedure used to sample the coin-flipping

problem. Similarly, increasing the number of trials leads to a correct reconstruction of the actual probability density function as shown in Fig. 6.

```
!       die_throwing for
!       Simulating the process of throwing a single die
!       M. Ragheb
        program dice_throwing
        dimension x(6),pdf(6),cdf(6),freq(6)
        integer :: trials = 1000000
        real :: c=1.0/6.0
        real pdf, cdf, freq
!       Initialize frequency distribution
        do i=1,6
                freq(i)=0.0
        end do
!       Initialize possible outcomes of random variable
        do i=1,6
                x(i)=i
        end do
!       Initialize probability density function
        do i=1,6
                pdf(i)=c
        end do
!       write(*,*) pdf
!       Generate cumulative distribution function
        cdf(1)=pdf(1)
        do i=2,6
                cdf(i)=cdf(i-1)+pdf(i)
        end do
!       write(*,*) cdf
!       Open output file
        open(44, file = 'random_out')
!       Sample random variable and accumulate scores in frequency distribution
        do i= 1, trials
                call random(rr)
                if(rr.LE.cdf(1))then
                        freq(1)=freq(1)+1.0
                end if
                do j=1,5
                        if((rr.GT.cdf(j)).AND.(rr.LE.cdf(j+1)))then
                                freq(j+1)=freq(j+1)+1.0
                        end if
                end do
        end do
!       Normalize frequency distribution
        do i=1,6
                freq(i)=freq(i)/trials
        end do
!       Write results to output file
        do i=1,6
                write (44,100) x(i), freq(i)
                write(*,*) x(i),freq(i)
        end do
100     format (2e14.8)
        end
```

Figure 6. Procedure for the Monte Carlo simulation of single die throwing.

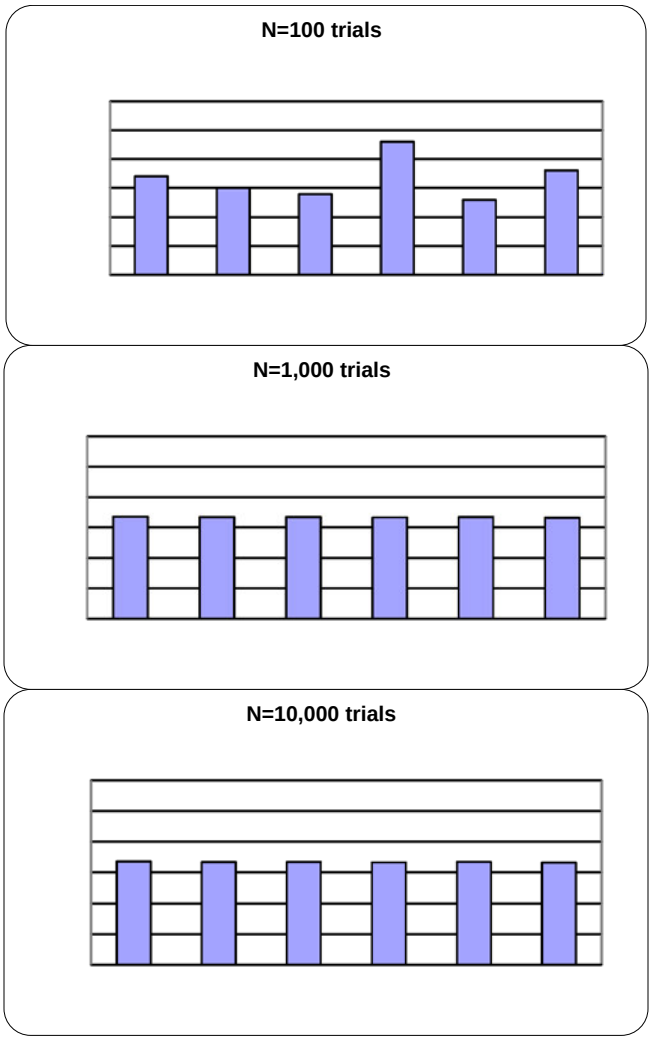

Figure 7. Frequency distribution for the increased number of trials for the Monte Carlo simulation of the single die throwing random variable.

8. INTERPOLATION OF SAMPLED DATA

When tabulated data is used as an approximation to an otherwise continuous distribution, interpolation between the tabulated values can improve the sampled values. The interpolation process can be undertaken in the following way:

1. The interval x_i in which the random variable lies is determined as in the case of discrete variables.
2. Linear interpolation can be used to get an improved value of the variable x from:

$$\frac{x_i - x(\rho)}{x_i - x_{i-1}} = \frac{C(x_i) - \rho}{C(x_i) - C(x_{i-1})} \tag{23}$$

The interpolation process is shown in Fig. 8 by considering the two similar triangles ABE and CDE with:

$$\frac{CE}{AE} = \frac{CD}{AB}$$

Equation 23 leads to an interpolated value:

$$x(\rho) = x_i - \frac{C(x_i) - \rho}{C(x_i) - C(x_{i-1})}(x_i - x_{i-1}) \tag{24}$$

Table lookout helps us reduce the number of operations needed for sampling a complicated function, but places a burden on the fast storage capacity. A balance must be struck between the needed sampling accuracy and the storage capacity requirements.

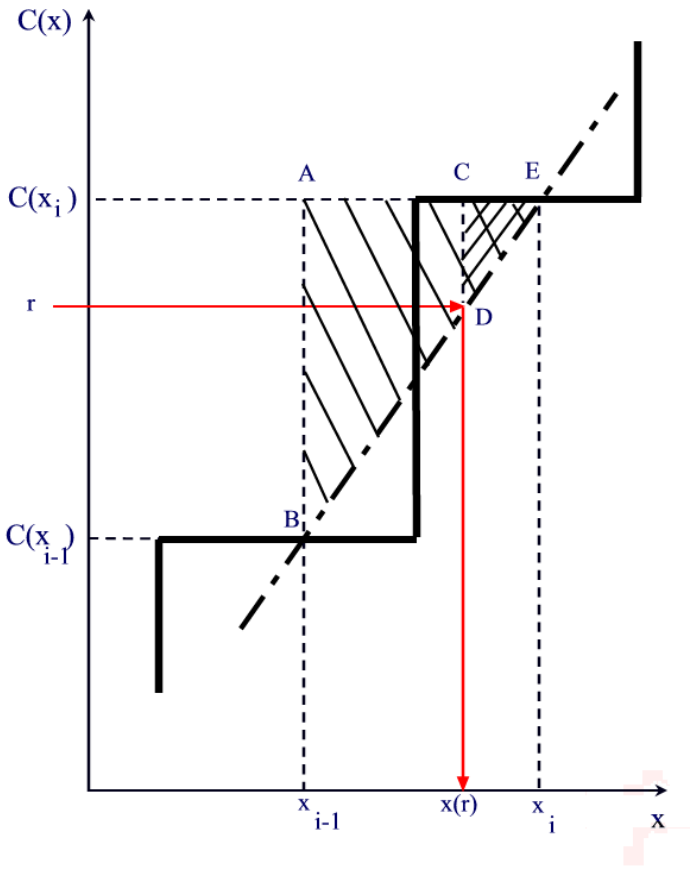

Figure 8. Linear interpolation over a cumulative distribution function (cdf) in histogram or table form.

9. SAMPLING CONTINUOUS DISTRIBUTIONS BY INVERSION

The probability that the random variable x has a value between x and x + dx is p(x)dx. The function p(x) is called the probability density function (pdf) or the differential distribution function. The probability C(x) that the random variable x' is less than x is called the cumulative distribution function (cdf) or integral distribution function.

The selection of a random variable distributed according to a given probability is of central importance in Monte Carlo simulations. Let ρ be a pseudo random number uniformly distributed over the unit interval:

$$\rho \in [0,1].$$

If we equate the cumulative distribution function to the generated pseudo random number such that:

$$C(x) = \rho, \tag{25}$$

then for each ρ there is a corresponding x, and the variable x is distributed according to p(x), and we can obtain x from the simple inversion of the cumulative distribution function:

$$x = C^{-1}(\rho) \tag{26}$$

This approach yields analytical expressions suitable for the sampling process and is normally used whenever the probability density functions lend themselves to simple mathematical manipulations.

10. SAMPLING PARTICLE TRANSPORT

We have deduced the cumulative distribution function for the transport kernel of a particle diffusing in an infinite homogeneous medium in Eqn. 16 as:

$$C(x) = \int_0^x p(x')dx' = \int_0^x \Sigma_t e^{-\Sigma_t x'} dx' = 1 - e^{-\Sigma_t x}$$

To generate values of x distributed according to the given cumulative distribution function, it is equated to a pseudo random number then inverted according to Eqns. 25 and 26:

$$C(x) = 1 - e^{-\Sigma_t x} = \rho,$$
$$e^{-\Sigma_t x} = 1 - \rho,$$
$$x(\rho) = -\frac{1}{\Sigma_t}\ln(1-\rho)$$

Since ρ is distributed in the same way as (1 - ρ) over the unit interval we could save a computational step by using the relationship:

$$x(\rho_i) = -\frac{1}{\Sigma_t}\ln \rho_i \ , \ i = 1, 2, \ldots, n \, , \qquad (27)$$

where now we are generating n sampled points using a sequence of n pseudo random numbers.

```
!       distance for
!       Simulating the process of a particle diffusing in a medium
!       M. Ragheb
!       probability density function:
!       sigma_total*exp(-sigma_total*x)
!       distance between collisions:
!       x = - mean_free_path * ln (rho)
!       where rho is a random number over the interval [0,1]
!       M. Ragheb, Univ. of Illinois
        program distance
        dimension x(100), freq(100)
        real :: sigma_total = 1.0
        integer :: trials = 100000
        real x,dist,freq,mean_free_path
!       Calculate mean_free_path
        mean_free_path = 1.0/sigma_total
!       Initialize frequency distribution
        do i=1,100
                freq(i)=0.0
        end do
!       Initialize distance_travelled bins in mean free paths
!       width=1.0
!       here in tenths of mean free paths
        width=0.1
!       here in hundredths of mean free paths
!       width=0.01
        do i=1,100
                x(i)=i * mean_free_path * width
        end do
!       open output file
        open(44, file = 'random_out')
!       Sample distances travelled and accumulate scores in frequency distribution
        do i= 1, trials
                call random(rr)
                dist = - mean_free_path * log (rr)
                if(dist.LE.x(1))then
                 freq(1)=freq(1)+1.0
```

```
                end if
                do j=1,99
                        if((dist.GT.x(j)).AND.(dist.LE.x(j+1)))then
                        freq(j+1)=freq(j+1)+1.0
                        end if
                end do
        end do
!       Normalize frequency distribution
        do i=1,100
                freq(i)=freq(i)/trials
!       turn into histogram
                freq(i)=freq(i)/width
        end do
!       Write results to output file
        do i=1,100
                write (44,100) x(i), freq(i)
                write(*,*) x(i),freq(i)
        end do
100     format (2e14.8)
        end
```

Figure 9. Procedure for the Monte Carlo simulation of particle transport in an infinite homogeneous medium.

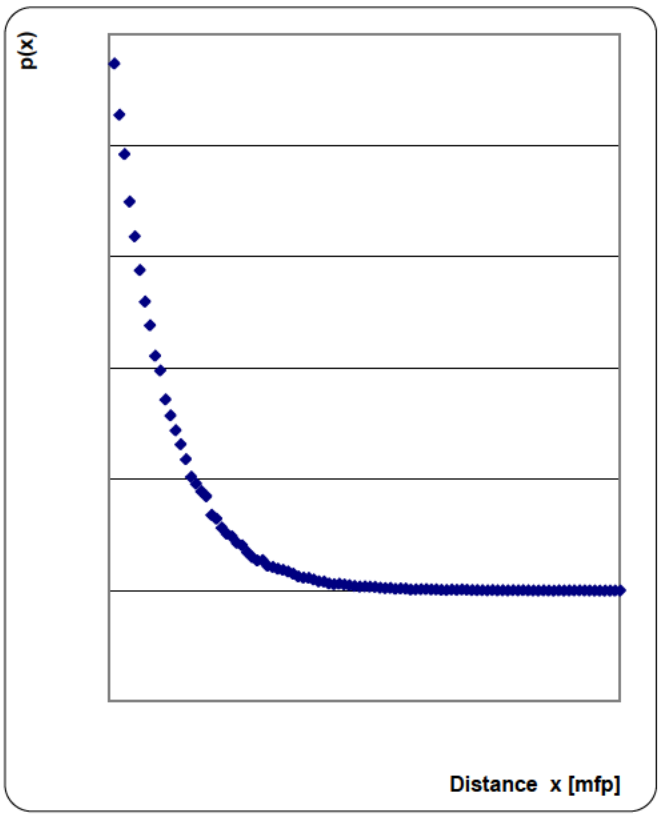

Figure 10. Sampled probability density function for the simulation of a particle diffusing in an infinite homogeneous medium with macroscopic total cross section = 1 $[cm]^{-1}$.

A procedure for the sampling the transport kernel of a particle diffusing in an infinite homogeneous medium is shown in Fig. 9. Notice the simpler structure of the algorithm compared to the case of a discrete distribution, where the construction of a table of discrete values and a process of table lookout and search is needed. In Fig. 10, the sampled distribution is shown for a unit macroscopic total cross section and mean free path.

11. REJECTION SAMPLING METHOD

If a bounded function f(x) cannot be easily inverted analytically, it can be sampled using the rejection method, due to von Neumann. If f(x) vanishes outside the interval [a,b], and we wish to construct samples from the density f(x), the following steps are used, with reference to Fig.11:

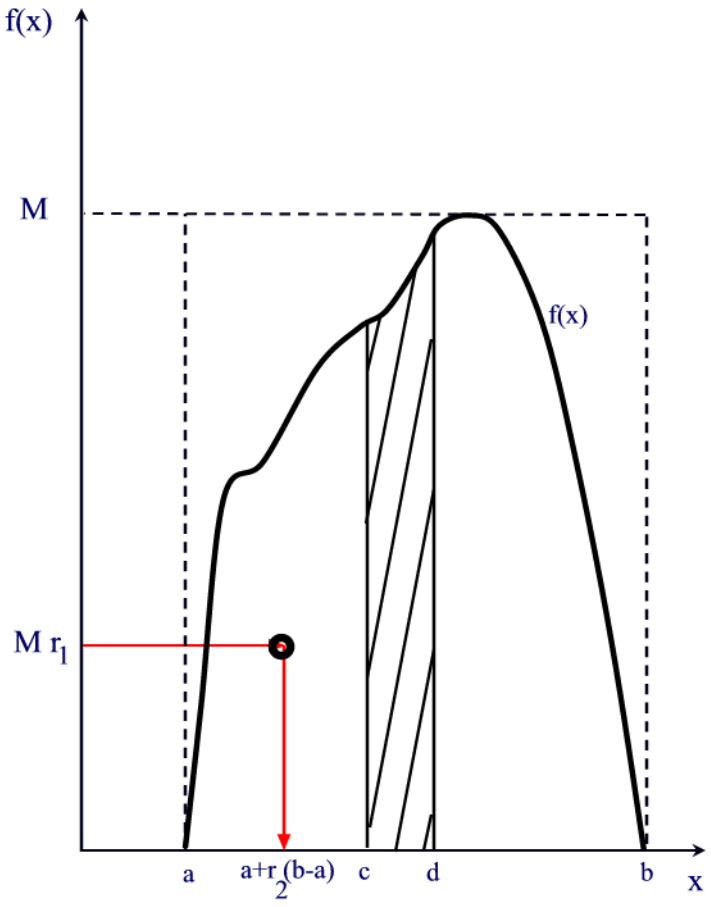

Figure 11. Sampling by the rejection method the probability density function f(x).

1. Determine the maximum value M that the function can reach over the interval [a,b].

2. Generate a pair of pseudo random numbers ρ_1 and ρ_2.
3. Interpret the point:

$$\{a + \rho_2(b-a), M\rho_1\} \tag{28}$$

as a point in the rectangle with base (b-a) and height M.

4. If this point falls below the graph of f(x):

$$M\rho_1 \leq f[a + \rho_2(b-a)], \tag{29}$$

then f [a + ρ_2(b-a)] is accepted as a valid sample from f(x).

5. If the point does not fall below the graph of f(x), reject the sample and repeat the sampling process until a sample has been determined to be valid.

The efficiency of such a technique, as measured by the fraction pairs of pseudo random numbers (ρ_1 , ρ_2) which are not rejected, is just the ratio of the area under the curve f(x) to the area of the enclosing rectangle.

To prove that this technique indeed samples the probability density function f(x), we recognize that the probability of a sampled point to fall below the curve f(x), and hence not to be rejected is equal to the ratio of the areas:

$$\frac{\int_a^b f(x)dx}{M(b-a)} = \frac{1}{M(b-a)} \tag{30}$$

The probability for the sampled point to fall below the curve in the interval:

$$c < x < d,$$

is also equal to:

$$\frac{\int_c^d f(x)dx}{M(b-a)}.$$

Hence the fraction of sampled values in the interval [c , d] among all sampled points is the ratio:

$$\frac{\left(\frac{\int_c^d f(x)dx}{M(b-a)}\right)}{\left(\frac{1}{M(b-a)}\right)} = \int_c^d f(x)dx \qquad (31)$$

which proves that we are sampling in this way the probability density function f(x).

```
!       fission_spectrum.for
!       Sampling the fission spectrum
!       probability density function:
!       x(e)=0.453*exp(-1.036E)*sinh(sqrt(2.29*E))
!       Sampling using the Rejection Method
!       M. Ragheb
!       program fission
        dimension e(100),freq(100),cumulative(100)
        integer :: trials = 1000000
        real x, y, max, e, freq, score, cumulative
!       Calculate maximum value of pdf at most probable energy e= 0.73 MeV
        max = (0.453*exp(-1.036*0.73))*(exp(sqrt(2.29*0.73))-
        & exp(-sqrt(2.29*0.73)))/2.0
        write(*,*) max
        total_score = 0.0
!       Initialize frequency distribution
        do i=1,100
                freq(i)=0.0
        end do
!       Initialize energy bins in MeV/10
        do i=1,100
                xi=i
                e(i)=xi/10.0
        end do
!       open output file
        open(44, file = 'random_out')

!       Sample distribution using rejection method
        do i= 1, trials
!       Sample x coordinate from zero to 10 MeV
                call random(rr)
                x = 10.0*rr
!       Estimate pdf at sampled x coordinate
        pdf = (0.453*exp(-1.036*x))*(exp(sqrt(2.29*x))-
        & exp(-sqrt(2.29*x)))/2.0
!       Sample y coordinate
                call random(rr)
                y = max*rr
!       Compare value of pdf to y coordinate, and score
                if(y.LE.pdf) then
                        score = 1.0
                        total_score=total_score+1.0
                else if(y.GT.pdf) then
```

```
                    score = 0.0
                end if
!       Construct sampled probabibility density function
                if(x.LE.e(1))then
                 freq(1)=freq(1)+score
                end if
                do j=1,99
                    if((x.GT.e(j)).AND.(x.LE.e(j+1)))then
                    freq(j+1)=freq(j+1)+score
                    end if
                end do
        end do
!       Normalize frequency distribution
        do i=1,100
!       Regerenerate probability density function
                freq(i)=freq(i)/total_score
        end do
!       Regenerate cumulative distribution function
        cumulative(1)=freq(1)
        do i=2,100
                cumulative(i)=cumulative(i-1)+freq(i)
        end do
!       Write results to output file
        do i=1,100
                write (44,100) e(i),freq(i),cumulative(i)
                write(*,*) e(i),freq(i),cumulative(i)
        end do
100     format (3e14.8)
        end
```

Figure 12. Procedure using the rejection method for sampling the fission neutron spectrum or Watt's curve.

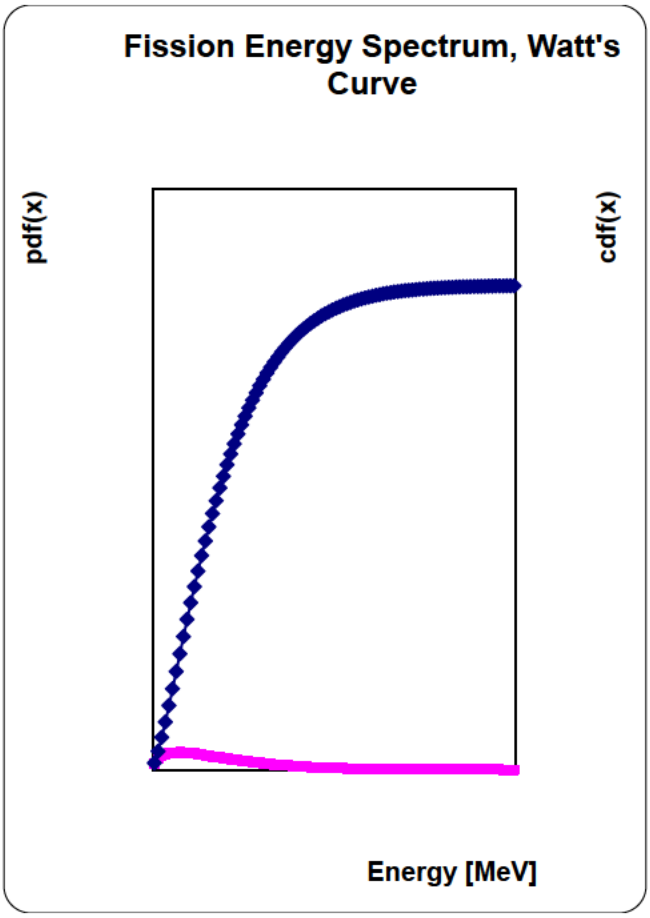

Figure 13. Probability density function and cumulative distribution function of the neutron fission spectrum: Watt's curve, using the rejection sampling method.

As an example of the use of the rejection method for sampling probability density functions that are not discrete and whose analytical expressions are difficult to invert analytically, we consider the sampling of the fission spectrum curve also designated as the Watt's curve. This curve is used in particle transport simulations to sample the energy of a neutron emerging from the fission process.

A procedure showing the application of the rejection method is displayed in Fig. 12, and uses the available expression for the fission spectrum. The result of the sampling process is displayed in Fig. 13. Both the probability density distribution function and the cumulative distribution function are shown in this graph.

12. PREVENTING BIAS IN SAMPLING

An unbiased sampling scheme can be devised to eliminate the bias from a sampling scheme that is suspected of being biased. A twist of the rejection method can be used to eliminate bias in sampling. Considering the Monte Carlo simulation of the random variable of flipping a coin:

$$\xi_{unbiased\ coin} : \begin{pmatrix} H & T \\ p(H) & p(T) \end{pmatrix} \tag{32}$$

where: H, T refer to heads and tails, respectively, p(H)=p(T)=0.5 are the probabilities of obtaining H or T for an unbiased coin.

If a suspicion exists that the coin is biased, an unbiased sampling method can be devised as follows:

1. Coin is flipped twice.
2. If two similar consecutive outcomes are obtained, HH or TT, the sample is rejected.
3. The coin is flipped twice again.
4. If two consecutive outcomes are different, the sample is accepted as H if the outcome is HT, and accepted as T for a TH outcome.
5. The probabilities of these two consecutive incomes HT or TH are the same, even if the coin is biased.

For instance let us consider the situation of a biased coin with:

$$p(H) \neq p(T) \neq 0.5 \tag{33}$$

The probability of obtaining any two consecutive outcomes a AND b is expressed by the probability theory multiplication principle:

$$p(a\ AND\ b) = p(a).p(b \mid a) = p(b).p(a \mid b) \tag{34}$$

where: $p(x \mid y)$ is the conditional probability of obtaining x given that y has occurred.

If the occurrences of x and y are independent, then:

$$p(x \mid y) = p(x) \tag{35}$$

If the two events a, b are independent, it ensues that:

$$p(a\ AND\ b) = p(a).p(b) = p(b).p(a) \tag{36}$$

Now we consider the random variable representing a biased coin:

$$\xi_{biased\ coin} : \begin{pmatrix} H & T \\ 0.4 & 0.6 \end{pmatrix}$$

For independent occurrences of H, T, application of the multiplication principle, two consecutive values of HH, TT yield the biased results with unequal probabilities:

$$p(HH) = p(H).p(H) = 0.4\,x\,0.4 = 0.16$$
$$p(TT) = p(T).p(T) = 0.6\,x\,0.6 = 0.36$$

Whereas, two different independent consecutive occurrences of H, T yield an unbiased result with equal probabilities:

$$p(HT) = p(H).p(T) = 0.4\,x\,0.6 = 0.24$$
$$p(TH) = p(T).p(H) = 0.6\,x\,0.4 = 0.24$$

This shows that an unbiased sampling scheme can be devised to eliminate the bias from a sampling scheme that is suspected of being biased.

13. REGRESSION TO THE MEAN AND THE GAMBLER'S FALLACY

In the sampling of the coin flipping random variable, even if the sampling scheme is unbiased, some odd occurrences can occur. Similar notions exist for roulette wheels, dice, investments, sports scores, corporate earnings and weather forecasting.

A surprising number of runs of consecutive values of H or T do in fact occur. If one keeps track of the proportion of time that the number of occurrences of H exceeds or is exceeded by the number of occurrences of T, one would notice the counter intuitive observation that it is rarely close to the expected value of one half. As a ratio, coins behave nicely and in fact, the ratio of H/T tends in the limit of a large number of samples toward one half. On the other hand, in terms of absolute numbers, they behave badly, where the difference between the number of H and T occurrences could get bigger as we continue flipping the coin, and the switching between leading H to T or vice versa becomes rare.

This leads people to think that the greater the deviation from the mean, the greater the restoring force toward the mean, as if there were a spring or rubber band bringing back the observations towards the expected mean. This gambler's fallacy is the mistaken belief that because a long string of H's has occurred, it follows that there exists a greater probability for T to occur *on the next flip*, or vice versa. This suggests to a gambler that he should bet on the occurrence of H after a string of T occurrences, or bet on T after a string of H occurrences. However, the coin does not know anything about any mean or any rubber band. It does not respond to the gambler's wishes either, and cannot read his mind. If it landed as H more than T, or more T than H, this difference is as likely to grow, as it is likely to shrink.

This gambler's fallacy should be distinguished from the phenomenon of regression to the mean, which is a true phenomenon. If the coin is continued to be flipped *a large number of times*, it is more likely that the proportion of H to T will approach the mean value of 0.5 but only in the long run.

Random fluctuations appearing as clumps, runs and patterns are a characteristic of random sequences, and should be expected to occur. They have been studied, and can to some extent be evaluated and predicted.

PROBLEMS

1. Estimate the variance of the random variable representing the diffusion of a particle in an infinite homogeneous medium with macroscopic total cross section: Σ_t .

2. Modify the coin flipping procedure to display the ratio of heads to tails as a function of the number of trials N. Calculate the value after each random number generation, and discuss the patterns that you can observe.

3. Modify the procedure for the simulation of die throwing to calculate the mean value or mathematical expectation of the outcomes, and its variance. Plot the results as a function of the number of throws N.

$$mean\ value: \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{variance}: \sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \mu^2$$

Compare the sampled mean value and variance to their exact theoretical values.

4. Study the effect of the width of the bins in which the frequency of the sampled distribution is scored, on the correct representation of the sampled probability density function. For instance, in the particle diffusion simulation, take the width of the bin as one, one tenth, and one hundredth mean free path. What do you observe in these cases?

5. Compare the generated probability density functions for particle diffusion in media with different diffusion media. Take for instance, the value of the total macroscopic cross section to be:

$$\Sigma_t = 0.1, 1.0, 10.0 \ \ [cm]^{-1}.$$

6. Write a rejection sampling procedure that would sample the Maxwellian distribution for particles energy, velocity or temperature.

Check that you are using a normalized probability density function.

Plot the sampled probability density function and cumulative distribution function.