Reliability Engineering and System Safety 149 (2016) 187–203

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Advantages of variance reduction techniques in establishing confidence intervals for quantiles

Dave Grabaskas [a,1,*], Marvin K. Nakayama [b,2], Richard Denning [c,3], Tunc Aldemir [c,4]

[a] Ohio State University, 201 W 19th Ave, Columbus, OH 43210, USA
[b] New Jersey Institute of Technology, Computer Science Department, Newark, NJ 07102, USA
[c] Ohio State University, USA

ARTICLE INFO

Article history:
Received 23 July 2013
Received in revised form
18 December 2015
Accepted 19 December 2015
Available online 4 January 2016

Keywords:
Confidence intervals
Quantiles
Nuclear regulation
Best-estimate
Uncertainty

ABSTRACT

Over the past two decades, U.S. nuclear power plant regulation has increasingly depended on best-estimate plus uncertainty safety analyses. As a result of the shift to best-estimate analyses, the distribution of the output metric must be compared against a regulatory goal, rather than a single, conservative value. This comparison has historically been conducted using a 95% one-sided confidence interval for the 0.95-quantile of the output distribution, which is usually found following the technique of simple random sampling using order statistics (SRS-OS). While SRS-OS has certain statistical advantages, there are drawbacks related to the available sampling schemes and the accuracy and precision of the resulting value. Recent work has shown that it is possible to establish asymptotically valid confidence intervals for a quantile of the output of a model simulated using variance reduction techniques (VRTs). These VRTs can provide more informative results than SRS-OS. This work compares SRS-OS and the VRTs of antithetic variates and Latin hypercube sampling through several experiments, designed to replicate conditions found in nuclear safety analyses. This work is designed as an initial investigation into the use of VRTs as a tool to satisfy nuclear regulatory requirements, with hope of expanded analyses of VRTs in the future.

1. Introduction and background

In its efforts to transition from conservative regulatory models to best-estimate plus uncertainty analyses, the Nuclear Regulatory Commission (NRC) has developed safety analysis guidelines that require quantification of the impact of uncertainties on the output of accident simulations [1]. While different methods to meet this quantitative requirement have been discussed [2], the most common approach is to calculate a confidence interval for a quantile of the output distribution. An NRC-approved method of accomplishing this task has been the technique of simple random sampling using order statistics (SRS-OS) [2]. While this method has many positive aspects for nuclear safety analysts, it has certain drawbacks related to the available sampling schemes and the accuracy and precision of the resulting value. Recent work has shown that it is possible to establish asymptotically valid confidence intervals for a quantile of the output of a model simulated using variance reduction techniques (VRTs), such as Latin hypercube sampling (LHS) [3]. These VRTs can possibly provide more informative results than SRS-OS. The current work compares SRS-OS and the VRTs of antithetic variates and LHS through several experiments, designed to replicate conditions found in nuclear safety analyses. These tests include a simple nonlinear equation system, a design-basis accident analysis of a nuclear power plant using a response surface surrogate for the thermal-hydraulic code RELAP5 [4], and a beyond-design-basis accident analysis conducted using the severe-accident analysis computer code MELCOR [5]. This work was designed as an initial investigation into the use of VRTs as a tool to satisfy nuclear regulatory requirements, with the hope of expanded analyses of VRTs in the future.

Section 1 begins with an overview of regulatory history and a quick description of hypothesis testing, which is used to frame the issue of regulatory compliance in a more-rigorous fashion. This is followed in Section 2 by a review of the statistical methods that are later compared through a series of example problems in Section 3. The conclusions are reviewed in Section 4.

* Corresponding author.
E-mail addresses: Grabaskas.2@osu.edu (D.o.s. Grabaskas), Marvin@njit.edu (M.K. Nakayama), Denning.8@osu.edu (R. Denning), Aldemir.1@osu.edu (T. Aldemir).
[1] Tel.: +1 630 252 2195.
[2] Tel.: +1 973 596 3398.
[3] Tel.: +1 614 292 2544.
[4] Tel.: +1 614 292 4627.

http://dx.doi.org/10.1016/j.ress.2015.12.015
0951-8320/

Nomenclature

95/95	95% One-sided Confidence Interval for 0.95-quantile
AV	Antithetic Variates
BFD	Backward Finite-Difference
CDF	Cumulative Distribution Function
CFD	Central Finite-Difference
CI	Confidence Interval
CLT	Central Limit Theorem
LHS	Latin Hypercube Sampling
LOCA	Loss of Coolant Accident
NRC	Nuclear Regulatory Commission
rLHS	Replicated Latin Hypercube Sampling
SRS	Simple Random Sampling
SRS-OS	Simple Random Sampling Order Statistics
VRT	Variance Reduction Technique
Run	A single execution of a computer code
Case	A collection of runs for the rLHS method
Trial	A complete experiment that an analyst would conduct; for the rLHS method, it would consist of multiple cases.
Accuracy	A measure of the expected distance between the correct quantile and the upper endpoint of a one-sided CI that arises in the 95/95 analysis
Precision	The variance of possible upper endpoints of one-sided CIs

1.1. Regulatory background

The initial approach to the treatment of modeling uncertainties in regulatory analysis was to use non-mechanistic, conservative models. In the implementation of the Part 50 Appendix K of the Code of Federal Regulations [6], which describes a prescription for the conservative treatment of uncertainties in the analysis of loss-of-coolant accidents (LOCAs), it became apparent that what was thought to be conservative might not be conservative in all cases, and that conservative regulatory models could be misleading with regard to the improvement of reactor safety. The transition to best-estimate plus uncertainty regulatory requirements began with an amendment to 10 CFR 50.46 [1] in 1988, which allowed for realistic modeling of LOCAs. While this rule-change signaled an advancement in regulatory safety analysis, the statistical requirements of the output result were vague, stating only that there should be a "high level of probability that the criteria would not be exceeded."

In 1989, the NRC issued RG1.157 [7], which helped clarify the procedure for performing a best-estimate calculation relating to the design bases for essential safety systems. It set the standard for the handling of computational uncertainty for nuclear safety applications by stating that a 95% probability level is considered acceptable to the NRC staff for comparison of best-estimate predictions to safety limits. However, the ambiguity of the term "95% probability level" remained an issue for the analyst.

The most obvious solution to the "95% probability" requirement was to estimate the 0.95-quantile of the output distribution. One method to do this was to perform a large number of simple random sampling (SRS) computer code runs using Monte Carlo sampling and simply order and count the results until 95% of the runs fell below that value. Then this point estimate of the 0.95-quantile would be compared to the safety limit. The large number of runs required by SRS to obtain sufficient accuracy represented a major problem for safety analysts, due to minimal computing power and extended code run times. There was also the question of just how many runs would be necessary for an analyst to be able to claim that the estimate of the 0.95-quantile was sufficiently accurate.

Response-surface methods [8] were initially proposed as a way of reducing runs and increasing knowledge of the overall behavior of the parameters of interest. An advantage of this approach is that it employs a fixed matrix of runs to be conducted to estimate the desired surface. This property not only gives the analyst a plan to provide to the regulator, but also produces a level of understanding about the impact of different input parameters. However, like the large-sample SRS case, run designs often needed to be very large to capture input interactions and nonlinearities, and the only way around this was to group input parameters based on the analyst's judgment [2]. In response to these considerations, methods were developed that required a smaller number of runs, but which could satisfy the regulatory guidelines.

Both AREVA [9] and Westinghouse [2] developed approaches for the use of simple random sampling using order statistics (SRS-OS) for their regulatory LOCA analyses. While the method of SRS-OS was first considered for use in the nuclear industry in the 1970's [10], it was not until the NRC published NUREG-1475 [11], a guide to applying statistics, in 1994 that the NRC provided a more comprehensive picture of its use for regulatory requirements. Gesellschaft fürAnlagen-und Reaktorsicherheit (GRS) helped bring SRS-OS to the thermal hydraulic and safety fields soon after [12]. Major steps forward occurred in 2003 and 2004 with publications by Guba, Pál, and Makai [13], and Nutt and Wallis [14]. These works not only expanded on how SRS-OS could be used in safety analyses, but also demonstrated how it could be applied to satisfy the 95% probability reporting requirement. The solution provided by Guba, Pál, and Makai [13] and Nutt and Wallis [14] to this question was to report a 95% upper one-sided confidence interval for the 0.95-quantile of the output distribution. Based on the works of Wilks [15] and Wald [16], this method simulates the model using SRS and specifies a particular order statistic as an endpoint of a 95% tolerance interval with 95% confidence. This method was considered acceptable by the NRC in regards to the 95% probability requirement [2], and is discussed in detail in Section 2.1.

While the acceptance of the 95% confidence interval for the 0.95-quantile has been adopted by the NRC for satisfying design-basis accident requirements, there are other safety applications for which less stringent requirements may be appropriate, such as for the analysis of beyond-design-basis events. For the analysis of these events, similar, but less stringent limits could be established, such as the use of the 95% confidence level with a lower quantile.

1.2. Comparison to safety requirement within the context of hypothesis testing

The process of using a confidence interval for a quantile to compare to a regulatory safety limit can be explained more rigorously using hypothesis testing. For example, assume there is a regulatory safety goal with value G, that represents a prescribed limit that the true 0.95-quantile ($\xi_{0.95}$) of the output of a safety analysis cannot exceed. In this case, we define a hypothesis test, with null hypothesis H_0:$\xi_{0.95} \geq G$ and alternative hypothesis H_1:$\xi_{0.95} < G$. This framework puts the burden of proof on H_1, which hypothesizes that the true 0.95-quantile value of the output falls below the prescribed limit. Hypothesis testing uses a statistic to make a decision about a parameter. Since the true 0.95-quantile $\xi_{0.95}$ of the system, a parameter, is unknown, it needs to be estimated. Define a 95/95 value to be the upper confidence limit of an

Table 1
Hypothesis testing outcome possibilities.

	H_0 is true	H_1 is true
Accept H_0	Correct decision	Incorrect decision type-II error
Accept H_1	Incorrect decision type-I error	Correct decision

upper-one-sided 95% confidence interval (the first "95" in 95/95) for the 0.95-quantile (the second "95" in 95/95), and assume that this 95/95 value will be used as the basis for a hypothesis test at level $\alpha = 0.05$. If the 95/95 value lies below the safety goal G, then H_0 is rejected (i.e., H_1 is accepted); otherwise, H_0 is accepted. An example of such a regulatory hypothesis test is the LOCA analysis referenced in the previous section, where it is established, with at least 95% confidence, that the 0.95-quantile of the analysis metric, such as the peak cladding temperature, must be below the regulatory requirement of $G = 2200$. From there, four outcomes of the hypothesis test are possible, as seen in Table 1.

As the table shows, there are two types of errors possible during the analysis. Type-I errors, or false positives, accept the alternative hypothesis when in actuality it is not true. In our example, this would be the system appearing to satisfy the safety limit when the true quantile does not. This will occur when the resulting 95/95 value of the analysis is below the regulatory limit, but the true 0.95-quantile exceeds it. Type-II errors, or false negatives, are when the null hypothesis is accepted as true, but in actuality, the alternative hypothesis is true. In our example, this would be the system appearing to fail the safety limit, with the 95/95 value at a higher value than the limit, even though the true quantile lies below the limit value.

Both errors have negative impacts. Type-I errors would appear to be the more serious error since a system is being approved that should not be. However, Type-II errors also have drawbacks, since a system will be viewed as failing when it should not. This type of error could mean that time and resources will be dedicated to fixing a non-existent problem, when they could have been applied more productively. The goal of the safety analysis should be to reduce the probability of committing both Type-I and Type-II errors. This not only helps reduce false positives, but helps assure that safety measures are most effectively addressing true safety issues.

Beyond avoiding errors during regulatory analysis, there are other reasons utilities and regulators would like to increase the accuracy of the resulting 95/95 value. The margin from the resulting value to the safety limit is also of use, and increasing the accuracy of the 95/95 value can provide a clearer picture of the magnitude of this margin. Significant margin may allow utilities to increase reactor temperature or power, increasing profit. As Westinghouse has stated, "The quantification and tracking of the margin is most often requested by both the plant operator and the regulator,…" [2].

2. Methods

Before describing the methods examined to establish confidence levels, it is important to review the process of estimating a quantile from empirical data. To state the problem more formally, suppose there is a system that has as its input a random vector X, and (univariate) output Y with cumulative distribution function[5] (CDF) F. The components of the input random vector X may have different marginal distributions and may be dependent. By inverting F, the p-quantile ξ_p can be found:

$$\xi_p = F^{-1}(p) \tag{1}$$

where $F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$. Thus, if $p = 0.95$, $\xi_{0.95}$ is the true 0.95-quantile. If the goal were to estimate ξ_p using simulation with SRS, a sample of independent and identically distributed (i.i.d.) univariate outputs $Y_1, Y_2, \ldots, Y_n$ from distribution F would be generated. From there, the empirical cumulative distribution function $\hat{F}_n$ can be computed:

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) \tag{2}$$

where $I(A)$ is the indicator function of an event A, which assumes value 1 on A and 0 on the complement A^C. The p-quantile estimator $\hat{\xi}_{p,n}$ is then computed by inverting $\hat{F}_n$, i.e., $\hat{\xi}_{p,n} = \hat{F}_n{}^{-1}(p)$. The quantile estimator $\hat{\xi}_{p,n}$ can also be calculated using order statistics, where the sample $Y_1, Y_2, \ldots, Y_n$ would be sorted in ascending order $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$, where $Y_{(i)}$ is the ith smallest output. Then $\hat{\xi}_{p,n} = Y_{(\lceil np \rceil)}$ where $\lceil \bullet \rceil$ is the round-up function. While these techniques result in a point estimate of the quantile, they do not provide any measure of the error in the estimate.

2.1. Simple random sampling using order statistics

As stated earlier, there are several key properties that make SRS-OS appealing to nuclear safety analysts. The biggest benefits of the SRS-OS method are that it is nonparametric and non-asymptotic. Nonparametric means that the method is independent of the output's probability distribution, as long as it is continuous. Since it is non-asymptotic, the validity of the confidence statement holds exactly for certain finite sample sizes n and does not depend on n growing toward infinity. The SRS-OS method first fixes an integer $r \geq 1$ and then determines the number n of runs necessary so that the r-th largest output of the n runs is a valid 95/95 value. Then the 95/95 criterion is verified by checking if the 95/95 value lies below the safety limit. With this method, it is also possible to find the number n of runs necessary to construct a valid confidence interval for any quantile.

The required value for n, when $r = 1$, can be determined as follows. Suppose that n i.i.d. runs are performed, giving n i.i.d. outputs, and consider the true p-quantile ξ_p. Each of the n outputs has probability p of lying below ξ_p, so the probability that all n outputs are less than ξ_p is p^n. Thus, the probability that at least one output is larger than ξ_p is $1 - p^n$, so the probability that the largest of the n outputs is greater than ξ_p is

$$\beta = 1 - p^n \tag{3}$$

Setting $\beta = p = 0.95$ and solving for n in Eq. (3) results in $n = 59$. Thus, if 59 SRS runs are conducted, then the largest (i.e., $r = 1$) of the 59 outputs is a 95/95 value.

Since so few SRS runs are conducted when $r = 1$, the variance of the resulting value can be high. To obtain a more *accurate* 95/95 value, the value of r can be increased, which necessitates a larger run size n. Here, *accuracy* is defined as the expected distance from the 95/95 value to the true quantile ξ_p, and *precision* is the variance of possible 95/95 values. For $r \geq 1$, the argument used to obtain Eq. (3) can be generalized to show that the probability that the r-th largest of the n outputs is larger than ξ_p is given by the binomial probability

$$\beta = 1 - \sum_{i=n-r+1}^{n} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \tag{4}$$

[5] While techniques for using SRS-OS for the evaluation of multiple output metrics from a single function (or computer code) have been proposed by both Pál and Makai [42] and Nutt and Wallis [14], the focus here is on the case with a single output metric Y, as this work is seen as a first step toward the use of new and innovative analysis methods. Future work will consider the case of multiple output metrics.

see Guba, Pál, and Makai [13] and Nutt and Wallis [14]. Now set $\beta = p = 0.95$ and fix $r \geq 1$ in Eq. (4). Then solving for n numerically gives the number of runs needed to ensure that the r-th largest output of the n runs is a valid 95/95 value. For example, if $r = 1$, then Eq. (4) reduces to Eq. (3), resulting in $n = 59$, as before. If $r = 3$, then $n = 124$, so the third largest output from the 124 runs is a valid 95/95 value.

While SRS-OS is straightforward and easy to implement, there are several potential drawbacks. SRS-OS requires i.i.d simulation input values *across runs*, such as those chosen when using SRS sampling, in order to satisfy the requirements of the binomial distribution, which forms the basis of the SRS-OS approach. (As noted earlier, the inputs *within* a run may have different marginal distributions and may be dependent.) Therefore, other sampling schemes, like the variance reduction techniques discussed in the following section, cannot be used in conjunction with SRS-OS since the input values *across runs* are no longer independently chosen. This drawback was one of the major motivations for the current work and influences the variance of the result in the following two ways.

The first issue arises not from the resulting SRS-OS values being invalid, but from the variance and conservatism of the results. At lower run levels, the 95/95 value will, on average, be overly conservative. This can be seen in Fig. 1, based on a similar plot by Nutt and Wallis [14], which shows the probability density of the 95/95 value of a SRS-OS analysis for the various values of r, where the corresponding sample size n is chosen to be the smallest integer so that β is at least 0.95. The probability density is computed by taking the derivative of Eq. (4) with respect to p and is shown as a function of the cumulative probability p; see Nutt and Wallis [14]. Thus, the figure gives an indication of the location of the determined 95/95 value when using SRS-OS at different sampling levels. For example, when 59 runs are conducted ($r = 1$, $n = 59$ runs) the determined 95/95 value is more likely to fall in an interval near the 1.0-quantile than an interval near the 0.95-quantile. Even when 1008 runs are used ($r = 40$, $n = 1008$ runs), the 95/95 value is more likely to be near the 0.96-quantile than the 0.95-quantile.

Second, as stated above, since SRS sampling is used, the variance of the 95/95 value can be very high when r (and subsequently n) is small, so the likely range of 95/95 values is large. In certain cases, this can mean that even though only ~5% of trials will fall below the actual quantile (due to the 95% confidence), there is a not-insignificant chance that they could fall well below. Thus, these trials not only pose the threat of a Type-I error, but by violating the regulatory limit by a wide margin, they may approach, or exceed, the true capacity of the system.

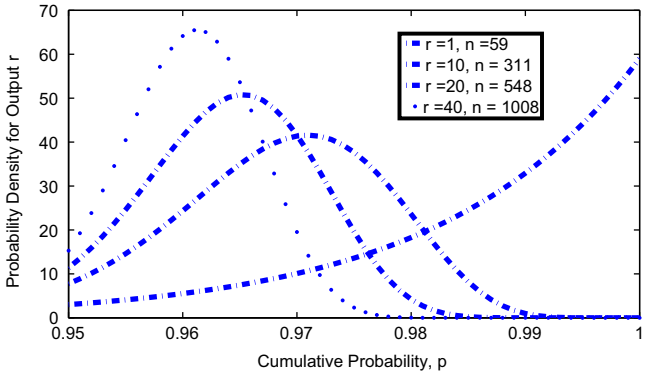

Fig. 1. Dependence of the probability density for the predictions on the order selected to represent the 0.95-quantile.

2.2. Asymptotic confidence intervals

In contrast to the SRS-OS method, which states *a priori* a set number of runs which must be conducted to establish a valid confidence interval (CI) for a quantile, it is also possible to establish other confidence intervals by proving a central limit theorem (CLT) as the number of runs grows large. This method has long been known when using SRS sampling [17], but until recently, has not been proven when using variance reduction techniques (VRTs). VRTs are methods of simulation-based sampling that seek to reduce the variance of the output result by either by using features of the model to correct or adjust outputs, or by reducing the variability of the inputs [18]. The most common VRTs are stratified sampling, Latin hypercube sampling (LHS), antithetic variates (AV), importance sampling, and control variates [18]

This section reviews asymptotic confidence intervals for SRS sampling, and discusses the recent work to expand their applicability to VRTs. For this work, only LHS and AV were investigated because these methods can generally be applied without relying on prior knowledge of characteristics of the system. Using a VRT that relies on detailed knowledge about the system to adjust sampling methods or outputs, such as importance sampling and some methods of control variates, may be difficult to apply in practice unless one has a good understanding of the dynamics of the underlying model. It is important to note that assumptions are still needed about the system when using LHS and AV to guarantee they reduce variance, but these assumptions are likely to be satisfied in great generality. Both methods are essentially guaranteed to reduce the variance of the output if the system is a monotone function of the inputs, meaning increasing an input value will lead to the output either always decreasing or always increasing. It is still possible to get variance reduction if this is not true, but it is not ensured [19]

Suppose we can represent the univariate output Y from our simulation model as

$$Y = g(U_1, U_2, \ldots, U_d) \tag{5}$$

where g is a given (deterministic) function having a fixed number d of arguments and $U_1, U_2, \ldots, U_d$ are i.i.d. uniform[0, 1] random variables. The function g, which takes the d i.i.d. uniforms and transforms them into a single output Y, can be quite complicated, and it may not be possible to express g in closed-form. For example, a LOCA simulation might have s input random variables $X_1, X_2, \ldots, X_s$ with some joint distribution G_0. Thus, $X_1, X_2, \ldots, X_s$ may not necessarily be independent nor identically distributed, so if X_j has marginal distribution G_j, then $G_1, G_2, \ldots, G_d$ may be different. Then $X_1, X_2, \ldots, X_s$ are fed into a detailed computer code, which then computes an output Y. In this case, the function g transforms the d i.i.d. uniforms $U_1, U_2, \ldots, U_d$ into observations of the s input variables $X_1, X_2, \ldots, X_s$, runs the computer code with these inputs, and produces a univariate output Y. (In many settings, we have $s = d$, and each input variable X_j is sampled from its marginal distribution G_j via inversion, i.e., $X_j = G_j^{-1}(U_j)$. In this case, we have that $X_1, X_2, \ldots, X_s$are independent but not necessarily identically distributed. More generally, our framework allows for dependence among $X_1, X_2, \ldots, X_s$.) Let F be the CDF of Y, so for $0 < p < 1$, the p-quantile is $\xi_p = F^{-1}(p)$.

2.2.1. Review of SRS

We now review how to use SRS to estimate and construct an asymptotically valid confidence interval for the p-quantile ξ_p based on a CLT when Y has the form in Eq. (5). We can generate n i.i.d. copies of Y by first generating nd i.i.d. uniform[0, 1] random variables $U_{i,j}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, d$, where d is as defined in Eq.

(5). We arrange the uniforms in an $n \times d$ array

$$\begin{matrix} U_{1,1} & U_{1,2} & \cdots & U_{1,d} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ U_{n,1} & U_{n,2} & \cdots & U_{n,d} \end{matrix} \tag{6}$$

where the ith row is used to generate the ith output Y_i, i.e.,

$$\begin{aligned} Y_1 &= g(U_{1,1}, U_{1,2}, \ldots, U_{1,d}), \\ Y_2 &= g(U_{2,1}, U_{2,2}, \ldots, U_{2,d}) \\ &\vdots \\ Y_n &= g(U_{n,1}, U_{n,2}, \ldots, U_{n,d}) \end{aligned}$$

Each Y_i has the distribution F because the d entries in the ith row of Eq. (6) are i.i.d. uniforms, as required by Eq. (5). Moreover, $Y_1, Y_2, \ldots, Y_n$ are independent by the independence of the n rows in Eq. (6). Then we compute the SRS p-quantile estimator as $\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p)$, where $\hat{F}_n$ is defined in Eq. (2).

To establish a CI for ξ_p based on the SRS point estimator $\hat{\xi}_{p,n}$, we first want to show that $\hat{\xi}_{p,n}$ satisfies a CLT as the sample size n grows large. One way of establishing this is by first proving that $\hat{\xi}_{p,n}$ satisfies a so-called *Bahadur representation*; see [20]. Let f denote the derivative, when it exists, of F, and assume that $f(\xi_p) > 0$. Now consider the following heuristic argument. When n is large, it is reasonable to expect that $\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p) \approx F^{-1}(p) = \xi_p$. Because $F(\xi_p) = p$ by definition, we see that $F(\hat{\xi}_{p,n}) \approx p$, so a Taylor approximation yields

$$\begin{aligned} p &\approx F(\hat{\xi}_{p,n}) \\ &\approx F(\xi_p) + f(\xi_p)(\hat{\xi}_{p,n} - \xi_p) \\ &\approx \hat{F}_n(\xi_p) + f(\xi_p)(\hat{\xi}_{p,n} - \xi_p) \end{aligned}$$

where the last approximation holds because $\hat{F}_n \approx F$. Rearranging terms leads to $\hat{\xi}_{p,n} = \xi_p + [p - \hat{F}_n(\xi_p)]/f(\xi_p)$, which approximates a quantile estimator by a linear transformation of a CDF estimator.

Bahadur [20] makes this argument mathematically rigorous. In particular, suppose that the second derivative F'' of F exists and is bounded in a neighborhood of ξ_p, and that $f(\xi_p) > 0$. Then Bahadur proves that

$$\begin{aligned} \hat{\xi}_{p,n} &= \xi_p + \frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)} + R_n', \text{ where } R_n' \\ &= O\left(n^{-3/4} \log n\right) \text{ as } n \to \infty \text{ with probability 1,} \end{aligned} \tag{7}$$

where the statement "$A_n = O(g(n))$ as $n \to \infty$ with probability 1" means that there exists an event Ω_0 such that $P(\Omega_0) = 1$ and for each $\omega \in \Omega_0$, there exists a constant $B(\omega)$ such that $|A_n(\omega)| \le B(\omega)\, g(n)$ for all n sufficiently large. Eq. (7) is known as a Bahadur representation.

Under weaker conditions, Ghosh [21] establishes a variant of a weaker version of Eq. (7), which will be useful and sufficient for our needs. Specifically, let p_n be a perturbed value of p converging to p as $n \to \infty$, and let $\hat{\xi}_{p_n,n} = \hat{F}_n^{-1}(p_n)$. (Working with a perturbed p_n rather than a fixed p will allow us to construct an asymptotic CI for ξ_p when applying VRTs.) Also, let $\Rightarrow$ denote convergence in distribution (Section 1.2.4 of [22]), which is weaker than convergence with probability 1 and is equivalent to convergence in probability when the limit is deterministic. Ref [21] shows that if $f(\xi_p) > 0$, then

$$\hat{\xi}_{p_n,n} = \dot{\xi}_{p_n} + \frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)} + R_n \tag{8}$$

with

$$\sqrt{n} R_n \Rightarrow 0 \text{ as } n \to \infty, \tag{9}$$

where

$$\dot{\xi}_{p_n} = \xi_p + \frac{p_n - p}{f(\xi_p)} \tag{10}$$

when $p_n = p + O(\frac{1}{\sqrt{n}})$. If f is also continuous in a neighborhood of ξ_p, then Eqs. (8) and (9) hold for all $p_n \to p$ with

$$\dot{\xi}_{p_n} = F^{-1}(p_n) \tag{11}$$

The results in Eqs. (8) and (9) ensure that the SRS quantile estimator $\hat{\xi}_{p,n}$ satisfies a CLT. To show this, fix $p_n = p$ in Eq. (8) so $\dot{\xi}_{p_n} = \xi_p$, rearrange terms and scale by $\sqrt{n}$ to obtain

$$\sqrt{n}(\hat{\xi}_{p,n} - \xi_p) = \sqrt{n}\left(\frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)}\right) + \sqrt{n} R_n \tag{12}$$

From Eq. (2), we see that $\hat{F}_n(\xi_p)$ is the sample average of i.i.d. indicator functions $I(Y_i \le \xi_p)$, $i = 1, 2, \ldots, n$, each of which has mean p and variance $0 < p(1-p) < \infty$. Hence, the first term on the right side of Eq. (12) satisfies a CLT (see p. 28 of [22]), with limit $N(0, \frac{p(1-p)}{f^2(\xi_p)})$ as $n \to \infty$, where $N(a, b^2)$ denotes a normal random variable with mean a and variance b^2. The second term on the right side of Eq. (12) vanishes (in distribution) as $n \to \infty$ by Eq. (9), so Slutsky's theorem (p. 19 of [22]) ensures that $\sqrt{n}(\hat{\xi}_{p,n} - \xi_p) \Rightarrow N(0, \frac{p(1-p)}{f^2(\xi_p)})$ as $n \to \infty$, or equivalently,

$$\frac{\sqrt{n}}{\sqrt{p(1-p)}\lambda_p}(\hat{\xi}_{p,n} - \xi_p) \Rightarrow N(0,1) \text{ as } n \to \infty \tag{13}$$

where

$$\lambda_p = \frac{1}{f(\xi_p)} \tag{14}$$

which is known as the *sparsity function* [23] or the *quantile density function* [24]. One interpretation of the CLT is that the left of Eq. (13) will have approximately a standard (i.e., mean 0, variance 1) normal distribution for large n.

The CLT in Eq. (13) illustrates one reason why a Bahadur representation is useful. The latter shows that a quantile estimator can be approximated as a linear transformation of a CDF estimator, which typically is a sample average so it satisfies a CLT. Thus, a Bahadur representation provides insight into why a quantile estimator, which is not a sample average, satisfies a CLT.

Once the CLT in Eq. (13) has been established, we can then unfold it to obtain a confidence interval for ξ_p. Let $z_\beta = \Phi^{-1}(1-\beta)$ for any $0 < \beta < 1$, where Φ is the CDF of $N(0,1)$. Then to derive a $100(1-\alpha)\%$ confidence interval for ξ_p, note that

$$\begin{aligned} 1 - \alpha &= P - z_{\alpha/2} \le N(0,1) \le z_{\alpha/2} \\ &\approx P\left\{ -z_{\alpha/2} \le \frac{\sqrt{n}}{\sqrt{p(1-p)}\lambda_p}(\hat{\xi}_{p,n} - \xi_p) \le z_{\alpha/2} \right\} \\ &= P\left\{ \hat{\xi}_{p,n} - z_{\alpha/2}\frac{\sqrt{p(1-p)}\lambda_p}{\sqrt{n}} \le \xi_p \le \hat{\xi}_{p,n} + z_{\alpha/2}\frac{\sqrt{p(1-p)}\lambda_p}{\sqrt{n}} \right\}, \end{aligned}$$

where the approximation holds for large n by the CLT. Hence,

$$\left[\hat{\xi}_{p,n}-z_{\alpha/2}\frac{\sqrt{p(1-p)}\lambda_p}{\sqrt{n}},\hat{\xi}_{p,n}+z_{\alpha/2}\frac{\sqrt{p(1-p)}\lambda_p}{\sqrt{n}}\right]\equiv\left[\hat{\xi}_{p,n}\pm z_{\alpha/2}\frac{\sqrt{p(1-p)}\lambda_p}{\sqrt{n}}\right] \tag{15}$$

is an asymptotically valid (two-sided) $100(1-\alpha)\%$ confidence interval for ξ_p. Since λ_p is unknown, for the CI in Eq. (15) to be implementable in practice, we need to replace λ_p with a *consistent* estimator $\hat{\lambda}_{p,n}$; i.e., $\hat{\lambda}_{p,n}\Rightarrow\lambda_p$ as $n\to\infty$. If we have such an estimator, then

$$J_n=\left[\hat{\xi}_{p,n}\pm z_{\alpha/2}\frac{\sqrt{p(1-p)}\hat{\lambda}_{p,n}}{\sqrt{n}}\right]$$

is another asymptotic two-sided $100(1-\alpha)\%$ CI for ξ_p, which is *asymptotically valid* in the sense that $P\xi_p\in J_n\to 1-\alpha$ as $n\to\infty$.

Now the key issue is that of constructing a consistent estimator $\hat{\lambda}_{p,n}$ of λ_p from Eq. (14). Since $\lambda_p=1/f(\xi_p)=\frac{d}{dp}F^{-1}(p)=\lim_{h\to 0}[F^{-1}(p+h)-F^{-1}(p-h)]/(2h)$ by the chain rule of differentiation, a natural estimator for λ_p is the (central) finite difference

$$\hat{\lambda}_{p,n}=\frac{\hat{F}_n^{-1}(p+h_n)-\hat{F}_n^{-1}(p-h_n)}{2h_n} \tag{16}$$

where $h_n>0$ is a user-specified (small) parameter known as the *bandwidth* or *smoothing parameter*. (See Section VII.1 of [25] or Section 7.1 of [18] for overviews of finite-difference estimators.) If $h_n\to 0$ and $nh_n\to\infty$ as $n\to\infty$, then [26] and [27] prove the consistency of $\hat{\lambda}_{p,n}$ as $n\to\infty$. More detail on this estimator is provided in Section 2.2.5.

Rather than a two-sided CI for ξ_p, we can also develop an asymptotic upper one-sided $100(1-\alpha)\%$ CI for ξ_p as

$$\left(-\infty,\hat{\xi}_{p,n}+z_{\alpha}\frac{\sqrt{p(1-p)}\hat{\lambda}_{p,n}}{\sqrt{n}}\right) \tag{17}$$

Setting $\alpha=0.05$ (so $z_\alpha=1.645$), we then get that the upper endpoint of Eq. (17) is an asymptotically valid 95/95 value for SRS as $n\to\infty$ when $p=0.95$.

2.2.2. Variance reduction techniques

The goal of VRTs is to produce an estimator with smaller (asymptotic) variance than the standard estimator using SRS. This will result in an asymptotically shorter confidence interval. For the case when applying VRTs, Chu and Nakayama [28] have developed methods for constructing asymptotically valid CIs for the p-quantile ξ_p. Let $\tilde{F}_n$ be an estimate of the CDF F, where $\tilde{F}_n$ is obtained by simulating using a VRT with sampling budget n. We will give examples in Sections 2.2.3 and 2.24 for specific VRTs. Then a VRT p-quantile estimator is $\tilde{\xi}_{p,n}=\tilde{F}_n^{-1}(p)$. The asymptotic validity of the method in [28] for constructing a CI for ξ_p based on $\tilde{\xi}_{p,n}$ relies on showing that the VRT quantile estimator satisfies a Bahadur representation analogous to the SRS version in Eqs. (8) and (9). Specifically, let $\tilde{\xi}_{p_n,n}=\tilde{F}_n^{-1}(p_n)$ be the VRT p_n-quantile estimator, with p_n a perturbed value of p, and assume that $f(\xi_p)>0$. Then Chu and Nakayama develop a set of general conditions [28] on the VRT CDF estimator $\tilde{F}_n$ to ensure that

$$\tilde{\xi}_{p_n,n}=\dot{\xi}_{p,n}+\frac{p-\tilde{F}_n\left(\xi_p\right)}{f\left(\xi_p\right)}+\tilde{R}_n \tag{18}$$

with

$$\sqrt{n}\tilde{R}_n\Rightarrow 0 \text{ as } n\to\infty \tag{19}$$

where $\dot{\xi}_{p,n}$ is as in Eq. (10) when $p_n=p+O(1/\sqrt{n})$. If f is further assumed to be continuous in a neighborhood of ξ_p and a condition in [28] is slightly strengthened, then Eqs. (18) and (19) hold with $\dot{\xi}_{p,n}$ defined in Eq. (11) for all $p_n\to p$ as $n\to\infty$. Chu and Nakayama [28] show that their set of general conditions hold for importance sampling, combined importance sampling and stratification, antithetic variates and control variates; Nakayama [3] establishes the same for a type of Latin hypercube sampling. We later discuss the cases of AV and LHS.

As in the case of SRS in Section 2.2.1, the Bahadur representation in Eqs. (18) and (19) implies that the VRT p-quantile estimator satisfies a CLT

$$\frac{\sqrt{n}}{\kappa_p}\left(\tilde{\xi}_{p,n}-\xi_p\right)\Rightarrow N(0,1) \text{ as } n\to\infty \tag{20}$$

where

$$\kappa_p=\psi_p\lambda_p \tag{21}$$

ψ_p^2 is the asymptotic variance in the CLT

$$\sqrt{n}(p-\tilde{F}_n(\xi_p))\Rightarrow N(0,\psi_p^2) \tag{22}$$

for the VRT CDF estimator $\tilde{F}_n$ at ξ_p, and λ_p is defined in Eq. (14). The value of ψ_p depends on the particular VRT used and equals $\sqrt{p(1-p)}$ for SRS (compare Eqs. (13) and (20)). Our goal is to develop VRTs for which $\psi_p<\sqrt{p(1-p)}$, so that the asymptotic variance $\kappa_p^2=\left[\psi_p\lambda_p\right]^2$ in the CLT in Eq. (20) for the VRT p-quantile estimator is $\tilde{\xi}_{p,n}$ is smaller than the asymptotic variance $\left[\sqrt{p(1-p)}\lambda_p\right]^2$ in the CLT in Eq. (13) for the SRS p-quantile estimator $\hat{\xi}_{p,n}$. It turns out that developing a consistent estimator $\tilde{\psi}_{p,n}$ of ψ_p is often straightforward; Section 2.2.5 presents such estimators for specific VRTs. The value of λ_p is independent of the VRT applied, and we estimate it using a (central) finite difference

$$\tilde{\lambda}_{p,n}=\frac{\tilde{F}_n^{-1}(p+h_n)-\tilde{F}_n^{-1}(p-h_n)}{2h_n} \tag{23}$$

where $h_n>0$ is the bandwidth. Chu and Nakayama [28] prove that if their general set of conditions hold and $f(\xi_p)>0$, then

$$\tilde{\lambda}_{p,n}\Rightarrow\lambda_p \text{ as } n\to\infty \tag{24}$$

for $h_n=c/\sqrt{n}$ for any constant $c>0$. If we further assume f is continuous in a neighborhood of ξ_p, then Eq. (24) is true for bandwidths satisfying

$$h_n\to 0 \text{ and } \sqrt{n}h_n\to b \text{ for some } b\in(0,\infty] \text{ as } n\to\infty. \tag{25}$$

For example, $h_n=cn^{-\nu}$ satisfies Eq. (25) for constants $c>0$ and $0<\nu\le 1/2$. Thus, an asymptotically valid two-sided $100(1-\alpha)\%$ CI for ξ_p when applying a VRT is

$$\left[\tilde{\xi}_{p,n}\pm z_{\alpha/2}\frac{\tilde{\psi}_{p,n}\tilde{\lambda}_{p,n}}{\sqrt{n}}\right] \tag{26}$$

Also, an asymptotically valid upper one-sided $100(1-\alpha)\%$ CI for ξ_p is

$$\left(-\infty,\tilde{\xi}_{p,n}+z_{\alpha}\frac{\tilde{\psi}_{p,n}\tilde{\lambda}_{p,n}}{\sqrt{n}}\right) \tag{27}$$

whose upper endpoint is an asymptotically valid 95/95 value when applying the VRT with $\alpha=0.05$ and $p=0.95$.

2.2.3. Antithetic variates

Instead of generating independent outputs as in SRS, the method of antithetic variates (AV) generates outputs in negatively correlated pairs, which can reduce variance; (see Section V.3 of [25] for an overview of AV). We call (Y,Y') an *AV pair* if Y and Y' each have marginal distribution F and are negatively correlated. One way to simulate such a pair is to generate d i.i.d. uniform $[0,1]$ random variables $U_1,U_2,\ldots,U_d$, and then set $Y=g(U_1,U_2,\ldots,U_d)$ and $Y'=g(1-U_1,1-U_2,\ldots,1-U_d)$, where g is from Eq. (5). Clearly, Y has CDF F by Eq. (5), but Y' also does since each $1-U_j$ is also uniform $[0,1]$. If g is monotonic in each argument U_j, then Y and Y'

are guaranteed to be negatively correlated [29], which will ensure a variance reduction, as we shall see shortly. AV can still result in a variance reduction when g is not monotonic in each argument, but it may be difficult to prove.

To estimate ξ_p using AV, generate $n/2$ AV pairs (Y_i, Y_i'), $i=1,2,\ldots,n/2$, where n is even. We accomplish this by generating i.i.d. uniforms as in Eq. (6) but with $n/2$ rows rather than n. Then set $Y_i = g(U_{i,1}, U_{i,2}, \ldots, U_{i,d})$ and $Y_i' = g(1-U_{i,1}, 1-U_{i,2}, \ldots, 1-U_{i,d})$. We then compute the AV estimator $\tilde{F}_n$ of the CDF F as

$$\tilde{F}_n(y) = \frac{1}{\frac{n}{2}} \sum_{i=1}^{\frac{n}{2}} \frac{1}{2}\left[I(Y_i \le y) + I(Y_i' \le y)\right] \tag{28}$$

and the resulting AV estimator of ξ_p is $\tilde{\xi}_{p,n} = \tilde{F}_n^{-1}(p)$.

We now show that for each y, $\tilde{F}_n(y)$ has no greater variance than the SRS estimator $\hat{F}_n(y)$ in Eq. (2). Note that since the $n/2$ AV pairs are i.i.d.,

$$Var\left[\tilde{F}_n(y)\right] = \left(\frac{1}{n/2}\right)^2 \sum_{i=1}^{n/2} \frac{1}{4} Var[I(Y_i \le y) + I(Y_i' \le y)]$$

$$= \left(\frac{1}{n/2}\right)\frac{1}{4}\left(Var[I(Y \le y)] + Var[I(Y' \le y)] + 2Cov[I(Y \le y), I(Y' \le y)]\right)$$

$$= \left(\frac{1}{n/2}\right)\frac{1}{2}\left(F(y)(1-F(y)) + Cov[I(Y \le y), I(Y' \le y)]\right) \tag{29}$$

$$\le \frac{F(y)(1-F(y))}{n} = Var[\hat{F}_n(y)] \tag{30}$$

since the negative correlation of Y and Y' implies the same for $I(Y \le y)$ and $I(Y' \le y)$ because $q(x) = I(x \le y)$ is monotonic in x [29]. Thus, AV can reduce the variance of the estimator of $F(y)$ compared to SRS, which leads to the AV p-quantile estimator $\tilde{\xi}_{p,n}$ having smaller variance.

Avramidis and Wilson [30] develop this AV estimator of ξ_p, which they prove satisfies the CLT in Eq. (20), but they do not consider the estimation of the asymptotic variance κ_p^2 in the CLT in Eq. (20) to construct a CI for ξ_p. To address this issue, Chu and Nakayama [28] prove that the AV CDF estimator $\tilde{F}_n$ satisfies their general set of conditions. Thus, if $f(\xi_p) > 0$, then the AV quantile estimator satisfies the Bahadur representation in Eqs. (18) and (19) with $\dot{\xi}_{p_n}$ in Eq. (10) for $p_n = p + O(1/\sqrt{n})$. Moreover, Eqs. (18) and (19) hold with $\dot{\xi}_{p_n}$ in Eq. (11) for any $p_n \to p$ when f is also continuous in a neighborhood of ξ_p. In either case, this implies the AV p-quantile estimator $\tilde{\xi}_{p,n}$ satisfies the CLT in Eq. (20). To derive an expression for ψ_p^2 in Eq. (22) when applying AV, which is needed to determine κ_p in Eq. (21), note that by Eq. (28), $\tilde{F}_n(\xi_p)$ is the sample average of $n/2$ quantities $Z_i \equiv [I(Y_i \le \xi_p) + I(Y_i' \le \xi_p)]/2$. Since the Z_i, $i=1,2,\ldots,n/2$, are i.i.d. with finite variance, the CLT in Eq. (22) holds with $\psi_p^2 = 2Var[Z_i]$, where the factor of 2 is a result of the $n/2$ AV pairs. Hence, it follows from Eq. (29) that

$$\begin{aligned}\psi_p^2 &= Var[I(Y \le \xi_p)] + Cov[I(Y \le \xi_p), I(Y' \le \xi_p)] \\ &= p(1-p) + E[I(Y \le \xi_p) I(Y' \le \xi_p)] - E[I(Y \le \xi_p)]E[I(Y' \le \xi_p)] \\ &= p(1-2p) + P\{Y \le \xi_p, Y' \le \xi_p\}\end{aligned}$$

since $E[I(Y \le \xi_p)] = E[I(Y' \le \xi_p)] = p$. Chu and Nakayama [28] show that

$$\tilde{\psi}_{p,n}^2 = p(1-2p) + \frac{1}{\frac{n}{2}} \sum_{i=1}^{\frac{n}{2}} I(Y_i \le \tilde{\xi}_{p,n}, Y_i' \le \tilde{\xi}_{p,n}) \tag{31}$$

consistently estimates ψ_p^2, so $\tilde{\psi}_{p,n} \Rightarrow \psi_p$ as $n \to \infty$. Substituting Eq. (31) into (26) and (27) then results in asymptotically valid two-sided and one-sided $100(1-\alpha)\%$ CIs for ξ_p when applying AV.

Computing the AV quantile estimator and constructing the corresponding CI require inverting the AV CDF estimator in Eq. (28), which can be accomplished as follows. Define $A_{2i-1} = Y_i$ and $A_{2i} = Y_i'$ for $i=1,2,\ldots,n/2$. Let $A_{(1)} \le A_{(2)} \le \cdots \le A_{(n)}$ be the order statistics of the A_j, $j=1,2,\ldots,n$. Then for any $0 < q < 1$, we can compute $\tilde{F}_n^{-1}(q) = A_{(\lceil nq \rceil)}$.

2.2.4. Latin hypercube sampling

Originally developed by [31] and further analyzed by [32], LHS has been frequently applied in nuclear engineering (see [32]), although not presently for the calculation of 95/95 values in uncertainty analyses of LOCAs. LHS can be thought of as an extension of stratified sampling (Chapter 5 of [33]) in multiple dimensions, and it induces correlations among the outputs, which can reduce variance. Avramidis and Wilson [30] develop LHS quantile estimators, but they do not develop CIs based on the estimators. Nakayama [3] shows how the general framework of [28] applies to a type of replicated LHS (rLHS)[6], which is called combined multiple-LHS in [1], thus allowing the construction of an asymptotically valid CI for a quantile when using rLHS.

Rather than generating a single LHS sample of size n, the basic idea of rLHS in [3] is to generate the $n = mt$ samples as m independent LHS samples, each of size t. For each independent LHS sample $k=1,2,\ldots,m$, let $U_{i,j}^{(k)}$, for $1 \le i \le t$ and $1 \le j \le d$, be td i.i.d. uniform $[0,1]$ random variables, which we arrange as a $t \times d$ array

$$\begin{matrix} U_{1,1}^{(k)} & U_{1,2}^{(k)} & \cdots & U_{1,d}^{(k)} \\ U_{2,1}^{(k)} & U_{2,2}^{(k)} & \cdots & U_{2,d}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ U_{t,1}^{(k)} & U_{t,2}^{(k)} & \cdots & U_{t,d}^{(k)} \end{matrix}$$

Then let $\pi_j^{(k)} = (\pi_j^{(k)}(i) : i = 1,2,\ldots,t)$ for $1 \le j \le d$ and $1 \le k \le m$ be dm independent permutations of $(1,2,\ldots,t)$, which are also independent of the $U_{i,j}^{(k)}$. Thus, $\pi_j^{(k)}(i)$ is the value to which i is mapped in the permutation $\pi_j^{(k)}$. Then define

$$V_{i,j}^{(k)} = \frac{\pi_j^{(k)}(i) - 1 + U_{i,j}^{(k)}}{t} \text{ for } 1 \le i \le t \text{ and } 1 \le j \le d.$$

For each $1 \le k \le m$, arrange the $V_{i,j}^{(k)}$ into a $t \times d$ array

$$\begin{matrix} V_{1,1}^{(k)} & V_{1,2}^{(k)} & \cdots & V_{1,d}^{(k)} \\ V_{2,1}^{(k)} & V_{2,2}^{(k)} & \cdots & V_{2,d}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ V_{t,1}^{(k)} & V_{t,2}^{(k)} & \cdots & V_{t,d}^{(k)} \end{matrix} \tag{32}$$

It is straightforward to show that each $V_{i,j}^{(k)}$ has a uniform $[0,1]$ distribution. Moreover, by the independence of the permutations $\pi_1^{(k)}, \pi_2^{(k)}, \ldots, \pi_d^{(k)}$, the columns of Eq. (32) are independent. Thus, defining

$$\begin{aligned} Y_1^{(k)} &= g(V_{1,1}^{(k)}, V_{1,2}^{(k)}, \ldots, V_{1,d}^{(k)}), \\ Y_2^{(k)} &= g(V_{2,1}^{(k)}, V_{2,2}^{(k)}, \ldots, V_{2,d}^{(k)}), \\ &\vdots \\ Y_t^{(k)} &= g(V_{t,1}^{(k)}, V_{t,2}^{(k)}, \ldots, V_{t,d}^{(k)}), \end{aligned} \tag{33}$$

we have that each $Y_i^{(k)}$ has CDF F by Eq. (5) because $V_{i,1}^{(k)}, V_{i,2}^{(k)}, \ldots, V_{i,d}^{(k)}$ are i.i.d. uniform[0, 1]. But the rows in Eq. (32) are dependent because all of the entries in column j depend on the same permutation $\pi_j^{(k)}$, so $Y_1^{(k)}, Y_2^{(k)}, \ldots, Y_t^{(k)}$ are dependent. We call $Y_1^{(k)}, Y_2^{(k)}, \ldots, Y_t^{(k)}$ an *LHS case* of run size t. Now independently replicating this for

[6] Replicated LHS was first introduced by [43].

$k=1,2,\ldots,m$, leads to

$$\begin{matrix} Y_1^{(1)} & Y_1^{(2)} & \cdots & Y_1^{(m)} \\ Y_2^{(1)} & Y_2^{(2)} & \cdots & Y_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ Y_t^{(1)} & Y_t^{(2)} & \cdots & Y_t^{(m)} \end{matrix} \tag{34}$$

where each column in Eq. (34) corresponds to one LHS case of run size t, as in Eq. (33). The columns in Eq. (34) are independent since the m LHS cases are generated independently, but the entries within a column are dependent since they are generated using LHS. We call the $n=mt$ values in Eq. (34) an rLHS sample with m cases, each with run size t.

The rLHS estimator of the CDF F is then

$$\tilde{F}_{m,t}(y)=\frac{1}{mt}\sum_{k=1}^{m}\sum_{i=1}^{t} I(Y_i^{(k)}\le y) \tag{35}$$

and the rLHS p-quantile estimator is $\tilde{\xi}_{p,m,t}=\tilde{F}_{m,t}^{-1}(p)$. Nakayama [3] proves that if $f(\xi_p)>0$, then the following Bahadur representation holds:

$$\tilde{\xi}_{p_m,m,t}=\dot{\xi}_{p_m}+\frac{p-\tilde{F}_{m,t}\left(\xi_p\right)}{f\left(\xi_p\right)} +R_{m,t} \text{ with } \sqrt{m}R_{m,t}\Rightarrow 0 \text{ as } m\to\infty \text{ with } t \text{ fixed}, \tag{36}$$

where $\dot{\xi}_{p_m}=\xi_p+(p_m-p)/f(\xi_p)$ when $p_m=p+O(1/\sqrt{m})$. If in addition f is continuous in a neighborhood of ξ_p, then Eq. (36) holds with $\dot{\xi}_{p_m}=F^{-1}(p_m)$ for all $p_m\to p$ as $m\to\infty$. (These results are established in [3] by proving that the general set of conditions from [28] hold for rLHS.) It then follows that $\tilde{\xi}_{p,m,t}$ satisfies the CLT

$$\frac{\sqrt{m}}{\kappa_{p,t}}\left(\tilde{\xi}_{p,m,t}-\xi_p\right)\Rightarrow N(0,1) \text{ as } m\to\infty \text{ with } t \text{ fixed}, \tag{37}$$

where $\kappa_{p,t}$ has the form of κ_p in Eq. (21). To construct a CI for ξ_p based on Eq. (37), we need to develop an estimator for $\kappa_{p,t}=\psi_{p,t}\lambda_p$, where $\psi_{p,t}^2$ is the asymptotic variance in the CLT $\sqrt{m}\left(p-\tilde{F}_{m,t}\left(\xi_p\right)\right)\to N\left(0,\psi_{p,t}^2\right)$ as $m\to\infty$ with t fixed.

As before, the Bahadur representation in Eq. (36) allows us to develop a consistent estimator for λ_p. If $f(\xi_p)>0$, then

$$\tilde{\lambda}_{p,m,t}=\frac{\tilde{F}_{m,t}^{-1}(p+h_m)-\tilde{F}_{m,t}^{-1}(p-h_m)}{2h_m} \tag{38}$$

satisfies $\tilde{\lambda}_{p,m,t}\Rightarrow\lambda_p$ as $m\to\infty$ with t fixed when $h_m=c/\sqrt{m}$ for any constant $c>0$. If f is also continuous in a neighborhood of ξ_p, then $\tilde{\lambda}_{p,m,t}\Rightarrow\lambda_p$ as $m\to\infty$ for fixed t for any $h_m\ne 0$ satisfying $h_m\to 0$ and $\sqrt{m}h_m\to b$ for some $b\in(0,\infty]$ as $m\to\infty$.

To derive an expression for $\psi_{p,t}^2$ note that $\tilde{F}_{m,t}(y)=\frac{1}{m}\sum_{k=1}^{m}W^{(k)}(y)$, where

$$W^{(k)}(y)=\frac{1}{t}\sum_{i=1}^{t} I(Y_i^{(k)}\le y).$$

Now $W^{(1)}(\xi_p),W^{(2)}(\xi_p),\ldots,W^{(m)}(\xi_p)$ are i.i.d. with finite variance since $0\le W^{(k)}(\xi_p)\le 1$. Thus, $\tilde{F}_{m,t}(\xi_p)$ satisfies the CLT in Eq. (22) with

$$\psi_{p,t}^2=Var[W^{(k)}(\xi_p)]$$

Nakayama [3]develops
as a consistent estimator (as $m\to\infty$ with t fixed) of $\psi_{p,t}^2$, where

$$\overline{W}_m=\frac{1}{m}\sum_{k=1}^{m}W^{(k)}(\tilde{\xi}_{p,m,t}).$$

Substituting $\tilde{\xi}_{p,m,t}$, $\tilde{\lambda}_{p,m,t}$, and $\tilde{\psi}_{p,m,t}$ for $\tilde{\xi}_{p,n}$, $\tilde{\lambda}_{p,n}$, and $\tilde{\psi}_{p,n}$, respectively, in Eqs. (26) and (27) then results in asymptotically valid two-sided and one-sided $100(1-\alpha)\%$ CIs for ξ_p when applying rLHS.

Constructing the CIs requires inverting the rLHS CDF estimator in Eq. (35), which we can do as follows. Define $B_{(k-1)m+i}=Y_i^{(k)}$ for $i=1,2,\ldots,t$, and $k=1,2,\ldots,m$. Let $B_{(1)}\le B_{(2)}\le\cdots\le B_{(mt)}$ be the order statistics of the B_j, $j=1,2,\ldots,mt$. Then for any $0<q<1$, we can compute $\tilde{F}_{m,t}^{-1}(q)=B_{(mtq)}$.

Using rLHS with m cases of size t leads to some loss of statistical efficiency compared to a single LHS with $n=mt$ samples; however, [19] notes the degradation when estimating a mean is small if t/d is large. Also, there is a tradeoff between the amount of variance reduction from rLHS and the rate of the convergence of the confidence interval's coverage, where coverage is the likelihood that the true value of the parameter being estimated lies within the bounds of the confidence interval, and should approach $100(1-\alpha)\%$ as m grows larger (this definition of coverage is different than what has been used in past works by Nutt and Wallis [14]). If an analyst takes many cases of a small run size, meaning large m but small t, then the asymptotics will converge more quickly, since the large m will help satisfy the CLT in Eq. (37). But the small run sizes will not reduce the quantile estimator's variance by as much as large run sizes would. As the run size t increases, the quantile estimator will have lower variance, but to remain at the same number n of total runs, the number m of cases must be reduced, and the coverage can suffer.

2.2.5. Finite-difference estimator and bandwidth selection

Sections 2.2.1–2.2.4 considered central finite-difference (CFD) estimators Eqs. (16), (24) and (38) to estimate the derivative λ_p in Eq. (14). Implementing these estimators in practice requires the user to specify the bandwidth h_n (or h_m), and the particular choice for h_n can have a large impact on the quality of the sparsity estimators. Previous work with asymptotic SRS provides some guidance on the choice for h_n. For example, [26] and [27] show that under certain conditions, taking $h_n=c_1n^{-1/5}$ for some constant c_1 asymptotically minimizes the mean-square error of the CFD estimator of λ_p. Also, the coverage error of CIs can be asymptotically minimized by taking $h_n=c_2n^{-1/3}$ for some constant c_2; see [34]. The values of c_1 and c_2 depend on the CDF F and p, and these papers provide data-based methods for estimating c_1 and c_2.

The CFD estimators in Eqs. (16), (24) and (38) are each symmetric in the sense that the inverse of the estimated CDF is evaluated at perturbed values that are symmetric about p. However, the symmetric CFD estimator often overestimates λ_p when $p\approx 1$, as when we estimate the 0.95-quantile. To see why, suppose that the CDF $F(y)$ has a density $f(y)$ that is differentiable and strictly decreasing for all y sufficiently large. (This is true for many common distributions, including the normal, lognormal, gamma and Weibull.) Thus, the density's derivative $f'(y)<0$ for all sufficiently large y. Then defining $Q(p)=F^{-1}(p)$ as the quantile function, we see that its first derivative $Q'(p)=\lambda_p=1/f(\xi_p)>0$ and its second derivative $Q''(p)=-f'(\xi_p)/f^3(\xi_p)>0$ for $p\approx 1$. Hence, as shown in

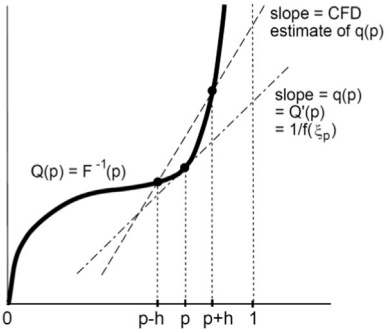

Fig. 2. Overprediction of derivative using CFD.

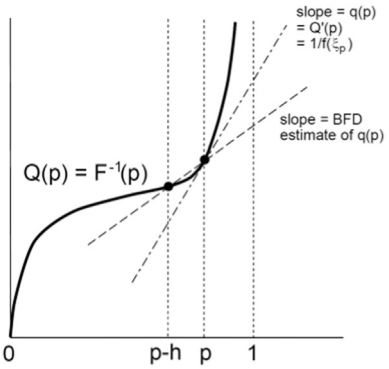

Fig. 3. Underprediction of derivative using BFD.

Fig. 2, $Q(p)$ is increasing and convex for $p \approx 1$. This leads to the symmetric CFD typically overestimating λ_p because $Q(p)$, whose derivative $Q'(p) = \lambda_p$ we are trying to estimate with the finite difference, has slope that is increasing as p increases. Similarly, the backward finite-difference (BFD) estimator $[\tilde{F}_n^{-1}(p) - \tilde{F}_n^{-1}(p-h_n)]/h_n$ will often underestimate λ_p, as shown in Fig. 3.

This suggests that we may more accurately estimate λ_p by using an *asymmetric* CFD estimator $[\tilde{F}_n^{-1}(p+h_n) - \tilde{F}_n^{-1}(p-h_n')]/(h_n+h_n')$, where $h_n' \neq h_n$. From Fig. 2, we see that choosing $h_n' > h_n > 0$ may be beneficial because the slope of $Q(p)$ increases as p grows. While we have carried out some experiments testing such an estimator, the results are not presented here in favor of more generic methods. (Other estimators of λ_p are alsok possible; for example, Fal [35] develops a kernel estimator of λ_p for SRS, and Nakayama [36] considers another type of kernel estimator when using importance sampling.)

The Appendix includes results when λ_p is estimated using a symmetric CFD and also when we use an "exact" value for λ_p, which was determined using a large-run SRS trial. This allows us to measure the effect on the CIs of having to estimate λ_p.

3. Experiments

References [3] and [28] run experiments on a small stochastic activity network (SAN) to study the behavior of the CIs in Eq. (26). However, since the present work focuses on the use of these methods in nuclear safety analysis, we devised experiments that would more closely mimic common safety-analysis situations. This included starting with a simple nonlinear equation, moving to a response-surface surrogate for the thermal-hydraulic computer code RELAP5, and using a large severe-accident analysis computer code to analyze a beyond-design-basis accident. The results presented here will focus on the comparison between SRS-OS and rLHS, since rLHS generally offers greater reduction in variance than AV, but the results when using AV and asymptotic SRS are also presented.

3.1. Nonlinear equation

The first test conducted used an equation found in previous literature on sampling schemes [37]. This equation is simply a statistical test; it has no physical meaning. It is used due to its complexity and since it is difficult to model accurately with a second-order response surface. This equation was chosen as a first step to see how the proposed methods would perform with a nonlinear equation, since many such equations are found in large severe-accident computer codes. The equation is defined as

$$Y = 5 + (2+9X_1)^{0.7}\ln\left(2+2X_3+X_3^2\right) + (1+2X_3)^{1.2}e^{X_4^2} + X_2^2 \qquad (39)$$

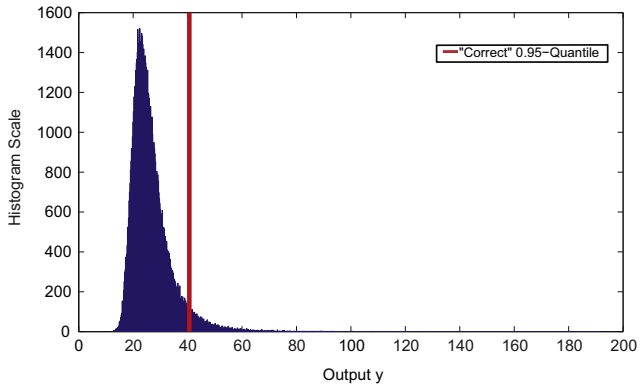

Fig. 4. Histogram of 10^5 run SRS trial – nonlinear equation, truncated normal inputs.

where the uncertain independent input parameters $0 \le X_1, X_2, X_3, X_4 \le 2$ (the equation was designed in [37] for input values between 0 and 2), and Y is considered the output of interest, which would be compared to a prescribed safety limit. The following two experiments will use different distributions for X_1, X_2, X_3, X_4.

3.1.1. Normal inputs

For this test, we assumed the four inputs to be independent truncated normal random variables with mean 1.0 and standard deviation 0.22639. We chose these distribution parameters in order for 99.999% of the non-truncated normal distribution to fall between 0 and 2. First, we conducted a SRS experiment with 10^8 runs in order to estimate the correct[7] 0.95-quantile of the system. The 0.95-quantile was chosen to see how these methods would perform when trying to satisfy the 95/95 criterion. The result was an estimated 0.95-quantile of 40.6457, which we considered as the "correct" 0.95-quantile.

We found this quantile in order to determine the distance between the calculated 95/95 values and the "correct" quantile. This distance would be considered a measurement of the accuracy of the 95/95 values. It is important to point out that poor accuracy, as defined here, does not mean that the 95/95 values are not valid, but that a 95/95 value that falls at a much higher value than the correct 0.95-quantile is not desired and could potentially lead to Type-II errors, or incorrect safety-analysis decisions. Fig. 4 shows a histogram of the outputs Y for a 10^5-run SRS trial, which is shown simply to give the reader an idea of the range of possible outputs. Most of the mass of the distribution is concentrated at lower levels, but there is a long tail to the right, so the p-quantile for $p \gg 0.95$ is considerably larger than the 0.95-quantile. As mentioned in Section 2.1, the resulting 95/95 value of an analysis conducted using SRS-OS is more likely to be near a higher quantile value when r is small. Thus, it is likely that the 95/95 value found using SRS-OS will fall at a significantly higher value than the 0.95-quantile.

Each method to calculate a 95/95 value was repeated for 10^4 trials. Here, a trial is a complete experiment that would be undertaken during a safety analysis. For example, one SRS-OS trial may consist of 59 computer code runs to obtain a single 95/95 value as the largest of the 59 outputs. For each method, 10^4 trials were conducted so that the spread of possible 95/95 values could be found. This gives information about the precision of each method. Once again, poor precision does not mean that the 95/95

[7] The 0.95-quantile found by the large SRS experiment is considered "correct," as it is numerically correct based on the input assumptions. It is not considered "true" as there may be epistemic uncertainties that affect the shape or range of the input distributions, which are not examined here.

Table 2
MSE error results[a].

	v	1/2											
	c	0.5			0.4			0.3			0.2		
	t	10	20	30	10	20	30	10	20	30	10	20	30
n	59			X			X	140%	142%	X	85%	84%	X
	93				157%	151%	157%	81%	81%	77%	67%	72%	62%
	124	169%	170%	172%	82%	83%	83%	62%	60%	58%	58%	56%	51%
	311	40%	36%	37%	36%	33%	32%	37%	33%	33%	39%	39%	38%
	548	27%	26%	24%	28%	26%	24%	30%	28%	28%	34%	34%	33%
	1008	20%	20%	19%	22%	21%	20%	24%	24%	23%	30%	27%	27%
	2004	16%	15%	15%	18%	17%	17%	20%	20%	19%	25%	24%	23%
	4000	13%	13%	12%	14%	14%	14%	17%	16%	16%	20%	20%	20%
		1/3											
		0.2			0.17			0.15			0.1		
		10	20	30	10	20	30	10	20	30	10	20	30
n	59			X	145%	139%	X	139%	142%	X	88%	85%	X
	93	157%	154%	157%	81%	157%	77%	77%	78%	78%	68%	74%	59%
	124	86%	84%	81%	83%	85%	83%	62%	63%	59%	58%	56%	51%
	311	40%	39%	37%	37%	34%	33%	37%	34%	31%	37%	36%	34%
	548	28%	27%	25%	27%	26%	24%	27%	26%	24%	29%	29%	28%
	1008	20%	18%	18%	20%	18%	18%	21%	19%	19%	24%	22%	22%
	2004	15%	14%	13%	15%	14%	14%	16%	15%	14%	19%	18%	18%
	4000	11%	11%	10%	12%	11%	11%	12%	12%	11%	14%	14%	14%

[a] Blank cells indicate that the combination of v and c would return a value above the 1.00-quantile, and "X" represents a combination of m and t that would result in $m = 2$, which is not desired.

values are invalid, but a technique that provides these values over a large range is undesirable, especially if decisions are to be made about the system based on the results.

Each method was tested at several different total run levels. These run levels are based off the results for n in Eq. (4) for SRS-OS. They start with the lowest possible run level using SRS-OS ($r = 1$), and increase from there. For rLHS, several representative values were chosen for the run size t for which $mt \approx n$. This gives information about the tradeoff between the sizes of m and t, as noted at the end of Section 2.2.4.

The derivative λ_p, as described in Eq. (14) of Section 2.2.5, is estimated using a CFD estimator, and determining the proper bandwidth h_m is not trivial. Small changes in the bandwidth parameters c and v in $h_m = c(mt)^{-v}$ can greatly impact the calculation of the CI, especially when $n = mt$ is small. Also, while there has been some guidance provided on the selection of these parameters when using the CFD to establish asymptotic CIs during a SRS simulation, there is less direction when using VRTs since asymptotic CIs have only recently been mathematically proven. Nakayama [3] provides some insight into the selection of v from limited experiments with a SAN example, which appeared to show $v = 1/2$ was efficient for estimating quantiles close to 1. For this experiment, a range of values for v and c were chosen in order to find the combination that presented the most accurate estimate of λ_p. For each of these combinations, the mean-square error (MSE) of the estimated value of $\lambda_p$was found using 10^4 trials. These results can be seen in Table 2, which show that the combination of $v = 1/3$ and $c = 0.15$ provided the smallest MSE over most run sizes. The MSE was calculated using a "correct" value for λ_p that was found using a CFD and a large (10^8) SRS trial. The number n of runs conducted was based on levels for SRS-OS, and it may not have been possible for AV (which needs an even number of runs) and rLHS (which used several different values for t in this work) to achieve that exact number. Therefore, run values for AV and rLHS were chosen as close as possible to those needed by SRS-OS. For rLHS, this was done by dividing the number of runs necessary for SRS-OS by the number chosen for t, and then taking the closest whole number for the value for m. For example, if 59 SRS-OS runs were conducted, then for the rLHS method with $t = 10$, $59/10 = 5.9$, so $m = 6$ is chosen, resulting in 60 total runs. Once again, representative values were chosen for t, but these are not the only options. Certain run levels were not conducted for some sizes of t, since the value for m would have been equal to two, and too low to properly satisfy the CLT. That is why there are several "X" cells on the table at low run levels.

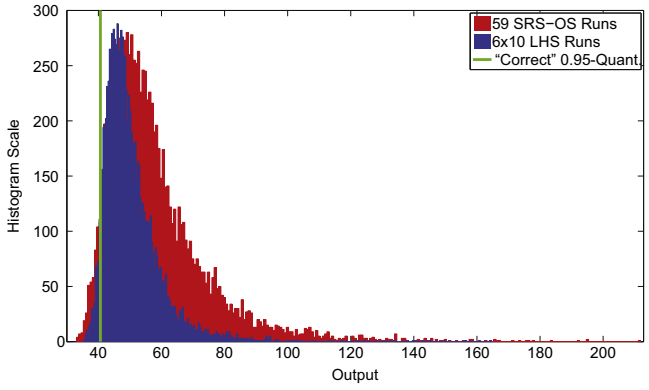

Fig. 5. Comparison of 95/95 value histograms for 10^4 trials – nonlinear equation, truncated normal inputs.

These results are important because at low run levels, small changes in the selected value for c can cause completely different qualitative results. This is due to the coarseness of the estimated CDF, since it is constructed with relatively small sample size. So a small change in the value of c can mean the values selected from the inverse CDF for the CFD estimator could differ by a wide margin. As the number of runs grows large (> 500), the selection of c has less impact (although not trivial) since the estimated CDF is more accurate. The optimal choices for c and v depend on the particular system being tested, the goal is to try to determine values which work well for many systems, and that are not problem-specific. Therefore, the values of $v = 1/3$ and $c = 0.15$ were used for all of the following experiments. For each

Table 3
Non-linear equation with truncated normal inputs Comparison of 95/95 values for 10^4 trials of rLHS and SRS-OS ("correct" 0.95-quantile = 40.6457).

	6×10 rLHS	59 SRS-OS
Mean of 10^4 95/95 values	50.13	57.79
S.D. of 10^4 95/95 values	8.52	16.52
% below "correct"	5.53%	5.10%

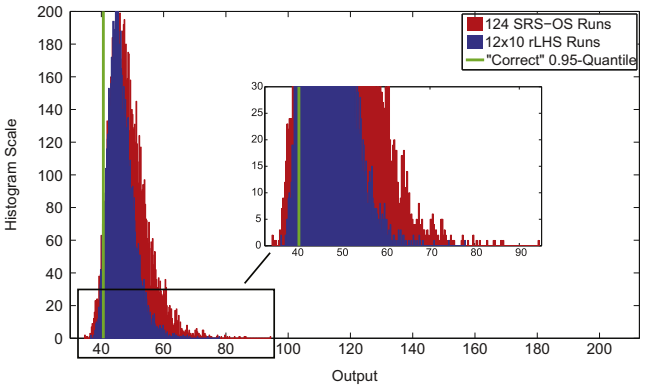

Fig. 6. Comparison of 95/95 value histograms for 10^4 trials – nonlinear equation, truncated normal inputs.

Table 4
Non-linear equation with truncated normal inputs Comparison of 95/95 values for 10^4 trials of rLHS and SRS-OS ("correct" 0.95-quantile = 40.6457).

	12×10 rLHS	124 SRS-OS
Mean of 10^4 95/95 values	45.33	48.80
S.D. of 10^4 95/95 values	3.59	6.14
% below "correct"	6.88%	4.90%

experiment, a detailed look at particular run levels is presented first, followed by the complete numerical results.

As stated, the first run-level conducted was based on the lowest value for which a 95/95 value could be found using SRS-OS. Fig. 5 shows a comparison of the histograms of the 10^4 95/95 values for SRS-OS at 59 runs and rLHS at 60 runs ($m=6$, $t=10$). The numerical results are in Table 3. It is important to note that these are not histograms of the output Y of the system, but rather for the 95/95 values for 10^4 complete trials. This means that 59 SRS-OS runs were conducted for each trial, and each trial resulted in one 95/95 value. As the results in Table 3 show, the rLHS method was not only more accurate, with a mean of 50.13 compared to a mean of 57.79 using SRS-OS, but also more precise, with about half the standard deviation. Both methods had ~5% of trials fall below the "correct" quantile[8] (called % below "correct" in Table 3), which is to be expected with a 95% confidence level. These results mean there is less chance that rLHS will cause Type-II testing errors, since rLHS overestimates the correct quantile by a smaller margin in the remaining 95% of trials.

Fig. 6 and Table 4 show the same results, but now for 124-run SRS-OS trials and 120-run rLHS trials ($m=12$, $t=10$). The scale of the horizontal axis in Fig. 6 is kept the same as in Fig. 5 to show the reduction in variance that naturally occurs with increased run size. The trend of rLHS being both more accurate and more precise continues at this higher run level, with rLHS having a mean value closer to the correct quantile and a smaller standard deviation than SRS-OS.

[8] The "correct" 0.95-quantile is 40.6457, as found in Fig. 4.

Before viewing the complete numerical results, there are several important points to note. First, the run levels on based on those needed for SRS-OS, as explained above. Second, five metrics are provided for the numerical results of the asymptotic SRS, AV, and rLHS methods, which are computed from the 10^4 trials. These include:

1. The mean of the 95/95 values over all trials.
2. The standard deviation of the 95/95 values over all trials.
3. The percentage of trials that fell below the "correct" quantile (this is expected to be ~5%).
4. The coverage, or the percent of trials where the "correct" quantile falls within the constructed 90% two-sided confidence interval for ξ_p (this confidence interval is defined in Eq. (15) for SRS, and Eq. (26) for AV and rLHS). Ideally, its value is ~90%.
5. The average value for the derivative estimator $\hat{\lambda}_{p,n}$ or $\tilde{\lambda}_{p,m}$ over all trials.

For SRS-OS, only the first three values are given since no derivative estimation is necessary (as shown in Section 2.1), and only a one-sided CI was computed (although two-sided CIs are also possible). Lastly, all tables include a comparison between asymptotic SRS, AV, and rLHS using a central finite-difference estimator for the derivative λ_p and an exact value for λ_p. This is presented in order to measure the effect of having to estimate the derivative λ_p.

The complete results for the experiment for calculating 95/95 values with the nonlinear equation using truncated normal inputs, can be found in Table 5.

As the table shows, SRS-OS and asymptotic SRS approach approximately the same results as the number of runs grows large. This is to be expected, as the two methods are asymptotically equivalent. When applied with the exact value for λ_p at the lowest run level ($n=59$), asymptotic SRS has fairly poor coverage. This may seem odd since the only part of Eq. (17) left to estimate is the quantile. However, at this low run level, the quantile estimator may be biased (*i.e.*, its expectation is not the true quantile). This is compounded by the use of the round-up function (see the start of Section 2), which has a larger impact at low run levels and may center the confidence interval at a biased estimate. So the coverage appears poor. When using asymptotic SRS with an estimated λ_p at this run level, the central finite difference overestimates λ_p (see Section 2.2.5 and Fig. 2); this causes the CI to increase in size, compensating for the effects of the poor quantile estimation.

The complete table of results also shows the tradeoff between run size t and number m of cases when using rLHS. The more cases (*i.e.*, large m, which corresponds to small t), the quicker the result approaches the proper coverage, since it is m that grows large in the asymptotics for the CLT. However, the larger the run size t, the more variance reduction will be seen when compared to SRS. As the table shows, the results when $t=10$ tend to converge the fastest, but the accuracy and precision of the result improves when using $t=20$ and $t=30$. Lastly, AV does appear to show a very slight variance reduction at large n when compared to SRS and SRS-OS, but not to the extent of rLHS.

3.1.2. Non-normal distributions

Next, the experiment with Eq. (39) was repeated, but the truncated-normally distributed inputs were replaced with a variety of distributions

$$\begin{aligned} &X_1 - \text{exponential}(0.1639) \\ &X_2 - N\left(1.0,\ 0.22639^2\right) \\ &X_3 - \text{lognormal}(-0.5, 0.3068) \\ &X_4 - \text{uniform}(0, 2) \end{aligned} \tag{40}$$

This was done to remove any possible influence from the use of normal distributions. Once again, the exponential and normal

Table 5
Comparison of 95/95 values for 10^4 trials, non-linear equation with truncated normal inputs ("correct" 0.95-quantile = 40.6457).

n	SRS-OS	CFD for λ_p					Exact λ_p					
		SRS	AV	rLHS			SRS	AV	rLHS			
				$t=10$	$t=20$	$t=30$			$t=10$	$t=20$	$t=30$	
59	57.79	54.11	51.38	50.13	48.62		49.60	46.78	45.99	44.80		**Mean**
	16.52	11.57	10.15	8.52	7.66		5.33	4.04	3.16	3.05		**S.D.**
	5.10	5.39	7.84	5.53	5.29		0.95	2.76	2.25	6.41		**% Below**
		89.81	90.90	93.46	93.64		85.69	93.13	93.34	90.68		**Covg.**
		253.50	259.52	267.12	278.67		156.60	156.60	156.60	156.60		**Avg.** $\tilde{\lambda}_p$
93	50.96	49.05	49.97	47.46	45.48	45.98	46.86	47.71	45.83	44.18	44.53	
	8.45	6.69	7.10	5.14	4.16	4.01	3.74	3.74	2.89	2.65	2.63	
	5.48	6.14	4.98	4.36	8.14	4.76	2.07	2.07	1.25	8.22	5.65	
		90.70	93.07	92.29	89.28	89.37	89.69	89.69	89.89	87.44	84.72	
		215.25	215.51	212.36	210.13	221.60	156.60	156.60	156.60	156.60	156.60	
124	48.80	46.88	46.92	45.33	44.39	44.22	45.47	45.52	44.39	43.61	43.44	
	6.14	5.01	4.89	3.59	2.97	2.89	3.03	3.03	2.24	2.11	2.15	
	4.90	7.95	7.07	6.88	7.79	8.65	3.81	2.72	2.90	7.79	9.55	
		89.76	91.14	91.28	88.88	87.90	90.71	91.82	92.67	86.37	85.30	
		200.21	200.92	194.48	199.35	200.08	156.60	156.60	156.60	156.60	156.60	
548	43.45	43.22	43.18	42.72	42.24	42.15	43.04	43.02	42.56	42.14	42.07	
	1.89	1.80	1.78	1.38	1.10	1.04	1.49	1.43	1.10	0.95	0.93	
	5.62	6.65	6.45	4.87	6.26	6.96	4.36	3.51	3.17	5.26	6.03	
		90.06	90.50	90.73	89.90	89.20	89.47	90.61	89.82	89.33	88.87	
		168.31	168.03	169.90	168.33	167.97	156.60	156.60	156.60	156.60	156.60	
1008	42.62	42.52	42.49	42.10	41.81	41.75	42.39	42.38	42.03	41.77	41.71	
	1.27	1.28	1.25	0.96	0.77	0.74	1.08	1.06	0.82	0.70	0.67	
	5.54	6.35	6.41	5.44	6.00	6.21	4.76	4.14	3.66	5.07	5.44	
		90.01	89.85	90.42	90.09	89.76	89.75	90.37	89.89	89.35	89.29	
		165.09	164.74	165.59	163.58	164.93	156.60	156.60	156.60	156.60	156.60	

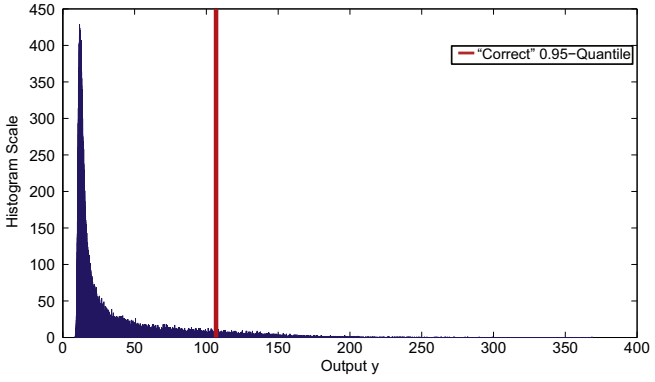

Fig. 7. Histogram of 10^5 run SRS trial–nonlinear equation, non-normal inputs.

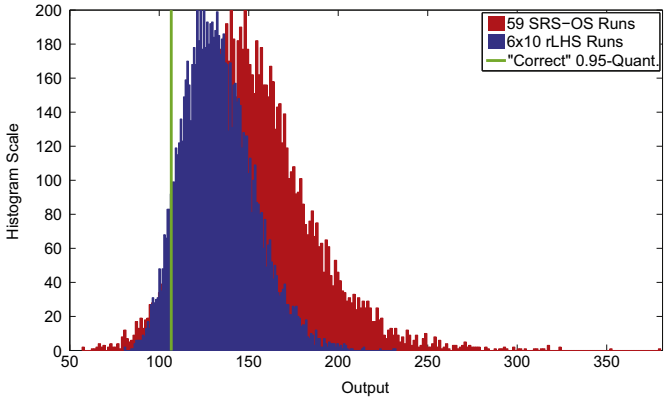

Fig. 8. Comparison of 95/95 value histograms for 10^4 trials – nonlinear equation, non-normal inputs.

distributions were truncated at 0 and 2, and the parameters were chosen so that 99.999% of the non-truncated distributions fall within that interval. Since the lognormal distribution does not extend below 0, 99.995% of the non-truncated distribution falls below 2, and 100% of the uniform distribution is between those bounds. As before, a 10^8-run SRS experiment was conducted to determine the "correct" quantile. Here, the 0.95-quantile was found to be 106.727. Fig. 7 shows the histogram of the outputs Y from a 10^5-run SRS trial. Once again, this is done to give the reader an idea of the output-distribution shape. Compared to the previous example (Fig. 4), this output has a fatter right tail, with values extending beyond y values of 200. This results in the 0.95-quantile lying at a much higher value than the median of the distribution. As in the example in Section 3.1.1, the expectation is that SRS-OS will result in a 95/95 value that will fall significantly higher than the 0.95-quantile, although the distributions were not purposely chosen for this overestimation to occur.

The same procedure as the previous example was followed, starting with a comparison between 10^4 trials of 59 SRS-OS runs and 60 rLHS runs ($m=6$, $t=10$). Fig. 8 and Table 6 have these results. The trend continues with rLHS being more precise and accurate. However, at this level over 6% of rLHS trials fell below the "correct" quantile. This could again be a sign that the asymptotics have not converged yet for proper coverage, or that the values for c and v are not appropriate. Nevertheless, even though the rLHS has a greater percentage of 95/95 values falling below the "correct"

Table 6
Non-linear equation with non-normal inputs Comparison of 95/95 values for 10^4 trials of rLHS and SRS-OS ("correct" 0.95-quantile = 106.727).

	6×10 rLHS	59 SRS-OS
Mean of 10^4 95/95 values	133.74	153.12
S.D. of 10^4 95/95 values	19.97	31.63
% below "correct"	6.74%	4.82%

Table 7
Comparison of 95/95 values for 10^4 trials, non-linear equation with non-normal inputs ("correct" 0.95-quantile = 106.727)

n	SRS-OS	CFD for λ_p					Exact λ_p					
		SRS	AV	rLHS			SRS	AV	rLHS			
				$t=10$	$t=20$	$t=30$			$t=10$	$t=20$	$t=30$	
59	153.12	150.60	138.87	133.74	128.26		145.13	132.92	129.10	123.36		**Mean**
	31.63	28.37	25.45	19.97	16.89		21.11	19.31	13.94	12.93		**S.D.**
	4.82	4.93	9.39	6.74	7.17		3.71	9.23	3.37	7.78		**% Below**
		88.97	88.17	91.65	91.47		89.34	89.12	94.02	90.11		**Covg.**
		862.83	877.21	884.15	912.75		745.73	745.73	745.73	745.73		**Avg.** $\tilde{\lambda}_p$
93	139.86	137.51	140.53	131.28	122.13	125.49	134.28	138.22	129.58	119.33	123.39	
	21.43	20.48	21.05	16.50	13.61	13.35	16.81	16.14	12.49	12.90	11.23	
	5.01	6.24	4.72	5.10	11.40	5.56	5.25	2.52	1.80	16.12	5.24	
		88.74	91.54	91.23	85.79	87.56	90.14	94.48	91.96	80.99	86.86	
		834.88	821.54	812.72	825.51	839.91	745.73	745.73	745.73	696.73	745.73	
124	133.98	130.08	130.10	124.18	119.62	119.08	127.74	128.14	123.14	118.48	118.07	
	17.25	16.67	16.41	12.96	11.06	10.71	14.22	14.17	10.41	9.83	9.62	
	5.01	7.84	7.18	7.66	10.82	11.08	7.36	6.75	4.27	11.57	11.44	
		88.08	89.61	89.86	85.45	84.98	90.13	90.53	92.43	83.68	83.42	
		815.79	807.92	796.62	813.70	818.72	745.73	745.73	745.73	745.73	745.73	
548	118.18	118.03	117.93	115.61	113.17	112.92	117.59	117.69	115.39	112.98	112.67	
	7.27	7.37	7.22	5.81	4.66	4.46	7.02	6.79	5.13	4.30	4.20	
	5.74	6.19	5.93	5.83	7.88	7.74	6.22	5.07	4.03	7.23	7.54	
		89.32	90.16	89.20	88.24	88.65	89.70	90.83	90.06	88.27	87.78	
		767.65	768.22	770.07	769.98	770.27	745.73	745.73	745.73	745.73	745.73	
1008	115.26	115.07	115.00	113.08	111.51	111.28	114.96	114.79	113.02	111.44	111.16	
	5.34	5.41	5.28	4.07	3.34	3.17	5.19	5.04	3.75	3.09	3.04	
	5.69	6.18	5.46	5.53	7.19	7.16	5.60	5.25	4.51	6.38	6.87	
		88.93	89.62	90.03	88.87	89.02	89.75	90.18	89.98	89.21	88.51	
		761.28	762.31	761.94	762.57	764.50	745.73	745.73	745.73	745.73	745.73	

quantile, these values still do not fall as low as values when using SRS-OS.

The complete numerical results are in Table 7. As the table shows, the rLHS method using $t=10$ has $>5\%$ of results "below correct" at $n=59$, but is closer to 5% at the next highest run level of $n=93$. However, at the $n=124$ level, over 7% of trials fall below the correct quantile even though the coverage is near 90%. A closer inspection shows that this is also the case for AV and asymptotic SRS, and by the next highest run level ($n=548$), the percent below correct is back closer to 5%.

Also, Table 7 shows that while the rLHS method with larger run sizes t provides variance reduction, it takes much longer to converge to the proper coverage level and proper percent below the correct quantile value than rLHS with smaller t. The coverage level converges to ~90% fairly quickly, but the percent below the correct quantile stays above 5% even at higher run levels.

3.2. Loss of Coolant Accident (LOCA) response surface

The next system analyzed was a second-order response surface, developed by French [38], which models the peak clad temperature of a nuclear power reactor during a LOCA. Historically, the large LOCA has represented the most extreme challenge as the design basis for a plant's emergency core cooling system. For best-estimate plus uncertainty analysis, a 95/95 criterion has been imposed on the entire spectrum of LOCA break sizes. Because the likelihood of a very large pipe break is extremely small, risk-informed requirements for emergency core cooling systems may be appropriate. In this approach, a transition break size would be established based on the expected frequency of breaks as a function of size. Below the transition break size, the 95/95 criterion would still be imposed. Above the transition break size, less conservatism would be applied (such as a 95/75 criterion). The LOCA response curve was chosen as a surrogate for the RELAP5 [4] plant deck from which it was created. While it is a relatively simple

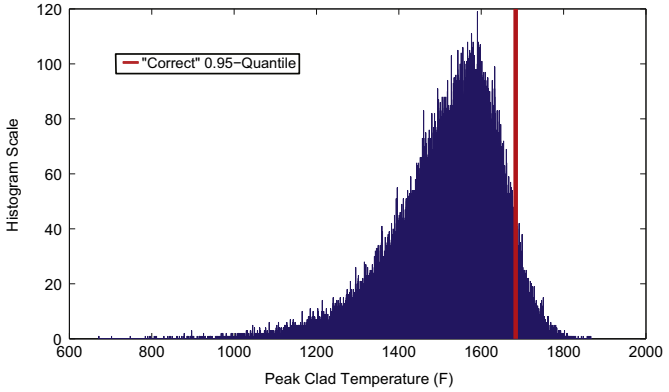

Fig. 9. Histogram of 10^5 run SRS trial – LOCA response surface.

equation, it is a step towards a realistic nuclear safety analysis,

$$Y = a + b_1X_1 + \ldots + b_{11}X_{11} + c_1X_1^2 + \ldots + c_{11}X_{11}^2 + d_1X_1X_2 + \ldots + d_{55}X_{11}X_{10}\, X_1, \ldots, X_{11} \sim N\left(0.5, 0.1^2\right) \quad (41)$$

where a, b_i, c_i, and d_i are constant coefficients and the eleven inputs $X_1, \ldots, X_{11}$ are independent normal random variables which are truncated at 0 and 1. The inputs represent certain reactor properties that were normalized between 0 and 1. The output Y is peak cladding temperature (PCT) in degrees Fahrenheit.

The result of a 10^8-run SRS experiment yielded a "correct" 0.95-quantile of 1683.65 °F. Fig. 9 shows the histogram of the output Y from a 10^5-run SRS trial. In this case, the upper tail decays very quickly, so the 0.95-quantile is relatively close to the mean.

For this example, only the complete list of results in Table 8 are given. As the table shows, for 59-run SRS-OS trials and 60-run rLHS trials ($m=6$, $t=10$), the rLHS method continues to show better performance than SRS-OS, but less so than in the previous examples. This is most likely a result of the output distribution shape. Unlike the previous examples, the tail of the output

Table 8
Comparison of 95/95 values for 10^4 trials, LOCA response surface ("correct" 0.95-quantile = 1683.63 °F).

n	SRS-OS	CFD for λ_p					Exact λ_p					
		SRS	AV	rLHS			SRS	AV	rLHS			
				$t=10$	$t=20$	$t=30$			$t=10$	$t=20$	$t=30$	
59	1730.89	1727.25	1715.88	1715.81	1712.99		1720.44	1708.39	1708.63	1707.53		**Mean**
	31.83	28.40	24.40	23.42	26.29		20.19	17.83	16.59	22.91		**S.D.**
	4.73	4.56	7.84	6.85	10.37		2.96	7.71	6.25	13.31		**% Below**
		89.57	90.04	90.95	86.26		88.70	90.04	91.24	82.74		**Covg.**
		841.04	857.06	863.77	866.08		695.70	695.70	695.70	739.92		**Avg.** $\tilde{\lambda}_p$
93	1717.33	1714.03	1717.00	1710.26	1706.22	1707.43	1710.18	1713.55	1707.18	1703.63	1704.68	
	21.67	20.69	21.04	18.06	20.16	19.75	16.02	15.48	13.95	15.75	15.92	
	4.80	5.92	4.11	5.62	11.83	8.97	4.68	2.05	3.74	9.90	8.54	
		89.24	92.77	89.67	83.26	82.51	89.30	93.98	90.26	84.93	81.35	
		802.00	797.65	784.91	774.27	788.84	695.70	695.70	695.70	695.70	695.70	
124	1711.62	1706.48	1706.84	1703.62	1702.21	1701.25	1703.95	1704.08	1701.51	1700.47	1699.77	
	17.90	16.37	16.43	14.36	15.01	15.61	13.42	13.00	11.66	12.09	12.99	
	4.69	7.32	7.23	7.02	9.55	11.76	6.32	5.06	5.61	8.23	10.51	
		89.49	89.50	90.14	86.38	82.93	89.90	91.59	90.78	87.31	83.95	
		779.75	775.82	765.40	763.86	762.51	695.70	695.70	695.70	695.70	695.70	
548	1694.88	1694.42	1694.32	1693.12	1692.40	1692.25	1694.09	1694.00	1692.23	1692.00	1691.93	
	7.10	7.12	6.98	6.09	6.01	6.03	6.49	6.29	6.11	5.46	5.49	
	5.51	5.81	6.05	5.54	6.68	7.07	5.57	4.79	7.56	6.02	6.44	
		89.90	89.63	89.89	89.39	88.68	89.74	90.63	86.81	89.63	88.67	
		719.18	720.44	721.75	720.37	719.95	695.70	695.70	655.47	695.70	695.70	
1008	1691.85	1691.59	1691.45	1690.56	1690.06	1690.02	1691.40	1691.26	1690.42	1689.89	1689.88	
	5.10	5.16	5.03	4.37	4.30	4.31	4.80	4.68	4.06	3.98	3.99	
	5.25	5.88	5.78	5.14	6.63	6.73	5.38	5.21	4.41	5.61	5.82	
		89.38	89.64	90.27	89.43	89.01	89.52	89.85	90.13	89.56	89.12	
		712.91	711.76	714.20	709.38	713.76	695.70	695.70	695.70	695.70	695.70	

distribution decays fairly quickly at the higher quantiles, which means SRS-OS is less susceptible to greatly overestimating the 0.95-quantile. Also, the rLHS analysis has $>6\%$ of trials falling below the "correct" quantile versus the 5% expected. Once again, this could be related to convergence, or to the selection of bandwidth parameters. It may have been possible to select bandwidth parameters that resulted in exactly 5% of trials falling below the "correct" quantile, but again the challenge is finding bandwidth parameters that are applicable to a variety of systems and sample sizes.

3.3. MELCOR large break LOCA

The next analysis was conducted to compare the methods using an actual nuclear power plant severe accident computer code. The code used for this analysis was MELCOR 2.1, developed by Sandia National Lab [5]. This code was chosen not only because it is used in real nuclear safety analyses, but also because it represents a "large and complex" model. Here, a large model is one requiring significant amounts of human, computational, or other resources in its construction and operation [39]. Complex means the system is made up of a large number of parts that interact in a nontrivial way [40]. Morgan, Henrion, and Small [39] actually use NRC "general purpose regulatory model" computer codes as an example of a large and complex system [39]. The scenario chosen was based on a MELCOR demonstration problem presented in NUREG/CR-6119 [41]. It represents a large break LOCA at the now retired Zion Nuclear Power Plants (ZNPP) near Chicago.

Both units at ZNPP are Westinghouse four-loop pressurized water reactors (PWRs) with large, dry containments. The plant is nodelized into two loops. The first loop represents the single loop in the plant with the pressurizer, and the other loop represents a combination of the other three loops of the plant. More detailed information regarding the nodelization of the plant and core can be found in [41]. The scenario represents a double-ended guillotine rupture of the pressurizer loop at the reactor coolant pump inlet. Three Emergency Core Cooling Systems (ECCS) should activate: high pressure injection (HPI), which is provided by the charging pumps, intermediate pressure injection (IPI), which is provided by the safety injection pumps, and low pressure injection (LPI), which is provided by the residual-heat-removal pumps. However, in this scenario, their flowrates are considered uncertain, and the time of activation of LPI is delayed and uncertain. Table 9 contains a full list of uncertainties.

The break occurs at time 0 s, with LPI activation occurring anywhere from 300 to 1100 s after. The analysis ends shortly after the activation of LPI, since even its minimum flow condition in this experiment is sufficient to temporarily cool the core.

The output of interest is again the PCT of the core. It was compared to the NRC limit of 2200°F [1]. Due to the long run-times of the MELCOR analysis, only SRS-OS, asymptotic SRS, and rLHS were evaluated. First, the "correct" 0.95-quantile of the system was calculated using a 5000-run SRS experiment. This returned a "correct" 0.95-quantile of 1293.16 °F. The empirical CDF in Fig. 10 shows the shape of the distribution. What is interesting to note, from this figure, is the slope of the distribution near the higher quantiles. While the slope is fairly constant until the 0.90-quantile, it quickly flattens, and there is almost a 1000 °F range between the 0.90-quantile and highest points of the distribution. This sensitivity is due to the delay in LPI activation, which can be substantial in some scenarios, and the heat released from zirconium oxidation, which increases exponentially with temperature.

Due to the time burden when running a large, complex code like MELCOR, unlike the previous examples, 10^4 independent trials of each method could not be performed. Instead, for SRS-OS and asymptotic SRS, a large 5000-run trial was conducted. For the analysis here, a random trial of the 5000 outputs were drawn as a trial output. For example, at $n=59$, 59 of the 5000 outputs were chosen at random, and treated as a separate trial. This process was repeated for 10^4 trials. Obviously, this introduces some correlation in the results, since the same output value will be used more than once, and not every output from the support is possible. However,

the likelihood of pulling the exact same 59 outputs is extremely small. Using 5000 outputs, this means there are 8.88×10^{137} different permutations of 59 runs possible. The repeated use of trials has a bigger effect on SRS-OS than asymptotic SRS due to the way the CIs are calculated. Using asymptotic SRS, the sample variance is used to calculate the CI, meaning many of the output results are used in the calculation of the 95/95 value. However, using SRS-OS, simply the highest value of the 59 outputs is taken as the 95/95 value. This means the highest values of the 5000 outputs will be repeatedly chosen as the 95/95 values. This could influence the results, and tests were conducted to quantify this influence, as will be discussed at the end of this section.

For rLHS, 500 cases of LHS with size 10 were conducted. If the desired run level was 60, then 6 of these 500 cases would be chosen at random in order to produce a trial. For rLHS, the number of permutations is smaller at 2.11×10^{13} since the choice is 6 out of 500, instead of 59 out of 5000. Like asymptotic SRS, since the CI for rLHS is determined by calculating a sample variance, the effect of repeating results is less than with SRS-OS.

The results for 10^4 trials of SRS-OS and asymptotic SRS at 59 runs, and rLHS at 60 runs ($t = 10$, $m = 6$) are shown in Table 10. Here, the coverage and "percent below correct" results appear to show that the asymptotic methods are, at the very least, close to convergence, with results of 5.65% and 92.33%. Also, rLHS results in a 95/95 value that is, on average, about 150 °F closer to the actual 0.95-quantile than those 95/95 values found with SRS-OS. This could mean a very significant gain in margin for an operating power plant. The values for the standard deviation should be viewed with caution since, as mentioned above, by using repeated output values, certain results may occur multiple times (especially in the case of SRS-OS) and influence the range of 95/95 values.

In order to confirm that the repeated use of output values did not significantly bias the results, another analysis was performed, but without using the same output result more than once across all trials. This was done by performing only 80 trials, which was determined by finding the largest value k such that $59 \times k < 5000$ (since 5000 SRS runs were initially performed), and $6 \times k < 500$ (since 500 LHS run were performed). This would prevent output results from being used multiple times across all trials. These results are in Table 11. As the table shows, the results are nearly identical to the repeated trial results in Table 10 (since only 80 trials are being performed, the statistical sample is not that large, so the methods may not have exactly 5% of trials "below correct"). It appears that the repeated trial results from above are accurate, and that the gain when using rLHS is real.

This analysis was repeated for the $n = 93$ run level. Table 12 shows the results for 10^4 trials of 93 SRS-OS and asymptotic SRS runs, and 90 rLHS runs ($t = 10$, $m = 9$), using the repeated trial method described above. While SRS-OS at this run level shows a substantial improvement in accuracy compared to the $n = 59$ run level, with over a 100 °F gain in accuracy, it still results in a 95/95 value that, on average, is approximately 100 °F higher than the resulting value when using rLHS.

Once again, this run level was examined without using repeated output values, by conducting only 50 trials. The results for this experiment are shown in Table 13, and like the previous example, they show very little variation from the repeated trial

Table 9
MELCOR analysis uncertainties.

	Uncertainty	Distribution[a]
1	HPI flowrate	Beta(2,5)
2	IPI flowrate	Beta(2,5)
3	LPI flowrate	Beta(2,5)
4	LPI activation time	Uniform(300,1100)
5	Decay heat multiplier	Normal(0.0,2.57)
6	Accumulator temperature	Uniform(3250,3350)
7	Accumulator pressure	Uniform(0.0706,0.0716)
8	Accumulator volume	Uniform(24.07,26.07)
9	Refueling water storage tank volume	Uniform(3150,3250)
10	Reactor power	Uniform(3.25e9,3.35e9)

[a] Many of the uncertainties are not the distribution of the actual parameter, but of a scaling factor or part of a larger formula

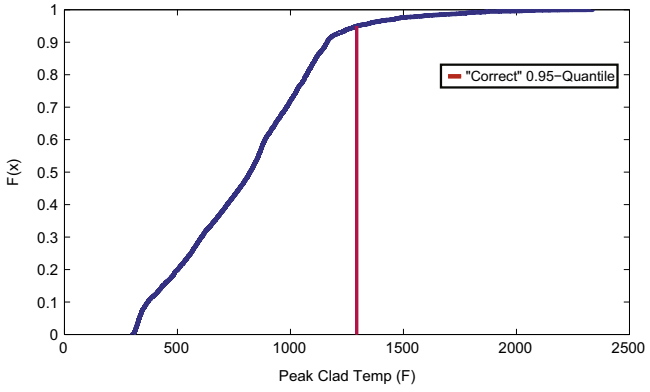

Fig. 10. Empirical CDF of peak clad temperature – 5000 runs – MELCOR analysis.

Table 10
MELCOR LOCA Comparison of 95/95 values for 10^4 trials of rLHS, SRS, and SRS-OS ("correct" 0.95-quantile = 1293.16 °F).

	6 × 10 rLHS	59 SRS	59 SRS-OS
Mean of 10^4 95/95 values	1595.52	1668.53	1740.14
S.D. of 10^4 95/95 values	207.16	254.90	275.38
% below "correct"	5.65%	5.61%	5.10%
Coverage	92.33%	87.30%	

Table 11
MELCOR LOCA Comparison of 95/95 values for *80 trials* of rLHS, SRS, and SRS-OS ("correct" 0.95-quantile = 1293.16 °F).

	6 × 10 rLHS	59 SRS	59 SRS-OS
Mean of 80 95/95 values	1608.58	1671.72	1742.31
S.D. of 80 95/95 values	224.26	235.92	270.82
% below "correct"	6.25%	5.06%	6.32%
Coverage	91.25%	91.14%	

Table 12
MELCOR LOCA Comparison of 95/95 values for 10^4 trials of rLHS, SRS, and SRS-OS ("correct" 0.95-quantile = 1293.16 °F).

	9 × 10 rLHS	93 SRS	93 SRS-OS
Mean of 10^4 95/95 values	1530.49	1557.42	1622.69
S.D. of 10^4 95/95 values	160.70	190.40	209.85
% below "correct"	5.96%	6.81%	4.92%
Coverage	89.66%	89.18%	

Table 13
MELCOR LOCA Comparison of 95/95 values for 50 *trials* of rLHS, SRS, and SRS-OS ("correct" 0.95-quantile = 1293.16 °F).

	9 × 10 rLHS	93 SRS	93 SRS-OS
Mean of 50 95/95 values	1518.55	1561.58	1623.14
S.D. of 50 95/95 values	168.78	183.86	218.62
% below "correct"	6.00%	6.00%	6.00%
Coverage	84.00%	92.00%	

results (again, the important value is the mean, since the statistical sample for the "% below correct" is small).

These results show a potentially large improvement in accuracy by using rLHS instead of SRS-OS. This may be caused, in part, by the shape of the output distribution. As Fig. 20 showed, the higher quantiles of the output distribution spanned over 1000 °F. As explained in Section 2.1, since SRS-OS at low run levels will often return a 95/95 value close to the 0.99- or 0.98-quantile, it returns a significantly more conservative result than rLHS.

3.4. Discussion of results

These experiments indicate that rLHS can provide more accurate and precise confidence intervals for quantiles than SRS-OS. This would mean a reduction in the probability of Type-II errors and the possibility of reducing Type-I errors. However, the rLHS method is not without its faults. As several results showed, at low run levels, the method may not have converged. This can result in significantly more than 5% of the trials returning a 95/95 value falling below the actual quantile. An interesting point though is that even when this did occur, the rLHS trial results still did not fall as far below the "correct" quantile as some SRS-OS trials. So it is not possible to say whether this would result in more Type-I errors than SRS-OS without knowing the actual location of the safety limit, which will be examined in future work.

There may be ways to help resolve the convergence issue. Additional experiments on different types of systems can lead to more guidance about the proper selection of the parameters of the derivative estimator. Also, it is possible to improve coverage of the constructed rLHS confidence interval by replacing the normal critical point with a critical point from a Student-t distribution with m-1 degrees of freedom, where m is the number of LHS cases. Since the Student-t distribution has somewhat heavier tails than a normal distribution, this results in slightly wider and more conservative CIs. It may help to ensure that the number of trials falling below the true quantile does not exceed 5%, but this will also reduce the accuracy. Lastly, as explained in Section 2.2.4, it appears that conducting more cases of a smaller size (increase m, decrease t) aids in the convergence, since the validity of the CLT requires that the number m of cases grows large.

Specifically, more work should be conducted on the derivative estimation and the CFD bandwidth. One interesting note on this point though is that a better derivative estimator at low run levels does not necessarily mean better coverage. As the derivative estimation improves, the width of the CI typically decreases since the CFD often overestimates the true sparsity, especially when n is small. This means the result will be closer to the quantile estimation, which can have large bias at low run levels, so it may dominate the error and disrupt the coverage level. Even though this quantile estimation error is present when the derivative is overestimated, the increased width of the CI tends to negate the errors caused by the quantile estimation bias.

Finally, it should be noted that this work in no way challenges the validity of SRS-OS for the calculation of confidence intervals for quantiles. Conversely, all the experiments carried out here demonstrated that the results of the SRS-OS did have ~95% confidence of exceeding the desired quantile. However, when estimating a 0.95-quantile, SRS-OS is vulnerable to returning a 95/95 value that considerably overestimates the true quantile when the output distribution has a fatter tail at these higher quantiles. This is a result of SRS-OS only using a single output value to derive a 95/95 value. The further the extremes of the output distribution (i.e. the 0.99-quantile) are from the 0.95-quantile, the greater the probability that the SRS-OS 95/95 value will be a greater distance from the 0.95-quantile.

4. Conclusions

Nakayama [3] and Chu and Nakayama [28] have developed a method to establish asymptotically confidence intervals for quantiles when applying VRTs, such as AV and LHS. The present paper compared these methods to the NRC-approved method of SRS-OS through a series of experiments meant to replicate systems found in a nuclear safety analysis. The results appear to show that rLHS, and to a lesser extent AV, can provide more accurate and precise results than SRS-OS. This could mean that analysts using these methods may commit fewer Type-I and Type-II errors in hypothesis testing. In other words, it could reduce the over-conservativeness of SRS-OS, while also lessening the extent of the nonconservative results, or the ~5% of results that will fall below the true quantile.

The application of VRTs in this analysis is not without its own drawbacks. The fact that these are asymptotic methods means that there is a possibility of non-convergence, especially at low run levels, which are, of course, the most desired by analysts for computational reasons. While it is unlikely that any tactic could completely resolve the issue of non-convergence, future work may help provide guidance on optimal utilization of these methods, and adjustments to ensure over-conservatism at low levels may be possible.

Current research is investigating the use of other VRTs, such as control variates and importance sampling, for use in regulatory analyses. These methods can provide much greater variance reduction than even rLHS. As mentioned at the beginning of Section 2.2, there are additional constraints which must be considered when using these methods in relation to if/how previous knowledge is used. However, if acceptable approaches are found, these methods may be able to offer great benefits in reducing regulatory error, increasing margin to safety limits, and generally improving the knowledge and characterization of the output distribution of safety analyses.

Acknowledgment

This work has been supported in part by the National Science Foundation under Grants nos. CMMI-0926949, CMMI-1200065, and CMMI-1537322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] 10 CFR Part 50.46. Acceptance criteria for emergency core cooling systems for light-water nuclear power reactors, Code of Federal Regulations; 2007.
[2] Frepoli C. Overview of Westinghouse realistic large break LOCA evaluation model. Sci Technol Nucl Install 2008;2008:498737.
[3] Nakayama MK. Asymptotically valid confidence intervals for quantiles and values-at-risk when applying Latin hypercube sampling. Int J Adv Syst Meas 2011;4:86–94.
[4] RELAP5-3D Code Development Team. RELAP5-ED Code Manual, Idaho National Laboratory; 2005.
[5] Sandia National Laboratories. MELCOR computer code manual – Vol. 1: primer and users' guide, NUREG/CR-6199; 2011.
[6] 10 CFR Park 50 Appendix K. ECCS Evaluation Models, Code of Federal Regulations; 2000.
[7] U.S. Nuclear Regulatory Commission. Best-estimate calculations of emergency core cooling system performance, RG 1.157; 1989.
[8] Box G, Wilson K. On the experimental attainment of optimum conditions. J R Stat Soc 1951;13:1.
[9] Martin R, O'Dell L. AREVA's realistic large break lOCA analysis methodology. Nucl Eng Des 2005;235:1713–25.
[10] Jaech J. On the use of tolerance intervals in acceptance sampling by attributes. J Qual Technol 1972;4:2.

[11] U.S. Nuclear Regulatory Commission. Applying statistics, NUREG-1475; 1994.
[12] Glaeser H, Pochard R. Review of uncertainty methods for thermal hydraulic computer codes. In: Proceedings of internatonal Conference on new trends in nuclear systems, vol. 1; 1994. p. 447–455.
[13] Guba A, Makai M, Pal L. Statistical aspects of best estimate method – I. Reliab Eng Syst Saf 2003;80(3):217–32.
[14] Nutt WT, Wallis GB. Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. Reliab Eng Syst Saf 2004;83(1):57–77.
[15] Wilks S. Order statistics. Bull Am Math Soc 1948;54(1), part 1:1.
[16] Wald A. An extension of wilks method for setting tolerance limits. Ann Math Stat 1943;14(1):14.
[17] Siddiqui MM. Distribution of quantiles in samples from a bivariate population. J Res Natl Bur Stand B 1960;64:145–50.
[18] Glasserman P. Monte Carlo methods in financial engineering. New York: Springer; 2004.
[19] Stein M. Large sample properties of simulations using Latin hypercube sampling. Technometrics 1987;29(2):143–51.
[20] Bahadur RR. A note on quantiles in large samples. Ann Math Stat 1966;37:577–80.
[21] Ghosh JK, New A. Proof of the Bahadur representation of quantiles and an application. Ann Math Stat 1971;42:1957–61.
[22] Serfling RJ. Approximation theorems of mathematical statistics. New York: John Wiley & Sons; 1980.
[23] Turkey JW. Which part of the sample contains information. Proc Natl Acad Sci USA 1965;53:127–34.
[24] Parzen E. Density quantile estimation approach to statistical data modeling. In: Gasser Th, Rosenblatt M, editors. Smoothing techniques for curve estimation. Berlin: Springer; 1979.
[25] Asmussen S, Glynn P. Stochastic simulation: algorithms and analysis. New York: Springer; 2007.
[26] Bloch DA, Gastwirth JL. On a simple estimate of the reciprocal of the density function. Ann Math Stat 1968;39:1083–5.
[27] Bofinger E. Estimation of a density function using order statistics. Aust J Stat 1997;17:1–7.
[28] Chu F, Nakayama MK. Confidence intervals for quantiles when applying variance-reduction techniques. ACM Trans Model Comput Simul, 36; 2012. Article 7 (25 pages plus 12-page online-only appendix).
[29] Ross S. Simulation. San Diego: Academic Press; 1997.
[30] Avramidis AN, Wilson JR. Correlation-induction techniques for estimating quantiles in simluation. Oper Res 1998;46:574–91.
[31] McKay MD, Beckham RJ, Conover WJ. A comparison of three methods for selecting input variables in the analysis of output from a computer code. Technometrics 1979;21:239–45.
[32] Helton J, Davis F. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab Eng Syst Saf 2003;81:23–69.
[33] Cochran WG. Sampling techniques. New York: Wiley; 1977.
[34] Hall P, Sheather SJ. On the distribution of a studentized quantiles. J R Stat Soc 1988;50:381–91.
[35] Falk M. On the estimation of the quantile density function. Stat Probab Lett 1986;4:69–73.
[36] Nakayama MK. Asymptotic properties of kernel density estimators when applying importance sampling. In: Proceedings of the 2011 winter simulation conference; 2011. p. 556–568.
[37] Rasmuson D, Anderson D, Mardekian J. Use of the 3n parallel flats fractional factorial designs in compute code uncertainty analysis. Department of Energy; 1970.
[38] French A. Response surface modeling of a large break loss of coolant accident (MS Thesis). Ohio State University; 2008.
[39] Morgan M, Henrion M, Small M. Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. New York: Cambridge University Press; 1992.
[40] Simon HA. Sciences of the artificial. 2nd ed. Cambridge: MIT Press; 1982.
[41] Sandia National Laboratories. MELCOR computer code manuals – Vol. 3: demonstration problems, NUREG/CR-6119; 2001.
[42] Pal L, Makai M. Remarks on statistical aspects of safety analysis of complex systems, Technical Report, arXiv:physics/0308086v2 [physics.data-an]; 2003.
[43] Iman RL. Statistical methods for including uncertainties associated with the geologic isolation of radioactive waste which allow for a comparison with licensing criteria. In: Kocher DC, editor. Proceedings of the symposium on uncertainties associated with the regulation of the geologic disposal of high-level radioactive waste. Gatlinburg, TN: Washington, DC: US Nuclear Regulatory Commission, Directorate of Technical Information and Document Control; 1981. p. 145–57.