Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

SELECTING INPUT MODELS AND RANDOM VARIATE GENERATION

Russell C. H. Cheng

School of Mathematics
University of Wales College of Cardiff
Senghennydd Road
Cardiff CF2 4 AG Wales

ABSTRACT

Discrete-event simulation almost invariably makes uses of random quantities drawn from given probability distributions to model chance fluctuations. This introductory tutorial discusses the basic ideas and techniques used to obtain such random variates. The two main points addressed are how appropriate distributions should be selected to model different quantities like arrival and service times, and how the variate values should actually be generated from the selected distributions.

1 INTRODUCTION

This tutorial is based on the tutorial that I gave at WSC92 (Cheng, 1992). The main points that I gave then are unchanged, but to make this tutorial self contained I shall discuss them here again. To avoid being overly repetitive, I shall try to emphasise and focus on different aspects where possible, so that this tutorial can be taken as a supplement to the earlier one.

The first point is that random variate generation, at least in discrete-event simulation, can be readily understood provided that the user is clear about the basic statistical ideas; most specifically about what constitutes a random variable and what is its probability distribution. Moreover most computer libraries contain subroutines for generating variate values from a wide range of distributions. Thus it is not absolutely necessary for a user interested mainly in the simulation modelling to become too concerned with the technical details of variate generation. However as will be shown later, there are still potential traps and dangers in too unquestioning a use of such routines, and it is as well to have some idea of what problems can arise so that they can be recognised and handled properly.

There are two main aspects to be considered. The first is the problem of selecting appropriate distributions, or input models as they are often called, to represent the various random quantities to be used in the simulation model. This is the more interesting problem from the point of view of the practitioner; and it is also the more difficult one. The main difficulty is that there are no hard and fast rules that can be invariably followed in selecting distributions. Guidelines can be set down, but the precise steps to be followed may vary in any particular instance and these steps will need to be chosen in the light of the experience of the practitioner. As already mentioned, a fundamental requirement is that the user does need to have a clear understanding of basic facts about random variables.

The other important aspect is the mechanics of generating random variates from given distributions once these have been selected. This is usually fairly straightforward in that use can be made of known methods. Many excellent texts exist which cover the basic methods of variate generation in far greater detail than would be possible in this tutorial. I shall try to give a brief guide which will serve as a simple introduction to more advanced texts. It is worth pointing out that, as far as discrete event simulation is concerned, relatively few theoretical distributions are used in practice. The main continuous distributions are the uniform, normal, exponential, gamma, lognormal, Weibull, beta of the first and second kinds, the triangular and the inverse Gaussian. The main discrete distributions are the discrete uniform, Bernoulli, binomial, geometric and negative binomial. To these should be added those distributions which simply mimic the sampled distributions observed in past data. The idea here is that past records may exist of actual service times, say. If it is believed that the service time distribution has not changed, then an easy way of generating service times is to sample them so that they resemble such past records.

The next section reviews the basic definitions of random variables that the user should be familiar with already. Section 3 discusses how to choose input distributions, and Section 4 discusses how to generate random numbers and random variates.

Good basic references which I have found useful are Law and Kelton (1991, 2nd Ed.), Lewis and Orav (1989) and Morgan (1984). I stress below the dangers of uncritical use of random number generators. Though rather advanced to be recommended as an elementary introduction, the warnings given by l'Ecuyer (1992) are well worth taking on board.

Finally it is perhaps worth repeating that variate generation in the context of simulation in statistics is a much more varied topic than when it is confined to discrete event simulation. This is because non-standard distributions are much more likely to arise, requiring in consequence more advanced techniques. Lewis and Orav focus on such problems, and the books by Devroye (1986) and Ripley (1987) are a rich mine of information on these statistical aspects.

2 RANDOM VARIABLES

This section reviews elementary facts about random variables that a user really needs to know if he/she is to use random variates meaningfully in a simulation.

A random variable, usually denoted by X, is simply a quantity that varies with the different possible outcomes of some experiment. For instance in a single server queue, X might be the total number of customers served in a given period, or it might be the waiting time of a given customer, or it might be the time of arrival of a given customer. Notice that X can be an input quantity (one used to drive the simulation) or it can be an output quantity (one that results from the experiment). In either case X will vary according to the outcome of the experiment, and because different outcomes occur with different probabilities, this means that X will take on different values also with different probabilities. We cannot say beforehand what the outcome of the experiment will be, and so we cannot predict the precise value of X that will be observed. The only thing that we can do is to say what the probability of occurrence of any particular value will be. In effect this is the *only* question that can be asked of a random variable: what is its distribution?

The cumulative distribution function (cdf), $F(x)$, is a very convenient way of defining probabilities:

$$F(x) = P(X \le x) \quad -\infty < x < \infty \tag{1}$$

where $P(X \le x)$ means the probability that, as a result of the experiment, X takes on a value less than or equal to x. If X can take on a continuous range of values then it is called a *continuous* random variable, and $F(x)$ is written as:

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{2}$$

where $f(x)$ is called the probability density function.

The value of f at x is always positive and is a measure of the relative chance of X taking a value close to x. The larger f is the greater this chance. An important instance is the behaviour of $f(x)$ when x is large. This determines how often large values of X are likely to occur - the so called 'tail' behaviour of the distribution.

The equation

$$p = F(x) \quad -\infty < x < \infty \tag{3}$$

can be viewed in two ways. It gives us the probability (often called the p-value) that X will be less than a given x. However it can be used in its inverse form

$$x_p = F^{-1}(p) \quad 0 < p < 1 \tag{4}$$

where F^{-1} is the function inverse to F, to determine x for a given p value. The resulting x_p is called the pth quantile or percentage point. An example is the exponential distribution which has cdf $F(x) = 1 - e^{-\alpha x}$, $x > 0$, when (4) becomes

$$x_p = -\alpha^{-1}\log(1-p). \tag{5}$$

Variate generation in cases like this where the inverse function can be written in closed form turns out to be easy.

If X can only take a fixed set of prescribed values $x_0, x_1, x_2, ...$,(for example when X is the number of customers served in a given period) then it is called a *discrete* random variable. In this case (2) becomes

$$F(x) = \sum_{x_i \le x} p_i \tag{6}$$

where p_i =probability that X equals x_i. The quantile is not uniquely determinable for all p values in this case. If $p = \sum_{i=0}^{j} p_i$ for some j then a range of x-values satisfies (6). However if we set

$$x_p = x_i \quad whenever \quad F(x_{i-1}) < p \le F(x_i) \tag{7}$$

with $F(x_{-1}) = 0$, then this fixes x_p uniquely for all $0 < p < 1$ so that (7) may be regarded as the analogue of (4). An example is the Bernoulli random variable defined by $X = 1$ with probability θ, $X = 0$ otherwise. Here

$$x_p = 0 \quad if \quad 0 < p \leq 1 - \theta,$$

$$x_p = 1 \quad if \quad 1 - \theta < p \leq 1. \tag{8}$$

The above definitions extend to random samples of observations. If the sample is ordered $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, then the empirical distribution function (edf) is defined as

$$F_n(x) = 0 \quad if \quad x < x_{(1)}$$

$$F_n(x) = \frac{i}{n} \quad if \quad x_{(i)} \leq x < x_{(i+1)}$$

$$F_n(x) = 1 \quad if \quad x_{(n)} \leq x \tag{9}$$

and this is the analogue of (6). The analogue of (7) is

$$x_p = x_{(i)} \quad if \quad \frac{i-1}{n} < p \leq \frac{i}{n}. \tag{10}$$

3 SELECTING INPUT MODELS

3.1 Input models

Input models or distributions are the probability distributions of random variables used to drive the simulation. As already remarked, the selection of appropriate distributions to represent quantities like interarrival times between customers and service times of customers, as far as model building is concerned, is the more interesting as well as difficult step, compared with their actual generation.

There are four main cases. Consider first the situation where substantial data already exists recording a quantity of interest such as a set of service times. This might be available from past records, or it might be gathered specifically for the simulation. This data can then be used directly in the simulation. An example is the simulation of electricity demand. This is known to depend heavily on daily air temperature. The observed daily temperature from past records can be used in the simulation. This is clearly a good approach especially if a period of observations is selected containing temperatures of special interest to the investigation, like periods of unusually low temperatures, say. This method is useful if comparison is to be made of the performance of different versions of a system.

The second case is when available data is sparse, but the user still wishes to make use of it in the simulation. Here the so called 'boot-strap' method can be used. The data is used to define the edf (9), and this is treated as being the population distribution. Variate values can then be generated using the inverse transform described in Section 4.2 below.

The third possibility is to fit a theoretical distribution to the data. This can be done whether the data is abundant or sparse. A distribution is selected that is fixed apart from a few unknown parameters. These parameters are then estimated from the data using some appropriate statistical estimation technique. An advantage of this method is that the theoretical distribution can be selected to have characteristics known to be possessed by the distributions to be modelled. For example, experience shows that many service time distributions are positively skewed, so it makes sense to select a theoretical distribution that is known to be positively skewed.

Using a theoretical distribution whose parameters have been estimated raises the question of how inaccuracies in the estimates affects the results. One possibility is to find confidence intervals for the unknown true parameter values and then carry out simulations at the upper and lower limits of these confidence intervals.

Finally, this last technique can be extended to situations where no past data is available at all. This situation occurs if a hypothetical system is being analysed. Theoretical distributions should be selected with enough flexibility to model the range of behaviour that might occur. Simulations can then be run at different parameter value settings. The appropriate way of reporting results in this case, is to make *conditional* statements: '*if* the conditions are like this, *then* the system behaves like this'.

The two main aspects of fitting distributions are how to estimate parameters and how to assess the accuracy of the estimates, and we discuss them a little more fully.

3.2 Fitting Distributions

Suppose we have selected some particular type of distribution to be an input model. We can denote its pdf by $f(x, \theta)$, where θ represents a vector of unknown parameters. A specific instance is the Weibull distribution with density $f(x, \theta) = \alpha\beta^{-\alpha}x^{\alpha-1}\exp(-(x/\beta)^{\alpha})$, where $\theta = (\alpha, \beta)$. This is a useful distribution for representing failure times in that it can be skewed in either direction depending on the value of α. We wish to estimate θ using a sample of observed failure times: $x_1, x_2, \ldots, x_n$. An extremely powerful way of doing this is the method of *maximum likelihood* (ml). In many well-known cases it is equivalent to the more elementary least-squares technique, but it is really more general. The likelihood is simply the product of the pdf's evaluated at the observed values, and then treated as a function of θ. It is usually easier to work with the logarithm of the likelihood (loglikelihood):

$$L(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta)$$

The maximum likelihood estimate, $\hat{\theta}$, is simply the value of θ which maximizes the loglikelihood. In some cases an explicit formula for $\hat{\theta}$ can be obtained by setting the derivative of $L(\theta)$ to zero. However I find that it is usually easier to use a general search procedure that evaluates $L(\theta)$ at different θ and then progressively improves the value of L based on comparisons of these different values.

Under general conditions the ml estimate has the important property that its distribution is asymptotically normal as the sample size n becomes large. This allows confidence intervals to be constructed which contain the unknown true parameters with prescribed degree of confidence. Various different ways can be used to do this. I find that the most reliable is based on the result that

$$2(L(\hat{\theta})-L(\theta_{true})) \qquad (11)$$

is approximately chi-squared with p degrees of freedom, where p is the number of unknown parameters being fitted. This result means that a given θ is unlikely to be the true one if we find that the difference $2(L(\hat{\theta})-L(\theta))$ is unusually large compared with a chi-squared variate with p degrees of freedom. We can thus form a $\alpha 100\%$ confidence region for θ_{true} by taking all those values of θ for which (11) is less than $\chi^2_p(\alpha)$, this last quantity being the upper α quantile of the chi-squared distribution with p degrees of freedom.

3.3 Goodness of Fit Tests

Law and Kelton give helpful descriptions of the general shape of different pdfs and in what applications they might be useful. Even if carefully chosen, when fitting a theoretical distribution to observed samples, it is usually advisable to check that the selected distribution matches the general characteristics of the sample.

The simplest non-technical methods are graphical ones. These compare the fitted and empirical distribution functions by plotting one against the other. These are known as quantile-quantile (Q-Q) or probability-probability (P-P) plots, depending on the precise quantities being compared. In either case if the two functions are similar in shape the resulting plot will be a 45^0 straight line. Differences show up as deviations from this straight line. See Law and Kelton or Lewis and Orav for more details.

A more formal method is to use a so-called goodness-of-fit test. A good review of the many such tests available is given by D'Agostino and Stephens (1986). The most well-known tests are the chi-squared and the Kolmogorov-Smirnov tests. The former is in effect a comparison of observed and fitted densities and is usually easy to apply. Moreover the test makes an allowance for the case when parameters of the theoretical distribution have been fitted. The latter compares the difference between the empirical and fitted cdf's. Though simple it tends to be rather sensitive to differences in the tails. It is also less easy to allow for the when parameters are estimated than the chi-squared goodness-of-fit test.

The Anderson-Darling test also compares cdf's, but it gives particular weight to differences in the tails. It is a sensitive test and easy to apply. However as with the Kolmogorov-Smirnov test it is not so easy to make allowance for the situation when parameters are estimated.

4 RANDOM VARIATE GENERATION

4.1 Random Numbers

The term random number is always used to mean a random variate uniformly distributed in the interval (0,1), thus it has pdf $f(x) = 1$ if $0 < x < 1$, $f(x) = 0$ otherwise. In computer simulations random variates from other distrbutions are always generated by taking random numbers and then converting these in some manner to the required variate. Thus random numbers play a fundamental role in random variate generation, and methods for producing them has been the subject of extensive study and review. Random numbers will be denoted by U , U_1 or U_2.

Most computer library routines contain what are called pseudo random number generators. These produce a sequence of numbers (lying between 0 and 1) using a fixed formula, but which give the appearance of being true random numbers. The definition of true randomness is difficult and this is reflected in the fact that there is no obvious single test that can be applied to check if a particular pseudo random number generator is always satisfactory or not. In fact, as such a generator is based on a deterministic formula, it is clear that one can *always* come up with a test (simply the formula itself!) that will predict the next number and which therefore shows that the numbers are not truly random. The big danger is that, in some application, one is unwittingly transforms the numbers in such a way that 'unravels' the generator so that the quantities produced are very obviously non-random. Ripley (1987) and Lewis and Orav (1988) give examples which show that these problems are not just mathematical niceties which can be ignored in practice. My advice would be to follow the suggestion of Ripley (1990) and check the following key features of a generator before relying on it too heavily.

One obviously desirable property is that the generator should produce numbers with little computational effort, i.e. it should be fast.

A good test of randomness is to treat consecutive blocks of k-numbers produced by a generator as being a k-dimensional point. Such points should be uniformly and independently distributed in the k-dimensional hypercube. There is clear evidence (see for example l'Ecuyer, 1992) that most generators produce points that are too uniformly distributed, and that insufficient chance clustering occurs. Moreover in high dimensions the points no longer appear to be independent.

All pseudo random number generators produce a sequence of numbers which ultimately repeat. The number of values that one gets before the numbers start to repeat is called the *period* of the generator. To be useful the period has to be large. The first most widely used generators were of the linear congruential type. Such generators have periods that are of the order of several thousand million. With the increased power of modern computers simulations to use up all the numbers of such a sequence very quickly. The most modern generators, particularly those of bit-shift register type, produce sequences which are several orders of magnitude longer and should probably now be preferred.

If there is any doubt about the performance of a generator - and my advice based on personal experience is to be very skeptical of library generators - a routine should be used that one has selected oneself, taking into account the above considerations.

Assuming that a pseudo random number generator is available, we now consider how it can be used to produce variates from other distributions. There are only three methods of any generality, and we consider each in turn.

4.2 The Inverse Transform Method

This method can be applied to both discrete and continuous variables. Equations (4) and (7) give the pth quantile x_p in terms of the probability value p. If p is set equal to a random number, U, then x_p becomes a random variable and its cdf is

$$\begin{aligned} P(X \leq x) &= P(F^{-1}(U) \leq x) \qquad (from(4)) \\ &= P(U \leq F(x)) \\ &= F(x). \end{aligned} \tag{12}$$

Thus X produced in this way has exactly the distribution that we want. In the case of the Weibull distribution this gives the explicit formula

$$X = \beta^{-1}[-\log(1-U)]^{1/\alpha}$$

Note that the special case $\alpha = 1$, corresponding to equation (5), gives an exponential random variable. Note also that $(1-U)$ can be replaced by U as both have the same $U(0,1)$ distribution.

The same result works for a discrete variable using (7) instead of (4). The Bernoulli case (8) gives:

$$X = 0 \qquad if \quad 0 < U < 1-\theta$$

$$X = 1 \qquad if \quad 1-\theta < U < 1$$

A very useful version of this result allows sampling from an observed data set. Clearly equation (10) is the sample analogue of (7), so use of (10) with p set equal to a $U(0,1)$ random number is all that is required.

4.3 Composition Method

If the cdf of interest can be written as

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

where $\{p_j\}$ is a discrete probability distribution and the $F_j(x)$ are cdf's, then $F(x)$ is called a mixture distribution. The mixture can arise naturally in the context of the problem or it can simply be an artificially constructed decomposition. In either case X can be generated with cdf $F(x)$ by selecting the jth distribution with probability p_j, generating a variate X_j with distribution F_j, and setting $X = X_j$.

Usually one tries make F_1 easy to generate from, whilst at the same time making p_1 as large as possible, so that F_1 is chosen frequently. For instance in the Marsaglia and Bray (1964) 'convenient' method of generating normal variates over 86% of the time a linear combination of three uniforms is taken, the remainder of the time more elaborate cdf's have to be used.

4.4 Acceptance-Rejection Method

This is most easily applied to continuous variables though ingenious methods exist for certain discrete cases. Suppose that (i) we wish to generate variates with pdf $f(x)$, (ii) we have a method of generating variates with pdf $g(x)$, and (iii) we can find a scaling factor K for which

$$e(x) = Kg(x) > f(x) \quad all \quad x.$$

Thus $e(x)$ is an envelope whose graph lies completely above the graph of $f(x)$. If we generate X so that it has pdf $g(x)$ and take $Y = Ue(X)$ where U is a random number independent of X, then (X, Y), treated as a point, will always lie underneath the graph of $e(x)$ and in fact it is uniformly distributed in the area under this graph. If now we generate such points (X, Y) and discard those lying above the curve $f(x)$, then the remaining points will have the property that the number with a given x value will be proportional to $f(x)$. Thus the sequence of X's corresponding to these accepted (X, Y) points will come from the distribution with the required pdf $f(x)$.

The efficiency of the method is dependent on the factor K. In fact K gives the average number of points (X, Y) needed for each X actually accepted.

An interesting example is the gamma distribution with pdf $f(x) = x^{\alpha-1}e^{-x}/\Gamma(\alpha)$, $x > 0$. Many acceptance-rejection methods have been suggested for this distribution. The method given by Fishman (1976), valid for $\alpha > 1$, is one of the easiest to implement. It uses a (negative) exponential envelope. In this case K is proportional to $\alpha^{\frac{1}{2}}$, and the method becomes increasingly inefficient as α increases; but it is effective if α is less than 3, say. The method GB using a log-logistic envelope which I suggested (Cheng, 1977) has $K \leq 1.47$ for all $\alpha \geq 1$; so it is satisfactory for larger α.

4.5 Discrete Distributions

Various methods that we have discussed work well for discrete distributions. Indeed we have already mentioned that the method based on (10) is precisely the inverse transform method applied to an empirical (discrete) distribution. Because of the way it operates in practice, it is sometimes called the table-look-up method.

A very neat and powerful general method for discrete distributions with a finite range is the *alias* method introduced by Walker(1977) and improved by Kronmal and Peterson (1979). A modified acceptance technique is used which tests a uniform variable. Depending on the outcome one or other of two variate values is produced so that no rejection ever occurs. A good description is given by Law and Kelton.

5 FINAL COMMENTS

Library routines usually implement fast, accurate methods with compactness considerations of secondary importance. A point worth bearing in mind is that many methods require the setting up of certain constants which actually depend on parameter values of the distribution. If such 'constant' values need changing for every variate produced then the cost of setting them up needs to be taken into account in assessing variate generation speed. A more important point is that, in complex simulations, the cost of generating variates is usually very small compared with the rest of the simulation. Thus speed is usually not of first importance. In this respect simple compact methods are very tempting for the user building 'one-off' simulation models, especially if they are reasonably efficient.

It should be realised that, though the general methods that we have considered are reasonably useful, many of the best techniques do make use of special properties of a distribution to achieve maximum speed and compactness. For instance a triangular distribution can be obtained as the sum of two uniforms, a beta variate can be generated as a ratio involving two gamma variates. A celebrated instance of this is the Box-Muller method of generating normal variates in pairs.

REFERENCES

Bratley, P., Fox, B.L. and Schrage, L.E. 1983. *A Guide to Simulation*. New York: Springer-Verlag.

Cheng, R.C.H. 1977 The generation of gamma variables with non-integral shape parameter, *Applied Statistics*, **26**, 71-75

Cheng, R.C.H. 1992. Distribution fitting and random number and random variate generation. In *Proceedings of the 1992 Winter Simulation Conference* (ed. J.J. Swain, D. Goldsman, R.C. Crain and J.R. Wilson), IEEE Press 74-81.

Devroye, L. 1986. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Fishman, G.S. 1976. Sampling from the gamma distribution on a computer. *Communications of the ACM*, **19**, 407-409.

Kronmal, R.A. and Peterson, A.V.Jr. 1979. On the alias method for generating random variables from a discrete distribution, *American Statistician*, **33**, 214-218.

Law, A.M. and Kelton, W.D. 1991. *Simulation Modeling and Analysis 2nd Edition.* New York: McGraw-Hill.

L'Ecuyer, P. 1992. Testing random number generators. In *Proceedings of the 1992 Winter Simulation Conference* (ed. J.J. Swain, D. Goldsman, R.C. Crain and J.R. Wilson), IEEE Press 305-313.

Lewis, P.A.W. and Orav, E.J. 1989. *Simulation Methodology for Statisticians, Operations Analysts and Engineers*, Vol. 1, Pacific Grove: Wadsworth and Brooks/Cole.

Marsaglia, G. and Bray, T.A. 1964. A convenient method for generating normal variables. *SIAM Review*, **6**, 260-264.

Morgan, B.J.T. 1984. *The Elements of Simulation.* London: Chapman and Hall.

Ripley, B.D. 1987. *Stochastic Simulation.* New York: Wiley.

Ripley, B.D. 1990. Thoughts on pseudorandom number generators. *Journal of Computational and Applied Mathematics*, **31**, 153-163.

Walker, A.J. 1977. An efficient method of generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software*, **3**, 253-256.

AUTHOR BIOGRAPHY

Russell C.H. Cheng is Reader in Statistics at the University of Wales, College of Cardiff. He obtained a B.A. in 1968, and the Diploma in Mathematical Statistics in 1969, from Cambridge University, England. He obtained his Ph.D. in 1972 from Bath University working on computer models of industrial chemical plants with ICI. He joined the Mathematics Department of the University of Wales Institute of Science and Technology in 1972 and was appointed Reader in 1988. He is General Secretary of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society and President of the Cardiff Branch of the Mathematical Association. His research interests include: variance reduction methods and parametric estimation methods.