Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

STEADY-STATE CONFIDENCE INTERVAL METHODOLOGY: A FORUM ON THEORY, PRACTICE, AND PROSPECTS

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

PARTICIPANTS

Averill M. Law
Department of Management Information Systems
University of Arizona
Tucson, Arizona 85721

Bruce W. Schmeiser
School of Industrial Engineering
Purdue University
West Lafayette, Indiana 47907

Richard W. Andrews
School of Business Administration
The University of Michigan
Ann Arbor, Michigan 48109

Peter D. Welch
IBM Watson Research Center
P.O. Box 218
Yorktown Heights, New York 10598

Peter W. Glynn
Department of Industrial Engineering
University of Wisconsin
Madison, Wisconsin 53706

Donald L. Iglehart
Department of Operations Research
Stanford University
Stanford, California 94305

Lee W. Schruben
School of Operations Research
and Industrial Engineering
Cornell University
Ithaca, New York 14853

INTRODUCTION

For the past two decades, much of the attention in the statistical methodology of computer simulation has been motivated by and directed toward what is essentially a single goal: To construct a valid confidence interval for a steady-state parameter of a stochastic process by means of simulation. During this time, a number of approaches have been developed in an attempt to meet this goal, and the purpose of this two-session forum is to gather several of these ideas together for exposition and discussion. Each participant in this forum is actively involved in research in this area, and has agreed to present or co-present one of six methods: Replication, batch means, time series, spectral methods, regenerative methods, and standardized time series. The order chosen represents an attempt to follow the chairperson's impression of the chronology of development.

Several definitions of the basic problem are possible (and are often equivalent), perhaps the most direct of which is as follows. Let $\{X_i, i \in \{1, 2, \dots\}\}$ be a discrete-time stochastic process and assume that $\mu = \lim_{i\to\infty} E(X_i)$ exists. The goal is to construct a $100(1 - \alpha)\%$ confidence interval for μ from the simulation output, i.e. to form an interval $[L, U]$ where L and U are statistics observable from the simulation output, such that $P\{L \leq \mu \leq U\} = 1 - \alpha$.

Alternatively, we may wish to observe a continuous-time process $\{X_t, t \in [0, \infty)\}$ and construct a confidence interval for $\mu = \lim_{t\to\infty} E(X_t)$. A different approach is to consider the time-dependent distribution functions $F_i(x) = P\{X_i \leq x\}$ (or $F_t(x)$ in the continuous-time case), and assume that $\lim_{i\to\infty} F_i(x) = F(x)$ where $F(x)$ is the distribution function of some random variable X, called the steady-state random variable; the goal is to form a confidence interval for $E(X)$ or (more generally) $E[g(X)]$ where g is a measurable function and may be chosen, for example, to define moments of X or indicate whether X falls in some interval.

In the remainder of this paper, each participant presents a short exposition and discussion of the method he has been asked to represent.

REPLICATION -- Averill M. Law

Steady-state simulations are often appropriate when one wants to determine the "long-run" behavior of manufacturing systems, computer systems, etc. The method of replication/deletion for steady-state analysis proceeds by making independent replications of the simulation, with the initial portion of each run not actually being used in the analysis. We discuss graphical (see Welch [1983]) and statistical techniques (see Schruben [1982]) for deciding on the length of the initial data deletion.

The method of replication/deletion is important because it is very widely used in practice, is similar to the approach used for terminating simulations, and easily accommodates multiple measures of performance.

References

Schruben, L., "Detecting Initialization Bias in Simulation Output," Operations Research, 30, 1982.

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

BATCH MEANS -- Bruce W. Schmeiser

Consider a steady-state stochastic process $\{X\}$ having mean μ and finite second moment. The classical batch means confidence interval for μ is $\bar{X} \pm H_{\alpha,k}$, where $H_{\alpha,k} = t_{\alpha/2,k-1} S_k/\sqrt{k}$ and $S_k^2 = (\sum_{i=1}^{k} \bar{X}_i^2 - k\bar{X}^2)/(k-1)$. The batch means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ may be averages of adjacent discrete observations (such as customer wait times), $\bar{X}_i = (\sum_{j=(i-1)m+1}^{im} X_j)/m$ with $\bar{X} = (\sum_{j=1}^{n} X_j)/n$; the integral of a continuous-time process (such as number in the system), $\bar{X}_i = (\int_{(i-1)t}^{it} X(s)\, ds)/t$ with $\bar{X} = (\int_0^T X(s)\, ds)/T$; or a count process average (such as number of blocked customers), $\bar{X}_i = (N(it)-N((i-1)t))/t$ with $\bar{X} = N(T)/T$. The number of batches is $k = \lfloor n/m \rfloor$ in the first case and $k = \lfloor T/t \rfloor$ in the second and third cases.

Theoretical Foundation

We discuss the classical batch means method assuming the first case of n discrete observations. Analogous statements can be made for the continuous cases almost directly by replacing summations with integrations.

For a constant sample size n and assuming steady state, $\bar{X}$ is an unbiased estimator for μ and $V\{\bar{X}\} = (R_0 + 2\sum_{i=1}^{n-1} (1-(i/n))R_i)/n$, where R_i is the i^{th} lag covariance between X_j and X_{j+i}.

We say the confidence interval $\bar{X} \pm H_{\alpha,k}$ is **valid** if $P\{|\mu-\bar{X}| \leq H_{\alpha,k}\} = 1-\alpha$, which is true if $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ are independent and identically normally distributed (NIID). (For reassurance that NIID conditions occur as $n \to \infty$ and $m \to \infty$, see Brillinger [1973]). When the batch size m is large enough to provide an essentially NIID sequence of batch means, then $E\{S_k^2/k\} \cong V\{\bar{X}\}$ and $\mathrm{Cov}\{\bar{X}, S_k^2\} \cong 0$, resulting in a valid interval.

Practical Implementation Issues

The key issue for implementation is the selection of the batch size m, or equivalently for fixed n the number of batches k, to provide a valid confidence interval procedure with good statistical properties. The usual good properties are that $E\{H_{\alpha,k}\}$, $V\{H_{\alpha,k}\}$, $CV\{H_{\alpha,k}\}$, and $P\{|\mu_1-\bar{X}| \leq H_{\alpha,k}\}$ for $\mu_1 \neq \mu$ are small. Since the magnitude of the bias $|V\{\bar{X}\}-E\{S_k^2/k\}|$ is increasing in k while the variance $V\{S_k^2/k\}$ is decreasing in k, the selection is not easy. Validity of the procedure calls for a small number of batches and the other properties call for a large number of batches.

Although these other properties are the reasonable measure of performance for confidence interval procedures, they cannot be estimated by a practitioner, so typically batch means algorithms choose k (in a variety of ways) based on statistical tests of independence of the batch means (Fishman [1978], Law and Carson [1979], Mechanic and McKay [1966], and Schriber and Andrews [1979]).

For discussion purposes, I will take the following position:

1. Based on familiarity with the actual system or with the simulation model during debugging and validation, practitioners usually have a good idea of the number of observations, n_0, or the length of time, t_0, required for approximate independence. Determining n_0 or t_0 typically involves answering the question "Does knowing the state of the system tell you anything about the state of the system n_0 customers (t_0 time units) later?" for various potential values of these variables.

2. Using the classical batch means algorithm with $k = (n/n_0)^f$ or $k = (T/t_0)^f$ with f a constant somewhere between 0.5 and 1.0 results in almost valid confidence intervals with reasonable statistical properties.

Item 1 is simply my opinion. I will quickly grant that there are some practitioners who cannot, or will not, provide reasonable values for n_0 or t_0. Also note that an exception in which the simulation modeler usually has very poor notions about n_0 and t_0 is simple queueing systems with heavy traffic, but then almost no one has observed such systems.

Item 2 could take many forms, and I certainly don't know the best value for f in any sense, but a value such as f = 0.5 (which is easy to calculate) causes many batches to be used only in a very conservative manner, which is consistent with the following point:

> Ten batches is enough for most purposes and thirty is almost as good as k = n.

We support this point in the remainder of this section following arguments in Schmeiser (1982).

Consider the coefficient of variation (CV) and mean of confidence interval half width for k = 2, 10, 30, and 121 batches (or replications) and levels of significance $\alpha = 0.10$ and 0.01. Assuming batch sizes of n/121 are large enough to provide NIID means, the following results from Schmeiser (1982, Table 1) apply:

k	$CV\{H_{\alpha,k}\}$	$E\{H_{0.10,k}\}$	$E\{H_{0.01,k}\}$
2	0.76	5.04	50.8
10	0.24	1.78	3.16
30	0.13	1.68	2.73
121	0.06	1.65	2.61

where $H_{\alpha,k}$ is the half width of a confidence interval with level of significance α and based on k batches. The units of $E\{H_{\alpha,k}\}$ are $(V\{\bar{X}\})^{0.5}$, which is a function of neither α nor k. Therefore, the results in the table are valid for any number of observations n, underlying correlation structure, and variance.

Clearly k = 121 is better than k = 2, 10, or 30, since the coefficient of variation and expected half width are less. However the difference between k = 30 and k = 121 is not great and is incurred at some risk. For batch means the risk is that the k = 121 batch means will be less normal and independent than k = 2, 10, or 30 batch means, resulting in poorer probability of covering the mean. For replications the risk is that k = 121 replications will be less normal and more biased (due to the initial transient) than k = 2, 10, or 30 replication means. Since similar behavior occurs in terms of coverage probabilities, and since additional batches require additional memory and computation, there seems little reason for k to be greater than 30.

On the other hand, using fewer than $k = 10$ batch or replication means results in substantially shorter confidence intervals than when $k < 10$, again under the assumptions of normality and independence. Thus, using a sample size n large enough to allow batches or replications of size n/10 to be almost normal and independent has a reward beyond the reduction of $V\{\bar{X}\}$.

Three points should be noted:

1. Increasing k while keeping the batch or replication length constant reduces $V\{\bar{X}\}$ approximately inversely with k since the number of observations is then proportional to k. The expected half width then decreases inversely with $k^{0.5}$. Schmeiser (1982) does not conclude otherwise, since only a fixed value of n is considered, although misinterpretation of the paper seems common.

2. The "fixed value of n" may arise in either a fixed or sequential batching algorithm. In the context of a sequential algorithm, the result is that when the algorithm is ready to terminate and calculate the confidence interval to be returned to the user, using more than 30 batches adds little improvement.

3. As noted in Law and Carson (1979), hundreds of batches may be necessary to determine adequately the degree of non-normality and dependence. However, after determining that a confidence interval is to be calculated, then (as in point 2 above) no more than about 30 batch means should be used.

Future Directions

The future of batch means in practice is to continue being the most used method other than independent replications for statistically analyzing simulation output, because batch means requires few assumptions, is easy to understand and is easy to implement. Inclusion in languages and packages remains necessary for widespread application of any output analysis method.

The future of batch means in research follows several directions, four of which are briefly mentioned here.

Seila (1984) considers batching discrete observations using equal time intervals, which results in a ratio estimator.

Bischak and Kelton (1984) are examining deletion of observations between batches to decrease batch mean dependence, which is thought to be more crucial than normality, which such a procedure hinders. They find "... the best coverage resulted from the strategy of deleting a large percentage of observations from each of a small number of batches, but the smallest half-length for a given run length resulted from batching observations into many small batches and averaging without deleting data." These conclusions are consistent with the discussion in the previous section.

Meketon (1980) studied the variance time curve $V(t) = \lim_{s \to \infty} V\{Z(s+t) - Z(s)\}$, where Z(s) is the cumulative (sum or integral) process observed at time s. V(t) is valuable in our context because $V\{\bar{X}\} = V(T)/T^2$, where $\bar{X} = Z(T)/T$ and an estimate of V(t) is the sample variance time curve

$$\hat{V}_T(t) = \int_0^{T-t} (Z(s+t) - Z(s) - tZ(T)/T)^2 ds)/ (T - t),$$

which is an estimator based on overlapping batch means.

Kang (1984) considered properties of batch means when the underlying process is autoregressive moving average. His analytic and numerical procedures provide insight into the structure of batch means in a wide (although certainly not all-inclusive) setting. In addition Monte Carlo studies can be performed efficiently since the batch means can be generated directly rather than by aggregating the underlying process.

Acknowledgment

This paper benefitted from discussions with Marc Meketon and from the time and resources provided by AT&T Bell Laboratories, where the author spent the summer of 1984 with the Operations Research Department.

References

Bischak, D.P. and W.D. Kelton, personal communication, Department of Industrial and Operations Engineering, The University of Michigan, 1984.

Brillinger, D.R., "Estimation of the Mean of a Stationary Time Series by Sampling," J. Appl. Prob., 10, 419-431, 1973.

Fishman, G.S., "Grouping Observations in Digital Simulation," Management Science, 24, 510-521, 1978.

Kang, K., "Confidence Interval Estimation via Batch Means and Time Series Modeling," School of Industrial Engineering, Purdue University, 1984.

Law, A.M. and J.S. Carson II, "A Sequential Procedure for Determining the Length of a Steady State Simulation," Operations Research, 27, 1011-1025, 1979.

Mechanic, H. and W. McKay, "Confidence Intervals for Averages of Dependent Data in Simulations II," Technical Report ASDD 17-202, IBM Corp., Yorktown Heights, New York, 1966.

Meketon, M.S., "The Variance Time Curve: Theory and Application," School of Operations Research and Industrial Engineering, Cornell University, 1980.

Schriber, T.J. and R.W. Andrews, "Interactive Analysis of Simulation Output by the Method of Batch Means," Proceedings of the Winter Simulation Conference, 513-525, 1979.

Schmeiser, B.W., "Batch Size Effects in the Analysis of Simulation Output," Operations Research, 30, 556-568, 1982.

Seila, A.F., "Batch Ratios in Discrete Event Simulation," Working Paper Series 84-150, College of Business Administration, The University of Georgia, 1984.

TIME SERIES -- Richard W. Andrews

The objective of this report is to describe briefly the Autoregressive Moving-Average (ARMA) confidence interval for the mean of a stationary stochastic process. A complete description of this methodology is contained in Schriber and Andrews (1984, referred to as S&A). The report will consist of answers to five questions. The five questions are:

1. What are the **assumptions** of the methodology?
2. What **settings** must be made by the user?
3. Is the methodology **robust** to the assumptions?
4. Does the methodology provide the type of **answer** desired?
5. How well does the methodology **perform**?

Before answering these five questions a few general comments concerning confidence intervals is appropriate. The concept of a confidence interval has as its origin the seminal paper of Neyman (1934). Neyman's approach is based on the foundations of statistical inference referred to as sampling theory. Sampling theory evaluates an inference procedure by how well it does in repeated samples.

In the simulation output analysis literature, the presentation of a new confidence interval methodology usually includes an empirical study. As part of the empirical study the coverage of the confidence interval method is reported. That coverage is found by using repeated samples; therefore the sampling theory approach is used. The likelihood principle provides an alternative foundation from which to view a confidence interval. In the closing section on Future Work the likelihood principle will be discussed. The ensuing answers to the five questions assume that only the sampling theory foundation of statistical inference is appropriate.

Assumptions

The output random variable, X, is assumed to be described by a stationary ARMA model:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} +$$
$$\theta_0 + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}$$

$$\epsilon_t \sim N(0,\sigma^2), \text{ for all } t$$

$$E(\epsilon_t \epsilon_s) = \begin{cases} \sigma^2 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases}$$

$$\text{Cov}(\epsilon_t, X_s) = 0 \text{ if } t > s.$$

For this model to be stationary all roots (solutions for B) of the equation

$$1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 0$$

must lie outside the unit circle in the complex plane. The normality assumption for the shock term, ϵ_t, is necessary in order to test hypotheses concerning the model parameters and in order to provide a distribution for the construction of a confidence interval. A further comment concerning the normality assumption will be given in the section on Robustness.

The choice of an ARMA model for an output random variable is compelling because of the following theorem.

Predictive Decomposition Theorem (Wold 1954)

Any stationary stochastic process, X_t, with finite variance, can be expressed as

$$X_t = \Phi_t + \Psi_t$$

for which

(a) the two components Φ_t and Ψ_t are uncorrelated,

(b) Φ_t is purely deterministic and Ψ_t is purely indeterministic (see Cox and Miller [1965 p. 287]),

(c) the purely deterministic component can be linearly predicted on the basis of past observations, and

(d) the purely indeterministic component allows a representation of the form

$$\Psi_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots$$
$$\text{with } \Sigma\theta_t^2 < \infty \text{ and Cov}(\epsilon_t, X_s) = 0 \text{ if } t > s.$$

This theorem is the foundation for the extensive use of the ARMA model; however it does not give any limits to the values of p and q. The procedures for determining the values for p and q are discussed in the Settings section. The procedures are based on the principle of parsimony. It is stated in Box and Jenkins (1976 p. 17) that "It is important, in practice, that we employ the **smallest possible** number of parameters for adequate representation." Furthermore, practice has shown that ARMA models with small values of p and q ($\leq$ 3) adequately model observed processes.

The output random variable is assumed to be observed at discrete equally spaced intervals of time. With simulation output this is not a restrictive assumption since the process can be observed whenever desired. Using the ARMA model a confidence interval will be constructed for the process mean, given by

$$\mu = \theta_0 (1 - \sum_{i=1}^{p} \phi_i)^{-1}.$$

The ARMA procedure for constructing a confidence interval for μ is (for a full discussion see S&A):

i. The values of p and q are determined.

ii. The parameters $(\phi_1, \phi_2, \ldots, \phi_p)$, $(\theta_1, \theta_2, \ldots, \theta_q)$, and σ^2 are estimated by using the conditional likelihood (see Box and Jenkins [1976 pp. 209-210]).

iii. Hypotheses tests determine if the candidate model adequately describes the data.

iv. The variance of the sample mean is estimated by using the ARMA spectral density function with its parameters set to time domain estimated values.

v. The normality assumption provides a t distribution for the confidence interval statistic.

Settings

The three settings that will be discussed are:

1. the number of observations, n;

2. the range of p and q; and

3. the significance levels for the hypothesis tests.

1. For any fixed sample size confidence interval procedure, the amount of data as measured by the sample size must be specified. The ARMA confidence interval methodology uses the raw data in two ways. First, the sample autocorrelations are formed and used in identifying the order (p,q). Second, in the estimation stage the likelihood for $\tilde{\Phi}$, $\tilde{\theta}$, and σ^2 is evaluated.

The ARMA process has been used extensively with econometric data where observations are much more difficult to obtain than simulated observations. In many cases 50 observations have been used to identify and estimate an ARMA model. In S&A the number of observations range from 100 to 400 and provide excellent results for tailor made processes.

2. In the identification stage the values of p and q must be set. One method of determining the values of p and q is by inspecting the autocorrelation function (ACF) and the partial autocorrelation function (PACF). In doing so the range of p and q is limited by the number of lags included in the ACF and PACF.

There are also various automatic identification procedures available. In S&A a procedure due to Gray, Kelly, and McIntire (1978) is used. All combinations of p and q are considered for p = 1, 2, 3 and q = 0, 1, 2, 3. As reported in S&A the automatic identification procedure correctly identifies both p and q in 74% of the replications. A higher percent of correct identifications were observed with the largest (n = 400) sample size.

3. In the diagnostic stage each of the coefficients of the candidate model is tested for significance. In addition, an overall test for lack of fit (Ljung and Box [1978]) is performed. In S&A the significance level is set at .05 for all tests. If the diagnostic tests are ignored and the estimated models are used, the coverage of the resulting confidence intervals does not change significantly.

Robustness

Because of the generality of Wold's theorem, the only two restrictive assumptions are the limits on p and q and the normality of the shock terms. The ARMA confidence interval methodology has been tested on only one queuing process, an M/M/1 queue. The output random variable was the number in system. The results are: 1) the restriction on the size of p and q is not a problem because the test statistic emphatically identified values for p and q well within the range chosen, and 2) a correlation between the estimator of μ and the standard error of the estimator shows that the normality assumption is violated.

Answer

One of the advantages of the ARMA confidence interval methodology is that the final answer provides more than an interval estimate of μ. An estimated model is available as part of the final answer. That model can be used to make additional inferences. In fact, it sometimes can be used as a surrogate for the simulated process.

Performance

In S&A the ARMA methodology was used to obtain 2800 confidence intervals. Most of the processes used to generate the observations were tailor made (Schriber and Andrews [1981]) ARMA processes. The methodology worked well on tailor made processes. The coverage was close to nominal and the properties of the half-width (relative mean and standard deviation) were appropriate. On the limited runs made with output from a simulated queue the results were not as favorable. The coverage for a 95% confidence interval ranged from 71% at n = 100 to 81% at n = 400.

Future Research

It is my belief that further confidence interval developments under the foundations of sampling theory is not the fruitful direction for research. The likelihood principle states that for any inference about a parameter, all of the information from the sample data is contained in the likelihood function (Barnett [1973 pp. 196-197]). The idea of repeated samples is irrelevant. The samples we did not observe are not important. The sample we have in hand is important. Furthermore, the likelihood foundation predates Neyman's sampling theory approach (for an interesting discussion see Good [1983 pp. 34-36]).

One of the important aspects of using the ARMA model is that the likelihood is known. For other confidence interval methodologies it is not clear what the likelihood is and it seems that some methods are justified only for repeated samples. Comparisons of confidence interval methodologies should be comparisons of likelihoods, not comparisons of coverage. Therefore, an important avenue for future research with the ARMA model is an investigation of the likelihood in terms of μ.

References

Box, G.E.P. and G.M. Jenkins. Time Series Analysis: Forecasting and Control. Rev. Ed., Holden-Day, San Francisco, 1976.

Cox, D.R. and H.D. Miller. The Theory of Stochastic Processes. Methuen and Co., London, 1965.

Good, I.J. Good Thinking: The Foundations of Probability and its Applications. University of Minnesota Press, Minneapolis, 1983.

Gray, H.L., G.D. Kelley, and D.D. McIntire, "A New Approach to ARMA Modeling," Communications in Statistics, B7, 1-77, 1978.

Ljung, G.M. and G.E.P. Box, "On a Measure of Lack of Fit in Time Series Models," Biometrika, 65, 297-303, 1978.

Neyman, J., "On Two Aspects of the Representative Method," J. Royal Statist. Soc., 97, 558-606, 1934.

Schriber, T.J. and R.W. Andrews, "A Conceptual Framework for Research in the Analysis of Simulation Output," Communications of the ACM, 24, 218-232, 1981.

Schriber, T.J. and R.W. Andrews, "ARMA-Based Confidence Intervals for Simulation Output Analysis," American J. Math. and Management Sciences, to appear.

Wold, H. A Study in the Analysis of Stationary Time Series, Almqvist & Wiksell, Stockholm, 1938.

SPECTRAL METHODS -- Peter D. Welch

In the references cited below a spectral method for confidence interval generation for steady-state simulations is proposed, developed and analyzed. In this panel discussion the speaker will discuss the advantages and disadvantages of this method, compare it as much as possible with alternatives, and suggest topics for additional research.

References

Heidelberger, P. and P.D. Welch, "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations," Communications of the ACM, 24, April 1981.

Heidelberger, P. and P.D. Welch, "Adaptive Spectral Methods for Simulation Output Analysis," IBM J. Research and Development, 25, November 1981.

Moeller, T.L. and P.D. Welch, "A Spectral Method for Generating Confidence Intervals from Simulation Outputs," Proceedings of the 1977 Winter Simulation Conference, 1977.

REGENERATIVE METHODS -- Peter W. Glynn and Donald L. Iglehart

Introduction

The regenerative method is a mathematically rigorous procedure for obtaining confidence intervals for steady state parameters. In order to properly assess the regenerative method, it is necessary to discuss those characteristics that make a confidence interval "good."

Qualitative Structure of Confidence Intervals

Given a parameter μ, a confidence interval for μ is generally based on a limit theorem of the form

$$(r_t - \mu)/v_t \Rightarrow L \tag{1}$$

as $t \to \infty$, where L is a finite random variable (r.v.) with a continuous distribution function; the parameter t measures the simulation effort required to obtain r_t and v_t. The processes r_t and v_t will be called a **point estimate** (for μ) and a **normalizing process**, respectively; we shall always assume v_t is positive. To obtain an approximate $100(1-\alpha)\%$ confidence interval for μ, select z_1, z_2 such that

$$P\{z_1 \leq L < z_2\} = 1 - \alpha.$$

Then, for large t,

$$[r_t - z_2 v_t, r_t - z_1 v_t] \tag{2}$$

contains μ with probability $1-\alpha$. The following hierarchy of properties largely determines the quality of the confidence interval.

a.) consistency of r_t: If r_t is not consistent, v_t does not tend to zero, and confidence interval half-length does not shrink to zero.

b.) asymptotic mean square error of r_t: In general, r_t is asymptotically normal. Then, there exists a non-negative σ such that

$$t^{1/2}(r_t - \mu) \Rightarrow \sigma N(0,1). \tag{3}$$

Squaring and taking expectations through (3), we observe that $\mathrm{MSE}(r_t) \sim \sigma^2/t$. Consequently, one wants to choose r_t so that σ^2 is as small as possible.

c.) expected half-width of confidence interval: By (2), the half-width of the confidence interval is $(z_2 - z_1)Ev_t$. In general, when asymptotic normality holds, $(z_2 - z_1)Ev_t \sim v/t^{1/2}$ for some v; the goal is to minimize v.

d.) Variability of half-width of confidence interval: The variance of the half-width is given by $(z_2 - z_1)^2 \mathrm{var}\, v_t$. Under quite general conditions, $(z_2 - z_1)^2 \mathrm{var}\, v_t \sim \alpha/t$; the goal is to minimize α.

e.) Approximation error: Let

$$\Delta_t = |P\{z_1 \leq (r_t - \mu)/v_t < z_2\} - P\{z_1 \leq L < z_2\}|$$

be the coverage error for the confidence interval. Berry-Esseen considerations suggest that, in general, $\Delta_t \sim \beta/t^{1/2}$; minimization of β is desirable.

The Regenerative Method

Loosely speaking, a regenerative process is one which looks like a sequence of independent and identically distributed (i.i.d.) r.v.'s, when viewed on an appropriate random time scale. More precisely, $X = \{X(t): t \geq 0\}$ is a regenerative process with regeneration times $0 = T_0 < T_1 < \cdots$ if $\{\tau_k, X(s): T_{k-1} \leq s < T_k\}$ is a sequence of i.i.d. random elements, where $\tau_k = T_k - T_{k-1}$. For examples of such processes, see Crane and Lemoine (1977). Given a real-valued function defined on the state space of X,

$$r_t = t^{-1} \int_0^t f(X(s))ds \longrightarrow r \quad \text{a.s.} \tag{4}$$

under mild assumptions on X and f. The goal of a steady state simulation is to produce confidence intervals for r.

If $N(t) = \max\{k \geq 0: T_k \leq t\}$ and $Y_i = \int_{T_{i-1}}^{T_i} f(X(s))ds$, then

$$r_t \approx \bar{Y}_{N(t)}/\bar{\tau}_{N(t)} \tag{5}$$

where $\bar{Y}_n$, $\bar{\tau}_n$ are the sample means of the Y_i's and τ_i's, respectively. Regenerative structure ensures that $\{(Y_i, \tau_i): i \geq 1\}$ is a sequence of i.i.d. random vectors, so that (4) and (5) together suggest that $r = EY_1/E\tau_1$. Then, by (5),

$$r_t - r \approx \bar{Z}_{N(t)} / \bar{\tau}_{N(t)}$$

where $Z_k = Y_k - r\tau_k$ has mean zero. Standard central limit theory arguments prove that

$$t^{1/2}(r_t - r) \Rightarrow \sigma N(0,1)$$

where $\sigma^2 \triangleq \sigma^2(Z_1)/E\tau_1$, if $E(Y_1^2 + \tau_1^2) < \infty$. Furthermore, $\eta_t \to \sigma$ a.s., where $\eta_t^2 = s_{N(t)}^2 / \bar{\tau}_{N(t)}$ and

$$s_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - (\bar{Y}_n/\bar{\tau}_n)\tau_i)^2.$$

We conclude that

$$(r_t - r) / v_t \Rightarrow N(0,1)$$

where $v_t = |\eta_t|/\sqrt{t}$ is the normalizing process for the regenerative method. The qualitative structure of the regenerative confidence interval can be summarized as follows:

a.) r_t is consistent for r

b.) $MSE(r_t) \sim \sigma^2(Z_1) / (E\tau_1 t)$ (note that <u>any</u> confidence interval method using the sample mean r_t as a point estimate will have the same MSE)

c.) $(z_2 - z_1)Ev_t \sim 2z(\alpha)\, \sigma(Z_1)/\sqrt{E\tau_1 t}$, where $z(\alpha)$ solves $P\{N(0,1) \le z(\alpha)\} = 1 - \alpha/2$

d.) $t(z_2 - z_1)^2$ var $v_t \to 0$ (in fact, $(z_2 - z_1)^2$ var $v_t \sim \alpha^2/t$; see Glynn and Iglehart [1984])

e.) $\beta$is currently unknown

Note that β is a reflection of approximation error due to the bias of r_t, and skewness/kurtosis effects. It is to be anticipated that the i.i.d. structure associated with the regenerative viewpoint can be used to reduce these errors. For example, Meketon and Heidelberger (1982) developed a point estimate which is asymptotically equivalent to r_t, but which significantly reduces bias. Also, Glynn (1982) proposed a procedure for reducing β in the closely related problem of estimating r on the time scale of regenerative cycles.

As discussed above, the regenerative method is a theoretically sound procedure for the steady state confidence interval problem. The main advantages of the method are:

i.) its good asymptotic properties (for example, $\sigma^2(v_t) = O(1/t^2)$ indicates the accurate "variance constant estimation" possible with the regenerative method)

ii.) the ability to make small-sample corrections, to reduce approximation error

iii.) the i.i.d. structure allows one to develop procedures for a host of other estimation problems (e.g. comparison of stochastic systems; see Heidelberger and Iglehart [1979])

iv.) no prior parameters are needed as input for the method, other than run length

The main disadvantages of the method are:

i.) the requirement to identify regeneration times means that the method is hard to "black box"

ii.) the method may behave unsatisfactorily if the expected time between regenerations is long

Acknowledgment

Preparation of this paper was supported by the U.S. Army Research Office under Contract DAAG29-84-K-0030 and the National Science Foundation under Grant NSF MCS 8203483.

References

Crane, M.A. and A.J. Lemoine. <u>An Introduction to the Regenerative Method for simulation Analysis</u>. (Lecture Notes in Control and Information Sciences.) Springer-Verlag, New York, 1977.

Glynn, P.W., "Asymptotic Theory for Nonparametric Confidence Intervals," Technical Report 19, Dept. of Operations Research, Stanford University, 1982.

Glynn, P.W. and D.L. Iglehart, "The Joint Limit Distribution of the Sample Mean and Regenerative Variance Estimator," forthcoming technical report, Dept. of Operations Research, Stanford University.

Heidelberger, P. and D.L. Iglehart, "Comparing Stochastic Systems using Regenerative Simulations with Common Random Numbers," <u>Adv. Appl. Prob.</u>, 11, 804-819, 1979.

Meketon, M.S. and P. Heidelberger, "A Renewal Theoretic Approach to Bias Reduction in Regenerative Simulations," <u>Management Science</u>, 26, 173-181, 1982.

STANDARDIZED TIME SERIES -- Lee W. Schruben

Standardizing a time series is conceptually the same as classical standardization of a scalar statistic. Here the entire series is standardized by scaling the partial sums of deviations about the sample mean. This series, under some mild assumptions, will converge in distribution to a Brownian bridge stochastic process. The theoretical properties of this limiting process are used for statistical inference in the same way that normal or t random variables are used in scalar inference. This permits the testing of hypotheses and the construction of confidence intervals for parameters of the original output series. This technique appears to be a promising approach to many of the problems in simulation output analysis. Papers on this topic are given below.

References

Goldsman, D., Ph.D. Thesis, School of Operations Research and Industrial Engineering, Cornell University, 1984.

Goldsman, D. and L. Schruben, "Comparison of some Confidence Intervals for Simulation Output," to appear in <u>Management Science</u>.

Schruben, L., "Detecting Initialization Bias in Simulation Output," <u>Operations Research</u>, 30, 569-590, 1982.

Schruben, L., "Confidence Interval Estimation Using Standardized Time Series," <u>Operations Research</u>, 31, 1090-1108, 1983.