Sampling

There exist several sampling schemes to address the reliability problem. In fact, sampling is equally applicable to component and system reliability problems. For this reason, the failure region is identified by the symbol Ω_f, where

$$\Omega_f = \{g \leq 0\} \tag{1}$$

for component problems and defined in terms of several limit-state functions for system reliability problems, as described in the document on system reliability analysis. In this notation, the reliability problem reads

$$p_f = P\left(\Omega_f\right) = \int\cdots\int_{\Omega_f} f(\mathbf{x})\,d\mathbf{x} \tag{2}$$

where $f(\mathbf{x})$ is the joint PDF for the random variables, which are collected in the vector $\mathbf{x}$.

Monte Carlo Sampling

Several sampling schemes are available to estimate the failure probability, p_f. The simplest and most popular approach is called Monte Carlo sampling, in which samples of the random variables are generated according to the distribution $f(\mathbf{x})$. Monte Carlo sampling is derived in two steps. First, the indicator function is introduced:

$$\int\cdots\int_{\Omega_f} f(\mathbf{x})\,d\mathbf{x} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} I(\mathbf{x})f(\mathbf{x})\,d\mathbf{x} \tag{3}$$

where I=1 for realizations inside Ω_f and 0 elsewhere. It is observed that p_f is the expectation of the indicator function $I(\mathbf{x})$ with respect to the distribution $f(\mathbf{x})$. Next, to obtain a workable expression for the joint probability distribution the integral is transformed into the space of standard normal random variables:

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} I(\mathbf{x})f(\mathbf{x})\,d\mathbf{x} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} I(\mathbf{y})\varphi(\mathbf{y})\,d\mathbf{y} \tag{4}$$

This leads to the following algorithm for Monte Carlo sampling:

1. Generate an outcome $\mathbf{y}_i$ of the n-dimensional random vector $\mathbf{y}$ according to the joint standard normal PDF:

$$\varphi(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y}\right) \tag{5}$$

2. Transform the realization $\mathbf{y}_i$ into the original space of random variables $\mathbf{x}_i$
3. Evaluate the limit-state function(s) $g(\mathbf{x}_i)$ and ultimately the indicator function $I(\mathbf{x}_i)$
4. Update the following average, i.e., expectation:

$$p_f = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{x}_i) \tag{6}$$

where N is the number of samples, which is unrelated to the number of random variables, n.

5. Repeat from Step 1 until sufficiently many samples are analyzed

Compared with this procedure, which aims at computing the failure probability associated with the domain Ω_f it is even simpler to compute some statistics of a response or a limit-state function, or to display a histogram of it.

Importance Sampling

Monte Carlo sampling requires a high number of samples to obtain an accurate estimate of p_f if p_f is small. If a FORM analysis has preceded the sampling analysis then a far more efficient sampling scheme is obtained by utilizing the design point from FORM as the centre for the sampling distribution. The use of a sampling distribution that is centred closer to the failure region than the mean of the random variables is called importance sampling. To derive it, reconsider Eq. (4) and multiply the integrand by the auxiliary unit fraction $h(\mathbf{y})/h(\mathbf{y})$.

$$p_f = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} I(\mathbf{y})\phi(\mathbf{y})\,d\mathbf{y} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(I(\mathbf{y})\frac{\varphi(\mathbf{y})}{h(\mathbf{y})}\right)h(\mathbf{y})\,d\mathbf{y} \tag{7}$$

It is observed that the failure probability is now the expectation of

$$q(\mathbf{y}) \equiv I(\mathbf{y})\frac{\varphi(\mathbf{y})}{h(\mathbf{y})} \tag{8}$$

with respect to the distribution $h(\mathbf{y})$, which is the new distribution that $\mathbf{y}$ are sampled from. In importance sampling around the design point from FORM, $h(\mathbf{y})$ is usually selected as the shifted standard normal distribution

$$h(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{y}^*)^T(\mathbf{y}-\mathbf{y}^*)\right) \tag{9}$$

where $\mathbf{y}^*$ is the design point coordinates. Otherwise the Monte Carlo sampling procedure holds valid also here, with the following expression for the failure probability:

$$p_f = \frac{1}{N}\sum_{i=1}^{N} q_i \tag{10}$$

where q_i shorthand notation for $q(\mathbf{y}_i)$, which is defined in Eq. (8).

Coefficient of Variation of the Sampling Result

The coefficient of variation of the failure probability, denoted δ_{pf}, is monitored to gauge when the sampling analysis has reached a sufficient level of accuracy. A coefficient of variation around 2-5% is usually considered acceptable. Higher values may draw the accuracy of p_f into question. The coefficient is defined as

$$\delta_{pf} = \frac{\sigma_{pf}}{\mu_{pf}} \tag{11}$$

The derivation of δ_{pf} starts with the expression for the failure probability in Eq. (10). As noted in the document on analysis of functions, expectation is a linear operator, thus the mean of p_f is

$$\mathrm{E}[p_f] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[q_i] = \frac{1}{N}\cdot N\cdot \mathrm{E}[q] = \mathrm{E}[q] \tag{12}$$

where q_i is a general shorthand notation for $I(\mathbf{x}_i)$ in Eq. (6) and the variance of p_f is

$$\mathrm{Var}[p_f] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[q_i] = \frac{1}{N^2}\cdot N\cdot \mathrm{Var}[q] = \frac{1}{N}\cdot \mathrm{Var}[q] \tag{13}$$

Substitution of Eqs. (12) and (13) into Eq. (11) yields

$$\delta_{pf} = \frac{1}{\sqrt{N}}\cdot\frac{\sqrt{\mathrm{Var}[q]}}{\mathrm{E}[q]} \tag{14}$$

For Monte Carlo sampling, Var[q] and E[q] are determined from the fact that q is a discrete random variable that can take on the values 0 and 1. The probability of q=1 equals p_f. Consequently:

$$\begin{aligned} \mathrm{E}[q] &= \sum_{i=1}^{2} q_i\cdot p(q_i) = 1\cdot p_f + 0\cdot(1-p_f) = p_f \\ \mathrm{Var}[q] &= \sum_{i=1}^{2}(q_i-\mu_q)^2\cdot p(q_i) \\ &= (1-p_f)^2\cdot p_f + (0-p_f)^2\cdot(1-p_f) = p_f\cdot(1-p_f) \end{aligned} \tag{15}$$

Substitution of Eq. (15) into (14) yields the coefficient of variation of p_f from Monte Carlo sampling:

$$\delta_{pf} = \frac{1}{\sqrt{N}}\cdot\frac{\sqrt{p_f\cdot(1-p_f)}}{p_f} = \sqrt{\frac{1-p_f}{N\cdot p_f}} \tag{16}$$

Solving for N yields the necessary number of samples to achieve a prescribed coefficient of variation for a specific target failure probability:

$$N = \frac{1}{\delta_{pf}^2}\left(\frac{1-p_f}{p_f}\right) \tag{17}$$

For example, if a failure probability around 10^{-3} is expected then around 399,600 samples are needed to achieve a 5% coefficient of variation and around 2,497,500 samples are needed to achieve a 2% coefficient of the failure probability. The phrase “variance reduction techniques" are often applied to sampling methods that reduce the variance of the sampling result, i.e., the variance in Eq. (14) more rapidly than Monte Carlo sampling.

The Concept of Subset Sampling

This sampling approach employs the multiplication rule of probability to subset events. Consider an event E2 that is a subset of the event E1. Furthermore, let E3 be a subset of E2, let E4 be a subset of E3, and so forth. This situation is visualized in Figure 1 in two ways. The left picture shows a Venn diagram in which the event E_{m+1} is the subset of the event E_m. The right-hand side diagram shows the same situation for a reliability problem, in which the failure domain of the limit-state function G_{m+1} is the subset of the failure domain of the limit-state function G_m.

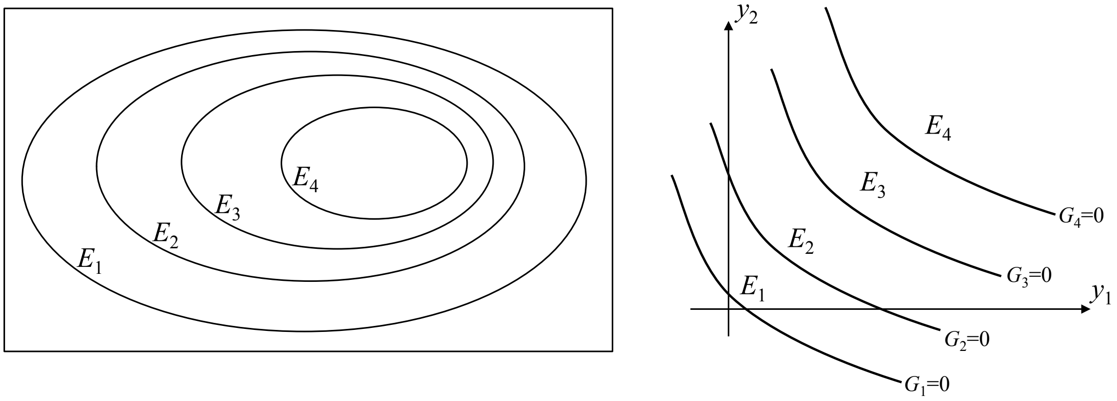

Figure 1: Subset events.

Because the event E_{m+1} is the subset of the event E_m the probability P(E_{m+1}) equals the intersection probability P($E_{m+1}E_m$). Furthermore, application of the multiplication rule of probability yields

$$P\left(E_{m+1}\right)=P\left(E_{m+1}E_n\right)=P\left(E_{m+1}\middle|E_m\right)\cdot P\left(E_m\right) \tag{18}$$

Recursive use of this equation for the examples visualized in Figure 1 yields

$$P\left(E_4\right)=P\left(E_4\middle|E_3\right)\cdot P\left(E_3\middle|E_2\right)\cdot P\left(E_2\middle|E_1\right)\cdot P\left(E_1\right) \tag{19}$$

The reason why this leads to a more efficient sampling method than Monte Carlo sampling is described with reference to the right-hand side of Figure 1. Let G_4 be the actual limit-state function for which the failure probability is sought. Subtract a constant to this limit-state function so that its mean value is approximately zero. In other words, a constant is subtracted so that G_4 turns into G_1 in Figure 1. Later modification of the constant will yield G_2 and G_3. It is clear that the origin-centred Monte Carlo approach will provide quite accurate results for G_1, i.e., P(E_1) with relatively few samples. Next, assume that a sampling distribution $h(\mathbf{y})$ could be constructed that generated realizations inside the failure domain E_1. Sampling with that distribution would lead to quite accurate results for the probability P($E_2|E_1$) with relatively few samples. The sampling distribution is subsequently updated to address the other conditional probabilities P(E_{m+}1|E_m) for sequentially increasing n. Finally, Eq. (19) is evaluated by multiplying the results. The key difficulty with subset sampling is the construction of a sampling distribution, $h(\mathbf{y})$ that generates samples only within specified regions E_n in the space of random variables.