

STATISTICAL ANALYSIS OF SIMULATION OUTPUT DATA: THE PRACTICAL STATE OF THE ART

Averill M. Law

Averill M. Law & Associates, Inc.
4729 East Sunrise Drive, #462
Tucson, AZ 85718, USA

ABSTRACT

One of the most important but neglected aspects of a simulation study is the proper design and analysis of simulation experiments. In this tutorial we give a state-of-the-art presentation of what the practitioner really needs to know to be successful. We will discuss how to choose the simulation run length, the warmup-period duration (if any), and the required number of model replications (each using different random numbers). The talk concludes with a discussion of three critical pitfalls in simulation output-data analysis.

1 INTRODUCTION

In many “simulation studies” a great amount of time and money is spent on model development and “programming,” but little effort is made to analyze the simulation output data appropriately. As a matter of fact, a very common mode of operation is to make a single simulation run of somewhat arbitrary length and then to treat the resulting simulation estimates as the “true” model characteristics. Since random samples from probability distributions are typically used to drive a simulation model through time, these estimates are just particular realizations of random variables that may have large variances. As a result, these estimates could, in a particular simulation run, differ greatly from the corresponding true characteristics for the model. The net effect is, of course, that there could be a significant probability of making erroneous inferences about the system under study.

We now describe more precisely the random nature of simulation output. Let Y_1, Y_2, \dots be an output stochastic process (see, for example, section 4.3 in Law 2015) from a *single* simulation run. For example, Y_i might be the delay in queue for the i th job to arrive at a single-server queueing system. Alternatively, Y_i might be the total cost of operating an inventory system in the i th month. The Y_i 's are random variables that will not, in general, be independent or identically distributed (IID). Thus, many of the formulas from classical statistics (see Section 2) will not be *directly* applicable to the analysis of simulation output data.

Example 1. For the queueing system mentioned above, the delays in queue will not be independent, since a large delay for one customer waiting in queue will tend to be followed by a large delay for the next customer waiting in queue. Suppose that the simulation is started at time zero with no customers in the system, as is usually the case. Then the delays in queue at the beginning of the simulation will tend to be smaller than later delays and, thus, the delays are not identically distributed.

Let $y_{11}, y_{12}, \dots, y_{1m}$ be a realization of the random variables Y_1, Y_2, \dots, Y_m resulting from running the simulation with a particular set of random numbers u_{11}, u_{12}, \dots . If we run the simulation with a different set of random numbers u_{21}, u_{22}, \dots , then we will obtain a different realization $y_{21}, y_{22}, \dots, y_{2m}$ of the random variables Y_1, Y_2, \dots, Y_m . (The two realizations are not the same since the different random numbers used in the

two runs produce different samples from the input probability distributions.) In general, suppose that we make n independent replications (runs) of the simulation (i.e., different random numbers are used for each replication, each replication uses the same initial conditions, and the statistical counters for the simulation are reset at the beginning of each replication) each of length m , resulting in the observations:

$$\begin{array}{c} y_{11}, \dots, y_{1i}, \dots, y_{1m} \\ y_{21}, \dots, y_{2i}, \dots, y_{2m} \\ \vdots \quad \quad \quad \vdots \\ y_{n1}, \dots, y_{ni}, \dots, y_{nm} \end{array}$$

The observations from a particular replication (row) are clearly not IID. However, note that $y_{1i}, y_{2i}, \dots, y_{ni}$ (from the i th column) are IID observations of the random variable Y_i , for $i = 1, 2, \dots, m$. More generally, each entire replication is independent of any other replication, and each replication's observations have the same (joint) distribution. This *independence across runs* is the key to relatively simple output-data analysis that is discussed in later sections of this paper. Then, roughly speaking, the goal of output-data analysis is to use the observations y_{ji} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$) to draw inferences about characteristics of the random variables Y_1, Y_2, \dots, Y_m .

Example 2. Consider a bank with five tellers and one queue, which opens its doors at 9 A.M., closes its doors at 5 P.M., but stays open until all customers in the bank at 5 P.M. have been served. Assume that customers arrive with IID exponential inter-arrival times with mean 1 minute, that service times are IID exponential random variables with mean 4.5 minutes, and that customers are served in a first-in, first-out (FIFO) manner. We made $n = 50$ independent replications of the bank assuming that no customers are present initially and Table 1 shows two typical output statistics from the first 5 replications. Note that results from different replications can be quite different. For example, the average delay in queue was 10.37 minutes on the first replication but only 2.17 minutes on the second. Thus, one run clearly does not produce the “answers.” See Example 10 for further analysis of this bank model.

Table 1: Results for 5 Independent Replications of the Bank Model.

Replication	Number of customers served	Average delay in queue (in minutes)
1	494	10.37
2	464	2.17
3	479	3.83
4	466	6.01
5	497	4.52

Our goal in this paper is to discuss methods for statistical analysis of simulation output data and to present the material with a practical focus. Section 2 of this paper reviews formulas from classical statistics based on IID data, which we will find useful later in this paper. In Section 3, we discuss the two main types of simulations with regard to output-data analysis, namely, terminating and non-terminating. Statistical methods for analyzing each type are given in Sections 4 and 5, respectively. Finally, we give a summary of this tutorial and three fundamental pitfalls in output-data analysis in Section 6.

Portions of this paper are based on chapters 4 and 9 of Law (2015). Other references on output-data analysis are Alexopoulos and Kelton (2017), Banks et al. (2010), and Nakayama (2008).

2 REVIEW OF CLASSICAL STATISTICS

Suppose that X_1, X_2, \dots, X_n are IID random variables with population mean and variance μ and σ^2 , respectively. Then unbiased point estimators for μ and σ^2 are given by

$$\bar{X}(n) = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

and

$$S^2(n) = \frac{\sum_{i=1}^n [X_i - \bar{X}(n)]^2}{n-1} \quad (2)$$

(An estimator is *unbiased* if its expected value is equal to the target population characteristic.) Furthermore, an approximate $100(1-\alpha)$ percent ($0 < \alpha < 1$) confidence interval for μ is given by

$$\bar{X}(n) \pm t_{n-1, 1-\alpha/2} \sqrt{S^2(n)/n} \quad (3)$$

where $t_{n-1, 1-\alpha/2}$ is the upper $1-\alpha/2$ critical point for a t distribution with $n-1$ degrees of freedom. If the sample size n is “sufficiently large,” then the confidence interval given by Expression (3) will have a coverage probability arbitrarily close to $1-\alpha$. Alternatively, if the X_i 's are normally distributed, then the coverage probability will be exactly $1-\alpha$. In practice, if the distribution of the X_i 's is reasonably symmetric, then the coverage probability will be close to $1-\alpha$ (see pp. 233-237 in Law 2015). If we increase the sample size from n to $4n$, then the half-length of the confidence interval, $t_{n-1, 1-\alpha/2} \sqrt{S^2(n)/n}$, will decrease by a factor of approximately 2, since there is an n in the denominator under the square-root sign.

As stated above, the Y_i 's from one simulation run are not IID and, thus, Expressions (1), (2), and (3) are not *directly* applicable to their analysis. However, if we take comparable output statistics from different independent replications of a simulation model, then these observations *are* IID and the three expressions are applicable.

Example 3. For the bank simulation of Example 2, the five average delays in queue from column 3 of Table 1 are IID and, thus, Expressions (1), (2), and (3) could legitimately be used for their analysis.

3 TYPES OF SIMULATIONS WITH REGARD TO OUTPUT ANALYSIS

The options available for designing and analyzing simulation experiments depend on whether the simulation of interest is terminating or non-terminating, which depends on whether there is an obvious way for determining the simulation run length.

A terminating simulation is one for which there is a “natural” event E that specifies the length of each run (replication). Since different runs use independent random numbers and the same initialization rule, this implies that comparable random variables are IID. The event E might occur at a time point that has one of the following properties:

- The system is “cleaned out”

- Beyond which no useful information is obtained
- Specified by management.

The event E is specified before any runs are made, and the time of occurrence of E for a particular run may be a random variable. Since the initial conditions for a terminating simulation generally affect the desired measures of performance, these conditions should be representative of those for the actual system.

Example 4. A retail/commercial establishment (e.g., a bank) closes each evening. If the establishment is open from 9 to 5, the objective of a simulation might be to estimate some measure of the quality of customer service over the period beginning at 9 A.M. and ending when the last customer who entered before the doors closed at 5 P.M. has been served. In this case, $E = \{8 \text{ hours of simulated time have elapsed and the system is empty}\}$, and the initial conditions for the simulation should be representative of those for the bank at 9 A.M.

Example 5. Consider a military ground confrontation between a blue force and a red force. Relative to some initial force strengths, the goal of a simulation might be to determine the (final) force strengths when the battle ends. In this case, $E = \{\text{either the blue force or the red force has "won" the battle}\}$. An example of a condition that would end the battle is one side losing 30 percent of its force, since this side would no longer be considered viable. The choice of initial conditions for the simulation, e.g., the number of troops and tanks for each force, is generally not a problem here, since they are specified by the military scenario under consideration.

Example 6. A company that sells a single product would like to decide how many items to have in inventory during a planning horizon of 12 months. Given some initial inventory level, the objective might be to determine how much to order each month so as to minimize the expected average cost per month of operating the inventory system. In this case $E = \{12 \text{ months have been simulated}\}$, and the simulation is initialized with the current inventory level.

A *non-terminating simulation* is one for which there is no natural event E to specify the length of a run. This often occurs when we are designing a new system or modifying an existing system, and we are interested in the behavior of the system in the long run when it is operating "normally." Unfortunately, "in the long run" doesn't naturally translate into a terminating event E .

Consider the output stochastic process Y_1, Y_2, \dots for a simulation model. Let $F_i(y|I) = P(Y_i \leq y|I)$ for $i = 1, 2, \dots$, where y is a real number and I represents the initial conditions used to start the simulation at time 0. [The conditional probability $P(Y_i \leq y|I)$ is the probability that the event $\{Y_i \leq y\}$ occurs given the initial conditions I .] For a manufacturing system, I might specify the number of jobs present, and whether each machine is busy or idle, at time 0. We call $F_i(y|I)$ the *transient distribution* of the output process at (discrete) time i for initial conditions I . Note that $F_i(y|I)$ will, in general, be different for each value of i and each set of initial conditions I . For fixed y and I , the probabilities $F_1(y|I), F_2(y|I), \dots$ are just a sequence of numbers. If $F_i(y|I) \rightarrow F(y)$ as $i \rightarrow \infty$ for all y and any initial conditions I , then $F(y)$ is called the *steady-state distribution* of the output process Y_1, Y_2, \dots . Note that the steady-state distribution $F(y)$ does *not* depend on the initial conditions I .

A measure of performance for a non-terminating simulation is said to be a *steady-state parameter* if it is a characteristic of the steady-state distribution of some output stochastic process Y_1, Y_2, \dots . If the random variable Y has the steady-state distribution, then we are typically interested in estimating the steady-state mean $\nu = E(Y)$.

Example 7. Consider a company that is going to build a new manufacturing system and would like to determine the long-run (steady-state) mean hourly throughput of their system after it has been running long enough for workers to know their jobs and for mechanical difficulties to have been worked out. The system will operate continuously 24 hours a day for 7 days a week. Let N_i be the number of parts manufactured in the i th hour. If the stochastic process N_1, N_2, \dots has a steady-state distribution with corresponding random variable N , then we are interested in estimating the steady-state mean $\nu = E(N)$.

Example 8. Suppose that a military organization is going to employ a new inventory system during (a long) *peacetime* and would like to determine the long-run mean monthly cost of operating their system. Let C_i be the cost of operating the inventory system in the i th month. If the output process C_1, C_2, \dots has a steady-state distribution with corresponding random variable C , then they might be interested in estimating the mean $\nu = E(C)$.

4 STATISTICAL ANALYSIS FOR TERMINATING SIMULATIONS

Suppose that we make n independent replications of a terminating simulation each terminated by the event E . Let X_j be an output random variable defined over the j th replication, for $j = 1, 2, \dots, n$; it is assumed that the X_j 's are comparable for different replications. Then the X_j 's are IID random variables. For the bank of Example 4, X_j might be the average delay $\sum_{i=1}^N D_i / N$ over a day from the j th replication, where N (a random variable) is the number of customers served in a day and D_i is the delay in queue of the i th arriving customer. (See columns 2 and 3 in Table 1.) For the combat model of Example 5, X_j might be the number of red tanks destroyed on the j th replication. For the inventory system of Example 6, X_j could be the average cost per month over the 12-month planning horizon.

Suppose that we would like to obtain a point estimate and confidence interval for the mean $\mu = E(X)$, where X is a random variable defined on a replication as described above. Make n independent replications of the simulation and let X_1, X_2, \dots, X_n be the resulting IID random variables. Then, by substituting the X_j 's into Expressions (1), (2), and (3), we get that $\bar{X}(n)$ is an unbiased point estimator for μ , and an approximate $100(1-\alpha)$ percent confidence interval for μ is given by

$$\bar{X}(n) \pm t_{n-1, 1-\alpha/2} \sqrt{S^2(n)/n}$$

Example 9. A small factory consists of a machine and an inspector, as shown in Figure 1. Unfinished parts arrive to the factory with exponential inter-arrival times having a mean of 1 minute. Processing times at the machine are uniformly distributed on the interval $[0.65, 0.70]$ minute, and subsequent inspection times at the inspector are uniformly distributed on the interval $[0.75, 0.80]$. (The assumption of uniformity is for ease of exposition, and is not likely to be valid in a real-world application.) Ninety percent of inspected parts are “good” and leave the system immediately; 10 percent of the parts are

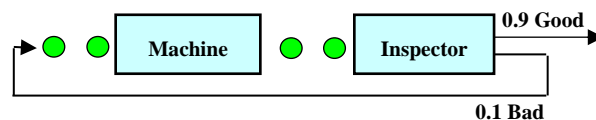


Figure 1: Small Factory.

“bad” and are sent back to the machine for rework. (Both queues are assumed to be of infinity capacity.) The machine is subject to randomly occurring breakdowns. In particular, a new (or freshly repaired) machine will break down after an exponential amount of *calendar* time with a mean of 6 hours. Repair times are uniform on the interval [8, 12] minutes. If a part is being processed when the machine breaks down, then the machine continues where it left off upon the completion of repair. Assume that the factory is initially empty and idle.

The factory gets an order to produce 2000 parts and, thus, a simulation of this system can be considered to be terminating with $E = \{2000 \text{ parts have been completed}\}$. Let T be the time required to complete the required 2000 parts. Then the company would like a point estimate and a 95 percent confidence interval for the mean $\mu = E(T)$.

We made 10 independent replications of the simulation and obtained the following observed values for T (in hours):

$$\begin{aligned} T_1 &= 32.62, T_2 = 32.57, T_3 = 33.51, T_4 = 33.29, \\ T_5 &= 32.10, T_6 = 34.24, T_7 = 32.70, T_8 = 33.49, \\ T_9 &= 33.36, T_{10} = 34.61 \end{aligned}$$

Substituting the T_j 's into Expressions (1), (2), and (3), gives the following results:

$$\bar{T}(10) = 33.25, S^2(10) = 0.606$$

and an (approximate) 95 percent confidence interval for $\mu = E(T)$ is given by

$$33.25 \pm 0.56 \quad \text{or} \quad [32.69, 33.81]$$

Thus, we are approximately 95 percent confident that μ is between 32.69 and 33.81 hours. (If 100 people performed this experiment independently, then we would expect that about 95 out of the 100 confidence intervals to contain the true μ .) Note also that the interval is quite precise, with the half-length of the confidence interval being less than 2 percent of the point estimate.

We have discussed above how to estimate the mean of the random variable T . However, in some cases we might also be interested in getting a measure of how large a value T can take on. For example, we may also want to estimate the 95th percentile, $x_{0.95}$, of the distribution of the random variable T , i.e., T is larger than $x_{0.95}$ only 5 percent of the time.

Example 10. For the bank of Example 2, we made $n = 50$ independent replications of length “one day.” Let X be the average delay over a day, which is defined as follows (see the notation in the first paragraph of Section 4):

$$X = \frac{\sum_{i=1}^N D_i}{N}$$

Then we would like a point estimate and 95 percent confidence interval for the mean $\mu = E(X)$. From the 50 X_j 's produced in Example 2, we get

$$\bar{X}(50) = 5.92, S^2(50) = 9.95$$

and an (approximate) 95 percent confidence interval for $\mu = E(X)$ is given by

$$5.92 \pm 0.90 \quad \text{or} \quad [5.02, 6.81]$$

Thus, subject to the correct interpretation, we are approximately 95 percent confident that μ is between 5.02 and 6.81 minutes. However, the interval is not very precise, with the half-length of the confidence interval being approximately 15 percent of the point estimate. If we want to reduce the half-length from 0.9 to, say, 0.3, then a *total* of approximately 450 ($= 9 \times 50$) replications will be required.

5 STATISTICAL ANALYSIS FOR NONTERMINATING SIMULATIONS

Let Y_1, Y_2, \dots be an output stochastic process from a single run of a non-terminating simulation. Suppose that we want to estimate the steady-state mean $\nu = E(Y)$, which is also defined by

$$\nu = \lim_{i \rightarrow \infty} E(Y_i)$$

where $E(Y_i)$ is the *transient mean* at time i . Thus, the transient means converge to the steady-state mean. However, $E(Y_i) \neq \nu$ for “small” i because we generally don’t know how to choose the initial conditions I to be representative of “steady-state behavior.” This causes the sample mean $\bar{Y}(m)$ to be a biased estimator of ν for all finite values of m . The problem that we have just described is called the *problem of the initial transient* in the simulation literature.

The technique most often suggested for dealing with this problem is called *warming up the model*. The idea is to delete some number of observations from the beginning of a run and to use only the remaining observations to estimate ν . In particular, given the observations Y_1, Y_2, \dots, Y_m , we would use

$$\bar{Y}(m, l) = \frac{\sum_{i=l+1}^m Y_i}{m-l}$$

($1 \leq l \leq m-1$) rather than $\bar{Y}(m)$ as an estimator of ν . In general, one would expect $\bar{Y}(m, l)$ to be less biased than $\bar{Y}(m)$, since the observations near the “beginning” of the simulation may not be very representative of steady-state behavior due to the choice of initial conditions.

The question naturally arises as to how to choose the *warmup period* (or deletion amount) l . We would like to pick l (and m) such that $E[\bar{Y}(m, l)] \approx \nu$. If l and m are chosen too small, then $E[\bar{Y}(m, l)]$ may be significantly different than ν . On the other hand, if l is chosen larger than necessary, then $\bar{Y}(m, l)$ will probably have an unnecessarily large variance.

The simplest and most general technique for determining l is a graphical technique due to Welch (1983) (see also pp. 513-520 in Law 2015). Its specific goal is to determine l such that $E(Y_i) \approx \nu$ for $i > l$, where l is the warmup period. This is equivalent to determining when the transient-mean curve $E(Y_i)$ “flattens out” at level ν . In general, it is difficult to determine l from a single replication due to the inherent variability of the process Y_1, Y_2, \dots . As a result, Welch’s procedure is based on making multiple replications of the simulation in a pilot study.

5.1 The Replication/Deletion Approach

In this section, we discuss how to construct a point estimate and confidence interval for ν . Suppose that the warmup period has been determined by Welch’s procedure or by using “engineering judgment.” Make n independent replications of the output process Y_1, Y_2, \dots each of length m , where m should be much larger than l . (There is no definitive way of picking the run length m here, as there was for terminating simulations.) Let Y_{ji} be the i th observation from the j th replication, for $j = 1, 2, \dots, n$ and $i = 1, 2, \dots, m$. Let

$$X_j = \sum_{i=l+1}^m Y_{ji} / (m-l) \quad \text{for } j = 1, 2, \dots, n$$

Note that $i = l + 1$ is where we think that “steady state” begins. Then the X_j ’s are IID random variables. Furthermore, $E(X_j) \approx \nu$ since $Y_{j,l+1}, Y_{j,l+2}, \dots, Y_{j,m}$ each have approximate mean ν . Then, by substituting the X_j ’s into Expressions (1), (2), and (3), we get that $\bar{X}(n)$ is an (approximately) unbiased point estimator for ν , and an approximate $100(1 - \alpha)$ percent confidence interval for ν is given by

$$\bar{X}(n) \pm t_{n-1, 1-\alpha/2} \sqrt{S^2(n) / n}$$

We call the above method for constructing a point estimate and confidence interval for ν the *replication/deletion method*. One criticism that has been levied against this method historically is that l observations must be discarded from each of the n replications. However, given the availability of fast PCs and cloud computing, this is no longer an issue for most steady-state analyses.

Example 11. Consider a manufacturing system with a receiving/shipping station and five workstations (see Figure 2), as described in chapter 14 of Law (2015). Assume that there are 4, 2, 5, 3, and 2 machines in stations 1 through 5, respectively. The machines in a particular station are identical, but machines in different stations are dissimilar. Jobs arrive to the system with exponential inter-arrival times with a mean of (1/15)th of an hour. Thus, 15 jobs arrive in a typical hour. There are three types of jobs, and jobs are of types 1, 2, and 3, with respective probabilities 0.3, 0.5, and 0.2. Job types 1, 2, and 3 require 4, 3, and 5 operations to be done, respectively, and each operation must be done at a specified workstation in a prescribed order. Each job begins at the receiving/shipping station, travels to the workstations on its routing, and then leaves the system at the receiving/shipping station. For example, the routing for a type 1 job is 3, 1, 2, 5.

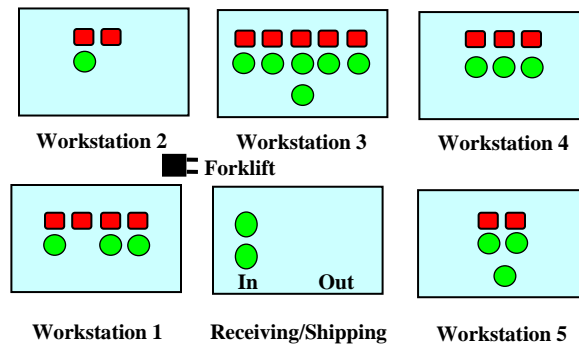


Figure 2: Factory with five workstations.

A job must be moved from one station to another by a forklift truck, which moves at a speed of 5 feet per second, and two forklifts are available. When a forklift becomes available, it processes requests by jobs using a shortest-distance-first dispatching rule. If more than one forklift is idle when a job requests transport, then the closest forklift is used. When a forklift finishes moving a job to a workstation, it remains at that station if there are no pending job requests. Each station has a single FIFO queue. The time to perform an operation at a particular machine is a gamma random variable with a shape parameter of 2, whose mean depends on the job type and the station to which the machine belongs. For example, the mean service time for a type 1 job at station 3 (the first station on its routing) is 0.25 hour. When a machine finishes processing a job, the job blocks that machine (i.e., the machine cannot process another job) until the job is removed by a forklift.

The factory is open 8 hours a day, and thus the arrival rate is 120 jobs per day. The system configuration described here is called system design 3 in Law (2015).

Let N_1, N_2, \dots be the output stochastic process corresponding to daily throughputs. Then we are interested in obtaining a point estimate and 90 percent confidence interval for the *steady-state* mean daily throughput $\nu = E(N)$. Since the simulation starts out with no jobs present at time zero, the throughput will tend to be “small” during the early part of a run and a warmup period is needed. Using Welch’s graphical procedure, we determined that a reasonable warmup period for this output process is $l = 15$ days (see chapter 14 in Law 2015). We made $n = 10$ (production) replications of length $m = 115$ days, and used a warmup period of $l = 15$ days. Let

$$X_j = \frac{\sum_{i=16}^{115} N_{ji}}{100}$$

where N_{ji} is the throughput in the i th day of the j th replication.

Substituting the X_j 's into Expressions (1), (2), and (3), we get the following point estimate and approximate 90 percent confidence interval for $\nu = E(N)$:

$$\hat{\nu} = \bar{X}(10) = 120.29$$

and

$$120.29 \pm 0.63 \quad \text{or} \quad [119.66, 120.92]$$

Thus, subject to the correct interpretation, we are approximately 90 percent confident that the steady-state mean daily throughput is between 119.66 and 120.92 jobs per day. Note that this confidence interval contains 120, which should be the mean daily throughput if the system has enough machines and forklifts because the arrival rate is 120 jobs per day. (In a real application, ν would not, of course, be known.)

Note also that the confidence interval is quite precise, with the half-length being less than 1 percent of the point estimate. Also, since X_j is the average of 100 N_{ji} 's, it should be approximately normally distributed by a central-limit-theorem type effect. This suggests that the coverage of the confidence interval should be close to the desired coverage probability of 0.9.

5.2 An Ill-Advised Approach to Steady-State Analysis

Some people have attempted to construct a confidence interval for a steady-state mean by applying Expressions (1), (2), and (3) directly to the output data from a *single* simulation run, which will almost always be positively correlated. To illustrate the *general problem* with this approach, consider an $M/M/1$ queue with an arrival rate of $\lambda = 1$ per minute and a service rate of $\omega = 10/9$ per minute, so that the utilization factor is $\rho = \lambda / \omega = 0.9$. Suppose that we want to construct a 95 percent confidence interval for the steady-state mean *total* time in system (in queue plus in service) w , which is given by $w = 1 / (\omega - \lambda) = 9$ minutes (see Ross 2019). We made one simulation run of length 15,000 minutes and used a warmup period of length 5000 minutes, which resulted in 10,000 minutes of data actually being used to construct the confidence interval for w . We then checked whether the resulting confidence interval based on Expressions (1), (2), and (3) contained $w = 9$, and we repeated this whole experiment 500 times each with different random numbers. To our amazement, only 36 of the resulting 500 confidence intervals did, in fact, contain 9, so that the estimated coverage probability was a shockingly low 7.2 percent as compared to the desired 95 percent! Of course, we would not know the true value of the steady-state mean for a real-world simulation model, but the problem of overstating the confidence level still remains. Unfortunately, this confidence-interval approach is a choice in one or more commercial simulation products.

6 SUMMARY AND PITFALLS IN OUTPUT-DATA ANALYSIS

We have seen that both terminating and non-terminating analyses can be performed easily by making independent replications of the simulation model and by using Expressions (1), (2), and (3), which come from a first undergraduate course in statistics. In the case of steady-state parameters, we also have to determine a warmup period, but this can be reliably addressed using Welch's graphical approach. The method of replication can also be easily applied to comparing alternative system configurations (see, for example, chapters 10 and 11 in Law 2015) and to estimating multiple measures of performance. Moreover, multiple replications can be made simultaneously on computers having multiple cores or by using a cloud-computing service.

The following are three major pitfalls in output-data analysis:

- Analyzing simulation output data from one run using formulas that assume independence, which might result in a gross underestimation of variances and overly-optimistic confidence intervals.
- Failure to have a warmup period for steady-state analyses
- Failure to determine the statistical precision of simulation output statistics by the use of a confidence interval, which can be accomplished easily using the replication approach.

REFERENCES

- Alexopoulos, C., and W. D. Kelton 2017. "A Concise History of Simulation Output Analysis". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 115-130. Piscataway, New Jersey: Institute of Electrical and Electronic Engineers.
- Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2010. *Discrete-Event System Simulation*, 5th ed. Upper Saddle River, New Jersey: Prentice-Hall.
- Law, A. M. 2015. *Simulation Modeling and Analysis*, 5th ed. New York: McGraw-Hill.
- Nakayama, M. K. 2008. "Statistical Analysis of Simulation Output". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 62-72. Piscataway, New Jersey: Institute of Electrical and Electronic Engineers.
- Ross, S. M. 2019. *Introduction to Probability Models*, 12th ed. San Diego: Academic Press.

Welch, P. D. 1983. "The Statistical Analysis of Simulation Results". In *The Computer Performance Modeling Handbook*, edited by S. S. Lavenberg. New York: Academic Press.

AUTHOR BIOGRAPHY

AVERILL M. LAW is President of Averill M. Law & Associates, Inc., a company specializing in simulation seminars, simulation consulting, and software. He has presented more than 580 simulation and statistics short courses in 20 countries, including onsite seminars for AT&T, Australian Department of Defence, Boeing, Caterpillar, Coca-Cola, GE, GM, IBM, Intel, Lockheed Martin, Los Alamos National Lab, NASA, NATO (Netherlands), NSA, Sasol Technology (South Africa), 3M, UPS, U.S. Air Force, U.S. Army, U.S. Navy, and Verizon. Dr. Law has been a simulation consultant to more than 50 organizations including Booz Allen & Hamilton, ConocoPhillips, Defense Modeling and Simulation Office, Kimberly-Clark, M&M/Mars, Oak Ridge National Lab, U.S. Air Force, U.S. Army, U.S. Marine Corps, and U.S. Navy. He has written or coauthored numerous papers and books on simulation, operations research, statistics, manufacturing, and communications networks, including the book *Simulation Modeling and Analysis* that has been cited more than 20,800 times and is widely considered to be the "bible" of simulation. He developed the ExpertFit® distribution-fitting software and also several videotapes on simulation modeling. He was awarded the INFORMS Simulation Society Lifetime Professional Achievement Award in 2009. Dr. Law wrote a regular column on simulation for *Industrial Engineering* magazine. He has been a tenured faculty member at the University of Wisconsin-Madison and the University of Arizona. He has a Ph.D. in industrial engineering and operations research from the University of California at Berkeley. His e-mail address is <averill@simulation.ws> and his website is <www.averill-law.com>.