Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

OUTPUT ANALYSIS PROCEDURES FOR COMPUTER SIMULATIONS

David Goldsman

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.

Gamze Tokol

Earley Corporation
130 Krog St.
Atlanta, GA 30307, U.S.A.

ABSTRACT

This paper concerns the statistical analysis of output from discrete-event computer simulations. In particular, we discuss problems involving terminating simulations, the initialization of simulations, steady-state point and confidence interval estimation for various system parameters, and comparison among competing system designs.

1 INTRODUCTION

Since the input processes driving a simulation are usually random variables (e.g., interarrival times, service times, and breakdown times), a prudent experimenter must also regard the output from the simulation as random. Thus, runs of the simulation only yield *estimates* of measures of system performance (e.g., the mean customer waiting time). Of course, these estimators are themselves random variables, and are therefore subject to sampling error. This sampling error must be taken into account in a rigorous way if we are to make valid inferences or decisions concerning the performance of the underlying system.

The fundamental problem is that simulations almost never produce raw output that is independent and identically distributed (i.i.d.) normal data. For example, consecutive customer waiting times from a complicated queueing system

- Are not independent — typically, they are serially correlated. If one customer at the post office waits in line a long time, then the next customer is also likely to wait a long time.
- Are not identically distributed. Customers showing up early in the morning might have a much shorter wait than those who show up just before closing time.
- Are not normally distributed — they are usually skewed to the right (and are certainly never less than zero).

These facts of life make it difficult to apply "classical" statistical techniques to the analysis of simulation output. And so our purpose in this survey is to give practical methods to perform statistical analysis of output from discrete-event computer simulations.

In order to facilitate the presentation, we identify two types of simulations with respect to output analysis.

1. *Terminating (or transient) simulations*. Here, the nature of the problem explicitly defines the length of the simulation run. For instance, we might be interested in simulating a bank that closes at a specific time each day.
2. *Nonterminating (steady-state) simulations*. Here, the long-run behavior of the system is studied. Presumedly this "steady-state" behavior is independent of the simulation's initial conditions. An example is that of a continuously running production line for which the experimenter is interested in some long-run performance measure.

Techniques to analyze output from terminating simulations are primarily based on the method of independent replications, discussed in §2. Additional problems arise for steady-state simulations. One must now worry about the problem of starting the simulation — how should it be initialized at time zero, and how long must it be run before data representative of steady state can be collected? Such initialization problems are considered in §3. Then §4 deals with methods of point and confidence interval estimation for steady-state simulation performance parameters. §5 concerns the problem of comparing a number of competing systems, i.e., which is the "best" system? §6 presents conclusions and provides the interested reader with additional references. Finally, we note that parts of this paper follow the discussions in Goldsman (1992) and Goldsman and Tokol (1997).

2 TERMINATING SIMULATIONS

Suppose we are conducting a terminating simulation, i.e., we simulate some system of interest over a finite time horizon. For now, assume that we obtain *discrete* simulation output $Y_1, Y_2, \ldots, Y_m$, where the number of observations m can be a constant or a random variable. For example, the experimenter can specify the number m of customer waiting times $Y_1, Y_2, \ldots, Y_m$ to be taken from a queueing simulation; or m could denote the random number of customers observed during a specified time period $[0, T]$. Alternatively, we might observe *continuous* simulation output $\{Y(t)|0 \le t \le T\}$ over a specified interval $[0, T]$; for instance, if we are interested in estimating the time-averaged number of customers waiting in a queue during $[0, T]$, the quantity $Y(t)$ would be the number of customers in the queue at time t.

Suppose for simplicity that the goal of the experiment is to estimate the expected value of the sample mean of the observations,

$$\theta \equiv \mathsf{E}[\bar{Y}_m],$$

where the sample mean in the discrete case is

$$\bar{Y}_m \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i$$

(with a similar expression for the continuous case). For example, we might be interested in estimating the expected average waiting time of all customers at a shopping center during the period 10 a.m. to 2 p.m.

Although $\bar{Y}_m$ is an unbiased estimator for θ, a proper statistical analysis requires that we also provide an estimate of $\mathsf{Var}(\bar{Y}_m)$. Since the Y_i's are not necessarily i.i.d. random variables, it is may be that $\mathsf{Var}(\bar{Y}_m) \ne \mathsf{Var}(Y_i)/m$ for any i, a case not covered in elementary statistics textbooks. For this reason, the familiar sample variance, $S^2 \equiv \sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2/(m-1)$, is likely to be highly *biased* as an estimator of $m\mathsf{Var}(\bar{Y}_m)$; and therefore one should *not* use S^2/m to estimate $\mathsf{Var}(\bar{Y}_m)$.

The way around the problem is via the method of *independent replications*. This method estimates $\mathsf{Var}(\bar{Y}_m)$ by conducting b independent simulation runs (replications) of the system under study, where each replication consists of m observations. It is easy to make the replications independent — just reinitialize each replication with a different pseudo-random number seed.

Let us denote the sample mean from replication i by

$$Z_i \equiv \frac{1}{m}\sum_{j=1}^{m} Y_{i,j},$$

where $Y_{i,j}$ is observation j from replication i, for $i = 1, 2, \ldots, b$ and $j = 1, 2, \ldots, m$. If each run is started under the same operating conditions (e.g., all queues empty and idle), then the replication sample means $Z_1, Z_2, \ldots, Z_b$ are *i.i.d.* random variables, and an obvious point estimator for $\mathsf{Var}(\bar{Y}_m) = \mathsf{Var}(Z_i)$ is

$$\hat{V}_R \equiv \frac{1}{b-1}\sum_{i=1}^{b}(Z_i - \bar{Z}_b)^2,$$

where the grand mean is defined as

$$\bar{Z}_b \equiv \frac{1}{b}\sum_{i=1}^{b} Z_i.$$

Notice how closely the forms of $\hat{V}_R$ and S^2/m resemble each other. But since the replicate sample means are i.i.d., $\hat{V}_R$ is usually much less biased for $\mathsf{Var}(\bar{Y}_m)$ than is S^2/m.

In light of the above, we see that $\hat{V}_R/b$ is a reasonable estimator for $\mathsf{Var}(\bar{Z}_b)$. If the number of observations per replication, m, is large enough, a central limit theorem tells us that the replicate sample means are approximately i.i.d. *normal*. Then we have an approximate $100(1-\alpha)\%$ two-sided confidence interval (CI) for θ,

$$\theta \in \bar{Z}_b \pm t_{\alpha/2,b-1}\sqrt{\hat{V}_R/b}, \tag{1}$$

where $t_{\alpha/2,b-1}$ is the $1-\alpha/2$ quantile of the t-distribution with $b-1$ degrees of freedom.

EXAMPLE Suppose we want to estimate the expected average waiting time for the first 5000 customers in a certain queueing system. We will make five independent replications of the system, with each run initialized empty and idle and consisting of 5000 waiting times. The resulting replicate means are:

i	1	2	3	4	5
Z_i	3.2	4.3	5.1	4.2	4.6

Then $\bar{Z}_5 = 4.28$ and $\hat{V}_R = 0.487$. For level $\alpha = 0.05$, we have $t_{0.025,4} = 2.78$, and (1) gives $[3.41, 5.15]$ as a 95% CI for the expected average waiting time for the first 5000 customers.

Independent replications can be used to calculate variance estimates for statistics other than sample means; and then the method can be used to get CI's for quantities other than $\mathsf{E}[\bar{Y}_m]$, e.g., quantiles. See any of the standard texts cited in §6 for additional uses of independent replications.

3 INITIALIZATION PROBLEMS

Before a simulation can be run, one must provide initial values for all of the simulation's state variables. Since the experimenter may not know what initial values are appropriate for the state variables, these values might be chosen somewhat arbitrarily. For instance, we might decide that it is "most convenient" to initialize a queue as empty and idle. Such a choice of initial conditions can have a significant but unrecognized impact on the simulation run's outcome. Thus, the *initialization bias* problem can lead to errors, particularly in steady-state output analysis.

We give several examples of problems concerning simulation initialization.

- Visual detection of initialization effects is sometimes difficult — especially in the case of stochastic processes having high intrinsic variance such as queueing systems.
- How should the simulation be initialized? Suppose that a machine shop closes at a certain time each day, even if there are jobs waiting to be served. One must therefore be careful to start each day with a demand that depends on the number of jobs remaining from the previous day.
- Initialization bias can lead to point estimators for steady-state parameters having high mean squared error, as well as CI's having poor coverage.

Since initialization bias raises important concerns, one must ask how to detect and deal with it? We first list methods to detect the presence of bias.

Attempt to detect the bias visually by scanning a realization of the simulated process. This might not be an easy task, since visual analysis can miss bias that happens to be present. Further, in a simulation study with a large number of runs, a visual scan can be very tedious. In order to make the visual analysis a bit more efficient, one might transform the data (e.g., by taking logs or square roots), smooth it (see Welch 1981, 1983), average it across several independent replications, or construct CUSUM plots (Schruben 1982).

Conduct statistical tests for initialization bias. Kelton and Law (1983) give an intuitively appealing sequential procedure to detect bias. Goldsman, Schruben, and Swain (1994), Schruben (1982), and Schruben, Singh, and Tierney (1983) present tests that check to see whether the initial portion of the simulation output contains more variation than latter portions.

If initialization bias is detected, one may want to do something about it. Two simple methods for dealing with bias have been suggested.

Truncate the output by allowing the simulation to "warm up" before data are retained for analysis. The experimenter hopes that the remaining data are representative of the steady-state system. Output truncation is probably the most popular method for dealing with initialization bias; and all of the major simulation languages have built-in truncation functions. But how can one find a good truncation point? If the output is truncated "too early," significant bias might still exist in the remaining data. If it is truncated "too late," then good observations might be wasted. Unfortunately, Wilson and Pritsker (1978ab) find that simple rules to determine truncation points do not seem to perform well in general. A common practice is to average observations across several replications, and then visually choose a truncation point based on the averaged run; see Welch (1983) for a nice visual/graphical approach.

Make a very long run to overwhelm the effects of initialization bias. This method of bias control is conceptually simple to carry out and may yield point estimators having lower mean squared errors than the analogous estimators from truncated data (see, e.g., Fishman 1978). However, a problem with this approach is that it can be wasteful with observations; for some systems, an excessive run length might be required before the initialization effects are rendered negligible.

4 STEADY-STATE ANALYSIS

We henceforth assume that we have on hand stationary (steady-state) simulation output, $Y_1, Y_2, \ldots, Y_n$. Suppose our goal is to estimate some parameter of interest, possibly the mean customer waiting time or the expected profit produced by a certain factory configuration. As in the case of terminating simulations, it is of the utmost importance to accompany the value of any point estimator with a measure of its variance.

A number of methodologies have been proposed in the literature for conducting steady-state output analysis. We will examine the two most popular: batch means and independent replications. (As discussed in §2, confidence intervals for *terminating* simulations usually use independent replications.)

4.1 Batch Means

The method of batch means is often used to estimate $\mathsf{Var}(\bar{Y}_n)$ or calculate CI's for the steady-state process mean μ. The idea is to divide one long simulation run into a number of contiguous *batches*, and then appeal to a central limit theorem to assume that the resulting batch sample means are approximately i.i.d. normal. In particular, suppose that we partition $Y_1, Y_2, \ldots, Y_n$ into b nonoverlapping, contiguous batches, each consisting of m observations (assume that $n = bm$). Thus, the ith batch consists of the random variables

$$Y_{(i-1)m+1}, Y_{(i-1)m+2}, \ldots, Y_{im},$$

$i = 1, 2, \ldots, b$. The ith batch mean is the sample mean of the m observations from batch i,

$$Z_i \equiv \frac{1}{m}\sum_{j=1}^{m} Y_{(i-1)m+j},$$

$i = 1, 2, \ldots, b$. Similar to independent replications (as described in §2), we define the batch means estimator for $\mathsf{Var}(Z_i)$ as

$$\hat{V}_B \equiv \frac{1}{b-1}\sum_{i=1}^{b}(Z_i - \bar{Z}_b)^2,$$

where

$$\bar{Y}_n = \bar{Z}_b \equiv \frac{1}{b}\sum_{i=1}^{b} Z_i$$

is the grand sample mean. If m is large, then the batch means are approximately i.i.d. *normal*, and (as in §2) we obtain an approximate $100(1-\alpha)\%$ CI for μ,

$$\mu \in \bar{Z}_b \pm t_{\alpha/2,b-1}\sqrt{\hat{V}_B/b}.$$

This equation is very similar to (1). Of course, the difference here is that batch means divides one long run into a number of batches, whereas independent replications uses a number of independent shorter runs.

To illustrate, we refer the reader to the example from §2 with the understanding that the Z_i's must now be regarded as batch means (instead of replicate means); then the same numbers carry through the example.

The technique of batch means is intuitively appealing and easy to understand. But problems can come up if the Y_j's are not stationary (e.g., if significant initialization bias is present), if the batch means are not normal, or if the batch means are not independent. If any of these assumption violations exist, poor confidence interval coverage may result — unbeknownst to the analyst. To ameliorate the initialization bias problem, the user can truncate some of the data or make a long run as discussed in §3. In addition, the lack of independence or normality of the batch means can be countered by increasing the batch size m. For more information on batch means, the reader should see Chien, Goldsman, and Melamed (1997), Schmeiser (1982), or the various textbooks cited herein.

4.2 Independent Replications

Of the difficulties encountered when using batch means, one can argue that the possibility of correlation among the batch means is the most troublesome. This problem is explicitly avoided by the method of independent replications, described in the context of terminating simulations in §2. The replicate means are independent by their construction. Unfortunately, since *each* of the b replications has to be started properly, initialization bias presents more trouble when using independent replications than when using batch means. Further, as in the case of batch means, we cannot guarantee the normality of the replicate means (although this is not usually a serious problem). These problems luckily disappear as the replicate size m becomes large. Nevertheless, for moderate m, Law and Kelton (1984) are concerned enough about the possibility of initialization bias in each of the replications to recommend the use of batch means over independent replications.

4.3 Other Methods

There are several other methods for obtaining variance estimators for the sample mean and CI's for the steady-state process mean μ.

Spectral Estimation. The spectral method estimates $\mathsf{Var}(\bar{Y}_n)$ (as well as the analogous CI's for μ) in a manner completely different from that of batch means. In particular, this approach operates in the so-called *frequency domain*, whereas batch means uses the *time domain*. Spectral estimation sometimes takes a little effort, but it works well enough to suggest that the reader consult the relevant references, e.g., Heidelberger and Welch (1981, 1983). Meketon and Schmeiser (1984) developed the method of *overlapping batch means*, which links together the best aspects of the batch means and spectral approaches.

Regeneration. Many simulations can be broken into i.i.d. blocks or groups that probabilistically "start over" at certain *regeneration* points. An elementary example of this phenomenon is an M/M/1 queue's waiting time process, where the i.i.d. blocks are defined by groups of customers whose endpoints have zero waiting times. The method of regeneration uses this i.i.d. structure and, under certain conditions, produces excellent estimators for $\mathsf{Var}(\bar{Y}_n)$ and CI's for μ. The method effectively eliminates any initialization problems. On the other hand, it may be difficult to define natural regeneration points, and it is frequently the case that *extremely* long simulation runs are needed to obtain a reasonable number of i.i.d. blocks. For more details, see the fundamental references Crane and Iglehart (1975) and Crane and Lemoine (1977).

Standardized Time Series. One often uses the central limit theorem to standardize i.i.d. random variables into an (asymptotically) normal random variable. Schruben (1983) generalizes this idea by using a *process* central limit theorem to standardize a stationary simulation process into a *Brownian bridge* process. Properties of Brownian bridges are then used to calculate a number of good estimators for $\mathsf{Var}(\bar{Y}_n)$ and CI's for μ. This method is easy to apply and has some asymptotic advantages over batch means (see

Goldsman and Schruben 1984); however, long simulations may be needed before the necessary asymptotics kick in (Sargent, Kang, and Goldsman 1992).

5 COMPARISON OF SYSTEMS

One of the most important uses of simulation output analysis regards the comparison of competing systems or alternative system configurations. For instance, we might want to evaluate two different "re-start" strategies that an airline can evoke following a major traffic disruption such as a snowstorm in the Northeast — which policy minimizes a certain cost function associated with the re-start?

Simulation is uniquely equipped to help the experimenter conduct this type of comparison analysis. The current section discusses four techniques that are useful for the problem of selecting the best of a number of systems: (i) classical statistical CI's, (ii) common random numbers, (iii) antithetic variates, (iv) and ranking, selection, and multiple comparisons procedures.

5.1 Classical Confidence Intervals

With our airline example in mind, we will use $Z_{i,j}$ in this section to denote the cost from the jth simulation replication of strategy i, $i = 1, 2$, $j = 1, 2, \ldots, b_i$. We shall assume that $Z_{i,1}, Z_{1,2}, \ldots, Z_{1,b_i}$ are i.i.d. normal with unknown mean μ_i and unknown variance, $i = 1, 2$.

How can we justify these assumptions? As in §2,

- We obtain independent data by controlling the random numbers between replications.
- We get identically distributed costs between replications by performing the replications under identical conditions.
- We end up with approximately normally distributed data by adding up (or averaging) many sub-costs to get overall costs for both strategies.

Our goal is to obtain a $100(1-\alpha)\%$ CI for the difference $\mu_1 - \mu_2$. To do so, we also assume for now that the $Z_{1,j}$'s are independent of the $Z_{2,j}$'s and define the respective sample means and variances as

$$\bar{Z}_{i,b_i} \equiv \frac{1}{b_i}\sum_{j=1}^{b_i} Z_{i,j}, \quad i = 1, 2,$$

and

$$S_i^2 \equiv \frac{1}{b_i - 1}\sum_{j=1}^{b_i}(Z_{i,j} - \bar{Z}_{i,b_i})^2, \quad i = 1, 2.$$

An approximate $100(1-\alpha)\%$ CI is

$$\mu_1 - \mu_2 \in \bar{Z}_{1,b_1} - \bar{Z}_{2,b_2} \pm t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{b_1} + \frac{S_2^2}{b_2}}$$

where the (approximate) degrees of freedom ν is given in any standard statistics text.

Suppose, as in the airline example, that small cost is good. Then we can interpret the above CI as follows: If the interval lies entirely to the left [right] of zero, then system 1 [2] is better; if the interval contains zero, then the two systems must be regarded, in a statistical sense, as about the same.

An alternative classical strategy is to use a CI that is analogous to a paired-t test. Here we take b replications from *both* strategies and set the difference $D_i \equiv Z_{1,j} - Z_{2,j}$ for $j = 1, 2, \ldots, b$. Then calculate the sample mean and variance of the differences:

$$\bar{D}_b \equiv \frac{1}{b}\sum_{j=1}^{b} D_j \quad \text{and} \quad S_D^2 \equiv \frac{1}{b-1}\sum_{j=1}^{b}(D_j - \bar{D}_b)^2.$$

The resulting $100(1-\alpha)\%$ CI is

$$\mu_1 - \mu_2 \in \bar{D}_b \pm t_{\alpha/2,b-1}\sqrt{S_D^2/b}.$$

These paired-t intervals are amenable to an efficiency trick if $\text{Corr}(Z_{1,j}, Z_{2,j}) > 0$, $j = 1, 2, \ldots, b$ (where we still assume that $Z_{1,1}, Z_{1,2}, \ldots, Z_{1,b}$ are i.i.d. and $Z_{2,1}, Z_{2,2}, \ldots, Z_{2,b}$ are i.i.d.). In that case, it turns out that (see §5.2)

$$\text{Var}(\bar{D}_b) < \frac{1}{b}[\text{Var}(Z_{1,j}) + \text{Var}(Z_{2,j})].$$

One can also show that if $Z_{1,j}$ and $Z_{2,j}$ had been simulated *independently*, then we would have had an *equality* in the above expression. Thus, our trick may result in relatively small S_D^2 and, hence, small CI length. So how do we evoke the trick?

5.2 Common Random Numbers

The idea behind §5.1's trick is to use *common random numbers*, i.e., use the same pseudo-random numbers in exactly the same ways for corresponding runs of each of the competing systems. For instance, we could use the same customer arrival times when simulating different proposed configurations of a job shop. By subjecting the alternative systems to identical experimental conditions, we hope to make it easy to distinguish which systems are best even though the respective estimators are subject to sampling error.

Consider the case in which we compare two queueing systems, A and B, on the basis of their expected customer transit times, θ_A and θ_B — the smaller θ-value corresponds to the better system. Suppose we have at our disposal estimators $\hat{\theta}_A$ and $\hat{\theta}_B$ for θ_A and θ_B, respectively. We will declare A as the better system if $\hat{\theta}_A < \hat{\theta}_B$. If $\hat{\theta}_A$ and $\hat{\theta}_B$ are simulated independently, then the variance of their difference,

$$\mathsf{Var}(\hat{\theta}_A - \hat{\theta}_B) \;=\; \mathsf{Var}(\hat{\theta}_A) + \mathsf{Var}(\hat{\theta}_B),$$

could be very large; in this case, our declaration might lack conviction. If we could reduce $\mathsf{Var}(\hat{\theta}_A - \hat{\theta}_B)$, then we could be much more confident about our declaration. By using common random numbers, we can sometimes induce a high positive correlation between the point estimators $\hat{\theta}_A$ and $\hat{\theta}_B$. Then we have

$$\begin{aligned}\mathsf{Var}(\hat{\theta}_A - \hat{\theta}_B) &= \mathsf{Var}(\hat{\theta}_A) + \mathsf{Var}(\hat{\theta}_B) - 2\mathsf{Cov}(\hat{\theta}_A, \hat{\theta}_B) \\ &< \mathsf{Var}(\hat{\theta}_A) + \mathsf{Var}(\hat{\theta}_B),\end{aligned}$$

and we obtain a savings in variance.

5.3 Antithetic Random Numbers

Alternatively, if we can induce *negative* correlation between two unbiased estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, for some parameter θ, then the unbiased estimator $(\hat{\theta}_1 + \hat{\theta}_2)/2$ might have low variance.

Most simulation texts (see §6) give advice on how to run the simulations of the competing systems so as to induce positive or negative correlation between them. The consensus is that, if conducted properly, common random numbers and antithetic random numbers can lead to tremendous variance reductions.

5.4 Ranking, Selection, and Multiple Comparisons

Ranking, selection, and multiple comparisons methods form another class of statistical techniques used to compare alternative systems. Here, the experimenter is interested in selecting the best of a number (≥ 2) of competing processes. In addition, the experimenter might want to correctly select the best process with a certain high probability, especially if the best process is significantly better than its competitors. These methods are simple to use, fairly general, and intuitively appealing. There is a great deal of literature on the subject; the reader could start by looking at the discussions in Bechhofer, Santner, and Goldsman (1995) and Law and Kelton (2000) before graduating to the more mathematical references cited therein.

6 CONCLUSIONS

Proper analysis of output is one of the most important aspects of any simulation study. Since simulation output is never i.i.d. normal, the experimenter must be careful when making conclusions about such data. Indeed, the purpose of this tutorial has been to inform the experimenter about some of the issues and techniques relevant to conducting valid analyses.

There are many interesting sides of output analysis that we have not had space to discuss in this paper, e.g., multivariate parameter estimation, sequential methods, and other variance reduction techniques. Fortunately, a number of excellent general resources are available that devote substantial discussion to the subject, e.g., Banks (1998), Banks, Carson, and Nelson (1995), Bratley, Fox, and Schrage (1987), Fishman (1978), Law and Kelton (2000), and any *Proceedings* of the annual Winter Simulation Conference — for instance, Alexopoulos and Seila (2000) in this volume.

ACKNOWLEDGMENT

David Goldsman's work was supported by National Science Foundation Grant No. DMI-96-22269.

REFERENCES

Alexopoulos, C., and A. F. Seila. 2000. Output analysis for simulations. In *Proceedings of the 2000 Winter Simulation Conference* (ed. J. A. Joines, R. R. Barton, P. Fishwick, and K. Kang), Piscataway, New Jersey: The Institute of Electrical and Electronics Engineers.

Banks, J. 1998. *Handbook of simulation*. New York: John Wiley and Sons.

Banks, J., J. S. Carson, II, and B. L. Nelson. 1995. *Discrete-event system simulation*, 2d Ed. Englewood Cliffs, New Jersey: Prentice-Hall.

Bechhofer, R. E., T. J. Santner, and David Goldsman. 1995. *Design and analysis of experiments for statistical selection, screening and multiple comparisons.* New York: John Wiley and Sons.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A guide to simulation*, 2d Ed. New York: Springer-Verlag.

Chien, C., D. Goldsman, and B. Melamed. 1997. Large-sample results for batch means. *Management Science* 43:1288–1295.

Crane, M. A., and D. L. Iglehart. 1975. Simulating stable stochastic systems: III. Regenerative processes and discrete-event simulations. *Operations Research* 23:33–45.

Crane, M. A., and A. J. Lemoine. 1977. *An Introduction to the Regenerative Method for Simulation Analysis.* Berlin: Springer-Verlag.

Fishman, G. S. 1978. *Principles of discrete event simulation.* New York: John Wiley and Sons.

Goldsman, D. 1992. Simulation output analysis. In *Proceedings of the 1992 Winter Simulation Conference* (ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson), 97–103, Piscataway, New Jersey: The Institute of Electrical and Electronics Engineers.

Goldsman, D., and L. W. Schruben. 1984. Asymptotic properties of some confidence interval estimators for simulation output. *Management Science* 30:1217–1225.

Goldsman, D., L. W. Schruben, and J. J. Swain. 1994. Tests for transient means in simulated time series. *Naval Research Logistics* 41:171–187.

Goldsman, D., and G. Tokol. 1997. Output analysis procedures for computer simulation. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics — Computational Cybernetics and Simulation*, 1802–1807, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Heidelberger, P., and P. D. Welch. 1981. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* 24:233–245.

Heidelberger, P., and P. D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* 31:1109–1144.

Kelton, W. D., and A. M. Law. 1983. A new approach for dealing with the startup problem in discrete event simulation. *Naval Research Logistics Quarterly* 30:641–658.

Law, A. M., and W. D. Kelton. 1984. Confidence intervals for steady-state simulations, I: A survey of fixed sample size procedures. *Operations Research* 32:1221–1239.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling and analysis*, 3d Ed. New York: McGraw-Hill.

Meketon, M. S., and B. W. Schmeiser. 1984. Overlapping batch means: Something for nothing? In *Proceedings of the 1984 Winter Simulation Conference* (ed. S. Sheppard, U. W. Pooch, and C. D. Pegden), 227–230, Piscataway, New Jersey: The Institute of Electrical and Electronics Engineers.

Sargent, R. G., K. Kang, and D. Goldsman. 1992. An investigation of finite-sample behavior of confidence interval estimators. *Operations Research* 40:898–913.

Schmeiser, B. W. 1982. Batch size effects in the analysis of simulation output. *Operations Research* 30:556–568.

Schruben, L. W. 1982. Detecting initialization bias in simulation output. *Operations Research* 30:569–590.

Schruben, L. W. 1983. Confidence interval estimation using standardized time series. *Operations Research* 31:1090–1108.

Schruben, L. W., H. Singh, and L. Tierney. 1983. Optimal tests for initialization bias in simulation output. *Operations Research* 31:1167–1178.

Welch, P. D. 1981. On the problem of the initial transient in steady state simulations. Technical Report, IBM Watson Research Center, Yorktown Heights, New York.

Welch, P. D. 1983. The statistical analysis of simulation results. In *The Computer Performance Modeling Handbook* (ed. S. Lavenberg), Orlando, Florida: Academic Press.

Wilson, J. R., and A. A. B. Pritsker. 1978a. A survey of research on the simulation startup problem. *Simulation* 31:55–58.

Wilson, J. R., and A. A. B. Pritsker. 1978b. Evaluation of startup policies in simulation experiments. *Simulation* 31:79–89.

AUTHOR BIOGRAPHIES

DAVID GOLDSMAN is a Professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include simulation output analysis and ranking and selection. He was the Program Chair for the 1995 Winter Simulation Conference. His e-mail and web addresses are `<sman@isye.gatech.edu>` and `<www.isye.gatech.edu/~sman/>`.

GAMZE TOKOL is a software consultant in Atlanta, GA, specializing in simulation and logistics modeling and software reliability. She received her Ph.D. in operational research from the Middle East Technical University in Ankara, Turkey. Her research interests lie in simulation output analysis and applied operations research. Her e-mail address is `<gamze@mindspring.com>`.