# Scalability on LHS Samples for Use in Uncertainty Analysis of Large Numerical Models

Barón, J.H. and Núñez Mac Leod, J.E.

Presentado en: International Conference on Safety and Reliability – ESREL ´99
Munich, Alemania, 13-17 setiembre 1999

# SCALABILITY ON LHS SAMPLES FOR USE IN UNCERTAINTY ANALYSIS OF LARGE NUMERICAL MODELS

Barón, J.H.
Nuclear Regulatory Authority and Institute CEDIAC, University of Cuyo

Núñez Mac Leod, J.E.
Institute CEDIAC, University of Cuyo

Argentina

## ABSTRACT

The present paper deals with the utilization of advanced sampling statistical methods to perform uncertainty and sensitivity analysis on numerical models. Such models may represent physical phenomena, logical structures (such as boolean expressions) or other systems, and various of their intrinsic parameters and/or input variables are usually treated as random variables simultaneously. In the present paper a simple method to scale-up LHS samples is presented, starting with a small sample and duplicating its size at each step, making it possible to use the already run numerical model results with the smaller sample. The method does not distort the statistical properties of the random variables and does not add any bias to the samples. The result is a significant reduction in numerical models running time can be achieved (by re-using the previously run samples), keeping all the advantages of LHS, until an acceptable representation level is achieved in the output variables.

## 1. INTRODUCTION

The uncertainty and sensitivity studies on the behavior on numerical models have become more and more relevant in recent years. These studies allow for the analysis of the model response to input variables as well as for internal parameters. The numerical models may represent physical problems, logical equations or other structures. Typically, a selected set of input variables or intrinsic parameters are treated simultaneously as random variables, which are properly sampled using Montecarlo, or Latin Hypercube Sampling (LHS) techniques. In this paper, a modification of LHS is proposed, that considers the Scalability of the samples (LHS-S or Scalable Latin Hypercube Sampling).

Additionally to the selection of variables to be treated, another problem appears on the determination of the proper sampling size. With no specific recommendations, the usual case is that the sample size is usually either sub- or over-dimensioned. This is not a big problem when generating the samples, but becomes a real one when running the numerical model, which may imply very large computing costs. A certain sample of size N which is considered adequate during the sampling stage, may be too small when performing the sensitiviy and uncertainty analysis on the output variables of the model, for example, to allow for distinction on the influence of each input variable on the model response. Being that the case, another larger sample should be used, without the possibility of using the already run samples of the smaller attempt. Both Montecarlo as well as LHS do not have an answer to this problem. However, the LHS-S has been designed to overcome this problem.

## 2. LATIN HYPERCUBE SAMPLING (LHS)

The Montecarlo sampling technique consists on the generation a sample of a random variable, choosing aleatory numbers over the accumulated probability distribution functions of that variable,

in order to obtain the corresponding variable values, which constitute the sample itself. This technique requires a quite large number or samples (or sample size) in order to get an adequate representation of the distribution function (typically thousands or tens of thousands).

Once the sample is generated, the numerical model is run once for each sample value, and the model output variables are then characterized from the statistical point of view, based on the principle that *the transformation of random variables is also a random variable.*

The whole process must be repeated for each input variable of interest, making the multi-variable study an extremely costly process, when the numerical models are moderately complex.

To reduce the needed sample size, Stratified Montecarlo methods were developed. These methods allow for a good representation using sampling techniques with one sample value for each sector (or strate) of a variable. Besides, to allow for the simultaneous multi-variable analysis, the Lating Hypercube Sampling (LHS) technique was developed (Iman & Shortencarrier, 1984). In LHS, besides the stratified generation of samples, a random coupling between the samples of different variables is performed.

The LHS method consists on the selection of the proper parameters and variables to be sampled, the assignation of probability distributions for each (that may be based on theoretical or experimental evidence), the subdivision of each probability distribution in a number of *a priori* defined equi-probable intervals, the generation of a random sample inside each interval and for each variable, and finally the random coupling between the input variables and parameters. As a result of the sampling method, a series of input vectors is obtained, being the number of vectors the number of sampling intervals, and being the size of each vector the number of variables to study.

When the input vectors are obtained, the numerical model is run, once for each input vector. This means that the numerical model is required to be run as many times as intervals are *a priori* assumed in the interval definition, and this that not depend on the number of variables that are sampled. Usually this technique allows for a reduction on the sample sizes on one or more orders of magnitude (depending of the model itself), in order to obtain a certain representativity, when compared to a classic Montecarlo technique.

Howevere, there is a problem, already indicated in the cited paper of Iman & Shortencarrier. The problem is that the representativity of the results can be observed only *a posteriori* of the numerical model runs. In case it is not satisfactory, a completely new sample of a larger size should be generated, by increasing the interval number. For the larger sample, the previously obtained numerical model results cannot be used, because they cause conflicts in the coupling stage, and distort the statiscal characteristics of the results.

This limitation of LHS becomes a very serious one when running complex models (i.e., that imply cpu-hours for each run), and this is the reason why a method that overcomes this limitation is useful.

## 3. SCALABLE SAMPLE GENERATION

The developed procedure is in its first stages similar to LHS. The probability distribution functions for each variable to be sampled are used. Over these distribution functions the ordinate axis is stratified in order to obtain equi-probable intervals, in disjoint adjacent sectors, and a random sampling is performed, generating one sample value for each interval. The corresponding variable values are obtained on the x-axis. Then the samples generated for all the variables of interest, are randomly coupled, checking the process by looking at the correlation factors between variables, which should be acceptably small. Two generations of random number have been used up to now, one for the sampling and other for the coupling.

This is the classic LHS approach, and then the numerical model is run for the input vectors. The model output variables are than studied from the statistical point of view.

In the case that this study is not satisfactory (i.e., that the statistical properties such as mean, variance, percentiles, etc, do not have a good representation), a new cycle is initiated, increasing the sample size. This is the initiation of the scaling process.

In the new cycle, the already sampled sectors are taken again and subdivided in adjacent disjoint equi-probable sectors (i.e., dividing each sector by two). A recognition of those sectors that do not have a sample value is performed, and for those new sample values are generated, randomly.

In Figure 1 (Values selected from each interval) the stratification and sampling process is indicated for a generic variable x, dividing the y-axis in four equi-probable intervals and generating four sample values (x1, x2, x3, x4). To scale-up the sample size, each interval is divided in two (Figure 2, New values selected for each non-sampled interval) and samples are generated in those intervals (x5, x6, x7, x8) where a previous sample value does not exist (allowed intervals in the right side figure).

Next in the process, the new sample values obtained for all the variables are randomly coupled, forbidding the combination with values that belong to the previous sample size. In such way, a new set of input vectors, the same size that the previous one, is obtained, and the numerical model is run again.

In Figure 2, the random coupling for two variables of the initial sample of size 4 is indicated. Initially, a set of four samples is obtained (1, 2, 3, 4). When scaling the process, four new samples are obtained (5, 6, 7, 8), which are also randomly paired and aggregated to the initial set, conforming a unique set of eight input vectors.

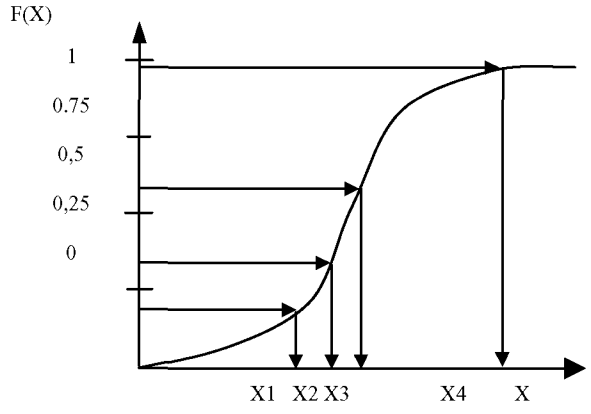



**Figure 1.** Values selected from each interval.

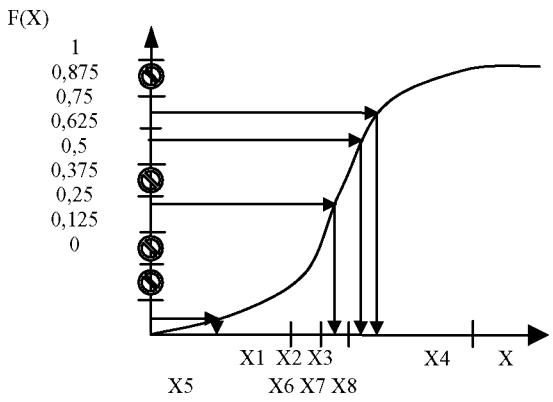



**Figure 2.** New values selected from each non-sampled interval.

The input vectors initially obtained and the obtained in the new cycle form the set of the actual input vectors, and the numerical model results initially obtained, merged with the presently obtained, constitute the actual output set. This output set is again analyzed from the statistical point of view and characterized. If found to be necessary, the sample size duplication process is repeated, until the representativity of the output set is satisfactory.

With this scaling process, the need for a large number of runs for the numerical model is reduced, by using the previous obtained results as part of the present set of results.

## 4. EXAMPLE AND COMPROBATION FOR TWO RANDOM VARIABLES

Next, an example on the use of the technique is presented. In this example, two random variables are combined. These variables are assumed to have gaussian distributions with mean 0 and variance 1.

Those probability distributions are sampled with the beforementioned technique, subdividing each accumulated probability distribution in ten equi-probable intervals. The sample values are indicated in Figure 3 for both variable 1 and 2. The random coupling for both variables is indicated also in Figure 3, with variable 1 in the x-axis and variable 2 in the y-axis.

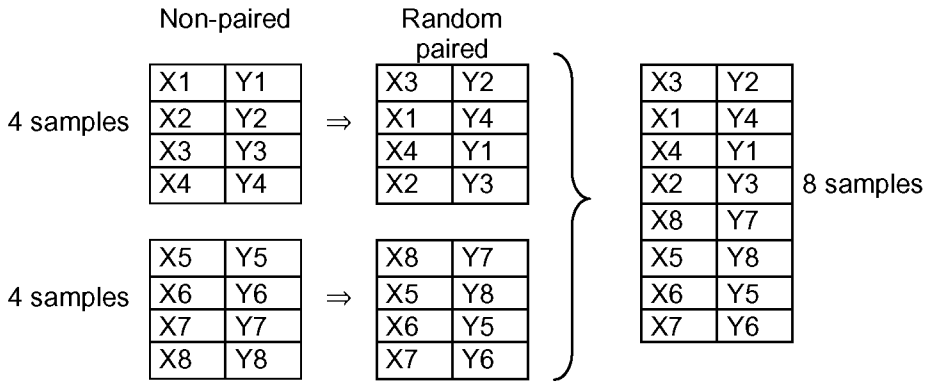



**Figure 3.** Pairing of samples.

If it is supposed that the representativity degree is not sufficient, each interval is divided in two, generating and coupling new samples for the empty intervals. The results obtained for variable 1 when using 20, 40, 80 and 160 intervals are indicated in Figure 5.

The results of the successive couplings between the two variables can be seen in Figure 6.

Next, on Table 1, the correlation values for each sample size are indicated

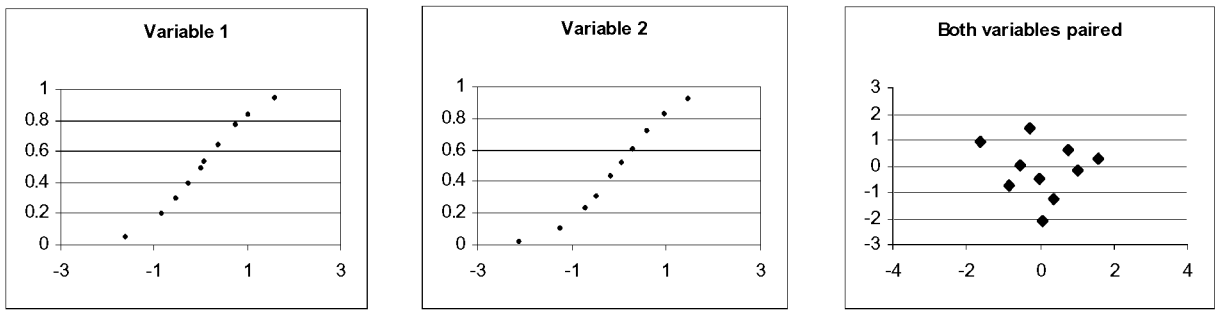
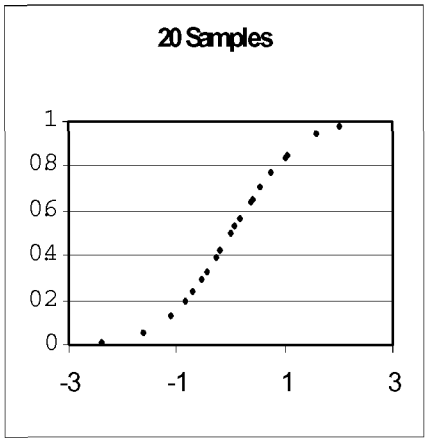
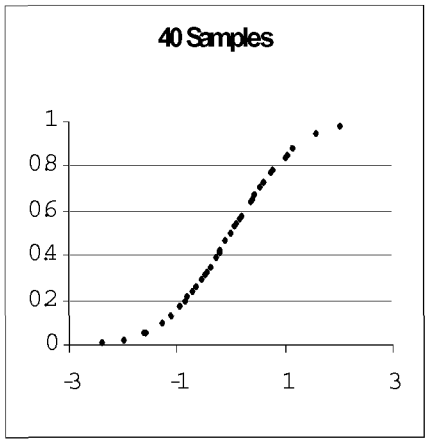
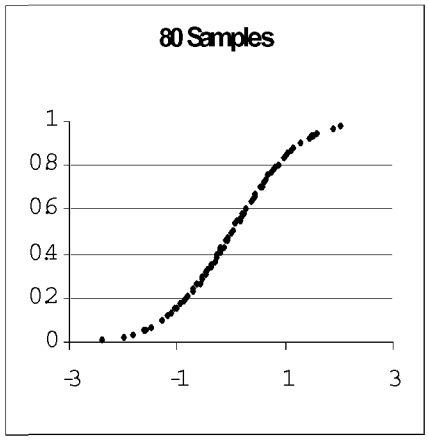



**Figure 4.** Samples values and paired for variables 1 and 2.



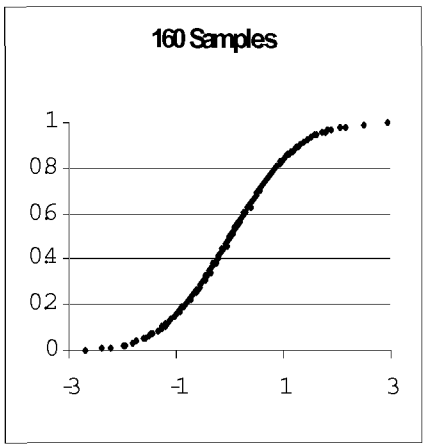
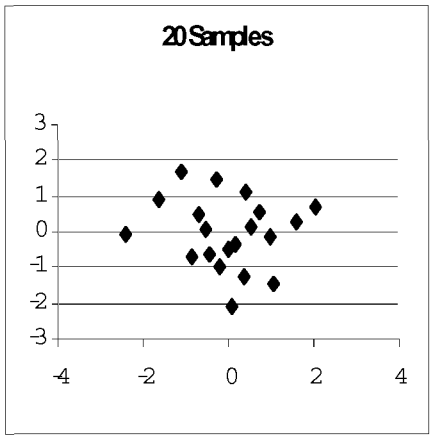
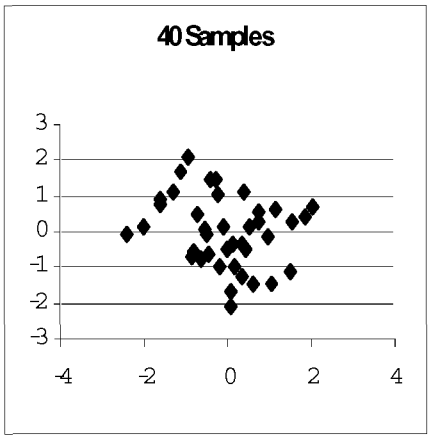
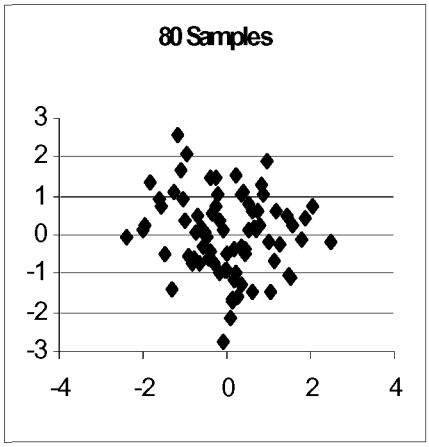




**Figure 5.** Value samples for variable 1.



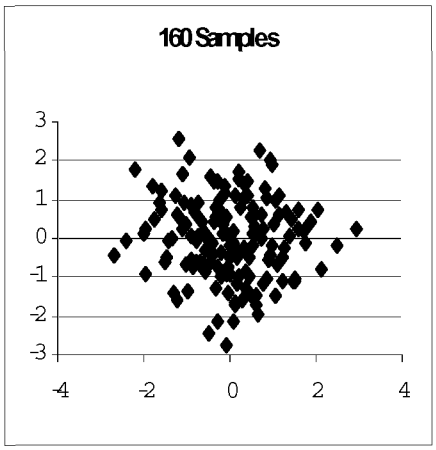




**Figure 6.** Paired for different samples size.

**Table 1.** Correlation Coefficients, covariance and variance for both variables.

| Samples | Correl. Coeff. | Covariance | Variance Var. 1 | Variance Var. 2 |
|---|---|---|---|---|
| 10 | -0.133 | -0.118 | 0.869 | 1.113 |
| 20 | -0.123 | -0.123 | 1.119 | 0.978 |
| 40 | -0.118 | -0.125 | 1.056 | 0.951 |
| 80 | -0.113 | -0.127 | 1.032 | 0.980 |
| 160 | -0.060 | -0.067 | 1.027 | 0.995 |

It is observed that the scaling process does not disturb the indicated probability distributions, and in all cases the correlation level is kept sufficiently low, in order to guarantee the random nature of the method. The covariance also indicates the sample independence.

## 5. CONCLUSIONS

A simple method to scale-up multi-variable samples generated with the LHS technique has been presented, starting with a small sample size and gradually increasing its size.

The method allows for the use of the sample vectors obtained in the previous step, in order to reduce the needs for numerical model runs. This is done without distorting the statistical properties of the samples.

This method is particularly valid for the cases when the numerical model is complex, and involves large running times.

The method does not impose any condition on the numerical model to analyze, which can be treated as a *black box* for the uncertainty and sensitivity analysis.

## REFERENCIAS

Iman, R. & M. Shortencarier 1984, *A FORTRAN 77 program and user's guide for the generation of Latin Hypercube and random samples for use with computers models*, NUREG/CR-3624, U.S. Nuclear Regulatory Commission.

Volver