Quasi-Monte Carlo Methods
Combating Complexity in Cost Risk Analysis

Blake Boswell
Booz Allen Hamilton

ISPA / SCEA Conference
Albuquerque, NM
June 2011

Table Of Contents

Introduction: Probabilistic Methods for Uncertainty Analysis

- Cost analyses rely on probabilistic numerical methods to estimate the impact of risk and uncertainty associated with systems' technical definitions and cost estimating methodologies Such methods involve
  - Modeling risk and uncertainty as probabilistic distributions
  - Applying iterative sampling techniques using Monte Carlo (MC) methods
  - Deriving risk and uncertainty adjusted statistical measures
- The accuracy of statistical measures resulting from probabilistic analyses is directly related to the number of samples considered
- A trade-off exits between accuracy of results and computational complexity
- To combat computational complexity in probabilistic numerical models, systematic sampling approaches have been designed with the goal of achieving accurate statistical measures while using fewer samples than traditional MC methods

Mote Carlo Methods: Quick Review

- MC methods are based on the analogy between probability and volume
  - Probability can be measured as the volume of a set of outcomes relative to that of a universe of possible outcomes

- MC methods calculate the volume of a set by interpreting the volume as probability
  - The **Law of Large Numbers** ensures that the MC approximation converges to the correct value as the number of draws increases
  - The **Central Limit Theorem** provides information about the likely magnitude of the error associated with a finite number of draws

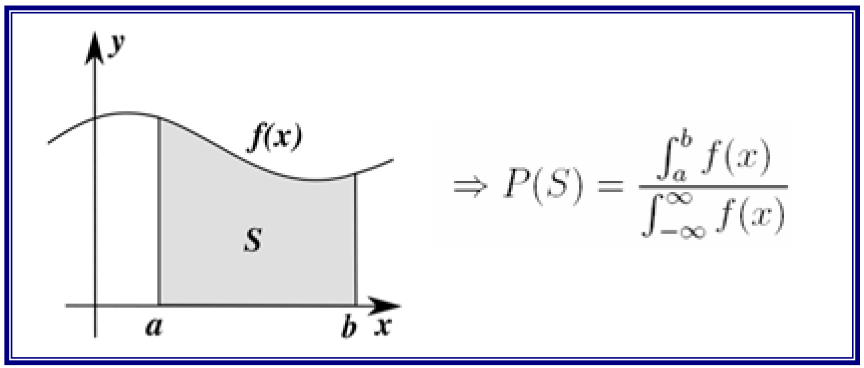

Figure 2: Analytical Derivation of Probability

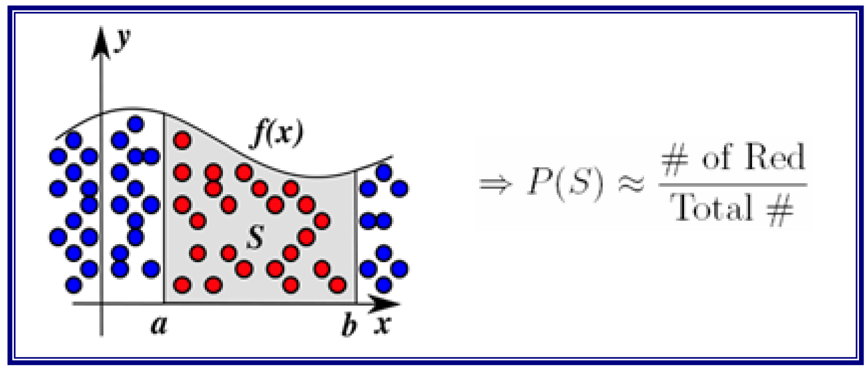

Figure 3: Monte Carlo Approximation of Probability

Monte Carlo Methods: Cost Uncertainty Simulation

- To measure the impact of cost uncertainty, MC methods are used to generate sets of trial costs which are used to quantify the non-deterministic properties inherent to the model. The standard results of such analyses are:
  - Probabilistic cost estimates or S-Curves
  - Descriptive statistics such as Mean, Mode, Variance, Standard Deviation, and CV
- A typical algorithm for performing MC cost uncertainty simulation is outlined in *Figure 4*:

For $i = 1$ to n [
    Generate RVs: $\{u_1, u_2, \cdots, u_d\} \in U(0,1)$;
    Map u_j to desired PDF: $x_j = \mathrm{CDF}_j^{-1}(u_j)$ for $j = 1...d$;
    Calculate model: $cost_i = f(x_1, x_2, \cdots x_d)$;
    Save $cost_i$]
Return $C = \{cost_1, cost_2, \cdots, cost_n\}$

Figure 4: Typical MC Cost Risk Simulation Algorithm

Monte Carlo Methods: Excel Implementation

- Excel and its associated programming language, Visual Basic for Applications (VBA), provide a powerful platform on which to conduct MC simulation
- The generation of U(0,1) RVs is performed by Random Number Generators (RNG)
  - The function "=Rand() " in Excel versions 2003 and later is a quality RNG in that it passes the DIEHARD tests as well as tests developed by the NIST
  - Open source RNG VBA routines are also available: *Mersenne Twister* [1], *HQRND* [2]
- Mapping U(0,1) RVs to PDFs is typically accomplished via the inverse CDF technique
  - Excel has CDF^{-1} for *Normal*, *Log Normal*, *Student's t*, *Chi-Squared*, *Beta*, *Gamma*
  - VBA programming allows for custom inverse CDF routines: *Triangular*, *Weibull*
- The Excel platform alone is insufficient for the modeling of interdependencies among stochastic elements
  - Open source VBA code enable modeling of RV interdependence in Excel

1. Ronchi, Mariano, http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/VERSIONS/BASIC/mt19937arVBcode.txt
2. ALGLIB, www.alglib.net

Monte Carlo Methods: Modeling RV Interdependence in Excel

- Modeling RV interdependence is achieved through rank correlation simulation (RCS)
  - RCS involves the re-ordering of RV draws to mimic relationships among stochastic elements

- Published methods for RCS rely on matrix algebra routines not available in Excel
  - Iman-Conover Method [1] is applicable to any distribution and sampling scheme
  - Matrix routines are available in open source VBA numerical libraries such as AlgLib [2]

- RCS is necessary for modeling stochastic interdependence and increases simulation complexity significantly

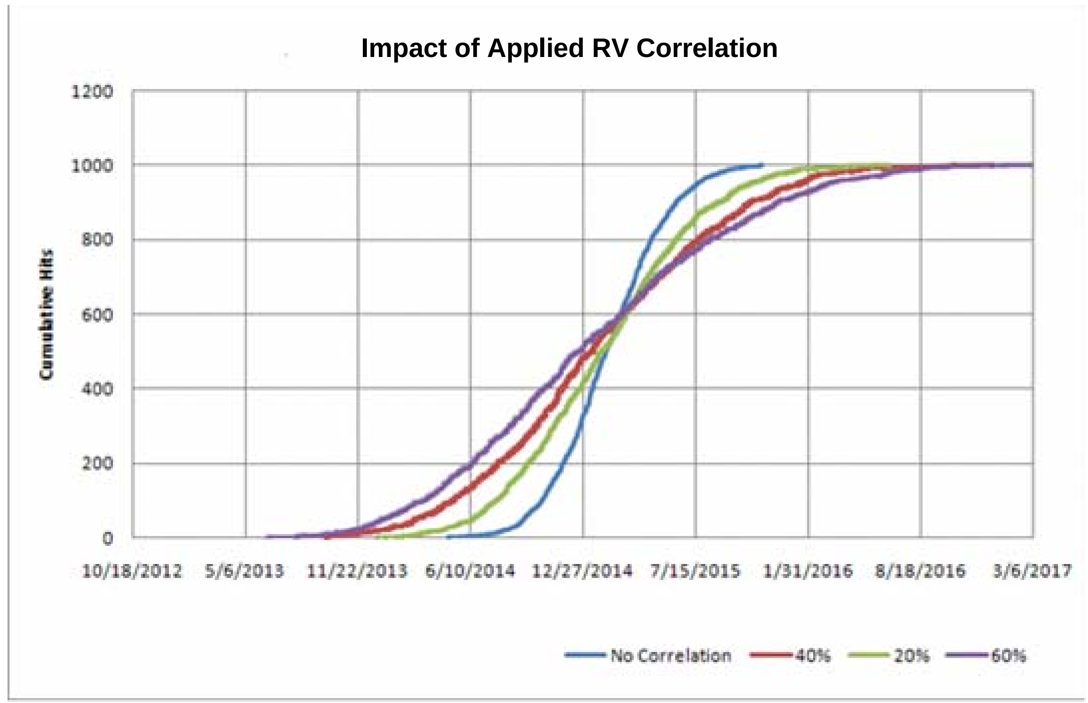

Figure 5: Impact of Applied RV Correlation

1. R. Iman, W. Conover , “A distribution-free approach to inducing rank correlation among input variables “, 1982.
2. ALGLIB, www.alglib.net

Booz | Allen | Hamilton

Reducing Simulation Complexity: Stratified Sampling Methods

- **Stratified Sampling** *refers broadly to any sampling mechanism that constrains the fraction of observations drawn from specific subsets of the sample space* [1]
  - *Figure 6* depicts a comparison of stratified and MC sampling of RVs~N(1000,400) with 1000 trials in one dimension
  - A major problem with stratified sampling is defining the strata and calculating their associated probabilities. As the dimensions of the sample space increase, stratified sampling becomes less tractable
- **Latin Hypercube Sampling (LHS)** is a compromise in that it relies on random pairings of stratified samples, thereby incorporating desirable elements from both MC and stratified sampling

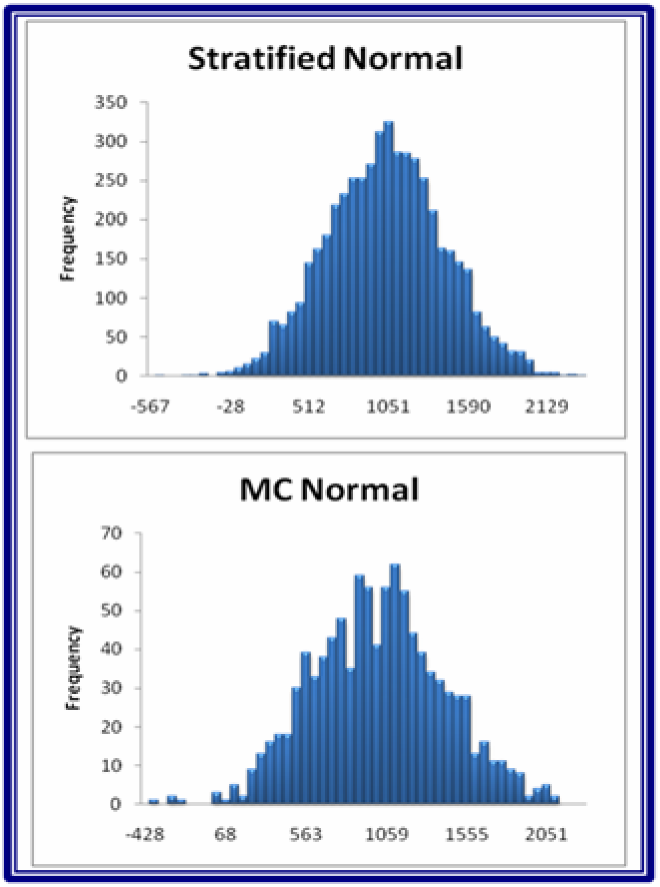

Figure 6: Stratified vs MC

1. P. Glasserman, "Monte Carlo Methods in Financial Engineering", 2003.

Booz | Allen | Hamilton

Reducing Simulation Complexity: Quasi-Monte Carlo Methods

- Quasi-Monte Carlo (QMC) methods differ from traditional MC in that they make no attempt to mimic randomness
  - QMC methods seek to increase sampling efficiency by generating points that are too evenly distributed to be random
  - The mathematical underpinnings of QMC are number theory and abstract algebra rather than probability and statistics
- QMC methods are prevalent in financial engineering for options pricing
  - Options prices are formulated as expectations, which are approximated by QMC as follows

$$E[f(U_1, U_2, \ldots U_d)] = \int_{[0,1)^d} f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

  - Where x_i are **deterministically chosen points** in the unit hypercube
- This process is analogous to uncertainty analysis in cost estimating
  - The expected value is not the only statistic of interest to cost estimators. The set of all $f(x_i)$ for $i = 1..n$ corresponds to the set *C* returned by the algorithm in *Figure 4*

Latin Hypercube Hammersley Sequence

- **Latin Hypercube Hammersley Sequences (LHHS)** is a relatively new hybrid sampling method proposed by Wang, Diwekar, and Gregoire Padro [1]

- **Hammersley Sequences (HS)**
  - HS is a low-discrepancy design for placing *n* points on a *d*-dimensional hypercube
  - HS points are formed through a process of representing integers as a set of binary fractions, *X*, and deriving corresponding values, *Y*, by reversing the binary digits of *X*
  - Exhibit good uniformity properties over *k*-dimensional hypercube for *k* >2

- **Latin Hypercube Sampling**
  - In LHS, the range of each input uncertainty variable is divided into *n* disjoint intervals of equal probability and one value is selected at random from each interval
  - The *n* values for variable 1 are **randomly paired** with the *n* values of variable 2. These *n* pairs are then combined with the *n* values for variable 3 and so on to form *n* *k*-tuplets
  - Has good uniformity in 1 dimension, but has low multidimensional uniformity due to random pairing

1. R. Wang, U. M. Diwekar, C. E. Gregoire Padro. "Efficient Sampling Techniques for Uncertainties in Risk Analysis", Environmental Progress, Vol 23, Issue 2, June 2004, pp. 141 – 157.

Latin Hypercube Hammersley Sequence

- LHHS incorporates desirable elements from both LHS and HS by generating samples via LHS to realize better 1 dimensional uniformity and pairing the samples using HS to achieve better multidimensional uniformity

- Pairing correlated LHS and LHHS samples is accomplished via the Iman-Conover method

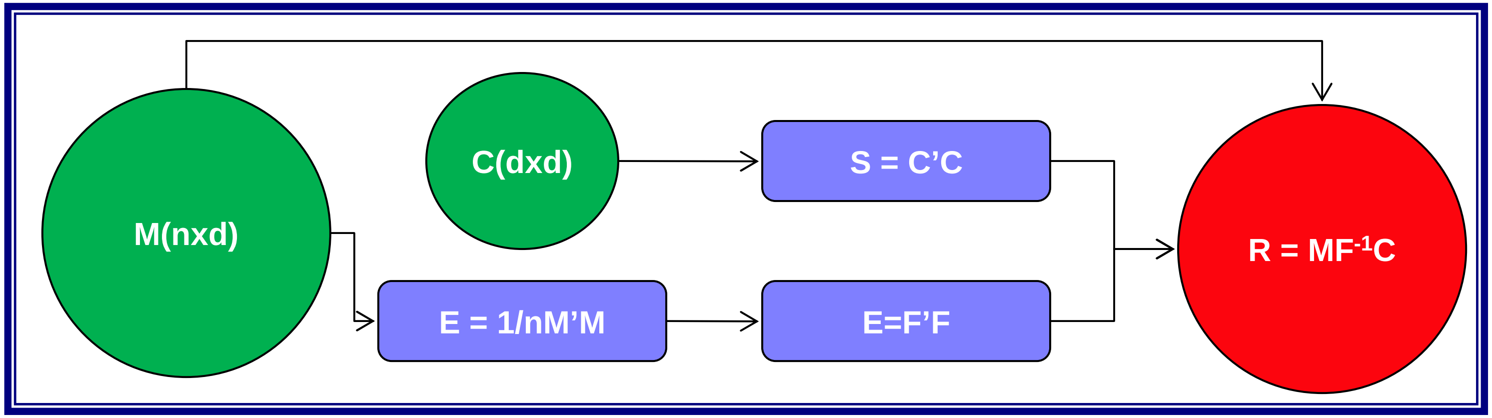

Figure 7: Inducing Correlation on Input Variables

- LHHS augments the conventional LHS pairing process through an alternative calculation of the *Score* matrix *M*(*nxd*)

Latin Hypercube Hammersley Sequence

- The Score matrix, *M*, in LHS pairing is a Van der Waerdan score matrix
  - A column vector, *A*, containing elements: $a_i = \Phi^{-1}\left(\frac{i}{n+1}\right) \quad i = 1, \ldots, n$ is generated and scaled to have a standard deviation 1
  - *A* is replicated *d* times to produce an *n* x *d* matrix, *M*, and each column is randomly shuffled
- LHHS pairing replaces the Van der Waerdan score matrix with the *n* x *d* HS matrix
- LHHS is a good candidate for uncertainty simulation in cost analysis
  - The underlying sampling technique, LHS, is prevalent within the field
  - Correlation is accomplished through a process similar to MC and LHS
  - Exhibits fast convergence and uniformity for low and high dimensions
- Implementing LHHS in Excel is possible through open source C++ routines
  - Routines can be compiled as Dynamic Link Libraries (DLLs) and called from user defined functions in excel

Sobol Sequence

- Sobol Sequence (SS) is a very efficient estimators for solving problems in low dimensions
  - Uses simple base 2 integers for generation of points in all dimensions
  - Operates by a complex reordering process that relies on the coefficients of irreducible primitive polynomials of modulo 2
- In general, SS are not efficient for high dimensional problems
  - Each dimension consist of a reordering of the same elements, creating strong interdependencies between dimensions
  - As dimensionality increases, SS loses its uniformity and exhibits clustering
- SS have proven tractable for certain high dimensional problems in financial engineering
  - Strategic assignment of sources of randomness to initial point coordinates has improved accuracy in *specific* high dimension problems
- SS are not great option for cost uncertainty analysis
  - Overly complex, lack robustness

LHHS: Excel Implementation

- Excel Demo

Contact Info

Blake Boswell
Consultant

Booz | Allen | Hamilton

Booz Allen Hamilton Inc.
8283 Greensboro Drive
McLean, VA 22102
Tel 202.412.7516
boswell_james@bah.com