Sampling Size Issue in PRA Uncertainty Analysis

Chunrui Deng
Nuclear Power Institute of China, Chengdu, China

Abstract: In risk-informed decision making process, mean values of the risk metrics are used to be compared to the acceptance criteria; when the risk metrics are highly uncertain, the quantiles are often used to characterize the dispersion of the risk metrics. A formal uncertainty propagation process is often necessary to obtain reliable estimate of these quantities. This paper discusses the sample size issue during sampling. The mathematical basis and illustrative examples are given in the paper. It can be concluded that if a small deviation is desired, several tens of thousands of samples may be necessary, and more samples are needed for quantile estimation than for mean estimation.

Keywords: PRA, Uncertainty, Sample size

1. INTRODUCTION

In risk-informed applications, the existence of uncertainty is natural and unavoidable. During decision making, the impact of uncertainty must be taken into account. RG 1.174[1] states that "the appropriate numerical measures to use in the initial comparison of the PRA results to the acceptance guidelines are mean values", and specially highlights the importance of state-of-knowledge correlation (SOKC). NUREG-1855[2], EPRI report 1016737[3] and 1026511[4] give practical guidance to deal with uncertainty. In these references, uncertainties are categorized into parameter uncertainties, model uncertainties and completeness uncertainties. When dealing with parameter uncertainties, a formal uncertainty propagation process may be needed. The most popular method to propagate uncertainty is the simple Monte Carlo sampling due to its simplicity and good sample properties (e.g. independent and identically distributed). It is often confusing how many samples are necessary. This paper discusses the sampling size issue. The mathematical basis is given and illustrative examples are provided.

The specific problems discussed are: 1) Relationship between point estimate and the sample mean; 2) Sample size for mean value estimate; 3) Sample size for quantile estimate. Section 2 describes the math behind these problems. Section 3 gives some examples. Section 4 is a simple conclusion. The parameterization of distributions mentioned is given in the appendix.

2. MATHEMATICAL BASIS FOR SAMPLING

2.1. Assumptions

The focus in this paper is level 1 PRA, assuming that the core damage frequency (CDF) can be obtained through applying rare event approximation, i.e., $CDF \approx \sum_{i=1}^{N} CDF_i$, where CDF is the total CDF, and CDF_i is the CDF for minimal cut set i, N is the number of minimal cut sets.

The common cause model used in this paper is the alpha factor model, assuming that the alpha factors are independent except $\alpha_1(\alpha_1 = 1 - \sum_{i>1} \alpha_i)$, where α_is are alpha factors.

2.2. Relationship between Point Estimate and the Sample Mean

It is not uncommon to see that when doing sampling in PRA, if the inputs are means of uncertain parameters, the point estimate of CDF is almost the same with sample mean without SOKC accounted (e.g. sampling use the event sampling option rather than the parameter sampling in RiskSpectrum). EPRI report 1016737[3] says "If all the events in the cutest were statistically independent, (i.e., based on independent data that is not pooled or correlated in any way), the output point estimates would

themselves be mean values". In mathematical language, we can present the fact as follows. A minimal cut set can be expressed as $\prod_i X_i$, where X_i s are basic events constituting the cut set. If X_i s are independent between each other, then according the property of mathematical expectation, $\mathrm{E}\left(\prod_i X_i\right) = \prod_i EX_i$. That is the mean value of the cut set $\mathrm{E}\left(\prod_i X_i\right)$ equals to the point estimate with mean value inputs. The total CDF is a linear combination of CDFs of all the minimal cut sets, so the same phenomenon presents.

In fact, independent variables are parameters rather than basic events. Point estimate of a basic event with mean values of parameters inputs is not strictly equals to the mean value of the basic event]. That is, if a basic event X is a function of parameter λ, then $\mathrm{E}[\mathrm{f}(\lambda)] \neq \mathrm{f}(\mathrm{E}(\lambda))$ for most cases. Jensen inequality says that, for a convex function g in R (R denotes the real domain), that is for arbitrary $0 < p < 1$,

$$\mathrm{g}(\mathrm{px} + (1-\mathrm{p})\mathrm{y}) \leq \mathrm{pg}(\mathrm{x}) + (1-\mathrm{p})\mathrm{g}(\mathrm{y}) \tag{1}$$

the following is valid

$$\mathrm{g}(\mathrm{EX}) \leq \mathrm{Eg}(\mathrm{X}) \tag{2}$$

The equality relationship holds if and only if X is constant or g is a linear function.

For popular case in PRA, the probability of a basic event can be approximated to a linear function of input parameters. This illustrate why the sample mean approximately equals to point estimate.

For cases in which SOKC has to be considered, the deviation between point estimate and the mean value can be large. Consider a cut set contains basic events with SOKC, then $\mathrm{E}\left(\prod_i X_i\right) = \prod_i EX_i$ never holds. Take a cut set with 2 totally correlated events as an example, the mean value of the cut set is $\mathrm{E}(\mathrm{X}^2)$. According to probability theory,

$$\mathrm{E}\left[\mathrm{X}^2\right] = (\mathrm{E}[\mathrm{X}])^2 + \mathrm{Var}(\mathrm{X}) \tag{3}$$

A variance item appears in equation (3). The standard deviation is usually several times larger than the mean, so the variance item cannot be neglected. So the impact of SOKC may be important. When SOKC is dominant, we cannot use point estimates to substitute mean values. Examples in subsection 3.1 show the impact of SOKC.

2.3. Sample Size for Mean Value Estimate

For simple random sampling, the problem of mean value estimate of the CDF is one of the most popular problem in statistical inference. Here, we use X denote random variable, μ and σ^2 denote the mean and variance of X. $\overline{\mathrm{X}}$ and S^2 denote the unbiased estimators for the population mean and variance, and the lower case notations denote specific values of a sample.

Central limiting theorem says that, for independent and identically distributed random variables $\mathrm{X}_1, \mathrm{X}_2, \cdots$ with limited variance, $\sqrt{\mathrm{n}}(\overline{\mathrm{X}}_\mathrm{n} - \mu)/\sigma$ asymptotically follows standard normal distribution, e.g. the limitation of cumulated distribution function of which is

$$\lim_{\mathrm{n}\to\infty} \mathrm{G}_\mathrm{n}(\mathrm{x}) = \int_{-\infty}^{\mathrm{x}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\mathrm{y}^2/2} \mathrm{dy} \tag{4}$$

For a moderate sample size (e.g. several tens), the normal approximation can be accepted. We can use central limiting theorem to construct interval estimation for the mean value.

With population variance known, an interval estimate of mean with $1-\alpha$ confidence level is

$$\left\{\mu{:}\overline{\mathrm{x}} - \mathrm{z}_{\alpha/2}\frac{\sigma}{\sqrt{\mathrm{n}}} \leq \mu \leq \overline{\mathrm{x}} + \mathrm{z}_{\alpha/2}\frac{\sigma}{\sqrt{\mathrm{n}}}\right\} \tag{5}$$

Where, $\mathrm{z}_\mathrm{a} = \Phi^{-1}(1-\mathrm{a})$, Φ is the cumulated distribution function of the standard normal distribution.

With population variance unknown, according to Slutsky theorem (see [5] for details), the following holds

$$\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n} \to n(0,1) \tag{6}$$

Where → denote approaching.

So for large samples, we can use sample variance to substitute population variance in equation (5), the $1-\alpha$ confidence interval for mean value is then

$$\left\{\mu:\bar{x}-z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+z_{\alpha/2}\frac{s}{\sqrt{n}}\right\} \tag{7}$$

The length of the interval is

$$2z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8}$$

If we use a relative deviation to the true value of mean to measure the accuracy of the estimate, and further we require the acceptable interval length is small portion q of the mean, that is

$$2z_{\alpha/2}\frac{s}{\sqrt{n}} \le q\mu \tag{9}$$

With μ unknown, use $\bar{x}$ instead. We can obtain

$$n \ge \frac{4z_{\alpha/2}^2 s^2}{q^2\bar{x}^2} \tag{10}$$

We can see that the required sample size is proportion to the square of the ratio of standard deviation and mean.

Subsection 3.2 gives practical examples, and illustrate the variation of the required size when multiple cut sets are added.

2.4. Sample Size for Quantile Estimate

Quantiles are important parameters to characterize the dispersion of a random variable. In PRA, the sample quantiles are naturally used as estimators of the population quantiles. Interval estimate for the quantiles is complex, and more advanced statistical knowledge such as order statistics and approximation theory has to be used. Reference [6] gives a $1-\alpha$ confidence interval as follows.

$$\frac{k_{1n}}{n} = p - z_{\alpha/2}\sqrt{p(1-p)/n} \tag{11}$$

$$\frac{k_{2n}}{n} = p + z_{\alpha/2}\sqrt{p(1-p)/n} \tag{12}$$

Where p is the cumulative probability at the quantile. The interval between the k_{1n}th and the k_{2n}th sample when ascending ordered is the required confidence interval.

The length of the interval is

$$C(X) = \frac{2z_{\alpha/2}\sqrt{p(1-p)}}{f(\xi_p)\sqrt{n}} \tag{13}$$

Where $f(x)$ is the density function of X, ξ_p is the pth quantile.

If a similar relative deviation is used for assessing the quantile estimate as for the mean estimate, then for a deviation portion q, there exists,

$$\frac{2z_{\alpha/2}\sqrt{p(1-p)}}{f(\xi_p)\sqrt{n}} \le q\xi_p \tag{14}$$

We can obtain

$$n \geq \frac{4z_{\alpha/2}^2}{q^2}\frac{p(1-p)}{\xi_p^2 f^2(\xi_p)} \tag{15}$$

The required sample size depends on the quantile and the density at the quantile, we can approximate to use sample values for substitution.

Compared with the case for mean estimate, we can find that the sample size required for quantile estimate depends on local property of the distribution. The sample size may be larger than that for mean estimate intuitively.

Subsection 3.3 gives examples about quantile estimate.

3. EXAMPLES

3.1. Relationship between Point Estimate and the Sample Mean

Firstly, the deviation between point estimate and the mean of a basic event is examined. Take the basic event a motor-driven pump fail to run as an example. In the mission time $T_{mission}$, the failure probability is $1 - e^{-\lambda T_{mission}}$, where λ is the failure rate which follows a gamma distribution with parameters α and β. Now we see if the following equation holds.

$$E(1 - e^{-\lambda T_{mission}}) \approx 1 - e^{-E(\lambda)T_{mission}} \tag{16}$$

It is easy to obtain the exact value of the left side of equation (16) is

$$E\left(1 - e^{-\lambda T_{mission}}\right) = 1 - \left(\frac{\beta}{\beta + T_{mission}}\right)^{\alpha} \tag{17}$$

Using the data from NUREG/CR-6928 [7] and $T_{mission} = 24h$, the mean (calculated with (17)) and the point estimate (the right hand side of equation (16)) are calculated and shown in the column 4 and 5 row 1 of table 1. It can be found, the deviation between the two is quite small.

In fact, if $\lambda T_{mission}$ is very small (e.g. <0.01), the following approximation exists.

$$1 - e^{-\lambda T_{mission}} \approx \lambda T_{mission} \tag{18}$$

The probability is approximately a linear function of the parameter λ.

The probability of another popular failure mode in PRA fail on demand is also linear function of the parameter (either expressed in the form of failure probability p or in the form of $\frac{1}{2}\lambda_{standby}T_{test}$ where $\lambda_{standby}$ is the standby failure rate and T_{test} is the test interval). According to Jensen inequality (equation (2)), the point estimate and the mean are very close.

Secondly, for cases with SOKC, still using the motor-driven pump as example. Here, double redundancy is assumed. What to be examined is if the following equation is valid.

$$E\left(1 - e^{-\lambda T_{mission}}\right)^2 \approx \left(1 - e^{-E(\lambda)T_{mission}}\right)^2 \tag{19}$$

Similarly, the mean can be calculated exactly or approximated using equation (18). Column 7 and 8 row 2 of table 1 give the value of left and right side of equation (19). It can be found the deviation between the two is large. It is easy to calculate that the difference between the two is $\mathrm{Var}\left(1 - e^{-\lambda T_{mission}}\right)$ which is shown in column 9 row 1 of table 1.

The results for another component with higher failure rate EDG (emergency diesel generator) are also shown in table 1. The similar conclusion is obtained. So when SOKC is not important, the point estimate can substitute the mean; and when SOKC is important, we cannot do that, a formal uncertainty propagation is necessary.

Table 1: Mean and point estimate with and without SOKC

	Failure rate λ with a gamma distribution with parameters α and β			Mean failure probability of 1 component	Point estimate of the failure probability of 1 component	Mean failure probability of 2 components	Point estimate of the failure probability of 2 components	Difference between the mean and point estimate of the failure probability of 2 components
	λ	α	β	$E(1-e^{-\lambda T})$	$1-e^{-E(\lambda)T}$	$E(1-e^{-\lambda T})^2$	$(1-e^{-E(\lambda)T})^2$	$Var(1-e^{-\lambda T})$
MDP fail to run	5.80E-6/h	0.5	8.62E+04	1.39E-04	1.39E-04	5.81E-08	1.94E-08	3.88E-08
EDG fail to run	8E-4/h	2	2.50E+03	1.89E-02	1.90E-02	5.32E-04	3.62E-04	1.84E-04

3.2. Sample Size for Mean Value Estimate

Samples from a lognormal distribution with mean 1E-5 and EF (error factor) 10 is used to illustrate the mean estimate issue. According to equation (10), at 90% confidence level, if the required relative deviation is 2%, then the necessary sample size is about 170,000. Figure 1 which is produced with the statistical package R gives the accuracy varied with sample size.

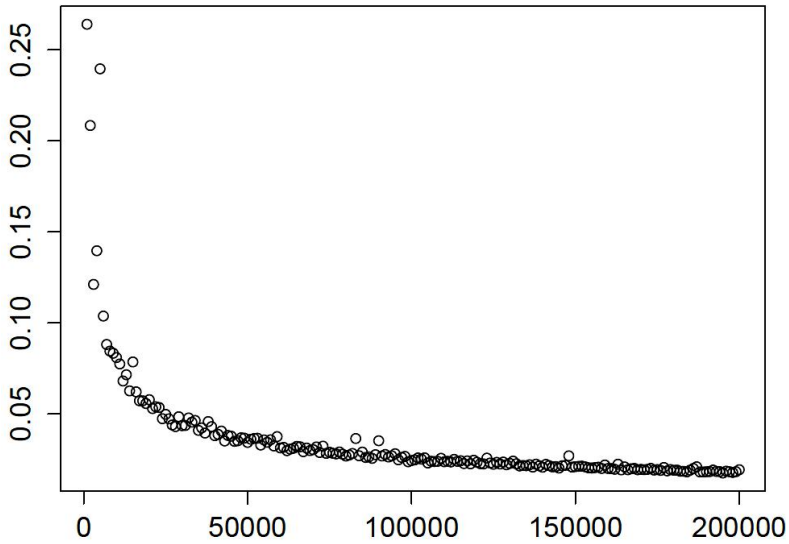

Figure 1: Accuracy vs sample size

Reference [3] recommends a sample size 25000. For our relative disperse distribution, sample size 25000 can ensure a deviation within 5% (see figure 1) at a confidence level 90%.

For comparison, the process above is repeated with lognormal distribution with mean 1E-5 and EF 3 and lognormal distribution with mean 1E-6 and EF 3. For the same confidence level and accuracy, the necessary sample sizes are both 16,500. It can be seen that if the EF keeps constant, the required sample size keeps constant too.

For actual reactor case, the question is that if the advanced reactor with smaller CDF requires more samples than the former design with larger CDF. Table 2 examines the case in which failure probability is reduced through increasing redundancy. Again, the MDP is used, and for CCF, NUREG/CR-5497 data are used.

Table 2: σ/μ for different cases

	Mean μ	Variance σ^2	σ/μ
Single	$E(1-e^{-\lambda t})$	$Var(1-e^{-\lambda t})$	1.41
	1.39E-4	3.88E-8	
Double redundancy	$E^2(1-e^{-\lambda t})$	Note 1	2.0

	1.94E-8	1.50E-15	
Double redundancy (with SOKC)	$E(1-e^{-\lambda t})^2$	$Var(1-e^{-\lambda t})^2$	3.27
	5.81E-8	3.61E-14	
Double redundancy (CCF)	Note 2	Note 2	1.60
	9.28E-6	2.21E-10	
Triple redundancy	$E^3(1-e^{-\lambda t})$	Note 1	5.10
	2.70E-12	1.89E-22	
Triple redundancy (with SOKC)	$E(1-e^{-\lambda t})^3$	$Var(1-e^{-\lambda t})^3$	6.72
	4.05E-11	7.41E-20	
Triple redundancy (CCF)	Note 2	Note 2	1.52
	1.65E-5	6.26E-10	

Note 1: The equation $Var(X_1X_2) = Var(X_1)Var(X_2) + E^2(X_1)Var(X_2) + E^2(X_2)Var(X_1)$ for independent random variable X_1 and X_2 is used.

Note 2: For CCF group size 2, $\frac{2\alpha_2}{\alpha_1+2\alpha_2}(1-e^{-\lambda t})$; For CCF group size 3, $\frac{3\alpha_2+3\alpha_3}{\alpha_1+2\alpha_2+3\alpha_3}(1-e^{-\lambda t})$

It can be seen from table 2: for the probability of cut sets with CCF, the required sample size is no more than the independent cases; SOKC increases the sample size moderately (e.g. for double redundancy case, 2.67 times larger than the independent case); redundancy decrease the mean value, but increases the sample size greatly (e.g. for triple redundancy case, 6.5 times larger than the double redundancy case).

For each single cut set, we can obtain the required sample size. The next concern is for the summation of all the cut sets how the sample size varies.

For summation of independent random variables, $X_1, X_2, \cdots$, the following equations are valid.

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2 \tag{20}$$

$$\mu = \sum_{i=1}^{N} \mu_i \tag{21}$$

For a simple 2 cut set case, assuming that $\frac{\sigma_1}{\mu_1} > \frac{\sigma_2}{\mu_2}$, then for the summation,

$$\frac{\sigma^2}{\mu^2} = \frac{\sigma_1^2+\sigma_2^2}{(\mu_1+\mu_2)^2} < \frac{\sigma_1^2+\sigma_2^2}{\mu_1^2+\mu_2^2} = \frac{\sigma_1^2}{\mu_1^2}\frac{\left(1+\frac{\sigma_2^2}{\sigma_1^2}\right)}{\left(1+\frac{\mu_2^2}{\mu_1^2}\right)} < \frac{\sigma_1^2}{\mu_1^2} \tag{22}$$

Equation (22) shows that for the summation of cut sets, σ^2/μ^2 is smaller than the maximum σ_i^2/μ_i^2. That is, for the summation, the required sample size is no more than the required sample size for each single cut set.

Additional, for the 2 cut set case, assuming that $\mu_1 \gg \mu_2$, then

$$\frac{\sigma_1^2+\sigma_2^2}{(\mu_1+\mu_2)^2} = \frac{\frac{\sigma_1^2}{\mu_1^2}+\frac{\sigma_2^2}{\mu_1^2}}{(\frac{\mu_2}{\mu_1}+1)^2} \approx \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_1^2} = \frac{\sigma_2^2}{\mu_2^2}\frac{\mu_2^2}{\mu_1^2} + \frac{\sigma_1^2}{\mu_1^2} \approx \frac{\sigma_1^2}{\mu_1^2} \tag{23}$$

That is the sample size is mainly determined by the dominant cut sets (e.g. Cut sets with larger means) which is consistent with intuition.

For multiple cut sets summation, $\frac{\sigma_1^2+\sigma_2^2}{(\mu_1+\mu_2)^2}$ may much smaller than σ_i^2/μ_i^2 (check the case that all the σ_i^2 and μ_i are equal). The CDF consists of many dominant cut sets, so the sample size for mean CDF estimate is smaller than single dominant cut set, which reflect the fact that when the accuracy of the CDF is ensured, the accuracy of cut sets mean estimate may not attain the same accuracy with the same sample size on the other side.

3.3. Sample Size for Quantile Estimate

For quantile estimate, the required sample size can be easily obtained using equation (15). Using the same lognormal distribution cases as in subsection 3.2 as example, the required sample sizes are listed in table 3, and for comparison, the required sample sizes for mean estimate are also listed in table 3.

Table 3: Required sample size for mean and quantile estimate

Lognormal distribution parameters	Sample size for mean estimate	Sample size for 95% quantile estimate
$\text{mean} = 1\text{E}-5, \text{EF} = 10$	170,000	236,900
$\text{mean} = 1\text{E}-5, \text{EF} = 3$	16,500	54,232
$\text{mean} = 1\text{E}-6, \text{EF} = 3$	16,500	54,232

It can be seen from table 3 if the same accuracy is desired, larger sample size is required for quantile estimate than for mean estimate.

Note that, the required sample size for quantile estimate depends on probability density function for which an accurate estimate is difficult.

4. CONCLUSIONS

The paper studies the sample size issue when doing sampling in uncertainty propagation process. The intent is to show the mathematical basis behind sampling, which is often confusing. We can draw several conclusions through the study: 1) SOKC can significantly impact the mean, a formal uncertainty propagation is recommended when available; 2) The required sample size for mean value estimate may be several tens of thousands, more samples are needed for quantile estimation than for mean estimation; 3) The dominant cutsets also dominate the sample size.

References

[1] U.S. Nuclear Regulatory Commission. An Approach for Using Probabilistic Risk Assessment in Risk-Informed Decisions on Plant-Specific Changes to the Licensing Basic. Regulatory Guide 1.174. Revision 2. May 2011.
[2] U.S. Nuclear Regulatory Commission. Guidance on the Treatment on Uncertainties Associated with PRAs in Risk-Informed Decisionmaking. NUREG-1855. Revision 1. March 2017.
[3] Treatment of Parameter and Model Uncertainty for Probabilistic Risk Assessment. EPRI, Palo Alto, CA:2008. 1016737
[4] Practical Guidance on the Use of Probabilistic Risk Assessment in Risk-Informed Applications with a Focus on the Treatment of Uncertainty. EPRI, Palo Alto, CA:2012. 1026511
[5] George Casella, Roger L. Berger, Statistical Inference, 2nd Edition, 2002.
[6] Robert J. Serfling. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc. New York, 1980.
[7] S.A. Eide, etc. Industry-Average Performance for Components and Initiating Events at U.S. Commercial Nuclear Power Plants. NUREG/CR-6928, February 2007.
[8] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[9] F.M. Marshall, etc. Common-Cause Failure Parameter Estimations. NUREG/CR-5497. October, 1998.

Appendix

The parameterizations of distributions used in the paper are list as follows.

Normal distribution: $\text{f}(\text{x}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\text{x}-\mu}{\sigma}\right)^2\right]$

Lognormal distribution: $f(y) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2\right]$

Gamma distribution: $f(t) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp\,(-t\beta)$

Beta distribution: $f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$