COPPEAD/UFRJ

RELATÓRIO COPPEAD Nº 167

# SOURCES OF VARIATION OF SIMULATION ESTIMATES: AN EMPIRICAL STUDY*

Eduardo Saliby **

September 1986

* Trabalho apresentado na EUROPEAN CONFERENCE ON OPERATIONAL RESEARCH - EURO VIII, realizada em Lisboa em setembro de 1986.

** Professor Adjunto da COPPEAD/UFRJ - Brasil.

# SOURCES OF VARIATION OF SIMULATION ESTIMATES : AN EMPIRICAL STUDY

Keywords: Monte Carlo, Simulation, Variance reduction techniques.

Author: Eduardo Saliby, COPPEAD/UFRJ (Federal University of Rio de Janeiro), C. Postal 68514, CEP 21945, Rio de janeiro, Brasil.

## ABSTRACT

Simple random sampling, the most commonly used sampling procedure in simulation, can be viewed as a double source of randomness: one related to the set of input values and the other to their random sequence.

Following this approach and using factorial experiments, the contribution of both sources of variation was studied for three different simulation problems: a PERT network, a queueing system and an inventory system.

The results showed that part of the set effect contribution can usually be isolated and explained in terms of a regression model, named the linear response model, which takes sample deviations into account.

Apart from providing a new approach for interpreting the variability of simulation estimates, this study brought another result. It contributed to the proposal of a new and more efficient sampling procedure in simulation: descriptive sampling.

# INTRODUCTION

In simulation, simple random sampling is the standard sampling procedure used to represent the stochastic behaviour of the input random variables. It is used in order to achieve two different results:

- to generate sample values in correct proportion to the represented distribution,

- to reproduce a pattern of randomly sequenced values.

As a consequence of the use of simple random sampling, it follows that simulation estimates will vary between different runs. Thus, simulation is an imprecise procedure.

One approach for improving the precision of simulation estimates is through the use of variance reduction techniques (1,2,3). With proper use of such techniques, it is possible to produce more precise estimates without changing the run length.

However, it is not always obvious how the many different variance reduction techniques do their job; therefore, the choice of the most appropriate approach for a particular problem can be a rather difficult task.

The purpose of the present work is to evaluate the contribution of each of the two sources of variability associated with a random sampling process:

- the random selection of a set of values,
- the random selection of their permutation.

Following an experimental approach similar to the one already used by Ehrenfeld and Ben-Tuvia (4), a common feature was observed in most of the simulation studies: the individual contribution of the variability due to the set of values in use.

Moreover, it was possible to identify a common pattern of variability in random sampling estimates, described by a regression model, named the Linear Response Model.

The Linear Response Model explains the individual contribution of the set effect as a result of the observed deviations between input sample moments and the corresponding theoretical values.

Once the sources of variation of simulation estimates are known, two courses of action can be followed:

- a more efficient use of the variance reduction techniques,
- the removal of one source of variation, the set effect, which is, in fact, unnecessary in the simulation context. This can be done with the use of a newly proposed sampling procedure, named descriptive sampling (5,6).

# THE VARIABILITY OF SIMULATION ESTIMATES

Without loss of generality, consider a simple simulation problem with only one output random variable for which the distribution parameters

$$\Theta_j \ , \quad j = 1,\ldots,K$$

are to be estimated. Consider too, for simplicity, that there is only one input random variable, X .

As far as the present work is concerned, each simulation run will produce one estimate for each of the unknown parameters. As a consequence, a simulation run may be seen as the numerical evaluation of a complex function defined by

$$Y_j = F_j \, ( X_1 \, , \ldots , X_{NN} ) \, ,$$

where

$Y_j$ is an estimator for the unknown parameter,

$F_j$ is the simulation function, usually defined by means of a computer program,

$X_i$ , i = 1,...,NN , is the input sample for variable X.

Adopting the mean square error criterion and assuming that estimate Y is unbiased, it follows that Var (Y) should be kept to a minimum.

One obvious way to reduce this variance is to increase NN, the run length. Another approach is to use a suitable variance reduction technique in order to produce a lower variance estimator without increasing the sampling effort.

Up to now, the choice of the most appropriate variance reduction technique for a specific simulation problem has been nearly guesswork.

However, after studying the variability of simulation estimates, it will be possible to evaluate, for each problem, the amount of variance that can be reduced by controlling the sources of variability and, as a result, to make a more appropriate choice of the variance reduction techniques for each case.

Another important contribution from the study of the sources of variability of simulation estimates was to show that it is possible to eliminate one of them and, as a consequence, to produce lower variance estimates. This improvement can be achieved if a new sampling procedure, called descriptive sampling, is used instead of simple random sampling.

## A MODEL FOR THE VARIABILITY OF SIMULATION ESTIMATES

As far as simulation is concerned, simple random sampling may be seen as a double source of randomness: one related to the sampled set of values and the other to their random sequence.

For example, consider the simple random sample

U = ( .21, .15, .95, .83, .07 )

from a standard uniform distribution. Sample U may be defined by the set of ordered values,

SET = ( .07, .15, .21, .83, .95 )

and their random permutation, corresponding to the order which those values occur in the sample.

Thus, whenever a simple random sample is generated in a simulation, both features - set of values and their sequence - are randomly selected.

A simulation estimate, as a function of the input samples, is also determined by both a set of values and their sequence. Therefore, if simple random sampling is used, a simulation estimate will be affected by two and only two sources of variability:

- The set of input sample values, and
- Their sequence.

According to this approach, the variability of simulation estimates may be described by the following variance components model:

$$\sigma^2_{Y} = \sigma^2_{SET} + \sigma^2_{SEQ} + \sigma^2_{SET \times SEQ},$$

where,

$\sigma^2_{Y}$ : is the total variance for estimate Y.

$\sigma^2_{SET}$ : is the variance component of Y due only to the variability of the set of input values, thus excluding the sequence effect.

$\sigma^2_{SEQ}$ : is the variance component of Y due only to the variability of the sequence of input values, thus excluding the set effect.

$\sigma^2_{SET \times SEQ}$ : is the variance component of Y due to the interaction between both effects which cannot be explained by either of them in isolation.

It is worth noticing that no residual error is present in the above model. This happens because a simulation estimate is fully determined by a set of values and their sequence.

The central idea behind the use of this variance components model for the study of the variability of random sampling simulation estimates is that, for large input samples, some global features like their mean, standard deviation and autocorrelations should have a preponderant effect on the resulting estimates over the individual values.

To evaluate the contribution of each variance component to the total variance of Y, a factorial experiment was designed, varying both factors.

For each of three simple simulation problems - a PERT network, a M/M/I queue and an inventory system - $N_{SET}$ different input sample values were randomly taken and stored in ascending order; also, for each problem, $N_{SEQ}$ random permutations were generated, thus defining different sequences.

For each set of input values and for each sequence, a simulation run was carried out, giving a total of $N_{SET} \times N_{SEQ}$ runs for each studied problem. Each run, corresponding to a simple random sampling simulation, produced an estimate which was used as data for a two-way analysis of variance model in order to estimate the desired variance components.

## EMPIRICAL RESULTS

### PERT NETWORK

The first experiment to be reported concerns a simple PERT network already studied by Kleindorfer (7) and Sculli (8). As shown in figure 1, this network has eight activities, with independent and identically distributed durations (d) following a discrete uniform distribution defined by

$$f(d) = 0.2 \quad \text{for} \quad d = 1,\ldots,5 .$$

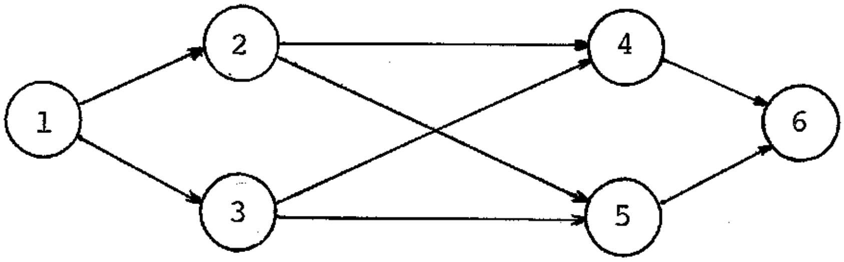


Figure 1. The simulated PERT network

Each run was defined by a set of N = 50 observations for the total project duration, DT, giving a pair of estimates associated with this response variable distribution:

$$\overline{DT} = \sum_{i=1}^{N} DT_i \;/\; N$$

and

$$S_{DT} = \sum_{i=1}^{N} \left[ (DT_i - \overline{DT})^2 / (N - 1) \right]^{1/2}$$

For this experiment, $N_{SET}$ = 10 different sets and $N_{SEQ}$ =10 different sequences were sampled, giving a total of 100 simulation runs. Table 1 presents the ANOVA table for estimate $\overline{DT}$, while Table 2 presents the ANOVA table for $S_{DT}$.

Table 1. ANOVA Table for estimate $\overline{DT}$

| Source of Variation | Sum of Squares | Degrees of freedom | Mean Square | F Value |
|---|---|---|---|---|
| Between sets | 2.994 | 9 | .333 | 12.95 |
| Between sequences | .387 | 9 | .043 | 1.67 |
| Unexplained | 2.081 | 81 | .026 | – |
| Total | 5.462 | 99 | .055 | |

Table 2. ANOVA Table for estimate $S_{DT}$

| Source of Variation | Sum of Squares | Degrees of freedom | Mean Square | F Value |
|---|---|---|---|---|
| Between sets | .387 | 9 | .043 | 1.52 |
| Between sequences | .365 | 9 | .041 | 1.44 |
| Unexplained | 2.276 | 81 | .028 | – |
| Total | 3.027 | 99 | .031 | |

For this problem, the results showed that the set effect had a strong individual influence over the variability of $\overline{DT}$, but not for $S_{DT}$. It also became clear that the contribution of the sequence effect could not be isolated.

After pooling the sum of squares due to different sequences with that due to unexplained variation, the variance components and their relative contribution for $\overline{DT}$ were estimated and displayed in Table 3.

Table 3. Variance components and their relative contribution for estimate $\overline{DT}$.

| Source of variation | Absolute Contribution | Relative Contribution |
|---|---|---|
| Set effect | .0306 | 53 % |
| Sequence effect | – | – |
| Interaction effect | .0274 | 47 % |
| Total | .0580 | 100 % |

Looking at the results, it can be said that, when using simple random sampling, 53 % of the total variability of estimate $\overline{DT}$ is only due to the set variation; the remaining 47 % is due to the interaction between both effects.

## QUEUEING PROBLEM

The second simulation problem studied concerned a M/M/I queue with two estimates produced for each run:

$\overline{SIZE}$ : the mean queue size, and

$S_{SIZE}$ : the standard deviation for the queue size.

Table 4 presents the relative contribution of each variance component for estimate $\overline{SIZE}$ for 6 different traffic intensity values. Table 5 presents the corresponding values for the estimate $S_{SIZE}$.

Table 4. Relative contribution of each source of variation for 6 different M/M/I queue simulations for estimate $\overline{SIZE}$.

| Traffic intensity | Relative contribution: Set | Sequence | Interaction |
|---|---|---|---|
| .15 | 96.3 | 3.4 | 0.2 |
| .30 | 86.1 | 13.4 | 0.5 |
| .45 | 78.8 | 20.5 | 0.7 |
| .60 | 69.1 | 29.5 | 1.4 |
| .75 | 62.9 | 32.5 | 4.6 |
| .90 | 57.0 | 30.8 | 12.2 |

Table 5. Relative contribution of each source of variation for 6 different M/M/I queue simulations for estimate $S_{SIZE}$.

| Traffic intensity | Relative contribution: Set | Sequence | Interaction |
|---|---|---|---|
| .15 | 76.9 | 18.5 | 4.6 |
| .30 | 50.0 | 47.8 | 2.2 |
| .45 | 47.3 | 51.0 | 1.7 |
| .60 | 41.0 | 56.6 | 2.4 |
| .75 | 33.3 | 59.8 | 6.9 |
| .90 | 32.6 | 59.4 | 8.0 |

Concerning the results from Tables 4 and 5, it is worth noting that:

a) The individual contribution of the set effect was always high for both estimates, although higher for $\overline{SIZE}$.

b) The individual contribution of the sequence effect was also high for both estimates, but higher for $S_{SIZE}$.

c) For both responses, the relative contribution of the set effect decreases with the traffic intensity, unlike the sequence effect contribution which increases with the traffic intensity.

d) For both responses, the amount of unexplained variation increases with the traffic intensity.

In summary, it was possible to explain a substantial amount of the variability associated with the two M/M/I queue estimates, in terms of the individual effects contribution. Once known this property, it will be possible to make better use of variance reduction techniques.

## INVENTORY SYSTEM

The third set of results concerns an inventory system, a problem already studied by Ignall (9) and Naylor and Hunter (10). It consists of a reorder point inventory system, with a fixed order quantity and a periodic review (daily); the demand is probabilistic, while the lead time is constant.

Costs associated with storage, placing an order and shortage are computed daily. At the end of each period run of N = 1000 days, the mean and standard deviation of the daily cost were computed.

Three different system configurations, defined by the operating parameters

Q , the fixed order quantity, and
ROP , the reorder point,

were studied, producing the results shown in Table 6.

Table 6. Relative contribution of each source of variation for 3 different inventory problem configurations.

| System configuration | Estimate | Relative contribution (%) SET | SEQUENCE | INTERACTION |
|---|---|---|---|---|
| A) ROP = 50 , Q = 150 | $\overline{COST}$ | 49 | – | 51 |
| | $S_{COST}$ | 43 | | 57 |
| B) ROP = 75 , Q = 200 | $\overline{COST}$ | 29 | – | 71 |
| | $S_{COST}$ | 35 | | 65 |
| C) ROP = 40 , Q = 150 | $\overline{COST}$ | 45 | – | 55 |
| | $S_{COST}$ | 33 | | 67 |

Here also, the set effect had an important individual contribution to the variance of simple random sampling estimates, while the individual sequence effect contribution was not statistically distinguished.

## UNDERSTANDING THE SET EFFECT VARIABILITY

Although only three simulation problems were studied, the individual contribution of the set effect to the total variability was shown to be a common feature in simulation studies. On the other hand, the sequence effect individual contribution was shown to be highly problem dependent and not so often detectable.

The commonly observed set effect contribution does suggest that simulation estimates could follow a standard pattern of variation depending upon some global set features like the input sample mean and standard deviation.

Following this idea, it was possible to explain most of the set effect contribution by means of a regression model, named the Linear Response Model, or simply the LRM.

The LRM assumes the existence of a relationship between the input sample moments, usually the mean and standard deviation, and the corresponding estimates from each simulation run. In the simplest case with only one sampled input variable, the LRM would have the form

$$Y_j = \Theta_j + a(\bar{X} - \mu_X) + b(S_X - \sigma_X) + \varepsilon_j ,$$

where,

$Y_j$, is the simulation estimate for the unknown parameter $\Theta_j$,

$\bar{X}$ and $\mu_X$ are, respectively, the sample mean and the distribution mean for variable X.

$S_X$ and $\sigma_X$ are, respectively, the sample standard deviation and the distribution standard deviation for variable X.

Each simulation run, producing one estimate for each parameter under study, provides an observation for the LRM regression equation. By conducting a set of N independent runs based on the use of simple random sampling, the regression parameters can be estimated, in particular the R2 coefficient measuring the amount of variability on Y explained by the observed variability on the input sample moments.

For a subset of the three simulation problems already studied, Table 7 presents the R2 coefficient from the LRM analysis together with the corresponding set effect relative contribution previously estimated. A close agreement between the two values is clear.

Table 7 . Amount of variability explained by the LRM and the corresponding set effect relative contribution.

| Problem | Estimate | R2 (%) | Set effect contribution (%) |
|---|---|---|---|
| PERT network | $\overline{DT}$ | 57 | 53 |
| | $S_{DT}$ | 9 | – |
| M/M/I Queue | $\overline{SIZE}$ | 62 | 63 |
| | $S_{SIZE}$ | 29 | 33 |
| Inventory System | $\overline{COST}$ | 48 | 49 |
| | $S_{COST}$ | 36 | 43 |

The results from Table 7 allow us to affirm that the set effect variabiliby is usually explained by the observed deviations between the input sample mean and standard deviation and the corresponding theoretical values.

Therefore, an easier way of evaluating the set effect contribution is to carry out a group of independent random sampling simulation runs and, after recording the estimates and the input samples means and standard deviations, to perform a regression analysis to estimate the R2 coefficient.

As a final remark concerning the LRM, it is worth noting the similarities between this model and the use of control variates (3), a variance reduction technique. There is, however, an important difference between the two approaches:

The LRM, unlike control variates, assumes a causal relationship between the input sample parameters and the corresponding estimates. Therefore, the LRM is a more informative and useful framework for the analysis and interpretation of simulation estimates.

The main idea behind the use of the LRM is that, instead of the specified model, a simulation estimate refers to a slightly different system configuration, defined not by the already known moments for the input distributions, but by the observed sample moments.

As such deviations are expected to be of small magnitude, a linear approximation is usually enough to explain their effect over the corresponding estimates, as stated in the LRM.

## CONCLUSIONS

The results presented in this study point to two different courses of action to follow in order to improve the precision of simulation estimates:

(1) The first suggested approach concerns a better use of the variance reduction techniques. Now, since more is known about the sources of variability of simulation estimates, the role of each technique should become easier to understand, as for example, in the case of antithetic variates ( 5 ). As a consequence, a more efficient use of such techniques is now possible. In such a case, the use of control variates, following the LRM structure, is strongly recommended.

(2) The second suggested approach is a less traditional one. Here, instead of reducing the variability of each source of variation, we are suggesting that, whenever the input sample size is known in advance, the set effect variability can be completely eliminated, without introducing any significant bias into the estimates.

As will be presented in a forthcoming work, the only source of randomness that it is really necessary to includ in a simulation study concerns the sequence effect variability, while the set effect can be excluded.

A full control over the set effect variability can be achieved through the use of descriptive sampling instead of simple random sampling in simulation. Descriptive sampling is a method based on a deterministic selection of the set of input values in order to fit, as closely as possible, the desired sampled distribution. Their sequence, however, is left to vary and is defined by a random permutation of the set values.

Using descriptive sampling, the variability of simulation estimates will be reduced to only the sequence effect contribution. Any variance reduction technique which controls the set effect variability will now become useless, unless it also controls the sequence effect. As shown in Saliby ( 5 ), antithetic variates are typically a case where only the set effect is controlled and, hence, has no more use if descriptive sampling is adopted as a standard procedure in simulation.

## REFERENCES


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

2) James, B. A. P. (1985), "Variance Reduction Techniques". J. Opl Res. Soc., vol 36, 525-530.

3) Kleijnen, J. P. C. (1974), "Statistical Techniques in Simulation, Part I. Marcel Dekker, New York.

4) Ehrenfeld, S. and Ben-Tuvia, S. (1962), "The Efficiency of Statistical Simulation Procedures". Technometrics, vol 4, 257-275.

5) Saliby, E. (1980), "A Reappraisal of Some Simulation Fundamentals". Ph.D. Thesis, University of Lancaster.

6) Pidd, M. (1984), "Computer Simulation in Management Sciences". Wiley, London.

7) Kleindorfer, G. B. (1971), "Bounding Distributions for Stochastic Acyclic Networks". Opns Res., vol 19, 1586-1601.

8) Sculli, D. (1983), "The Completion Time of PERT Networks". J. Opl Res. Soc., vol 34, 155-158.

9) Ignall, E. J. (1972), "On Experimental Designs for Computer Simulation Experiments". Management Sci., vol 18, 384-388.

10) Naylor, T. H. and Hunter, J. S. (1969), "Experimental Design". In The Design of Computer Simulation Experiments, T. H. Naylor (ed.), Duke University Press, Durham N.C., 39-58.