STATISTICAL METHODS FOR IMPROVING SIMULATION EFFICIENCY

Donald P. Gaver, Jr.

Management Sciences Research Group
Graduate School of Industrial Administration
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Westinghouse Research Laboratories

Abstract

This paper describes a variety of statistical devices for improving the effectiveness of computer simulations of random processes. The methods are illustrated by examples from a queueing problem that is inadequately treated by conventional analytical theory.

1. INTRODUCTION

In a great many situations encountered in systems engineering and management science one is confronted with a problem that, figuratively speaking, falls into some familiar category of models, and yet is too complex for the mathematical techniques thereof. The queueing problems encountered in job-shop studies and computer system and traffic (road, air, elevator) analyses are frequently in this category. So also are problems connected with the completion time of certain projects involving inter-related activities, as in PERT networks, where these activities have uncertain duration. There are undoubtedly other examples, but these will suffice.

After an initial attempt to develop information about such problems by use of strictly analytical or mathematical methods, the analyst is typically forced to adopt computer simulation techniques for their treatment. When this is done it is also common to drop further analytical investigations, and merely to use the results of the simulations. I wish to argue in this paper for an intermediate, synthetic, approach: one that incorporates both "oversimplified", "unrealistic" analytically obtained results for simple models, and the output or observations from a simulation experiment. The evidence for the possible value of such techniques will be presented in terms of certain specific examples, and hence will be largely empirical.

2. SIMULATION, AND MONTE CARLO METHODS

A brief statement of the simulation procedure is as follows. We are concerned with a system, and a particular response variable, W, which is influenced by

several other variables, X, Y, ...; we denote these collectively by $\underline{X}$. For example, W might be the waiting time of an aircraft at an airport, and then $\underline{X}$ represents the interarrival times of the planes appearing previously on that day (or portion thereof), their runway occupancy times, etc. The modelling process involves relating W to $\underline{X}$; the latter are often taken to be random variables. We then investigate the distribution of W in terms of that of $\underline{X}$ and are interested in figures of merit such as the expected value of W, the probability that W will exceed some value, etc. That is, we seek to find characteristics of the probability distribution of

$$W = f(\underline{X}) \qquad (2.1)$$

where $f(\cdot)$ is presumed known, but is usually a complicated function. We now outline briefly some procedures for studying the distribution of W, or certain of its characteristics such as its mean, $E[W]$

(a) Straightforward Sampling

In order to obtain a sample value of W, we first obtain a sample value of $\underline{X}$, and then compute W by means of (2.1). More specifically, $\underline{X} = (X, Y, \ldots)$ may be found by first selecting a vector of pseudo random numbers, and converting these to realizations or samples of X, Y, ... utilizing the probability integral transformation or an equivalent, i.e. solving

$$X = F_X^{-1}(R) \qquad (2.2)$$

$F_X(\cdot)$ is the distribution function of X, and R represents a random number, uniformly distributed over (0, 1). A set of k independent realizations of W being at hand, denoted by $\left\{ W^{(i)} \;\; i = 1, 2, \ldots, k \right\}$, the latter may then be averaged to obtain an unbiased estimator of $E[W]$:

$$\widehat{E[W]} = \frac{1}{k} \sum_{i=1}^{k} W^{(i)} \qquad (2.3)$$

having variance

$$\mathrm{Var}\left\{ \widehat{E[W]} \right\} = \frac{1}{k} \mathrm{Var}[W]. \qquad (2.4)$$

Apparently the estimate can be brought closer to $E[W]$ by increasing k, the number of realizations. However, the latter brute force approach is apt to be expensive in terms of computer time, and alternatives are worth exploring. A number of these are familiar, and will be briefly reviewed.

(b) Antithetic Variables

Frequently the response, W, is consistently positively or associated with one or more of the input variables, say X. That is, if inter-arrival times in a queue realization are larger than normal, the waiting times will be smaller than normal, etc. Suppose then that the random number R creates a realization X and X is "large"; then 1-R will tend to create a corresponding realization X' that is "small". At small cost in programming one can then generate antithetic realizations $W^{(i)}$ and $W'^{(i)}$, where the latter are formed from the antithetic samples $\underline{X}^{(i)}$ and $\underline{X}'^{(i)}$, in turn the result of $\underline{R}$ and $1-\underline{R}$. The latter antithetic realizations are then averaged to obtain the final estimate. Permutations of the random numbers among the components of $\underline{X}$ to obtain further antithetic realizations is also profitable occasionally. E. S. Page [6] describes the use of this technique in queueing problems. The writer and J. Burt have experimented with it, and found it to be effective in queueing and PERT network simulations.

(c) Stratification

This method in a sense extends the antithetic idea. Again in brief, we can segment the unit interval over which R ranges into, say, three equal parts:

$r_1 = (0, \frac{1}{3})$, $r_2 = (\frac{1}{3}, \frac{2}{3})$, $r_3 = (\frac{2}{3}, 1)$. We then select a subrange, r_i at random, using one random number. Within r_i, a value for R(1) is selected in accordance with a random number uniform over $(0, \frac{1}{3})$; here R(1) denotes the random number that generates X(1), a variable associated with Realization Number One. To obtain X(2), the corresponding variable for Realization Number Two, it is only necessary to add $\frac{1}{3}$ to R(1), thereby obtaining R(2),-- possibly a subtraction of unity will be required to locate

R(2) in the range(0, 1). From R(2) X(2) results. Another addition of $\frac{1}{3}$ to R(2), together with a subtraction of unity of necessary, generates R(3), and hence X(3) for Realization Number Three. Note that this stratification procedure may be carried out for each variable in $\underline{X}$, and that two independent random numbers, in the case above, generate three parallel realizations. Actually, six realizations can be generated by the above procedure provided the second uniform random number, over $(0, \frac{1}{3})$ is treated antithetically inside that interval.

It can be seen that stratification tends to force an equal distribution of X across companion realizations, and hence a negative correlation of the corresponding values $W^{(j)}(1)$, $W^{(j)}(2)$, and $W^{(j)}(3)$. The average

$$\widehat{E[W]} = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{W^{(j)}(1) + W^{(j)}(2) + W^{(j)}(3)}{3}\right]$$

$$= \frac{\overline{W}(1)+\overline{W}(2)+\overline{W}(3)}{3} \qquad (2.5)$$

thus tends to have a variance smaller than that obtained from 3k independent realizations.

3. CONTROL AND CONCOMITANT VARIABLES THE USE OF APPROXIMATE MODELS

The techniques just described are useful for reducing the sampling variability of simulations, but they fail to employ extra information that may exist concerning the approximate behavior of a system. I will now describe several procedures that involve the simultaneous use of simulation with approximate models and concomitant information, and will illustrate them by means of relatively simple examples.

(a) Control Variates

A classical and useful estimating procedure that involves the use of an approximate model operates as follows. We desire to estimate E [W], where W is related to $\underline{X}$ by (2.1). We are able to calculate (analytically or numerically) the expectation of W* relatively easily; W* is the variable of a model approximating that giving W. We might have either

$$W^* = f^*(\underline{X}) \qquad (3.1)$$

or

$$W^* = f(\underline{X}^*)$$

or even $$W^* = *(X^*)\,; \qquad (3.2)$$

an asterisk will generally be used to denote an approximation.

The important intuitive requirement is that the distributions of W and W* be similar. We then simulate both W and W*, utilizing the same random numbers $\underline{R}$. That is, comparing (3.1) and (2.1), the values of $\underline{X}$ are identical accross realizations to as great a degree as possible. This implies that W and W* will be correlated. We now estimate E [W] as follows

$$\widetilde{E[W]} = E[W^*] + \frac{1}{k}\sum_{j=1}^{k} W^{(j)} - \frac{1}{k}\sum_{j=1}^{k} W^{*(j)} \qquad (3.4)$$

$$= E[W^*] + \overline{W} - \overline{W^*}$$

If expectations are taken it is seen that

$$E\{\widetilde{E[W]}\} = E[W^*] + E[\overline{W}] - E[W^*] = E[W] \qquad (3.5)$$

and so the estimate is unbiassed. Owing to the built-in correlation between W and W* we have (3.6)

$$\mathrm{Var}\{\widetilde{E[W]}\} = \frac{1}{k}\{\mathrm{Var}[W] + \mathrm{Var}[W^*] - 2\,\mathrm{cov}[W, W^*]\}$$

Consequently, if the quantity in brackets on the right-hand side is smaller than Var [W] then an improvement has been achieved over straightforward simulation. This is equivalent to requiring that the control variable, W* exhibit the property $\frac{\mathrm{cov}[W, W^*]}{\mathrm{var}[W^*]} > \frac{1}{2}$. (3.7)

It will, of course, not always be easy to see that (3.7) is satisfied in advance. However, if the results of several realizations are available one can simply compare the empirically determined variances of a straightforward and a control variate estimate to assess the contribution of the latter.

(b) Control and Regression

The form of (3.7) suggests another possibility for improving precision, namely that of a correction of the form

$$\widetilde{E[W]}_r = \overline{W} + \beta(\overline{W^*} - E[W^*]) \tag{3.8}$$

where β is selected to minimize the variance of the estimate $\widetilde{E[W]}_r$. Since (3.9)

$$\mathrm{Var}\{\widetilde{E[W]}_r\} = \mathrm{Var}[\overline{W}] + 2\beta\ \mathrm{cov}[\overline{W},\ \overline{W^*} - E[W^*]] + \beta^2 \mathrm{Var}[\overline{W^*}] ,$$

simple differentiation and straightforward simplification yields for the optimum β the value (3.10)

$$\beta_0 = \frac{\mathrm{cov}[W,W^*]}{\mathrm{Var}[W^*]} = -\ \mathrm{correlation}\ [W,W^*]\sqrt{\frac{\mathrm{Var}[W]}{\mathrm{Var}[W^*]}}$$

If this value of β is utilized, the resulting optimal regression adjusted estimate has variance (3.11)

$$\mathrm{Var}\{\widetilde{E[W]}_{r,0}\} = \frac{1}{k}\left\{\mathrm{Var}[W] - \frac{\mathrm{cov}^2[W,W^*]}{\mathrm{Var}[W^*]}\right\} = \frac{1}{k}\mathrm{Var}[W]\{1 - (\mathrm{correlation}[W,W^*])^2\}.$$

Notice that the variance of this estimate is always at least as small as $\mathrm{Var}[\overline{W}]$, and hence will, in theory, always be an improvement over simple estimates, while ordinary control variates need not have this property; see (3.7). On a practical note, however, we remark that the required covariance will not be known, and hence must be estimated from data. Since the control variable model has been chosen for its analytical tractability, $\mathrm{Var}[W^*]$ is presumably known. We are led to the use of the estimated optimal β

$$\hat{\beta}_0 = -\frac{\frac{1}{k}\sum_{j=1}^{k}(W^{(j)} - \overline{W})(W^{*(j)} - E[W^*])}{\mathrm{Var}[W^*]} , \tag{3.12}$$

and it is now clear that the realistic estimate is

$$\widetilde{E[W]}_{r,\hat{0}} = \overline{W} + \hat{\beta}_0(\overline{W^*} - E[W^*]) \tag{3.13}$$

which is no longer unbiassed, although the bias decreases as the sample size, k, increases. In order to decrease the bias for finite k an application of the Quenouille-Tukey "jacknife" method is perhaps worth a try. The latter may actually reduce the mean-squared error of the estimate. There is, of course, no need to restrict attention to linear corrections; estimates of the form (3.14)

$$\widetilde{\widetilde{E[W]}}_r = \overline{W} + \alpha + \beta_1(\overline{W^*} - E[W^*]) + \beta_2(\overline{W^*} - E[W^*])^2$$

may also be worth investigation.

(c) Concomitant Variables

Suppose that realizations of the random variables $\underline{X}$ are used to create realizations of W, where

$$W^{(i)} = f(\underline{X}^{(i)}) . \tag{3.15}$$

Quite commonly $W^{(i)}$ and $X_i^{(i)}$ are monotonically related, and we can put

$$\mathrm{cov}[W^{(j)},\ X_i^{(j)}] = c_i \tag{3.16}$$

where c_i is either positive or negative. Furthermore, we actually know $E[X_i^{(j)}] = E[X_i]$, since $\underline{X}$ is a given specified input. Since sampling only k times will naturally mean that the realized X-values deviate from their means, then a linear correction to the simple average suggests itself:

$$E[W]_c = \overline{W} + \sum_{i=1}^{I} \gamma_i\ (\overline{X}_i - E[X_i]) . \tag{3.17}$$

Again γ_i can be estimated in terms of the covariance of W and X_i, and the resulting estimate is unbiassed and consistent (tends in probability to $E[W]$) asympototically as k, the sample size, increases. There is no restriction to a linear correction.

4. QUEUEING EXAMPLES

The methods just described are well illustrated by consideration of a very simple queueing problem. It is well-known that the waiting time, W_n, of the n-th arrival to a single-server facility may be written as

$$W_n = \max[W_{n-1} - A_n + S_{n-1}, 0] \qquad (4.1)$$

where A_n is the inter-arrival period elapsing between the $n-1^{st}$ and n-th addition to the queue (or entrance to the server), and S_n is the service time of the n-th customer. If $\{A_n\}$ and $\{S_n\}$ are mutually independent sequences of independent and identically distributed random variables with $E[A_n] = E[A] > E[S_n] = E[S]$, and if other moments exist as required, then a stationary or "steady-state" distribution for W_n exists as $n \longrightarrow \infty$. This distribution can sometimes be described in simple analytical form but exact formulas are formidable in most cases. In the "heavy traffic" situation, when $E[A]$ is only a little larger than $E[S]$ and queues tend to be large, then neat approximations based on diffusion equation solutions are available: see Gaver [3], and Newell [5]. When $E[A] < E[S]$ the queue tends to grow, and little information is available. Intuition suggests that in the latter case the annoying boundary necessitating the "max" in (4.1) is eventually of no importance, and the distribution of W_n approaches that of W_n^*, where

$$W_n^* = W_{n-1}^* - A_n + S_{n-1} \qquad (4.2)$$

as becomes large. Of course the latter is approximately normal (if A_n and S_n have finite variances) and so we feel that W_n is approximately normal as n increases, with mean

$$E[W_n] \approx (n-1)(E[S] - E[A]) \qquad (4.3)$$

provided that $W_1 = 0$. But a more refined analysis indicates that the adequacy of (4.3) depends on the variances of A and S, and it may be desirable to estimate $E[W_n]$ by simulation for small to moderate n, both when

$E[A] < E[S]$ -- the over-saturated case--

and in the steady-srate case when $E[A] > E[S]$.

We shall display the effects of applying various of the variance-reduction methods described to estimate $E[W_n]$ for selected values of n. In particular we focus on the control and concomitant variables approaches. For a control variable it is quite natural to select the simple boundary-less random walk W_n^* when $E[A] < E[S]$. For a concomitant variable it is tempting to select

$$\sum_{i=1}^{n} A_i, \qquad \text{and also} \qquad \sum_{\ell=1}^{n-1} S_\ell,$$

for intuitively speaking an increase in the former is associated with a small W_n, while an increase in the latter induces an increase in W_n. Also,

$$E[A_n] = nE[A] \quad \text{and} \quad E[S_{n-1}] = (n-1)E[S].$$

Numerical examples for the following cases will now be presented. In each of these, service times are taken to be exponentially distributed with mean μ^{-1}:

$$P\{S_n \le x\} = 1 - e^{-\mu x} \qquad x \ge 0$$

$$= 0 \qquad x < 0$$

The inter-arrival times are taken to be either constant (regular arrivals), or exponentially distributed; in each case the mean is unity.

5. DISCUSSION

Tables 1 and 2 illustrate the value of some of the proposed estimating procedures. The service system selected for study is quite simple: customers arrive (regularly in Table 1, in Poisson fashion in Table 2) at a single server, where their service times are exponential with mean 10/9. Even so, the transient response of such a system is not easily characterized mathematically, and so simulation suggests itself. An alternative is the diffusion approximation; see [3] and [5].

Rows (1) and (2) of the tables show the results obtained if 25 independent realizations are averaged to obtain $\bar{W}_n$ as an estimate of $E[W_n]$. Then using the same random numbers we simulated the process again antithetically and averaged to obtain the antithetic estimate $\bar{W}_n(a)$. Comparison of the variances in rows (2) and (4)

indicates that the antithetic device produces an improvement even after the labor of simulating a total of 50 realizations is taken into account. The improvement is smaller for Table 2 than for Table 1, because of the added variability contributed by the random arrivals.

Next the simple control variable device is applied; see rows (6) and (7). The control is the boundary-free random walk.

According to the variances computed, this control estimate seems to perform somewhat better than the antithetic estimate for large n (customer numbers), and less well for intermediate n, although the small differences observed may be due to sampling errors. Certainly one is led to explore further the comparative values of "antithetic" and "control" as inherent process variability builds up: control seems better than antithetic in Table 2 than in Table 1. A combination of antithetic control might be profitable.

Rows (8) and (9) display the effect of adjusting a straightforward estimate (see row (1)) in accordance with the concomitant variable that equals the sum of the first n service times in Table 1; Table 2 considers both service and arrivals times as concomitant variables. The latter device behaves in a manner comparable to antithetics and to control. Rows (10) and (11) exhibit the results of applying concomitant variables to the components of the antithetic estimate of (3) and (4). This adjusted estimate seems to be more effective than the others for the present problems. Rows (12) and (13) indicate the value of a regression-adjusted control procedure; one should compare the variances of (7) with those of (13).

A diffusion approximation to $E[W_n]$, valid for large n, involves adding a constant term to $E[W_n^*]$; see [3]. The result in this case is

$$E[W_{n,\text{diff.}}] \sim n(E[S]-E[A]) + \frac{\text{Var}[S]+\text{Var}[A]}{2(E[S]-E[A])} . \qquad (5.1)$$

If arrivals are regular (Table 1) then Var[A] = 0, and the constant correction equals

$$\frac{\text{Var}[S]}{2(E[S]-E[A])} = \frac{(\frac{10}{9})^2}{2(\frac{10}{9}-1)} = 5.6: \qquad (5.2)$$

while for Table 2 the constant equals 10. The diffusion approximation is tabulated in row (12). Its values agree quite closely with the control variables (row (6)) and regression-adjusted estimates (rows (8) and (10)) for large n.

The methods reported here are also of use in improving the efficiency of simulation studies of stochastic networks of the PERT/CPM type. Certain simple networks may be analyzed analytically, provided node-to-node (link) times are taken to be exponentially distributed. These simple networks may be used to supply control variables for actual networks that do not have exponential links. Some numerical illustrations of this procedure are available, and more are in the process of construction.

REFERENCES

(1) Cox, D. R., and Smith, W. L., Queues; London, Methuen and Co., Ltd., 1961

(2) Ehrenfeld, S., and Ben-Tuvia, S.; The efficiency of statistical simulation procedures; TECHNOMETRICS, Vol. 4, No. 2, May 1962; pp. 257-275.

(3) Gaver, D. P., Diffusion approximations and models for certain congestion problems; J. of Applied Probability, 5, 1968; pp. 697-623.

(4) Hammersley, J., and Handscomb, D., Monte Carlo Methods; London, Methuen and Co. Ltd; 1964.

(5) Newell, G. F. Queues with time-dependent arrival rates (I, II, III), J. of Applied Probability, 5, 1968; pp. 436-451, 579-590, 591-606.

(6) Page, E. S., On Monte Carlo methods in congestion problems: II; Operations Research, Vol. 13, No. 2, March-April, 1965; pp. 300-305.

Table 1. Estimated Wait of n-th Customer

(Regular Arrivals, Exponential Service at Single Server)

Based on 25 Realizations

$A_n = 1$, $ES_n = 1.111$

			n: 5	10	25	50	100	150	200
Straightforward:	(1)	Mean	1.19	1.87	4.51	7.85	11.43	17.04	23.60
	(2)	Variance	.124	.128	.605	1.356	2.758	6.795	10.678
Antithetic:	(3)	Mean	1.35	2.51	5.11	8.67	14.50	19.58	24.88
	(4)	Variance	.054	.141	.256	.367	.622	1.370	1.770
Random Walk:	(5)	Mean	.44	.99	2.67	5.44	11.00	16.55	22.11
Control: (β=1)	(6)	Mean	1.48	2.91	5.92	9.86	16.01	22.12	28.00
	(7)	Variance	.039	.102	.373	.763	.963	1.296	1.427
Straightforward & regression:	(8)	Mean	1.42	2.27	5.49	9.48	14.62	22.50	29.08
	(9)	Variance	.019	.042	.144	.366	.446	1.715	3.864
Antithetic & regression:	(10)	Mean	1.38	2.10	5.16	8.78	15.01	21.72	27.46
	(11)	Variance	.014	.106	.165	.093	.093	.671	1.350
Control:	(12)	Mean	1.39	2.43	5.33	9.03	14.54	20.94	27.13
(β estimated)	(13)	Variance	.016	.034	.126	.198	.445	.733	.832
Diffusion (Asymptotic; (5)+Constant:	(14)	Mean	5.9	6.5	8.2	10.9	16.5	22.1	27.6

Table 2. Estimated Expected Wait of n-th Customer
(Poisson Arrivals, Exponential Service at Single Server)
Based on 25 Realizations

$EA_n = 1 \quad ES_n = 1.111$

			n: 5	10	25	50	100	150	200
Straightforward:	(1)	Mean	2.15	2.77	6.92	11.93	19.84	25.65	29.28
	(2)	Variance	.225	.295	1.119	1.832	2.692	6.596	11.446
Antithetic:	(3)	Mean	1.95	3.33	6.62	11.22	17.30	24.33	29.55
	(4)	Variance	.062	.107	.271	.558	1.079	1.964	3.239
Random Walk:	(5)	Mean	.44	.99	2.67	5.44	11.00	16.55	22.11
Control:	(6)	Mean	1.63	3.05	6.13	9.81	15.70	21.55	27.23
(β =1)	(7)	Variance	.125	.225	.477	1.026	1.278	1.396	1.393
Straightforward & regression:	(8)	Mean	1.69	2.86	6.04	10.79	17.93	22.46	27.64
	(9)	Variance	.148	.087	.162	.443	1.201	1.565	2.307
Antithetic & regression:	(10)	Mean	1.73	3.22	6.15	10.83	17.61	24.09	29.57
	(11)	Variance	.025	.054	.074	.216	.368	1.016	.974
Control:	(12)	Mean	1.84	2.92	6.41	10.65	17.03	22.38	27.41
(β estimated)	(13)	Variance	.051	.078	.196	.397	.866	1.040	1.297
Diffusion (Asymptotic; (5)+constant:	(14)	Mean	10.4	11.0	12.7	15.4	21	26.6	32.1

BIOGRAPHY

Donald P. Gaver, Jr.

Mr. Gaver received his B. S. degree from M. I. T. in 1950 and his Ph. D in Mathematics in 1956 from Princeton University. He has been active in Operations Research and Statistical Analysis as applied to industrial problems since the early 1950's. Mr. Gaver has numerous publications in the areas of queueing problems, statistics and probability. He is presently a Professor of Mathematical Statistics at Carnegie-Mellon University and also is an Advisory Mathematician to Westinghouse Research Labs. He is a membor of IMS, ORSA, and the American Statistical Association.