Proceedings of the 1987 Winter Simulation Conference
A. Thesen, H. Grant, W. David Kelton (eds.)

MODELING AND GENERATING INPUT PROCESSES

Mark E. Johnson
Statistics Group, MS F600
Los Alamos National Laboratory
Los Alamos, NM 87545, U.S.A.

ABSTRACT

This tutorial paper provides information relevant to the selection and generation of stochastic inputs to simulation studies. The primary area considered is multivariate but much of the philosophy at least is relevant to univariate inputs as well.

1. INTRODUCTION

An important aspect of conducting simulation studies is the selection of the stochastic inputs (probability distributions and their parameters) and their subsequent generation. This problem is particularly interesting and challenging in the context of multivariate statistical simulation studies, and differs in flavor and tactics somewhat from selecting multivariate inputs in process simulations (which are in the majority at the Winter Simulation Conferences). This paper will concentrate on multivariate inputs but much of the philosophy in this realm will carry over to the univariate situation. A much more comprehensive coverage of the multivariate simulation area is given by Johnson (1987), so that in a sense, the current paper can be viewed as yet another unabashed advertisement for that book. As a token defense, let me add that I am doing this job against my will at the request of the program chair (just kidding, David).

2. UNIVARIATE PRELIMINARIES

Before delving into a superficial treatment of multivariate simulation, the problem of univariate distribution selection and generation must be addressed. The first difficulty involves uniform 0-1 generation, the essential component in generating random variates. Previously, I have capriciously assumed away the problem in hopes that the recommended generalized-feedback-shift-register generator (using the primitive trinomial $x^{98}+x^{27}+1$) would stand the test of time. Recently, I was delighted and relieved to learn that Ripley (1987, p. 46) prefers this generator as well. Hence, until I discover some repugnant aspect of the generator, I will continue to use and to recommend it, and let others pursue new, improved uniform generators. I must add that I have yet to encounter a situation in which the uniform generator was the culprit in producing anomalous results in a simulation.

For non-uniform generation, my two favorite sources are Devroye (1986) and Bratley, Schrage and Fox (1983). Devroye offers an encyclopedic account of univariate generation and also has approximately sixty pages on multivariate generation. This book is an essential reference for graduate students interested in variate generation research, as it is a source of many clever ideas. The Bratley et al. book is notable for considerable coverage of univariate distributions and includes FORTRAN listings for many of these distributions. I am grateful for both books in that I can comfortably refer them to readers needing univariate help, and I can avoid the overwhelming temptation to generously paraphrase (lift) material from these sources.

3. UNIVARIATE MODELING ASPECTS.

As a statistician, I cannot help but admit some disquiet in the distribution fitting process as espoused in some undergraduate simulation texts. The usual cozy paradigm goes something like this:

1. Hypothesize a distributional model (e.g., if arrival times, assume that the Poisson process holds).
2. Estimate parameters.
3. Apply goodness-of-fit tests.
4. Proceed on our merry way.

My objections are many-fold (distinguish from manifold). The data collection aspect is restrictive in that the distributions may arise under less than

ideal conditions. For example, a system badly in need of improvement may yield poor estimates of service times (bad morale on the part of the servers) and arrival times (perhaps many more customers would arrive if the service were not so pathetic). Although one of the most compelling arguments for simulation is its ability to handle potential system configurations, I rarely see this data-collection concern addressed. A question that could be asked that could lead to meaningful simulation runs is what arrival rates can be tolerated before the system produces unsatisfactory performance. Also, what service rates are necessary to accommodate some specified arrival pattern.

Another objection to the paradigm is the reliance on goodness-of-fit tests to substantiate the distributions. It is not a forgone conclusion that these tests will always pass or that the practioner will be able to find some model for which the tests yield positive (not negative) results. Goodness-of-fit tests are presently being carefully scrutinized in the statistical literature and many results are being published which are relevant to this situation. Diaconis and Efron (1986) have addressed the situation in which a test fails--how far from the baseline are we? Adapting this approach to fitting continuous data, Pederson and Johnson (1987) note necessary adjustments to chi-square to account for varying numbers of bins to allow comparison of data sets of different sample sizes. The bottom line in this diatribe is the recommendation to consult your friendly neighborhood statistician in data fitting situations.

4. MULTIVARIATE STORIES

In contrast to the univariate realm where there are many standard distributions which can be used (provided you believe the goodness-of-fit tests), the multivariate arena is not so well characterized. There are a variety of distinct distributions with varying degrees of parameter estimation capability, but the list of reasonable candidates is short. The following annotated list of my favorite ought-to-be-included multivariate distributions is offered:

a. Multivariate Normal. Of course, here I mean the "usual" multivariate normal distribution. In n dimensions the distribution can arise as follows: start with n independent normal variates (n calls to one's normal generator) and apply an affine transformation, say $\underline{A}\underline{x}+\underline{b}$, where $\underline{x}$ is the vector of generated normal variates, A is an n×n matrix and $\underline{\mu}$ is a location vector. Frequently, A is a lower triangular matrix so that the product AA′ is the covariance matrix of $\underline{X}$. It is thus very easy to generate this multivariate normal distribution.

b. Normal mixtures. In keeping with the robustness crowd, probabilistic mixtures of multivariate normals are easy to generate and of interest especially in such statistical applications. For the simplest case, generate from one multivariate normal distribution with probability p and from another multivariate normal with probability 1-p. This one is easy, also.

c. Johnson's translation system. This system consists of the usual multivariate normal distribution augmented with the possible component transformations:

exp(x) yielding lognormal variates

$[1+\exp(x)]^{-1}$ yielding logit-normal variates

$\sinh(x)$ yielding $\sinh^{-1}$-normal variates.

This system of distributions introduced by Johnson (1949, no relation) has had an enviable track record both in empirical modelling and in some multivariate simulation studies. About the only difficulty in using the system in simulation work is that there are many parameters to specify so it takes some thought to design a study (this is a much better "problem" than being stuck with little or no flexibility).

d. Elliptically contoured distributions. The Pearson Type II and Type VII (including the multivariate Cauchy) distributions are useful representatives of this class of distributions. An unlabeled contour plot of the density function for a bivariate Pearson Type II or VII is indistinguishable from that for a bivariate normal. A difference can be detected if one considers the distribution of $\underline{X}'\underline{X}$, which of course is chi-square with n degrees of freedom for $\underline{X}$ multivariate normal with zero mean vector and identity covariance matrix. The touted Type II and VII are easy to generate once the distribution of $\underline{X}'\underline{X}$ is derived, but this is easy.

e. Anisotropic distributions. Getting back to robustness applications, useful distributions in the simulation context are those that are painless to generate and yet capture some interesting departure from a baseline distribution (usually the normal). If we ponder multivariate distributions that are not spherically symmetric (a special case of elliptically symmetric) but which have $\underline{X}'\underline{X}$ as chi-square with n

d.f., we are led to anisotropic distributions introduced by Nachtsheim et al. (1987). These distributions have considerable flexibility and are easy to generate.

f. Burr-Pareto-logistic distributions. This class of distributions is being pushed by two camps. One group recognizes the distribution as arising from a general survival model from which results can be obtained (see, for example, Hougaard, 1986). Another group (Cook et al., 1981 and 1986) has been exploiting a particular parametric form from the class to fit uranium mineralization data. The case with normal marginal distributions has some nice features (non-constant conditional variance) which implore its use in simulation applications. As you might expect by now, this distribution is again easy to generate.

5. CRYSTAL BALL EXTRACTIONS

The list given in the previous section was provided to introduce the multivariate (continuous type) distributions I have found most useful in simulation applications. The details have been glossed over but honestly, the mentioned distributions are easy to generate. How will this list likely change in the next few years? The current "hot" topic in constructing multivariate distributions involves the so-called copulas, a basic functional component leading to many distributions including some disparate ones. For references on this topic, see Genest and MacKay (1986, 1987). Other research directions that are not yet obsolete can be found in Chapter 11 of Johnson (1987).

REFERENCES

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

Cook, R. D. and M. E. Johnson (1981). "A Family of Distribution for Modelling Non-elliptically Symmetric Multivariate Data," Journal of the Royal Statistical Society, Series B, 43, 210-218.

Cook, R. D. and M. E. Johnson (1986). "Generalized Burr-Pareto-Logistic Distributions With Applications to a Uranium Exploration Data Set," Technometrics, 28, 123-131.

Devroye, L. (1986). Non-uniform Random Variate Generation, New York: Springer-Verlag.

Diaconis, P. and B. Efron (1985). "Testing for Independence in a Two-way Table: New Interpretations of the Chi-square Statistic," (with discussion), The Annals of Statistics, 13, 845-913.

Genest, C. and J. MacKay (1986). "The Joy of Copulas: Bivariate Distributions with Uniform Marginals," The American Statistician, 40, 280-283.

Genest, C. and J. MacKay (1987). "More Joy of Copulas," personal communication.

Hougaard, P. (1986). "A Class of Multivariate Failure Time Distributions," Biometrika, 73, 671-678.

Johnson, M. E. (1987). Multivariate Statistical Simulation, New York: John Wiley.

Johnson, N. L. (1949). "Bivariate Distributions Based on Simple Translation Systems," Biometrika, 36, 297-304.

Nachtsheim, C. J. and M. E. Johnson (1987). "New Anisotropic Distributions with Applications to Hotelling's T^2," submitted to JASA.

Pederson, S. P. and M. E. Johnson (1987). "Discrepancy Measures for Grouped Continuous Data," submitted to Technometrics.

Ripley, B. (1987). Stochastic Simulation. New York: John Wiley and Sons, Inc.

Rubinstein, R. (1981). Simulation and the Monte Carlo Method. New York: John Wiley and Sons, Inc.