Optimal Sample Size Determination based on Bayesian Reliability and Value of Information

Wei Xiang
Graduate Student, Dept. of Civil and Environmental Engineering, The University of Western Ontario, London, Ontario, Canada

Wenxing Zhou
Associate Professor, Dept. of Civil and Environmental Engineering, The University of Western Ontario, London, Ontario, Canada

ABSTRACT: In the structural reliability analysis, the probabilistic distributions of basic random variables may contain uncertainties arising from the imperfect knowledge from which the distributions are elicited. It subsequently introduces uncertainty into the calculated failure probability P_f, which may affect the decision-making. To reduce the uncertainty of the failure probability estimation, it is desirable to collect samples of the basic random variables and use these samples to update the corresponding probability distributions. In this work, the relationship between the sample size of the basic random variable and variance of the estimated failure probability is derived by using the Bayesian pre-posterior analysis, based on which the optimal sample size criterion is established. To make the pre-posterior analysis and criterion applicable to a wide range of distributions, continuous random variables are discretized at first. The probability mass functions of the discretized random variables are then assigned Dirichlet prior distributions. The total probability theorem is employed to express P_f in terms of PMFs of the discretized variables and conditional failure probabilities corresponding to given values of discretized variables. Then the prior, posterior and pre-posterior analysis of P_f are carried out. The optimal sample size criterion to maximize the expected net gain of sampling is developed based on the result of the pre-posterior analysis of P_f and quadratic loss function. An example of determining the optimal number of burst tests for collecting the samples of model error of the burst capacity model for corroded pipelines is used to illustrate the proposed criterion. Moreover, the sensitivity analysis indicates that the optimal sample size is insensitive to the discretization of the basic random variables, but sensitive to the equivalent sample size of the prior Dirichlet distribution.

1. INTRODUCTION

The structural reliability analysis of engineering structures generally involves estimating the failure probability, P_f, as follows:

$$P_f = \int_{\Omega_f} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \quad (1)$$

where $f_{\mathbf{X}}(\mathbf{x})$ denotes the joint probability density function (PDF) of a vector of basic random variables $\mathbf{X}$ such as dimensions of the structural members, material properties and magnitudes of loads and model errors, and Ω_f denotes the failure domain that is typically defined through one or more so-called limit state functions. Since $f_{\mathbf{X}}(\mathbf{x})$ is often elicited from imperfect information such as expert opinions and databases with limited sample sizes, there are uncertainties associated with $f_{\mathbf{X}}(\mathbf{x})$. The epistemic uncertainties can be taken into account in the analysis by considering the distribution parameters of basic random variables to be uncertain (Der Kiureghian 1989, Hong 1996). This introduces uncertainty in P_f, which may affect the decision making based on P_f. It is therefore desirable to gather sufficient samples of $\mathbf{X}$ to reduce the uncertainties in $f_{\mathbf{X}}(\mathbf{x})$. The determination of appropriate sample sizes for $\mathbf{X}$ is

a challenging yet often-encountered task in the design and assessment of engineering structures; for instance, gathering soil property data in the design of foundations (Goldswarthy 2007), proof-load testing quasi-identical multi-components structural systems (Nishijima and Faber 2007; Shafieezadeh and Ellingwood 2012), collecting corrosion defect data for the integrity management of buried oil and gas pipelines (Caleyo et al. 2014) and measuring the wall thickness of deteriorating piping systems in nuclear reactors (Higo and Pandey 2016). Since the cost of sampling is in general high, the sample size should be determined by balancing between the cost and associated benefit. This is known as the problem of the sample size determination (SSD).

The Bayesian pre-posterior analysis (Raiffa and Schlaifer 1961) is a viable approach to deal with SSD. Pham and Turkkan (1992) employed the pre-posterior analysis to study SSD for the parameter of the binomial distribution. Assuming the parameter to have a beta prior distribution and exploiting the conjugacy of the beta-binomial pair, the authors derived analytical expressions for the expectations of the posterior mean and variance of the binomial parameter with respect to the outcome of sampling with a given sample size. The appropriate sample size can then be determined by using one of three criteria: limiting the posterior variance and Bayes risk to pre-determined allowable values, respectively, and maximizing the expected net gain of sampling (ENGS). Adjock (1992) extended Pham and Turkkan's approach to investigate SSD for parameters of the multinomial distribution by assuming the prior distribution of the parameters to be the Dirichlet distribution and utilizing the conjugacy of the Dirichlet-multinomial pair. Based on the pre-posterior analysis and value of information (VoI) concept, Higo and Pandey (2016) derived an analytical expression for the optimal number of wall thickness measurements for nuclear piping systems by assuming the wall thickness to follow a normal distribution. The aforementioned studies address SSD for parameters of specific distributions; however, there is a lack of a general framework that can deal with SSD for a wide range of probability distributions by considering the impact of uncertainties in $f\mathbf{x}(\mathbf{x})$ on P_f.

In this study, a methodology that is based on the Bayesian pre-posterior analysis of P_f is developed to deal with SSD. The methodology starts by discretizing the basic variables for which sample sizes need to be determined. The probability mass functions (PMFs) of the discretized variables are then assigned Dirichlet prior distributions. The total probability theorem is employed to express P_f in terms of PMFs of the discretized variables and conditional failure probabilities corresponding to given values of discretized variables. This facilitates the pre-posterior analysis of P_f based on those of the discretized variables. The criterion of determining optimal sample size based on the result of the pre-posterior analysis of P_f is then proposed by maximizing ENGS. The proposed SSD criterion is illustrated through an example involving determining the optimal number of burst tests for collecting the samples of model error of the burst capacity model for corroded pipelines.

2. FORMULATION

2.1. Pre-posterior analysis of the PMF

Let Y denote a discrete random variable with m states y_i (i = 1, 2, …, m). The PMF of Y is represented by an m-dimensional vector $\mathbf{W}_Y = \{W_{Y,1}, W_{Y,2}, \ldots, W_{Y,m}\}$ ($\sum_{i=1}^{m} W_{Y,i} = 1$). Consider that $\mathbf{W}_Y$ is uncertain and hence a random vector. The Dirichlet distribution is often assigned as the prior distribution of $\mathbf{W}_Y$ in the literature (Spiegelhalter et al 1993); that is, $\mathbf{W}_Y \sim \mathrm{Dir}(\boldsymbol{\alpha}_Y)$, where "~" denotes the assignment of a probability distribution, and $\boldsymbol{\alpha}_Y = \{\alpha_{Y,1}, \alpha_{Y,2}, \ldots, \alpha_{Y,m}\}$ is the m-dimensional parameter vector of the Dirichlet distribution. The prior joint PDF of $\mathbf{W}_Y$, $f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y)$, is given by,

$$f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y) = \frac{\Gamma(\alpha_{Y0})}{\prod_{i=1}^{m}\Gamma(\alpha_{Y,i})}\prod_{i=1}^{m}(w_{Y,i})^{\alpha_{Y,i}-1} \quad (0 < w_{Y,i} < 1 \text{ and } \alpha_{Y,i} > 0;\ i = 1, 2, \ldots, \mathrm{m}) \tag{2}$$

where $\mathbf{w}_Y = \{w_{Y,1}, w_{Y,2}, \ldots, w_{Y,m}\}$ is the value of $\mathbf{W}_Y$; $\Gamma(\bullet)$ is the gamma function, and $\alpha_{Y0} = \sum_{i=1}^{m} \alpha_{Y,i}$ is known as the equivalent sample size of the Dirichlet distribution (Johnson and Kotz 1972).

The prior mean and variance of $W_{Y,i}$ ($i = 1, 2, \ldots, m$), $\mu^{\pi}_{W_{Y,i}}$ and $\xi^{\pi}_{W_{Y,i}}$, respectively, are given by,

$$\mu^{\pi}_{W_{Y,i}} = \frac{\alpha_{Y,i}}{\alpha_{Y0}} \tag{3}$$

$$\xi^{\pi}_{W_{Y,i}} = \frac{\alpha_{Y,i}(\alpha_{Y0}-\alpha_{Y,i})}{(\alpha_{Y0})^2(\alpha_{Y0}+1)} \tag{4}$$

Throughout the paper, the symbols $\mu_\bullet$ and $\xi_\bullet$ are used to denote the mean and variance of a random variable $\bullet$, respectively, whereas superscripts π and p are used to denote prior and posterior statistics, respectively. $W_{Y,i}$ and $W_{Y,j}$ ($i, j = 1, 2, \ldots, m$; $i \neq j$) are correlated with the corresponding covariance, $\omega^{\pi}_{W_{Y,ij}}$, given by

$$\omega^{\pi}_{W_{Y,ij}} = \frac{-\alpha_{Y,i}\alpha_{Y,j}}{(\alpha_{Y0})^2(\alpha_{Y0}+1)} \ (i \neq j) \tag{5}$$

Now suppose that a set of samples $\mathbf{n}_Y = \{n_{Y,1}, n_{Y,2}, \ldots, n_{Y,m}\}$ are obtained from the outcome space of Y, where $n_{Y,i}$ ($n_{Y,i} \geq 0$; $i = 1, 2, \ldots, m$) represents the number of samples lying in the i-th state. These samples can be used to update the prior distribution of $\mathbf{W}_Y$. The likelihood of $\mathbf{n}_Y$, $L(\mathbf{w}_Y|\mathbf{n}_Y)$, is of the multinomial form as follows:

$$L(\mathbf{w}_Y|\mathbf{n}_Y) = \frac{n_{Y0}!}{\prod_{i=1}^{m} n_{Y,i}!}\prod_{i=1}^{m}(w_{Y,i})^{n_{Y,i}} \tag{6}$$

where $n_{Y0} = \sum_{i=1}^{m} n_{Y,i}$, i.e. the total number of samples. Given the conjugacy between the multinomial and Dirichlet distributions, the posterior distribution of $\mathbf{W}_Y$ is also the Dirichlet distribution with the corresponding PDF, $f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y, \mathbf{n}_Y)$, given by,

$$f(\mathbf{w}_Y|\boldsymbol{\alpha}_Y, \mathbf{n}_Y) = \frac{\Gamma(\alpha_{Y0}+n_{Y0})}{\prod_{i=1}^{m}\Gamma(\alpha_{Y,i}+n_{Y,i})}\prod_{i=1}^{m}(w_{Y,i})^{\alpha_{Y,i}+n_{Y,i}-1} \tag{7}$$

It follows that the parameter vector of the posterior Dirichlet distribution of $\mathbf{W}_Y$ is $(\boldsymbol{\alpha}_Y + \mathbf{n}_Y)$. The posterior mean, variance and covariance of $\mathbf{W}_Y$ are then given by,

$$\mu^{p}_{W_{Y,i}} = \frac{\alpha_{Y,i}+n_{Y,i}}{\alpha_{Y0}+n_{Y0}} \tag{8}$$

$$\xi^{p}_{W_{Y,i}} = \frac{(\alpha_{Y,i}+n_{Y,i})(\alpha_{Y0}+n_{Y0}-\alpha_{Y,i}-n_{Y,i})}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)} \tag{9}$$

$$\omega^{p}_{W_{Y,ij}} = \frac{-(\alpha_{Y,i}+n_{Y,i})(\alpha_{Y,j}+n_{Y,j})}{(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+n_{Y0}+1)} \ (i \neq j) \tag{10}$$

If a decision is made to draw a total of n_{Y0} samples but the actual sampling process has not been carried out, the potential sample count in the i-th state ($i = 1, 2, \ldots, m$) is now uncertain, denoted by a random variable $N_{Y,i}$ (the total sample count n_{Y0} is a constant). The posterior statistics of $\mathbf{W}_Y$ then depend on the realization of the random vector $\mathbf{N}_Y = \{N_{Y,1}, N_{Y,2}, \ldots, N_{Y,m}\}$. This is the pre-posterior analysis (Raiffa and Schlaifer 1961). The marginal (or compound) distribution of $\mathbf{N}_Y$ is the so-called Dirichlet-multinomial distribution (Johnson and Kotz 1972). Replacing $n_{Y,i}$ and $n_{Y,j}$ in Eqs. (8) through (10) by random variables $N_{Y,i}$ and $N_{Y,j}$, one can then evaluate the expectations of the posterior mean, variance and covariance of $\mathbf{W}_Y$ with respect to the distribution of $\mathbf{N}_Y$, respectively, as follows:

$$\mathrm{E}_N\left[\mu^{p}_{W_{Y,i}}\right] = \frac{\alpha_{Y,i}}{\alpha_{Y0}} \tag{11}$$

$$\mathrm{E}_N\left[\xi^{p}_{W_{Y,i}}\right] = \frac{\alpha_{Y0}}{\alpha_{Y0}+n_{Y0}}\xi^{\pi}_{W_{Y,i}} \tag{12}$$

$$\mathrm{E}_N\left[\omega^{p}_{W_{Y,ij}}\right] = \frac{n_{Y0}\alpha_{Y,i}\alpha_{Y,j}(\alpha_{Y0}+n_{Y0})-\alpha_{Y,i}\alpha_{Y,j}(\alpha_{Y0})^2(\alpha_{Y0}+1)-(n_{Y0})^2\alpha_{Y,i}\alpha_{Y,j}(\alpha_{Y0}+1)-2n_{Y0}\alpha_{Y,i}\alpha_{Y,j}\alpha_{Y0}(\alpha_{Y0}+1)}{(\alpha_{Y0})^2(\alpha_{Y0}+n_{Y0})^2(\alpha_{Y0}+1)(\alpha_{Y0}+n_{Y0}+1)} \tag{13}$$

where $\mathrm{E}_N[\bullet]$ denotes the expectation with respect to $\mathbf{N}_Y$. Note that the expectation of the posterior mean (Eq. (11)) coincides with the prior mean (Eq. (3)).

2.2. Pre-posterior analysis of the P_f

Consider that Y is one element in the vector of basic random variables $\mathbf{X}$ in Eq. (1) and that samples of Y are needed to reduce the epistemic uncertainty and the corresponding sample size is to be determined. To apply the methodology in Section 2.1, Y is first discretized into m states, where the i-th state is a continuous interval and

denoted by $(y_i, y_{i+1}]$. One can rewrite Eq. (1) using the total probability theorem as follows:

$$P_f = \sum_{i=1}^{m} \Pr(\text{Failure}|Y \in (y_i, y_{i+1}])W_{Y,i} \quad (14)$$

where $\Pr(\text{Failure}|Y \in (y_i, y_{i+1}])$ is the failure probability conditioned on $Y \in (y_i, y_{i+1}]$; $W_{Y,i}$ is the PMF associated with $(y_i, y_{i+1}]$. For the sake of brevity, $\Pr(\text{Failure}|Y \in (y_i, y_{i+1}])$ is denoted as $p_{f,i}$ henceforth. Given that PMF of Y is considered a random, Eq. (14) suggests that P_f is also a random variable, for which the prior mean value and variance are given by Eqs. (15) and (16) respectively,

$$\mu_{P_f}^{\pi} = \sum_{i=1}^{m} p_{f,i}\mu_{W_i}^{\pi} \quad (15)$$

$$\xi_{P_f}^{\pi} = \sum_{i=1}^{m} p_{f,i}^2 \xi_{W_i}^{\pi} + \sum_{1\le i\le m} \sum_{1\le j\le m, j\ne i} p_{f,i}p_{f,j}\omega_{W_{ij}}^{\pi} \quad (16)$$

where $\mu_{W_i}^{\pi}$, $\xi_{W_i}^{\pi}$ and $\omega_{W_{ij}}^{\pi}$ can be computed by Eqs. (3) through (5).

Once $\mathbf{W}_Y$ is updated by a set of samples of Y, the posterior statistics of P_f can be obtained as follows:

$$\mu_{P_f}^{p} = \sum_{i=1}^{m} p_{f,i}\mu_{W_i}^{p} \quad (17)$$

$$\xi_{P_f}^{p} = \sum_{i=1}^{m} p_{f,i}^2 \xi_{W_i}^{p} + \sum_{1\le i\le m} \sum_{1\le j\le m, j\ne i} p_{f,i}p_{f,j}\omega_{W_{ij}}^{p} \quad (18)$$

where $\mu_{W_i}^{p}$, $\xi_{W_i}^{p}$ and $\omega_{W_{ij}}^{p}$ can be computed by Eqs. (8) through (10).

Equations (17) to (18) imply that $\mu_{P_f}^{p}$ and $\xi_{P_f}^{p}$ are functions of the number of samples distributed in the entire sample space of Y, i.e. $[y_1, y_2]$, $(y_2, y_3]$, …, $(y_m, y_{m+1}]$. Given a prescribed total sample size n_{Y0} of Y, the expectations of the posterior mean and variance of P_f with respect to the sampling outcome in the entire space of Y are as follows,

$$\mathrm{E}_N\left[\mu_{P_f}^{p}\right] = \sum_{i=1}^{m} p_{f,i}\mathrm{E}_N\left[\mu_{W_i}^{p}\right] \quad (19)$$

$$\mathrm{E}_N\left[\xi_{P_f}^{p}\right] = \sum_{i=1}^{m} p_{f,i}^2 \mathrm{E}_N\left[\xi_{W_i}^{p}\right] + \sum_{1\le i\le m} \sum_{1\le j\le m, j\ne i} p_{f,i}p_{f,j}\mathrm{E}_N\left[\omega_{W_{ij}}^{p}\right] \quad (20)$$

where $\mathrm{E}_N\left[\mu_{W_i}^{p}\right]$, $\mathrm{E}_N\left[\xi_{W_i}^{p}\right]$ and $\mathrm{E}_N\left[\omega_{W_{ij}}^{p}\right]$ can be computed by Eqs. (11) through (13).

3. THE CRITERION OF OPTIMAL SAMPLE SIZE DETERMINATION

As described in Section 2.2, P_f is a random variable due to the epistemic uncertainty on the basic random variable. The decision is made based on an estimate of P_f, p_f^e. The loss caused by the discrepancy between p_f^e and P_f is modeled by the following quadratic loss function (Morris 1968),

$$L(P_f, p_f^e) = C(p_f^e - P_f)^2 \quad (21)$$

where C is the parameter of the quadratic loss function. Since generally accepted rules to quantify C is scarce in the literature, we determine the magnitude of C based on the following simple heuristic. Equation (21) suggests that the loss increases as the discrepancy between P_f and p_f^e increases. The upper bound of $(p_f^e - P_f)^2$ is equal to unity, which represents the worst estimation of the failure probability and leads to the maximum loss. It is reasonable to assume that the cost of failure of the structure, C_F, is the maximum loss. It then follows that C is equal to C_F.

It has been shown that $p_f^e = \mu_P^{\pi}$ is the optimal prior estimate in the sense of minimizing the expectation of $L(P_f, p_f^e)$ with respect to the distribution of P_f (Morris, 1968), based on which the prior expected loss is,

$$\mathrm{E}_{P_f}[L] = \int C\left(\mu_{P_f}^{\pi} - p_f\right)^2 f_{P_f}^{\pi}(p_f)\mathrm{d}p_f = C\xi_{P_f}^{\pi} \quad (22)$$

where $f_{P_f}^{\pi}(p_f)$ is the prior PDF of P_f. Equation (22) is also known as the expected value of perfect information (EVPI). Once $\mathbf{W}_Y$ and P_f are updated by a set of samples, the posterior expected loss is given by,

$$\mathrm{E}_{P_f}[L|\mathbf{n}_Y] = \int C\left(\mu_{P_f}^{p} - p_f\right)^2 f_{P_f}^{p}(p_f)\mathrm{d}p_f = C\xi_{P_f}^{p} \quad (23)$$

where $f_{P_f}^{p}(p_f)$ denotes the posterior PDF of P_f, and the result of Eq. (23) is known as the conditional value of perfect information (CVPI).

It follows that the conditional value of sampling information (CVSI) is given by,

$$\text{CVSI} = \text{E}_{P_f}[L] - \text{E}_{P_f}[L|\mathbf{n}_Y] = C\xi_{P_f}^{\pi} - C\xi_{P_f}^{p} \quad (24)$$

Given a prescribed sample sizes n_{Y0} of Y, the expectation of CVSI with respect to the sampling outcome in the entire space of Y is the expected value of sampling information (EVSI) as follows,

$$\text{EVSI}(n_{Y0}) = \text{EVPI} - \text{E}_N\left[C\xi_{P_f}^{p}\right] \quad (25)$$

ENGS can then be calculated by,

$$\text{ENGS}\,(n_{Y0}) = \text{EVSI}(n_{Y0}) - n_{Y0}C_S \quad (26)$$

where C_S is the unit cost of sampling. The value of n_{Y0} that maximizes ENGS is the optimal sample size, $n_{Y0\text{-opt}}$.

Note that Eqs. (21) through (26) formulate EVPI, EVSI and ENGS by considering the influence of the epistemic uncertainty on the failure probability evaluation of a single structure. If the epistemic uncertain influences the failure probability evaluation of a group of individual structures, the total EVPI (EVSI) is equal to the sum of EVPI (EVSI) associated with individual structures.

4. NUMERICAL EXAMPLE

4.1. General information

The numerical example considers the reliability evaluation of a corroded pipeline. The buried pipeline segment has a nominal outside diameter D_n = 508 mm, a nominal wall thickness w_{tn} = 5.40 mm and a nominal operating pressure p_n = 5.5 MPa. The pipe is made of API 5L Grade X52 steel with the specified minimum yield strength (SMYS) of 359 MPa. It is assumed that the pipeline segment contains 100 corrosion defects that have been detected and sized by a recently conducted inline inspection (ILI). For simplicity, the ILI-reported defect sizes (depth and length) of different defects are assumed to be identical. The probability of burst of the pipeline at every detected defect is calculated. The burst failure at a given corrosion defect is defined by the following limit state function,

$$g = r_b - p \quad (27)$$

where r_b is the remaining burst pressure capacity of the pipe at the defect calculated by the B31G Modified model (Kiefner and Vieth 1989),

$$r_b = \kappa \frac{2w_t(\sigma_y + 68.95)}{D}\left[\frac{1-0.85\frac{d}{w_t}}{1-0.85\frac{d}{Mw_t}}\right] \quad (28)$$

where d is the actual defect depth; D is the actual outside diameter; w_t is the actual pipe wall thickness; σ_y is the actual yield strength; κ denotes the model error associated with the B31G Modified model, and M is Folios bulging factor which is a function of D, w_t and actual defect length l. d and l are normally distributed with the mean values equal to the ILI-reported depth and length, respectively, and standard deviations equal to $0.078w_{tn}$ and 7.8 mm, respectively (Zhou et al. 2016). The probabilistic properties of the considered random variables are given by Table 1.

Table 1: Probabilistic characteristics of random variables for the numerical example

RV	*Distribution*	*μ*	*COV*	*σ*
d	*Normal*	$0.4w_{tn}$	-	$0.078w_{tn}$
l	*Normal*	*75 mm*	-	*7.8 mm*
D/D_n	*Deterministic*	*1.0*	-	-
w_t/w_{tn}	*Normal*	*1.0*	*0.015*	-
p/p_n	*Gumbel*	*1.0*	*0.03*	-
$\sigma_y/SMYS$	*Lognormal*	*1.1*	*0.035*	-
κ	*Lognormal*	*1.297*	*0.258*	-

RV: random variable
μ: mean value of the random variable
σ: standard deviation of the random variable

The distribution of κ given in Table 1 was estimated mainly from the data of burst tests on pipelines with isolated small defects. However, suppose that the majority of the defects considered in this example are clustered corrosion defects; the probabilistic characterization of κ given in Table 1 does not capture entirely the uncertainty of the burst model for such defects. Therefore, the distribution of κ given in Table 1 is considered as a prior distribution containing epistemic uncertainty. Given the failure probability is highly sensitive to the probabilistic property of κ (Zhou 2010), it is desirable to perform a number of burst tests on naturally

corroded pipelines with clustered defects to collect the samples of κ and update the distribution of κ. The proposed SSD criterion is applied to determine the optimal number of burst tests. In practice, the cost of each full-scale burst test, C_S, is approximately \$100,000. Generally, the loss caused by a pipeline failure, C_F, can be extreme, and is assumed to be $500C_S$ in this study. Therefore, the relative magnitudes of C_F and C_S are 500 and 1, respectively.

4.2. The results of SSD

The prior distribution of κ is discretized into 40 states and the corresponding PMF are plotted in Fig. 1. Then, the PMF is modeled by a prior Dirichlet distributions $\mathbf{W} \sim \text{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \{\alpha_{\kappa,1}, \alpha_{\kappa,2}, \ldots, \alpha_{\kappa,40}\}$. The equivalent sample size, $\alpha_{\kappa 0} = \sum_{i=1}^{40} \alpha_{\kappa,i}$, of the prior Dirichlet distribution is assumed to be unity, which is commonly assumed in the literature (Zhou et al. 2016). $\mu_{P_f}^{\pi}$ and $\xi_{P_f}^{\pi}$ associated with the failure at each defect are calculated to be 0.0068 and 0.0019, respectively. $p_{f,i}$ in Eqs. (15) and (16) is calculated using the simple Monte Carlo (MC) simulation with 1,000,000 trials. Note that in the MC simulation to calculate $p_{f,i}$, κ is sampled from the prior lognormal distribution truncated beyond the boundaries of the state $(\kappa_i, \kappa_{i+1}]$.

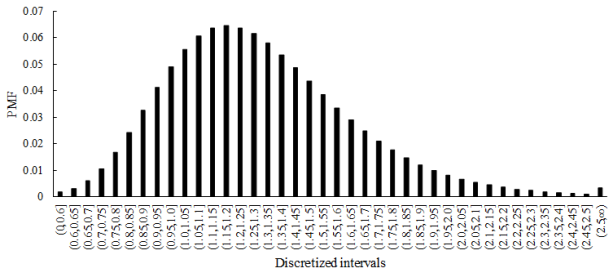


Fig. 1 Discretization and PMFs of κ

EVPI is calculated to be 97, which is the upper bound of EVSI. According to Eq. (26), the EVSI associated with any sample size large than 97 cannot outweigh the associated sampling cost, thus leads to a negative ENGS. Let the sample size $n_{\kappa 0}$ vary from 1 through 100, and the corresponding EVSI and ENGS are calculated and plotted in Fig. 2. This figure indicates that EVSI increases as the sample size increases. However, the contribution from a unit sample to EVSI decreases as the sample size increases. The peak value of ENGS indicates that the optimal number of burst tests is 9.

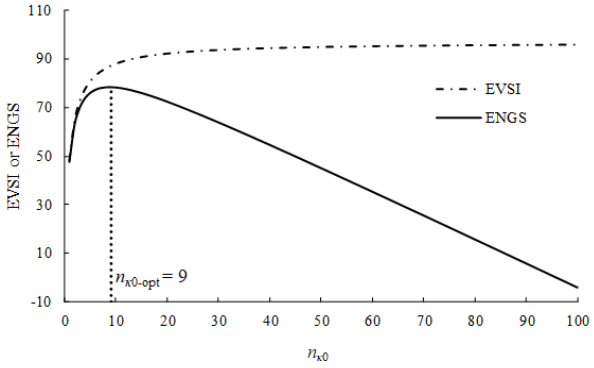
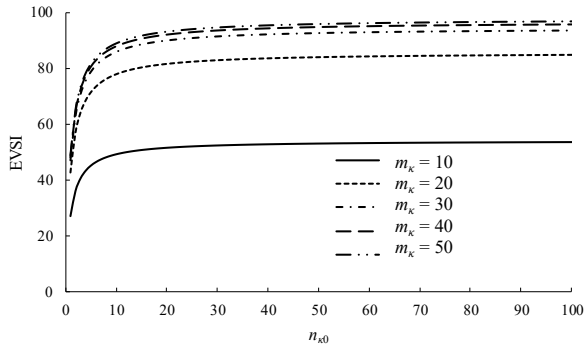


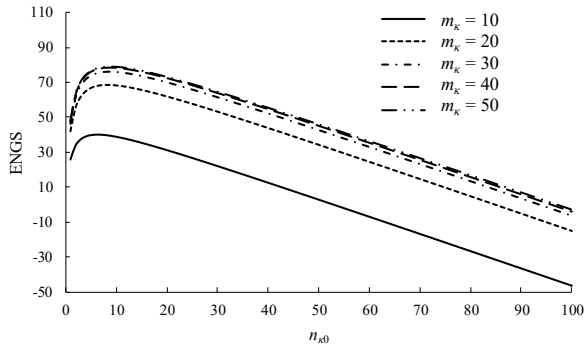
Fig. 2 The results of EVSI and ENGS

4.3. Sensitivity analysis

The sensitivity of the SSD result to relevant parameters is investigated in this section. First, the sensitivity of the optimal sample size to the number of discretization states of κ is investigated. All else being equal, the distribution of κ is discretized into 10, 20, 30, 40 and 50 states, respectively, and the corresponding EVSI and ENGS are plotted in Figs. 3(a) and 3(b), respectively. These figures indicate slight differences between EVSI (ENGS) associated with different discretization cases where m_κ is equal to or larger than 30. That is, the discretization in Fig. 1 is adequate for the SSD problem. Therefore, when the continuous distribution is discretized into a fairly large number of states, the SSD result is insensitive to the number of discretization states. The sufficient number of discretization states for different problems can be determined by the approach of trial-and-error.



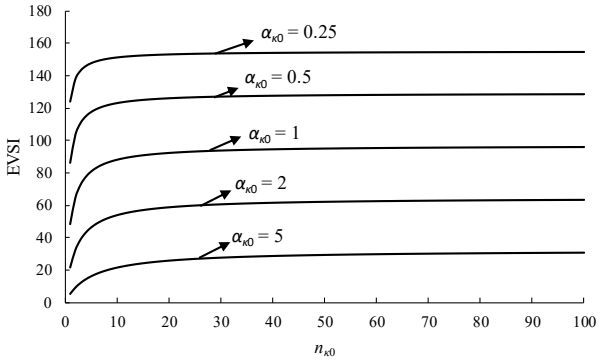
(a) EVSI



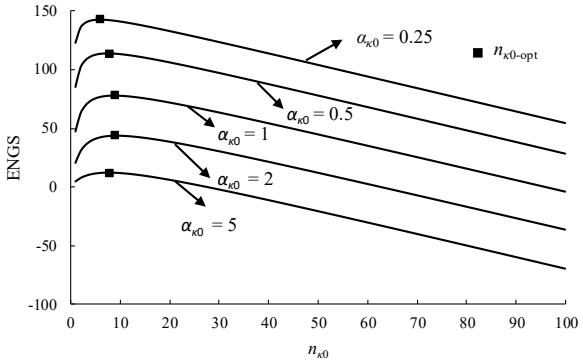
(b) ENGS

Fig. 3 Sensitivity of EVSI and ENGS to m_κ

Then, the sensitivity of the optimal sample size to the equivalent sample size $\alpha_{\kappa 0}$ of the prior Dirichlet distribution is investigated. All else being equal, $\alpha_{\kappa 0}$ is set to 0.25, 0.5, 1, 2 and 5, respectively. The corresponding EVPI equal to 154, 129, 97, 65 and 32, respectively, indicates that EVPI decreases as $\alpha_{\kappa 0}$ increases. The reason is that a larger $\alpha_{\kappa 0}$ implies less uncertainty on the prior Dirichlet distributions as well as P_f. Since EVSI and ENGS depend on EVPI as defined by Eqs. (25) and (26), EVSI and ENGS depicted by Figs. 4(a) and 4(b), respectively, decrease too as $\alpha_{\kappa 0}$ increases. However, the same trend does not hold for the optimal sample size. As $\alpha_{\kappa 0}$ varies through 0.25, 0.5, 1, 2 and 5, the optimal sample size increases at first, but starts to decrease after reaching the greatest value at $\alpha_{\kappa 0}$ equal to 1 and 2. This trend suggests that a relatively small and large $\alpha_{\kappa 0}$ both tend to lead to a smaller optimal sample size, which is explained by the trade-off between two influencing factors, the magnitude of EVSI and sensitivity of EVSI to the sample size. Figure. 4(a) indicates that EVSI increases as $\alpha_{\kappa 0}$ decreases from 5 to 0.25, which tends to lead to a larger optimal sample size according to Eq. (26). On the other hand, as $\alpha_{\kappa 0}$ decreases, the sensitivity of EVSI to the sample size increases. When $\alpha_{\kappa 0}$ is small (i.e. less than unity), a relatively small sample size already enables EVSI to approach its upper bound, which renders the contribution from the subsequent samples to EVSI too small to outweigh their sampling cost. This trend tends to lead to a smaller optimal sample size as $\alpha_{\kappa 0}$ decreases. Therefore, as $\alpha_{\kappa 0}$ decreases, the ultimate change of the optimal sample size depends on which one of the two influencing factors mentioned above dominates.



(a) EVSI



(b) ENGS

Fig. 4 Sensitivity of EVSI and ENGS to $\alpha_{\kappa 0}$

5. CONCLUSIONS

This paper establishes an SSD methodology for collecting samples to reduce the epistemic uncertainties in the distributions of basic random variables involved in the failure probability evaluation. The ENGS maximization approach

determines the optimal sample size by balancing between the sampling cost and associated benefit. The discretization of the continuous variables makes the method applicable to a wide range of distributions.

The effectiveness the proposed method is demonstrated by a numerical example. The following conclusions are drawn from the sensitivity analysis. First, the SSD result is insensitive to the discretization of the basic random variable if the continuous random variable is discretized into a reasonably large number of states. Second, the SSD result is highly sensitive to the equivalent sample size of the prior Dirichlet distribution. Relatively small and large equivalent sample sizes both lead to a small optimal sample size. A further study of eliciting the equivalent sample size from the prior information is desirable in the future.

6. ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), Ontario Trillium Scholarship and Faculty of Engineering at the University of Western Ontario.

7. REFERENCES

Adcock, C.J. (1992) “Bayesian approaches to the determination of sample sizes for binomial and multinomial sampling-some comments on the paper by Pham-Gia and Turkkan” *The Statistician*, 41: 399-404.

Caleyo, F., Valor, A., Alfonso, L. et al. (2015). “Bayesian analysis of external corrosion data of non-piggable underground pipelines” *Corrosion Science*, 90: 33-45.

Der Kiureghian, A. (1989). “Measures of structural safety under imperfect states of knowledge.” *Journal of Structural Engineering*, 115(5): 1119-1140.

Goldsworthy, J.S., Jaksa, M.B., Fenton, G.A., et al. (2007) "Effect of sample location on the reliability based design of pad foundations." *Georisk*, 1(3): 155-166.

Higo, E., and Pandey, M. D. (2016). “Value of information and hypothesis testing approaches for sample size determination in engineering component inspection: a comparison.” In *ASME 2016 Pressure Vessels and Piping Conference* (pp. V005T10A009-V005T10A009). American Society of Mechanical Engineers.

Hong, H.P. (1996) "Evaluation of the probability of failure with uncertain distribution parameters." *Civil Engineering Systems*, 13(2): 157-168.

Jonson, N. L., and Samuel Kotz. (1972) "Distributions in Statistics: Continuous Multivariate Distributions." Wiley, New York, USA.

Kiefner, J.F., and Paul H.V. (1989) “A modified criterion for evaluating the remaining strength of corroded pipe.” No. PR-3-805. Battelle Columbus Div., OH, USA.

Morris, W.T. (1968). “Management science: a Bayesian introduction” Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA.

Nishijima, K., and Faber M.H. (2007) "Bayesian approach to proof loading of quasi-identical multi-components structural systems." *Civil Engineering and Environmental Systems*, 24 (2): 111-121.

Pham-Gia, A., and Turkkan, T. (1992). “Sample size determination in Bayesian analysis” *The Statistician*, 41: 105–112.

Raiffa, H., and Schlaifer, R. (1961). “Applied statistical decision theory.” Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA, USA.

Shafieezadeh, A., and Ellingwood, B. R. (2012). “Confidence intervals for reliability indices using likelihood ratio statistics.” *Structural Safety*, 38: 48-55.

Spiegelhalter, D., Dawid, D., Lauritzen, S., et al. (1993). “Bayesian analysis in expert systems.” *Statistical Science*, 8: 219-282.

Zhou, W. (2010). “System reliability of corroding pipelines.” *International Journal of Pressure Vessels and Piping*, 87(10): 587-595.

Zhou, Y., Fenton, N., and Zhu, C. (2016). “An empirical study of Bayesian network parameter learning with monotonic influence constraints.” *Decision Support Systems*, 87: 69-79.

Zhou, W., Xiang, W., and Cronin, D. (2016). “Probability of rupture model for corroded pipelines.” *International Journal of Pressure Vessels and Piping*, 147: 1-11.