Proceedings of the 1997 Winter Simulation Conference
ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

DESCRIPTIVE SAMPLING: AN IMPROVEMENT OVER LATIN HYPERCUBE SAMPLING

Eduardo Saliby

COPPEAD/UFRJ
Universidade Federal do Rio de Janeiro
Caixa Postal 68514
21949-900 - Rio de Janeiro - RJ, BRAZIL

ABSTRACT

Descriptive Sampling (DS), a Monte Carlo sampling technique based on a deterministic selection of the input values and their random permutation, represents a deep conceptual change on how to carry out a Monte Carlo application. Abandoning the paradigm that a random selection of sample values would be necessary in order to describe random behavior, DS is a rather polemical idea. An interesting issue related to DS are the similarities between it and Latin Hypercube Sampling (LHS) to be discussed in this paper. After a brief description of both methods, it is shown how close DS and LHS are. As such, DS can be seen as a limiting case of LHS and also as an improvement over it. An experiment and a set of empirical results illustrating the relationship between DS and LHS are also presented.

1 INTRODUCTION

Proposed as an alternative approach to Monte Carlo simulation, Descriptive Sampling - DS - (Saliby,1980 and 1990) is based on a fully deterministic selection of the input sample values and their random permutation. As such, DS avoids set variability of the input values and leads to more precise simulation estimates. However, since DS is based on a non-random selection of input sample values, it also represents an important conceptual change on how to sample in any Monte Carlo application. The DS proposal questions the paradigm that a random selection of sample values would be necessary in order to describe random behavior, stating that a fully deterministic selection of such values would be more appropriate. The usefulness of DS was confirmed by several comparisons already carried out (Saliby, 1989 and 1990), showing that the estimates it produces are, in principle, unbiased and with lower variance than the classical use of Simple Random Sampling.

DS is justified by the fact that in any Monte Carlo application, the sampled distribution must be assumed known in advance. As such, the sampling context is not inferential, where one wants to acquire information about a population but descriptive, where the purpose is just to describe an information already known (the assumed probability distribution).

In spite of some theoretical results already available, DS still lacks an adequate theoretical development. One way towards this goal is to back this development over the theory already available for Latin Hypercube Sampling (LHS) which, as will be shown here, is rather similar to DS and, as such, presents the same kind of challenges for the supporting theory. Like DS, LHS is based on a highly controlled selection of the input values and their random permutation. The unique difference between both methods is that, unlike DS, LHS still preserves a minimum random variability on the sample values selection, which is completely eliminated with DS.

2 DESCRIPTIVE SAMPLING

This section presents a basic introduction to the idea of DS. A more extensive description is given in papers by Saliby (1990) and Saliby and Paul (1993). Descriptive sampling was proposed in order to avoid the set variability in simulation studies (Saliby, 1980). When using the standard Simple Random Sampling (SRS) or Monte Carlo approach, two kinds of variation are present in a randomly generated sample - one related to the set of values and the other to their sequence. But, of these two kinds of variability, only the sequence variability is really inevitable, while the set variability is, according to the author, in fact unnecessary. Symbolically, the two sampling methods can be represented as

Simple random sampling
=
random set x random sequence,

whilst

Descriptive sampling
=
deterministic set x random sequence.

The only additional requirement to use DS instead of SRS is to know, in advance, the input sample size, which, as stressed in Saliby and Paul (1993), must be related to a full simulation run. Once the sample size is known, at least approximately, the set values are defined for each input random variable X_j, j=1,..,k, using the inverse transform method, so that

$$xd_{j,i} = F^{-1}[(i-0.5)/n_j)],\ i=1,..,n_j,\ j=1,..,k \quad (1)$$

where

$$F^{-1}(R),\ R \in (0,1)$$

is the inverse transform for the particular input distribution.

Although in some applications the sample size may be the same for all input variables, there are cases where different n_j applies for different inputs.

Also, when the inverse distribution is not available, numerical or functional approximations can be used, like, for example, the Ramberg and Schmeiser (1972) approximation for the inverse Normal distribution.

Completing the DS generation process, each of the k sets of input values is used in a random sequence in each simulation run. Now, unlike with SRS, set values are the same for all replicated runs in a simulation experiment. This random shuffling process is easily accomplished by sampling without replacement the descriptive set of values (Saliby,1990).

Undoubtedly, the main issue concerning DS is the fact that it follows from the assertion that, in any Monte Carlo application, instead of being draw at random, sample values should be carefully chosen in order to achieve the closest possible fit with the represented distribution. This follows from another assertion: that randomness in Monte Carlo sampling is essentially a sequence feature, thus not being improved by a random selection of the input sample values. Although surprising at first sight, the idea of controlling the input values is being widely used nowadays, almost to the same extent as imposed with DS. This is the case of Latin Hypercube Sampling, which turns out to be a very close idea to DS.

3 LATIN HYPERCUBE SAMPLING

A contemporary development to DS, Latin Hypercube Sampling (McKay et al. 1979) was suggested as a Variance Reduction Technique, but also seen as a screening technique, in which the selection of sample values is highly controlled, although still letting them to vary. The basis of LHS is a full stratification of the sampled distribution with a random selection inside each stratum. Like with DS, sample values are randomly shuffled among different variables.

Using LHS, an input sample will be also generated based on the inverse transform method, and given by

$$xh_{j,i} = F^{-1}[(i-1+R_i)/n_j)],\ i=1,..,n_j,\ j=1,..,k, \quad (2)$$

where R_i stands for an independent random uniform on [0,1] , i=1,.., n_j , and, like with DS, $F^{-1}(R)$, $R \in (0,1)$ is the inverse transform for the particular input distribution.

After the original paper (McKay et al. 1979) where it was proposed, LHS has been widely used both in engineering in what some authors call deterministic simulation for computer experimentation as well as in risk analysis. Both situations can be seen as terminating simulations in which a set of k input variables generates a set of r output variables. Among the main theoretical results about LHS, one is that LHS estimates are unbiased (McKay et al. 1979) and another that the estimates variance is asymptotically lower than with simple random sampling (Stein, 1987). Further developments were also presented by Iman and Conover (1980 and 1982), Owen (1992), Iman and Helton (1991) and Loh (1996).

4 CLOSENESS BETWEEN DS AND LHS

Since both DS and LHS are based on a random permutation of the input values, the only difference between both methods relies on how those values are selected inside each of the n stratum. As seen from (1) and (2) above, with both DS and LHS, a random sequence or permutation is defined for the input values. Assuming for simplicity that $n_j = n$, j=1,.., k, this permutation can be defined as

$$\mathbf{P} = (P_1, P_2, \ldots, P_k)$$

where $P_j = (p_{j1}, p_{j2}, \ldots, p_{jn})$ defines a random permutation of (1, 2, ... , n) for variable X_j, j=1,.., k.

Geometrically, **P** defines a particular choice of n k-dimensional minicubes of size n-k in the unitary k-hypercube in which each input value stratum appears once and only once, thus leading to a sort of latin design. Given **P** and using DS, the centre of each of the n minicubes is deterministically chosen, while, when using LHS for the same **P**, a point is still randomly draw inside each of the same set of n minicubes.

Being Y a general simulation estimate, YD a DS

estimate and YL a LHS estimate, conditioning Y on the (n!)k equally likely sequences for the k input values, it follows that

$$Var(Y) = Var_{SEQ}\{E[Y/SEQ]\} + E_{SEQ}\{Var(Y/SEQ)\},$$

or, simply,

$$Var(Y) = \sigma^2_{SEQ} + \sigma^2_R ,$$

where the first term (σ^2_{SEQ}) accounts for the variance component of Y due to the sequence variability, while the second term (σ^2_R) reflects the remaining variance component of Y due the set variability conditioned on the $(n!)^k$ sequences.

Now, based on several empirical studies and some theoretical results, we observed that:

- no matter the sampling method (SRS, LHS or DS), σ^2_{SEQ} is always of order $o(n^{-1})$, while the second component (σ^2_R) will depend upon the sampling method;
- for the standard Simple Random Sampling (SRS) method, σ^2_R is also of order $o(n^{-1})$ and usually the dominating term;
- using LHS, the residual term σ^2_R will be of order $o(n^{-a})$, with $a \geq 2$. This implies that, for LHS, σ^2_R will decrease faster than the first term, so that there will be a sample size n after which the first term will dominate the estimate variance $Var(Y_L)$. This is equivalent to say that, for LHS,

$$\sigma^2_R/Var(Y_L) \rightarrow 0 \text{ , as n increases;}$$

- since there is no set variability left when using DS ($\sigma^2_R = 0$),

$$Var(Y_D) = Var_{SEQ}\{E[Y_D/SEQ]\} = Var\{Y_D(SEQ);$$

- even for moderate n values,

$$E[Y_L/SEQ] \simeq Y_D(SEQ),$$

so that

$$Var_{SEQ}\{E[Y_L/SEQ]\} \simeq Var(Y_D);$$

- finally, as n increases, LHS becomes practically equivalent to DS.

Some of those properties are illustrated by the following experiment and results.

5 EXPERIMENT AND RESULTS

Although we could have used any simulation problem to compare DS with LHS, we preferred to use a simple problem for such comparison: the study of the response variable

$$D = (X^2 + Y^2)^{1/2},$$

where both X and Y are independent standard normal distributed random variables. Our purpose was to estimate μ_D. Incidentally, we have that μ_D =1.2533 and that σ_D = 0.4292.

Now, since both DS and LHS are based on the same kind of stratification of the input distributions and on the use of the input values in a random order, we first sampled a random permutation $\mathbf{P} = (P_1, P_2, \ldots, P_k)$ for scrambling the input values. Then, for each permutation, we carried out a Descriptive Sampling run and NL LHS runs, so that all LHS runs were based on the same permutation **P** as the DS run. This procedure was repeated NP times, once for each different random permutation. As such, we were able to isolate the sequence generation from the set generation in LHS and also to have the DS run as a sort of control group.

In this experimental design, each sampled sequence (permutation **P**) represented a factor level randomly sampled, for which NL independent LHS runs were conducted. As such, we carried out a one factor (sequence) random effects experiment, from which we could estimate the variance components (σ^2_{SEQ} and σ^2_R) of Y_L. Thus, for each value of n in (10,30,50,100,250,500,1000) a full experiment based on NP = 1000 random permutations, NL=100 LHS runs and the corresponding DS control run were carried out for each of the NP permutations. A summary of the results for the testing problem is presented in table 1.

Table 1. Summary Results for the Testing Problem

n	$nVar(Y_D)$	$n\ Var(Y_L)$	$n\sigma^2_R$	$\sigma^2_R/\sigma^2_{SEQ}$
10	0.01110	0.01173	0.06342	5.4072
30	0.01149	0.01175	0.01595	1.3574
50	0.01131	0.01153	0.00853	0.7398
100	0.01162	0.01171	0.00350	0.2989
250	0.01203	0.01211	0.00112	0.0925
500	0.01160	0.01170	0.00046	0.0393
1000	0.01144	0.01149	0.00020	0.0174

From the above results, we notice that:

- $Var(Y_D)$ always remained very close to $Var(Y_L) = Var(E(Y_L/SEQ))$ and both were of order $o(n^{-1})$;

- As n increases, ratio $\sigma^2_R/\sigma^2_{SEQ}$ decreases. Thus, with LHS, the set variability contribution to $Var(Y_L)$ decreases with n, so that, as n gets larger, the sequence variability becomes the dominant variance component of Y_L. In other words, LHS turns out to be equivalent to DS as n increases!

Finally, in order to check the closeness between DS and LHS, we computed for each experiment the Pearson correlation coefficient between Y_D and the estimated value of $E(Y_L/SEQ)$. In this case, even for n values as low as n = 10, correlation coefficients were above 0.99 (NP=1000 cases), thus confirming that

$$E(Y_L/SEQ) \simeq Y_D(SEQ) .$$

6 CONCLUSIONS

Although a simple example was used to evaluate the variance components of LHS estimates, in terms of the sequence x set decomposition, the same kind of result is expected for any other simulation problem. Of course, the benefits from using DS or LHS may vary from problem to problem, but the gain achieved with DS will always establish the upper limit for the LHS gain.

As such, DS represents an improvement over LHS, being more efficient both in statistical terms as well as in computing terms, since it avoids the unnecessary step of randomly sample the set values.

REFERENCES

Iman, R. L., and W. J. Conover. 1980. Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment. *Communications in Statistics: Theory and Methods* A 9: 1749-1874.

Iman, R. L., and W. J. Conover. 1982. A Distribution-free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics* B 11: 311-334.

Iman, R. L., and J. C. Helton. 1991. The Repeatability of Uncertainty and Sensitivity Analyses for Complex Probabilistic Risk Assessments. *Risk Analysis* 11: 591-606.

Loh, W. L. 1996. On Latin Hypercube Sampling. *The Annals of Statistics* 24: 2058-2080.

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

Owen, A. B. 1992. A Central Limit Theorem for Latin Hypercube Sampling. *Journal of the Royal Statistical Society Ser. B* 54: 541-551.

Ramberg, J. S., and B. W. Schmeiser. 1972. An Approximate Method for Generating Symmetric Random Variables. *Communications of the ACM* 15: 987-990.

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

Saliby, E. 1989. *Rethinking Simulation: Descriptive Sampling*. Sao Paulo: Atlas/EDUFRJ. (In Portuguese).

Saliby, E. 1990. Descriptive Sampling: A Better Approach to Monte Carlo Simulation. *Journal of the Operational Research Society* 41: 1133-1142.

Saliby, E., and R. J. Paul. 1993. Implementing Descriptive Sampling in Three-Phase Discrete Event Simulation Models. *Journal of the Operational Research Society* 44: 147-160.

Stein, M. 1987. Large Sample Properties of Simulations Using Latin Hypercube Sampling. *Technometrics* 29: 143-151.

AUTHOR BIOGRAPHY

EDUARDO SALIBY is a full professor at COPPEAD/UFRJ, Graduate Business School, Federal University of Rio de Janeiro, Brazil. He received a B.S. degree in industrial engineering from the University of Sao Paulo in 1971, a M.S. degree in industrial engineering and operations research from the Federal University of Rio de Janeiro in 1974, and a Ph.D. in Operations Research from The University of Lancaster, UK, in 1980. His research interests are simulation methodology, with special emphasis on simulation sampling, simulation practice and simulation software. He is a member of SOBRAPO (Brazilian O. R. Society), ORS (UK) and SCS.