Interactive Analysis of Simulation Output by the Method of Batch Means

Thomas J. Schriber
Richard W. Andrews

The University of Michigan

ABSTRACT

An interactive FORTRAN subroutine is presented for use with ongoing simulations to determine and collect the sample size needed to estimate the mean of a process with a specified level of statistical precision. The subroutine can be used with simulation models written in a variety of languages, e.g., FORTRAN, GASP, GPSS, SIMSCRIPT. The subroutine partitions a sequence of observations on the random variable of interest into a series of consecutive batches, finding those batch sizes whose batch means are independent. The classical iid method is then applied to build a confidence interval on the mean. Under interactive user control, the subroutine then goes back to the simulation model as often as may be necessary to extend sample size to the point that the confidence interval satisfies the user's needs.

This paper complements an earlier paper presenting software for interactive autoregressive analysis of simulation output [1]. The present paper reports on the use of both techniques to analyze data produced by data models for which analytic results are known. The method of batch means is not successful in identifying the batch size for which the batch means are known to be independent in one of these data sets. This raises serious questions about the procedure used to test for independence of batch means, and points out the need for further research in this area.

I. INTRODUCTION

The purpose of building a stochastic simulation model is to imitate a process which is too complex to model analytically. For example, suppose a manager wants to estimate the hourly cost associated with the in-context use of an expensive piece of materials handling equipment being considered for purchase. This hourly cost consists not just of the uniform capital recovery cost of the equipment, but also of various potential delay costs attributable to the equipment's ability or inability to successfully move material from point to point in timely fashion. Because of the many interdependent elements characteristic of the equipment-usage context, it is impossible to build an analytic model with which to estimate the hourly cost. The alternative is to build a computer-based stochastic simulation model to imitate the process of interest. The overall purpose of such a model is to simulate the hour-by-hour performance of the proposed piece of equipment, and thereby estimate the magnitude of the costs associated with its use.

In general, the output from such a stochastic simulation model consists of a sequence of observations made on one or more dependent random variables whose behavior is of interest to a decision maker. In the materials-handling problem, for example, the important output variable is cost per hour. Throughout this paper, we will be concerned with such a single output variable from a simulation model. The associated random variable will be denoted by X.

We assume that the simulation of interest has been run for a long enough time to eliminate transient responses and bring about a steady state of operation. This means that observations obtained from the simulation are realizations of a stationary stochastic process. Since we want to use the observations to make inferences about the process, it is reasonable to begin by saying a few things about such a stochastic process itself.

We designate the n random variables for which our observations are realizations by X(1), X(2), ... , X(n). Because the stochastic process is assumed to be stationary, these random variables are identically distributed with a common mean, μ, and a common variance, σ^2. This sequence of random variables differs from an independent and identically distributed sequence because it has a non-zero autocovariance function R(s), where for any i,

$$R(s) = \text{Covariance}[X(i),X(i+s)],$$

s = 0,1,2,... The non-zero autocovariance function indicates that the random variables are not independent. For example, in the materials-handling problem, it will very likely be the case that the

CH1437-3/79/0513-0525$00.75

1979 Winter Simulation Conference

cost observed in a given hour will have an effect on the cost observed in one or more following hours.

In this paper, our sole inference objective is to use the observed values of X to make inferences about the process mean, μ. A point estimator of μ is given by $\bar{X}$, where

$$\bar{X} = [X(1) + X(2) + \ldots + X(n)]/n.$$

However, our interest is in obtaining an interval estimate for the mean, not just a point estimate. We might naively compute a confidence interval as...

$$\bar{X} \pm t(\alpha/2;\ n-1)s/n^{1/2}$$

where

$$s^2 = \sum_{i=1}^{n} [X(i) - \bar{X}]^2/(n-1)$$

and $t(\alpha/2; n-1)$ is the t statistic with n-1 degrees of freedom at a $100(1-\alpha)\%$ confidence level. This method of confidence interval construction is based on two assumptions which are usually unfounded for simulation output. One assumption is that the X values are realizations from a normal distribution. This assumption is often not a critical one, however, because the procedure for estimating the confidence interval is capable of giving satisfactory results even when the X's depart from normality. The second assumption is that the X's are independent. This assumption is very critical and, when not satisfied, can result in a distorted confidence interval which is either too wide or too narrow, depending on the autocovariance function R(s).

Some way is needed either to incorporate the autocovariance function into the analysis, or to circumvent its existence. Two important papers addressing this problem have appeared, [2] and [4], both by Fishman. In [4], the correlation structure in simulation output is modeled by finding the order of an autoregressive representation which best fits the data. Then the parameters of the chosen autoregressive model are estimated, and these parameters are used to build the desired confidence interval. This methodology can be described as direct, because the correlation structure of the data is exploited through the use of an autocorrelated model.

In [2], an attempt is made to circumvent the existence of the autocovariance function by dividing simulation output into batches of such size that consecutive batch means are independent of each other. In contrast with the autoregressive approach, the batch-means approach can be described as indirect in that it attempts to work around the correlation structure by finding subsets of the overall data set which have uncorrelated batch means. These independent batch means can then be used to build the desired confidence interval for the mean of the process.

In [1], Andrews and Schriber presented an interactive FORTRAN subroutine implementing the autoregressive methodology presented in [4]. That subroutine, named AUTOR, is easily coupled with a simulation model to automate the autoregressive analysis of the simulation output. And, because of its interactive nature, the AUTOR subroutine can be conveniently used by a decision maker to extend sample size to the point needed to produce a confidence interval adequate to support the decision-maker's needs.

The purpose of the present paper is to present an interactive FORTRAN subroutine implementing a variation of the batch-means methodology presented in [2]. Named BMEAN, the subroutine is analogous to AUTOR in that it is easily used to automate the batch-means analysis of output from a simulation model, and can be conveniently used by a decision maker to produce confidence intervals adequate for the needs at hand.

The next section describes the batch-means model on which the BMEAN subroutine is based. In Section III, the BMEAN subroutine itself is then described, and its use in conjunction with a GPSS model is demonstrated in Section IV. Section V returns to the question of the differences between BMEAN and AUTOR and reports on an experimental investigation of their performance, both in relative and in absolute terms. Finally, conclusions are presented in Section VI.

II. THE BATCH MEANS MODEL

This section explains the underlying statistical model on which the batch-means method for confidence interval construction is based. An illustrative numeric example is developed to support the discussion. As its context, the numeric example uses the materials-handling problem described in Section I. Hence, denoting the sequence of values produced by the stationary stochastic process of interest as X(1), X(2), ... , X(n), it is then possible to interpret X(i) as the dollar cost incurred during the i-th simulated steady-state hour and attributable to the materials-handling equipment being considered.

Suppose that a simulation of the materials-handling system has reached steady state and then has proceeded for another 50 simulated hours, thereby providing 50 identically distributed hourly-cost observations. These 50 hypothesized observations are given in column 2 of Table 1. (These observations were actually generated by a simulation model corresponding to the quantified materials-handling problem to be described in Section IV.) Employing the batch-means method of analysis, we now want to group these 50 observations into batches, with each batch being of identical size, i.e., containing an identical number of observations. Let k denote the number of batches, and m denote the batch size. For example, with 50 observations, we might work with 50 batches of batch size 1; or with 25 batches of batch size 2; or with 16 batches of batch size 3; and so on. (Note that some batching schemes fail to use all of the observations in the overall data set. For example, the 16 batches of batch size 3 would use only the first 48 of the 50 values in Table 1.)

Now for a given choice of batch size m and the cor-

Table 1

Batch Means Corresponding to Various Batching Schemes for 50 Observations from the Materials-Handling Problem

Hour	Cost Observed During Hour	Batch Means for Various Batch Sizes					
		m=1	m=2	m=3	m=4	m=5	m=6
1	20	20					
2	28	28	24				
3	23	23		23.67			
4	35	35	29		26.5		
5	46	46				30.4	
6	33	33	39.5				30.83
7	21	21		38			
8	20	20	20.5				
9	35	35		25.33	30		
10	36	36	35.5			29	
11	32	32					28.67
12	28	28	30	32	32.75		
13	20	20					
14	20	20	20			24.8	
15	24	24		21.33			
16	24	24	24		22		
17	27	27					22.5
18	20	20	23.5	23.66			
19	23	23				26.4	
20	38	38	30.5		27		
21	82	82		47.67			
22	46	46	64				38.5
23	20	20					
24	22	22	21	29.33	42.5	39.2	
25	26	26					
26	22	22	24	22.67			
27	20	20			23		
28	24	24	22			21.4	22.17
29	21	21		21.67			
30	20	20	20.5				
31	33	33			26.5		
32	32	32	32.5	28.67			
33	21	21				26.6	
34	20	20	20.5				26.5
35	27	27			23.5		
36	26	26	26.5	24.33			
37	20	20					
38	37	37	28.5	31.33			
39	37	37			35.5	33.6	
40	48	48	42.5				31.5
41	20	20		31.67			
42	27	27	23.5				
43	29	29			24	23.8	
44	20	20	24.5	24			
45	23	23					
46	23	23	23				34.8
47	20	20		45	39.5		
48	92	92	56			55.4	
49	72	72					
50	70	70	71				

esponding number of batches k, we proceed to compute the batch mean for each of the k batches. Let Y(1), Y(2), ... , Y(k) denote those batch means. For example, in Table 1, with m = 2 and k = 25, there are 25 batch means having the values 24, 29, 39.5, ... , and 71 (4th column in Table 1).

Because we are at steady state, the expected value of each X is the same. We have denoted that value by μ. Also, since each Y is an average of a subset of the X's, the mean of the Y's is also μ. Our objective is to build a confidence interval for μ, the process mean. We can do this quite easily, in concept, as follows: (a) First, determine a batch size (if any) for which the batch means are independent; (b) Then compute the sample variance of

the batch means; (c) Finally, form the confidence interval as

$$\bar{Y} \pm t(\alpha/2;\ k-1)s/k^{1/2}$$

where s^2 is the sample variance of the batch means, and $t(\alpha/2; k-1)$ is the t statistic with k-1 degrees of freedom at a $100(1-\alpha)\%$ confidence level. The two assumptions inherent in the above computational procedure are those described earlier (see Section I). The important thing to note here is that the assumption of independence has been satisfied via (a) above. In addition, the assumption that the Y's are normally distributed becomes increasingly better satisfied as batch size m grows, via the Central Limit Theorem.

In the Table 1 example, we potentially have the choice of building a confidence interval based on six alternative batching schemes, corresponding to m = 1, 2, 3, 4, 5, or 6, with the respective k's being 50, 25, 16, 12, 10, and 8. (Batches of size 7 and larger are excluded from consideration here for a reason to be indicated shortly.) The choice of which particular batching scheme (if any) to use in building a confidence interval for the process mean will be determined by testing the hypothesis, batch-size by batch-size, that the batch means are independent.

The testing procedure used to test the hypothesis of batch-mean independence requires the use of at least 8 batches. This explains why batch sizes larger than 6 aren't possible in the Table 1 example.

For the technical details of the hypothesis test itself, see [2].

It is suggested in [2] that the hypothesis test be used sequentially, starting with a batch size of 1. If the null hypothesis of independence is rejected, [2] suggests doubling the batch size and running the test again. This double-and-test procedure is continued until either the independence hypothesis is accepted (at which point a confidence interval is constructed), or until the number of batches is less than 8 (at which point no confidence interval can be constructed).

We use the same test of hypothesis in our work as described in [2], but our procedure for determining which batching scheme to use in building a confidence interval differs from that in [2]. We test the independence hypothesis for every possible batch size. That is, setting the Type I error at 0.10, we test the following hypothesis structure for every value of m from 1 up to the largest batch size which will provide 8 batches:

H_0: The batch means are independent.

H_a: H_0 is false.

There are then two possible overall results. Either (a) the hypothesis of independence will be rejected for every batching scheme, or (b) some of the batching schemes will be accepted as providing independent batch means. In the (a) case, we recommend to the user that more observations be generated by the simulation model. In the (b) case, we choose that batching scheme for which it is the most probable that the batch means are independent. Since the standardized test statistic has a normal distribution, this means we choose that batching scheme whose test statistic is closest to zero.

Table 2 displays pertinent information which results from applying batch-means analysis to the various feasible batching schemes possible with the 50 Table 1 observations. Inspection of Table 2 shows that in our procedure, we would proceed to build a confidence interval using the batching scheme corresponding to 10 batches of size 5, because the test statistic for this scheme, 0.23, is closer to zero than any of the other batching-scheme test statistics. In contrast, if the procedure in [2] were

Table 2

Selected Results from Analysis of the Table 1 Data Set by the Method of Batch Means

	m=1	m=2	m=3	m=4	m=5	m=6
Value of the Normalized Test Statistic	3.73	1.73	-0.33	-1.15	0.23	-1.22
Confidence Interval Half-Width	*none**	*none**	4.29	4.30	7.14	4.73
Sample Variance of the Batch Means	265.49	190.47	65.00	45.88	99.63	32.02

**These batching schemes failed to produce independent batch means.*

followed and a Type I error of 0.10 were used, the result would be to build a confidence interval using 12 batches of size 4.

Table 2 provides some interesting insights into the relationship between the number of batches and the sample variance of the batch means. It might be thought that as the batch size increases, the sample variance of the batch means would get smaller. This would indeed be the case if the X's were independent. However, in the case of an autocorrelated sequence, the sample variance of the batch means will depend not only on the variance in the underlying population, but also on the autocorrelation structure. This means it is quite possible for the variance of the batch means to increase with increasing batch size. This phenomenon occurs in Table 2, in fact. As batch size increases from 4 to 5, the variance of the batch means increases from 45.88 to 99.63. One possible explanation for this is that the data may harbor autocorrelation of order 5.

It is also interesting to consider how the number of batches, k, enters into the computation of the width of the confidence interval. As k gets larger, the t statistic becomes smaller (leading to a narrower confidence interval), and for a given sample variance, the variance of the sampling distribution of the batch means becomes smaller (as reflected by the appearance of the square root of k in the denominator of the confidence interval expression based on $\bar{Y}$, and given earlier), which also leads to a narrower confidence interval. As Table 2 shows, however, the sample variance in autocorrelated data may not follow a predictable pattern as k changes. It is not necessarily the case, then, that one wants to choose a batch size which maximizes the value of k while satisfying the condition that the batch means are independent. As indicated earlier, we recommend basing the confidence interval on that batching scheme which best supports the test for independent batch means, regardless of the relative value of the associated sample variance.

Our work departs from the methodology suggested in [2] in one other way. For large data sets, it proved impossible to work directly with the batch means themselves when large batch sizes were involved. Trouble occurred because as batch size increases, the variance of the batch means approaches zero. This is true because the variance of the batch means is of the order of 1/m. Letting Y(j) be the j-th batch mean, this fact is evident upon inspection of the expression for the variance of Y(j),

$$\text{Variance}[Y(j)] = \frac{\sigma^2 + 2\sum_{s=1}^{m-1}(1 - s/m)R(s)}{m}$$

where, for any i,

$$R(s) = \text{Covariance}[X(i),X(i+s)],$$

s = 0,1,2,... This problem was handled by using a standardized version of the batch means, denoted by Y* and defined as the square root of the batch size times the batch mean, i.e.,

$$Y^*(j) = m^{1/2}Y(j)$$

This standardized batch mean has the following variance.

$$\text{Variance}[Y^*(j)] = \sigma^2 + 2\sum_{s=1}^{m-1}(1 - s/m)R(s)$$

where R(s) is as defined above. Note that the variance of this standardized batch mean is not of the order of 1/m.

III. THE SUBROUTINE BMEAN

The subroutine BMEAN is designed to carry on a running dialog both with a simulation model, and with an interactive user. The dynamics and some of the options involved in using the subroutine will now be sketched out.

The action starts with the simulation model, which can be thought of as a main program. (The main program discussed in the next section is a GPSS model but it could, of course, take other forms, such as that of a conventional FORTRAN main program). In general, the simulation model must move itself first of all through transient conditions and into a steady state of operation (with the simulated time required to bring this about, if any, having been determined by the modeler through earlier experimentation).

The main program then calls BMEAN which, detecting that this is the first call on it, requests that the user input values for (1) the confidence level which is to apply to the confidence interval on the mean; and (2) the number of observations to be taken initially on the random variable of interest. Control is then passed back and forth between BMEAN and the simulation program the number of times needed to collect this initial sample size. (Each time the simulation model has control returned to it by BMEAN, it proceeds with the simulation until one additional observation has been collected on the random variable, and then passes this value to BMEAN. Experience shows that this approach makes it relatively easy to convert a simulation model not originally developed for BMEAN use to one that can use BMEAN quite naturally.)

BMEAN then performs the batch-means analysis on this initial sample, and reports relevant results, including a confidence interval, to the user. If the user is not satisfied with this confidence interval, s/he can indicate what width interval is desired. BMEAN then reports to the user what the estimated additional sample size required to achieve this confidence interval would be. The user is then given three options: (1) terminate the simulation; or (2) specify the amount by which the existing sample is to be supplemented by the taking of additional observations; or (3) provide a new specification on the desired width of the confidence interval, in response to which BMEAN will report the estimated additional sample size required to achieve this newly specified size of confidence interval.

If the user selected option 2, BMEAN then interacts accordingly with the main program to extend the total number of observations as requested, analyzes and reports out the confidence interval associated with the total sample, and again gives the user the three options indicated above. This back and forth process typically continues until the user chooses to

terminate the simulation, or until BMEAN terminates the simulation because the total sample size has reached 15,000. (BMEAN cannot handle samples consisting of more than 15,000 observations, although this restriction could easily be relaxed by introducing modest changes in BMEAN.)

Another condition which might come about results from the possibility that no batching scheme leads to acceptance of the hypothesis that the batch means are independent. In this case, the user is given the option of terminating the simulation, or of extending the sample size in the hope that one or more of the batching schemes will result in independent batch means for the larger sample size.

Except for the facts pointed out in Section II that BMEAN considers all feasible batching schemes and uses standardized batch means, the subroutine presented here is statistically identical to a subroutine presented in [3], and written in SIMSCRIPT. The subroutine in [3] is, however, set up for batch mode use. One of the obvious advantages of BMEAN is that because it is interactive, the user can conduct an analysis of output relatively quickly and efficiently. This contrasts with a batch mode approach, in which the user might have to engage in a potentially lengthy sequence of experimental runs to determine the number of observations needed to build a confidence interval having the desired or required width. In addition, observations from earlier runs that might be awkward to save and re-use in a batch context are continuously taken into account as the user extends his or her sample size in the interactive mode investigation. Finally, the present implementation puts tested software for analysis of output within easy reach of the casual user of FORTRAN, GASP, and GPSS simulation models. (Copies of the subroutine at the source level will be distributed to interested persons on request. See the end of this paper, however, for some reservations concerning the test for independence of batch means which BMEAN implements.)

IV. AN APPLICATION OF BMEAN

The following problem, which will be used to illustrate the use of BMEAN in context, is a modified version of a problem in [5]. (This same problem was also used to illustrate the use of AUTOR in [1].)

"A certain materials-handling unit is used to transport goods between producing centers in a job shop. Calls for the materials-handling unit to move a load come essentially at random (i.e., according to a Poisson input process) at a mean rate of two per hour. The total time required to move a load has an exponential distribution with an expected time of 15 minutes. The total equivalent uniform hourly cost (capital recovery cost, plus operating cost) for the materials-handling unit is $20. The estimated cost of idle goods (waiting to be moved, or in transit) because of increased in-process inventory is $10 per load per hour. Furthermore, the scheduling of the work at the producing center allows for just 1 hour from the completion of a load at one center to the arrival of that load at the next center. Therefore an additional $5 per load per hour of delay (including transit time) after the first hour is to be charged for lost production. Using simulation, estimate the mean cost per hour and report a 95% confidence interval for this mean hourly cost."

A key decision which must be made in modeling the materials-handling system concerns the approach to follow in making observations on the cost-per-hour random variable of interest. Suppose it is agreed to make an observation at the end of each simulated hour. Two choices are then available for computing the cost incurred during the hour: (1) the cost could be based only on those jobs which have left the system during the hour; or (2) the cost could be based on all jobs which have been (and perhaps still are) in the system at any time during the course of the hour. Of course, total costs incurred in the long run will turn out to be the same either way. The alternative cost-accounting choices outlined above differ only in the timing with respect to which cost information is accumulated. Should a job in the system be viewed as contributing continuously to the value of the cost-per-hour random variable (choice 2 above), or should its total cost contribution be taken into account in one lump sum, at the time the finished job leaves the system (choice 1 above)?

Following either choice, the cost-per-hour random variable has one and the same expected value. Does it then matter which choice is made in deciding how to take readings on the random variable? Well, if one observational method has a smaller variance associated with it than the other, then it is the better method to follow. The reason is that the smaller variance leads to construction of a narrower confidence interval for a given sample size, other things being equal. Now, it is perhaps intuitively clear that costs accumulated continuously (method 2) will show less variation on an hour-to-hour basis than costs which are based only on jobs completed during the hour. For example, if a given job is resident in the system part of hour 1, all of hour 2, and part of hour 3, then by method 2 part of its delay cost will be accounted for in hour 1, another part in hour 2, and the remaining part in hour 3, whereas by method 1, instead of spreading out its delay cost across hours 1, 2, and 3, all of its delay cost will be taken into account in one lump sum in hour 3. This means that hourly costs observed under method 2 will be "smoother" (fluctuate less) than those observed under method 1. A GPSS model which follows method 2 is shown in detail in [1]. Despite its advantages, that model is necessarily considerably more complicated and logically demanding than a GPSS model which follows method 1. For variety, a GPSS model following method 1 is used in the present paper. (Copies of the GPSS model will be distributed to interested persons on request. The model has been liberally documented with comments so that the person who knows GPSS will be able to come quite easily to a complete understanding of its design and operation.)

The beginning portion of a run made using the GPSS model coupled with BMEAN is shown in Appendix A. Information typed in by the user has been underlined (after the fact) in Appendix A to make it readily

distinguishable from information typed out by BMEAN. The following features of the run can be followed by referring to Appendix A:

(1) The user specifies a confidence level of 95%, and an initial sample size of 500. (The three alternative confidence levels which BMEAN supports are 90%, 95%, and 99%. As for initial sample size, the only restriction on it is that it not be less than 50, and not more than 15,000.)

(2) BMEAN reports back a confidence interval on the mean hourly cost of ($27.34, $33.02) and indicates that this case was developed based on 14 batches, with 34 observations per batch. (In coming to this conclusion, BMEAN considered 62 alternative batching schemes, corresponding to batch sizes ranging from 1 through 62. This fact is not evident in the output.)

(3) Dissatisfied with the width of this confidence interval, the user indicates that a confidence interval with a half-width of $1 is desired. BMEAN responds that this will require the taking of an estimated 3,364 additional observations.

(4) Not willing to take this many additional observations at this time, the user then backs off, indicating that a confidence interval with a half-width of $1.5 would be of interest. BMEAN replies that this will require an estimated 1,230 additional observations.

(5) The user decides, conservatively, simply to have an additional 500 observations taken. The underlying thinking is that the initial sample size may have been too small to provide a very accurate estimate of the population variance (and this estimated variance, in turn, is being used by BMEAN to estimate the number of additional observations needed to shrink the width of the confidence interval to a range which satisfies the user.)

(6) Based on what is now a total sample of 1,000 observations, BMEAN reports out a confidence interval of ($28.63, $32.83). Eleven batches of size 90 were used in this analysis.

(7) Noting that the confidence interval still exceeds $3 in width, the user again specifies that the desired half-width is $1.5. BMEAN then recommends that 942 additional observations be taken in an attempt to achieve this desired half-width.

(8) This time the user takes 1,000 more observations, bringing the total number of observations to 2,000. The resulting confidence interval which BMEAN reports is ($29.35, $32.08), based on 47 batches of size 42. This confidence interval meets the $1.5 half-width criterion and so the user, now satisfied, terminates the simulation.

The available analytic solution for the materials-handling problem indicates that the expected hourly cost is $30.68. The confidence interval reported in (8) above does cover the mean, as it would ideally be expected to do 95% of the time.

V. EXPERIMENTAL PERFORMANCE OF BMEAN AND AUTOR

This section reports on an experimental investigation of the performance of BMEAN and AUTOR (or, more precisely, reports on an experimental investigation of the statistical methodologies which these two subroutines implement). The experimental investigation takes place in terms of three alternative data sets produced by three distinct data-generating models.

The first data set consists of independent observations produced from a trivariate normal distribution model. The idea of "independent trivariate" observations would seem to be a contradiction in terms, and requires some explanation. Consider the first, second, and third observations in this set, vs. the fourth, fifth, and sixth observations. Because they come from a trivariate distribution, observation two depends on observation one, and observation three, in turn, depends on observations one and two. Similarly, observation five depends on observation four, and observation six depends on observations four and five, because these observations also come from the trivariate distribution in question. But, by design, observation four does not depend on observation three. Hence, the first set of three observations, although exhibiting dependency within the set, is independent of the second set of three observations, and so on. In spite of the awkwardness of the phrase, then, "independent trivariate data set" will henceforth be used to refer to this particular data set. Precise details of the sampling design used to produce the "independent trivariate" observations are spelled out below.

The second data set consists of autocorrelated observations produced by an autoregressive model of order 2. The third data set consists of hourly cost observations taken from the simulation model for the materials-handling problem introduced in the preceding section.

The purpose of using BMEAN and AUTOR to analyze data produced by these three data-generating models should be obvious. The observations produced by the independent trivariate scheme should be ideally suited for analysis by BMEAN. (In fact, BMEAN should report that batches of size 3 in this data set are independent.) Similarly, the observations generated by the 2nd order autoregressive model should be ideally suited for analysis by AUTOR. (For these data, AUTOR should report that a 2nd order autoregressive model fits the data well.) Finally, the observations generated from the materials-handling problem provide a realistic usage context both for BMEAN and AUTOR. In any event, by using BMEAN and AUTOR to analyze each of these three data sets, their performance can be considered under a variety of circumstances.

Each data set consists of 19,200 observations. These data sets, in turn, are each partitioned into 200 consecutive subsets of data, with each subset containing 96 observations. Each subset of data is analyzed by BMEAN and AUTOR independent of the other 199 subsets in the corresponding data set. This means that each 96-observation data subset can be viewed as a replication, and each overall data set can be thought of as consisting of data produced by replicating the data-generating process 200 times.

Let's now indicate how observations in the three data sets were produced. For the independent trivariate normal, a data set was generated using a mean vector (1000,1000,1000), and variance-covariance matrix...

$$V = \begin{bmatrix} 10,000 & 1,000 & 8,000 \\ 1,000 & 10,000 & 1,000 \\ 8,000 & 1,000 & 10,000 \end{bmatrix}$$

For a given replication, a value of X(1) was first determined by sampling from a normal distribution with mean 1,000 and variance 10,000. As for X(2), it can be shown that under the above trivariate normal specifications, the conditional distribution of X(2) given X(1) is normal with mean equal to 0.1*X(1) + 900, and with variance 9,900. And so a value of X(2) was determined simply by sampling from the normal distribution with those parameters. Finally, it can be shown further that, given X(1) and X(2), X(3) is normally distributed with mean [2X(2) + 79X(1) + 18,000]/99, and variance 356,000/99. Hence, determining a value for X(3) again involves sampling from a normal distribution with the appropriate parameters.

The method used to generate X(1), X(2), and X(3) was then repeated to generate X(4), X(5), and X(6); X(7), X(8), and X(9); and so on, until the overall data set of 19,200 observations had been produced. Note then that X(4), X(5), and X(6) are independent of X(1), X(2), and X(3), and so on for consecutive 3-observation sequences within the overall data set, as was indicated earlier in a qualitative way.

Now let's discuss the 2nd order autoregressive data-generating model. The overall autoregressive data set was developed by generating 19,302 observations; then the first 102 of these were discarded, leaving the 19,200 observations desired. The first 3 observations in the set of 19,302, namely, X(1), X(2), and X(3), were produced by the trivariate normal method just described. It was then assumed that X(4) depends on X(3) and X(2), with X(4) being normally distributed with mean equal to [2X(3) + 79X(2) + 18,000]/99 and variance equal to 356,000/99. Similarly, X(5) depends on X(4) and X(3) in the same way that X(4) depended on X(3) and X(2), and so on. Another way to express this is to say that the autoregressive observations were generated from an autoregressive process of the form:

$$X(i) = [2X(i-1) + 79X(i-2) + 18,000] + \varepsilon$$

where ε is normally distributed with mean 0 and variance 356,000/99, and i = 3,4,5, ... , 19,302.

As for the data set produced by the model simulating the materials-handling system, it was a simple matter to write a FORTRAN subroutine to force the simulation model to produce 19,200 steady-state observations, and then to form 200 consecutive subsets of these observations, each of size 96.

Before reporting the results of using BMEAN and AUTOR to analyze these three overall data sets, it is useful to state what results we expected from the analysis. When analyzing the 200 replications of the independent trivariate observations, BMEAN should conclude for most of the replications that batch means based on batches of size 3 are independent. Similarly, when analyzing the 200 replications of the 2nd order autoregressive observations, AUTOR should conclude for most of these replications that a 2nd order autoregressive representation fits the data. If BMEAN and AUTOR fail to draw such conclusions, then we have reason to question whether the underlying statistical methodologies which they implement (in particular, the hypothesis test for batch mean independence; and the hypothesis test for autoregressive order) are as effective as they should ideally be.

As for what the results might be when BMEAN is used to analyze the autoregressive data, and when AUTOR is used to analyze the independent trivariate data, and finally when both BMEAN and AUTOR are used to analyze the observations coming from the materials-handling simulation model, there is no theoretical basis for speculation.

Tables 3 through 8 summarize the results of the various analyses, each of which was conducted at a 95% confidence level. While inspecting each of these tables, the following questions should be kept in mind.

(1) Is the batch size or the autoregressive order that which might have been expected?
(2) Did the analysis produce a confidence interval?
(3) What is the average half-width of the confidence intervals produced?
(4) Does the confidence interval cover the known mean of the data-generating process?

Table 3 summarizes the results of analyzing the independent trivariate observations with BMEAN. For all 200 replications, a confidence interval was reported out by BMEAN. However, the anticipated batch size of 3 was reported in only 23 of the runs (11.5%). Also, 20 of the runs (10%) reported that the observations were independent. The reported batch sizes fell into 12 categories (corresponding to batch sizes ranging from 1 to 12), with the batch-size frequencies evidencing no particular pattern. We are forced to conclude that the hypothesis test in BMEAN chooses batch size ineffectively, at least in the case of this particular data set.

93% of the confidence intervals produced by BMEAN cover the process mean, which compares favorably with the 95% confidence level at which the analyses were performed. Coverage itself should not be discussed, however, without discussing the matter of confidence interval half-width. In the third column of Table 3, we see that the average half-width produced by BMEAN for the 200 cases was 27.66, and that the average half-widths for the various batch sizes were relatively stable from batch size to batch size, ranging from 20.4 (batch size 1) to 33.36 (batch size 11).

Table 4 summarizes the results of analyzing the autoregressive data with AUTOR. In 198 of the 200 replications, AUTOR determined an order and reported out a confidence interval. In 168 of the 198 cases, the 2nd order model which AUTOR should ideally accept was indeed accepted. When an order other than 2 was chosen, the order was variable over the range from 1 to 25. It is worth noting that the order zero (the order corresponding to uncorrelated observations) was not chosen by AUTOR for any of the 200 cases. The reported confidence intervals covered

Table 3

Selected Results from BMEAN Analysis of an Independent Trivariate Data Set

Batch Size	Number of Occurrences	Number of Intervals Covering the Mean	Cover Percentage	Average Half-Width
1	20	20	100	20.40
2	3	3	100	22.24
3	23	21	91	26.14
4	11	10	91	26.14
5	19	18	95	26.62
6	26	24	92	28.75
7	17	16	94	28.80
8	20	18	90	27.65
9	17	17	100	30.41
10	10	9	90	33.36
11	24	22	92	33.36
12	10	8	80	29.58
ALL	200	186	93	27.66

the true process mean in 183 of the 198 cases (92%). For the 168 2nd order cases, 157 cover the mean (93%). These coverages correspond closely to the expected 95% coverage.

As indicated in the last Table 4 column, average half-width of the confidence intervals reported by AUTOR varies widely, ranging from 11.01 for the single 19th order case to 507 for the single 10th order case. This large variability in half-width is disturbing, and should be investigated. The overall average half-width was 147.17.

Table 5 summarizes the results of analyzing the autoregressive data with BMEAN. In 22 cases out of 200, no batch size was found to produce independent batch means. For these data, there is no basis for expecting that any particular batch size will result. When batch sizes other than 1 were accepted, they tended to be large (9, 10, 11, or 12). This does not seem unreasonable. However, a very disturbing outcome of these analyses is that 42 of them (out of 200) accepted a batch size of 1, implying that the observations are independent (which we know is not true). This provides further evidence (in addition to that in Table 3) that BMEAN does not choose batch size well. In contrast, when these same cases were analyzed by AUTOR (Table 4), none of the analyses concluded that the observations were independent, which speaks well for AUTOR.

In Table 5, 173 of the 178 confidence intervals, or 69% of them, covered the process mean. The greatest coverage occurred when the batch size was largest (39 out of 47, or 83%, when the batch size was 12). In any event, the coverage falls far short of expectation. The average half-width of the 173 confidence intervals was 35.36, which is considerably smaller than the confidence intervals developed when these same observations were analyzed with AUTOR.

Table 6 summarizes the results of analyzing the independent trivariate data with AUTOR. In 4 cases out of 200, no autoregressive order in the range from 0 to 25 was accepted. No particular order was anticipated; however, in 160 of the 200 cases, AUTOR reported that the observations were uncorrelated, which is obviously false. It is worth noting that 15 of the 200 cases reported an order of 2. The reported confidence intervals covered the process mean 170 out of 196 times (87%), with an average half-width of 24.24. However, these average half-widths range from 7.42 to 430.61 in value, depending on the autoregressive order.

Tables 7 and 8 respectively show the BMEAN and AUTOR analyses of observations taken from the materials-handling simulation model. Table 7 indicates that the BMEAN analyses resulted in a coverage of only 76% and produced confidence intervals with a stable half-width of 5.12. It was concluded in only 1 of the 200 cases that the observations were independent. Apart from this, the procedure chose batch sizes in the range from 2 to 12 with approximately uniform frequency.

Table 8 indicates that the AUTOR analyses of the materials-handling observations resulted in a coverage of only 67%, with a widely varying confidence interval half-width averaging 4.46. A potentially disturbing fact appearing in Table 8 is that in 161 cases out of 200, AUTOR reported no autocorrelation in the data (autoregressive order = 0). BMEAN anslysis of the same observations resulted in the no-autocorrelation conclusion in only 1 case, so that BMEAN and AUTOR are at wide variance with respect to each other here.

The coverage rate produced both by BMEAN and AUTOR for the materials-handling simulation is very low, especially for such an uncongested queuing system.

Table 4

Selected Results from AUTOR Analysis of a 2nd Order Autoregressive Data Set

Order Reported	Number of Occurrences	Number of Intervals Covering the Mean	Cover Percentage	Average Half-Width
0	0	--	--	--
1	1	0	0	11.08
2	168	157	93	139.87
3	1	1	100	80.60
4	0	--	--	--
5	0	--	--	--
6	1	1	100	55.99
7	0	--	--	--
8	0	--	--	--
9	0	--	--	--
10	1	1	100	507.00
11	1	1	100	33.69
12	2	2	100	26.75
13	1	1	100	200.89
14	0	--	--	--
15	0	--	--	--
16	1	1	100	11.01
17	1	1	100	79.20
18	0	--	--	--
19	2	2	100	32.59
20	2	1	50	325.32
21	4	3	75	199.50
22	0	--	--	--
23	3	3	100	278.64
24	1	0	0	17.02
25	8	8	100	147.17
ALL	198*	183	92	147.17

Table 5

Selected Results from BMEAN Analysis of a 2nd Order Autoregressive Data Set

Batch Size	Number of Occurrences	Number of Intervals Covering the Mean	Cover Percentage	Average Half-Width
1	42	21	50	19.98
2	0	--	--	--
3	1	1	100	22.81
4	0	--	--	--
5	3	1	33	19.85
6	0	--	--	--
7	6	3	50	26.31
8	7	4	57	27.35
9	16	11	69	34.71
10	16	10	63	33.96
11	40	33	83	43.35
12	47	39	83	46.60
ALL	178*	123	69	35.36

*For 2 replications of the Table 4 data, no autoregressive order was found which provided an acceptable model for the replication.

For 22 replications of the Table 5 data, there were no batching schemes for which the batch means were found to be independent.

Table 6

Selected Results from AUTOR Analysis of an Independent Trivariate Data Set

Order Reported	Number of Occurrences	Number of Intervals Covering the Mean	Cover Percentage	Average Half-Width
0	160	137	86	19.99
1	5	4	80	23.52
2	15	15	100	35.18
3	2	2	100	27.09
4	2	2	100	21.23
5	0	--	--	--
6	1	0	0	14.96
7	0	--	--	--
8	0	--	--	--
9	1	1	100	430.61
10	0	--	--	--
11	0	--	--	--
12	0	--	--	--
13	1	1	100	7.42
14	0	--	--	--
15	0	--	--	--
16	0	--	--	--
17	0	--	--	--
18	0	--	--	--
19	1	0	0	9.54
20	1	1	100	39.81
21	3	3	100	63.55
22	1	1	100	16.04
23	1	1	100	36.95
24	1	1	100	11.08
25	1	1	100	53.55
ALL	196*	170	87	24.24

*For 4 replications of the Table 6 data, no autoregressive order was found which provided an acceptable model for the replication.

Table 7

Selected Results from BMEAN Analysis of a Data Set from the Materials-Handling Problem

Batch Size	Number of Occurrences	Number of Intervals Covering the Mean	Cover Percentage	Average Half-Width
1	1	1	100	3.11
2	22	18	82	3.43
3	18	12	67	3.63
4	17	16	94	4.23
5	15	14	93	4.12
6	11	5	45	4.20
7	17	15	88	6.02
8	20	15	75	4.86
9	22	16	73	6.29
10	15	9	60	4.64
11	21	16	76	6.80
12	21	15	71	6.19
ALL	200	152	76	5.02

Table 8

Selected Results from AUTOR Analysis of a Data Set from the Materials-Handling Problem

Order Reported	Number of Occurrences	Number of Intervals Covering the Mean	Cover Percentage	Average Half-Width
0	161	97	60	2.90
1	30	29	97	8.57
2	3	3	100	20.60
3	1	1	100	59.54
. . .	*No autoregressive order between 4 and 20, inclusive, was found for any replications of the data.*			
21	1	1	100	1.12
22	1	1	100	25.13
23	0	--	--	--
24	0	--	--	--
25	2	2	100	7.57
ALL	199*	134	67	4.46

*For 1 replication of the data, no autoregressive order was found which provided an acceptable model for the replication

VI. CONCLUSIONS

In summary, these experimental investigations of BMEAN and AUTOR's performance strongly suggest that the statistical methodologies which these subroutines implement both have serious deficiencies. Both methodologies readily accept the hypothesis of independence when used to analyze observations known to be autocorrelated. BMEAN does not appear to have a satisfactory way of choosing batch size, and AUTOR reports confidence intervals with half-widths which vary over drastically wide ranges. Further experimental testing of these methodologies against data sets with known properties is required. If the methodologies fail to provide satisfactory analyses in these cases, additional work should be directed toward finding or developing better statistical procedures for testing the key hypotheses which these routines implement (the independence hypothesis in the case of BMEAN, and the autoregressive order hypothesis in the case of AUTOR). We will be pleased to provide copies of our software to others who, like us, are interested in pursuing these issues.

VII. REFERENCES

[1] Andrews, Richard W., and Thomas J. Schriber, "Interactive Analysis of Output from GPSS-Based Simulations," Proceedings of the 1978 Winter Simulation Conference (Association for Computing Machinery, New York, 1978), pp. 267-278.

[2] Fishman, George S., "Grouping Observations in Digital Simulation," Management Science, Vol. 24, No. 5, 1978, pp. 510-521.

[3] Fishman, George S., Principles of Discrete Event Simulation, Wiley-Interscience, 1978.

[4] Fishman, George S., "Estimating Sample Size in Computing Simulation Experiments," Management Science, Vol. 18, No. 1, 1971, pp. 21-38.

[5] Hillier, Frederick S., and Gerald J. Lieberman, Introduction to Operations Research, 2nd Edition, Holden-Day, 1974.

Appendix A

An Example of Interactive Use of the BMEAN Subroutine with a Simulation Model for the Materials-Handling Problem

```
#EXECUTION BEGINS

***ALWAYS USE A DECIMAL POINT WHEN
   ENTERING VALUES FOR THIS PROGRAM***

ENTER THE PERCENT CONFIDENCE LEVEL AS
99., 95., OR 90.
?
95.

ENTER THE INITIAL SAMPLE SIZE
(50. <= ENTRY <= 15000.)
?
500.

   BATCH MEAN ANALYSIS PROGRAM INITIALIZED
   ======================================
                CONFIDENCE LEVEL =    95%
  INITIAL NUMBER OF OBSERVATIONS =   500


      STATISTICAL RESULTS OF SIMULATION
      =================================

  SAMPLE MEAN +/- HALF-WIDTH =      30.1806 +/-

        CONFIDENCE INTERVAL = (    27.3403 ,

                SAMPLE SIZE =       500

             NO. OF BATCHES =        14

     OBSERVATIONS PER BATCH =        34

IS THIS CONFIDENCE INTERVAL SATISFACTORY?
  IF YES, TYPE 2.0 TO TERMINATE THE RUN;
  IF NO, TYPE 3.0 TO SPECIFY A NEW HALF-WIDTH
?
3.

ENTER THE DESIRED HALF-WIDTH
OF THE CONFIDENCE INTERVAL
?
1.

YOUR CONFIDENCE INTERVAL HALF-WIDTH REQUIRES
APPROXIMATELY       3364 ADDITIONAL OBSERVATIONS.
AT THIS STAGE YOU HAVE THREE OPTIONS:
  (1) TYPE 1.0 TO TAKE MORE OBSERVATIONS; OR
  (2) TYPE 2.0 TO TERMINATE THE RUN; OR
  (3) TYPE 3.0 TO SPECIFY A (NEW) HALF-WIDTH
?
3.

ENTER THE DESIRED HALF-WIDTH
OF THE CONFIDENCE INTERVAL
?
1.5

YOUR CONFIDENCE INTERVAL HALF-WIDTH REQUIRES
APPROXIMATELY       1230 ADDITIONAL OBSERVATIONS.
AT THIS STAGE YOU HAVE THREE OPTIONS:
  (1) TYPE 1.0 TO TAKE MORE OBSERVATIONS; OR
  (2) TYPE 2.0 TO TERMINATE THE RUN; OR
  (3) TYPE 3.0 TO SPECIFY A (NEW) HALF-WIDTH
?
1.

ENTER NUMBER OF ADDITIONAL OBSERVATIONS
?
500.
```

Because of space limitations, the values of the half-widths and of the upper confidence points have been trimmed from this sample session in the three places where STATISTICAL RESULTS OF SIMULATION are shown. For example, under the STATISTICAL RESULTS OF SIMULATION appearing immediately below, the pre-trimmed session appeared as follows:

SAMPLE MEAN +/- HALF-WIDTH = 30.7272 +/- 2.0956
CONFIDENCE INTERVAL = (28.6317 , 32.8228)

```
      STATISTICAL RESULTS OF SIMULATION
      =================================

  SAMPLE MEAN +/- HALF-WIDTH =      30.7272 +/-

        CONFIDENCE INTERVAL = (    28.6317 ,

                SAMPLE SIZE =       1000

             NO. OF BATCHES =         11

     OBSERVATIONS PER BATCH =         90

IS THIS CONFIDENCE INTERVAL SATISFACTORY?
  IF YES, TYPE 2.0 TO TERMINATE THE RUN;
  IF NO, TYPE 3.0 TO SPECIFY A NEW HALF-WIDTH
?
3.

ENTER THE DESIRED HALF-WIDTH
OF THE CONFIDENCE INTERVAL
?
1.5

YOUR CONFIDENCE INTERVAL HALF-WIDTH REQUIRES
APPROXIMATELY        942 ADDITIONAL OBSERVATIONS.
AT THIS STAGE YOU HAVE THREE OPTIONS:
  (1) TYPE 1.0 TO TAKE MORE OBSERVATIONS; OR
  (2) TYPE 2.0 TO TERMINATE THE RUN; OR
  (3) TYPE 3.0 TO SPECIFY A (NEW) HALF-WIDTH
?
1.

ENTER NUMBER OF ADDITIONAL OBSERVATIONS
?
1000.

      STATISTICAL RESULTS OF SIMULATION
      =================================

  SAMPLE MEAN +/- HALF-WIDTH =      30.7197 +/-

        CONFIDENCE INTERVAL = (    29.3564 ,

                SAMPLE SIZE =       2000

             NO. OF BATCHES =         47

     OBSERVATIONS PER BATCH =         42

IS THIS CONFIDENCE INTERVAL SATISFACTORY?
  IF YES, TYPE 2.0 TO TERMINATE THE RUN;
  IF NO, TYPE 3.0 TO SPECIFY A NEW HALF-WIDTH
?
2.

***SIMULATION TERMINATED***
```

DR. STEVEN E. SHLADOVER has been a Staff Engineer, specializing in Transportation Systems, at Systems Control, Inc. (Vt) in Palo Alto, California, since October 1978. His work spans a broad range of system analysis disciplines, including automatic control systems, dynamic system modeling, operations research, economics, urban studies, and transportation systems analysis.

Dr. Shladover received the SB and SM degrees in Mechanical Engineering from the Massachusetts Institute of Technology in February 1974, as part of the Mechanical Engineering Honour Course, and completed all the requirements for the Sc.D. degree in Mechanical Engineering in September 1978 (also satisfying all the course requirements for a doctoral degree in Transportation Systems from the Dept. of Civil Engineering). Between February and September 1978, he served as both Lecturer and Research Assistant in the M.I.T. Dept. of Mechanical Engineering. He was co-principal investigator of the research project which served as his doctoral thesis, a multidisciplinary study of the operation of automated guideway transit (AGT) vehicles in dynamically reconfigured trains and platoons. The paper Dr. Shladover is presenting at this conference describes one part of that research project.

Also at M.I.T., Dr. Shladover was the technical coordinator of a separate research project on the lateral control of AGT vehicles in collaboration with two professors, and had responsibility for overseeing the work of two graduate students. He also collaborated with two other students on the development and calibration of an M/G/1 queuing model of the operation of gasoline service stations during the 1973-74 petroleum shortage. This model was used to evaluate the impact on wait times and queue lengths of the various policy options which were used and suggested for use to manage the shortage.

Dr. Shladover's work experience prior to joining SCI, during the years 1968-1974, included summer employment at the U.S. Department of Transportation, Transportation Systems Center (TSC) and Urban Mass Transportation Administration (UMTA), and at Grumman Aerospace Corp. He was a National Science Foundation Graduate Fellow from 1972 to 1975, and is a member of the Pi Tau Sigma, Tau Beta Pi, and Sigma Xi honorary societies. His professional society memberships include American Society of Mechanical Engineers (ASME), Transportation Research Board (TRB), and Advanced Transit Association (ATRA).

He has presented numerous papers at technical conferences in the United States and Europe, he gave guest lectures on automated guideway systems by invitation at the University of Tokyo, and he has published works in several discipline areas.

Dr. Shladover's publications have appeared in Transportation Research, High Speed Ground Transportation Journal, Computers and Operations Research, and the ASME Transactions, Journal of Dynamic Systems, Measurement and Control.

Dr. Steven E. Shladover
Systems Control, Inc. (Vt)
1801 Page Mill Road
Palo Alto, California 94304
(415) 494-1165, Ext. 186