Proceedings of the 1985 Winter Simulation Conference
D. Gantz, G. Blais, S. Solomon (eds.)

A DECOMPOSITION APPROACH TO VARIANCE REDUCTION

Barry L. Nelson
Department of Industrial and Systems Engineering
The Ohio State University
1971 Neil Avenue
Columbus, Ohio 43210

For analyzing stochastic models, simulation trades the tractability problems of analytical techniques for the problem of sampling variability. Variance reduction techniques (VRTs) attack this problem by transforming the simulation experiment in a way that makes it more statistically efficient. Unfortunately, VRTs are infrequently used, even though significant reductions are possible in practical problems. This tutorial introduces some basic concepts of variance reduction, and uses a new taxonomy of VRTs as the basis for an algorithm to select appropriate VRTs for general simulation experiments.

1. Introduction

Simulation is conceptually the simplest methodology for analyzing dynamic, stochastic systems: A model is defined by probability distributions that characterize the uncertain, or uncontrollable elements in the system, and by an algorithm that mimics the behavior or response of the system, given values of the uncertain elements. A simulation experiment is performed by sampling values of the uncertain elements, exercising the algorithm, and observing the resulting behavior. Almost any system that can be modeled can be simulated, and there are now many computer simulation languages that facilitate sampling from probability distributions and representing logical relationships as algorithms. In addition, behavior is often automatically summarized by statistics, and will soon be routinely observed via animation.

Of course, all of the problems associated with modeling in general -- validation, for example -- are also problems in simulation. However, tractability, the primary curse for analytical analysis, is not an issue. By sampling system behavior simulation trades the tracability problem for the problem of sampling variability; if a longer realization of system behavior is generated, then almost certainly new behavior will be observed. Even with increasing computer speeds it is not always possible to observe the system long enough to ensure a representative sample. In fact, the availability of faster computers, rather than diminishing the problem, has spurred interest in experimentation that was previously unmanageable. For example, the use of simulation to optimize stochastic models and in conjunction with real-time control systems are two applications for which available computer budgets and computer speed, respectively, are not adequate.

One measure of the uncertainty inherent in sampling is the variance of the estimators (statistics). Variance reduction techniques (VRTs) reduce the population variance of estimators based on sampling, without increasing the computer (sampling) burden. VRTs succeed by increasing and making better use of the information generated by the simulation experiment. Although historically VRTs have been applied to estimators of unknown system performance parameters, they will become more important with the increasing use of animation. Animation necessarily implies that only a brief realization of system behavior will be observed. Thus, to make reliable decisions a representative sample is essential.

Unfortunately, VRTs are seldom used, despite the fact that significant (one to two orders of magnitude) variance reductions are possible in practical experiments. There are at least two reasons for this situation: First, VRTs were originally developed for survey sampling (Cochran [1]) and Monte Carlo estimation (Hammersley and Handscomb [2]), and it is difficult to adapt techniques appropriate for sampling from static populations and evaluating definite integrals to simulating dynamic stochastic processes. Secondly, there is no unifying theory of variance reduction, making it difficult to select a VRT that will work from among the multitude that have been developed. Researchers have attacked the first problem in recent years, primarily by restricting attention to classes of models, and there now exists both theory and computational experience for variance reduction in simulation. Nelson and Schmeiser [3, 4] have addressed the second problem by proposing a parsimonious taxonomy of VRTs. One purpose of this tutorial is to combine these two research efforts into a guide for selecting VRTs and finding available information about using them.

The tutorial is divided into two parts. The first part (sections 2 - 4) introduces some basic concepts of variance reduction and presents an overview of the taxonomy of Nelson and Schmeiser. While not an exhaustive survey of VRTs (we reference several good ones later), we do discuss five VRTs as illustrations and list several others. The second part (sections 5 - 7) is an algorithm for selecting potentially useful VRTs and a guide to the literature on them. A practitioner with indepth knowledge of variance reduction will likely do a better job of selecting VRTs than the algorithm presented here. However, the possibility of such an algorithm demonstrates that the use of variance reduction need not be limited to such practitioners. Additional research is needed to develop an effective, automated procedure to both select and apply VRTs in general simulation experiments, but the approach presented here may suggest what part of such a procedure would be like.

2. Some Principles of Variance Reduction

Before discussing any particular VRT or the taxonomy of VRTs, we cite some basic results that are the underlying principles of many VRTs.

Suppose we are estimating some unknown scalar parameter, θ, using an estimator that is a function of the sequence of random variables $Y = (Y_1, Y_2, \ldots, Y_I)$, where $E[Y_i] = \theta$, $Var[Y_i] = \sigma^2$ for all i, and $Cov[Y_i, Y_j] = \phi_h$ when $|i - j| = h$ (i.e. Y is covariance stationary). Let $Z = \Sigma Y_i / I$ be the sample mean, and let X be some other random variable with cumulative distribution function (cdf) F(x). The following are well-known results in statistics (see for instance, Bickel and Doksum [5]):

$$Var[Z] = \frac{\sigma^2}{I} + \frac{2}{I} \sum_{h=1}^{I-1} (1 - h/I)\phi_h \qquad (1)$$

If the Y_i are independent, then (1) reduces to σ^2/I. Now let X be a scalar random variable with variance σ_X^2, b a constant, and $Z = Y_i \pm bX$. Then

$$Var[Z] = \sigma^2 + b^2 \sigma_X^2 \pm 2bCov[Y_i, X] \qquad (2)$$

Finally, (3) expresses the expected value of Y in as an average over the random variable X.

$$\theta = \int E[Y_i \mid X = x]dF(x) \qquad (3)$$

Result (1) shows that there are three components that determine the variance of a sample mean: σ^2, ϕ_h, and I. Decreasing σ^2 and ϕ_h, or increasing I can reduce Var[Z]. Result (2) shows that the combination of Y with another random variable X may yield a random variable with smaller variance, provided the covariance between them is large enough and has the correct sign. Variance reduction involves changing or transforming a simulation experiment in ways suggested by these results. Of course, in the process of making any changes we need to take care to preserve Z as an estimator of θ.

To change an experiment in a useful way often depends upon being able to express the unknown parameter θ in terms other than $E[Y_i]$. Result (3) is one of the most useful alternative expressions. In particular, if we have prior knowledge about either $E[Y_i|X]$ or F(x) we may be able to use it to estimate θ with smaller variance than using Y directly.

It is important to stress that *variance reduction* refers to reducing the population variance of an estimator of θ, where θ may be a variance. Variance reduction does not necessarily affect the variability of the simulated stochastic process.

3. A Taxonomy of Variance Reduction

In this section we give an overview of the taxonomy of VRTs proposed by Nelson and Schmeiser [3, 4]. The taxonomy characterizes VRTs as transformations from one simulation experiment to another, and decomposes VRTs into combinations of transformations from six elemental classes. Thus, a definition of simulation experiments is needed to make the concept of a transformation precise.

For our purposes, a simulation experiment is a collection of interrelated random variables. Given a source of randomness (usually independent, identically uniformly distributed random variables on [0,1], denoted U(0,1)), realizations of the simulation experiment can be generated. We partition the random variables into three subsets, *inputs*, *outputs*, and *statistics*, that can be described loosely as follows:

The *inputs*, denoted by X, are random variables defined by known (possibly conditional) probability distributions. Examples are the interarrival and service times in a queueing network simulation or the time until component failure in a reliability model. Another example is the demand per period in an inventory system whose distribution, conditional on the time of the year, is known. The distribution of the countably infinite set X is denoted F(x).

The *outputs*, denoted by Y, are random variables defined by known (possibly implicity) functions of the inputs. They include the observations of system performance. Examples are the delay experienced by a customer in a queueing simulation or the time between total system failures in a reliability model. For theoretical reasons (Nelson [3]) we restrict Y to the *essential* random variables defined by functions of X, in the sense that

all remaining random variables that are functions of X can be derived from Y.

In a simulation experiment there may be many sequences of outputs and (conceptually) we can sample indefinitely. Thus, we define a *sampling plan*, denoted by R_*, that specifies a stopping rule for the simulation experiment in terms of the lengths of the various output sequences. We write $Y = g(X;R_*)$. The function g embodies the operating logic of the system we are modeling.

The *statistics*, denoted by Z, are functions that aggregrate the outputs into point estimators of the system parameters of interest, θ. We write $Z = h(Y)$. Variance reduction refers to reducing the variance of Z, and not necessarily any of the elements of X and Y.

Suppose that Z and θ are scalars, and Z is an unbiased estimator of θ. Then

$$Var[Z] = \int [h(g(x;R_*)) - \theta]^2 dF(x) \qquad (4)$$

Nelson and Schmeiser [4] view VRTs as transformations that redefine F, g, R_* and/or h to reduce Var[Z], while holding θ and the sample space of X fixed. All possible transformations can be formed by compositions of members of six elemental classes. Loosely defined, the classes are:

1. *Distribution Replacement (DR):* Redefine the scalar marginal distributions of the inputs without altering any statistical dependencies among the inputs.

2. *Dependence Induction (DI):* Redefine the statistical dependencies among the scalar inputs without altering any marginal distributions of the inputs.

3. *Equivalent Allocation (EA):* Redefine the functions from inputs to outputs, g without altering the sampling plan, R_*.

4. *Sample Allocation (SA):* Redefine the sampling plan, R_* without altering the functions from inputs to outputs, g.

5. *Auxiliary Information (AI):* Redefine the argument (subset of Y) of the statistics without altering the functions from outputs to statistics, h.

6. *Equivalent Information (EI):* Redefine the functions, h, from outputs to statistics without altering the argument set of the statistics.

These six classes of transformations exhaust the ways to transform a simulation experiment to reduce (4); the rigorous definitions needed to prove this property are given in [3, 4], along with the proof. While there are other possible partitions of the transformations, this set is useful for studing variance reduction. In particular, VRTs can be decomposed into to their elemental transformations thereby facilitating selection and application of appropriate VRTs (see Nelson [6]).

4. Some Variance Reduction Techniques

In this section some specific VRTs are presented, emphasizing the basic principles they employ (section 2) and, their decomposition in terms of the taxonomy (section 3). The notation is the same as section 3. We do not discuss issues of implementation or effectiveness, but refer the reader to the references cited later. Since there is no universally accepted definition of any VRT, we present simple versions that are useful in a tutorial setting; see Nelson and Schmeiser [6] for a more complete development. Broadly defined variance reduction strategies are the subject of Nelson [7].

4.1 Antithetic Variates (AV)

Suppose we estimate θ by the sample mean $Z = \Sigma Y_i/I$. AV exploits (1) by inducing favorable covariance terms ϕ_h. The required covariances between the outputs are realized indirectly by inducing dependence among previously independent inputs (a transformation in DI). There are two major problems: First, it is not possible to make all the covariances negative. Thus AV often induces negative covariance between pairs of outputs, (Y_{2i-1}, Y_{2i}), leaving different pairs independent. The second problem is that inducing negative covariance between inputs does not guarantee negative covariance between outputs. A number of researchers have addressed these issues (see the references).

A fundamental result is that if X_{2i-1} and X_{2i} are scalar inputs with cdfs $F_1(x)$ and $F_2(x)$, respectively, then letting

$$X_{2i-1} = F_1^{-1}(U) \qquad (5)$$
$$X_{2i} = F_2^{-1}(1-U)$$

where $U \sim U(0,1)$, yields realizations of X_{2i-1} and X_{2i} with the correct marginal distributions and the minimal achievable covariance. This method of variate generation is known as the *inverse transform*. If the output transformation $Y = g(X;R_*)$ is monotone in the inputs then the negative correlation between the inputs is preserved in the outputs.

4.2 Common Random Numbers (CRN)

Suppose that, instead of θ being an absolute measure, $\theta = \alpha - \beta$, a difference or relative measure. This situation is common since we frequently compare the performance of system 1 (α) and system 2 (β) to determine which is better. Of course, in general we may want to compare several systems, not just two.

If sample means of the output sequences Y and Y′ are used to estimate α and β, respectively, then CRN exploits (2) (where X = Y′, and b = 1) by inducing positive covariance between pairs of outputs and using the difference of pairs ($Y_i - Y'_i$, i = 1, 2,..., I) as the basic observations. Again, the covariance is induced by inducing dependence between the inputs (a transformation in DI); the issue of preserving this covariance is also a factor in CRN, as in AV. In the worst case, the sign of the covariance between the outputs is reversed!

The maximum possible covariance between pairs of inputs is induced via (5) with 1 - U replaced by U (a "common" random number) in the second expression. CRN is the most intuitively appealing VRT because the idea of comparing systems under as nearly identical conditions (here, the inputs) as possible is easily accepted. In fact, if $F_1 = F_2$ then identical inputs drive both systems.

Practically speaking, preserving induced dependence requires the inverse transform approach to variate generation (5) and synchronizing the inputs between antithetic or common random number runs. Bratley, Fox and Schrage [8], pages 44-57, give an excellent discussion of, and practical suggestions for, synchronization.

4.3 Stratified Sampling (STRAT)

STRAT exploits (3) by dividing the range of X (usually an input) into nonoverlapping intervals (strata), sampling from each strata a fixed number of times, estimating the conditional expectation of Y in each strata separately, and then combining the estimators via (3). Denote the strata by L_j and let $p_j = Pr\{X \in L_j\}$, j = 1, 2,..., n.

Within each strata, $E[Y|X \in L_j]$ is estimated by the sample mean of I_j independent observations of $Y_{i,j}$, where $Y_{i,j}$ is the ith output when $X \in L_j$, and $\Sigma I_j = I$. Since the variance of a sample mean decreases as the number of observations increases by (1), an intelligent selection of the I_j (a transformation in SA) that allocates more observations to those strata with larger variance will reduce the overall variance of the combined estimator.

The new estimator (a transformation in EI) is

$$Z = \sum_{j=1}^{n} \left[\sum_{i=1}^{I_j} \frac{1}{I_j} Y_{ij} \right] p_j \qquad (6)$$

The term inside the brackets is an estimator of $E[Y|X \in L_j]$, and these estimators are weighted by $p_j = P\{X \in L_j\}$ and summed as in (3). Note that the p_j must be known, and they will be if X is an input.

Although widely used in survey sampling, STRAT has been less successful in simulation because it is often difficult to control the sampling plan of a dynamic stochastic process and the Y_i are dependent. Good stratification variables (X) that are sometimes available in simulation are initial conditions generated randomly at the beginning of independent simulation runs (e.g. the number of employees that show up for work at the beginning of the day).

4.4 Postratifying the Sample (PSTRAT)

One source of variance in estimating θ is that the empirical distribution of Y will almost surely not match the theoretical distribution. Of course, the distribution of Y is unknown in general, so there is no way to measure just how significant the deviation is. However, again using (3), if the distribution of X is known then we can measure how far its empirical distribution deviates from its theoretical distribution. STRAT fixes the number of observations from each strata, I_j, in the sampling plan. When predetermining the sampling plan in this way is not possible, then the I_j become random outputs of the simulation, and provide an empirical distribution for X.

We expect $I_j = Ip_j$. If $I_j > Ip_j$ then strata j is overrepresented probabilistically, and if $I_j < Ip_j$ it is underrepresented. Using functionally the same estimator (6) as STRAT, the PSTRAT estimator gives weight 1/I to each $Y_{i,j}$ only if the sample distributes itself proportionately (all $I_j = Ip_j$); otherwise it gives a smaller or larger weight to $Y_{i,j}$ depending on whether strata j is over or underrepresented, respectively. PSTRAT does not alter R_*, but rather uses the auxiliary outputs I_j (an AI transformation) and the estimator (6) (an EI transformation) to reduce variance.

4.5 Control Variates (CV)

PSTRAT uses an auxiliary variable to correct for disproportionate sampling. CV also employs AI and EI to adjust an estimator. We will discuss only the linear CV, but there are many other forms (Nelson [9]).

Suppose $E[X_i] = \mu$, and μ is known (as it would be for an input). Then CV exploits (2) by forming the estimator

$$Z = \bar{Y} - b(\bar{X} - \mu) \qquad (7)$$

where the bar over X and Y denotes sample means. For any constant b, (7) is unbiased for θ. The difference between X_i and its expectation is used to correct the observed value of Y_i, and (2) shows that if $Cov[Y_i,X_i]$ is large enough then the variance of the new estimator will be smaller. Since the outputs are functions of the inputs, correlation is likely to be present.

The value of b that minimizes the variance is $b^* = Cov[Y_i,X_i]/Var[X_i]$, which is seldom known. The issues involved in estimating b*, called the CV multiplier, have been a research topic for some time. The CV estimator (7) readily generalizes to multiple control variates and estimating multivariate θ.

AV and CRN are based on the potential to induce correlation. If X is an output from simulating a second system, similar to the system of interest but for which the analytic solution is known, then simulating the two systems using CRN may induce positive correlation between the corresponding outputs. The output of the second system can then be used as a control variate. This variation is called *external* control variates (ECV), and combines transformations from DI, AI and EI.

4.6 Summary

AV, CRN, STRAT, PSTRAT and CV are only five of many VRTs. Two additional VRTs that employ transformations from DR and EA are decomposed in [6]. A necessary condition for successfully applying any VRT is to have *prior knowledge*, which we define to be any knowledge either known with certainty or suspected, beyond what is needed to construct the original simulation experiment. Examples we have seen so far are a) monotonicity of g, the function from input to output, b) conditional relationships between random variables, and c) correlation between random variables. In the next section we present an algorithm for determining some of the available prior knowledge and using it to select VRTs in general simulation experiments.

5. An Algorithm for Selecting VRTs

This section contains an (informal) algorithm for selecting potentially useful VRTs for general simulation experiments. The three major steps (and the three subsections of this section) are 1) express the simulation experiment in terms of inputs, outputs, sampling plan and statitics (section 3), 2) determine the available prior knowledge, and 3) select a VRT from those completely or partially decomposed into elemental transformations in Figure 3. A list of references organized by VRT (section 6) can then be consulted for specific details.

The philosophy of this algorithm is to select a single VRT, since great care must be taken to combine VRTs so that they do not conflict. We also assume that a VRT is sought for a dynamic simulation model, rather than for survey sampling or evaluation of an explicit integral (see [1] and [2], respectively, for these applications).

To lay the foundation for any simulation experiment, the experimenter should carefully consider what parameters are important to estimate. This is especially true when contemplating variance reduction, because VRTs that improve the estimators of some parameters may make it more difficult to estimate other parameters. Another important decision is whether unbiased estimators are required, since some VRTs introduce bias. In many cases the variance reduction is so large that a certain amount of bias is tolerable, but the decision should be made in advance. Finally, the biggest payoff from variance reduction comes when the simulation experiment evaluates multiple configurations of a basic system model. When only a single simulation model is analyzed then careful consideration of the experimenter effort required to use a VRT is necessary.

5.1 Definition of the Experiment

Use the following steps to partition the random variables in the experiment into inputs, outputs and statistics as described in section 3:

1. List the system performance measures to be estimated. These are the parameters of interest, θ. Note that θ will include the parameters of interest of all variations of the system that will be simulated.

2. List all random variables whose realizations will be generated from known distributions. These are the inputs, X. Note that even though the distribution of a random variable may be the result of fitting a family of distributions to data, once a distribution is selected it is considered "known" in the context of our definition. An input may also be a value selected from a distribution known implicitly by the experimenter.

3. List all other random variables in the experiment. Since they are not inputs they must be functions of the inputs. These are tentatively the outputs, Y. Although for theoretical reasons (Nelson and Schmeiser [3]) we defined an essential output set in section 3, for the algorithm presented here it is better to make this set too large rather than too small.

4. List the conditions that determine when the simulation experiment will end (how many outputs will be generated). If the stopping rule depends on a sequential procedure, list the test for stopping. This is the sampling plan, R*. Note that the sampling plan is based on the outputs, but is not an output.

5. Among those random variables tentatively identified as outputs, list the ones that are point estimators of the parameters of interest. These are the statistics. There should be only one point estimator for each parameter. The remaining random variables are the outputs.

5.2 Determining the Available Prior Knowledge

The set of questions that follow determine some of the prior knowledge the experimenter has and converts it to knowledge about a particular class of models and/or the six classes of transformations defined in section 3. The abbreviations DR, DI, EA, SA, AI and EI are used to designate the classes of elemental transformations. The DI and AI classes are further divided into DI+ and DI- (relative versus absolute measure) and AI.s and AI.f (statistical versus functional relationship). Positive answers to a question require recording a response.

0. Given: a simulation experiment expressed in terms of the definition of section 3. Further divide the outputs into those that are arguments of a statistic (outputs of interest) and those that are not arguments of a statistic (auxiliary outputs).

Classes of Models

1. Can the simulation model be considered a member of one or more of the following classes of models?

1.1 A stochastic network (see Figure 1)?

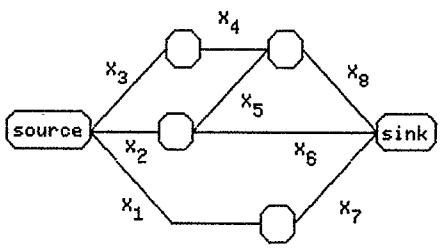

Figure 1: Stochastic Network

1.2 A Markov chain?

1.3 An (s,S)-type inventory system?

1.4 A reliability network?

1.5 A Jackson-type queueing network (see Figure 2)?

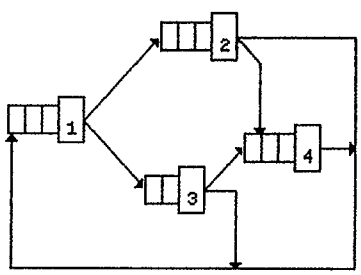

Figure 2: Queueing Network

If yes then record the model class.

Parameters of Interest

2. If the parameter of interest is a function of one or more other parameters of the system then record AI.f.

3. If the parameter of interest is the difference between parameters of two or more systems then record DI+.

4. If there are any constraints on the parameters (they must be probabilities, they must be positive, etc.) then record EI.

5. If the parameter of interest is a conditional expectation then record SA.

Inputs

6. If some inputs map one-to-one with an output of interest then record DI, AI.s.

7. If some inputs identified in 5 are independent and identically distributed, record DR.

8. If some inputs have the same distribution for alternative systems then record DI+.

9. If an input represents a system starting condition or initial state then record SA, AI.s.

10. If an input represents a demand on the system or the availability of a resource then record AI.s.

11. If some inputs can be generated via the inverse transform then record DI.

Outputs

12. If an output is monotonic in some input then record DI-.

13. If there is an approximate analytic model of the system then record DI+, EA, SA, AI.f.

14. If an output of interest is the result of a sequence of dependent events then record SA, AI.f, EI.

15. If the outputs of interest are independent then record DI.

16. If the auxiliary outputs are independent then record SA.

17. If the output of interest is the result of a rare event then record DR, SA.

18. If the experimenter has simulated a similar system then record SA, AI.s.

19. If the sampling plan can be specified in terms of the inputs then record SA.

Statistics

20. If the statistic is an average then record DI, EI.

5.3 Selection of VRTs

Consider the information recorded in section 5.2.

1. If the simulation model is an element of one of the standard classes (question 1 above) then go directly to the references for that class (section 6.3). The VRTs developed for these classes are frequently very effective.

2. Sometimes prior knowledge is only suspected to be true. Rank the knowledge recorded in section 5.2 as classes of transformations from most to least certain based on the experimenter's subjective evaluation.

3. Pick the top one or two classes from 2 above, and go to Figure 3. The selected classes designate a cell (you may have to try both row-column and column-row order) containing names of VRTs to consider. Note that the decomposition of the VRTs may not be complete, since many VRTs are composed of more than two classes of transformations. However, the decomposition given in Figure 3 is one likely to lead to a match with the results of section 5.2. The VRTs on the diagonal are those invoking only one class of elemental transformations.

4. After selecting a VRT from Figure 3, go to the references for that VRT (sections 6.4-6.13). If the VRT is not appropriate, repeat step 2.

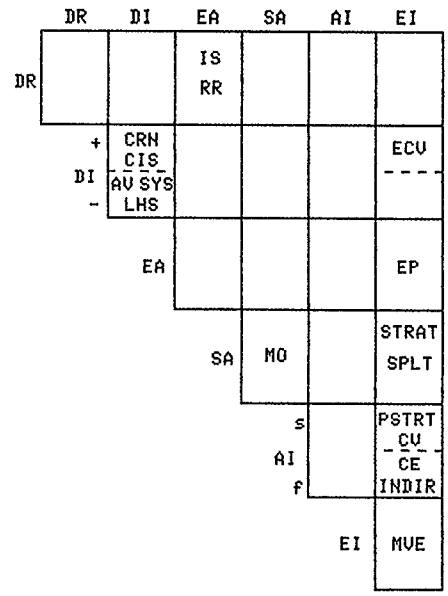

Figure 3: VRT Chart

6. Bibliography

This section contains a bibliography to be used in conjunction with the algorithm of section 5. It is not exhaustive, and contains almost entirely references that are available as books or in journals (as opposed to technical reports). Three excellent books, Bratley, Fox and Schrage, Kleijnen, and Law and Kelton, are cited once in the beginning and then repeatedly by page numbers only for VRTs that they cover. Brief descriptions are given for VRTs not discussed in section 3.

6.1 Textbook Treatments

Bratley, P., B.L. Fox and L.E. Schrage (1983), *A Guide to Simulation*, Springer-Verlag, NY.

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

6.2 Surveys

Kleijnen, J.P.C. (1974), *Statistical Techniques in Simulation*, Vol. I, Marcel Dekker, NY, Chap. 3.

McGrath, E. and D.C. Irving (1973), "Techniques for Efficient Monte Carlo Simulation, Vol. III, Variance Reduction," ORLC Report SAI-72-509-LJ.

Nelson, B.L. and B.W. Schmeiser (1985), "Decomposition of Some Well-Known Variance Reduction Techniques," *J. Statist. Comput. Simul.*, forthcoming.

Wilson, J.R. (1984), "Variance Reduction Techniques for Digital Simulation," *Am. J. Math. Mgt. Sci.*, 4, 3 & 4, 277-312.

6.3 Classes of Problems

6.3.1 Stochastic Networks

Burt, J.M. and M.B. Garman (1971), "Conditional Monte Carlo: A Simulation Technique for Stochastic Network Analysis," *Mgt. Sci.*, 18, 3, 207-217.

Burt, J.M. and M.B. Garman (1971), "Monte Carlo Techniques for Stochastic PERT Network Analysis," *INFOR*, 9, 3, 248-262.

Garman, M.B. (1972), "More on Conditioned Sampling in the Simulation of Stochastic Networks," *Mgt. Sci.*, 19, 1, 90-95.

Grant, F.H. (1983), "A Note on Efficiency of the Antithetic Variate Method for Simulating Stochastic Activity Networks," *Mgt. Sci.*, 29, 3, 381-384.

Grant, F.H. and J.J. Solberg (1983), "Variance Reduction Techniques in Stochastic Shortest Route Analysis: Application, Procedures and Results," *Math. Comput. Simul. XXV*, 366-375.

Fishman, G.S. (1985), "Estimating Network Characteristics in Stochastic Activity Networks," *Mgt. Sci.*, 31, 5, 579-593.

Fishman, G.S. (1985), "Estimating Critical Path and Arc Probabilities in Stochastic Activity Networks," *Naval Res. Log. Quart.*, 32, 2, 249-261.

Sigal, C.E., A.A.B. Pritsker and J.J. Solberg (1980), "The Use of Cutsets in Monte Carlo Analysis of Stochastic Networks," *Math. Comput. Simul. XXI*, 376-384.

Sullivan, R., J. Hayya and R. Schual (1982), "Efficiency of the Antithetic Variate Method for Simulating Stochastic Networks," *Mgt. Sci.*, 28, 5, 563-572.

6.3.2 Markov Chains

Bayes, A.J. (1972), "A Minimum Variance Sampling Technique for Simulation Models," *J. ACM*, 19, 4, 734-741.

Fishman, G.S. (1983), "Accelerated Accuracy in the Simulation of Markov Chains," *Op. Res.*, 31, 3, 466-487.

Fishman, G.S. (1983), "Accelerated Convergence in the Simulation of Countably Infinite State Markov Chains," *Op. Res.*, 31, 6, 1074-1089.

Heidelberger, P. (1977), "Variance Reduction Techniques for Simulating Markov Chains," *Proceedings of the 1977 Winter Simulation Conference*, IEEE, 161-164.

Ross, S.M. and Z. Schechner (1985), "Using Simulation to Estimate First Passage Distribution," *Mgt. Sci.*, 31, 2, 224-234.

6.3.3 Inventory Models

Ehrhardt, R. (1983), "Variance Reduction Techniques for (s,S)-Type Inventory Simulations," *Proceedings of the Conference on Simulation in Inventory and Production Control* (H. Bekiroglu, ed.), SCS, 12-16.

6.3.4 Reliability Networks

Easton, M.C. and C.K. Wong (1980), "Sequential Destruction Method for Monte Carlo Evaluation of System Reliability," *IEEE Trans. Rel.*, R-29, 1, 27-32.

Kumamoto, H., K. Tanaka and K. Inoue (1977), "Efficient Evaluation of System Reliability by Monte Carlo Method," *IEEE Trans. Rel.*, R-26, 5, 311-315.

Kumamoto, H., K. Tanaka, K. Inoue and E.J. Henley (1980), "Dagger-Sampling Monte Carlo for System Unavailability," *IEEE Trans. Rel.*, R-29, 2, 122-125.

Kumamoto, H., K. Tanaka, K. Inoue and E.J. Henley (1980), "State-Transition Monte Carlo for Evaluating Large, Repairable Systems," *IEEE Trans. Rel.*, R-29, 5, 376-380.

M. Mazumdar (1975), "Importance Sampling in Reliability Estimation," in *Reliability and Fault Tree Analysis*, SIAM, Philidelphia, 153-163.

6.3.5 Queueing Networks

Lavenberg, S.S. and P.D. Welch (1979), "Using Conditional Expectation to Reduce Variance in Discrete Event Simulation," *Proceedings of the 1979 Winter Simulation Conference*, IEEE, 291-294.

Lavenberg, S.S., T.L. Moeller and P.D. Welch (1982), "Statistical Results on Control Variables with Application to Queueing Network Simulation," *Op. Res.*, 30, 1, 182-202.

Wilson, J.R. and A.A.B. Pritsker (1984), "Variance Reduction in Queueing Simulation Using Generalized Concomitant Variables," *J. Statist. Comput. Simul.*, 19, 129-153.

Wilson, J.R. and A.A.B. Pritsker (1984), "Experimental Evaluation of Variance Reduction Techniques for Queueing Simulation Using Generalized Concomitant Variables," *Mgt. Sci.*, 30, 12, 1459-1472.

6.4 Antithetic Variates (AV) and Common Random Numbers (CRN)

See sections 4.1 and 4.2

Bratley, Fox and Schrage (1983), pp. 46-57, 273-277.

Cheng, R.C.H. (1981), "The Use of Antithetic Control Variates in Computer Simulations," *Proceedings of the 1981 Winter Simulation Conference*, IEEE, 313-318.

Cheng, R.C.H. (1982), "The Use of Antithetic Variates in Computer Simulations," *J. Opl. Res. Soc.*, 33, 229-237.

Fishman, G.S. and B.D. Huang (1983), "Antithetic Variates Revisited," *Comm. ACM*, 26, 964-971.

Gal, S., R.Y. Rubinstein and A. Ziv (1984), "On the Optimality and Efficiency of Common Random Numbers," *Math. Comput. Simul.*, 26, 6, 502-512.

Kleijnen (1974), pp. 187-240.

Law and Kelton (1982), pp. 350-357.

Mihram, G.A. (1974), "Blocking in Simular Experimental Designs," *J. Statist. Comput. Simul.*, 3, 4, 29-32.

Rubinstein, R.Y., Samorodnitsky, G. and M. Shaked (1985), "Antithetic Variates, Multivariate Dependence and Simulation of Complex Stochastic Systems," *Mgt. Sci.*, 31, 66-77.

6.4.1 Correlation Induction Strategies (CIS)

A technique for assigning random number streams to multiple simulation experiments when the simulation response can be represented as a general linear model.

Schruben, L.W. (1979), "Designing Correlation Induction Strategies for Simulation Experiments," in *Current Issues in Computer Simulation* (N.R. Adam, A. Dogramaci, eds.), Academic Press, NY, 235-256.

Schruben, L.W. and B.H. Margolin (1978), "Pseudorandom Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments," *J. ASA*, 73, 363, 504-520.

6.4.2 Latin Hypercube Sampling (LHS*) and Systematic Sampling (SYS)

Techniques for inducing dependence among more than two inputs for estimating an absolute performance measure. See also Fishman and Huang (1983) above.

*McKay, M.D., R.J. Beckman and W.J. Conover (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, 21, 2, 239-245.

Madow, W.G. and L.H. Madow (1944), "On the Theory of Systematic Sampling," *Ann. Math. Stat.*, 15, 1-24.

6.4.3 Variate Generation for Dependence Induction

The following references detail variate generation techniques for inducing dependence between inputs.

Bratley, Fox and Schrage (1983), Chap. 5.

Cheng, R.C.H. (1983), "Random Samples with Known Sample Statistics: With Application to Variance Reduction," *Proceedings of the 1983 Winter Simulation Conference*, IEEE, 395-400.

Cheng, R.C.H. (1985), "Generation of Multivariate Normal Samples with Given Sample Mean and Covariance Matrix," *J. Statist. Comput. Simul.*, 21, 39-49.

Fishman, G.S. and L.R. Moore (1984), "Sampling from a Discrete Distribution While Preserving Monotonicity," *Amer. Statistician*, 38, 3, 219-223.

Franta, W.R. (1975), "A Note on Random Variate Generators and Antithetic Sampling," *INFOR*, 13, 1, 112-117.

Kleijnen, J.P.C. (1974), "A Note on Sampling Two Correlated Variables," *Simulation*, 22, 2, 45-46.

6.5 Control Variates (CV, ECV*)

See section 4.5

Bratley, Fox and Schrage (1983), pp. 57-61, 277-280.

*Gaver, D.P. and G.S. Shedler (1971), "Control Variable Methods in the Simulation of a Model of a Multiprogrammed Computer System," *Naval Res. Log. Quart.*, 18, 435-450.

Kleijnen (1974), pp. 138-164.

Law and Kelton (1982), pp. 357-361.

Lavenberg, S.S. and P.D. Welch (1981), "A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations," *Mgt. Sci.*, 27, 3, 322-335.

Nozari, A., S.F. Arnold and C.D. Pegden (1984), "Control Variates for Multipopulation Simulation Experiments," *IIE Trans.*, 16, 159-169.

6.6 Conditional Expectations (CE)

Sometimes called conditional Monte Carlo, this VRT is based on (3) when $E[Y|X = x]$ is known for all x. It is often useful when the simulation is performed to observe rare events.

Bratley, Fox and Schrage (1983), pp. 66-70, 285.

Carter, G. and E.J. Ignall (1975), "Virtual Measures: A Variance Reduction Technique for Simulation," *Mgt. Sci.*, 21, 6, 607-616.

Lavenberg, S.S. and P.D. Welch (1979), "Using Conditional Expectation to Reduce Variance in Discrete Event Simulation," *Proceedings of the 1979 Winter Simulation Conference*, IEEE, 291-294.

Law and Kelton (1983), pp. 363-366.

6.7 Embedded Process (EP)

Sometimes simulating an embedded stochastic process directly, rather than the process that mimics the system of interest, may facilitate variance reduction. This is a relatively new area of research.

Fox, B.L. and P.W. Glynn (1985), "Discrete-time Conversion for Simulating Semi-Markov Processes," *Technical Report*, Departement d'informatique et de recherche operationelle, University de Montreal, Montreal, Canada.

6.8 Importance Sampling (IS)

Replacing F(x) with a distribution that biases sampling toward regions of interest, where interest is measured by likelihood and magnitude of the outputs in the region.

Bratley, Fox and Schrage (1983), pp. 63-66, 282-285.

Hammersley, J.M. and D.C. Handscomb (1964), *Monte Carlo Methods*, Chapman and Hall, London.

Jeruchim, M.C. (1984), "On the Application of Importance Sampling to the Simulation of Digital Satellite and Multihop Links," *IEEE Trans. Comm.*, 32, 10, 1088-1092?.

Kleijnen (1974), pp. 164-186.

6.8.1 Russian Roulette (RR)

Biases sampling by randomly killing off time paths (output sequences) in uninteresting regions.

Kahn, H. (1956), "Use of Different Monte Carlo Sampling Techniques," in *Symposium on Monte Carlo Methods* (H. Meyer, ed.), Wiley, NY, 146-190.

McGrath, E. and D.C. Irving (1973), "Techniques for Efficient Monte Carlo Simulation, Vol. III, Variance Reduction," ORLC Report SAI-72-509-LJ.

6.9 Indirect Estimators (INDIR)

Exploits functional relationships between parameters by estimating the parameter of interest as a function of the estimator of a second parameter. Applications have been in queueing simulation.

Carson, J.S. and A.M. Law (1977), "Conservation Equations and Variance Reduction in Queueing Simulations," *Proceedings of the 1977 Winter Simulation Conference*, IEEE, 187-189.

Carson, J.S. and A.M. Law (1977), "Conservation Equations and Variance Reduction in Queueing Simulations," *Op. Res.*, 28, 3, 535-546.

Law, A.M. (1975), "Efficient Estimators for Simulated Queueing Systems," *Mgt. Sci.*, 22, 1, 30-41.

Law and Kelton (1982), pp. 361-363.

Minh, D.L. and R.M. Sorli (1983), "Simulating the GI/G/1 Queue in Heavy Traffic," *Op. Res.*, 31, 5, 966-971.

6.10 Minimum Variance Estimator (MVE)

In some situations there exists a minimum variance estimator for θ given a fixed output sequence Y. This problem has been studied in mathematical statistics.

Bickel, P.J. and K.A. Doksum (1977), *Mathematical Statistics, Basic Ideas and Selected Topics*, Holden-Day, San Francisco.

6.11 More Observations (MO)

For any reasonable estimator the variance decreases as the number of observations increases (see (1)). If the variance is unacceptably high and there is additional computing budget, then the simplest VRT is MO. Unfortunately, the rate of reduction is generally slow (O(n)).

6.12 Postratifying the Sample (PSTRAT)

See section 4.4.

Cochran, W.G. (1977), *Sampling Techniques*, Wiley, NY, pp. 134-135.

Kleijnen, J.P.C. (1974), pp. 116-121.

Nelson, B.L. and B.W. Schmeiser (1985), "Decomposition of Some Well-Known Variance Reduction Techniques," *J. Statist. Comput. Simul.*, forthcoming.

6.13 Stratified Sampling (STRAT)

See section 4.3.

Bratley, Fox and Schrage (1983), pp. 61-63.

Cochran, W.G. (1977), *Sampling Techniques*, Wiley, NY, Chaps. 5, 5A.

Kleijnen (1974), pp. 110-133.

6.13.1 Splitting (SPLT)

A dynamic version of STRAT that prescribes multiple observations of Y when X falls in certain strata, rather than allocating observations of X itself. Useful when the output series is not independent, and for simulation of rare events.

Hopmans, A.C.M. and J.P.C. Kleijnen (1979), "Importance Sampling in Systems Simulation: A Practical Failure?", *Math. Comput. Simul XXI*, 2, 209-220.

Kahn, H. (1956), "Use of Different Monte Carlo Sampling Techniques," in *Symposium on Monte Carlo Methods* (H. Meyer, ed.), Wiley, NY, 146-190.

Kioussis, L.C. and D.R. Miller (1983), "An Importance Sampling Scheme for Simulating the Degradation and Failure of Complex Systems During Finite Missions," *Proceedings of the 1983 Winter Simulation Conference*, IEEE, 631-639.

McGrath, E. and D.C. Irving (1973), "Techniques for Efficient Monte Carlo Simulation, Vol. III, Variance Reduction," ORLC Report SAI-72-509-LJ.

7. Using Pilot Runs

Results from pilot simulation experiments can be used to acquire prior knowledge needed to answer the questions in section 5. Since a minimum of some debugging runs have to be made in any simulation study, this section will list some ways that the results of these runs, or pilot runs made expressly to gain prior knowledge, can be used.

One common debugging technique is to test the program under extreme conditions. For example, fixing some variables at their maximum or minimum values. The relative contribution of different simulation inputs to the outputs can be investigated in this way. The inputs that make the largest contribution are often the best auxiliary information. Fixing the values of some inputs can also make it possible to determine if the outputs are monotone in the inputs. Monotonicity facilitates VRTs based on dependence induction (DI).

Once the simulation is running, the results of pilot runs can be used to determine statistical properties of the outputs. For example, stepwise regression of outputs on inputs can determine which inputs are strongly correlated with which outputs, and

thus provide useful auxiliary information. As a side benefit, the regression coefficients can later provide estimates of the CV multipliers (see the references).

The potential effectiveness of STRAT can be evaluated by doing PSTRAT, which does not require fixing the sampling plan in advance. The empirical distribution of an output provides information needed for VRTs that bias sampling toward areas that contribute most to the variance.

8. Conclusions

It is interesting to note that there are more empty cells in Figure 3 than there are cells containing VRTs. This is partially due to excluding some of the more complex, and less frequently used, VRTs. However, there are some cells that contain no VRTs known to us. An open research question is why some cells are empty, and if they suggest new VRTs waiting to be discovered.

The questions that constitute the algorithm in section 5 are based on our assessment of the types of prior knowledge needed to invoke known results that ensure a VRT, formed from one or more of the six classes, will be effective. Others can, and are encouraged, to add questions to the list. The advantage of this approach, as opposed to a one-at-a-time search of VRTs, is that it identifies the available prior knowledge in terms of the six classes of transformations first, then yields all VRTs that make use of that knowledge. A one-at-a-time investigation of VRTs produces more dead ends and may overlook some useful prior knowledge.

Although a practitioner with indepth knowledge of variance reduction can likely decide on an appropriate VRT more quickly without the algorithm, the approach proposed here is the kind needed for automated variance reduction. We believe that only by incorporating automated variance reduction into general purpose simulation packages will all of the potential benefits of variance reduction be realized.

9. Acknowledgement

This research was partially supported by a University Seed-Grant from the Office of Research and Graduate Studies, The Ohio State University.

10. References

[1] Cochran, W.G., *Sampling Techniques*, Wiley, NY, 1977.

[2] Hammersley, J.M. and D.C. Handscomb, *Monte Carlo Methods*, Chapman and Hall, London, 1964.

[3] Nelson, B.L. and B.W. Schmeiser, "A Mathematical-Statistical Framework for Variance Reduction, Part I: Simulation Experiments," *Research Memorandum No. 84-4*, School of Industrial Engineering, Purdue University, West Lafayette, IN, 1984.

[4] Nelson, B.L. and B.W. Schmeiser, "A Mathematical-Statistical Framework for Variance Reduction, Part II: Classes of Transformations," *Research Memorandum No. 84-5*, School of Industrial Engineering, Purdue University, West Lafayette, IN, 1984.

[5] Bickel, B.J. and K.A. Doksum, *Mathematical Statistics, Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.

[6] Nelson, B.L. and B.W. Schmeiser, "Decomposition of Some Well-Known Variance Reduction Techniques," *J. Statist. Comput. Simul.*, forthcoming.

[7] Nelson, B.L., "A Perspective on Variance Reduction in Simulation Experiments," *Working Paper Series No. 1985-011*, Dept. of Industrial and Systems Engineering, The Ohio State University, Columbus, OH, 1985.

[8] Bratley, P., B.L. Fox and L.E. Schrage, *A Guide to Simulation*, Springer-Verlag, NY, 1983.

[9] Nelson, B.L., "On Control Variate Estimators," *Working Paper Series No. 1985-006*, Dept. of Industrial and Systems Engineering, The Ohio State University, Columbus, OH, 1985.

BARRY L. NELSON is an Assistant Professor in the Department of Industrial and Systems Engineering at the Ohio State University, and has a parttime appointment with the OSU Statistical Consulting Service. He received his Ph.D. in Industrial Engineering in 1983 from Purdue University. His research interests are the design and analysis of computer simulation experiments on models of stochastic systems.

Professional society affiliations include the Operations Research Society of America, The Institute of Management Sciences (TIMS), the American Statistical Association, the Society for Computer Simulation, and the Institute of Industrial Engineers. Dr. Nelson is assistant editor of the TIMS College on Simulation and Gaming Newsletter.

Dr. Barry L. Nelson
Department of Industrial and Systems Engineering
The Ohio State University
1971 Neil Avenue
Columbus, OH 43210
(614) 422-0610