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

MULTIVARIATE ESTIMATION IN SIMULATION

Andrew F. Seila
Department of Quantitative Business Analysis
College of Business Administration
University of Georgia
Athens, Georgia 30602

Methods for multivariate estimation in regenerative and nonregenerative discrete event simulation are discussed, along with practical considerations in applying these methods.

1. INTRODUCTION

Frequently, the type of information that systems analysts wish to learn about a system under study cannot be summarized in a single performance measure. Instead, the analysis must involve simultaneous computation or estimation of several parameters. When simulation is involved in the analysis, one is therefore confronted with the problem of jointly estimating the values of several performance measures.

Consider, for example, a queueing system which is being simulated in order to estimate the stationary mean waiting time per customer and the server utilization. To test whether both of these performance measures are in an acceptable range of values using data from a single simulation run, they must be estimated jointly. Considering the estimates separately would be erroneous and misleading.

As a second example, consider the same queueing system and the same two performance measures, but now suppose the system is being operated under two different service policies. We wish to determine if the two policies induce a significant difference in the system performance. Again, a test for each parameter separately would not be sufficient. Instead, we must test for a difference in the parameters jointly.

Finally, consider a service system in which there are five classes of customers. Suppose that we wish to estimate the mean waiting time for each class of customer. We may wish to do this by computing confidence intervals for each class' mean waiting time such that all five confidence intervals include their respective true parameter values simultaneously with probability .95. As in the first two examples, this is a problem in multivariate estimation which cannot be solved using a univariate approach.

In this paper, we discuss methods for jointly estimating a vector $\underline{r} = (r_1, r_2, \ldots, r_s)$ of s stationary performance measures using data from a discrete event simulation. The methods are applicable to any performance measure which is the stationary mean of a sequence of observations, in discrete or continuous time, and to any system which has a stationary, or steady-state, distribution. Informally, a stationary system is one in which the statistical properties show a pattern of long-term stability. In Section 2, we will discuss methodology applicable to regenerative systems; in Section 3, methods appropriate for stationary, nonregenerative systems will be presented. Finally, the last section discusses some practical problems with multivariate estimation in simulation.

2. MULTIVARIATE ESTIMATION IN REGENERATIVE SIMULATION

A regenerative simulation has the property that there is a random sequence of regeneration times such that, at each of these times, the future behavior of the system becomes statistically independent of the past, and probabilistically identical to the behavior after any other regeneration time. The most familiar example of a regenerative system is that of certain queueing systems in which the system "starts anew" each time a customer arrives to find all servers idle and all queues empty. The term "starts anew" means that future behavior is independent of the past and identical to the probabilistic behavior after any other empty and idle period.

A cycle is defined to consist of all observations between two consecutive regeneration times. The

important characteristic of regenerative simulations is that the cycles are statistically independent and identical replicas of one another, and all information about the values of stationary performance measures is contained in the properties of each cycle. As a result, the cycle, rather than the individual observation, is the fundamental unit of data in a regenerative simulation.

In our presentation, we will omit some technical details and concentrate on the general concepts and required computations. For a justification and more thorough discussion, see the papers by Seila (1982 and 1983).

Let the observations on s performance measures be represented by $\{(X_{i1}, N_{i1}), (X_{i2}, N_{i2}), \ldots, (X_{is}, N_{is}), i = 1, 2, \ldots, n\}$, where X_{ij} is the sum and N_{ij} is the number of observations on parameter j during cycle i. For simplicity of exposition, we will assume all observations are discrete--time; however, if this is not the case, the X_{ij}'s and N_{ij}'s can be replaced by appropriate integrals. It is well known that $r_j = E(X_{ij})/E(N_{ij})$, where r_j is the value of parameter j, and $E(X_{ij})$ and $E(N_{ij})$ are the means of X_{ij} and N_{ij}, respectively. The estimator for each r_j is

$$\hat{r}_j = \bar{X}_j/\bar{N}_j, \quad j = 1, 2, \ldots, s,$$

where $\bar{X}_j = n^{-1}\sum_{i=1}^{n} X_{ij}$ and $\bar{N}_j = n^{-1}\sum_{i=1}^{n} N_{ij}$. The vector $\underline{\hat{r}} = (\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_s)$ of these estimators is the multivariate estimator for $\underline{r} = (r_1, r_2, \ldots, r_s)$.

In order to compute confidence intervals or test hypotheses, the sampling distribution of $\underline{\hat{r}}$ must be known. When the sample size, n, is large, $\underline{\hat{r}}$ has approximately a multivariate normal distribution with mean $\underline{r}$. For each cycle, define

$$Z_{ij} = X_{ij} - r_j N_{ij}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, s,$$

and let $\sigma_{jk} = E(Z_{ij}Z_{ik})$ for j, k = 1, 2, ..., s. Since $E(Z_{ij}) = 0$, the values of σ_{jk} are the elements of the covariance matrix for the vector $\underline{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{is})$. The covariance matrix for $\underline{\hat{r}}$ is $\underline{V}/n$, where $\underline{V}$ is an s x s matrix whose (j,k)th element is $\sigma_{jk}/(E(N_{ij})E(N_{ik}))$. The values σ_{jk} can be estimated using

$$\hat{\sigma}_{jk} = n^{-1}\left\{\sum_{i=1}^{n} X_{ij}X_{ik} - \hat{r}_j\sum_{i=1}^{n} X_{ik}N_{ij} - \hat{r}_k\sum_{i=1}^{n} X_{ij}N_{ik} + \hat{r}_j\hat{r}_k\sum_{i=1}^{n} N_{ij}N_{ik}\right\}.$$

Define the statistic

$$T^2(\underline{r}) = n\sum_{j=1}^{s}\sum_{k=1}^{s}(\hat{r}_j - r_j)\bar{N}_j W_{jk}\bar{N}_k(\hat{r}_k - r_k),$$

where W_{jk} is the (j,k)th element of the inverse of the matrix $[\hat{\sigma}_{jk}]$. According to the multivariate normal theory, $((n-s+1)/ns)T^2(\underline{r})$ has approximately an F-distribution with s and n-s+1 degrees of freedom in the numerator and denominator, respectively. In particular, an approximate 100(1-α)-percent confidence interval for $\underline{r}$ is given by all vectors $\underline{x}$ for which

$$T^2(\underline{x}) \le (ns/(n-s+1))F_{\alpha;s,n-s+1},$$

where $F_{\alpha;v_1,v_2}$ is the 100(1-α)-percentage point of the F-distribution with v_1 and v_2 degrees of freedom.

One can also compute simultaneous confidence intervals for linear compounds of $\underline{r}$. Suppose that $\underline{\pi}_1, \underline{\pi}_2, \ldots, \underline{\pi}_\ell$ are ℓ vectors, and $\rho_j = \underline{\pi}_j'\underline{r}$ for j = 1, 2, ..., ℓ, are ℓ linear compounds of $\underline{r}$. If we define $\hat{d}_j^2 = \underline{\pi}_j'\hat{\underline{V}}\underline{\pi}_j/n$, where $\hat{\underline{V}}$ is a matrix whose (j,k)th element is $\hat{\sigma}_{jk}/(\bar{N}_j\bar{N}_k)$, then the confidence intervals $[\underline{\pi}_j'\underline{\hat{r}} - h_j, \underline{\pi}_j'\underline{\hat{r}} + h_j]$, j = 1, 2, ..., ℓ, where $h_j = \hat{d}_j^2[(ns/(n-s+1)F_{\alpha;s,n-s+1}]^{1/2}$, will have a simultaneous confidence coefficient exceeding 100(1-α)-percent. In particular, if $\underline{\pi}_j$ is a vector having a 1 in the jth component and zeros elsewhere, then $\underline{\pi}_j'\underline{\hat{r}} = r_j$ and

$$\hat{d}_j^2 = V_{jj} = \frac{\hat{\sigma}_{jj}}{\bar{N}_j^2 n}.$$

Thus, a set of 100(1-α)-percent simultaneous confidence intervals for $r_1, r_2, \ldots, r_s$ is

$$\left[\hat{r}_j - \sqrt{\frac{\hat{\sigma}_{jj}}{\bar{N}^2 n}\left(\frac{ns}{n-s+1}F_{\alpha;s,n-s+1}\right)}, \quad \hat{r}_j + \sqrt{\frac{\hat{\sigma}_{jj}}{\bar{N}^2 n}\left(\frac{ns}{n-s+1}F_{\alpha;s,n-s+1}\right)}\right],$$

for j = 1, 2, ..., s.

Note that $\hat{\sigma}_{jj}/\bar{N}^2 n$ is the estimate of the variance of $\hat{r}_j$ if r_j were being estimated by itself. Another approach to computing simultaneous confidence intervals uses Bonferroni intervals (Law and Kelton, 1982, p. 308).

3. MULTIVARIATE ESTIMATION IN STATIONARY NONREGENERATIVE SIMULATIONS

The structure of regenerative simulations allows the confidence intervals presented in Section 2 to be valid in large samples. One would naturally ask if similar large sample confidence intervals could be computed for systems that are stationary but not necessarily regenerative. It turns out that the answer is "yes."

In nonregenerative systems, collection of observations for computing estimates must be begun only after significant initialization bias has been removed. If the response is multivariate, this is a more difficult problem than with univariate output. Schruben (1981) presents one method for determining an appropriate truncation point for beginning data collection when the output is multivariate.

For the purpose of data analysis, we will group the simulation output into n batches. A batch could be defined by elapsed simulation time, or by the number of observations. To illustrate, consider the example discussed in Section 1 where we want to jointly estimate server utilization and mean waiting time per customer in a queueing system. Let $U_i(t)$ denote the observed portion of servers that are busy at time t during batch i, and W_{ki} denote the kth observed waiting time during batch i. If a batch consists of observations recorded during 100 hours of operation, say, then the ith batch mean for server utilization would be $\overline{U}_i = \int U_i(t)dt/100$, where the integral is taken over the duration of batch i. However, the number of customers to complete service during the ith batch would be a random variable, N_i. Then, the ith batch mean for customer waiting time would be $\overline{W}_i = \Sigma W_{ki}/N_i$, where the sum is taken over all waiting times in batch i. This is a ratio estimator, since the denominator is a random variable. Similarly, if we define a batch in terms of the number of waiting times observed, $\overline{W}_i$ would be an ordinary sample mean, since N_i would be fixed; however, $\overline{U}_i$ would be a ratio estimator because the elapsed time for the ith batch would now be a random variable. Therefore, generally, at least some of the batch means would be ratio estimators.

Let the number of observations on parameter j in batch i be denoted by K_{ij}, and let Y_{ij} denote the sum of observations on parameter j in batch i. If parameter j is observed in continuous time, then K_{ij} would be the length of batch i relative to parameter j, and Y_{ij} would be the integral over batch i of the "level" of parameter j. Define the ith batch ratio to be

$$\overline{Y}_{ij} = Y_{ij}/K_{ij} ,$$

and let $\underline{\overline{Y}}_i = (\overline{Y}_{i1}, \overline{Y}_{i2}, \ldots, \overline{Y}_{is})$. If the batches are sufficiently large, and some technical conditions are met, then the batch ratios, for distinct batches, are approximately uncorrelated. Thus, the simulation run can be represented by a sequence of (approximately) uncorrelated random vectors $(\underline{\overline{Y}}_1, \underline{\overline{Y}}_2, \ldots, \underline{\overline{Y}}_n)$. Standard multivariate estimation techniques can then be applied to this sample for the purpose of statistical inference.

The estimator for $\underline{r} = (r_1, r_2, \ldots, r_s)$ is $\underline{\bar{Y}} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_s)$, where

$$\bar{Y}_j = \frac{1}{n} \sum_{i=1}^{n} \overline{Y}_{ij} .$$

Let $\underline{S}$ be the estimated covariance matrix for $\underline{Y}_i$, whose (j,k)th component is given by

$$S_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (\overline{Y}_{ij} - \bar{Y}_j)(\overline{Y}_{ik} - \bar{Y}_k) .$$

Then, the statistic

$$\frac{n(n-s)}{s(n-1)} (\underline{\bar{Y}} - \underline{r})' \underline{S}^{-1} (\underline{\bar{Y}} - \underline{r})$$

$$= \frac{(n-s)}{s(n-1)} T^{*2}(\underline{r})$$

has approximately an F-distribution with s and n-s degrees of freedom if the number of batches, n, is large. Thus, a $100(1-\alpha)$-percent confidence region for $\underline{r}$ is given by all vectors $\underline{x}$ for which

$$T^{*2}(\underline{x}) \leq \frac{s(n-1)}{n-s} F_{\alpha;s,n-s} .$$

Methods for computing simultaneous confidence intervals can be found in standard textbooks on multivariate statistics, such as (Morrison 1967).

4. DISCUSSION

In this paper, we have presented methods for multivariate statistical inference in regenerative and nonregenerative simulations. These methods are valid only for sufficiently large samples. However, in practice, all samples are finite in size. One would therefore question how effective these procedures are.

Experimentation has shown that the validity of confidence intervals computed using the methodology in this paper depends on the sample size and the model which is generating the observations. Sample sizes required for multivariate confidence regions are generally much greater than those for univariate confidence intervals. Thus, the analyst would be well advised to use conservatively large sample sizes.

A second problem that appears in practice is one of bias. The estimators for both regenerative and nonregenerative simulations are ratios of correlated random variables, which are known to be biased estimators for finite samples. This bias could be reduced by applying the jackknife (Law and Kelton 1982, p. 312); however, this technique

unfortunately increases the variance of the estimator. Fishman (1978) has proposed an estimator for regenerative simulations which has $O(n^{-1})$ bias removed and has the same variance as the unadjusted estimator, to $O(n^{-2})$. Bias reduction in nonregenerative simulations remains an open research problem.

REFERENCES

Fishman, G. S. (1978), Principles of Discrete Event Simulation, John Wiley, New York.

Law, A. M. and Kelton, W. D. (1982), Simulation Modeling and Analysis, McGraw-Hill, New York.

Morrison, D. F. (1967), Multivariate Statistical Methods, McGraw-Hill, New York.

Schruben, L. W. (1981), Control of initialization bias in multivariate simulation response, Communications of the ACM, Vol. 24, pp. 246-252.

Seila, A. F. (1982), Multivariate estimation in regenerative simulation, Operations Research Letters, Vol. 1, No. 4, pp. 153-156.

Seila, A. F. (1983), Multivariate estimation of conditional performance measures in regenerative simulation, Technical Report, Department of Quantitative Business Analysis, College of Business Administration, University of Georgia, Athens, GA, September, 1983.