Using Conditional Expectation to Reduce Variance in Discrete Event Simulation

S. S. Lavenberg and P. D. Welch
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

Abstract

We review the method of using conditional expectation to reduce variance in discrete event simulation and present a new application to the simulation of queueing networks. This method can be particularly useful in reducing variance when estimating quantities associated with rare events.

1. INTRODUCTION

We will be concerned with a variance reduction technique based on the use of conditional expectation and its application to the discrete event simulation of stochastic models of systems. Such a variance reduction technique is not new and has been referred to by a variety of names including statistical estimation (by Kahn and Marshall (4) in the context of Monte Carlo calculations and by McGrath and Irving (10) in the context of discrete event simulation), virtual measures (by Carter and Ignall (2) in the context of discrete event simulation), conditional Monte Carlo (by Burt and Garman (1) in the context of simulating stochastic PERT networks) and simple conditioning swindle (by Simon (12) in the context of sampling experiments). The use of the term conditional Monte Carlo does not seem to be appropriate however, since this term usually refers to a much more complex variance reduction technique (e.g., see Hammersley and Handscomb (3)).

The basic idea is the following. Suppose we wish to estimate the expected value of a random variable X. The straightforward method is to generate observations of X and estimate E[X] by the sample mean. Suppose, however, that X depends on another random variable Y and that given the value of Y the expected value of X, E[X | Y], can be computed. Then by generating observations of Y, E[X] can be estimated by the sample mean of the computed conditional expectations. As applied to discrete event simulation Y will typically be the state of the system. As we will see this method is particularly useful for reducing variance when estimating quantities associated with rare events.

In the next section we review the use of conditional expectations and present two examples, a simple sampling experiment example due to Simon (12) and a complex discrete event simulation example due to Carter and Ignall (2). In Section 3 we present an application of conditional expectation to the simulation of queueing networks. This application is new. We present our conclusions in Section 4.

Other variance reduction techniques for discrete event simulation are discussed in the 1974 book by Kleijnen (5) which also contains an extensive bibliography. The 1978 paper by Lavenberg and Welch (9) contains more recent references as a supplement to Kleijnen's bibliography.

2. REVIEW OF METHOD

For any two random variables X and Y it is known that

$$E[E[X \mid Y]] = E[X] \tag{1}$$

and

$$\begin{aligned} \mathrm{Var}[E[X \mid Y]] &= \mathrm{Var}[X] - E[\mathrm{Var}[X \mid Y]] \\ &\leq \mathrm{Var}[X] \end{aligned} \tag{2}$$

with equality holding in equation (2) if and only if X is a deterministic function of Y. Suppose that we wish to estimate E[X] and that E[X | Y] is a known function of Y which we denote by g(Y). Then from equations (1) and (2) g(Y) has the same expected value as X and, as long as X is not a deterministic function of Y, g(Y) has smaller variance than X. This suggests a better way to estimate E[X] than the usual way.

Example (Simon (12)). X has a beta distribution with parameters Y and $Y^2 + 1$ where Y has a Poisson distribution with mean equal to 10. The usual way to estimate E[X] is to generate independent observations of Y, denoted $Y_1,\ldots,Y_N$, and from these to generate independent observations of X, denoted $X_1,\ldots,X_N$, where X_n is a sample from the beta distribution with parameters Y_n and $Y_n^2 + 1$. The usual unbiased estimate of E[X] is

$$\bar{X} = \sum_{n=1}^{N} X_n/N. \tag{3}$$

However, g(Y) = E[X | Y] is a known function of Y; in particular

$$g(Y) = Y/(Y^2 + Y + 1). \tag{4}$$

Then

$$\bar{g} = \sum_{n=1}^{N} g(Y_n)/N \tag{5}$$

CH1437-3/79/0291-0294$00.75

1979 Winter Simulation Conference

is an unbiased estimate of E[X] and since the observations of X are statistically independent as are the observations of Y, it follows from equations (2), (3) and (5) that $Var[\bar{g}]<Var[\bar{X}]$. Note that observations of X need not be generated to obtain $\bar{g}$, but that the function g has to be computed for each observation of Y.

In the above example it was obvious what the conditioning random variable Y should be and the observations of X were independent as were the observations of Y. In discrete event simulation, the choice of Y may not be so obvious and the observations obtained will typically not be independent. Nonetheless it may be possible to apply the method with dramatic results as the following example shows.

Example (Carter and Ignall (2)). A simulation model of fire department operations in a borough of New York City was considered. The model was used to measure the effectiveness of various policies for deploying fire fighting equipment in responding to serious fires. Serious fires occurred relatively rarely (about 1 in every 30 alarms was for a serious fire). In the model it was assumed that serious fires occur according to a homogeneous Poisson process. The state of the simulated system described the disposition of all fire fighting equipment. The simulated system was observed at a set of discrete times yielding the dependent sequence of observations (Y_j,Z_j,X_j), j=1,...,J, where for the j-th observation Y_j is the state, $Z_j = 1$ if a serious fire occurs and $Z_j = 0$ otherwise, and X_j is the time to respond to the serious fire if $Z_j = 1$ and $X_j = 0$ if $Z_j = 0$. It was assumed that for all j, (Y_j,Z_j,X_j) is distributed as (Y, Z, X). The quantity to be estimated was $E[X \mid Z=1]$, the expected response time to a serious fire. Let N denote the number of serious fires observed, i.e., the number of observations for which $Z_j = 1$. The usual estimate of $E[X \mid Z=1]$ is the average $\bar{X}$ of the N observed response times to serious fires. (Note that since N is a random variable, $\bar{X}$ will in general be biased.) For this model $g(Y)=E[X \mid Z=1,Y]$ can be computed as a function of Y, i.e., given the state of the system and given that a serious fire occurs the expected response time can be computed. In order to apply the method we could compute $g(Y_j)$ for all observations Y_j such that $Z_j = 1$ and average these N computed values. However, we can do even better! Due to the Poisson nature of the serious fires Z and Y are independent, i.e., $P\{Y=y \mid Z=1\} = P\{Y=y\}$ for all y. Thus,

$$E[X \mid Z = 1] = \sum_{y} E[X \mid Z = 1, Y = y]P\{Y = y\}. \qquad (6)$$

$g(Y_j)$ can be computed for all observations Y_j, not just those for which $Z_j = 1$, and averaged to obtain the estimate

$$\bar{g} = \sum_{j=1}^{J} g(Y_j)/J. \qquad (7)$$

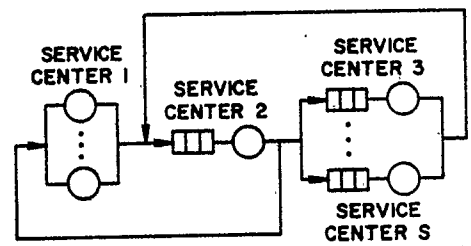
It follows from equation (6) that $\bar{g}$ is unbiased. The simulation is run only to obtain observations of the state of the system. For each observed state the conditional expected response time to a serious fire is computed whether or not a serious fire actually occurred. The observed response times to serious fires can be discarded. Note that the observations (Y_j,Z_j,X_j) are not modified by this procedure. The observations are merely processed in a different way.

Since $Var[g(Y)] < Var[X]$ and N, the number of observations of X, is typically much smaller than J, the number of observations of Y, we expect $\bar{g}$ to have a much smaller variance than $\bar{X}$ does. However, since the observations of X are not independent and the observations of Y are not independent we could not prove that $Var[\bar{g}]<Var[\bar{X}]$. Carter and Ignall found empirically, however, that variance was reduced by a factor of approximately 20. The cost involved in computing g(Y) doubled the cost of the simulation. Thus, variance was reduced by a factor of about 10 for the same total cost.

3. APPLICATION TO QUEUEING NETWORKS

Queueing networks are commonly used to model the contention for resources in job shop type systems and have been particularly important in recent years as models of computer systems and communication networks (e.g., see Kleinrock (6), Kobayashi (7), Reiser and Sauer (11)). The queueing network shown in Figure 1 is a simple model of an interactive multiprogrammed computer system which we will use to illustrate the application of conditional expectation. There are a fixed number K of customers in the network, each customer representing a user of the computer system. Service center 1 represents the terminals and a service time at this service center represents a user's "think" time. Service center 2 represents the processor and service centers 3,...,S represent secondary storage devices (e.g., drums and disks). A service time at service center 2 represents an interval of processing time for a user's task until either the task is completed, in which case the user starts another think time, or until data from a secondary storage device is required. A service time at any of service centers 3,...,S represents the time to access and transfer data from a secondary storage device to main memory. A customer leaving service center 2 next enters service center s with probability p_s, s=1,...,S, where $p_2 = 0$, independent of the state of the system. The queueing discipline for each service center is first come first served. The service times at service center s are independent random variables, each distributed as a nonnegative random variable T_s which has an arbitrary distribution with finite moments of all orders.

FIGURE 1
QUEUEING NETWORK



Suppose we wish to estimate the expected waiting time (not including the service time) in service center 3. Let the state of the network be given by the number of customers in each service center and the elapsed service time for each customer in service and let Y(t) denote the state of the network at time t. Let t_j, j = 1,2,..., denote the times at which customers complete service at service center 2 and let Y_j denote the state just prior to t_j, i.e., $Y_j = Y(t_j-)$. Let $Z_j = 1$ if the customer which completes service at t_j, next enters service center 3 and let $Z_j = 0$ otherwise. Let X_j denote the waiting time in service center 3 for the customer which completes service at t_j if $Z_j = 1$ and let $X_j = 0$ if $Z_j = 0$. We assume that for all j, (Y_j,Z_j,X_j) is distributed as (Y, Z, X). We wish to estimate $E[X \mid Z=1]$. The usual estimate would be the average of the observed waiting times in service

center 3, i.e., the average of those X_j for which $Z_j = 1$. Note that if p_3 is small the number of observed waiting times will be small. For this model $g(Y) = E[X \mid Z=1,Y]$ can be computed as a function of Y as follows. For state Y let n(Y) denote the number of customers in service center 3 and let T(Y) denote the elapsed service time of the customer in service in service center 3 if n(Y)>0 and let T(Y)=0 if n(Y)=0. Then, since the queueing discipline is first come first served

$$E[X \mid Z = 1,Y] = T(Y) + E[T_3] \max (n(Y) - 1,0). \quad (8)$$

Furthermore, since a customer completing service at service center 2 next enters service center 3 with probability p_3 independent of the state of the system, it follows that Z and Y are independent. Thus, equation (6) holds for this model. (It is necessary to integrate over the continuous-valued components of the state in equation (6).) $g(Y_j)$ can be computed for all the observations Y_j, not just those for which $Z_j = 1$. It follows from equation (6) that the average $\bar{g}$ of these computed values is an unbiased estimate for $E[X \mid Z=1]$. As in the fire department simulation example in the preceding section we expect that $Var[\bar{g}]<Var[\bar{X}]$ since $Var[g(Y)]<Var[X]$ and the number of observations of X is less than the number of observations of Y (much less if p_3 is small). However, we could not prove this since the observations of X are dependent and the observations of Y are dependent. Later in this section we will present empirical results to illustrate the variance reduction obtainable.

The expected waiting time in service center s, s=4,...,S, can be estimated in the same way. Note that each observation Y_j can be used to compute the conditional expected waiting times for each of service centers 3,...,S. It does not matter which service center was actually entered at time t_j. The expected waiting time in service center 2 can be estimated in a similar way by observing the sequence of the states just before customers complete service at any of service centers 1,3,...,S. Since these customers always enter service center 2 next Z will always equal 1. Thus, the number of observations of the state will not be greater than the number of observed waiting times as was the case for service centers 3,...,S. However, we still expect the variance to be reduced since $Var[g(Y)]<Var[X]$.

We can also estimate moments of the waiting time in a similar way. For service center 3, consider estimating $E[X^2 \mid Z = 1]$, the second moment of the waiting time. It is straightforward to show that

$$E[X^2 \mid Z = 1,Y] = T^2(Y) + (2T(Y)E[T_3] + E[T_3^2]$$

$$+ (n(Y) - 2)(E[T_3])^2) \max (n(Y) - 1,0). \quad (9)$$

The waiting time distribution function, i.e., $P\{X \le t \mid Z=1\}$, can also be estimated, although the computation of $g(Y)=P\{X \le t \mid Z=1,Y\}$ may not be easy.

We now present empirical results which illustrate the variance reduction obtainable using conditional expectation. For the networks we simulated all service times were exponentially distributed. In that case the elapsed service times can be dropped from the state, i.e., the state is simply the number of customers in each service center and equations (8) and (9) simplify. For example, equation (8) becomes

$$E[X \mid Z = 1,Y] = E[T_3]n(Y). \quad (10)$$

We simulated the 8 networks described in Table 1; S is the number of service centers and K is the number of customers. For all 8 networks $E[T_1] = 100$, $E[T_2] = 1$, $p_1 = .2$, $p_2 = 0$. For networks 1-4 $p_3 = .72$, $p_4 = .08$ and for networks 5-8, $p_3 = p_4 = .36$, $p_5 = p_6 = .04$. Both the mean and second moment of waiting times were estimated using conditional expectation and via the usual method. Table 2 contains estimates of the variance ratio $Var[\bar{g}]/Var[\bar{X}]$ obtained from 100 independent replications of a simulation. The variance ratio estimate used was the ratio of the sample variances. Table 3 contains similar results for estimating the second moment.

TABLE 1
NETWORKS SIMULATED

Network	S	K	$E[T_3]$	$E[T_4]$	$E[T_5]$	$E[T_6]$
1	4	15	1.39	12.5	-	-
2	4	25	1.39	12.5	-	-
3	4	15	.694	6.25	-	-
4	4	25	.694	6.25	-	-
5	6	15	2.78	2.78	25.0	25.0
6	6	25	2.78	2.78	25.0	25.0
7	6	15	1.39	1.39	12.5	12.5
8	6	25	1.39	1.39	12.5	12.5

TABLE 2
VARIANCE RATIOS FOR ESTIMATING
MEAN WAITING TIMES

Network	Service Center				
	2	3	4	5	6
1	.81	.72	.44	-	-
2	.79	.80	.59	-	-
3	.74	.57	.33	-	-
4	.91	.67	.37	-	-
5	.71	.53	.54	.46	.37
6	.78	.68	.68	.66	.68
7	.84	.48	.43	.28	.22
8	.86	.43	.51	.37	.43

TABLE 3
VARIANCE RATIOS FOR ESTIMATING
SECOND MOMENT OF WAITING TIMES

Network	Service Center				
	2	3	4	5	6
1	.67	.57	.33	-	-
2	.68	.73	.53	-	-
3	.53	.34	.20	-	-
4	.87	.47	.31	-	-
5	.51	.43	.42	.37	.27
6	.61	.52	.53	.66	.62
7	.73	.31	.33	.14	.15
8	.81	.29	.37	.26	.32

Note that variance reduction (variance ratio less than one) was always obtained using conditional expectation. Typically the variance reduction was greatest at the least frequently visited service center and the variance reduction was greater for estimating the second moment than for estimating the mean. The variance reductions were achieved with negligible extra computing cost.

4. CONCLUSIONS

We have reviewed the variance reduction method of using conditional expectation and presented a new application to queueing network simulation. In the empirical studies we conducted for queueing networks variance reduction was always achieved using the method and the variance reduction was greatest when estimating quantities associated with rare events, e.g., mean waiting time at an infrequently visited service center. Since the estimates obtained using the method are sample means (e.g., see equation (7)) it should not be any more difficult than usual to construct confidence intervals when using the method. This is not the case for some other variance reduction methods (e.g., see Lavenberg, Moeller, and Welch (8)).

REFERENCES

1. Burt, J. M. and M. B. Garman, "Conditional Monte Carlo: A Simulation Technique for Stochastic Network Analysis," *Manage. Sci.* 18, 207-217, 1971.
2. Carter, G. and E. J. Ignall, "Virtual Measures: A Variance Reduction Technique for Simulation," *Manage. Sci.* 21, 607-616, 1975.
3. Hammersley, J. M. and D. C. Handscomb, *Monte Carlo Methods*, Methuen and Co., Ltd., London, 1964.
4. H. Kahn and A. W. Marshall, "Methods for Reducing Sample Size in Monte Carlo Computations," *Oper. Res.* 1, 263-278, 1953.
5. Kleijnen, J.P.C., *Statistical Techniques in Simulation Part I*, Marcel Dekker, Inc., New York, 1974.
6. Kleinrock, L., *Queueing Systems Volume 2: Computer Applications*, John Wiley and Sons, New York, 1976.
7. Kobayashi, H., *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*, Addison-Wesley, Reading, Massachusetts, 1978.
8. Lavenberg, S. S., T. L. Moeller and P. D. Welch, "Statistical Results on Multiple Control Variables with Application to Variance Reduction in Queueing Network Simulation," IBM Research Report RC 7423, Yorktown Heights, New York, 1978.
9. Lavenberg, S. S. and P. D. Welch, "Variance Reduction Techniques," *Proc. 1978 Winter Simulation Conference*, 167-170, 1978.
10. McGrath, E. J. and D. C. Irving, "Application of Variance Reduction to Large Scale Simulation Problems," *Comput. and Ops Res* 1, 283-311, 1974.
11. Reiser, M. and C. H. Sauer, "Queueing Network Models: Methods of Solution and Their Program Implementation," in *Current Trends in Programming Methodology Volume III: Software Modeling*, K. M. Chandy and R.T. Yeh, Editors, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.
12. Simon, G., "Computer Simulation Swindles, with Applications to Estimates of Location and Dispersion," *Appl. Statist.* 25, 266-274, 1976.