Stochastic Simulation

BRIAN D. RIPLEY

Professor of Statistics
University of Strathclyde
Glasgow, Scotland

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

This Page Intentionally Left Blank

Stochastic Simulation

This Page Intentionally Left Blank

Stochastic Simulation

BRIAN D. RIPLEY

Professor of Statistics
University of Strathclyde
Glasgow, Scotland

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:

Ripley, Brian D., 1952-
Stochastic simulation.

(Wiley series in probability and mathematical statistics. Applied probability and statistics, ISSN 0271-6356)
Includes index.
1. Digital computer simulation. 2. Stochastic processes. I. Title. II. Series.

QA76.9.C65R57 1987 001.4'34 86-15728
ISBN 0-471-81884-4

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

This book is intended for statisticians, operations researchers, and all those who use simulation in their work and need a comprehensive guide to the current state of knowledge about simulation methods. Stochastic simulation has developed rapidly in the last decade, and much of the folklore about the subject is outdated or fallacious. This is indeed a subject in which "a little knowledge is a dangerous thing!" Although this *is* a comprehensive guide, most of the chapters contain explicit recommendations of methods and algorithms. (To encourage their use, Appendix B contains a selection of computer programs.) Thus, this book can also serve as an introduction, and no prior knowledge of the subject is assumed.

Simulation is one of the easiest things one can do with a stochastic model, which may help to explain its popularity. Although easy to perform, some of the "tricks" used are subtle, and the analysis of what has been done can be much more complicated than is apparent at first sight. Simulation is best regarded as mathematical experimentation, and needs all the care and planning that are regarded as a normal part of training in experimental sciences. The general mathematical level of this book is elementary, involving no more than a first course in probability and statistics. A notable exception is those parts of Chapter 2 that deal with the theoretical behavior of random-number generators, which contain a number of applications of number theory. All the necessary mathematics is developed there, but some prior knowledge of pure mathematics will help a great deal. Random-number generators are so fundamental that the reader should eventually tackle Chapter 2 unless he or she is *sure* that all the generators he or she uses are adequate (that is, have been checked by someone who understands that chapter). It might be disastrous to believe in your computer manufacturer!

Chapters 3 and 4 cover drawing realizations from standard probability distributions and stochastic processes. The emphasis is on methods that are easy to program (compact and with a simple logic, therefore easy to check). These are particularly suitable for personal computers. A small number of workers have specialized in developing faster and increasingly

more complex algorithms. These are referenced but, in general, not described in detail. The coverage of *methods* was comprehensive at the time of writing.

Even statisticians often fail to treat simulations seriously as experiments. Even more is possible in the way of design since the randomness was introduced by the experimenter and hence is under his or her complete control. Such techniques are described in Chapter 5 under the heading of "variance reduction." A general knowledge of the statistical design of experiments is helpful here and essential to a competent practitioner of simulation. The analysis of the output of many simulation experiments, for example queueing systems, is also more complicated than many users suppose, although not as difficult as the literature makes out! This topic is discussed in Chapter 6.

Chapter 7 discusses many novel uses of simulation. It can be used, for example, in optimizing designs of integrated circuits and in fundamentally new ideas in statistical inference.

The literature on simulation is vast, and I have made no attempt to cite comprehensively. There are several published bibliographies, but a lot of the work has been superseded or is misleading.

The exercises vary considerably in difficulty. Some are routine exercises in developing algorithms from general theory or in providing illustrative examples. Others are of an open-ended nature; they suggest experiments to be done and demand access to a computer (although the humblest personal computer would suffice).

Simulation has long been a cinderella subject, particularly in statistics. I hope this book shows that it raises fascinating mathematical and statistical problems that demand attention.

BRIAN D. RIPLEY

Glasgow
October 1986

Acknowledgments

I am indebted to everyone who has taught me about simulation or has been prepared to share their experiences with me, in particular, Anthony Atkinson and Luc Devroye. The manuscript was typed with great efficiency by Lynne Westwood. The figures were produced on equipment funded by the Science and Engineering Research Council.

B.D.R.

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Stochastic Simulation

This Page Intentionally Left Blank

CHAPTER 1

Aims of Simulation

The terminology of our subject can be confusing, with some authors insisting on shades of meaning that do not have widespread agreement. A dictionary definition of "to simulate" is

> Feign, . . . , pretend to be, act like, resemble, wear the guise of, mimic, . . . imitate conditions of (situation etc.) with model, for convenience or training . . . .
>
> *Concise Oxford Dictionary*, 1976 ed.

In everyday usage "simulated" has a derogatory ring, but the value of simulators in training pilots is also recognized. In its technical sense simulation involves using a model to produce results, rather than experiment with the real system under study (which may not yet exist). For example, simulation is used to the explore the extraction of oil from an oil reserve. If the model has a stochastic element, we have *stochastic simulation*, the subject of this monograph.

Another term, the *Monte-Carlo method*, arose during World War II for stochastic simulations of models of atomic collisions (branching processes). Sometimes it is used synonymously with stochastic simulation, but sometimes it carries a more specialized meaning of "doing something clever and stochastic with simulation." This may involve simulating a different system from that under study, perhaps even using a stochastic model for a deterministic system (as in Monte-Carlo integration). We will not use Monte Carlo except in the conventional terms "Monte-Carlo integration" and "Monte-Carlo test."

Simulation can have many aims, which makes it impossible to give universal guidelines to good practice. Tocher (1963) wrote one of the first texts on the subject. His title was *The Art of Simulation*, and simulation is still an art despite a much greater understanding of the simulator's toolkit. The aim of this volume is to display those tools in their most useful form with guidance about their use.

1.1. THE TOOLS

The first thing needed for a stochastic simulation is a source of randomness. This is often taken for granted but is of fundamental importance. Regrettably many of the so-called random functions supplied with the most widespread computers are far from random, and many simulation studies have been invalidated as a consequence.

Digital computers cannot easily be interfaced to a truly random phenomenon such as the electronic noise in a diode. All random functions in common use are in fact pseudo-random, which is to say that they are deterministic, but mimic the properties of a sequence of independent uniformly distributed random variables. Their essence is unpredictability. Consider for example the following sequence

$$13, 8, 1, 2, 11, 14, 7, 12, 13, 12, 17, 2, 11, 10, 3, \ldots$$

It is generated by a simple deterministic rule, but no one had guessed what the rule was or what the next number is at the time of writing. (Exercise 1.1 will give the game away, but try to guess first.) The algorithms commonly used are similar, and much mathematical analysis has gone into the question of how well they do mimic a random sequence.

Only occasionally does one want independent, uniformly distributed random variables. However, they are a useful source of randomness that can be turned into anything else. Chapters 3 and 4 consider tools to make samples of all the standard distributions and stochastic processes from this source of randomness.

Simulation for us is about sampling from stochastic models. Too much emphasis has been placed in the literature on producing the samples and too little on what is done with those samples. Any stochastic simulation involves observing a random phenomenon and so is a statistical experiment. Statisticians, even experts in the design of experiments, are notoriously bad at designing their own experiments! There is even more scope for designing a simulation experiment than a real one, for the randomness and the model are under our complete control. Thus techniques for the design and analysis of simulation experiments are important tools and still an under-researched area.

1.2. MODELS

A stochastic simulation is of a *model*, and the aims of simulation are closely connected to those of modeling. So, why model? Within the scope of statistics and operations research we can usefully identify two principal

reasons:

1. *To summarize data.* A very common example is the general linear model of statistics as used in regression and the analysis of variance.
2. *To predict observations.* A regression equation can be used to predict a response under new conditions or to find a combination of control variables giving an optimum response. This "what if" use of models is the basis of much of operations research.

It is also useful to consider two classes of a model. Models can either be *mechanistic* or *convenient*. For example, the general linear model is merely convenient whereas the models of genetics are thought to represent the actual mechanisms. The models of the physical world used by engineers are usually both deterministic and mechanistic, whereas most stochastic models are convenient. Either type of model can be used to help understand, to predict, or to aid decision-making. An example of the latter is the "convenient" models of errors in agricultural field trials which are used to help disentangle the true differences in fertility of plant varieties from the fertilities of the plots in which they were grown.

To make use of a model one has two choices:

1. To bring mathematical analysis to bear to try to understand the model's behavior. This is very easy for a general linear model but nigh impossible for a complex queueing system or for the equations of fluid flow in a complex structure such as a rock. The work involved is usually laborious (although if one is lucky it may already have been done). There are also likely to be necessary approximations and questionable assumptions.
2. To experiment with the model. For a stochastic model the response will vary, and we will want to create a number of *realizations* (sets of artificial data) for each set of parameters.

Sometimes one of these choices may be unfruitful. We might not be able to make progress by analytical means or might not have the resources to simulate the model. (It is almost always possible to simulate a well-defined model given sufficient resources.)

The choice of analysis or simulation will depend on the purpose of modeling. Simulation is good at answering specific "what if" questions whereas analysis almost always deepens understanding of the model. One neglected use of simulation is a hybrid approach: do a simulation experiment, analyze it to produce a "convenient" model, and use *this* model for predictions and decisions.

The cost analysis is rapidly tilting in favor of simulation as computer time becomes ever cheaper and mathematicians remain scarce. It may be incredible to younger readers that Cox and Smith (1961) reported a simulation

performed with the aid of a slide rule (a mechanical device to perform multiplications and evaluate standard functions) and a table of random numbers. Nowadays (1984/5) desktop computers are further revolutionizing the ease of mathematical experimentation.

1.3. SIMULATION AS EXPERIMENTATION

We have stressed that simulation is experimental mathematics and that simulation studies should be designed carefully, a process often termed *variance reduction* in this field. Their classification as experiments also has repercussions for the reporting of simulation studies. It is essential that enough details are given for the experiments to be repeated and the results checked. Hoaglin and Andrews (1975) gave some standards on reporting which seem to have been followed only exceptionally. In view of the preceding warnings on the deficiencies of certain pseudo-random-number generators, it is important to report the generator used.

Good design is the key to reducing the cost of the study when this is necessary. The cost of generating random variables and sampling from stochastic models is usually a tiny part of the cost of the study, so the main aim should be to make best use of a small number of replications.

The analysis of simulation experiments also needs care, because the observations may not be independent. This can either occur deliberately as part of the design or because one is simulating a stochastic process through time. (The problems of analyzing observations of a simulated stochastic process apply equally to observing real processes, but this is done much less intensely.) Chapter 6 considers various ways to include dependence in the analysis or to select independent sets of observations.

1.4. SIMULATION IN INFERENCE

Simulation has recently become popular as part of statistical inference. The advantages are again the need to make fewer approximations, although interpretation may be more difficult. Monte-Carlo tests compare the data with simulated data from the supposed model. The similarity of real and simulated data provides a test of goodness-of-fit. Bootstrap methods resample from the data, using the data as a reference distribution to assess the variability or bias of an estimator. Both are discussed in Chapter 7.

1.5. EXAMPLES

Checking Distribution Theory

"Student" (1908) when deriving his t distribution carried out a small simulation experiment. He had 3000 physical measurements on humans which were known to be approximately normally distributed. These were shuffled and divided into 750 sets of (X_1, X_2, X_3, X_4). From each sample of size four the t statistic was calculated, giving 750 realizations to compare with the theoretical density. (This was done for each of two measurements.)

We can repeat this experiment with very much less effort. Figure 1.1 shows a simple BASIC program to do so. The 750 numbers can be compared with a t distribution in any way we choose. Perhaps the simplest thing to do is to compare some moments with their population values. Each run of this program on a BBC microcomputer took 130 sec. (Appendix A gives details of the computers used in this work.)

Simulation is often useful to check theoretical calculations. For example, the author was asked to check the solution to Sylvester's problem (Kendall

```
10 DIM X(4)
20 FOR I%=1 TO 750
30 FOR J%=1 TO 3 STEP 2
40 U=2*RND(1)-1
50 V=2*RND(1)-1
60 W=U*U+V*V
70 IF W>1 THEN 40
80 C=SQR((-2*LN(W))/W)
90 X(J%)=C*U
100 X(J%+1)=C*V
110 NEXT J%
120 SUM=0
130 FOR J%=1 TO 4
140 SUM=SUM+X(J%)
150 NEXT J%
160 XBAR=SUM/4
170 SUM=0
180 FOR J%=1 TO 4
190 SUM=SUM+(X(J%)-XBAR)^2
200 NEXT J%
210 S=SQR(SUM/3)
220 T=SQR(4)*XBAR/S
230 PRINT T
240 NEXT I%
```

Figure 1.1. A BASIC program to repeat Student's simulations. The function RND(1) returns a pseudo-random number. Lines 40 to 100 code algorithm 3.9 to produce normal variates.

and Moran, 1963; Solomon, 1978). Four points are placed at random in a disc and their convex hull found. What is the probability that it is a triangle? The theoretical value is $35/12\pi^2$. A simulation study was performed with 100,000 replications. In 29,432 cases the convex hull was a triangle, giving a 95% confidence interval for the probability of (0.2915, 0.2971) and confirming the theoretical value, 0.2955. The whole study took half an hour, using a VAX11/782 (including programming).

Much of statistical practice is based on asymptotic distributions, and simulation is much used to check the accuracy of asymptotic results for small samples. Ripley and Silverman (1978) considered the distribution of d, the smallest distance between any pair of n random points in the unit square. Their asymptotic result is that $n(n-1)d^2$ has an exponential distribution with mean $2/\pi$ (see also Theorem 2.6). Large values of d provide the rejection region of a test of inhibition between points, so we will count the number of values of $T = n(n-1)d^2 \geqslant 1.907$, the 95% point of the asymptotic distribution. Figure 1.2 shows the program and Table 1.1 gives the results. The count has a binomial (10,000, 0.05) distribution on the asymptotic theory, so the acceptance region of a 5% test is (457, 543) (using a normal approximation). Thus our experiment gives us no reason to doubt the asymptotic theory even for sample sizes as small as $n = 10$.

```
 10 INPUT "N", N%
 20 DIM X(N%), Y(N%)
 30 INPUT "Reps", R%
 40 CNT = 0
 50 DC = 1.907/(N%*(N% - 1))
 60 FOR L% = 1 TO R%
 70 FOR I% = 1 TO N%
 80 X(I%) = RND(1)
 90 Y(I%) = RND(1)
100 NEXT I%
110 D = 2
120 FOR I% = 2 TO N%
130 X1 = X(I%): Y1 = Y(I%)
140 FOR J% = 1 TO I% - 1
150 DD = (X1 - X(J%))^2 + (Y1 - Y(J%)))^2
160 IF DD < D THEN D = DD
170 NEXT J%, I%
180 IF D > DC THEN CNT = CNT + 1
190 NEXT I%
200 PRINT "Count="; CNT
```

Figure 1.2. BASIC program to check exponential distribution for $n(n-1)d^2$.

Table 1.1. Results from Figure 1.2

n	CNT	out of	R%	Time (min)
10	516		10,000	103
15	516		10,000	227
20	509		10,000	405

This experiment was run overnight on a personal computer and so was free. Nevertheless we should still consider whether we could have obtained more information from the experiment. [In fact we only used the fact that at least one or no pairs (x, y) had $n(n-1)d(x, y) < 1.907$, so we could have stopped searching as soon as one was found.] Clearly we could have checked other percentage points with the same data. Could we make use of the actual values of T? One possibility is to assume that the tail of the distribution of T is exponential of unknown mean λ^{-1}, and to estimate $P(T > 1.907) = e^{-1.907\hat{\lambda}}$ for an estimate $\hat{\lambda}$ of λ, say obtained from the observations with $T > 1$. Exercise 1.4 shows that this idea is worthwhile only in the extreme tail.

Comparing Estimators

Andrews et al. (1972) report a large simulation experiment that used variance reduction very effectively. Consider a location-parameter estimation problem:

$$\text{Estimate } \theta \text{ in } \{ f(x-\theta) \mid x \in \mathbb{R} \} \quad \text{from } x_1, \ldots, x_n$$

The density f is symmetric and is similar to the normal density. The idea is to find estimators that perform well across a wide class of possible densities f. Some obvious estimators of θ are the sample mean and the sample median, and a trimmed mean (the mean of all except the r largest and r smallest values). Let $T(x)$ be such an estimator. All the estimators considered were location equivariant ($x_i \to x_i + c$ implies $T \to T + c$) and many were scale equivariant ($x_i \to sx_i$ implies $T \to sT$). Our examples are both location and scale equivariant.

The key to the variance reduction was that all simulations were done for f belonging to the so-called normal/independent family. That is, f is the density of $X = Z/S$, where $Z \sim N(0, 1)$ and $S > 0$ is independent of Z. Consider first conditioning on $S_1 = s_1, \ldots, S_n = s_n$. Then $X_i \sim N(0, 1/s_i^2)$,

and suitable statistics for the $\{X_i\}$ are $\hat{X}$ and $\hat{S}$ where

$$\hat{X} = \frac{\sum X_i s_i^2}{\sum s_i^2}$$

$$S^2 = \frac{\sum (X_i - \hat{X})^2 s_i^2}{n-1}$$

Define $C_i = (X_i - \hat{X})/S$. Then for a location and scale equivariant estimator T,

$$T(\mathbf{x}) = \hat{x} + sT(\mathbf{c})$$

The point here is that $T(\mathbf{c})$ is much less variable than $T(\mathbf{x})$. We will assume T is unbiased, so $E_\theta(T) = \theta$. Consider

$$E[(T-\theta)^2 | \mathbf{C} = \mathbf{c}, \mathbf{S} = \mathbf{s}] = v(\mathbf{c}, \mathbf{s})$$

say, so the expectation is merely over the location and scale of the sample. Conditionally, $\hat{X}$ and S are independent, with $\hat{X} \sim N(\theta, 1/\Sigma s_i^2)$ and $(n-1)S^2 \sim \chi^2_{n-1}$. Thus

$$\begin{aligned} v(\mathbf{c}, \mathbf{s}) &= E\{\hat{X} - \theta + ST(\mathbf{c})\}^2 \\ &= E(\hat{X} - \theta)^2 + E(\hat{X} - \theta)ST(\mathbf{c}) + E(S^2)T(\mathbf{c})^2 \\ &= \frac{1}{\sum s_i^2} + T(\mathbf{c})^2 \end{aligned}$$

where all expectations are conditional on $\mathbf{C} = \mathbf{c}, \mathbf{S} = \mathbf{s}$. Finally,

$$\text{var}(T) = E\left[\frac{1}{\sum S_i^2} + T(\mathbf{C})^2\right]$$

and this is found by a simulation experiment as an average over many samples $(X_1, \ldots, X_n)$ of the random variables. Almost no more work is needed than in calculating $T(\mathbf{X})$, but the estimate of var(T) obtained is much more accurate (see Table 1.2).

The essence of this transformation is to average analytically over as much of the variation as possible. The assumption on f is slightly restrictive, but includes Student's t distribution as well as the Cauchy, Laplace, and con-

Table 1.2. Estimates of $n \times \text{var}(T)$ Based on 200 Replications for Sample Size $n = 25$ for the Mean, Median, and Trimmed Mean ($r = 2$) Estimators T

	α				
	1.5	2	5	10	100
Mean					
Average	1.57	1.53	1.185	1.096	1.009
s.e.1[a]	0.14	0.18	0.10	0.095	0.084
s.e.2[a]	0.048	0.062	0.019	0.010	0.0017
Variance reduction	9	8	27	90	2,400
Median					
Average	2.17	1.96	1.67	1.55	1.60
s.e.1	0.19	0.23	0.14	0.14	0.11
s.e.2	0.11	0.093	0.064	0.052	0.051
Variance reduction	3	6	5	7	5
Trimmed mean					
Average	1.72	1.61	1.22	1.14	1.051
s.e.1	0.18	0.16	0.080	0.097	0.084
s.e.2	0.065	0.065	0.022	0.015	0.0052
Variance reduction	7	6	12	40	260

[a]The s.e.1 and s.e.2 are standard errors from direct estimation and conditional estimation. The distribution of S_i^2 was $\alpha^{-1} \times \text{gamma}(\alpha)$, so $X_i \sim t_{2\alpha}$.

taminated normal distributions. It is a small price to pay for a six-fold reduction in experimental replication. It should be stressed that negligible extra work is involved. Instead of for each replication

1. Sample $Z_1, \ldots, Z_n \sim N(0, 1)$, $S_1, \ldots, S_n$, set $X_i = Z_i/S_i$
2. Form $V = T(\mathbf{X})^2$

and averaging V, we

1. Sample $Z_1, \ldots, Z_n \sim N(0, 1)$, $S_1, \ldots, S_n$, set $X_i = Z_i/S_i$
2. Calculate $\hat{X}$, $\hat{S}$
3. Form $V = 1/\sum S_i^2 + \{T(\mathbf{X}) - \hat{X}\}^2/S^2$

and average V. The variance reduction is most when $T(\mathbf{c})$ is most nearly constant, but is always worthwhile.