Proceedings of the 1983
Winter Simulation Conference
S. Roberts, J. Banks, B. Schmeiser (eds.)

SIMULATION ANALYSIS

W. David Kelton
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, Michigan 48109

The use of a computer simulation model to learn about the system(s) under study must involve an analysis of the results from the simulation program itself. A classification of simulation types is given which provides a framework for a treatment of simulation analysis. A more detailed discussion of the most difficult class of simulation analysis is presented. Various goals of analyses are mentioned, together with a brief discussion of related topics.

1. INTRODUCTION

The study of a system by means of computer simulation involves a substantial effort in modeling, validation, coding, and debugging. Once these difficult tasks have been accomplished, the simulation program will presumably be used to study the behavior of the corresponding system, try out alternative systems specifications and designs, and aid in making recommendations and decisions. This paper focuses on such use of simulation models, and in particular how the model's output should be analyzed to enable the drawing of valid, accurate, and precise conclusions.

Section 2 gives one possible way of classifying computer simulations which is useful for discussing their analysis. In Section 3 the most difficult kind of simulation is discussed in more detail. Sections 4 and 5 briefly mention additional topics in simulation analysis, and a few conclusions are drawn in Section 6.

2. A CLASSIFICATION OF COMPUTER SIMULATIONS AND ANALYSIS TYPES

The structure of mathematical models of systems can be classified along several alternative dimensions. For models to be studied by simulation, it is useful to use a two-dimensional classification to delineate the different kinds of analyses which are appropriate. Along one of the dimensions, a model may be classified as either deterministic or stochastic. Along the other dimension, a model may be either static or dynamic.

2.1 Deterministic

A deterministic mathematical model assumes that there are no random or uncontrollable elements which enter in a meaningful way. In such a model, the exogenous inputs are assumed to be exact in the sense that there is no uncertainty associated with their values. Thus, a simulation evaluation under a fixed set of structural assumptions and parameter values will produce an exact, deterministic set of output responses and system performance measures, subject to numerical accuracy. In this case, the analysis is quite simple, at least conceptually: A single run of the simulation model produces the exact (again, up to roundoff) set of values desired. Further runs of the same model will, of course, produce the same results. There may well be other difficulties, however, since the model may still be quite expensive to run, and the analyst must choose which alternative model specifications should be evaluated, given a budget constraint.

Static. In a static model, there is no time element involved; there is simply no concept of the passage of time in the system being modeled. Some examples of deterministic, static models might include:

> Calculating the value of a dependent variable from a fitted regression for a particular set of independent variables which have not been observed
>
> Calculating the values of key financial variables under alternative conditions arising from alternative decisions at a specified time

Note that in each example, we can calculate (exactly) the quantities desired, since the inputs are deterministic.

Dynamic. Here, the flow of time is an important part of the model. The system being modeled

CH1953-9/83/0000-0159 $01.00 © 1983 IEEE

evidently changes over time, and we want to study and quantify this evolution. Some examples of deterministic, dynamic models are:

> Calculating growth and decay rates of biological populations interacting with each other over time as functions of their initial sizes, natural birth and death rates, and predator-prey rules, by means of numerical evaluation of differential or finite-difference equations chosen to model their dynamic evolution
>
> Calculating values pertaining to product demand or production levels into the future by simple use of a deterministic forecasting rule

Since the results from dynamic models may take the form of a discrete or continuous time path, some summary measures may be helpful, such as the final or average population size, or the average rate of increase of demand, over the time period of interest. Here again, we are able to calculate, exactly, the quantities desired.

2.2 Stochastic

As opposed to deterministic models, at least some of the inputs driving a stochastic model are random quantities whose exact values on an individual evaluation of the model are not known in advance. This appears to be far less desirable than the deterministic case, and from the standpoint of the analysis problem this is true. However, many systems are inherently stochastic and it is thus necessary to model them stochastically to obtain a level of validity sufficient to obtain the information desired.

Given the need, then, to use a stochastic simulation model, we generate observations on random variables (r.v.'s) from appropriate distributions to drive the model, using a random number generator and appropriate transformation techniques to obtain the desired distributions; see, for example, chapters 8 and 9 of Fishman (1978b), or chapters 6 and 7 of Law and Kelton (1982b). While we must know exactly the distributions from which to generate, we will not know the particular values of the observations that will be generated on a given execution of the model. Thus, unlike the deterministic case, two runs of the simulation under identical structural assumptions and parameter settings will not, in general, produce the same output values if we use independent streams of basic random numbers. It is very important to be aware of this, for it implies that simply running a stochastic simulation model once and noting the values of some point esimators is not an acceptable analysis technique. The output values from a stochastic simulation are properly thought of as r.v.'s themselves, which have their own probability distribution, mean, variance, etc. The purpose of a stochastic simulation is to learn something about these output distributions, such as estimation of means, forming confidence intervals (c.i.'s), performing hypothesis tests, or estimating distribution quantiles. Unfortunately, the output distributions, being a result of the simulation itself, will certainly not be known; if we could derive these distributions mathematically, there would have been no need to simulate in the first place. The use of the simulation output to infer something about the output distributions proceeds by means of statistical analysis of that output.

As with deterministic simulations, we distinguish between static and dynamic models.

Static. A static stochastic model has no time element; these models are sometimes referred to now as "Monte Carlo" models, although this use of the term is not completely standard. Many uses of these models are for purely mathematical or probabilistic problems which may not have any inherent random structure at all, but for which a paper-and-pencil analysis is intractable or numerically unstable. Some examples are:

> Estimate an analytically intractable integral by writing it as the expectation of some r.v. which is then repeatedly sampled and averaged (see Law and Kelton 1982b, pp. 49-50)
>
> Estimate the null distribution or critical points of the test statistic for a proposed hypothesis test by generating many independent data sets under the null hypothesis and calculating the value of the test statistic for each set

Note that instead of "calculating" quantities as in the deterministic case, we can only "estimate" quantities from a stochastic simulation, due to the random nature of the output. For stochastic, static simulations, the general analysis technique is relatively simple. The model is repeatedly executed under the same structural assumptions and parameter values, but using an independent stream of input random numbers each time; this we call replication. The results constitute independent and identically distributed (i.i.d.) random variables to which the techniques of standard "classical" statistical analysis may be applied; we think of each replication as producing one data value on an output variable. The only open question is that of determining the number of replications, and must be decided on the basis of desired accuracy and cost. In general, it is advisable to make as many replications as possible to obtain the greatest accuracy in the output estimates, and to justify the use of any normal-theory techniques (such as using the t distribution to form a c.i. by invoking the central limit theorem.

Some general references on Monte Carlo simulation are Hammersley and Handscomb (1964) and Rubinstein (1981). Of the many good books available as references for statistical analysis and experimental design, Box, Hunter, and Hunter (1978) provides a recent source. For more on response surface methodology and optimization, see Myers (1971), or Biles and Swain (1980).

Dynamic. A stochastic, dynamic model is a random model which evolves through time in a way which is not completely predictable. This is the kind of simulation to which many people now refer as "simulation." A few examples are:

Estimate the expected time a customer is delayed in queue and the expected number of customers in system in a fast-food restaurant

Estimate the probability that a customer's demand from an inventory cannot be immediately satisfied

The proper analysis of stochastic, dynamic simulations is far more complicated than for the above three cases, and will occupy the remainder of the paper.

A caveat that should be immediately issued in this case is that the mere existence or application of some statistical technique need not imply that it is appropriate or that it is valid. Indeed, inappropriate statistical methods will always produce some numerical results, but these numbers, if taken literally, can lead to serious misinterpretation and thus to the drawing of invalid conclusions. An important example of this is the calculation of a "sample variance" statistic, s^2, from successive customers' waits in a queueing simulation (which is an available option in some simulation languages' output reports); such a statistic can be heavily biased due to the presence of autocorrelation between the individual customers' waits (see, for example, Anderson 1971, p. 448). If the waits are positively correlated, which is usually the case, then s^2 will be biased low, i.e., $E(s^2)$ is less than the actual variance of a customer's wait. The result is that s^2 gives a deceptively small estimate of the uncertainty associated with the simulation output, leading in turn to a tendency to place more faith in the results than is justified. Thus, an improper analysis of a simulation may, in some sense, be worse than no analysis at all.

Over the past several years, the subject of the statistical analysis of the output from stochastic, dynamic simulations has received considerable attention in the operations research literature. In the remainder of this paper, an overview of some of the main problem types and procedures is given; space limitations prohibit a detailed, immediately usable treatment of the methodologies. For a more thorough treatment, the reader is referred to the books by Fishman (1978b) and Law and Kelton (1982b), as well as to the comprehensive, detailed surveys by Law (1983) and Welch (1983). These works also contain a large number of references to the original papers.

3. ANALYSIS OF STOCHASTIC, DYNAMIC COMPUTER SIMULATIONS

In this section, let us assume that we have a single simulation model of interest; Section 4 discusses the analysis of more than one system. Also, we focus on c.i.'s rather than hypothesis tests. An appropriately defined c.i. gives all the information embodied in the result of a hypothesis test as well as a quantitative measure of the degree of departure from the null hypothesis in case it is rejected.

3.1 Types of Measures

There may be several different measures of performance of interest for the model, reflecting different kinds of properties. Many of the main kinds of measures may be classified as means, probabilities, utilizations, variances, or quantiles.

Means. A mean is the expectation of an output r.v. For example, if D_i is the delay (in minutes) in queue of the ith exiting customer, we may want to estimate the expectation of D_{100}, or of

$$\sum_{i=1}^{500} D_i/500,$$

the average of the first 500 customers' delays. In the latter case, the average is the r.v. whose expectation is desired. As a second example, suppose Q_t is the number of jobs waiting at time t hours in a queue for a printer in a computer facility. We might want to estimate the mean number of jobs in this queue at time 8, i.e., $E(Q_8)$, or the expected time-average number of jobs in the queue over a day, i.e.,

$$E\left(\int_0^{24} Q_t dt/24\right).$$

A mean is a traditional (but not the only) measure of central tendency of a r.v.

Probabilities. A probability output measure is simply the probability that some condition in the output will occur. Following the same two examples as above, we might be interested in the probability that D_{100} is between 2 and 6 minutes, i.e., $P(2 \le D_{100} \le 6)$, or in the probability that the average of the first 500 delays is no more than 4 minutes. Probabilities can be thought of as (and measured by) the mean of an appropriate indicator r.v. For example, define the r.v. I to be equal to 1 if $2 \le D_{100} \le 6$, and define it to be zero otherwise. Then $P(2 \le D_{100} \le 6) = E(I)$, and can be estimated as such.

Utilizations. A utilization is the expected proportion of time some facility is in a particular one of its two possible states, usually referred to as "busy." In the printer queue example, let

$$B_t = \begin{cases} 1 \text{ if the printer is busy at time } t \\ 0 \text{ otherwise} \end{cases}$$

Then

$$E\left(\int_0^{24} B_t dt/24\right)$$

is the expected proportion of time in a day the printer is busy, which we call its utilization.

Variances. A variance is a measure of the dispersion, or degree of instability, of an output r.v. While this may certainly be of

interest in itself, the estimation of variances plays a central role in performing statistical analysis on and inference from simulation output, especially concerning inference on means. The variance of an output r.v. X is also an expectation of another appropriately defined r.v.: $\mathrm{Var}(X) = E\{[X - E(X)]^2\}$.

Quantiles. Unlike the above measures, a quantile is not an expectation. For $0 < q < 1$, the q-quantile of X is a number x_q satisfying

$$P(X < x_q) \le q \text{ and } P(X \le x_q) \ge q.$$

(If X has a cumulative distribution function which is strictly increasing, then x_q is defined by the simpler relation $P(X \le x_q) = q$.) Note that the 0.50-quantile is the median, which is a measure of central tendency that is sometimes used instead of the mean, especially if the distribution of X is highly skewed. Quantiles can be very important measures in simulation. For example, if the output measure X is the amount of oil arriving to a tank during a day, we might want to know what tank capacity is needed to give us a 0.95 probability of being able to store all the oil that arrives; this is the 0.95-quantile of X. The estimation of quantiles from simulation output tends to be more difficult than the other measures (typically involving some kind of sorting operation), and will not be explicitly treated here; see, for example, Heidleberger and Lewis (1981), Iglehart (1976), Seila (1981, 1982), and Welch (1983).

3.2 Measuring Several Quantities Simultaneously

In most cases, we will want to get more out of a simulation study than an estimate of only a single parameter. For example, we might want to estimate the expected average delay in queue, the expected time-average length of the queue, and the utilization of a particular server. All three estimates could be readily obtained and (by methods described below), 95% (say) c.i.'s could be placed on their expectations. However, we are making three separate statements about unknown parameters, each of which is at confidence level 0.95. By a simple result known as the Bonferroni inequality, the overall confidence level for all three c.i.'s, taken simultaneously, may not be 0.95, but can only be said to be at least 0.85: $0.85 = 1 - 3(1 - 0.95)$. (This may not seem too severe, but consider the consequences of having, say, twenty instead of three output measures.) This unpleasant phenomenon can be rectified by making the initial three c.i.'s at level 0.9833... instead of 0.95, which would produce an overall confidence level of at least 0.95. The cost is that the individual c.i.'s will be wider (and thus less precise), or that more data will have to be collected to get 98.333% c.i.'s as small as the original 95% c.i.'s.

In the statistics literature, the above difficulty is called the "problem of multiple comparisons," especially in reference to its version for hypothesis tests. It is important for the analyst to be aware of its gloomy consequences and avoid making unwarranted statements about overall levels of confidence.

3.3 Terminating vs. Steady-State Simulations

A fundamental issue in simulation analysis concerns the time horizon with respect to which the desired output measures are defined. In a sense, this issue could be dismissed almost as one of semantics, and the distinction may not be too important in many applied contexts; by simply rephrasing the goals of the study and the definitions of the quantities to be estimated, a terminating simulation could be transformed into a steady-state simulation, or vice-versa. However, the distinction is crucial in terms of the type of statistical analysis that should be done, as well as in terms of the basic approach to running and experimenting with the simulation model.

Terminating simulations. Following Law (1980), a simulation is called terminating if the quantities to be estimated are defined relative to specific starting and stopping conditions. For example, a manufacturing system could be simulated in a terminating mode by specifying the number of machines up at time zero (or, more generally, the distribution from which this number is to be independently drawn), the times since the last repairs of each machine, the number of production items present initially, their initial stages of production, etc. The stopping conditions might be simply that a fixed amount of simulated time has elapsed, or that a specified number of finished items have been produced. The choice of starting and stopping conditions is really a part of the modeling process, and can have a major impact on the values obtained in the output. Thus, a complete description of what is being measured by a terminating simulation must include a statement of the starting and stopping rules.

Steady-state simulations. As opposed to terminating simulations, a steady-state simulation is one in which the output distributions are defined with respect to a limit as the length of the simulation becomes infinite. There is no dependence on the initial conditions, nor is there any rule specified to stop the simulation (as indeed it must). The quantities to be measured are defined in a way which makes them independent of the initial conditions chosen, so theoretically one could initialize the simulation in any convenient way; practically, however, the initial conditions actually chosen for a steady-state simulation run can have a great impact, as we shall discuss in Section 3.5. The open-ended stopping of a steady-state simulation also poses great problems for the analyst, who must somehow choose a way to stop the run(s) when it somehow appears that they are "long enough." Thus it seems clear that the analysis problem is considerably more difficult in the case of steady-state simulations.

The question arises as to whether there really are any situations in which a steady-state analysis would be appropriate anyway. After all, time can never really "go to infinity." In addition, the analysis problem for terminating simulations is far simpler. Furthermore, it may be that the focus of much of the simulation analysis literature on the steady-state case is just a result of following in the footsteps of

mathematical queueing and stochastic process theory, where it is generally necessary to take a limit as time becomes infinite in order to get any tractable results. Nevertheless, there are some reasons for considering steady-state analysis. For example, an industrial operation may operate on a 24-hour, every-day schedule, and there is interest in observing how the system might behave after an initial period when machines are new, operators are inexperienced, etc. Another application is to study a system under a hypothetically indefinite period of peak-loading in an attempt to design conservatively for some kind of perceived worst-case scenario.

It is likely that both terminating and steady-state simulations find appropriate application. Since the available techniques for analyzing them are quite different, the following two subsections treat these problems separately.

3.4 Analysis for a Terminating Simulation

For a terminating simulation, the starting and stopping rules are embodied in the model specification, so there is no question about how to run the simulation. From a single run, or replication, we obtain one or several output measures, e.g., the average delay in queue of the first 500 customers, or the observed proportion of time a server was busy during this particular run. Since the goal is to estimate some property of the distributions of these measures, such as their expectations, we need to make several independent replications of the model to be able to use standard statistical analysis. This simple replication approach provides the basic means for statistical analysis of the output from a terminating simulation.

Let X_j denote an output measure obtained from the jth independent replication. For example, X_j could be the average of the first 500 customers' delays in the jth replication, or it might be the proportion of time a server was busy during the jth replication. X_j could also be an indicator r.v. indicating whether some condition in the output occurred, whose probability we seek to estimate. From n independent replications, we then obtain $X_1, X_2, \ldots, X_n$, which are i.i.d. r.v.'s since they result from observing the same quantity over independent runs of the same model. These X_j's form the basic units for statistical analysis.

It is important to note that one entire run of the simulation produces only a single data point, X_j, for analysis. We do not work directly with, for example, the individual customers' delays in a queueing model as the basic units of data analysis. Thus, in terms of the final statistical analysis, an entire replication of a terminating simulation constitutes a sample of size one. Statistical analysis is used only indirectly on the individual values computed within a single replication.

Given the i.i.d. X_j's, statistical inference proceeds in a fairly standard way. As an unbiased point estimator of $E(X_j)$, we use

$$\bar{X} = \sum_{j=1}^{n} X_j/n,$$

and as an unbiased point estimator of $\mathrm{Var}(X_j)$ we use

$$s^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2/(n-1). \qquad (1)$$

Under a normality assumption for the distribution of the X_j's, a $100(1-\alpha)\%$ c.i. for $E(X_j)$ is

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

where $t_{n-1,1-\alpha/2}$ is the upper $1-\alpha/2$ critical point of the t distribution with n-1 degrees of freedom. Although the X_j's will in general not be normal, the c.i. in (2) will be approximately valid if the distribution of the X_j's is not too skewed. Also, validity is enhanced for larger choices of n. Again, assuming that the X_j's are normally distributed, a c.i. can be formed for $\mathrm{Var}(X_j)$ using the chi-squared distribution, but its validity is more sensitive to the normality assumption. Welch (1983) gives an alternative c.i. for $\mathrm{Var}(X_j)$, based on the jackknifing technique, which is more robust to departures from normality of the X_j's. Finally, if the X_j's are indicator r.v.'s, an alternative c.i. for $E(X_j) = P(\text{the condition in question occurs})$, based on the binomial distribution, is possible; see, for example, Welch (1983).

One potential difficulty with the kinds of c.i.'s discussed in the preceding paragraph is that they are based on a fixed number of replications and may thus turn out to be too large to allow for making statements as precise as we wold like. In this case, one might consider a sequential procedure, in which we continue to replicate, re-forming the c.i. after each additional replication, until the c.i. is small enough, either in an absolute sense or in comparison with the point estimate, $\bar{X}$. This could require a large amount of computing if our precision demands are too stringent, since the required number of replications unfortunately grows (approximately) quadratically with the c.i. smallness criterion; i.e., to make the c.i. half as wide, we need about four times as many replications. For more on sequential procedures, see Law (1983) or Welch (1983).

The analysis problem in the terminating case is, at least conceptually, fairly straightforward. The only real difficulties are nonnormality of the X_j's and cost if the X_j's are highly variable or we need high precision. For a more complete treatment, see Law (1980).

3.5 Analysis for a Steady-State Simulation

The analysis problem in the steady-state case is much more problematic. In most steady-state simulations, we want to estimate a steady-state mean

$$\mu = \lim_{i\to\infty} E(Y_i),$$

where Y_i is an individual observation obtained

from within a run; for example, Y_i could be what we earlier called D_i, the delay in queue for the ith exiting customer. This is for a discrete-time process $\{Y_i, i = 1, 2, \ldots\}$; definition for a continuous-time process (such as our earlier Q_t and B_t) is similar.

One of the chief difficulties, which is immediately apparent, is that our goal is to estimate a quantity defined as a limit as the length of the simulation becomes infinite. However, we must obviously stop the simulation at some time, and it is not clear how to decide on a stopping point or rule. Not surprisingly, steady-state simulation runs tend to be long, and thus costly. A related difficulty is that the quantity to be estimated is defined to be independent of the initial conditions, so we are left with the practical problem of deciding how the run(s) should actually be started. Yet another (perhaps not so obvious) difficulty is that of estimating the variance of a point estimator of μ; point estimators are easier to obtain than estimators of their variances, and we need both for c.i. formation. The various methods for steady-state analysis described below take different approaches to coping with these difficulties.

The initial transient problem. By the definition of a limit, we can be sure that if we run the simulation long enough, the means of the output sequence will be arbitrarily close to μ. This says nothing, however, about the length of time we must wait for this to happen, and the duration of this "transient" or "warmup" period can depend heavily on the way the run was initialized. The "problem of the initial transient," perhaps one of the longest-standing questions in simulation methodology (going back at least twenty years to Conway 1963) has traditionally been viewed as one of identifying the extent of the transient period, relative to some practical criterion. The usual procedure is then to delete (or truncate) this initial period from consideration, and use only the Y_i's past the deleted portion in forming output measures. Presumably, this eliminates (or at least greatly reduces) the bias induced by the fact that the initial conditions cannot generally be chosen in accordance with steady-state behavior of the system. (If we knew the steady-state distributions involved, there would be no need to simulate.) Several methods have been developed for identifying the length of the transient; see the comprehensive survey of Gafarian, Ancker, and Morisaku (1978), and the more recently proposed methods in Schruben (1981,1982b), Welch (1983), and Kelton and Law (1983).

The extent of the transient period is certainly influenced by the initial consitions. Thus, to reduce the amount of initial output that must be discarded, it is probably worth giving some thought to choosing initial conditions which appear, at least, to bear some rough resemblance to anticipated steady-state behavior. Wilson and Pritsker (1978) evaluated tradeoffs between deletion and choosing "good" initialization, and concluded that the latter method is a more effective means of dealing with the initial transient problem.

Replication. This method of steady-state analysis is really the same as that for terminating simulations discussed above, except that the goal is now to estimate a steady-state parameter. The simulation is replicated, starting with the same initial conditions and stopping according to the same rule to produce i.i.d. X_j's, as before, which are used as the basic units for statistical analysis. In this case, however, it is not clear how we should initialize, or how long the replications should last. Due to the initial transient problem, the X_j's may not have expectation μ, i.e., they will generally be biased. This bias may invalidate the statistical analysis; e.g., the c.i. in (2) may have an actual probability of containing μ that is far below the desired level, $1-\alpha$.

Thus, the main drawback to the replication approach to steady-state analysis is the problem of the initial transient. Three approaches to dealing with it are (1) choose better initial conditions, (2) effectively delete the initial transient segments from each replication, and (3) make the replications very long; these three ideas could be used together. Quantitatively, however, implementing these ideas in practice can pose difficult questions; e.g., how long should a "long" run be? If the analyst feels that the initial transient problem has been effectively dealt with, then replication is attractive due to the built in independence of the runs (a property not enjoyed by most of the other steady-state analysis methods described below). However, the consequences of not dealing effectively with the initial transient problem can be severe in terms of the validity of the ensuing analysis. For example, c.i. coverage probabilities can actually worsen as more data are taken; see Law (1977). In any case, it is probably good practice to choose to make a few long replications rather than many short ones to avoid having to pass through the transient phase many times. It would appear that in choosing fewer but longer runs we lose efficiency (e.g., the c.i. widens) since n is smaller, but the longer runs will tend to produce less variable X_j's, having the opposite effect on efficiency.

Batch means. Since the main problem with replication is repeatedly having to pass through the transient, it seems promising to develop a method based on only a single run, presumably very long. The method of batch means (as well as all the other methods described below) takes this approach. By only having to pass through the transient once, its effect on the output is greatly diminished; it still may be desirable to delete some initial portion of the run's output, to be conservative. As a point estimator of μ, we use $\bar{X}$, the mean of all the (undeleted) output from the run, which should be unbiased, at least approximately. The difficult task is then estimating the variance of $\bar{X}$, which is needed for c.i. formation as well as other inferential goals. The source of the difficulty is that in making a single run rather than multiple replications, we have lost the independence of the output, which is crucial for unbiased variance esimation.

The method of batch means attempts to regain some degree of independence by breaking the output

record into subsequences called "batches." From a run of length m (after any deletion) Y_i's, form n adjacent batches of k successive Y_i's each (where m = nk), and let X_j be the average of the k Y_i's in the jth batch; these X_j's form the basic units of analysis. Noting that the average of the X_j's is always equal to $\bar{\bar{X}}$, we estimate the variance of $\bar{\bar{X}}$ by s^2 as given formally in (1), except that the X_j's are now batch means rather than replication means. Finally, the c.i. for μ is formed as in (2).

As in the case of terminating simulations, the c.i. in (2) may not be valid in this case since the X_j's are not normal. More seriously, the X_j's are not independent, having arisen from a single simulation run. Thus, s^2 is generally a biased estimator of $\text{Var}(\bar{\bar{X}})$, and is biased low if the Y_i's (and thus the X_j's) are positively correlated, which is typically the case in queueing-type simulations. This in turn can cause the c.i. to be too small and to have a lower-than-desired probability of containing μ. The idea behind batching is to make the batch means approximately uncorrelated (and, as a secondary goal, more symmetrically distributed than the underlying Y_i's through the averaging operation), which in turn should make s^2 approximately unbiased for $\text{Var}(\bar{\bar{X}})$, and enhance the c.i.'s validity. The reason for suspecting that the batch means may be nearly uncorrelated for most Y_i processes encountered in practice lies in the belief that only those Y_i's which are near to each other in time are likely to be heavily correlated. Thus, if the batch size is large, different X_j's will, at worst (if they are adjacent batches), be computed from only relatively few Y_i's which are close together, just on either side of the batch boundary. Thus, we would expect that X_j and X_{j+1} should be only weakly correlated, provided the batch size is chosen large enough.

Once again, the analyst is faced with the practical, quantitative problem of deciding just what "large enough batch size" means in a given application. Generally, it is good advice to opt for a few large batches rather than many small batches if a tradeoff must be made. There may be little to be gained in efficiency yet much to be lost in validity if one chooses to split the run into many batches which are too small; see Schmeiser (1982). For further papers concerning batch size selection rules and their evaluation, see Fishman (1978a), Law (1977), Law and Carson (1979), or Law and Kelton (1983).

Spectral analysis. The method of batch means attempts to obtain (nearly) uncorrelated observations to estimate the variance of the point estimator $\bar{\bar{X}}$. The method of spectral analysis, on the other hand, uses a relation between $\text{Var}(\bar{\bar{X}})$ and the autocorrelation structure of the Y_i's to form a variance estimate. Thus, there is no attempt to approximate an i.i.d. situation.

If the Y_i's are covariance stationary (i.e., the covariance between Y_i and Y_{i+k} depends only on k and not on i), then

$$\text{Var}(\bar{\bar{X}}) = [C_0 + 2\sum_{k=1}^{m-1}(1 - k/m)C_k]/m, \quad (3)$$

where C_k is the covariance between Y_i and Y_{i+k}. Thus, if we had good estimates of the C_k's, we could insert them into (3) to obtain a variance estimate. The standard estimators of the C_k's, however, do not have good statistical properties (e.g., they are highly variable for large k, and are themselves heavily correlated), and estimating many of them can be extremely time-consuming on a computer. The method of spectral analysis modifies (3) to obtain a statistically better variance estimator; for details see Law and Kelton (1983) and references there.

To form a c.i. for μ the degrees of freedom must be specified, which depends on the particular implementation. Recent extensions and improvements to the method have been undertaken by Heidelberger and Welch (1981a, 1981b).

Autoregressive representation. Like spectral analysis, the autoregressive method, developed by Fishman (1971, 1978b), attempts to use the autocorrelation structure in the Y_i's to estimate $\text{Var}(\bar{\bar{X}})$ and form a c.i. for μ, but in a different way. Since the output sequence of Y_i's can be viewed as a time series, this method assumes that it can be closely represented as a certain time series model, called an autoregressive model, in which each Y_i is a linear combination of the p previous Y_i's, plus a random disturbance term. Such a model is "fitted" to the output sequence, including an empirical determination of the autoregressive order, p. From the fitted model, a variance estimate, degrees of freedom, and c.i. are obtained. For details, see the original references given earlier in this paragraph, or the survey of Law and Kelton (1983), which also contains results on the performance of this method.

The regenerative method. One drawback of all of the above methods for steady-state analysis is that they are not as mathematically well-grounded as, say, classical statistics for i.i.d. data. The regenerative method, developed by Crane and Iglehart (1974a, 1974b, 1975) and by Fishman (1973, 1974) puts the analysis problem on a much firmer footing, and also eliminates the problem of the initial transient from many estimation goals. The price paid is that extra assumptions must be made concerning the Y_i process being simulated, and the simulation program itself must be modified to recognize certain conditions under which it must halt and tally data.

The extra assumption made is that the Y_i process is regenerative, which loosely means that at certain random but repeating points in simulated time called regeneration points, the process "starts over probibalistically" and is independent of the past. The evolution of the process between successive regeneration points is called a cycle (or tour), and what happens during a cycle is an i.i.d. replicate of what happens during any other cycle. Observations are collected within cycles and are then combined to

form point estimates and c.i.'s for μ in several alternative ways. For details, see the references in the preceding paragraph, Crane and Lemoine (1977), or Iglehart (1975).

Standardized time series. A new basic methodology to steady-state analysis is currently being developed by Schruben (1982a) and Goldsman and Schruben (1982) that is based on weak assumptions about the underlying Y_i process, but does not necessarily assume that it is regenerative. The output sequence is standardized to have mean zero and run duration on the unit interval, and a central limit theorem for this entire process is developed which leads to several ways of forming c.i.'s. This new approach to the steady-state analysis problem has the appeal of being well-grounded in probability theory, yet avoids making strong assumptions about the process or having to write the simulation program in a different way.

Sequential procedures. The discussion of the steady-state analysis techniques above was mostly in the context of a fixed sample size, i.e., the run lengths and (in the case of the method of replication) number of replications was assumed to be fixed and prespecified. However, the resulting c.i.'s may be too wide to be useful, so sequential procedures might be considered, as in the analysis of terminating simulations. In the steady-state case, sequential procedures generally increase the length of a (single) run. In addition to reducing c.i. width, a sequential procedure designed to drive the run length farther should also benefit the c.i. by making it more valid, i.e., having the desired probability of covering μ. A survey and evaluation of several steady-state sequential procedures appears in Law and Kelton (1982a), and other methods based on spectral analysis are proposed in Heidelberger and Welch (1981a, 1981b). In general, sequential procedures can be expected to produce c.i.'s which are more valid (and narrow) than their fixed-sample-size counterparts, but can result in extremely long runs if allowed to work in a purely automatic mode.

4. STUDYING SEVERAL SYSTEMS

Section 3 focused on analysis from the output of a single system design. In many (perhaps most) simulation projects, however, there is interest in several alternative system designs or specifications, and the goal of the study is to compare them, rank order them, select one or some of them as the best in some sense, or to conduct a program of experimentation designed to lead to an optimal specification. Statistical methodologies for accomplishing goals such as these are outlined in chapters 9 and 12 of Law and Kelton (1982b). More detailed treatments are Dudewicz and Koo (1981), Myers (1971), and Biles and Swain (1980).

Nearly all statistical methods designed to accomplish such goals assume the ability to collect i.i.d. data, with expectation equal to the system performance parameter with respect to which the comparison, ranking, or optimization is to be done. In the case of terminating simulations, this presents no problem, since we just use the output measure X_j obtained from the jth replication. For steady state simulations, however, there is not a clear choice. One possibility would be to use the method of replication, together with an effective initialization/deletion rule to deal with the initial transient problem. Another possibility would be to use batch means as an approximation to i.i.d. data. In either case it is important for the validity of the statistical methodology that the analyst take care that the basic units of data analysis (the X_j's) can be regarded as being independent and with the desired expectation.

5. VARIANCE REDUCTION TECHNIQUES

In most statistical experimentation, the measurements being taken have some inherent uncertainty associated with them, perhaps induced by natural forces. The precision of the results of the analysis is limited by this natural variability, and can usually be improved only by increasing the amount of data, which incurs additional cost. However, the source of variation in a stochastic computer simulation is the (pseudo) random number generator, which can be controlled by the simulator. Variance reduction techniques (VRT's) are, for the most part, schemes which exploit this ability to control the random number generator and carry out the simulation in something other than a straightforward manner to reduce the variability of the output at no (or very little) extra cost. If successful, a VRT can give us more precise results for the same computing effort, or (equivalently stated) give us the same precision at a reduced effort.

There are several different kinds of VRT's, such as common random numbers, antithetic variates, control variates, rotation sampling, conditional Monte Carlo, and indirect estimation, designed for different situations. Most of these rely on re-using random numbers, either directly or in a transformed form. For a survey of several of these, see chapters 2 and 3 of Fishman (1978b) or chapter 11 of Law and Kelton (1982b).

6. CONCLUSIONS AND PROSPECTS

Probably the most important point to be made is that an analysis should be an integral, planned part of any simulation study. Especially for stochastic simulations, the lack of a proper analysis leaves one with results that may be misleading and inaccurate. If one makes the substantial effort to validate, code, and debug a complex simulation model, it seems worthwhile to make some effort to use the model effectively and interpret its output appropriately. One impediment to proper simulation analysis has been the high cost of running and replicating a time-consuming computer simulation program. Fortunately, we are now seeing dramatic increases in computer speed, accompanied by a fall in the cost of computing. Thus, it seems likely that this impediment could be in the process of becoming far less binding. However, even if we had infinitely fast computers whose use were free, it would still be quite necessary to make

sure that an appropriate and valid analysis technique is used to insure the accuracy and reliability of the results.

REFERENCES

Anderson TW (1971), The Statistical Analysis of Time Series, Wiley, New York.

Biles WE, Swain JJ (1980), Optimization and Industrial Experimentation, Wiley, New York.

Box GEP, Hunter WG, Hunter JS (1978), Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, Wiley, New York.

Conway RW (1963), Some Tactical Problems in Digital Simulation, Mgmt. Sci., Vol 10, pp. 47-61.

Crane MA, Iglehart DL (1974a), Simulating Stable Stochastic Systems, I: General Multiserver Queues, J. Assoc. Comput. Mach., Vol 21, pp. 103-113.

Crane MA, Iglehart DL (1974b), Simulating Stable Stochastic Systems, II: Markov Chains, J. Assoc. Comput. Mach., Vol 21, pp. 114-123.

Crane MA, Iglehart DL (1975), Simulating Stable Stochastic Systems, III: Regenerative Processes and Discrete-Event Simulations, Opns. Res., Vol. 23, pp. 33-45.

Crane MA, Lemoine AJ (1977), An Introduction to the Regenerative Method for Simulation Analysis, Springer-Verlag, New York.

Dudewicz EJ, Koo JO (1981), The Complete Categorized Guide to Statistical Selection and Ranking Procedures, American Sciences Press, Columbus, Ohio.

Fishman GS (1971), Estimating Sample Size in Computer Simulation Experiments, Mgmt. Sci., Vol. 18, pp. 21-38.

Fishman GS (1973), Statistical Analysis for Queueing Simulations, Mgmt. Sci., Vol. 20, pp. 363-369.

Fishman GS (1974), Estimation in Multiserver Queueing Simulations, Opns. Res., Vol. 22, pp. 72-78.

Fishman GS (1978a), Grouping Observations in Digital Simulation, Mgmt. Sci., Vol. 24, pp. 510-521.

Fishman GS (1978b), Principles of Discrete Event Simulation, Wiley, New York.

Gafarian AV, Ancker CJ Jr., Morisaku, T. (1978), Evaluation of Commonly Used Rules for Detecting "Steady State" in Computer Simulation, Naval Res. Logist. Quart., Vol. 25, pp. 511-529.

Goldsman D, Schruben LW (1982), Asymptotic Properties of Some Confidence Interval Estimators, Technical Report No. 544, School of Operations Research and Industrial Engineering, Cornell University.

Hammersley JM, Handscomb DC (1964), Monte Carlo Methods, Methuen, London.

Heidelberger P, Lewis PAW (1981), Quantile Estimation in Dependent Sequences, IBM Research Report RC 9087, Yorktown Heights, New York.

Heidelberger P, Welch PD (1981a), A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations, Comm. Assoc. Comput. Mach., Vol. 24, pp. 233-245.

Heidelberger P, Welch PD (1981b), Adaptive Spectral Methods for Simulation Output Analysis, IBM J. Res. Develop., Vol. 25, pp. 860-876.

Iglehart DL (1975), Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators, Naval Res. Logist. Quart., Vol. 22, pp. 553-565.

Iglehart DL (1976), Simulating Stable Stochastic Systems, VI: Quantile Estimation, J. Assoc. Comput. Mach., Vol. 23, pp. 347-360.

Kelton WD, Law AM (1983), A New Approach for Dealing with the Startup Problem in Discrete Event Simulation, Naval Res. Logist. Quart., to appear.

Law AM (1977), Confidence Intervals in Discrete Event Simulation: A Comparison of Replication and Batch Means, Naval Res. Logist. Quart., Vol. 24, pp. 667-678.

Law AM (1980), Statistical Analysis of the Output Data from Terminating Simulations, Naval Res. Logist. Quart., Vol. 27, pp. 131-143.

Law AM (1983), Statistical Analysis of Simulation Output Data: The State of the Art, Opns. Res., to appear.

Law AM, Carson JS (1979), A Sequential Procedure for Determining the Length of a Steady-State Simulation, Opns. Res., Vol. 27, pp. 1011-1025.

Law AM, Kelton WD (1982a), Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures, Mgmt. Sci., Vol. 28, pp. 550-562.

Law AM, Kelton WD (1982b), Simulation Modeling and Analysis, McGraw-Hill, New York.

Law AM, Kelton WD (1983), Confidence Intervals for Steady-State Simulations, I: A Survey of Fixed Sample Size Procedures, Opns. Res., to appear.

Myers RH (1971), Response Surface Methodology, Allyn and Bacon, Boston.

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

Schmeiser BW (1982), Batch Size Effects in the Analysis of Simulation Output, Opns. Res., Vol. 30, pp. 556-568.

Schruben LW (1981), Control of Initialization Bias in Multivariate Simulation Response, Commun. Assoc. Comput. Mach., Vol. 24, pp. 246-252.

Schruben LW (1982a), Confidence Interval Estimation using Standardized Time Series, Technical Report No. 518, School of Operations Research and Industrial Engineering, Cornell University.

Schruben LW (1982b), Detecting Initialization Bias in Simulation Output, Opns. Res., Vol. 30, pp. 569-590.

Seila AF (1981), Percentile Estimation in Discrete Event Simulation, Dept. of Quantitative Business Analysis, University of Georgia.

Seila AF (1982), A Batching Approach to Quantile Estimation in Regenerative Simulations, Mgmt. Sci., Vol. 28, pp. 573-581.

Welch PD (1983), The Statistical Analysis of Simulation Results. In: Computer Performance Modeling Handbook, S Lavenberg (ed.), Academic Press, New York, pp. 267-329.

Wilson JR, Pritsker AAB (1978), Evaluation of Startup Policies in Simulation Experiments, Simulation, Vol. 31, pp. 79-89.