EFFECT OF INITIAL SEED AND NUMBER OF SAMPLES ON SIMPLE-RANDOM AND LATIN-HYPERCUBE MONTE CARLO PROBABILITIES (CONFIDENCE INTERVAL CONSIDERATIONS)

Vicente J. Romero
Sandia National Laboratories, Albuquerque, NM 87185*
vjromer@sandia.gov

Abstract

In order to devise an algorithm for autonomously terminating Monte Carlo sampling when sufficiently small and reliable confidence intervals (CI) are achieved on calculated probabilities, the behavior of CI estimators must be characterized. This knowledge is also required in comparing the accuracy of other probability estimation techniques to Monte Carlo results. Based on 100 trials in a hypothesis test, estimated 95% CI from classical *approximate* CI theory are empirically examined to determine if they behave as *true* 95% CI over spectrums of probabilities (population proportions) ranging from 0.001 to 0.99 in a test problem. Tests are conducted for population sizes of 500 and 10,000 samples where applicable. Significant differences between true and estimated 95% CI are found to occur at probabilities between 0.1 and 0.9, such that estimated 95% CI can be rejected as not being true 95% CI at less than a 40% chance of incorrect rejection. With regard to Latin Hypercube sampling (LHS), though no general theory has been verified for accurately estimating LHS CI, recent numerical experiments on the test problem have found LHS to be conservatively over an order of magnitude more efficient than SRS for similar sized CI on probabilities ranging between 0.25 and 0.75. The efficiency advantage of LHS vanishes, however, as the probability extremes of 0 and 1 are approached.

Introduction

Derived statistics such as means and probabilities from Monte Carlo (MC) sampling are themselves random variables that depend on (among other things) the number of samples taken, the sampling algorithm, and random number generator (RNG) initial seed. Beginning practitioners of uncertainty analysis (and I was certainly guilty of this!) sometimes do not realize that the magnitudes of these effects can be significant, and report and use calculated statistics at face value without considering the associated confidence intervals (CI). However, ignoring this context information introduces unrealized, unacknowledged, or unquantified uncertainty to downstream calculations, interpretations, and conclusions.

Here, an exploratory test problem is used to quantitatively examine the effects of initial seed, number of samples, and sampling algorithm (*i.e.* Simple Random sampling, SRS, and Latin Hypercube sampling, LHS (McKay *et al.*, 1979)) on calculated probabilities. Consideration of confidence intervals naturally arises. Conditions surrounding calculation of CI on SRS probabilities are frequently different from the "text book" conditions underlying exact CI theory. The accuracy of classical *approximate* CI theory for these more realistic conditions is empirically investigated here. Regarding LHS, no general theory has been verified for accurately estimating LHS CI, though Iman (1981) outlines an approach. The applicability of the approach can now be assessed against empirical data (Romero, 2000) which establishes (at a 0.1 level of significance over 100 trials) that LHS is over an order of magnitude more efficient than SRS on this test problem for similar sized 95% CI on probabilities ranging between 0.25 to 0.75. The efficiency advantage of LHS diminishes completely, however, as the probability extremes of 0 and 1 are approached.

---

*Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Exploratory Problem

Figure 1 shows the temperature histories of two critical components in a device that is nominally said to fail if the temperature of the "stronglink" (SL) component reaches its nominal failure temperature of 593.3 C before the "weaklink" (WL) component reaches its nominal failure temperature of 248.9 C. The reliability of the device is thus measured in terms of a "safety margin" S [=time_of_stronglink_failure - time_of_weaklink_failure].

As determined from the figure, the device has a nominal safety margin of $S \approx 2.5$ minutes at nominal failure temperatures. The figure also shows, however, that marginal uncertainty in the stronglink and weaklink failure temperatures results in large uncertainty in the value of the safety margin. When the uncertainty bands in the figure at 5% above and below the nominal SL and WL failure temperatures are considered, the corresponding ranges in failure times can produce a safety margin that varies from about -25.5 to +34.5 minutes. Thus, this problem exhibits high sensitivity to component failure temperature uncertainties.

If the component failure temperatures are treated as random variables then the safety margin S is a [dependent] random variable in this context. It is desired to calculate the corresponding "probability of failure" P_f that the actual safety margin S is zero or negative (*i.e.* the probability that the random variable S will lie below a certain threshold value S_t generally, where here S_t is equal to zero). Hence,

$$P_f = \text{Prob}(S \leq S_t) \; ; \; S_t = 0. \tag{1}$$

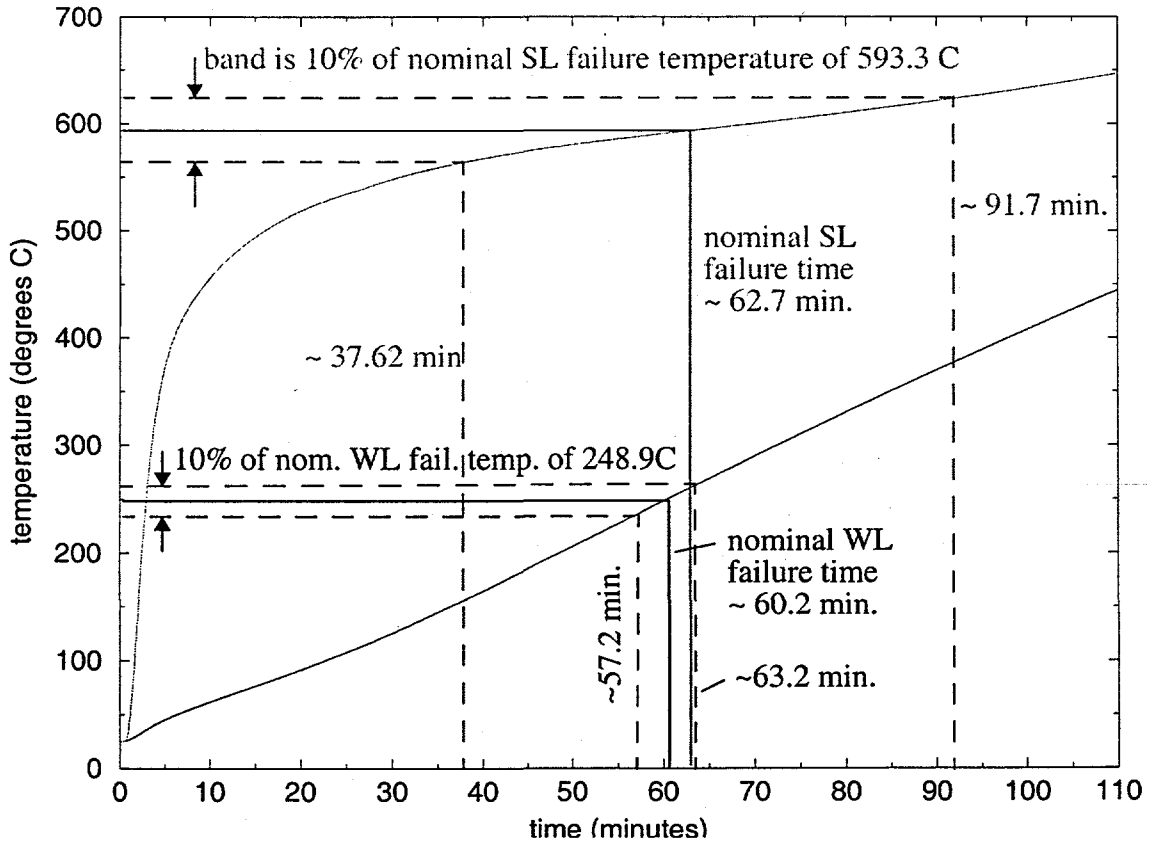

Figure 1. Sensitivity of stronglink (SL) and weaklink (WL) failure times, and hence safety margin *S*, to uncertain SL and WL failure temperatures. (Upper curve depicts SL temperature history, lower curve depicts WL temperature excursion.)

For this problem the SL and WL failure temperatures (random variables) are namely defined by normal distributions with means μ equal to the respective nominal failure temperatures of 593.33° C and 248.89° C, and standard deviations σ equal to 3% of the nominal values (*i.e.* 17.8° C and 7.47° C, respectively). (Here the distributions are actually truncated at 3σ above and below their means and then re-nomalized to unit area. Truncation of the SL distribution at $> 3\sigma$ requires a longer time scale than the one represented in the figure.) Sets of randomly paired weaklink and stronglink failure temperatures are generated from these distributions with the Monte Carlo sampling code of Iman and Shortencarrier (1984), and then a safety margin S_i is computed for each set by processing the data files containing the time-temperature data plotted in Figure 1. The probability of failure P_f is determined as the number of safety margin realizations S_i with value less than or equal to S_t, divided by the total number of samples.

Effect of Initial Seed and Number of Samples on LHS and SRS Prob. and CI

Figure 2 shows 21 calculated failure probabilities in each of four plots representing combinations of 500 or 10,000 samples with LHS or SRS. The 21 probabilities in each plot result from supplying 21 different initialization seeds[†] to the random number generator (RNG). For now, just looking at the point estimates and ignoring the associated error bars in the plots, the spread in probability estimates demonstrates and reinforces the fact that statistics (point estimates) derived from Monte Carlo sampling are in the class of random variables. (In fact, the basis of classical CI theory is that SRS means and probabilities are, assymptotically, normally distributed random variables.) The nontrivial variance in estimates here illustrates that we must not quote a Monte Carlo result in isolation, as though a deterministic quantity. Rather, we must be cognizant of the uncertainty associated with such results and always communicate and use them in context. Such context is expressly provided by confidence intervals as explained in the next section.

As expected, the predicted values from 500 samples vary much more widely over the 21 trials than the 10,000-sample results do; the signature of the specific RNG path taken from a given starting seed diminishes as more samples enter the population and decrease the effect of finite sampling and the specific path taken (assuming the RNG is a good random number generator with large period relative to the number of samples generated). The approximate 95% CI "error bars" about the point estimates of P_f in Figure 2 are considerably smaller for the 10,000-sample results than for the 500-sample results, nominally by a factor of $\sqrt{10{,}000/500}$ or ~ 4.5. It is also immediately evident from the figure that for the same number of samples the SRS results vary much more widely than the LHS results do. McKay *et al.* (1979) show that under fairly broad conditions the variance of statistical estimators derived from LHS is indeed less than that from SRS for the same number of samples. Under very restrictive special conditions, Iman and Conover (1980) analytically determine *how much* less the variance of LHS means is relative to SRS means, as a function of number of samples. Though no general theory has been verified for accurately estimating LHS CI under general conditions, recent numerical experiments (Romero, 2000) have empirically shown LHS to be (at a 0.1 significance level based on 100 trials) over an order of magnitude more efficient than SRS on the exploratory test problem for 95% CI on probabilities from 0.25 to 0.75. This finding is reflected in the approximate 95% CI LHS error bars in the figure, which are $1/\sqrt{10}$ or about 32% the size of the SRS error bars.

[†]These seeds are the first 21 5-digit pseudo-random numbers in the first column of Table 26.11 of *The Handbook of Mathematical Functions* (Abramowitz and Stegun, 1972).

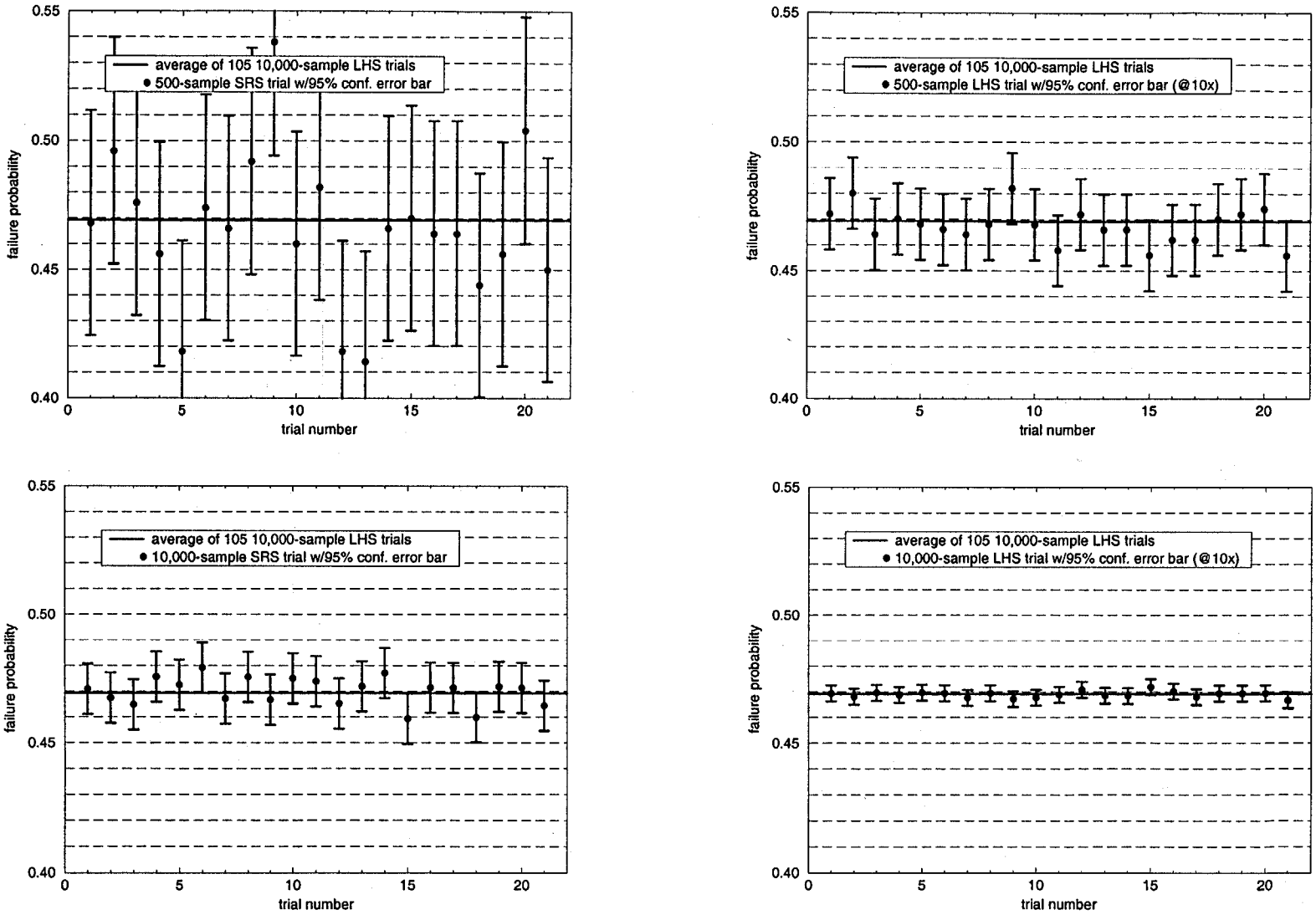

Figure 2. Differently seeded probability estims. with approx. 95% conf. intrvl. error bars for SRS and LHS at 500 and 10,000 samples.

Approximate Confidence Intervals for SRS Probabilities

The SRS data in Figure 2 is fitted with error bars representing *approximate* 95% CI from classical statistics theory: Any safety margin sample S_i that has a value greater than the specified threshold value S_t can be recoded to have the value 0, or the value 1 otherwise. With this coding, the number of 1's in N samples provides an indicator of the true proportion of the random variable S (*i.e.* the proportion of its probability density function) that lies below the threshold value S_t. Thus, $\hat{P}_f$, the ratio of the number of 1's to the total number N of samples, here represents an estimate of the true probability P_f of device failure in attaining a zero or negative safety margin (since $S_t = 0$ here). The total number of 1's that can be expected from N samples, or equivalently the probability estimate $\hat{P}_f$, follow binomial distributions. If enough samples are taken,[‡] the binomial distributions asymptote to normal distributions. In this case, the estimate $\hat{P}_f$ can be viewed as a random sample from a normal distribution centered on the true probability P_f. Thus, with $x\%$ likelihood or "confidence", the estimate $\hat{P}_f$ can be said to lie within a corresponding "confidence interval" range of the true result P_f. In the common case where the true probability value P_f is not known, but is to be inferred from the sampling, then *approximate* CI (ACI) are formed where the unknown value P_f is said to approximately lie within the following ACI of the probability estimate $\hat{P}_f$:

$$\left|\hat{P}_f - P_f\right| \leq F_{x\%}\sqrt{\frac{\hat{P}_f(1-\hat{P}_f)}{N}} \tag{2}$$

where the factor $F_{x\%}$ scales the interval size according to the relevant ($x\%$) degree of likelihood or confidence. Everything else being fixed, the wider the intervals the more likely they are to contain or envelop the actual value P_f, and *vice-versa*, the higher the % confidence with which the ACI are to contain the true result, the larger the intervals (thus $F_{x\%}$) must be. From classical statistics theory $F_{x\%}$=1.96 for 95% ACI, which are intervals that are approximately 95% likely to envelop the true result. For 99% ACI, $F_{x\%}$=2.576.

The error bars in the SRS plots in Figure 2 depict 95% ACI. If the approximation (2) is truly applicable, then on average no more than one in twenty such approximate 95% error bars will fail to contain the true probability.[**] However, five such failures and nearly a sixth are evident in the 42 SRS trials plotted, suggesting that equation (2) may not be strictly applicable. Accordingly, hypothesis tests are conducted to determine the significance level at which the validity of equation (2) can be rejected for the current application. Devore (1982) gives a concise five-page presentation of the methodology employed here. The null or default hypothesis H_0 claims that 95% ACI calculated from equation (2) fail 5% or less of the time to encompass the true probability. The alternative hypothesis H_a is that ACI bars so constructed fail more than 5% of the time to encompass the true result.

[‡] the criteria commonly stated in statistics books is that $N \cdot P_f \geq 5$ and $N(1-P_f) \geq 5$

[**] Here the "true probability" is actually a probability *range* given by 99% ACI calculated with equation (2), where $\hat{P}_f$ is taken to be the average of 105 10,000-sample LHS probability estimates (determined with 105 different initialization seeds in the spirit of footnote †), and N is taken to be 105×10^5 because, as cited previously, for this problem 10^4 LHS samples are at least equivalent to 10^5 SRS samples in the context of equation (2). Thus, to "contain the true probability" as asserted in the lead-in to this footnote, the 95% ACI error bars must completely encompass the 99% ACI error bars. The expanse of the 99% error bars is actually less than the thickness of the line marking the 105-trial average in the plots in Figure 2.

In one hypothesis test, 100 differently seeded 500-sample SRS trials yielded 12 failures or "outliers". Thus, the null hypothesis is rejected in favor of the alternative hypothesis when the "rejection region" is defined to contain 12 or more failures, and it is concluded in this case that 95% ACI from equation (2) allow more than 5% of estimates to be outliers, and thus are not truly CI at the 95% confidence level. The data contends that there is less than a one percent chance that H_0 is actually valid, *i.e.* that its rejection is incorrect. (H_0 is rejected at a continuity-corrected 0.0058 level of significance.)

Though a population of 500 samples easily meets the criteria quoted in footnote ‡, a similar experiment with 100 10,000-sample SRS populations was run to get another indication. Eight outliers occurred, whence H_0 can be rejected at a continuity-corrected 0.245 level of significance. Thus, equation (2) applies better under more sampling (larger populations), but even for N=10,000 samples its validity for deriving true 95% CI for this application can still be rejected with less than a 25% chance of incorrect rejection. Hence serious doubts are cast on the universal applicability of equation (2).

Related results documented in (Romero, 2000) consider nominal P_f levels (population proportions) ranging from 0.00053 to 0.99 corresponding to various threshold levels S_t in the exploratory problem. It is found that for 95% ACI on calculated probabilities between about 0.1 and 0.9, the validity of equation (2) can generally be rejected with less than a 40% chance of incorrect rejection. As the probability extremes of 0 and 1 are approached the evidence against the applicability of equation (2) diminishes considerably.

In conclusion, assuming validity of the well-pedigreed RNG used here, the classical "text book" expression for approximate SRS 95% CI appears not to be conservative. In particular, away from the probability extremes of 0 and 1 it should not be assumed that 95% ACI will on average contain the true probability 95% of the time. Still, the expression for 95% ACI is useful in *gauging* the chance that an estimate obtained will be an outlier.

References

Abramowitz, M., and I.A. Stegun, Eds. (9th printing, circa 1972), *Handbook of Mathematical Functions,* Dover Publications, New York.

Devore, J.L. (1982), *Probability & Statistics for Engineering and the Sciences*, Brooks/Cole Publishing Co., Wadsworth, Inc., Belmont, CA., pp. 99 - 104.

Iman, R.L. (1981), "Statistical Methods for Including Uncertainties Associated with the Geologic Isolation of Radioactive Waste which allow for a Comparison with Licensing Criteria," *Proceedings of the Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste*, Gatlinburg, Tennesssee, March 9-13.

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, Part A– Theory and Methods*, **17**, pp. 1749-1842.

Iman, R.L., and M.J. Shortencarier (1984), "A FORTRAN77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples to Use with Computer Models," Sandia National Laboratories report SAND83-2365 (RG).

McKay, M.D., R.J. Beckman, and W.J. Conover, (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.

Romero, V.J. (2000), "A Systematic Approach to Non-Deterministic Analysis with Large Computer Models," Sandia National Laboratories SAND report currently in preparation.