MANAGEMENT SCIENCE
Vol. 50, No. 7, July 2004, pp. 855–868
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5007 | 0855

informs

DOI 10.1287/mnsc.1040.0246

50th Anniversary Article

Stochastic Simulation Research in *Management Science*

Barry L. Nelson
Northwestern University, Evanston, Illinois 60208, nelsonb@northwestern.edu

When the simulation department of *Management Science* was created in 1978 it ushered in an era of significant methodological advances in stochastic simulation. However, the foundation for the field—not just the work that has been published in *Management Science*—was provided by two papers published long before simulation had its own department in the journal. We will review the seminal papers of Conway, Johnson, and Maxwell (1959) and Conway (1963), and then trace their impact through eight award-winning papers that appeared much later in *Management Science*.

Key words: simulation; experiment design; output analysis; variance reduction; regenerative model

1. Two Seminal Papers

> Our objective is to discuss the technique of digital system simulation. This procedure has already achieved a considerable stature in industrial and research organizations and promises to attain even greater importance in the future. Yet, with few exceptions, the published literature in the area consists of introductory expositions or of descriptions of the solution of particular problems. (Conway, Johnson, and Maxwell 1959, p. 92)

From these opening remarks on the state of published research in computer simulation, Conway et al. (1959), in conjunction with a later paper by Conway (1963), described a number of simulation problems that continue to occupy researchers to this day. In fact, it can be argued that these two papers defined the simulation research area that is now known as "analysis methodology" for the operations research and management science communities (Nance and Sargent 2002). In the following sections we will review the research issues identified by Conway et al. (1959) and Conway (1963) and then trace their impact through eight award-winning papers that have appeared in *Management Science*.

2. Strategic and Tactical Issues in Simulation

Defining "simulation" is a notoriously difficult task. However, the type of simulation that the readers of *Management Science* have most often encountered has three features:

(1) There are one or more stochastic input processes that are specified via the language of probability and from which synthetic realizations can be generated.

(2) There is a logical model that completely describes how the system of interest reacts to realizations of the stochastic inputs. The logical model is usually an algorithm that updates the system state upon the occurrence of some discrete events.

(3) There are output processes, typically ordered by some concept of time, that represent the system behavior of interest to the analyst.

A standard example is a queueing system simulation in which the input processes are customer-arrival and customer-service times, the logical model describes the order in which waiting customers are chosen for service, and the output processes include the delay in queue experienced by each customer and the number of customers in the queue as a function of time. The state of the system is the number of customers in the queue and the status of the servers, while the events that alter the system state are customer arrivals and customer departures. This type of simulation describes how a system evolves through time at a very detailed level, which has several implications according to Conway et al. (1959):

> Consideration of the atomistic characteristic of simulation reveals much about its properties. First, it suggests a condition for feasibility of simulation; a system, however complex, can be simulated if it can be broken down into a set of elements for which operating rules can be given. If the smallest elements into which we can divide the system are themselves unpredictable (even in a probabilistic sense) digital simulation is not feasible. Second, it is a mathematical model that is "run," rather than one that is "solved." It is not inherently optimizing; rather it is descriptive of the

> performance of a given configuration of the system. Optimization must be superimposed upon this model by varying the configuration in search of a maximum of performance. Third, the simulation does more than yield a numerical measure of the performance of the system. It provides a display of the manner in which the system operates. Finally, this discussion of the description of individual elements, the recording of individual events, and the frequent necessity of replication is rather suggestive of the reason why this form of system simulation was not widely used before the advent of the modern high-speed stored-program digital computer .... It is simply the tremendous volume of logical, numerical and bookkeeping operations that must be performed that makes this procedure a natural application for a digital computer. (Conway et al. 1959, pp. 94–95)

As a consequence of this analysis, Conway et al. (1959) concluded that simulation problems fall into two broad categories: the *construction* of a computer simulation, and the *use* of a computer simulation. It is probably fair to say that the construction problems that Conway et al. (1959) identified have been addressed; they include modular construction of simulation programs for easy updating, management of computer memory, controlling error due to discretization, specifying an effective time-advance mechanism, and managing active data files. The problems of simulation use, however, remain a source of significant research interest. Conway (1963) divided the "use" problems into those that are addressed by *strategic planning*—primarily the design of an experiment that will produce the desired information—and those addressed by *tactical planning*—determining how each of the simulations specified by the experiment design will actually be run. A present-day student of simulation reading Conway et al. (1959) or Conway (1963) would clearly recognize and understand these strategic and tactical issues. Perhaps more impressive, they would also recognize that the solutions they know are consistent with the proposals found in these papers.

Conway et al. (1959) and Conway (1963) viewed simulations as statistical experiments, so it was natural that they would bring to bear the statistical tools of the day for strategic planning, tools such as factorial experiment design. Tactical planning is focused on the efficiency of the simulation run, and here we find problems that are unique to simulation and about which little had been written at the time (Conway 1963, p. 48). Conway et al. (1959) and Conway (1963) identified three key tactical issues:

(1) *Establishing when a simulation run is in statistical equilibrium*. This is one aspect of the so-called "steady-state simulation problem." A stochastic simulation implicitly defines a stochastic process, which we call the output process. Executing the simulation generates sample paths. If the output stochastic process has a limiting distribution, then it may be of interest to estimate properties of that distribution as a summary of long-run behavior (this is not the only definition of steady-state simulation; see for instance Henderson 2000). Examples of steady-state performance measures include the long-run expected cycle time for products in a job shop, the long-run availability of a repairable system, and the long-run expected cost per period of an inventory policy. The tactical problem is determining how much simulated time should elapse before the probability law of the output process is sufficiently close to the limiting distribution to allow accurate (i.e., low bias) estimates to be obtained. A less strenuous requirement is that some specific property of the output process, such as its mean value, is close to the steady-state limit, and this is the sort of definition that is typically employed in practice. Deciding when statistical equilibrium has been (nearly) attained is difficult, so Conway (1963) described two approaches to reduce the impact of deciding poorly: Delete data from an initial period of each run, and choose starting conditions for the simulation state that nudge the output process closer to long-run conditions than an arbitrary initial state does. The data deletion rule in Conway (1963) is probably the first published algorithm and it was the standard against which new proposals were tested for many years (e.g., Gafarian et al. 1978).

(2) *Producing precise comparisons of alternatives*. Recognizing that most simulation studies are performed to compare competing system designs, Conway et al. (1959) and Conway (1963) observed that driving each simulated alternative with the same realizations of the input processes (when possible) would typically yield sharper comparisons because the observed differences should be due to structural differences in the systems, rather than chance differences in the stochastic inputs. They also recognized that using common stochastic inputs introduces correlation between the outputs from different alternatives, invalidating many of the available statistical tools for making comparisons (e.g., ANOVA). Statistical analysis under "common random numbers" has been the topic of many papers in *Management Science*, including Nelson and Matejcik (1995) and Kleijnen (1988).

(3) *Obtaining a valid measure of error for simulation-based estimates of equilibrium performance*. When simulating, it is always possible to generate independent and identically distributed (i.i.d.) replicates of any simulation-based estimator simply by running the model repeatedly with independent realizations of the input processes. The availability of i.i.d. replicates makes classical statistical analysis possible. However, Conway et al. (1959) and Conway (1963) realized that the difficulty of identifying an approximate statistical equilibrium, and the potential waste of

precious data from deleting outputs from each replication, argue against a multiple-replication approach. The alternative is a single long run. However, while the outputs from within a single run may be identically distributed (if approximate equilibrium has been attained), they are rarely independent. Lack of independence invalidates standard variance-estimation and confidence-interval procedures, sometimes dramatically, as Conway (1963) demonstrates. Both papers suggest partitioning the single output realization into nonoverlapping batches of outputs, computing summary statistics from within each batch, averaging these summary statistics to obtain an overall estimator, and calculating a variance estimator or confidence interval from the (typically far less dependent) batch statistics. Conway (1963) anticipates many papers in *Management Science* on batching-based methods (e.g., Steiger and Wilson 2002), time-series methods (Fishman and Kiviat 1967), and estimation of the so-called asymptotic variance (Goldsman et al. 1990). He also states a result that was later made rigorous by Schmeiser (1982): When batching to estimate the mean, it is rarely beneficial to form more than 20 batches no matter how much output data are available.

The tactical issues in Conway et al. (1959) and Conway (1963) were certainly known prior to the publication of these papers, but their work is responsible for bringing them to the attention of the operations research and management science communities. And Conway (1963) acknowledges that they are not the last words on these topics:

> While there are some theoretical bases and rational arguments for making these tactical decisions in simulation, much is still based on the experience and judgment of the investigator. Simulation on a digital computer is still very much an art, with success depending heavily on the skill of the artist. Particularly during these formative years for the technique, it is vital that its practitioners exchange information and experiences on every aspect of its use, so that previous mistakes will be less frequently repeated. (Conway 1963, p. 61)

In §4 we will review some particularly outstanding examples of the exchange of information that Conway desired.

3. The Simulation Department and the Award

In response to lobbying by The Institute of Management Sciences (TIMS) College on Simulation and Gaming, especially Robert Sargent (Syracuse University) and J. William Schmidt (Virginia Polytechnic Institute and State University), a simulation department was added to *Management Science* in 1978. George Fishman (University of North Carolina) was the first department editor, and his editorial policy for the new department can be found in the *Newsletter of the TIMS College on Simulation and Gaming* (Fishman 1980). In brief, the department was looking for contributions that describe (a) innovative ideas for modeling flow (state change) logic in simulated systems; (b) new probabilistic representations of underlying stochastic structures; (c) new and improved methodologies for analyzing simulation output and increasing statistical efficiency in estimation; and (d) unusual applications using existing or new methodological procedures. Papers in category (c) have been most prevalent, and Fishman set an important and influential standard for accepting such papers:

> Papers under topic c should describe how the proposed technique compares with past proposals for solving a particular statistical or decision making problem. Here comparison would include considerations of statistical performance, degree of generality, computational efficiency, ease of implementation and simplicity of concept. A methodological paper is expected to demonstrate its proposed technique and evaluate its performance relative to competing techniques. This demonstration should be accomplished with simulation models for which theoretical solutions are known. In this way a reader can assess both absolute and relative performance. In cases in which a paper presents a new methodology in an area where no alternatives exist, the absolute standard will provide the basis for an evaluation. A description of the testing procedure should specify an experimental design for each theoretical model considered. For example, if one is using a queueing model as the basis for testing, one would expect that studying performance of the proposed technique for a range of traffic intensities would be one element of the experimental design. (Fishman 1980, p. 5)

Fishman's requirement for empirical, as well as theoretical, evaluation of new methodologies using a sensible experiment design and thorough documentation of the experiment has been a consistent requirement of the department. James R. Wilson (North Carolina State University) succeeded Fishman as department editor, followed by Pierre L'Ecuyer (Université de Montréal), Paul Glasserman (Columbia University), and Perwez Shahabuddin (Columbia University). When Glasserman took over in 1998 the department was merged with applied stochastic models to become the stochastic models and simulation department.

In May of 1980 the TIMS College on Simulation and Gaming established an award to recognize the best paper on simulation appearing in the previous year's volume of *Management Science*. As stated in Fall 1980 Newsletter, "All papers and technical notes in a given volume of *Management Science* that have a major focus

Table 1 Recipients of the TIMS, then INFORMS, College on Simulation Publication Award

Award year	Authors	Title
1981	Lee W. Schruben	A coverage function for interval estimators of simulation response
1982	Stephen S. Lavenberg and Peter D. Welch	A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations
1983 (tie)	Mark S. Meketon and Philip Heidelberger	A renewal theoretic approach to bias reduction in regenerative simulations
1983 (tie)	Averill M. Law and W. David Kelton	Confidence interval procedures for steady-state simulations, II: A survey of sequential procedures
1985	James R. Wilson and A. Alan B. Pritsker	Experimental evaluation of variance reduction techniques for queueing simulation using generalized concomitant variables
1990	Philip Heidelberger, Xi-Ren Cao, Michael A. Zazanis, and Rajan Suri	Convergence properties of infinitesimal perturbation analysis estimates
1991	Ward Whitt	Planning queueing simulations
1996	Perwez Shahabuddin	Importance sampling for the simulation of highly reliable Markovian systems

on the theory or practice of simulation or gaming are eligible for this award" (p. 3). Nominations for a minimum of three papers were required for the award to be given in any year, and the recipient was selected by a vote of the College membership.[1] In 1985 the College changed the award to the "Outstanding Simulation Publication Award" and expanded eligibility to include any simulation publication copyrighted within the previous three years, with the recipient selected by a College committee (see the *Newsletter of the TIMS College on Simulation and Gaming* 1985). The award continues to be given by the INFORMS College on Simulation; more information can be found at www.informs-cs.org.

4. The Award-Winning Papers

Five papers received the College on Simulation publication award during the period in which it was restricted to papers appearing in *Management Science*, and three papers that were published in *Management Science* have received the award since the change. The award-winning papers are listed in Table 1 with complete bibliographic information in the references. Although this list does not include nearly all of the outstanding contributions to simulation found in *Management Science*, it does provide a peer-selected sample of the very best. For a list of simulation papers with over 50 citations, as compiled by Hopp (2004), see Table 2.

In the sections that follow, each award-winning paper is reviewed in the context of the strategic and tactical issues defined by Conway et al. (1959) and Conway (1963). Before describing the papers, we establish some basic background.

4.1. Notation and Definitions

A certain amount of background and notation is necessary to appreciate the work in the eight award-winning papers, and we present that here. We ask the reader to tolerate some mathematical looseness throughout this section and the reviews; more rigor would add complexity without adding insight. See the papers themselves for technically tight presentations.

Most of the award-winning papers focus on a single replication of a simulation output process, denoted $\{Y_i, i = 1, 2, \ldots, n\}$ for discrete-time processes and $\{Y(t), 0 \le t \le T\}$ for continuous-time processes. As a concrete example, Y_i might be the delay in queue of the ith customer to depart from a queueing system and $Y(t)$ might be the number of customers in the system at time t. For convenience we will sometimes let $\mathbf{Y}$ denote the entire output process from a simulation.

The purpose of the simulation experiment is usually to estimate some property of Y_i or $Y(t)$, often the steady-state mean denoted by θ. The standard estimators are

$$\overline{Y}(n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

for a discrete-time output process of n observations, and

$$\overline{Y}(T) = \frac{1}{T}\int_0^T Y(t)\,dt$$

Table 2 Most Cited Simulation Papers in the First 50 Years of *Management Science*

Conway, R. W. 1963. Some tactical problems in digital simulation. *Management Science* **10** 47–61.

Lavenberg, S. S., P. D. Welch. 1981. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science* **27** 322–335.

Naylor, T. H., J. M. Finger. 1967. Verification of computer simulation models. *Management Science* **14** B92–B101.

Van Horn, R. L. 1971. Validation of simulation results. *Management Science* **17** 247–258.

Suri, R., M. A. Zazanis. 1988. Perturbation analysis gives strongly consistent sensitivity estimates for the *M*/*G*/1 queue. *Management Science* **34** 39–64.

Glynn, P. W., D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* **35** 1367–1392.

[1] If there were only one or two nominations in a year, then the nominated papers carried over to the next year.

for a continuous-time output process up to time T. It is essential to have a measure of error for these estimators, which usually requires an estimator of their variances. Since Y_i and $Y(t)$ are stochastic processes—not the i.i.d. observations of classical statistics—estimating the variance is tricky, as Conway (1963) points out. However, under certain conditions the following limits exist:

$$\sigma_d^2 = \lim_{n\to\infty} n \operatorname{Var}(\overline{Y}(n))$$

for a discrete-time output process, and

$$\sigma_c^2 = \lim_{T\to\infty} T \operatorname{Var}(\overline{Y}(T))$$

for a continuous-time process. In either case the constant σ^2 is called the *asymptotic variance*. For large n or T, $\operatorname{Var}(\overline{Y}(n)) \approx \sigma_d^2/n$ and $\operatorname{Var}(\overline{Y}(T)) \approx \sigma_c^2/T$, respectively. The asymptotic variance plays a role in several of the award-winning papers.

A simulation output process is said to have regenerative structure if there is an increasing sequence of times $\{T_j, j = 0, 1, 2, \ldots\}$, with $T_0 = 0$, such that $\{Y_i,\ T_{j-1} \le i < T_j - 1\}$ are i.i.d. for $j = 1, 2, \ldots$, or in continuous time $\{Y(t),\ T_{j-1} \le t < T_j\}$, $j = 1, 2, \ldots$ are i.i.d. For instance, in some queueing systems letting T_j be the jth time (either customer number for Y_i or simulation clock time for $Y(t)$) that a customer arrives to find the queueing system empty and idle is a regeneration time. Thus, the T_j partition the simulation output process into i.i.d. (but random-sized) batches or cycles of output data. Appealing to regenerative structure (plus some additional conditions) is one way to establish the existence of the asymptotic variance. However, regenerative structure can also be exploited directly in simulation output analysis.

Let $Z_j = \sum_{i=T_{j-1}}^{T_j - 1} Y_i$ or $Z_j = \int_{T_{j-1}}^{T_j} Y(t)\,dt$, and let $C_j = T_j - T_{j-1}$ in either case. Abusing notation, let

$$N(t) = \max\{j\colon T_j \le t\},$$

allowing t to represent either discrete or continuous time. Thus, $N(t)$ corresponds to the number of completed cycles by time t. The standard regenerative estimator for the steady-state mean θ based on a simulation run of length T is

$$\hat{\theta}(N(T)) = \frac{\sum_{j=1}^{N(T)} Z_j}{\sum_{j=1}^{N(T)} C_j}. \tag{1}$$

Under certain conditions this estimator converges to θ with probability 1 as $T \to \infty$. More critically, the i.i.d. nature of the regenerative cycles provides an opening to construct a variance estimator for $\hat{\theta}(N(T))$; see, Glynn and Iglehart (1993), for instance. The drawback, of course, is the need to identify regeneration times. In the award-winning papers, regenerative analysis is used indirectly to prove results, and directly to form estimators.

4.2. The Coverage Function (Schruben 1980)

Motivated by Conway et al. (1959) and Conway (1963), much of the early research in analysis methodology focused on deriving valid confidence intervals for steady-state performance measures. Typically a procedure was proposed whose validity could be justified either heuristically or asymptotically (as the simulation run length goes to infinity), and then supported by an empirical analysis. Schruben (1980), the first paper to receive the TIMS College on Simulation and Gaming award, introduced a tool for evaluating the performance of a confidence-interval procedure (CIP) and reporting the results of an empirical study.

Let $R(\eta, \mathbf{Y})$ be the interval generated by a CIP given simulation data $\mathbf{Y}$ and nominal confidence level η. Ideally, $\Pr\{\theta \in R(\eta, \mathbf{Y})\} = \eta$ (often $\eta = 1 - \alpha$, where α is the allowable chance of error). The achieved coverage is

$$\eta^* \equiv \inf\{\eta \in [0, 1]\colon \theta \in R(\eta, \mathbf{Y})\}$$

the smallest confidence level at which the region just covers the parameter. Schruben noted that under mild conditions a valid CIP has the property that

$$F_{\eta^*}(\eta) \equiv \Pr\{\eta^* \le \eta\} = \eta$$

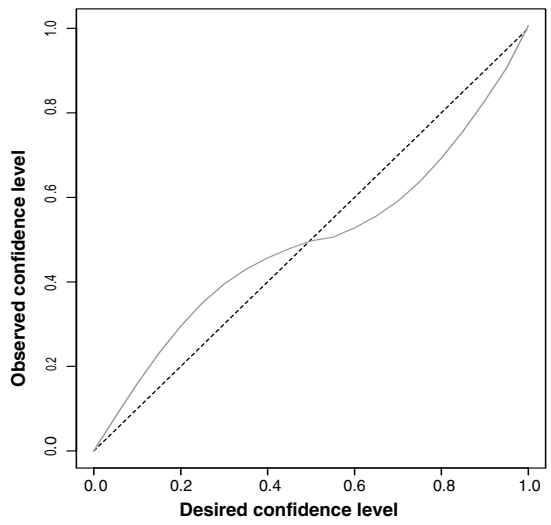
and he called $F_{\eta^*}(\eta)$ the *coverage function*. The beauty of the coverage function is that deviations of $F_{\eta^*}(\eta)$ from η, for $0 \le \eta \le 1$, provide a more comprehensive evaluation of the CIP than testing it at a small number of standard confidence levels (e.g., 0.9, 0.95, 0.99). In empirical studies on systems for which the true value of θ is known (as suggested by Fishman's editorial policy), an empirical coverage function based on m independent trials can be formed as

$$\widehat{F}_{\eta^*}(\eta) \equiv \frac{1}{m}\sum_{i=1}^{m} \mathcal{I}(\eta_i^* \le \eta) \tag{2}$$

where $\mathcal{I}$ is the indicator function.

In some simple cases (normally distributed data, known variance and correlation structure) $F_{\eta^*}(\eta)$ can be computed explicitly and Schruben (1980) did this to illustrate the deleterious effects of initial-condition bias and serial correlation about which Conway (1963) warned. He also presented an empirical study estimating the steady-state expected delay in an $M/M/1$ queue that demonstrated difficulties that the simplest version of the regenerative estimator (1) and associated confidence interval could encounter. Figure 1 is an empirical coverage function similar to the one in the paper. As Schruben explains, the overcoverage at low confidence levels and undercoverage at high levels results from two factors: a negative bias in the regenerative point estimator, which tends to cause the confidence interval to be centered too low, and a positive correlation between the regenerative point estimator and its associated variance estimator, which

Figure 1 **An Empirical Coverage Function Similar to Schruben (1980), Figure 4**



implies that point estimators that are too small tend to be associated with confidence intervals that are too short.[2]

Schruben's coverage function paper is an example of all-too-rare research on tools for conducting research, and it set important standards for a field in which empirical evaluation is critical. The coverage function concept was recently extended by Schmeiser and Yeh (2002).

4.3. The "Equilibrium" Problem (Meketon and Heidelberger 1982)

One of the key tactical issues for Conway et al. (1959) and Conway (1963) was mitigating the bias due to initial conditions in steady-state simulation. Meketon and Heidelberger (1982) developed a simple strategy for reducing the bias of the point estimator, a strategy that is most effective when it is needed most (i.e., when the run length is short).

To present their idea it is easiest to work with the continuous-time output process $Y(t)$ (e.g., number of customers in the queue). In §4.1 we introduced two estimators for the steady-state mean, $\overline{Y}(T)$, the time-average of $Y(t)$ through time T, and the regenerative estimator based on $N(T)$ regenerative cycles, $\hat{\theta}(N(T))$. In typical cases both estimators have bias that is $O(1/T)$ as T increases. Meketon and Heidelberger (1982) proposed the following simple refinement: Use $\hat{\theta}(N(T)+1)$ instead. In other words, run the simulation until time T *and then complete the current regenerative cycle before computing the point estimator*. Amazingly (until you understand the reason), the bias of $\hat{\theta}(N(T)+1)$ is $O(1/T^2)$, under very mild conditions, completely eliminating the first-order term in the bias (Meketon and Heidelberger 1982, Theorem 1).

Why does this work? As a consequence of the i.i.d. properties of regenerative cycles, $\{N(t), t \geq 0\}$ is a renewal process to which the so-called "inspection paradox" applies. Therefore, the regenerative cycle containing time T tends to be longer than the typical cycle. "Thus this last cycle contains significantly more information than a typical cycle and its inclusion in the ratio estimate guarantees a reduction in bias" (Meketon and Heidelberger 1982, p. 175). The proof exploits the fact that $N(T)+1$ is a stopping time, while $N(T)$ is not, and thus Wald's identity applies with the former but not the latter.

Meketon and Heidelberger (1982) is an elegant example of an approach that was not well developed at the time of Conway et al. (1959) and Conway (1963): View the simulation itself, or at least its output processes, as an instance of a well-studied family of stochastic processes and exploit knowledge about such processes to do simulation better. The stochastic processes perspective pays huge dividends because powerful analysis tools can be applied that may not depend on whether the simulation is of a queue, a supply chain, or a financial instrument. This approach turns up in nearly all of the award-winning papers.

4.4. Small-Sample Properties of Confidence-Interval Procedures (Law and Kelton 1982)

While Schruben (1980) was interested in establishing a methodology for evaluating CIPs, Law and Kelton (1982) sought to compare the performance of the leading competitors of the day. The subject of their investigation was sequential CIPs, which are procedures that terminate (stop simulating) when the stopping criterion is satisfied. They examined four procedures, two based on the regenerative method and two that were not. Two of the methods terminate when the half-length of their confidence interval achieves a prespecified relative width γ (half-length of the interval divided by the point estimator); the third terminates when the half-length of the interval achieves a given absolute width δ; and the fourth procedure did not enforce a precision criterion (for simplicity it will be omitted from the remainder of the discussion).

Law and Kelton (1982) were interested in determining which procedures could be expected to work well in practice. To that end they selected 10 system simulation models that were realistic, but for which the steady-state mean θ could be calculated: single queues, networks of queues, and an inventory system. Both discrete- and continuous-time output processes were evaluated. Because all of these models could be initialized in steady-state, or near steady-state, conditions, the coverage of each procedure reflected only

[2] We hasten to point out that these problems have been addressed by other researchers, including Meketon and Heidelberger (1982) discussed in §4.3.

the procedures' tolerance of dependent output data and not the impact of any residual initial-condition bias.

The CIPs were evaluated on their ability to achieve a nominal 90% coverage and on the average total run length required to terminate. Law and Kelton (1982) considered coverage to be the most important criterion and, among procedures that achieve the desired coverage, the one that is most efficient is best. This perspective has been pervasive in the simulation literature.

Law and Kelton (1982) is exemplary for its selection of cases, careful experiment design, and thorough documentation and analysis of the results. Although the stated goal of the paper was to determine which existing procedures would be useful in practice, it also provided a benchmark against which new procedures could be evaluated.

Why should we be interested in sequential procedures? Law and Kelton (1982) observed that a guarantee neither of coverage nor of precision can be provided if a fixed run length is specified arbitrarily. "It will often not be possible to know in advance even the order of magnitude of the run length needed to meet these goals in a given simulation problem, so some sort of procedure to increase iteratively this run length would appear to be in order" (Law and Kelton 1982, pp. 550–551). The challenge of determining a run length prior to doing any simulation is taken up by Whitt (1989) in a later award-winning paper.

4.5. Control Variates (Lavenberg and Welch 1981, Wilson and Pritsker 1984b)

Statistical efficiency was a paramount concern in Conway et al. (1959) and Conway (1963), which makes sense because, at the time, computing was relatively slow. Surprisingly, as computers have gotten faster, research on efficient simulation has increased, rather than diminished, because the availability of more computing horsepower has whetted the appetite of modelers to solve more complex problems, more often. Techniques that attempt to squeeze more precise estimators of system performance out of the same amount of computing effort, or equally precise estimators out of less effort, are called variance reduction techniques (VRTs). VRTs were originally developed for Monte Carlo estimation and survey sampling problems, not discrete-event stochastic simulation. Unfortunately, techniques that are effective in these other contexts do not always translate directly to the simulation of dynamic systems, especially when the VRT requires changing the simulation model itself. One technique that does translate well is the method of control variates (CVs). CVs are not invasive; the technique requires observing some additional concomitant variables that are generated during the course of the simulation and then using them to modify the standard or "crude" estimator after the run is completed.

In the late 1970s and early 1980s there were significant advances in the application of CVs to discrete-event, stochastic simulation, particularly for the simulation of queueing systems (e.g., Iglehart and Lewis 1979; Lavenberg et al. 1979, 1982). Lavenberg and Welch (1981) surveyed the state of knowledge at the time, resulting in one of the most-cited simulation papers ever published in *Management Science*.

Lavenberg and Welch (1981) described CVs for simulations employing multiple replications and for single-run, steady-state simulations that have regenerative structure. For brevity we will focus on the replication environment in which the goal is to estimate $\theta = \mathrm{E}(Y)$, and i.i.d. replicates $Y_1, Y_2, \ldots, Y_n$ can be generated by the simulation. For example, Y_i could be the average delay in queue of all customers served during replication i, so that θ is the expected average delay.

The standard or "crude" estimator in this case is $\overline{Y}(n)$, the sample mean, which has variance σ_Y^2/n. CVs attempt to produce, with essentially the same computational effort, an alternative estimator whose variance is smaller (ideally much smaller) than σ_Y^2/n. Although θ is unknown, there are many random quantities in any stochastic simulation whose true means are known, and some of these may be strongly correlated with Y. CVs exploit this relationship. For instance, the service-time random variables for a queue are typically input processes, in which case their distribution is completely specified by the modeler. Therefore, if X_i is the average of all of the service times sampled during the ith replication, then $\mu = E(X_i)$ is known, and (Y_i, X_i) will tend to be positively correlated because longer than expected service times will tend to be associated with longer than expected delays in queue, and vice versa. The random variable $(X_i - \mu)$ is called a "control." Lavenberg and Welch (1981) noted that simple functions of input random variables are a good source of controls.

Here is how controls are used. Let $\mathbf{X}_i$ be a $q \times 1$ vector of controls from the ith replication with known expectation $\boldsymbol{\mu}$. Then the control-variate estimator of θ is

$$\overline{Y}(\mathbf{b}) = \overline{Y}(n) - \mathbf{b}'(\overline{\mathbf{X}}(n) - \boldsymbol{\mu}) \tag{3}$$

where $\mathbf{b}$ is a $q \times 1$ vector of constants. Observe that this estimator is the intercept term in a linear regression of $\overline{Y}(n)$ on the control vector $\overline{\mathbf{X}}(n) - \boldsymbol{\mu}$ with slope coefficient vector $\mathbf{b}$. This is an unbiased estimator of θ whose variance is minimized if $\mathbf{b} = \boldsymbol{\beta} \equiv \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}Y}$, where $\boldsymbol{\Sigma}_{\mathbf{X}}$ is the variance-covariance matrix

of the controls, and $\boldsymbol{\Sigma}_{\mathbf{X}Y}$ is the $q \times 1$ vector of correlations between each control and Y. The minimum variance is

$$\mathrm{Var}(\overline{Y}(\boldsymbol{\beta})) = (1 - \rho^2_{Y\mathbf{X}})\frac{\sigma^2_Y}{n} \tag{4}$$

where $\rho^2_{Y\mathbf{X}}$ is the square of the multiple correlation coefficient between Y and $\mathbf{X}$. The message from (4) is clear: Stronger correlation means more variance reduction. And because $\rho^2_{Y\mathbf{X}}$ will never go down if you add more controls, it seems to make sense to use every control in sight. The flaw in this reasoning is that the optimal value of $\mathbf{b}$, $\boldsymbol{\beta}$, is almost never known, and an arbitrary choice of $\mathbf{b}$ can actually inflate the variance. When $\boldsymbol{\beta}$ is estimated, as it must be, then things change. See Lavenberg and Welch (1981) for details on the estimator $\hat{\boldsymbol{\beta}}$.

Lavenberg and Welch (1981) surveyed a wide range of issues, including the theory of CVs, the formation and selection of controls, and published applications. The greatest impact, however, came from establishing a framework in which the properties of $\overline{Y}(\hat{\boldsymbol{\beta}})$ can be derived. Suppose that $(Y, \mathbf{X})$ have a joint, nonsingular, multivariate normal distribution. Then Lavenberg and Welch (1981) showed that $\overline{Y}(\hat{\boldsymbol{\beta}})$ is an unbiased estimator of θ, a valid confidence interval for θ can be formed, and

$$\mathrm{Var}\big(\overline{Y}(\hat{\boldsymbol{\beta}})\big) = \left(\frac{n-2}{n-q-2}\right)\big(1 - \rho^2_{Y\mathbf{X}}\big)\frac{\sigma^2_Y}{n}. \tag{5}$$

The term $(n-2)/(n-q-2)$ became known as the "loss ratio" because it shows the penalty for having too many controls, and this penalty is independent of the controls' effectiveness. Expression (5) is perhaps the most well-known result on the theory of CVs, and it implies that it is very important to select a small number of effective controls.[3]

Lavenberg and Welch (1981) concluded their survey with a list of research directions for the future, including selection of controls (motivated by the loss ratio), use of batch means instead of replications for steady-state simulation, empirical studies of CVs using the regenerative method, and a plea for published applications of CVs. This list provided fodder for researchers for many years (roughly five years of the author's research life was spent working on the batching problem and remedies for nonnormality). An award-winning paper that picked up where Lavenberg and Welch (1981) left off is Wilson and Pritsker (1984b), which we discuss next.

[3] The loss ratio concept was later extended by Venkatraman and Wilson (1986) to the case of p output responses where it becomes $((n-2)/(n-q-2))^p$.

The multivariate-normal framework for CVs of Lavenberg and Welch (1981) provided a formal justification for the use of CVs and associated inference; thus, it makes sense to try to operate within that framework. Wilson and Pritsker (1984a, b) observed that the standard way of forming controls (averages of input random variables) was inconsistent with this goal. Suppose that the q controls $\mathbf{X} - \boldsymbol{\mu}$ are themselves sample means of stochastic input processes. Then as the run length increases, the variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$ becomes singular because the variance of the controls goes to 0; this will clearly cause problems when estimating $\boldsymbol{\beta}$, and it is disconcerting that difficulties should arise as sample sizes get larger. Wilson and Pritsker (1984a) proposed a framework in which "standardized" controls could be formed for queueing-network simulations that are asymptotically stable and satisfy the requirements of Lavenberg and Welch's (1981) multivariate-normal framework. For a queueing network with q stations, the standardized control associated with the kth station is

$$X_k(t) = \frac{\sum_{j=1}^{n_k(t)}(U_{jk} - \mu_k)}{\sigma_k\sqrt{n_k(t)}}$$

where $\{U_{jk}, j = 1, 2, \ldots\}$ are i.i.d. service times at station k, and $n_k(t)$ is the number of service times started at station k up to and including simulation time t, for $k = 1, 2, \ldots, q$. Under mild conditions the vector of controls $(X_1(t), X_2(t), \ldots, X_q(t))$ will be asymptotically multivariate normal as $t \to \infty$.

Wilson and Pritsker (1984b, the award-winning *Management Science* paper) provided a thorough empirical evaluation of these standardized controls in the spirit of Fishman's guidelines. They performed experiments on two variations of a closed queueing network and two variations of a mixed open and closed queueing network; and they considered estimation of the steady-state mean queue length, utilization, and response-time measures using up to four standardized controls. The evaluation was comprehensive, considering point-estimator variance (the primary goal of a VRT), confidence-interval half-length, and confidence-interval coverage. Both multiple-replication and single-replication regenerative experiment designs were employed, and the results were very encouraging. Although we have only described the contribution of this paper to research on CVs, Wilson and Pritsker (1984a, b) also contain what is perhaps the best treatment of poststratified sampling (another VRT) for discrete-event, stochastic simulation that has ever been published.

4.6. Gradient Estimation (Heidelberger et al. 1988)

In the 1980s there was a flurry of interest in going beyond estimating system performance via simulation to estimating the gradient of system performance

with respect to controllable input or decision variables. Gradients are useful in their own right as measures of sensitivity, and also as a key component of gradient-based optimization algorithms. A particularly active area of research was gradient estimation via perturbation analysis (PA), which had (and has) many variations, including infinitesimal perturbation analysis (IPA, see Heidelberger et al. 1988 for a long list of references, and especially Suri and Zazanis 1988, which is among the most-cited simulation papers in *Management Science*; see Fu 2001 for a current reference). IPA methods are intriguing because they make it possible to estimate gradients for multiple performance measures, and with respect to multiple-input variables, from a single simulation run. At the time, there were so many papers, talks, and applications of PA and IPA that it was difficult to keep track of exactly what was known and what was yet to be established. Of some importance to the analysis methodology community were the statistical properties of IPA estimators, specifically whether or not they converge to the true gradients as the simulation effort goes to infinity (this property is known as "consistency"). Heidelberger et al. (1988) developed a mathematical framework for establishing when IPA estimators are strongly consistent (converge with probability 1) that not only encompassed existing results, but also filled in gaps and revealed open research problems. The paper's rigorous treatment of the topic, while also managing to convey an intuitive understanding of the issues, made it an exceptional contribution to the research literature.

Suppose that a steady-state performance measure θ is a function of an input parameter x. For example, θ might be the long-run expected delay for customers in a queueing system, and x the arrival rate. We add a subscript to θ_x to denote this dependence. Along with an estimator of θ_x for any specific x, we would also like to estimate

$$\theta'_x \equiv \frac{d\theta_x}{dx}.$$

Estimating θ'_x presents a host of problems. Perhaps the most obvious approach is to create a finite difference estimate by making simulation runs at settings $x + \Delta x$ and x. Unfortunately, this introduces bias and requires at least $k + 1$ simulation runs if x is k-dimensional (we will only consider the scalar x case here, however). IPA is based on the idea that for small enough (say infinitesimally small) changes in x, the sample paths of the simulation change in small but predictable ways that can be tracked as the simulation at setting x progresses; therefore, no run at a perturbed value $x + \Delta x$ is required. IPA constructs a "sample path derivative," which is literally a derivative of the sample performance with respect to x.

Heidelberger et al. (1988) first considered systems with regenerative structure. From (1) the regenerative estimator of θ based on n complete regenerative cycles is denoted $\hat{\theta}(n)$. The paper sought to determine when the following holds with probability 1:

$$\theta'_x = \frac{d}{dx} \lim_{n\to\infty} \hat{\theta}_x(n) \stackrel{?}{=} \lim_{n\to\infty} \frac{d}{dx} \hat{\theta}_x(n)$$

where the quantity on the far right-hand side is the limit of the IPA estimator. Mathematical conditions that allow the exchange of the limit and differentiation are well known; the significant impact of the paper came from embedding them in a mathematical framework from which they could be interpreted for stochastic simulation problems.

Loosely speaking, IPA is strongly consistent if, for any sample path of the output process, there is a Δx that is small enough so that it is almost certain that the order in which events are executed in the simulation will not change. Yet for several specific cases in which this condition is not satisfied, IPA is nevertheless strongly consistent. A goal of the paper was to find out why.

Theorem 2.1 of Heidelberger et al. (1988) states "if and only if" conditions for strong consistency of IPA regenerative gradient estimators: Either the probability that events change order goes to zero faster than a critical rate as $\Delta x \to 0$, or two bias terms that represent what happens when events do change order precisely cancel each other. Of equal importance, the form of this result provides insight into which types of systems and performance measures will or will not yield strongly consistent estimators. The paper also presents results that do not depend on regenerative structure, focusing in particular on IPA throughput estimators, where throughput is the rate at which an event (such as departure of a class i customer from a queueing system) occurs over the long run. An empirical evaluation confirmed that consistency, or lack of it, is a key indicator of whether or not the IPA gradient estimator is useful in real problems.

Heidelberger et al. (1988), in concert with another *Management Science* paper by L'Ecuyer (1990) that presented a unified treatment of a number of gradient-estimation techniques, put *Management Science* in the forefront for establishing the foundations of this area.

4.7. Experiment Planning (Whitt 1989)

A central tactical issue in Conway (1963) is choosing the run length necessary to obtain performance estimates with adequate precision in steady-state simulation. To choose a run length you need a measure of the estimator variability. The statistics literature is replete with "known variance" procedures, often justified by past experience with a similar process. Simulators rarely find themselves with such experience.

Even if they did, the variance they need is not the marginal variance of an i.i.d. sequence of data, but (in the best case) the variance of the sample mean of a stationary time series of (generally dependent) data. Thus, as the earlier quote from Law and Kelton (1982) emphasized, if the goal is to determine the run length necessary to reach a prespecified precision, then sequential simulation procedures are usually required.

Fortunately, "usually" is not always. In the context of steady-state queueing simulation (including networks of queues), Whitt (1989) presented a method for approximating the run length required to achieve a prespecified absolute or relative width confidence interval for queueing performance measures *that can be applied prior to running any simulation of the model*. His approach applied heavy-traffic limits and associated diffusion approximations, along with adjustments based on tractable queueing models, to a large class of queueing systems.

In a nutshell, Whitt's (1989) idea was to develop simple approximations for the mean θ and the asymptotic variance σ^2 of the queueing output process of interest. Given these values, an approximate $(1-\alpha)100\%$ confidence interval for θ, as a function of the simulation run length t, has half-length $z_{1-\alpha/2}\sigma/\sqrt{t}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Therefore, given approximations for θ and σ^2, the value of t needed to achieve a given absolute or relative width can be determined.

To be concrete we will focus on the queue-length process $\{Y_\rho(t),\ 0 \le t \le T\}$. The additional subscript ρ denotes the traffic intensity of the queue (or bottleneck station in a network of queues), where the system is stable for $\rho < 1$, unstable for $\rho \ge 1$, and the congestion increases sharply as $\rho \to 1$. Let σ_ρ^2 denote the asymptotic variance of $\overline{Y}_\rho(T)$ as $T \to \infty$.

Whitt (1989) is a superb example of the stochastic processes approach to simulation output analysis. Supported by a large literature on heavy-traffic queueing analysis, Whitt took as his central approximation that as $\rho \to 1$ an appropriately standardized version of $Y_\rho(t)$ converges in distribution to a limiting stochastic process that depends on only two parameters. Even better, the asymptotic variance of this limiting process is a simple function of the two parameters. And best of all, *for many queueing systems the values of these two parameters can be determined prior to simulating, using only information that must be available to build the simulation model*.

The limiting process in Whitt's analysis is regulated Brownian motion (RBM), which is Brownian motion on the positive real line with negative drift, a, positive diffusion coefficient, d, and reflecting barrier at 0. For a queue-length process, the state of RBM corresponds to the number in queue and the barrier represents an empty queue. The particular limit that Whitt employed is

$$(1-\rho)Y_\rho(t(1-\rho)^{-2}) \stackrel{\rho\to 1}{\Longrightarrow} \mathcal{R}(t; a, d)$$

where $\mathcal{R}$ is RBM.

Stationary RBM is well studied: Its expected value is $d/(2|a|)$, and the asymptotic variance of the sample mean of RBM is $\sigma_R^2 = d^3/(2a^4)$. These results can be used to derive the approximations

$$\theta_\rho \approx \frac{d}{2|a|(1-\rho)} \tag{6}$$

$$\sigma_\rho^2 \approx \frac{d^3}{2a^4(1-\rho)^4}. \tag{7}$$

As a refinement, Whitt (1989, p. 1354) multiplied the right-hand sides of (6)–(7) by ρ^2 to make them match up better with known results for the $M/G/1$ queue. The paper also shows how a and d can be obtained for $GI/G/m$ and $G/G/1$ queues, queues with interrupted service and open queueing networks.

4.8. Importance Sampling (Shahabuddin 1994)

Similar to Lavenberg and Welch (1981) and Wilson and Pritsker (1984b), Shahabuddin (1994) addresses statistical efficiency in simulation experiments. The goal in this case is more specific: to estimate the mean time to failure (MTTF) or the steady-state unavailability (long-run fraction of time that the system is not usable) of a highly reliable system. The VRT is a customized version of importance sampling (IS), which we describe briefly below. In addition to the College on Simulation award, the paper also received the Nicholson Prize from the Institute for Operations Research and the Management Sciences as the best student paper in 1990 (see www.informs.org/Prizes/NicholsonPrize.html for additional information). IS is also the topic of one of the most-cited simulation papers in *Management Science*, Glynn and Iglehart (1989).

Imagine using simulation to estimate the probability that a really unlikely event occurs. Because the event is rare (e.g., say probability on the order of 10^{-9} of occurring), an excessive number of simulated trials would be required to observe even a small number of these events. Conducting so many trials would be impossible if each trial required even a moderate amount of simulation effort. But suppose that you could change the simulation model in such a way that the rare event occurred much more often, say exactly 1,000,000 times more often. Then you would observe the event more frequently, giving you a much better estimator of the probability that it occurs. Of course,

your estimator would be *wrong*, but you know exactly how wrong (1,000,000 times), so you can correct for the bias. This is essentially the idea behind IS, and the correction factor is called the likelihood ratio. Unfortunately, if the probability dynamics of the system are changed crudely to increase the frequency of the rare event, then the result is often a variance increase; hence, IS is much more difficult than it might first appear.

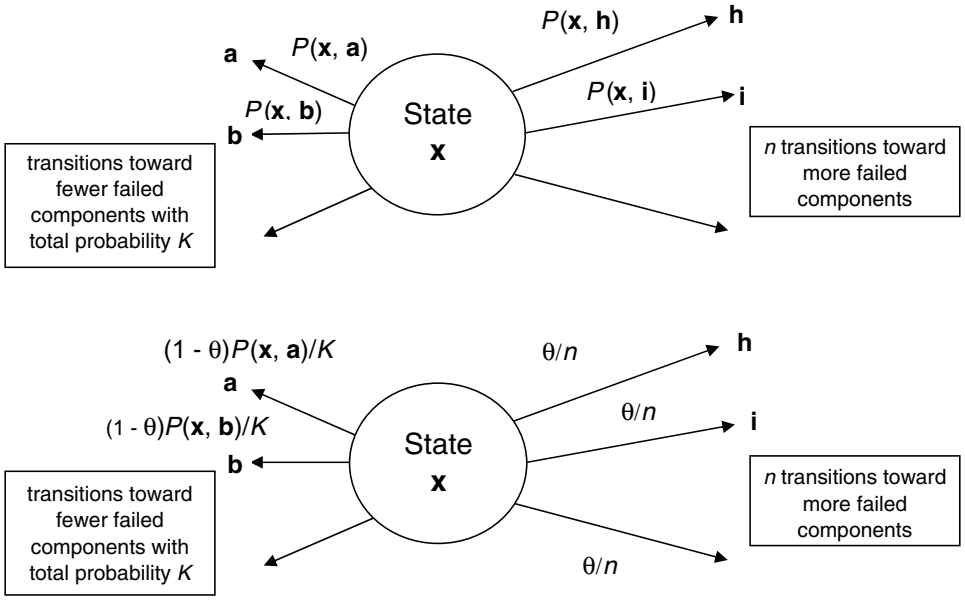
Shahabuddin (1994) considered systems of many components that are subject to random failure, but can also be repaired. When enough components in certain combinations fail, then the entire system is unavailable. A system is highly reliable if the component failure rates are tiny relative to the repair rates. Let $\mathbf{X}(t)$ be a $c \times 1$ vector of random variables representing the status of the components at time t. Shahabuddin (1994) considered systems for which $\{\mathbf{X}(t), t \geq 0\}$ is a continuous-time Markov chain (CTMC) whose generator $\mathbf{Q}$ satisfies some sensible conditions. The reason for using simulation instead of numerical methods is the very large dimension of $\mathbf{Q}$ for even moderate values of c, and the ease of simulating $\mathbf{X}(t)$ (see Shahabuddin 1994 for details).

IS for such problems is conceptually simple: Change the failure rates (and perhaps the repair rates) to make a system's failure much more likely (e.g., replace the generator $\mathbf{Q}$ by a different generator $\mathbf{Q}'$ that has larger component failure rates and smaller component repair rates). Of course, this alters the probability measure of the stochastic process, but the correction can be computed similarly to the simple example above. The idea is very powerful, but the reality is that finding a $\mathbf{Q}'$ (or more complex change of probability measure) that guarantees a variance reduction in a dynamic, stochastic simulation is not easy, and variance increases are possible. However, Shahabuddin's problem was not one of finding a good change of measure; certain "failure-biasing" heuristics (that make component failures more likely) had been observed to work well for CTMCs with certain characteristics (e.g., Goyal et al. 1992). Shahabuddin's question was why did they work, and which heuristic could be expected to work for which problems?

In this type of simulation it is well known that the CTMC itself need not be simulated, only the embedded discrete-time Markov chain (DTMC) with transition matrix $\mathbf{P}$ (Hordijk et al. 1976, Fox and Glynn 1986). Figure 2 illustrates one of the IS heuristics called "balanced failure biasing." The top figure shows a state $\mathbf{x}$ in the original DTMC along with transition probabilities $P(\mathbf{x}, \cdot)$ that move the process to states with more failed components and transitions that move the process to states with fewer failed components (for simplicity the figure does not show transitions to states with the same number of failed components). The lower figure shows the transition probabilities after balanced failure biasing. A fixed value of the parameter θ between 0 and 1 is used to make the failure transitions more likely ($\theta = 1/2$ is often used in practice).

Let μ be the minimum component repair rate, which without loss of generality can be taken to be 1; and let λ be the largest component failure rate. Shahabuddin's (1994) approach was to model the component failure rates as a function of λ; specifically, the ith component failure rate takes the form $c_i\lambda^{d_i}$, where $0 < c_i \leq 1$, $c_i \approx 1$, and $d_i > 0$. If the d_i param-

Figure 2 Balanced Failure Biasing for the Embedded DTMC



eters for all of the components are approximately 1, then the system is considered balanced; otherwise it is unbalanced. Shahabuddin (1994) then examined the performance of different failure-biasing heuristics as $\lambda \to 0$ (component failures become more and more rare). In particular, he established whether different failure-biasing heuristics lead to estimators of MTTF or unavailability with bounded relative error. Relative error is represented by the half-length of a confidence interval for the unknown parameter divided by the value of that parameter. The standard (no IS) estimator has unbounded relative error.[4] An IS heuristic that results in bounded relative error can be expected to work well in practice; in fact, it gets more efficient as the problem gets harder. The remarkable feature of this paper is the clever representation of the failure rates as a function of the largest failure rate, and how little else need be assumed about the structure of the CTMC to obtain such deep results. The conclusions of this and related work have been implemented in IBM's SAVE availability modeling package.

4.9. A Postscript

A reader of this review might incorrectly conclude that there was no simulation research published in *Management Science* between Conway's paper in 1963 and Schruben's paper in 1980. In fact, there was substantial activity including influential papers by Fishman and Kiviat (1967), Burt and Garman (1971), and many others. Research topics that would not be classified as analysis methodology have also appeared throughout the years, including papers on verification and validation of simulation models (e.g., Naylor and Finger 1967 and Van Horn 1971, both among the most-cited papers) and modeling paradigms (e.g., Sargent 1988). A search on the keyword "simulation" at pubsonline.informs.org provides a long list.

5. The Future

The field of analysis methodology in stochastic simulation has matured to the point where it is impossible for anyone to define the research problems for the next 40 years in the way Conway et al. (1959) and Conway (1963) did some 40 years ago. However, there are two directions that seem obvious, and one that is less obvious but that could (and should) develop.

1. The distinction between simulation analysis and applied probability research will become even less clear in the future. Selfishly, we might like to once again have a standalone simulation department in *Management Science*, but the single stochastic models and simulation department probably makes sense. Applied probability and simulation researchers are both interested in formulating and evaluating stochastic models, and the evaluation of increasingly complex models often requires mathematical analysis, numerical approximations, and stochastic simulation in various proportions. Nowhere is this more apparent than in the emerging area of financial engineering and quantitative finance (e.g., Glasserman et al. 2000). A key to making a single department work is having editors like Glasserman and Shahabuddin who have a broad perspective.

2. Recall the comment in Conway et al. (1959) cited in §2 that a simulation "is not inherently optimizing; rather it is descriptive of the performance of a given configuration of the system. Optimization must be superimposed upon this model by varying the configuration in search of a maximum of performance." The problem of optimization of simulation—optimizing the expected or long-run average performance of a system that is represented by a simulation model—is one in which experiment design, statistical efficiency and even data management all come into play. Realistic problems that are attacked with methods that do not account for the stochastic nature of simulation can yield solutions that are far from optimal. On the other hand, accounting for randomness inefficiently can exhaust the time that is available to solve the problem before much of the feasible space is explored. Research progress in this area has been slow because the approach has been to thoroughly examine small pieces of the problem (e.g., gradient estimation). We may now have enough pieces in place to begin to assemble rigorously justified, but practically workable, algorithms (e.g., Andradóttir 1996, Fu 2002). Practitioners want to optimize, so it is important that we do this.

3. Optimization of simulation is a problem that taxes computing resources, but there are many situations in which the scarce resource is not computer time, but rather decision-maker time. Simulation is popular because it can incorporate any details that are important, and the now common practice of animating simulations means that they have a face validity that a system of equations can never hope to achieve. Unfortunately, simulation can be a clumsy tool for planning when the work is done interactively, perhaps by a group of decision makers who need to consider political, as well as performance, issues. Even a few minutes per simulation run is too slow to allow what-if analysis in real time. Optimization of simulation does not solve this problem because an objective function must be formulated, which hinders the

[4] For the crude estimator of a probability p based on n i.i.d. samples, the relative error is

$$z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{p\sqrt{n}} \xrightarrow{p\to 0} \infty.$$

That is, as the probability of interest gets smaller, the relative error is unbounded.

decision maker's ability to consider trade-offs that are not easily quantified. As researchers we need to focus attention on the efficiency of obtaining useful simulation results, as well as on the efficiency of the simulation run itself. One approach is to use simulation to parameterize sophisticated metamodels that are easily explored or optimized with respect to the controllable decision variables. Although it might take days of simulation on multiple processors to build the metamodel, if it supports a productive one-hour meeting of highly compensated managers it is well worth it. The experiment design, metamodeling, and run-control tools to precisely map a large, complex response surface, without overfitting, do not yet exist, nor do the tools to query the result. They should.

Acknowledgments

The author gratefully acknowledges the help of George Fishman, Michael Fu, Dave Goldsman, Philip Heidelberger, Shane Henderson, David Kelton, Pierre L'Ecuyer, Richard Nance, Robert Sargent, Lee Schruben, Perwez Shahabuddin, Peter Welch, Ward Whitt, and James Wilson in preparing this paper.

References

Andradóttir, S. 1996. Optimization of the transient and steady-state behavior of discrete event systems. *Management Sci.* **42** 717–737.

Burt, J. M., M. B. Garman. 1971. Conditional Monte Carlo: A simulation technique for stochastic network analysis. *Management Sci.* **18** 207–217.

Conway, R. W. 1963. Some tactical problems in digital simulation. *Management Sci.* **10** 47–61.

Conway, R. W., B. M. Johnson, M. L. Maxwell. 1959. Some problems of digital simulation. *Management Sci.* **6** 92–110.

Fishman, G. S. 1980. *Newsletter of the TIMS College on Simulation and Gaming* **4**(3) 4–5.

Fishman, G. S., P. J. Kiviat. 1967. The analysis of simulation-generated time series. *Management Sci.* **13** 525–557.

Fox, B. L., P. W. Glynn. 1986. Discrete-time conversion for simulating semi-Markov processes. *Oper. Res. Lett.* **5** 191–196.

Fu, M. C. 2001. Perturbation analysis. S. Gass, C. Harris, eds. *Encyclopedia of Operations Research and Management Science*, 2nd ed. Kluwer Academic Publishers, 608–611.

Fu, M. C. 2002. Optimization for simulation: Theory vs. practice. *INFORMS J. Comput.* **14** 192–215.

Gafarian, A. V., C. J. Ancker, T. Morisaku. 1978. Evaluation of commonly used rules for detecting steady-state in computer-simulation. *Naval Res. Logist.* **25** 511–529.

Glasserman, P., P. Heidelberger, P. Shahabuddin. 2000. Variance reduction techniques for estimating value-at-risk. *Management Sci.* **46** 1349–1364.

Glynn, P. W., D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Sci.* **35** 1367–1392.

Glynn, P. W., D. L. Iglehart. 1993. Conditions for the applicability of the regenerative method. *Management Sci.* **39** 1108–1111.

Goldsman, D., M. Meketon, L. Schruben. 1990. Properties of standardized time series weighted area variance estimators. *Management Sci.* **36** 602–612.

Goyal, A., P. Shahabuddin, P. Heidelberger, V. F. Nicola, P. W. Glynn. 1992. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Trans. Comput.* **41** 36–51.

Heidelberger, P., X. Cao, M. A. Zazanis, R. Suri. 1988. Convergence properties of infinitesimal perturbation analysis estimates. *Management Sci.* **34** 1281–1302.

Henderson, S. G. 2000. Mathematics for simulation. J. A. Joines, R. R. Barton, K. Kang, P. A. Fishwick, eds. *Proc. 2000 Winter Simulation Conf.* IEEE, Piscataway, NJ, 137–146.

Hopp, W. 2004. Fifty years of *Management Science. Management Sci.* **50** 1–7.

Hordijk, A., D. L. Iglehart, R. Schassberger. 1976. Discrete-time methods for simulating continuous time Markov chains. *Adv. Appl. Probab.* **8** 772–778.

Iglehart, D. L., P. A. W. Lewis. 1979. Regenerative simulation with internal controls. *J. ACM* **26** 271–282.

Kleijnen, J. P. C. 1988. Analyzing simulation experiments with common random numbers. *Management Sci.* **34** 65–74.

Lavenberg, S. S., P. D. Welch. 1981. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Sci.* **27** 322–335.

Lavenberg, S. S., T. L. Moeller, C. H. Sauer. 1979. Concomitant control variables applied to the regenerative simulation of queueing systems. *Oper. Res.* **27** 134–160.

Lavenberg, S. S., T. L. Moeller, P. D. Welch. 1982. Statistical results on control variables with application to queueing network simulation. *Oper. Res.* **30** 182–202.

Law, A. M., W. D. Kelton. 1982. Confidence interval procedures for steady-state simulations, II: A survey of sequential procedures. *Management Sci.* **28** 550–562.

L'Ecuyer, P. 1990. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Sci.* **36** 1364–1383.

Meketon, M. S., P. Heidelberger. 1982. A renewal theoretic approach to bias reduction in regenerative simulations. *Management Sci.* **28** 173–181.

Nance, R. E., R. G. Sargent. 2002. Perspectives on the evolution of simulation. *Oper. Res.* **50** 161–172.

Naylor, T. H., J. M. Finger. 1967. Verification of computer simulation models. *Management Sci.* **14** B92–B101.

Nelson, B. L., F. J. Matejcik. 1995. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Sci.* **41** 1935–1945.

Newsletter of the TIMS College on Simulation and Gaming. 1980. **4**(3) 3–4.

Newsletter of the TIMS College on Simulation and Gaming. 1985. **9**(2) 3–4.

Sargent, R. G. 1988. Event graph modelling for simulation with an application to flexible manufacturing systems. *Management Sci.* **34** 1231–1251.

Schmeiser, B. 1982. Batch size effects in the analysis of simulation output. *Oper. Res.* **30** 556–568.

Schmeiser, B., Y. Yeh. 2002. On choosing a single criterion for confidence-interval procedures. E. Yücesan, C.-H. Chen, J. L. Snowdon, J. M. Charnes, eds. *Proc. 2002 Winter Simulation Conf.* IEEE, Piscataway, NJ, 345–352.

Schruben, L. W. 1980. A coverage function for interval estimators of simulation response. *Management Sci.* **26** 18–27.

Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable Markovian systems. *Management Sci.* **40** 333–352.

Steiger, N. M., J. R. Wilson. 2002. An improved batch means procedure for simulation output analysis. *Management Sci.* **48** 1569–1586.

Suri, R., M. A. Zazanis. 1988. Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue. *Management Sci.* **34** 39–64.

Van Horn, R. L. 1971. Validation of simulation results. *Management Sci.* **17** 247–258.

Venkatraman, S., J. R. Wilson. 1986. The efficiency of control variates in multiresponse simulation. *Oper. Res. Lett.* **5** 37–42.

Whitt, W. 1989. Planning queueing simulations. *Management Sci.* **35** 1341–1366.

Wilson, J. R., A. A. B. Pritsker. 1984a. Variance reduction in queueing simulation using generalized concomitant variables. *J. Statist. Comput. Simulation* **19** 129–153.

Wilson, J. R., A. A. B. Pritsker. 1984b. Experimental evaluation of variance reduction techniques for queueing simulation using generalized concomitant variables. *Management Sci.* **30** 1459–1472.