LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-739175

Confidence Intervals from Realizations of Simulated Nuclear Data

W. Younes, A. Ratkiewicz, J. J. Ressler

September 29, 2017

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Confidence intervals from realizations of simulated nuclear data

W. Younes, A. Ratkiewicz, J. J. Ressler

September 28, 2017

Lawrence Livermore National Laboratory, Livermore, CA 94551

Abstract

Various statistical techniques are discussed that can be used to assign a level of confidence in the prediction of models that depend on input data with known uncertainties and correlations. The particular techniques reviewed in this paper are: 1) random realizations of the input data using Monte-Carlo methods, 2) the construction of confidence intervals to assess the reliability of model predictions, and 3) resampling techniques to impose statistical constraints on the input data based on additional information. These techniques are illustrated with a calculation of the k_{eff} value, based on the $^{235}\text{U}\,(n,f)$ and $^{239}\text{Pu}\,(n,f)$ cross sections.

1 Introduction

The propagation of uncertainty in models that depend on data with measured values and correlations can be carried out using standard techniques from statistical inference theory. The goal is to quantify a level of confidence in a value predicted by the model, given the uncertainties and correlations in its inputs. This report presents some techniques to construct a confidence interval for the prediction of a model with known input values and correlations. Section 2 presents the methodology, which begins with a Monte-Carlo procedure to generate random realizations of correlated input data, and then constructs confidence intervals for the predictions of the model. Section 3 illustrates the methodology using calculations of the k_{eff} value based on random realizations of the energy-dependent $^{235}\text{U}\,(n,f)$ and $^{239}\text{Pu}\,(n,f)$ cross sections as input. In addition, appendix A reviews the confidence interval formalism, and appendix B gives a brief introduction to resampling techniques that can be used in case the input data need to be brought into better agreement with supplementary experimental information.

2 Methodology

2.1 Stochastic versus deterministic

In this section, we consider an uncertainty quantification (UQ) method for a model that depends on data with known uncertainties. In section 3 we give the example of k_{eff} calculations using evaluated (n, f) cross sections with a known energy-energy covariance matrix. The method we will use is based on a Monte-Carlo simulation of the input data in order to generate a corresponding random sample of outputs from the model. The model may be easy to compute, in which case a large random sample can be used, or computationally costly, in which case a smaller sample may have to be used. Regardless of of the size of the sample, it is always possible to generate a confidence interval, with a well-defined statistical interpretation for the model outputs.

In general, the error propagation for a model that is a function of uncertain parameters can be performed by either a deterministic or a stochastic method. An example of a deterministic error propagation method is given by the Kalman-filter technique described in [1]: the model is first approximated as a linear function of its parameters and then, given data values and their covariance matrix (and optionally a covariance matrix for the optimal parameter values of the model), a set of linear equations is derived to update the parameter values and their covariances. In a stochastic approach, random realizations of the model input data are generated in a Monte-Carlo process, and the model is executed to calculate outputs for each random input. After a sufficient number of Monte-Carlo trials, the output can be analyzed by various statistical techniques. In this report, we will focus on the stochastic approach to generate model outputs and on confidence intervals (CI) [3, 4] as the statistical technique used to interpret those outputs. The main advantages of stochastic UQ methods are that they are typically easier to implement than the deterministic ones and are not limited to a linear approximation for the model. The main disadvantage of stochastic methods is the computational cost, if a large number of Monte-Carlo simulations is required to achieve a required accuracy in the outputs.

2.2 Random realizations of input quantities

We consider a model M which is a function of a vector $\vec{x}$ of input values and which produces an output quantity y,

$$y = M(\vec{x}) \tag{1}$$

The problem can be generalized to a vector of output values $\vec{y}$, but the principle remains the same, and the application in section 3 is closer to the formulation in Eq. (1). We will not need to assume that the function $M(\vec{x})$ is linear, we will however assume that the components of $\vec{x}$ are random variables with a normal probability distribution function, and that their mean values $\vec{\mu}$ and covariance matrix Σ are provided. The assumption of a normal distribution is not essential, and the results can be generalized to other other types of distributions (see, e.g.,

section A.2.2 and reference [10]). There are well-known techniques for generating random realizations of $\vec{x}$ (see, e.g., [2] and references therein). First, a matrix C is determined such that

$$CC^T = \Sigma \tag{2}$$

where C^T is the transpose of matrix C, and Σ is the covariance matrix defined above. There are many ways of constructing the matrix C (e.g., Cholesky decomposition, matrix square root, principal component analysis [2]). Once C has been generated, a vector of (uncorrelated) values $\vec{x}_0$ is produced by drawing each of its components independently from a standard normal distribution (i.e., with mean 0 and standard deviation 1). A random realization of $\vec{x}$, whose components are correlated through the covariance matrix Σ is then given by the linear operation

$$\vec{x} = \vec{\mu} + C\vec{x}_0 \tag{3}$$

Note that different decomposition methods can produce different matrices C, but as long as they satisfy Eq. (2), then Eq. (3) will, after a sufficient number of Monte-Carlo trials, produce the same probability distribution of $\vec{x}$ with mean $\vec{\mu}$ and covariance matrix Σ.

When an experimental measurement of y and its uncertainty are available, they can be compared to the distribution of values calculated using Eq. (1) with random realizations of $\vec{x}$. If there is an unacceptable discrepancy between the measured and simulated y distributions, it may be possible to bring them into better agreement by resampling the realizations of $\vec{x}$. This technique is described in appendix B. In this sense, resampling can be used to impose additional experimental constraints on the input data. An example is given in section 3.

2.3 Confidence intervals

Using the technique described in section 2.2, random realizations of the input $\vec{x}$ are generated, and Eq. (1) can be used to calculate the corresponding output values y_i of the model, where $i = 1, \ldots, n$ for n random realizations. For example, the various $\vec{x}$ could be realizations of (n, f) cross sections as a function of incident energy and the y_i values could be corresponding calculated values of k_{eff}. We now want to make statistical inferences about the y_i values, and this can be done using confidence intervals. Appendix A summarizes the formalism needed to calculate CIs. Once we have determined the sample mean ($\bar{y}$) and standard deviation (s_y) of the set of y_i values using Eqs. (5) and (6), respectively, we can make the formal statement that the true (population) mean μ_y is contained in the interval

$$\bar{y} - t_{\alpha/2}\frac{s_y}{\sqrt{n}} \leq \mu_y \leq \bar{y} + t_{\alpha/2}\frac{s_y}{\sqrt{n}} \tag{4}$$

with $100 \times (1 - \alpha)\,\%$ confidence, for a given value of α (and where $0 \leq \alpha < 1$). The $t_{\alpha/2}$ value in Eq. (4) depends only on the underlying distribution and the α value (in particular, it does not depend on the sampled y_i values), and can be

calculated once and for all. Table 4 in appendix A gives $t_{\alpha/2}$ values for a normal population distribution. We can interpret Eq. (4) In the context of the random realizations described in section 2.2 as follows. Suppose we could generate a very large number of realizations of $\vec{x}$ (i.e., $n \to \infty$), then the sample mean $\bar{y}$ of the y_i values would tend to the population mean μ_y. If we repeatedly sampled smaller sets of size n from the very large set of y_i values (the one where $n \to \infty$), then for say $\alpha = 0.05$, we would expect that 95% of the CIs constructed using Eq. (4) would contain the population mean μ_y. We give concrete examples of this interpretation in section 3.

3 Application to $k_{\mathbf{eff}}$

We illustrate the methodology described in section 2 with a calculation of k_{eff} using random realizations of (n, f) cross sections. In a first example, 3984 random realizations of the $^{235}\text{U}\,(n, f)$ cross section were generated using the code Kiwi [5, 6, 2] and evaluated data from the ENDL2009.2 database. The random realizations are plotted in Fig. 1. For each cross section realization, a corresponding k_{eff} was calculated for Godiva, a bare sphere of highly enriched uranium, using two different transport codes: ARDRA, a deterministic code [7, 8], and Mercury, a Monte-Carlo code [9]. Figure 2 shows histograms of the calculated k_{eff} values obtained from the two codes. Table 1 gives the mean and standard deviations of the data plotted in Fig. 2. Note that the calculated standard deviations agree well with the uncertainty of the experimental k_{eff} value [12]. This may be purely coincidental, or it may be that the evaluated $^{235}\text{U}\,(n, f)$ cross section was adjusted at some point to match the experimental k_{eff} value. Whatever the origin of this agreement, it means that the cross-section realizations do not need to be resampled in order to match the experimental k_{eff} value.

	ARDRA	Mercury	Experiment
mean	1.00055	1.00035	1.0000
std dev	0.00271	0.00271	0.0026

Table 1: Mean and standard deviation of the calculated k_{eff} from the ARDRA and Mercury transport codes, compared to the measured experimental value and its uncertainty. These statistics correspond to the distributions in Fig. 2.

We can now construct and test confidence intervals for the k_{eff} value. To demonstrate that CIs can be reliably estimated even for small sample sizes, we sample 5 cross sections at a time with replacement from the 3984 realizations and their corresponding k_{eff} values obtained from ARDRA, and construct a confidence interval from those 5 cross sections using Eq. (4) and the $t_{\alpha/2}$ values in Table 4. We repeat this process 100000 times and count the fraction of times that the confidence interval contains the value $k_{\text{eff}} = 1.00055$ from Table 1. Table 2 compares requested CI for different confidence levels with the actual

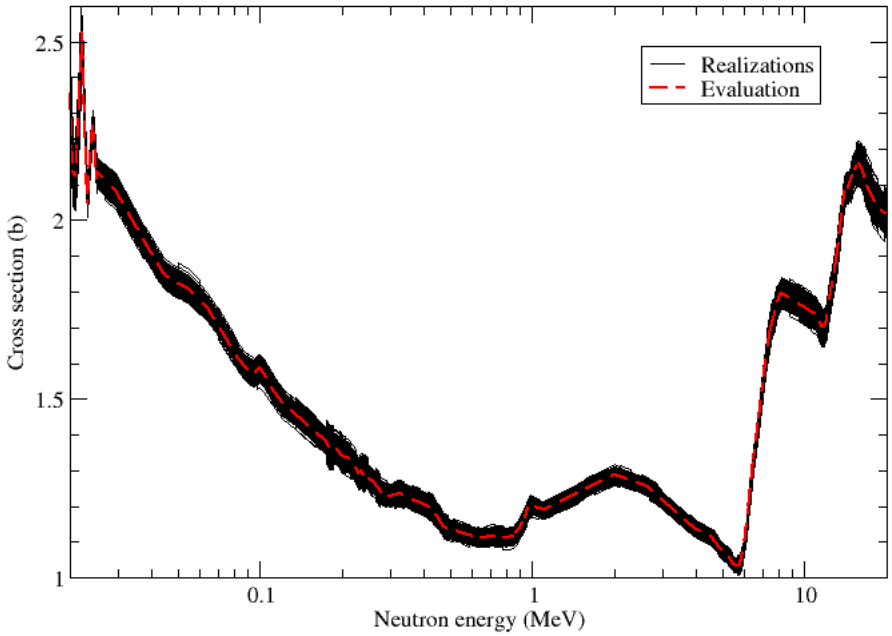

Figure 1: Plot of evaluated $^{235}\mathrm{U}\,(n,f)$ cross section from the ENDF2009.2 database (dashed line), and 3984 random realizations (solid lines) of this cross section generated using the Kiwi code [5, 6].

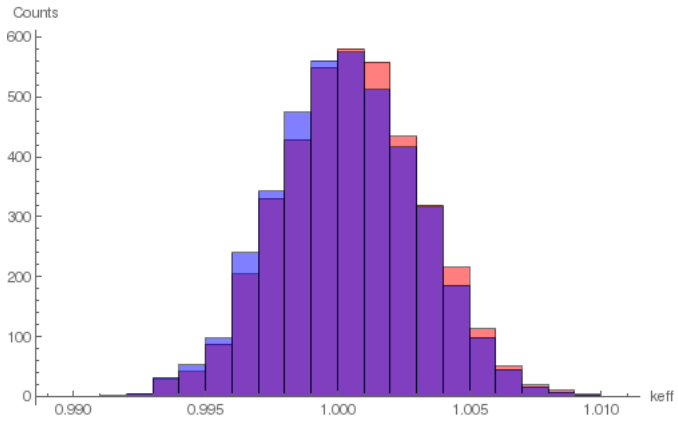

Figure 2: Histogram plots of the calculated k_{eff} values using the ARDRA (red bars) and Mercury (blue bars) codes. The area of overlap of the two histograms is shown in purple.

Requested (%)	Achieved (%)
50.0	50.000
68.0	67.903
80.0	80.063
90.0	90.032
95.0	94.978
99.0	99.036

Table 2: Requested and achieved confidence intervals for the data in Fig. 2.

coverage frequency achieved, and shows the excellent agreement between the two. In this example, it was possible to check the validity of the CIs by analyzing a large number of small samples. In other cases, the computational cost of evaluating the model may be too high and we may only be able to obtain a small number of sampled points. Table 2 shows that reliable CIs can be constructed even with only 5 sampled points in this example. Of course, the larger the sample size, the tighter the confidence interval will be, both because of the $1/\sqrt{n}$ factor in Eq. (4) and the decreasing behavior of the $t_{\alpha/2}$ values with increasing n, as can be seen in Table 4.

Finally, we noted in the discussion of Table 1 that the sample mean and standard deviation for the ARDRA and Mercury calculations were both already very close to the experimental values, however, we can improve this agreement by using the resampling technique discussed in appendix B. By resampling 100000 selections from the ARDRA results for example, we obtain a new sample mean of 0.999994 and a new sample standard deviation of 0.002578, which are both closer to the experimental values than the original values quoted in Table 1. Note that the size of the resampled data set is much larger than the size of the

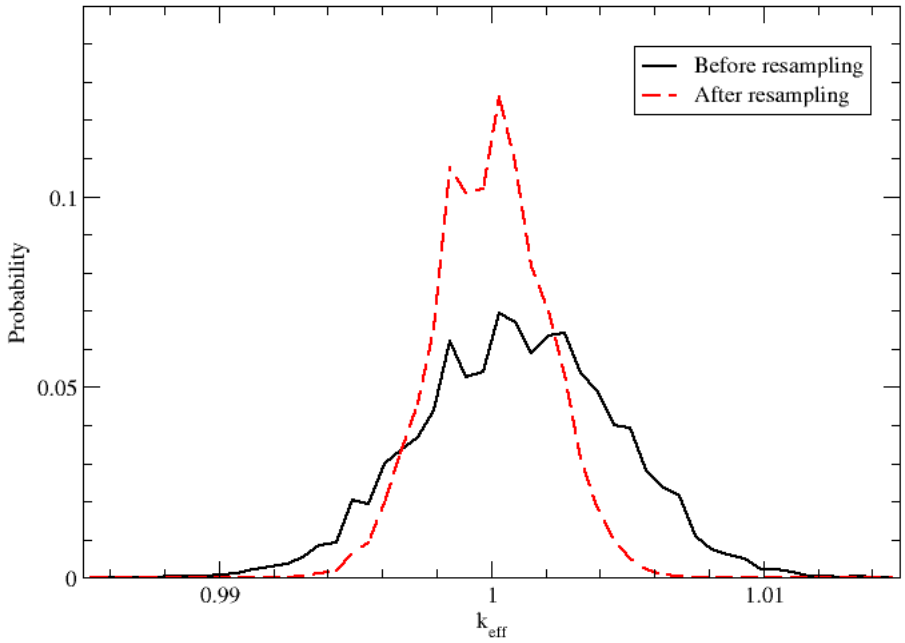

Figure 3: Probability distribution for k_{eff} obtained from ARDRA calculations with 4000 random realizations of the $^{239}\text{Pu}\,(n, f)$ cross section (solid black curve). The probability distribution obtained are resampling 10^5 in order to bring the k_{eff} mean value and standard deviation into agreement with experimental measurements is shown as a dashed red line.

initial sample (3984), and therefore some cross-sections and their corresponding k_{eff} value were sampled multiple times.

A similar application with ARDRA to the ^{239}Pu Jezebel critical assembly with 4000 realizations of the $^{239}\text{Pu}\,(n, f)$ cross section produces a mean value of 1.0010 with standard deviation of 0.0036, compared to the measured value 1.000 ± 0.002 [13]. Resampling 10^5 cross sections, we can bring the ARDRA result to a mean of 1.0000 and a standard deviation of 0.0020, in substantially better agreement with the experimental values. Figure 3 shows the effect of resampling the Pu cross section data on the probability distribution of k_{eff}. Although the mean value is not substantially affected by the resampling procedure, since it was already very close to the experimental one, the width of the distribution is noticeably narrower after resampling.

4 Conclusion

In this paper we have discussed several statistical tools that can be used together to quantify a level of confidence in predictions from a model that relies on uncertain input data, provided the uncertainties and correlations of the input data are known. The main techniques discussed are 1) random realizations of the input data, 2) the construction of confidence intervals and their interpretation, and 3) resampling techniques to compensate for inadequacies in the input data values and uncertainties. These statistical tools have been illustrated using a calculation of k_{eff}.

5 Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Security, LLC, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

A Calculation of confidence intervals for the mean of a distribution

Given a random sample of n numbers $\{x_1, x_2, \ldots, x_n\}$ drawn from a probability distribution with unknown mean μ and a standard deviation σ (which may or may not be known), we can estimate μ from the sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5}$$

In many applications, the sample standard deviation is also calculated,

$$s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{6}$$

and used as an estimate for σ, if it is unknown. Then the quantity $\bar{x} \pm s/\sqrt{n}$ is quoted as the estimate of the mean and its uncertainty. What this statement implies is that if repeatedly drew samples $\{x_1, x_2, \ldots, x_n\}$, and constructed the interval

$$\left[\bar{x} - \frac{s}{\sqrt{n}}, \bar{x} + \frac{s}{\sqrt{n}}\right] \tag{7}$$

each time, we would expect $\approx 68.27\%$ of those intervals to contain the true population mean μ. The interval in Eq. (7) is said to be a 68.27% confidence interval on the mean of the distribution. We can generalize this result and construct intervals for any desired level of confidence and for any underlying probability distribution.

A.1 Large sample size

If the sample size is sufficiently large (typically, $n \geq 30$ is considered large [3]) then, regardless of the form of the population distribution, the central limit theorem [4] ensures that the sample mean values $\bar{x}$ from repeated draws are approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. In that case, the quantity

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{8}$$

will have a normal distribution with mean 0 and standard deviation 1 (i.e., the standard normal distribution). The probability that z lies in some interval $[-c, c]$ is then given by the integral

$$\begin{aligned} P\left(-c \leq z \leq c\right) &= \int_{-c}^{c} d\xi \, \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} \\ &= \mathrm{erf}\left(\frac{c}{\sqrt{2}}\right) \end{aligned} \tag{9}$$

where erf is the error function. Constructing a confidence interval amounts to solving Eq. (9) for c, given a probability $P\left(-c \leq z \leq c\right)$. By convention, this probability equation is written as

$$P\left(-z_{\alpha/2} \leq z \leq z_{\alpha/2}\right) = 1 - \alpha \tag{10}$$

where the quantity $100 \times (1 - \alpha)$ is the confidence level expressed as a percentage (e.g., $\alpha = 0.05$ corresponds to a 95% confidence level). The number $z_{\alpha/2}$ is the value of z that leaves an area of $\alpha/2$ for $z \geq z_{\alpha/2}$ under the standard normal distribution curve. Likewise, $-z_{\alpha/2}$ is the value of z that leaves an area of $\alpha/2$ for $z \leq -z_{\alpha/2}$ under the standard normal distribution curve. From Eq. (9), this means we want

$$z_{\alpha/2} = \sqrt{2}\mathrm{erf}^{-1}\left(1 - \alpha\right) \tag{11}$$

The $z_{\alpha/2}$ value can be obtained by calculating Eq. (11) (Mathematica provides an InverseErf function for example), or from published tables (see, e.g., [3]). A few values of $z_{\alpha/2}$ are listed in Table 3 for different confidence levels. As we mentioned in the introduction to this appendix, the 68.27% confidence interval is just a special case where $z_{\alpha/2} = 1$.

Recalling the definition of z from Eq. (8), we therefore have

$$P\left(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha \tag{12}$$

or,

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{13}$$

which defines the $100 \times (1 - \alpha)\,\%$ confidence interval for the mean:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{14}$$

Confidence	α	$z_{\alpha/2}$
50%	0.5	0.67449
68.27%	0.3173	1
80%	0.2	1.28155
90%	0.1	1.64485
95%	0.05	1.95996
99%	0.001	2.57583

Table 3: Table of $z_{\alpha/2}$ values from Eq. (11) for various confidence levels.

A.2 Small sample size

The problem of calculating a confidence interval is more complicated when the sample size is small and the standard deviation σ is unknown. Perhaps surprisingly, it is still possible to construct a meaningful confidence interval, even with a single sampled value (provided something is known about the underlying probability distribution) [10, 11].

A.2.1 Normal population distribution

If the population distribution is normal (but with μ and σ not necessarily known), then the quantity

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{15}$$

which differs from Eq. (8) by the use of the sample standard deviation (Eq. (6)) in the denominator, is known to follow a Student t distribution with $\nu = n - 1$ degrees of freedom [3], which has the probability distribution function (pdf)

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \tag{16}$$

The small difference between Eqs. (8) and (15) is crucial: the sample standard deviation s is not a constant, but a random variable with its own pdf. In fact, for a sample of size n, the ratio of the sample variance to the population variance, s^2/σ^2, is known to have a $\chi^2_{n-1}/(n-1)$ distribution with $\nu = n - 1$ degrees of freedom, mean 1, and variance $2/(n-1)$ [3, 4]. In light of this fact, we can re-write Eq. (15) as

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \times \frac{1}{\sqrt{s^2/\sigma^2}} \\ &= \frac{z}{\sqrt{s^2/\sigma^2}} \end{aligned} \tag{17}$$

where z in the last line matches the definition in Eq. (8) and therefore has a standard normal distribution. The form of the t distribution in Eq. (16) then follows from the known distributions of the numerator and of the denominator in Eq. (17) [3, 4].

Confidence	α	$t_{\alpha/2}$ for $n = 5$	$t_{\alpha/2}$ for $n = 10$	$t_{\alpha/2}$ for $n = 100$
50%	0.5	0.740697	0.702722	0.676976
68.27%	0.3173	1.14165	1.05875	1.0051
80%	0.2	1.53321	1.38303	1.29016
90%	0.1	2.13185	1.83311	1.66039
95%	0.05	2.77645	2.26216	1.98422
99%	0.01	4.60409	3.24984	2.62641

Table 4: Table of $t_{\alpha/2}$ values from Eq. (23) for various confidence levels and different sample sizes n (and corresponding degrees of freedom $\nu = n - 1$).

As we did in Eq. (10), we now seek the value $t_{\alpha/2}$ for a given α such that

$$P\left(-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right) = 1 - \alpha \tag{18}$$

It is useful at this stage to introduce the cumulative distribution function (cdf),

$$F(t) \equiv \int_{-\infty}^{t} dt' \, f(t') \tag{19}$$

Then, Eq. (18) becomes

$$F\left(t_{\alpha/2}\right) - F\left(-t_{\alpha/2}\right) = 1 - \alpha \tag{20}$$

Because $f(t)$ in Eq. (16) is an even function of t, it follows from Eq. (19) that

$$F\left(-t_{\alpha/2}\right) = 1 - F\left(t_{\alpha/2}\right) \tag{21}$$

and therefore, combining Eqs. (20) and (21), the equation we need to solve is

$$F\left(t_{\alpha/2}\right) = 1 - \frac{\alpha}{2} \tag{22}$$

In principle we can invert the cdf for the t distribution numerically to obtain $t_{\alpha/2}$ as

$$t_{\alpha/2} = F^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{23}$$

We can also use the Mathematica function InverseCDF[StudentTDistribution[nu] , x], which returns $F^{-1}(x)$ for ν degrees of freedom. Table 4 gives the $t_{\alpha/2}$ values for different confidence levels and different sample sizes. Note that, as the sample size increases, the $t_{\alpha/2}$ values tend to the $z_{\alpha/2}$ values in Table 3.

A.2.2 Arbitrary population distribution

The t pdf in Eq. (16) gives the distribution of the sampled values $\{x_1, x_2, \ldots, x_n\}$ when they are drawn from a normal population distribution with unknown σ.

Suppose now that the sample is drawn from a distribution which is not necessarily normal (e.g., a uniform or a Cauchy distribution). In that case, we can no longer use the t distribution and the $t_{\alpha/2}$ values from Table 4. We must instead calculate equivalent $t_{\alpha/2}$ values that correspond to the actual probability distribution (e.g., uniform, Cauchy, Poisson, etc.). This calculation is also discussed in [10]. As in Eqs. (10) and (18), we need to solve the probability equation

$$P\left(-u_{\alpha/2} \leq u \leq u_{\alpha/2}\right) = 1 - \alpha \tag{24}$$

for the unknown $u_{\alpha/2}$ value where

$$u \equiv \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{25}$$

does not necessarily follow the t distribution. If we think of the set of sampled values $\{x_1, x_2, \ldots, x_n\}$ as the coordinates of a point in n-dimensional space, then the inequality on the left-hand side of Eq. (24),

$$-u_{\alpha/2} \leq u \leq u_{\alpha/2} \tag{26}$$

defines a region $\mathcal{R}$ in that space. If we know the population pdf $f(x)$, then Eq. (24) can be written as

$$\int \ldots \int_{\mathcal{R}} dx_1 \ldots dx_n \, f(x_1) \ldots f(x_n) = 1 - \alpha \tag{27}$$

Then, in principle, for a given value of $u_{\alpha/2}$, the region $\mathcal{R}$ is defined and the multiple integrals in Eq. (27) can be performed numerically giving a value for $1 - \alpha$. In this way, a table of $u_{\alpha/2}$ values and corresponding $100 \times (1 - \alpha)\,\%$ confidence levels can be constructed for any pdf $f(x)$, even if $f(x)$ does not have an analytical form (e.g., if it is given by a table of numerical values). A numerical algorithm to construct the $u_{\alpha/2}$ values is presented in [10].

B Importance resampling

Suppose we have a set of sampled values $\{x_1, x_2, \ldots, x_n\}$ drawn from a population pdf $f(x)$, and we wish to sample values $\{y_1, y_2, \ldots, y_n\}$ which satisfy a different pdf $g(x)$. For example, the x_i may represent k_{eff} values each one calculated from a random realization of an (n, f) cross section. Suppose we find that the sampled k_{eff} values have a standard deviation which differs from the experimental uncertainty, and we wish to enforce agreement with the experimental k_{eff} uncertainty. This particular example is discussed in section 3. One way to proceed is to resample the cross-section realizations with weights so that the corresponding k_{eff} values have a normal distribution with the desired standard deviation. The resampled cross sections form a set that is now consistent with the measured k_{eff} value and its uncertainty. This technique is known as importance resampling [14].

We start by defining the weights

$$w(x) \equiv \frac{g(x)}{f(x)} \tag{28}$$

for each sampled value and form the set $\{w(x_1), w(x_2), \ldots, w(x_n)\}$. We then calculate the sum of the weights

$$W = \sum_{i=1}^{n} w(x_i) \tag{29}$$

Next, we generate a pseudo-random number between 0 and W sampled from a uniform pdf,

$$r = U(0, W) \tag{30}$$

and we step through each x_i value incrementing the running sum

$$S_i = \sum_{j=1}^{i} x_i \tag{31}$$

The first x_i for which $S_i + w(x_i) \geq r$ is then selected and stored in the set of resampled values. This process is repeated until a sufficiently large set $\{y_1, y_2, \ldots, y_m\}$ of m resampled values has been constructed. Figure 4 shows an example of a standard normal distribution (mean $\mu = 0$, standard deviation $\sigma = 1$) resampled into a target normal distribution with mean 0.5 and standard deviation 1.1. For the original 10^6 sample points, we calculate a sample mean ($\bar{x}$) and standard deviation (s_x) of

$$\begin{aligned} \bar{x} &= 0.00040 \\ s_x &= 1.00175 \end{aligned} \tag{32}$$

while for the 10^6 points in the resampled distribution we have

$$\begin{aligned} \bar{y} &= 0.49962 \\ s_y &= 1.10044 \end{aligned} \tag{33}$$

as expected.

An important point to keep in mind is that, in general, the greater the difference between the original and desired pdfs, the larger the sample size will need to be. For example, if the target pdf $g(x)$ is a normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 2$, with 10^6 points in the original and resampled data we find

$$\begin{aligned} \bar{y} &= 0.81736 \\ s_y &= 1.81640 \end{aligned} \tag{34}$$

which are not in good agreement with the population mean and standard deviation.

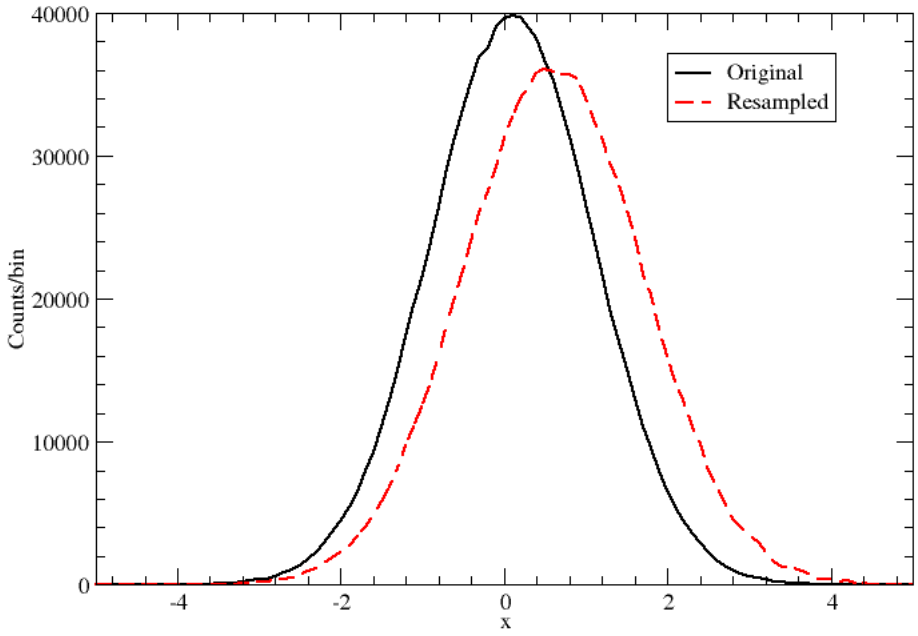

Figure 4: Histogram of points sampled from a normal distribution with mean 0 and standard deviation 1 (black solid curve), compared to a histogram of resampled data with a target normal distribution with mean 0.5 and standard deviation 1.1. Both the original and resampled distributions contain 10^6 sample points.

References

[1] M. Herman, M. T. Pigni, P. Oblozinsky, S. F. Mughabghab, C. M. Mattoon, R. Capote, Young-Sik Cho, A. Trkov, "Development of covariance capabilities in EMPIRE code", Tech. Rep. BNL-81624-2008-CP (2008).

[2] W. Younes, C. Mattoon, B. Beck, technical report in preparation.

[3] R. E. Walpole and R. H. Myers, "Probability and Statistics for Engineers and Scientists", 3rd ed., McMillan Publishing Company, New York (1985).

[4] S. Ross, "A first course in probability theory", 8th ed., Prentice Hall (2010).

[5] J. Pruet, "Kiwi: An Evaluated Library of Uncertainties in Nuclear Data and Package for Nuclear Sensitivity Studies", LLNL Tech. Rep. UCRL-TR_233400 (2007).

[6] C. M. Mattoon, D. Brown, and J. B. Elliott, EPJ Web of Conferences 27, 00002 (2012).

[7] P. N. Brown, "ARDRA, scalable parallel code system to perform neutron and radiation transport calculations" LLNL Tech. Rep. UCRL-TB-132078-REV-1 (1999).

[8] A. J. Kunen, "DOE HPCOR - ARDRA Project Review", LLNL Tech. Rep. LLNL-CONF-676935 (2015).

[9] P. S. Brantley, S. A. Dawson, M. S. McKinley, M. J. O'Brien, D. E. Stevens, "Recent Advances in the Mercury Monte Carlo Particle Transport Code", LLNL Tech. Rep. LLNL-PROC-609976 (2013).

[10] N. M. Blachman and R. E. Machol, IEEE Trans. Info. Theory 33, 373 (1987).

[11] G. J. Feldman and R. D. Cousins, Phys. Rev. D 57, 3873 (1998).

[12] R. D. Mosteller and J. M. Goda, Intl. Conf. Adv. Math., Saratoga Springs, NY (2009); LANL Tech. Rep. LA-UR-09-01007 (2009).

[13] J. A. Favorite and R. W. Brewer, NEA/NSC/DOC(95)03/I, (2013).

[14] A. F. M. Smith and A. E. Gelfand, "Bayesian Statistics without Tears: A Sampling-Resampling Perspective", Am. Stat. 46, 84 (1992).