International Conference
Nuclear Energy for New Europe 2002
Kranjska Gora, Slovenia, September 9-12, 2002
www.drustvo-js.si/gora2002

TRANSFORMATION OF BAYESIAN POSTERIOR DISTRIBUTION INTO A BASIC ANALYTICAL DISTRIBUTION

Romana Jordan Cizelj
"Jožef Stefan" Institute
Reactor Engineering Division
Jamova 39, SI-1000 Ljubljana, Slovenia
romana.jordan@ijs.si

Ivan Vrbanic
Nuclear Power Plant Krško
Vrbina 12, SI- 8270 Krško, Slovenia
ivan.vrbanic@nek.si

ABSTRACT

Bayesian estimation is well-known approach that is widely used in Probabilistic Safety Analyses for the estimation of input model reliability parameters, such as component failure rates or probabilities of failure upon demand. In this approach, a prior distribution, which contains some "generic" knowledge about a parameter is combined with likelihood function, which contains plant-specific data about the parameter. Depending on the type of prior distribution, the resulting posterior distribution can be estimated numerically or analytically. In many instances only a numerical Bayesian integration can be performed. In such a case the posterior is provided in the form of tabular discrete distribution. On the other hand, it is much more convenient to have a parameter's uncertainty distribution that is to be input into a PSA model to be provided in the form of some basic analytical probability distribution, such as lognormal, gamma or beta distribution. One reason is that this enables much more convenient propagation of parameters' uncertainties through the model up to the so-called top events, such as plant system unavailability or core damage frequency. Additionally, software tools used to run PSA models often require that parameter's uncertainty distribution is defined in the form of one among the several allowed basic types of distributions. In such a case the posterior distribution that came as a product of Bayesian estimation needs to be transformed into an appropriate basic analytical form. In this paper, some approaches on transformation of posterior distribution to a basic probability distribution are proposed and discussed. They are illustrated by an example from NPP Krško PSA model.

1 INTRODUCTION

Reliability of complex systems is today one of the most important areas of engineering research. Because the assessment demands evaluation of a large amount of information, which is often not exact or precisely known, mathematical assessment is based on probability theory and theory of statistics. A part of such assessment is determination of probability distributions, depending on mathematical and practical limitations and assumptions, as for example: quantity and quality of raw data, and limitations of used computer code. Often determination of probability distribution involves examining a random sample from some

unknown distribution in order to test the null hypothesis that the unknown distribution function is, in reality, a specified function.

Probabilistic Safety Assessment (PSA) is a systematical assessment of reliability and availability of complex system and is widely used for chemical, nuclear and petroleum industry. The well-known PSA approach for estimation of model input parameter is Bayesian parameter estimation[1, 2]. The prior distribution, which contains some "generic" knowledge about a parameter is combined with likelihood function, which contains specific data about the parameter[3]. Resulting posterior distribution can be estimated numerically or analytically[4, 5]. If numerical calculation is used, the posterior distribution would be obtained in a tabular discrete form. It may, however, become desirable to replace it with one of appropriate basic probability distributions, for example lognormal, gamma or beta distribution. After the transformation, the transfer of data would be easier and generally more suitable for a further analysis.

The numerically obtained tabular posterior distribution ought to be compared to the selected basic probability distribution to determine if it is reasonable to postulate that the latter is the true distribution function of the random sample considered. A good method to test the selected basic distribution is to compare cumulative distribution functions to see if there is good agreement.

In this paper, some approaches on transformation of numerically obtained posterior distribution to a basic probability distribution are proposed. Additionally, the Kolmogorov-Smirnov goodness-of-fit test and some new mathematical approaches for evaluation of agreement between discrete and basic probability distribution are introduced. The research was performed in the frame of PSA models development.

2 TRANSFORMATION OF POSTERIOR DISTRIBUTION TO A BASIC ANALYTICAL DISTRIBUTION

The term "posterior distribution" is in this paper used for probability distributions, which were obtained by Bayesian estimation with numerical calculation.

The term "basic analytical distribution" is used for probability distributions that are usually applied in analysis of component reliability and are allowed to be used in a PSA computer code, for example lognormal distribution, Weibull distribution, Gamma distribution or beta distribution.

Let a basic analytical probability distribution be a two-parameter probability distribution $f_A(x;p_1, p_2)$. The unknown parameters p_1 and p_2 are calculated on the basis of some selected values of known Bayesian posterior distribution $f_B(x)$. The choice of selected values depends on the application or on the use of analytical distribution.

The common approach is the so-called Method of Moments, which is based on point estimates of the distribution mean and variance[6, 7]:

$$mean_B = \int_0^\infty x \cdot f_B(x) \cdot dx = \bar{x}, \tag{1}$$

$$\mathrm{var}_B = \int_0^\infty (x - \bar{x})^2 \cdot f_B(x) \cdot dx, \tag{2}$$

where $mean_B$ is the first moment about the origin or mean of posterior distribution, var_B is the second moment about its mean or variance, and x_i random variable.

Other possible characteristic values, which could be selected for estimation of parameters of a basic analytical distribution are median and limits of a selected probability interval. To estimate characteristic values, one should numerically solve the following equation:

$$F_B(x_l) = \int_0^{x_l} f_B(x) \cdot dx\,, \tag{3}$$

where $F_B(x_l)$ is the selected value of cumulative posterior distribution and x_l the calculated value of random variable x. For example, if calculated value of random variable is median then l = 0,5 and $F_B(x_{0,5})$ = 0,5.

Some appropriate combinations of characteristic values for determination of 2-parameters basic analytical distribution can be:

- mean and variance ($mean_B$, var_B),
- mean and median ($mean_B$, $median_B$),
- the lower and the upper limit of the 90% probability interval ($x_{B0,05}$, $x_{B0,95}$),
- the upper limit of the 90% probability interval and mean($x_{B0,95}$, $mean_B$).

3 GOODNESS-OF-FIT EVALUATION

The so-called goodness-of-fit tests are used to determine whether a sample belongs to a hypothesized analytical distribution. The two procedures, very often used for this purpose, are the Chi-square and the Kolmogorov-Smirnov goodness-of-fit tests[6, 7].

The Chi-square test uses the statistic c^2, which (if applied to transformation of Bayesian posterior into a basic analytical distribution, as discussed in this paper) would be defined as:

$$c^2 = \sum_{i=1}^{I} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{I} \frac{\left[\left(F_B(x_i) - F_B(x_{i-1})\right) - \left(F_A(x_i) - F_A(x_{i-1})\right)\right]^2}{F_A(x_i) - F_A(x_{i-1})}\,, \tag{4}$$

where o_i is the observed frequency and e_i the expected frequency for each interval i, and I the number of nonoverlapping intervals of observed sample.

The c^2 statistics depends considerably on the selection of the lowest value of x. Namely, the ratio in equation (4) could be very high for small values of random variable, thus suggesting that the hypothesized distribution should be rejected. However, the probability distributions could often be very approximate for small values of random variable, because this inprecision has negligible influence on the application results.

For this reason, more suitable is Kolmogorov-Smirnov goodness-of-fit test, which measures the maximal difference between the cumulative distribution functions of the two distributions. Applied to the case considered, the Kolmogorov-Smirnov statistic K-S is defined as:

$$K - S = \underset{i}{Sup}\left|F_A(x_i) - F_B(x_i)\right|, \tag{5}$$

where the *Sup* indicates the supremum.

Besides the two tests mentioned above, one can define some other tests to decide if a hypothesized analytical distribution meets the selected requirements. These tests are based on analyst's judgment and their use depends on the application. Bellow, three possible tests are described.

First, one can define the maximum difference between the selected characteristic values of posterior and analytical distribution:

$$K_l \le \frac{x_B}{x_A} \le K_u\,, \tag{6}$$

where K_l and K_u are the lower and the upper limits of acceptability interval (for example K_l = 0,95 and K_u = 1,05), and x_B and x_A the selected characteristic values of posterior and hypothesized analytical distribution (for example mean, variance or median).

There are several other possibilities to compare the cumulative distribution functions of analytical distribution and Bayesian posterior distribution. One can compare the relative area between the two cumulative distribution functions:

$$\frac{\int F_A(x)\cdot dx - \int F_B(x)\cdot dx}{\int F_B(x)\cdot dx} \le K, \tag{7}$$

where K is the arbitrary selected value, for example K = 0,05.

One can compare the relative difference *RD* between the cumulative distribution functions of analytical distribution and Bayesian posterior distribution and limit it to some arbitrary selected value K_{RD}:

$$RD(x_i) = \frac{\left|F_A(x_i) - F_B(x_i)\right|}{F_B(x_i)} \quad \le \quad K_{RD}, \quad \forall x_i, \tag{8}$$

One can also define the maximum absolute difference *AD* between cumulative distribution functions of analytical distribution and Bayesian posterior distribution and limit it with some arbitrary selected value K_{AD}:

$$AD(x_i) = \left|F_A(x_i) - F_B(x_i)\right| \quad \le \quad K_{AD}, \quad \forall x_i. \tag{9}$$

4 EXAMPLE

For an example, the posterior distribution is used, calculated with the Bayesian numerical estimation procedure with the following analysis data:

- the prior probability density function is lognormal distribution with the mean equal to 3,00E-5 [1/hours] and the variance equal to 5,48E-9,
- the likelihood function is the Poisson distribution, modeling probability of component failure. Raw data for distribution parameters' calculation are: 1 failure of the component in 20.858 [hours] of operation.

A numerical calculation gives the following values for the posterior distribution mean and variance:

$$\begin{aligned} mean_B &= 3{,}22E-5, \\ \mathrm{var}_B &= 9{,}21E-10, \\ median_B &= 2{,}20E-5, \\ x_{B0,05} &= 3{,}70E-6, \\ x_{B0,95} &= 8{,}90E-5. \end{aligned} \tag{10}$$

For a hypothesized analytical distribution the lognormal distribution is chosen:

$$f_A(x) = \frac{1}{x\cdot s\cdot\sqrt{2p}}\cdot e^{-\frac{(\ln x - m)^2}{2\cdot s^2}}, \tag{11}$$

where s and m are the parameters of the lognormal distribution.

All four possibilities suggested in section 2 as appropriate combinations of characteristic values for determination of 2-parameters basic analytical distribution are taken into account in this example, thus leading to four different lognormal distributions $f_{Aj}(x)$. Calculation of distribution parameters is shown in Table 1.

Table 1: Calculation of parameters of hypothesized lognormal distribution

Characteristic values	j	m_j	s_j
$mean_B$, var_B	1	$\ln \frac{mean_B}{\sqrt{1+\frac{\mathrm{var}_B}{mean_B^2}}}$	$\sqrt{\ln(1+\frac{\mathrm{var}_B}{mean_B^2})}$
$mean_B$, $median_B$	2	$\ln(median_B)$	$\sqrt{2\cdot(\ln(mean_B - m_2)}$
$x_{B0,05}$, $x_{B0,095}$	3	$\frac{\ln(x_{B0,05}\cdot x_{B0,95})}{2}$	$\frac{\ln\left(\frac{x_{B0,95}}{x_{B0,05}}\right)}{3,29}$
$mean_B$, $x_{B0,095}$	4	$\ln(mean_B)-\frac{s_4^2}{2}$	$1,645+\sqrt{\ln\left(\frac{mean_B}{x_{B0,95}}\right)^2+1,645}$

The posterior distribution $f_B(x)$ is approximated with lognormal distributions $f_{Aj}(x)$, j = 1,…,4. The lognormal distributions $f_{Aj}(x)$ are shown in Table 2 and Figure 1. The differences of characteristic values of lognormal distributions are reasonably small. If a posterior distribution can be approximated with a distribution from a family of distributions, then the selected characteristic values for determination of an analytical distribution depend on the application, namely on the intended use of hypothesized analytical distribution.

Table 2: **Characteristic values of the lognormal distributions $f_{Aj}(x)$, j = 1,…,4**

		lognormal distribution				
pdf	characteristic values	5%	median	mean	95%	variance
$f_B(x)$	-	3,70E-06	2,20E-05	3,22E-05	8,90E-05	9,21E-10
$f_{A1}(x)$	mean, var	6,20E-06	2,30E-05	3,17E-05	9,60E-05	9,13E-10
$f_{A2}(x)$	mean, median	5,10E-06	2,20E-05	3,16E-05	1,01E-04	1,17E-09
$f_{A3}(x)$	5% in 95%	3,60E-06	1,80E-05	2,84E-05	1,01E-04	1,29E-09
$f_{A4}(x)$	mean, 95%	5,80E-06	2,30E-05	3,16E-05	9,90E-05	1,00E-09

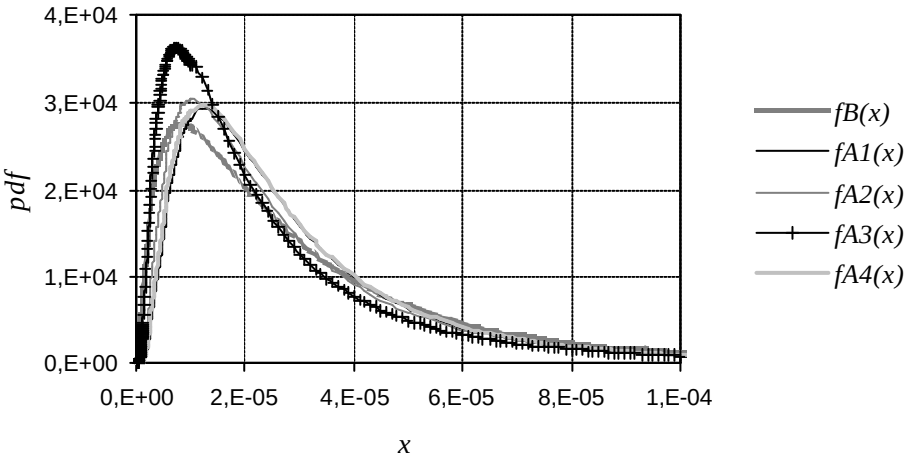

Figure 1: **The lognormal distributions determined on different assumptions**

The decision, if a hypothesized analytical distribution appropriately fits the posterior distribution, can be based on different tests. The approximation of the posterior distribution with characteristic values as shown in equation (10) with the four different lognormal distributions given in equation (11) and Table 1, is examined with four tests, introduced in section 3. Results are shown in Table 3.

Table 3: **Goodness-of-fit evaluation**

	goodness-of-fit-evaluation						
	x_B/x_A						
	mean	*variance*	Chi-square	Kolmogorov	Area	Relative	Absolute
	criteria						
	K_l=0,95; K_u=1,05		<800	0,041	K = 0,050	K_{RD} = 0,050	K_{AD} = 0,050
$f_{A1}(x)$	0,983	0,991	0,167	0,071	0,017	1,000	0,071
$f_{A2}(x)$	0,982	1,000	0,051	0,037	0,018	1,000	0,037
$f_{A3}(x)$	0,973	1,135	0,059	0,106	0,019	1,000	0,085
$f_{A4}(x)$	0,983	1,112	0,107	0,058	0,017	1,000	0,058

The deviation of the lognormal distributions $f_{Aj}(x)$ from posterior distribution $f_B(x)$ can be graphically represented with factors $RD(x_i)$ or $AD(x_i)$, namely relative or absolute difference between the lognormal cumulative distribution function $F_{Aj}(x)$ and posterior cumulative distribution function $F_B(x)$. The factor $AD(x_i)$ is shown in Figure 2.

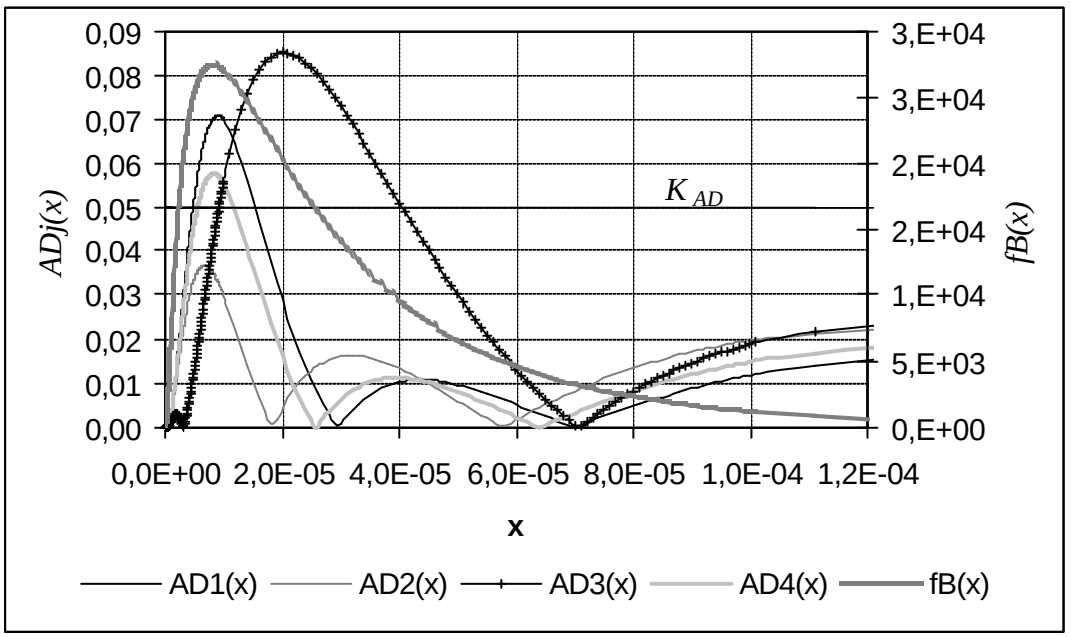

Figure 2: **The absolute difference between lognormal and posterior cumulative distribution function $AD(x_i)$**

The selection of test, which is to be used for estimation of appropriateness of a transformation from posterior distribution to some hypothesized analytical distribution depends on the application. Results of PSA calculations are often presented with mean and variance, so the best way to check the appropriateness of transformation from posterior to analytical

distribution is to limit the maximum difference between these values for two distributions, see equation (6). Additionally, the difference can be limited with factors K, K_{RD} or K_{AD}, equations (7), (8) and (9) respectively.

5 CONCLUSIONS

In the paper, the transformation of a known Bayesian posterior distribution to a hypothesized analytical distribution is discussed. The two well-known goodness-of-fit tests are represented and their use for probabilistic rather than statistical evaluation is disscussed. Some additional technics are proposed to evaluate goodness-of-fit between posterior and hypothesized analytical distribution.

In the example, the known posterior distribution was approximated with lognormal distribution. The main conclusions are:

- If a posterior distribution can be approximated with an analytical distribution from a family of distributions, then the selected assumptions for determination of a hypothesized analytical distribution depend on the application, namely on the use of analytical distribution.
- A number of different tests can be used to evaluate the appropriateness of a transformation from posterior distribution to some hypothesized analytical distribution. The selection of test depends on the application.
- The need to transform a Bayesian posterior distribution to a hypothesized analytical distribution appears often in PSA input parameters evaluation. Results of PSA calculations are often presented with mean and variance. Consequently, in the example considered it was decided, that the best way to ensure the appropriateness of transformation from a posterior to basic analytical distribution is to limit the maximum difference between their mean and variance values. Additionally, the absolute difference between the values of the two cumulative distribution functions was also limited.

6 REFERENCES

1. Hickman, J. W., and others. *PRA Procedures Guide, A Guide to the Performance of Probabilistic Risk Assessments for Nuclear Power Plants.* NUREG/CR-2300. USA: Nuclear Regulatory Commission, 1983.
2. *Procedures for Conducting Probabilistic Safety Assessments of Nuclear Power Plants (Level 1)*. Safety Series No. 50-P-4. Vienna, Austria: International Atomic Energy Agency, 1992.
3. Martz, Harry F., and Ray A. Waller. *Bayesian Reliability Analysis*. USA: John Wiley & Sons, Inc., 1982.
4. Jordan Cizelj, Romana, and Ivan Vrbanic. "Parameter Estimation of Component Reliability Models in PSA Model of Krško NPP. "*Nuclear Energy in Central Europe 2001*, Portorož, Slovenia. Ljubljana: Nuclear Society of Slovenia.
5. Vrbanic, Ivan, and Romana Jordan Cizelj. "Uncertainty Analysis of Component Failure Model Parameters in PSA: A Case Study. "*PSAM 2002*, San Juan, Puerto Rico, USA.
6. Modarres, Mohammad, Mark Kaminskiy, and Vasiliy Krivtsov. *Reliability Engineering and Risk Analysis, A Practical Guide*. New York, USA: Marcel Dekker, Inc., 1999.
7. Lawless, J. F. *Statistical Models and Methods for Lifetime Data.* New York, USA: John Wiley & Sons, 1982.