Appendix A: Latin Hypercube Sampling[1]

[1]R. L. Iman (1999). "Latin Hypercube Sampling,"
Encyclopedia of Statistical Sciences, Update Volume 3, Wiley, NY, 408-411.

Latin Hypercube Sampling

Background. While serving as a consultant to Los Alamos National Laboratory during the summer of 1975, W. J. Conover of Texas Tech University was asked to develop a method for improving the efficiency of simple Monte Carlo used to characterize the uncertainty in inputs to computer models. Conover's work resulted in the development of a stratified Monte Carlo sampling method called Latin hypercube sampling [3].

At the time of its development in 1975, LHS was applied to some computer modeling applications at Sandia National Laboratories (Steck, Iman, and Dahlgren, [14]). The first journal article on LHS appeared in *Technometrics* (McKay, Conover, and Beckman, [10]). Software to implement the LHS strategy was developed in 1975 by R. L. Iman. This software was refined and was first released formally in 1980 (Iman, Davenport and Zeigler, [8]). A later revision (Iman and Shortencarier, [9]) has been the most widely distributed mainframe version of the program. Commercial vendors of LHS software include @RISK® and Crystal Ball®.

Latin hypercube sampling is used worldwide in computer modeling applications related to performing safety assessments for geologic isolation of radioactive waste and safety assessments for nuclear power plants (Helton [5], Helton et al [6], Campbell and Longsine [2]). It is also used in computer modeling for reliability analyses for manufacturing equipment, particularly in optimization schemes for repairable equipment (Painton and Campbell [13]). Other applications include the petroleum industry (MacDonald and Campbell [11]); transmission of HIV (Blower and Dowlatabadi [1}); and subsurface stormflow modeling (Gwo, Toran, Morris, and Wilson [4]).

Latin Hypercube Sampling. LHS uses a stratified sampling scheme to improve on the coverage of the input space. The stratification is accomplished by dividing the vertical axis on the graph of the distribution function of a random variable X_j into n nonoverlapping intervals of equal length, where n is the number of computer runs to be made. Through $F^{-1}(x)$, these n intervals divide the horizontal axis into n equi-probable, but not necessarily equal length, intervals. Thus, the x-axis has been stratified into n equi-probable and nonoverlapping intervals. The next step in the LHS scheme requires the random selection of a value within each of the n intervals on the vertical axis. When these values are mapped through $F^{-1}(x)$, exactly one value will be selected from each of the intervals previously defined on the horizontal axis. Let **X** be an n x k matrix whose jth column contains the LHS for the random variable X_j. A random process must be used to ensure a random ordering of the values within each column of this matrix. This mixing process serves to emulate the pairing of observations in a simple Monte Carlo process.

To fully appreciate the value of the underlying structure of LHS, it is helpful to be familiar with computer models used in actual applications. Such models are usually characterized by a large number of input variables (perhaps as many as a few hundred) and usually only a handful of these inputs are important for a given response. In addition, the model response is frequently multivariate and time dependent. If the input values were based on a factorial design, each level of each factor would be repeated many times. Moreover, the experimenter usually has a particular response in mind when constructing the factorial design and this design may be totally ineffective with multiple responses. On the other hand, LHS ensures that the entire range of each input variable is completely covered without regard to which single variable or combination of variables might dominate the computer model response(s). This means that a single sample will provide useful information when some input variable(s) dominate certain responses (or certain time intervals) while other input variables dominate other responses (or time intervals). By sampling over the entire range, each variable has the opportunity to show up as important, if it indeed is important. If an input variable is not important, then the method of sampling is of little or no concern. Also, as will be shown, the LHS is more efficient that simple random sampling in a large range of conditions.

Variability of Estimates from Random Sampling and LHS. Simple random sampling enjoys widespread use in simulation applications, so it is of interest to compare it with LHS. One way to do this is to compare the variability of estimates obtained from the two procedures. Let X_1 and X_2 be two independent input variables for a computer model. Using order statistics, the expected probabilistic coverage of the joint input space for X_1 and X_2 under random sampling for a sample of size n is given as:

$$\left(\frac{n-1}{n+1}\right)^2 \tag{1}$$

On the other hand, LHS requires that one value be selected from each of the extreme intervals (0, 1/*n*) and [(*n*-1)/n, 1]. Using the expected values from these intervals gives the expected probabilistic coverage of the joint input space for two variables under LHS for a sample of size *n* is given as

$$\left(\frac{n-1}{n}\right)^2 \tag{2}$$

For $n \geq 2$ the expression in Eq. 2 is always greater than the expression in Eq. 1, so LHS can be expected to provide better probabilistic coverage of the input space than a simple random sample. For example, when $n = 10$, Eq. 1 gives 66.9% and Eq. 2 gives 81%.

The probabilistic coverage of the input space provides a useful comparison of LHS and simple random sampling, but it does not provide a direct answer as to which sampling scheme might be preferred. One way to address the preference issue is to compare the variability of the estimates obtained from each of the sampling schemes as this provides a measure of efficiency, and thereby, cost effectiveness. For example, in many applications it is desired to estimate the mean of the output. Which sampling scheme provides the most efficient estimate for the mean?

Variability can be measured by using replicated LHS. That is, rather than using one LHS with $n = 100$, five replicates of 20 each could be used. The replication approach works very well if the model is complex. On the other hand, if the model is simple, it may be possible to calculate the variability analytically. The following simple linear model is used to provide a direct comparison of LHS with random sampling:

$$Y = \sum_{i=1}^{k} b_i X_i \tag{3}$$

where the X_i are independent and uniformly distributed on the interval (0,1). Under simple random sampling the variance of the estimator of the mean for the model in Eq. 3 is

$$V_{RS}\left(\overline{Y}\right) = \frac{1}{12n}\sum_{i=1}^{k} b_i^2 \tag{4}$$

Iman and Conover (Eq. 2.43) [7] have shown that under LHS the variance of the estimator of the mean for the model in Eq. 3 is

$$V_{LHS}\left(\overline{Y}\right) = \frac{1}{12n^3}\sum_{i=1}^{k} b_i^2 \tag{5}$$

Note that the estimator in Eq. 5 is a factor of $1/n^2$ smaller than the estimator in Eq. 4, that is

$$V_{LHS}\left(\overline{Y}\right) = \frac{V_{RS}\left(\overline{Y}\right)}{n^2} \tag{6}$$

This means that a value of $n = 1000$ in Eq. 4 provides the same variance as a value of $n = 10$ in Eq. 5, or the cost savings in sampling is reduced by a factor of 100 when using LHS rather than simple random sampling to estimate the mean for the model in Eq. 3.

Confidence Intervals for the Mean for Random Sampling and LHS. To illustrate the results in Eqs. 3 to 6, two random samples and two LHSs of size $n = 10$ have been selected for independent random variables uniformly distributed on the interval (0,1). The samples are given below.

A 95% confidence interval for the population mean of Y for the model in Eq.3 with $k = 2$ and $\beta_1 = \beta_2 = 1$ using the random samples is: $0.702 \leq E(Y) \leq 1.436$. Iman and Conover (Eqs. 2.44 and 2.45) [7] show that under LHS if all β_i ¤ 0, the largest possible value of the sample mean when evaluating the linear model in Eq. 3 is

Random Samples			Latin Hypercube Samples		
Observation	X_1	X_2	Observation	X_1	X_2
1	.164	.257	1	.270	.963
2	.549	.136	2	.372	.611
3	.595	.021	3	.148	.520
4	.351	.629	4	.712	.313
5	.831	.565	5	.574	.052
6	.847	.622	6	.437	.453
7	.890	.769	7	.963	.822
8	.231	.135	8	.820	.122
9	.816	.820	9	.003	.226
10	.938	.528	10	.628	.747

$$\max\left\{\overline{Y}\right\} = \frac{n+1}{2n}\sum_{i=1}^{k} b_i \tag{7}$$

and that the smallest possible value of the sample mean is

$$\min\left\{\overline{Y}\right\} = \frac{n-1}{2n}\sum_{i=1}^{k} b_i \tag{8}$$

Combining the results in Eqs. 7 and 8 produces the following interval for the mean of Y as

$$\frac{n}{n+1}\overline{Y} \le E(Y) \le \frac{n}{n-1}\overline{Y} \tag{9}$$

Substitution of the n = 10 values obtained for the two LHSs in the above table into the function $Y = X_1 + X_2$ gives a sample mean of .976 and Eq. 9 provides the following interval: $0.887 \le E(Y) \le 1.084$.

The interpretation of the confidence interval based on random sampling is that there is 95% confidence that the true value of the population mean (1) is contained in the interval. However, the interval based on LHS has a much stronger interpretation — the true value of the mean *is contained in the interval*. Another way of stating this result is that Eq. 9 provides a 100% confidence interval for the population mean.

This last statement is worth considering further. Equation 7 gives the maximum value of the sample mean as 1.10 for this example. Substitution of this maximum value into Eq. 9 produces the following interval: $1.000 \le E(Y) \le 1.222$, which contains the E(Y) = 1. In an analogous manner Eq. 8 gives the minimum value of the sample mean as 0.90 and based on this value Eq. 9 produces the following interval: $0.818 \le E(Y) \le 1.000$, which also contains the E(Y) = 1. In other words, it is not possible for LHS to produce a sample mean that allows the E(Y) to fall outside of the interval given in Eq. 9. Hence, the reason for using the expression 100% confidence.

Eqs. 7 and 8 give the absolute minimum and maximum values for the sample mean under LHS as 0.9 and 1.1, respectively. While these bounds are quite narrow, a computer simulation (200 samples of size n = 10) provides more realistic estimates of the actual minimum and maximum values for the sample mean as 0.964 and 1.030, respectively. That is, the bounds 0.9 and 1.1 are extremely conservative (much wider than actual sampling will produce).

The overall mean of the 200 random samples in this simulation was 1.00338 while the standard error of the means was .1332208 (true values are 1 and .129101). The overall mean of all 200 LHSs is 1.000338 and the variance of the means was .013266. Other linear models will give similar results.

Transforming Models to Linearity. The linear model example illustrated an impressive reduction in the variance associated with the estimation of the E(Y) when LHS is used rather than simple random sampling. Stein [15] has shown that as long as the sample size n is large compared to the number of variables k, LHS gives an estimator of E(Y)

with lower variance than simple random sampling for any function Y(**X**) having finite second moment. In addition, the closer that Y(**X**) is to additive in the components of **X**, the more LHS "helps" relative to random sampling. Stein also proposes a procedure for transforming the variables so that the functions whose expectation is being estimated is more nearly additive in the transformed variables than in the original variables. (Note: see Owen [12] for an important correction to Stein's work.)

Estimate of the Variance of the Mean of a Function. McKay, Conover, and Beckman [10] and Stein [15] compare the variance of estimate of the mean of a function h(**X**) for LHS and random sampling. The expected value of h(**X**) is estimated as

$$\bar{h} = n^{-1} \sum_{j=1}^{n} h(X_j) \tag{10}$$

For simple random sampling, the estimator is unbiased and

$$\mathrm{var}(\bar{h}) = n^{-1} \mathrm{var}(h(X)) \tag{11}$$

If LHS is used, then the estimate of the mean is still unbiased, and

$$\mathrm{var}(\bar{h}) = n^{-1} \mathrm{var}(h(X)) + n^{-1}(n-1) \mathrm{cov}(h(X_1), h(X_2)) \tag{12}$$

where X_1 and X_2 represent any two LHS input vectors. Thus, LHS lowers the variance if and only if cov(h(X_1),h(X_2)) < 0. Stein shows that as $n \to \infty$, the covariance term is nonpositive. McKay, Conover, and Beckman [10] show the covariance is nonpositive if h(X) is a monotone function of the components of the vector X.

References

1. Blower, S. M. and Dowlatabadi, H. (1994). “Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: an HIV Model, as an Example,” *International Statistical Review*, Vol. 62, 229-243.
2. Campbell, J. E. and Longsine, D. E. (1990). “Application of Generic Risk Assessment Software to Radioactive Waste Disposal,” *Reliability and Systems Safety*, Vol. 30, 183-193.
3. Conover, W. J. (1975). "On a Better Method of Selecting Values of Input Variables for Computer Codes." Unpublished manuscript.
4. Gwo, J. P., Toran, L. E., Morris, M. D., and Wilson, G. V. (1996). “Subsurface Stromflow Modeling with Sensitivity Analysis Using a Latin Hypercube Sampling Technique,” *Ground Water*, Vol. 34, 811-818.
5. Helton, J. C. (1994). “Treatment of Uncertainty in Performance Assessments for Complex Systems,” *Risk Analysis*, Vol. 14, No. 4, 483-511.
6. Helton, J. C., Bean, J. E., Butcher, B. M., Garner, J. W., Schreiber, J. D., Swift, P. N., and Vaughn, P. (1996). “Uncertainty and Sensitivity Analysis for Gas and Brine Migration at the Waste Isolation Pilot Plant: Fully Consolidated Shaft,” *Nuclear Science and Engineering*, Vol. 122, 1-31.
7. Iman, R. L. and Conover, W. J. (1980). "Small Sample Sensitivity Analysis Techniques for Computer Models, with An Application to Risk Assessment," *Communications in Statistics*, A9(17), 1749-1842. "Rejoinder to Comments," 1863-1874.
8. Iman, R. L., Davenport, J. M., and Zeigler, D. K. (1980). "Latin Hypercube Sampling (Program User's Guide)." Technical Report SAND79-1473, Sandia National Laboratories, Albuquerque, NM.
9. Iman, R. L. and Shortencarier, M. J. (1984). "A FORTRAN 77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples for Use With Computer Models," Technical Report NUREG/CR-3624, SAND83-2365, Sandia National Laboratories, Albuquerque, NM.
10. McKay, M. D., Conover, W. J., and Beckman, R. J. (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.
11. MacDonald, R. C. and Campbell, J. E. (1986). “Hydrocarbon Economic Risk Analysis”, *Journal of Petroleum Technology*, Vol. 38, 57-69.
12. Owen, A. B. (1990). "Correction to: Large Sample Properties of Simulations Using Latin Hypercube Sampling." *Technometrics*, .32(3), 367.
13. Painton, L. A. and Campbell, J. E. (1995). “Genetic Algorithms in Optimization of System Reliability,” IEEE Transactions on Reliability, Vol. 44, No 2, 172-178.
14. Steck, G. P., Iman, R. L., and Dahlgren, D. A. (1976). “Probablistic Analysis of LOCA , Annual Report for 1976,” SAND76-0535, Sandia National Laboratories, Albuquerque, NM.
15. Stein, M. (1987). "Large Sample Properties of Simulations Using Latin Hypercube Sampling," *Technometrics*, 29(2), 143-151.