Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

DISTRIBUTION SELECTION AND VALIDATION

Stephen G. Vincent

School of Business Administration
University of Wisconsin–Milwaukee
P.O. Box 742
Milwaukee, Wisconsin 53201, U.S.A.

W. David Kelton

Department of Operations and Management Science
Carlson School of Management
University of Minnesota
Minneapolis, Minnesota 55455, U.S.A.

ABSTRACT

We have been concerned with the problem of specifying simulation input distributions for well over a dozen years. Over this time our attitudes and understanding of this problem have changed dramatically, and indeed continue to change at this time. The purpose of the *Proceedings* paper is to record the more philosophical aspects of our current thinking. We intend the conference talk to focus more on the practical topics that depend upon the material presented below.

1 INTRODUCTION AND GENERAL CONCEPTS

Developing a validated simulation model (Figure 1) entails three basic entities: the *real-world system* under consideration, a theoretical *model of the system*, and a *computer implementation* of the model. The activity of deriving the theoretical model from the real world system can be referred to as *simulation modeling*, and the activity whereby the computer implementation is derived can be referred to as *simulation programming*. The figure also shows the basic checks of *verification* and *validation* that are applied in the development of a simulation model. We shall assume that the concept of model validity is well-established in the reader's mind, and recommend, for example, the tutorial by Robert G. Sargent in this volume if refreshment is required.

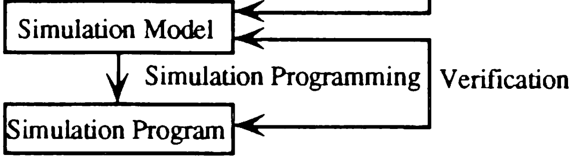

Figure 1: Overview of Simulation Model Development

One of the primary reasons for using simulation is that the model of the real-world system is too complicated for study using analytical methods. Major sources of complexity are the components of the model that "drive," or are inputs to, the logic of the model (the reasons that these components complicate the model will shortly become apparent). Examples of such inputs include arrivals of orders to a job shop, times between arrivals to a service facility, times between machine breakdowns, etc. Without loss of generality we can represent an "input" in the real-world using the notation $X = \{X_1, X_2, ...\}$, where the subscripts merely denote the time-ordered appearance of values. (Without loss of generality we have assumed that the process is discrete-time rather than continuous-time.) In most cases, there will be a corresponding representation of the input in the theoretical model, denoted $\hat{X} = \{\hat{X}_1, \hat{X}_2, ...\}$, as well as an implementation of $\hat{X}$, denoted Gen_X, in the simulation program; see Figure 2. At the most basic level, this tutorial is concerned with specifying appropriate representations $\hat{X}$, whereas the process of creating an appropriate implementation Gen_X is discussed in the review by Luc Devroye in this volume.

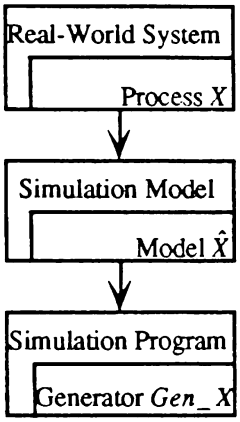

Figure 2: Role of Input Distributions

Typically, an $\hat{X}$ process is taken to be a *stochastic* process (i.e., there are some random components). The fundamental assumption that we shall make is that modeling X by an $\hat{X}$ that is stochastic is not inherently wrong; that is, the mere fact that $\hat{X}$ is stochastic does not immediately and completely invalidate our simulation model. Part of deriving a simulation model thus becomes determining an appropriate stochastic process model $\hat{X}$ for X. A fundamental choice that has to be made in modeling X is whether the model should be a *multivariate* or *univariate* stochastic process. (We are careful to indicate that assumptions are related to the models of the process and not to the process itself.) For example, a model of times between customer arrivals would typically be considered to be univariate since there is just the one value of interest. A model of arrivals of orders to a job shop, on the other hand, might be multivariate since the time of the order arrival as well as numbers of different products to be produced might be included in each $\hat{X}_k$ (see below). Note that in some cases we might be able to separate what appears to be a multivariate process into separate univariate (marginal) processes (such separation makes sense only when the components can be assumed to be independent of each other). The presence or lack of independence can arise in a number of ways for a process $\hat{X}$. Consider the arrival of orders to the job shop. Let $\hat{X}_k = \{\hat{t}_k, \hat{A}_k, \hat{B}_k, \ldots\}$ where $\hat{t}_k$ is the time of order arrival, $\hat{A}_k$ is the desired number of the first product, $\hat{B}_k$ is the desired number of the second product, etc. For a specific order, the numbers of products could be independent of each other, or could be correlated (e.g., a high value of $\hat{A}_k$ might typically be accompanied by a high value of $\hat{B}_k$). Further, there could be relationships between the numbers found on subsequent orders (e.g., for a single product a large amount on an order could be followed by smaller numbers on subsequent orders). The time of order arrivals $\hat{t}_k$ could be as simple as an integer period number or could represent a stochastic order-arrival time. The times between subsequent arrivals could be independent of each other or related (e.g., five orders arrive per month but at random points in the month).

Regardless of whether $\hat{X}$ is univariate or multivariate, we must specify the probability distribution of it, namely $F_{\hat{X}_k}(t) = \Pr\{\hat{X}_k \le t\}$. A significant simplification can occur both in notation and model complexity if the index does not matter, so that the $\hat{X}_k$ variables are taken to be identically distributed with $F_{\hat{X}}(t) = F_{\hat{X}_k}(t)$ for all k. The model can be simplified even further if the $\hat{X}_k$ variables can be assumed to be independent of each other. When both of these conditions hold, the process is called *IID* (independent and identically distributed). More is known about IID processes than any other type of process.

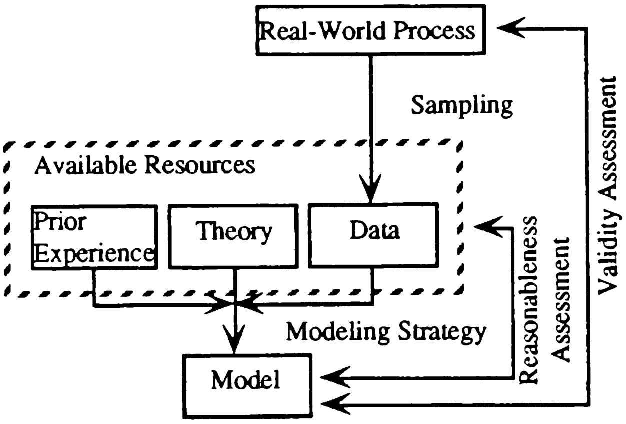

Figure 3: Modeling Input Distributions

An overview of the activity of modeling a real-world process is shown in Figure 3. An analyst will use general knowledge, relevant theory, and collected data (if available) as inputs to a modeling strategy that produces a model of the process. The phrase "modeling strategy" has been used to emphasize that the activity is typically more involved than application of a single statistical technique, and further, a good strategy may conditionally apply disparate methods. The checks shown in the figure address assessing model validity directly and indirectly. The (indirect) reasonableness assessment addresses the question of how well the model represents the data or other prior knowledge and is relatively easy to accomplish in practice. The (direct) validity assessment addresses the question of whether the model is a reasonable representation of reality. It is virtually impossible to accomplish this direct assessment in practice.

2 THE RELATIONSHIP BETWEEN INPUT MODELING AND OVERALL MODEL VALIDITY

It is important to emphasize that the overall goal in simulation modeling is to provide a model that is valid for a given context. The impact of the choice of a model $\hat{X}$ upon the overall simulation validity may range from crucial to virtual irrelevance (depending upon the system under consideration) and there is no definitive manner to ascertain it without application of formal validation methods. On occasion it has been suggested that a rough sensitivity analysis can be used to judge the severity of the impact. The logic of the analysis can be paraphrased as "try a number of representations and if the simulation results do not vary significantly, then the choice is not important." The fatal flaw in this logic is that the lack of variation does not establish the validity of any of the alternatives. (The logic *does* have its uses, but only toward the end of the simulation-development activity: once a completed model has been validated with a particular model $\hat{X}$, then a simpler form of $\hat{X}$ could be substituted for efficiency reasons, provided it does not substantially change the model results.) Because, in general, we don't know *a priori* the magnitude of the impact of the choice of a model $\hat{X}$ upon the overall simulation validity, we recommend the conservative (paranoid) approach of assuming that the impact is large. This implies that we should always attempt to create the most valid model $\hat{X}$ of X possible.

3 THE DIFFICULTY IN USING CLASSICAL STATISTICAL TECHNIQUES TO ASSESS INPUT MODEL VALIDITY

One issue arises when we consider what it means for a model $\hat{X}$ of a process X to be a valid representation. Heretofore we have not assumed that the X_k variables were themselves random, but rather we have assumed that it is not inappropriate for us to *model* them in that manner. Many simulation authorities have not made this distinction, and indeed assume that there is a true distribution function $F_{X_k}(t)$. For these authors, a primary method of assessing the model's validity is to use methods of classical statistics that are designed to test the closeness of theoretical and fitted models. In essence, for these authors there is no distinction between the direct and indirect validity assessments shown in Figure 3. We do not believe that this logic can be supported in a simulation context. Figure 4 represents the standard application of a goodness-of-fit test in classical statistics. An idealized application of such a test begins with a hypothesis of the form of the distribution function (e.g., exponential). A sampling plan is designed and carried out that produces a sample. Estimation procedures can be applied if the parameters of the distribution function are not known (e.g., specify the mean of the hypothesized exponential distribution). A formal goodness-of-fit test then tests whether the true underlying distribution is the hypothesized one by assessing the reasonableness of an assertion that the specified distribution could have produced the observed data. There are three problems with the use of such a test in a simulation context:

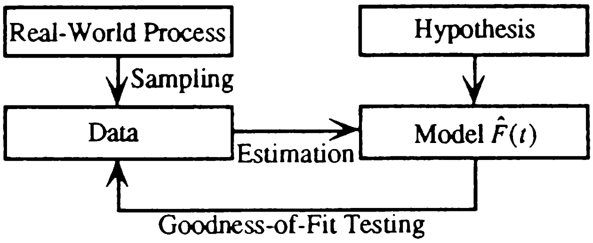

Figure 4: Goodness-of-Fit Testing

1. A simulationist will typically ***not*** know the form of the distribution function, but rather, might deduce the form from the data, which violates the validity of the test. Although much research has been performed on the use of goodness-of-fit tests, little has been done on how well these tests work when the form of the distribution is not known. There is a smaller body of knowledge on a related issue of selecting from a small group of candidates; that is, if we ***know*** that the form is one of a small number (typically two or three) of distributions, how can we determine which is the most likely parent?
2. Not only do we not assume knowledge of the form of the distribution, we are not necessarily assuming that reality is inherently stochastic. This problem is indicative of a more general difference between the aims of goodness-of-fit testing and the activity of modeling simulation input processes. Even if we ***do*** assume that reality is inherently stochastic, it is unlikely that we can deduce the true form of the distribution from the data. Further, to a large extent we are not interested in determining the ***exact*** form of the distribution function as we are in determining a form that provides a good approximation of it for purposes of producing useful input to the simulation via *Gen_X*.
3. Even when appropriate assumptions are made, the test does not necessarily give us the information we desire. This point can be demonstrated by considering the following scenario. Suppose that X is an IID process and further has a *known* distribution function $F(t)$. Suppose further that we were to observe and record with complete accuracy n observations of X. Due to the properties of random sampling, the recorded values may not "appear" to be distributed according to $F(t)$. Simply because of the randomness inherent in the data and the nature of test procedures, if we set the so-called level of the test at any value α, then $100\alpha\%$ of the time the test would indicate that we should reject the true hypothesis. The difficulty for the simulationist lies not in the rejection of the true hypothesis as much as the lack of any recommendation on how to then model the process.

4 THE DIFFICULTY IN EMPLOYING OVERALL SIMULATION VALIDATION TECHNIQUES TO THE PROBLEM OF ASSESSING INPUT MODEL VALIDITY

Formal and informal approaches for overall simulation model validation have been proposed (we again refer the interested reader to the tutorial by Robert G. Sargent in this volume). In either case, the general concept is to compare the "output" of the simulation model with the "output" or expectation of "output" from the real-world system. The unavailability of historical system data clearly precludes the application of any formal comparative method. When historical data *are* available, it is still difficult to compare formally the outputs due to their nature. Because the output data are correlated and not IID, most formal statistical comparative techniques are not directly applicable. Further, techniques that can be applied to the correlated data are not as powerful in detecting gross discrepancies as are the corresponding methods for IID data. A commonly used technique that increases our ability to detect gross discrepancies for correlated data is to compare the outputs produced from the real-world system with those produced when the *model* is driven with the historical input data. Although this technique allows for validation of the simulation logic, it precludes any validation of the models of the input processes. When the simulation program is driven by the input models, and we compare its results to historical data, detecting even gross discrepancies between the model and real-world outputs is difficult due to the differences in driving forces. Even when gross discrepancies are not found, there is no guarantee that the input models are valid. The lack of gross discrepancies in the outputs may be attributable to our inability to detect the errors, the masking effect of the simulation logic, the limited amount or quality of historical data against which to compare, or the randomness inherent in the simulation outputs themselves. Our conclusion is that formal validation techniques can detect only a portion of invalid input models, and we should therefore not rely upon them to validate the input models.

5 FURTHER COMPLICATIONS IN MODELING PROCESSES

Classical statistical methods assume that we have a representative sample of the process to use in specifying our model. This is certainly not true in all simulation projects. Even when data *are* available, the situation can differ from what is assumed in classical statistics; see Figure 5. Typically, the assumption in classical statistics is that data are collected in a planned and systematic manner *after* the analysis method has been chosen. In a simulation study, it is quite common to make use of data that were previously collected for purposes other than the determination of a simulation model. This situation can lead to a number of annoyances:

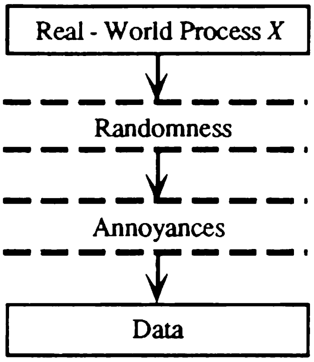

Figure 5: The Reality of Sampling

- Data can be grouped into intervals
- Data can be recorded with insufficient precision, perhaps even rounded to the closest integer even though the observations were real-valued
- Data can be full of obviously erroneous values, simply because the recorder did not anticipate the need for highly accurate data
- Data can be contaminated or mixed with observations of other processes
- Data can be representative values from a completely different process

Although not all of these annoyances will occur with each data set in a simulation study, they occur with a frequency that has led us to be slightly suspicious of the validity of any data set that is derived from historical records. Our skepticism concerning the accuracy and representativeness of data leads us to a slightly paradoxical position: Although a data sample is the best evidence available for use in specifying and evaluating a proposed model, it should not be taken *too* seriously. The impact of this conclusion is a mistrust of model-selection strategies that are overly reliant upon "clean" samples.

6 FINAL THOUGHTS

We can summarize the material presented so far in the following manner: We believe that simulationists should strive to create the most valid model of a process *X* that is possible, in the interest of overall model validity. It is not possible to perform a direct validation check on a proposed model using classical statistical techniques, and the results of validation procedures applied to the entire simulation model may not indicate problems with models of input processes. Data available to us can suffer from inaccuracies and can "look" different from the process that produced them. What modeling strategy should a simulationist employ? A general recommendation is that the strategy that produces from its generator *Gen_X* values that most closely resemble the values that can be produced by the real-world process. We believe for the simulation context that differences in modeled and true process values are more important than differences in modeled and true process distribution functions; we take the situation to be close to that of regression analysis where we wish to know how closely an estimate is to the true value. The conference tutorial shall consist to a large extent of practical advice on what strategies to employ, with recommendations based upon research findings and practical experience.

AUTHOR BIOGRAPHIES

STEPHEN G. VINCENT is an Assistant Professor in the School of Business Administration at the University of Wisconsin–Milwaukee, where he teaches courses in the areas of software engineering and simulation. He was Vice President of Simulation Modeling and Analysis Company in charge of software development until 1987 during which time he developed the UniFit software package with Averill Law. He received his Ph.D. in Management Information Systems from the University of Arizona and has B.S. and M.S. degrees in Industrial Engineering from the University of Wisconsin–Madison.

W. DAVID KELTON is Professor of Operations and Management Science in the Carlson School of Management at the University of Minnesota. His interests are in simulation methods, stochastic-process estimation, queueing, and statistical quality control. He serves as Area Editor for Simulation for the *ORSA Journal on Computing*, and is an Associate Editor for *Operations Research* as well as *IIE Transactions*. He was President of the TIMS College on Simulation from 1990 to 1992, and is the ORSA representative to the Winter Simulation Conference Board of Directors. In 1987, he served as Program Chair for the WSC, and in 1991 as General Chair.