# Overview of the Tolerance Limit Calculations with Application to TSURFER

Hany Abdel-Khalik [1,*], Dongli Huang [1,*], Ugur Mertyurek [2], William Marshall [2] and William Wieselquist [2]

1 School of Nuclear Engineering, Purdue University, West Lafayette, IN 47906, USA
2 Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA; mertyureku@ornl.gov (U.M.); marshallwj@ornl.gov (W.M.); wieselquiswa@ornl.gov (W.W.)
* Correspondence: abdelkhalik@purdue.edu (H.A.-K.); huang714@purdue.edu (D.H.)

**Abstract:** To establish confidence in the results of computerized physics models, a key regulatory requirement is to develop a scientifically defendable process. The methods employed for *confidence, characterization, and consolidation*, or $C^3$, are statistically involved and are often accessible only to avid statisticians. This manuscript serves as a pedagogical presentation of the $C^3$ process to all stakeholders—including researchers, industrial practitioners, and regulators—to impart an intuitive understanding of the key concepts and mathematical methods entailed by $C^3$. The primary focus is on calculation of tolerance limits, which is the overall goal of the $C^3$ process. Tolerance limits encode the confidence in the calculation results as communicated to the regulator. Understanding the $C^3$ process is especially critical today, as the nuclear industry is considering more innovative ways to assess new technologies, including new reactor and fuel concepts, via an integrated approach that optimally combines modeling and simulation and minimal targeted validation experiments. This manuscript employs intuitive, analytical, numerical, and visual representations to explain how tolerance limits may be calculated for a wide range of configurations, and it also describes how their values may be interpreted. Various verification tests have been developed to test the calculated tolerance limits and to help delineate their values. The manuscript demonstrates the calculation of tolerance limits for TSURFER, a computer code developed by the Oak Ridge National Laboratory for criticality safety applications. The goal is to evaluate the tolerance limit for TSURFER-determined criticality biases to support the determination of upper, subcritical limits for regulatory purposes.

**Keywords:** statistical tolerance limits; Bayesian inference; aleatory and epistemic uncertainties

## 1. Introduction and Mathematical Background

To establish confidence in the results of computerized physics models, a key regulatory requirement is to develop a scientifically defendable process, known as *model validation*. To ensure that code predictions can be trusted for a given application, such as the domain envisaged for code use, the regulatory process requires the consolidation of two independent sources of knowledge: (1) measurements collected from experimental conditions similar to those of the application, and (2) code predictions that model the same experimental conditions. This is equivalent to the idea of cross-referencing knowledge obtained from independent sources. The idea is that the consolidated knowledge will provide a higher level of confidence than the knowledge obtained from a single source. Drawing upon an example from criticality safety, which is the focus of this work, the analyst is interested in predicting a response, such as $k_{eff}$, for an arrangement of fuel assemblies in a shipping cask, serving as a representative application. The two sources of knowledge are (1) prior code predictions and (2) measurements for a number of critical experiments containing fuel similar in composition, geometry, and arrangement to that of the application. The premise is that, while both the code-calculated and experimentally measured responses have unavoidable uncertainties, the consolidation process could produce a better estimate of $k_{eff}$ with a higher level of confidence than an estimate obtained with the experiments or code predictions

alone. This process also demonstrates the applicability of the computational method to real-world measurements. Validation to real-world measurements is required by consensus standards and regulation promulgated by the US Nuclear Regulatory Commission (NRC).

This approach represents the core principle behind model validation, and it is designed to establish confidence in code predictions. Details on how the confidence in each knowledge source is characterized and how the knowledge consolidation process works are not prescribed by the regulators. Instead, the regulation process mandates a minimum level of confidence that the licensee must demonstrate in their code predictions, such as the 95%/95% criterion often employed by nuclear practitioners. This leaves enough freedom to the licensees to design their own technical basis for *confidence, characterization, and consolidation*, or $C^3$.

Many statistical methods have been developed over the years to describe the $C^3$ process: see for example Fisher 1950 [1], Hotel 1971 [2], Jaynes 1976 [3], Samaniego 2010 [4], Stanton 2017 [5], Bolstad 2017 [6], Bishop 2006 [7], Wakefield 2013 [8], and Thornber 1965 [9]. The theory has appeared in many textbooks and is presented from two distinct, heavily debated viewpoints—the frequentist and Bayesian statistics—with various renditions, such as parametric, nonparametric, and order-statistics. Despite the striking differences between the two viewpoints and their associated analysis methods, the final results could look very similar, making it very difficult for practitioners to see the value of all the theoretical subtleties. This work maintains that this classical presentation diminishes the value to engineering practitioners, who need an intuitive understanding of the how the $C^3$ process works, thus allowing for a transparent approach through which the values of different methods may be assessed. This manuscript presents an intuitive exposition of the $C^3$ process and its relevance to criticality safety problems. The presentation is supported by basic mathematical/statistical principles which are accessible to most engineering practitioners, as well as graphical aids delineating the process by which confidence can be established.

This presentation supports the development of a tolerance limit for bias calculations performed by ORNL's TSURFER code, which is used for criticality safety [10]. Currently, TSURFER only calculates bias and bias uncertainty. The objective of this work is to evaluate the tolerance limit for the calculated bias to support determination of upper subcritical limits for regulatory purposes. In support of criticality safety analysis licensing, the regulator requires development of a confidence characterization and consolidation process ($C^3$) when calculating quantities of interest, such as $k_{eff}$, for a given application condition. The consolidation process is mandated to develop confidence collected from two independent sources: physics-based calculations and experimental measurements. In practice, it is not feasible to experiment with every application condition, so a method is needed that can leverage similar experiments to complete the $C^3$ process. The TSURFER methodology attempts to achieve this goal by consolidating experimental measurements and calculational results from a number of experiments that are judged to be sufficiently similar to the application; this is accomplished by using a Bayesian-based generalized least-squares methodology [10]. The measurements and calculations for the experimental conditions are consolidated by identifying a common source of uncertainty across all of the experiments and the target application. In the TSURFER methodology, this common source is the nuclear cross sections, which are adjusted by TSURFER to improve the agreement between calculations and measurements. The adjusted cross sections serve as a basis for updating the value of $k_{eff}$ for the application (while the generalized least-squares methodology is based on the concept of cross section adjustments, TSURFER is able to directly calculate the impact on the response of interest in the form of calculational biases without having to store the adjusted cross sections.) The difference between the prior and adjusted $k_{eff}$ values is referred to as a *bias*.

To establish confidence in the calculated bias, TSURFER also calculates the standard deviation of the bias assuming the bias uncertainty is described by a normal distribution. In this effort, TSURFER was augmented with another measure of confidence, denoted by

the tolerance limit [11,12], an upper threshold that is a function of the mean value of the bias and its standard deviation in the form $\mu_t + K\sigma_t$. The $K$ is a coverage parameter calculated to ensure that there is a high probability that the true bias remains below the upper threshold value. This supports the calculation of the upper subcritical limit (USL) [13]. The regulator enforces the use of USL to ensure, with high confidence, that the true value of $k_{eff}$ for the given application remains upper-bounded. To support this goal, TSURFER must be able to calculate an upper threshold limit for the bias, also with high confidence. Typically, US NRC requires that there is 95% confidence that at least 95% of the population of bias values are covered.

Demonstration of this specific requirement has been an obstacle to implementation of the TSURFER methodology in the United States since its initial release. Since TSURFER represents a specific rendition of the $C^3$ process, it is important to first develop a general understanding of the value, interpretation, and calculational methods for tolerance intervals. This is important to establish before navigating through the various methods appearing in the literature, so that tolerance calculations from both the frequentist and Bayesian viewpoints can be managed. This manuscript is not intended to support or criticize the TSURFER methodology; instead, it provides clear guidance to practitioners on how to select a specific tolerance calculational method, as demonstrated herein, for the TSURFER methodology.

The $C^3$ process does not provide any details on the technical method used to mathematically characterize confidence nor does it provide instructions on how confidence can be algorithmically consolidated from multiple sources. To make such decisions, one must first determine the type of uncertainties present in the problem. The presence of uncertainty implies that one cannot make a deterministic statement about the outcome of an experiment: that is, one cannot say with 100% certainty that $k_{eff}$ is equal to 1.0. This results from two distinct types of uncertainty. The first type of uncertainty results from the inability to control every aspect of an experiment. For example, one cannot exactly know, with 100% confidence, the geometry, the composition, and other experimental conditions. If employing a code, the same situation occurs wherein the lack of knowledge—regarding the physics model itself and/or its associated input parameters—contributes to the overall uncertainty. The second type of uncertainty results from the physical nature of the quantity being measured, which may not assume a single value due to its inherent randomness. The first type of uncertainty is referred to as *epistemic* or *Bayesian uncertainty*, and the second type is associated with more conventional frequency-based interpretation of uncertainty, referred to as *aleatory uncertainty*. Both these types of uncertainties are mathematically described using probability density functions (PDFs), albeit with different interpretations and different consolidation approaches. For example, as explained below, aleatory uncertainties are consolidated via averaging of their PDFs, whereas epistemic uncertainties, interpreted as degree-of-belief, are consolidated via PDFs multiplication.

Despite the striking differences in the interpretations of these two types of uncertainties and their markedly different consolidation approaches, their final results, including confidence intervals, credible intervals, or tolerance intervals, often look very similar, with only subtle variations that are heavily debated by statisticians, while seeming mostly of unclear value to practitioners [3]. Statements like the following are typically made about a response such as $k_{eff}$ for a given application, for example, a shipping cask containing nuclear fuel: With 95% confidence, the probability that $k_{eff}$ will not exceed the stated threshold is 95%. Ideally, a straightforward interpretation of this statement would read as follows: there is 95% chance that the true value of $k_{eff}$ will not exceed the stated threshold. However, this simple interpretation is not 100% accurate; it is only 95% accurate. In layman terms, this is equivalent to stating: *You certainly need to get a college education to succeed in life, but not necessarily*, or *with no surprises, the surgery will be 99% safe*. Intuitive interpretation of these statements is troubling because they provide conflicting, ambiguous, probabilistic information. On the one hand, they make a positive probabilistic assertion for the outcome, but on the other hand, they hedge, thus reducing the confidence in that assurance. In other

words, they hedge twice, once by employing a probabilistic rather than deterministic assertion to describe the outcome, and then they hedge a second time against that very assertion. It is natural to ask, why cannot one hedge only once, resulting in a much simpler/intuitive interpretation? The goal of this work is to explain that this double-hedging strategy is a result of the theoretical methods employed to describe confidence. While this strategy affords mathematical convenience, it entangles the meanings of aleatory and epistemic uncertainty. However, simple interpretation can still be achieved using basic arguments and graphical aids, which is one of the key contributions of this work. Another goal is to explain how both epistemic and aleatory uncertainties are essential in establishing confidence. In many textbooks, the two viewpoints are discussed separately, as independent tracks in statistics [6]. However, all practical problems contain a mix of epistemic and aleatory uncertainties, necessitating a flexible approach that can transition seamlessly between the two viewpoints.

Methods used to determine the tolerance limits (tolerance intervals) are based on order statistics as defined by Wilks [14,15], in which the distribution of the proportion of the population between two-order statistics is independent of the sample population. Wald [16] extends Wilks' method and constructs the tolerance limits for normally distributed data. Somerville tabularized the tolerance limits for nonparametric tolerance limits to minimize the interpolation, and Patel [17] discusses the tolerance limits for various distributions in both parametric and nonparametric methods. In nuclear engineering, tolerance limits attracted attention because the US NRC's licensing process switched to best-estimate methods plus uncertainty (BEPU) analyses in 1988 [18]. Tolerance limits were introduced to nuclear systems by Pál [19] and Guba [20] for the application of safety analyses. They have been incorporated within the framework of standard uncertainty analysis methods like the GRS method [21], thus serving as measures of confidence. Many papers in nuclear literature apply tolerance limits and propose suitable data treatment in their specific studies. For example, nonparametric tolerance limits are applied on quantities like fission gas release for spent fuel assembly [22]. One- and two-sided tolerance limits are used to improve the precision of measurement uncertainty for engineering systems and are applied on a contaminant transport model [23]. The distribution-free Wilks tolerance limit is used to evaluate the nuclear data uncertainty impact on power and reactivity in rod-withdrawal transients for pressurized water reactor (PWR) mini-cores when the output is time dependent and its distribution varies across time [24]. Porter [25] provides a comprehensive discussion on the nonparametric tolerance limits, including derivations, statistical properties, application to computational tools, requirements imposed on the applied code, verification process, and higher-rank Wilks formula. The tolerance limits are tested with various orders to obtain recommendations on safety audit analysis for energy systems [26]. The Wilks formula is used to evaluate uncertainty quantification (UQ) results on a TRACE boiling water reactor (BWR) model to simulate a spray cooling experiment [27]. The upper tolerance limits of parametric and nonparametric methods under mixture of normal distributions for nuclear power plant safety (e.g., peak cladding temperature in accident scenario) are derived and calculated [28]. Wilks' method was compared to the Monte Carlo method for performing UQ on loss-of-coolant-accident (LOCA) application [29] and was found to be conservative and flexible at the 95/95 tolerance. Shockling [30] evaluated the number of thermal-hydraulics calculations required to meet the safety limits using nonparametric order statistics on peak cladding temperature. Suggestions on the order and number of random samples needed in the use of Wilks' formula for more appropriate nuclear safety applications are provided in the Nuclear Energy Agency Committee on the Safety of Nuclear Installations workshop proceedings from 2013 [31] and based on the authors' in-depth understanding. The appropriate quantitative "high level of probability", from a regulatory point of view, are explained in Martin and Nutt's article from 2011 [32], in which different assumptions and limitations of acceptance criteria are reviewed, and recommendations for LOCA applications regarding univariate and multivariate situations for UQ with order statistics are provided.

To set the stage for the $C^3$ process, the reader is strongly recommended to visit Appendix A before proceeding to the next section. Appendix A starts with a few preliminary definitions on two key concepts in probability theory—the frequency of occurrence and the degree of belief—both of which are dubbed as probability and described by PDFs. The goal is to show that despite the differences, which arguably appear to be purely philosophical to practitioners, the combined use of Bayesian and frequentist statistics is essential to establish/consolidate confidence in most practical problems. The presentation in Appendix A is meant to be pedagogical, to allow for a transparent dialogue about the effectiveness and value of various $C^3$ methods. For experienced statisticians, Appendix A may be skipped. For other readers, it is strongly recommended as it describes in detail the basic principles of the $C^3$ process which is essential before proceeding to the TSURFER application in the next section.

## 2. TSURFER Tolerance Limit Development

This section demonstrates how the upper threshold for a 95%/95% tolerance interval may be calculated for the TSURFER methodology. The goal is to ensure that the TSURFER bias will not result in a $k_{eff}$ value that exceeds the upper subcritical limit, as illustrated in Figure 1.
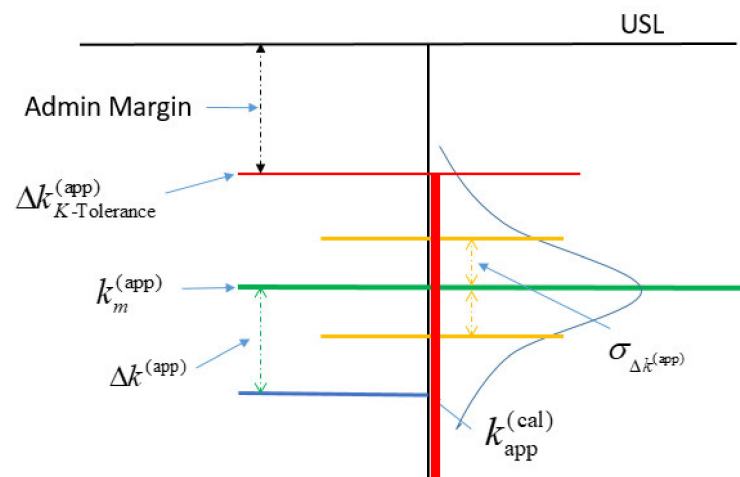


**Figure 1.** Relationship of USL and TSURFER Bias Tolerance Limit.

The reference value for $k_{eff}$ prior to the application of TSURFER is marked by the blue horizontal line $k_{app}^{(cal)}$. The bias $k^{(app)}$ calculated by TSURFER is marked by the double-pointing arrow, bringing the best-estimate of $k_{eff}$ to the green line. Since the TSURFER methodology is statistical, the confidence in the bias estimate is described by an epistemic PDF, as marked by the blue curve. This PDF is assumed to be normal in the TSURFER methodology. The standard deviation $\sigma_{\Delta k^{(app)}}$ of this PDF is calculated by TSURFER, as marked by the two double-pointing orange arrows around the best-estimate adjusted value. The upper threshold of the tolerance interval is marked by the red line, so the area underneath the PDF is at least 0.95, implying with 95% confidence that the true value for the bias will ensure a $k_{eff}$ value below the upper threshold limit. The interval upper-limited by the red line (marked by a solid red bar) is the tolerance interval for TSURFER's bias. Finally, an additional administrative margin is added to reach the USL [13].

### 2.1. TSURFER's Problem Definition

With the TSURFER theory well-documented [10], a simplified rendition is presented below, consistent with the mathematical treatment provided in Appendix A. Effectively, TSURFER applies a customized rendition of the $C^3$ process to update knowledge about the

target application for which no experimental data exist using two sources of knowledge—the application's model-predicted $k_{eff}$, given by a prior PDF of the form:

$$p_{\text{app}}^{(\text{cal})}\left(k_t^{(\text{app})}\middle|k_{\text{app}}^{(\text{cal})}\right) \propto e^{-\frac{\left(k_{\text{app}}^{(\text{cal})}-k_t^{(\text{app})}\right)^2}{2\sigma_t^{(\text{app,cal})2}}}, \tag{1}$$

and the measurements from $N$ experiments,

$$p_i^{(\text{exp})}\left(k_t^{(i)}\middle|k_i^{(\text{exp})}\right) \propto e^{-\frac{\left(k_i^{(\text{exp})}-k_t^{(i)}\right)^2}{2\sigma_t^{(\text{exp}_i)2}}}, \; i = 1, 2, \ldots, N, \tag{2}$$

consolidated with their associated model-predicted values,

$$p_i^{(\text{cal})}\left(k_t^{(i)}\middle|k_i^{(\text{cal})}\right) \propto e^{-\frac{\left(k_i^{(\text{cal})}-k_t^{(i)}\right)^2}{2\sigma_t^{(\text{exp}_i,\text{cal})2}}}, \; i = 1, 2, \ldots, N. \tag{3}$$

The distributions in Equations (1) and (3) are centered around the reference prior $k_{eff}$ values of $k_i^{(\text{cal})}|_{i=1}^N$ and $k_{\text{app}}^{(\text{cal})}$, which were calculated using the reference values for the cross sections, leaving the true values, $k_t^{(i)}|_{i=1}^N$ and $k_t^{(\text{app})}$, as unknown in the epistemic sense. Notwithstanding the aleatory nature of the interaction between a neutron and a target nucleus, the epistemic treatment is more adequate because the range of cross section variations that result from the random nature of the neutron-nucleus interaction is much smaller in magnitude than the uncertainties associated with the cross-section measurement and subsequent evaluation procedures. This situation is depicted in Figure A1 in the Appendix A, where the narrow PDF represents the true aleatory spread of the cross-section values, whereas the wider PDF represents the degree-of-belief about the mean values of the aleatory PDF.

All three sets of distributions in Equations (1)–(3) assume the normal shape because only the mean and standard deviations are employed to construct the PDFs. This approach was used based on a famous statistical result [33] stating that the least-informative PDF about a parameter is the normal distribution when the mean value and standard deviations are the only known features about the distribution. The true standard deviation of the calculated $k_{eff}$ value for the $N$ experiments—$\sigma_t^{(\text{exp}_i,\text{ cal})}|_{i=1}^N$ and the application, $\sigma_t^{(\text{app, cal})}$—are obtained by propagating the cross-sectional uncertainties from Evaluated Nuclear Data File (ENDF) libraries, which are available in the form of covariance information [34]. Arguably, the propagation process may result in non-normal distributions for the calculated $k_{eff}$ values. Assessing the adequacy of the normality assumption is outside the scope of this work. If indeed the prior distributions given in Equations (1) and (3) considerably deviate from the normal shape, then nonparametric methods are needed to determine the tolerance limits. Nonparametric methods do not make assumptions about the shape of the true PDF, so the shapes of the sampling distributions (e.g., normal distribution, $\chi^2$-distribution, $t$-distribution) cannot be evaluated analytically. Instead, sampling techniques that emulate the numerical results described in Appendix A.7 are performed to numerically integrate the resulting joint sampling distributions of the mean and standard deviations. However, this discussion is outside the scope of this study, which is limited to the assumptions made by TSURFER, whose justification may be found in earlier publications by TSURFER developers [10].

The measurement PDFs in Equation (2) are centered around the measured values, which are assumed to be a single measurement per experiment—$k_i^{(\text{exp})}|_{i=1}^N$ leaving the true $k_{eff}$ value as the unknown. The true standard deviations $\sigma_t^{(\text{exp}_i)}|_{i=1}^N$ are assumed to be known which greatly simplifies the calculation of the tolerance limit as will be shown later. The standard deviations are often based on conservative estimates of the various sources of

uncertainties that are believed to contaminate the experimental model, including composition, geometry, and experimental conditions, often referred collectively to as the benchmark uncertainties, where a benchmark denotes an experiment. This uncertainty can also be inflated to account for statistical uncertainties resulting from Monte Carlo simulation.

The mathematical procedure used to combine these uncertainties is selected by the experimentalist (the quadrature approach is often considered adequate to combine these uncertainties since they tend to be independent of each other; this discussion is, however, considered outside the scope of this work) which is typically unchallenged and may not be well documented. Essentially, it is left to the experimentalist to decide on the best approach to capture and combine the different types of uncertainties, often employing conservatism. When performing downstream analysis (e.g., TSURFER), the analyst employing these uncertainties must decide whether these uncertainties are aleatory or epistemic. TSURFER assumes that the reported experimental standard deviations are the true values and that the associated PDF is normal per Equation (2). Furthermore, it assumes that the PDF is epistemic and thus employs the concept of PDF multiplication for consolidation with the code-calculated PDFs shown in Equation (3). As shown below, the choice of epistemic treatment is straightforward for TSURFER because the true standard deviations for the consolidated PDFs are assumed to be known. With more explicit treatment of the uncertainties from the experimental procedure, the choice between aleatory and epistemic treatment would become more relevant.

### 2.2. TSURFER's $C^3$ Process

In the above formulation, the number of parameters for which knowledge is to be updated is equal to $N + 1$, representing the $k_{eff}$ values for the $N$ experiments $k_t^{(i)}|_{i=1}^N$ and the one application $k_t^{(\mathrm{app})}$. Each experiment has two sources of knowledge: one from the experimental measurements $k_i^{(\exp)}|_{i=1}^N$, and one from prior code calculations $k_i^{(\mathrm{cal})}|_{i=1}^N$, and, finally, the target application has a single knowledge source from code calculations $k_{\mathrm{app}}^{(\mathrm{cal})}|_{i=1}^N$. A straightforward application of the $C^3$ process, as outlined in Appendix A.4, only allows for updating knowledge about the true value for each experiment $k_t^{(i)}|_{i=1}^N$ independently of all other experiments, as obtained by consolidating the two respective PDFs for the $i$th experiment in Equations (2) and (3), as follows:

$$p_i^{(\mathrm{con})}\left(k_t^{(i)}\right) \propto e^{-\frac{\left(k_i^{(\exp)}-k_t^{(i)}\right)^2}{2\sigma_t^{(\exp_i)2}} - \frac{\left(k_i^{(\mathrm{cal})}-k_t^{(i)}\right)^2}{2\sigma_t^{(\exp_i,\mathrm{cal})2}}} \propto e^{-\frac{\left(k_i^{(\mathrm{con})}-k_t^{(i)}\right)^2}{2\sigma_t^{(\mathrm{con}_i)2}}},$$

where the respective mean values and the standard deviations of the consolidated PDFs are given by

$$k_i^{(\mathrm{con})} = \frac{\dfrac{k_i^{(\mathrm{cal})}}{\sigma_t^{2(\exp_i,\mathrm{cal})}} + \dfrac{k_i^{(\exp)}}{\sigma_t^{2(\exp_i)}}}{\dfrac{1}{\sigma_t^{2(\exp_i,\mathrm{cal})}} + \dfrac{1}{\sigma_t^{2(\exp_i)}}},$$

$$\frac{1}{\sigma_t^{(\mathrm{con}_i)}} = \frac{1}{\sigma_t^{2(\exp_i,\mathrm{cal})}} + \frac{1}{\sigma_t^{2(\exp_i)}},$$

where the superscript *con* refers to the consolidated values. This brute-force application of the $C^3$ process is not interesting, because it does not exploit the similarity between the experiments, and more importantly, it does not provide a way to update the application's PDF in Equation (1). If the PDFs in Equations (1)–(3) are multiplied directly, then $2N + 1$ would be unknown, and all PDFs would be treated as independent of each other, hence, providing no real value from consolidation. To overcome this, the unknown epistemic pa-

rameters must be common to all PDFs to realize the value of PDF multiplication. TSURFER achieves that by assuming the following:

$$k_t^{(\text{app})} = k_{\text{app}}^{(\text{cal})} + \vec{\theta}_{\text{app}}^T \vec{\delta}_t + \varepsilon, \tag{4}$$

$$k_t^{(i)} = k_i^{(\text{cal})} + \vec{\theta}_i^T \vec{\delta}_t + \varepsilon, \; i = 1, 2, \ldots, N. \tag{5}$$

These $N + 1$ equations imply that the true $k_{\text{eff}}$ values for the $N$ experiments and the one application all depend linearly on a common set of $n$ epistemic parameters represented by a vector $\vec{\delta}_t^T \in R^n = [\; \delta_{t1} \quad \delta_{t2} \quad \cdots \quad \delta_{tn} \;]$, where $t$ denotes the unknown true values. These epistemic parameters are selected in TSURFER to be the multigroup nuclear cross sections, and the coefficients of linearity, represented by the components of the vectors $\vec{\theta}_i$ and $\vec{\theta}_{\text{app}}$, are precalculated using a sensitivity (parametric) analysis, which estimates the first-order derivative of the calculated $k_{\text{eff}}$ with respect to each cross section.

It is important to note here that while these equations allow for application of the $C^3$ process, they are based on two key assumptions that require proper validation: (a) cross sections represent the key sources of the observed discrepancies between calculated and measured $k_{\text{eff}}$ values, and (b) the unexplained errors $\varepsilon$ have no structure and are purely aleatory, so they can be reduced with repeated measurements or many experiments. In reality, these unknown errors are likely a combination of epistemic and aleatory uncertainties, with the epistemic uncertainties resulting from other parameters not captured by the physics model, and with the aleatory uncertainties resulting from uncontrolled experimental conditions. With epistemic uncertainties present but unaccounted for, the error compensation phenomenon is likely to happen. Thus, analysis is required to ensure that these assumptions are adequate. However, this is outside the scope of the current work.

With a source of epistemic uncertainties that is common to all experiments, the $C^3$ process can consolidate all the PDFs in Equations (2) and (3) into a single updated PDF for the cross-sectional adjustments. Since the cross sections are described by a vector, the resulting consolidated PDF—$p\left( \vec{\delta}_t | k_i^{(\text{exp})} |_{i=1}^N, k_i^{(\text{cal})} |_{i=1}^N \right)$—is in the form of a multi-dimensional normal distribution,

$$p\left( \vec{\delta}_t | k_i^{(\text{exp})} |_{i=1}^N, k_i^{(\text{cal})} |_{i=1}^N \right) \propto e^{-\frac{1}{2}(\vec{\delta}_t - \vec{\delta}_m)^T \mathbf{C}_m^{(\delta)-1} (\vec{\delta}_t - \vec{\delta}_m)}. \tag{6}$$

Although it is more complicated in form than the simpler single-variable PDFs such as Equation (A15), it can be shown that this PDF is centered around a vector of mean values $\vec{\delta}_m$, and the spread (i.e., as measured by the variance or standard deviation) of the PDF along the $n$ dimensions is described by a squared $n \times n$ covariance matrix $\mathbf{C}_m^{(\delta)}$, both of which can be analytically determined. Expressions for $\vec{\delta}_m$ and $\mathbf{C}_m^{(\delta)}$ may be found in the TSURFER literature [10]. To support the pursuit for the tolerance interval, simpler expressions are presented here for $\vec{\delta}_m$ which may be written as:

$$\vec{\delta}_m = B \begin{bmatrix} k_1^{(\text{exp})} - k_1^{(\text{cal})} \\ k_2^{(\text{exp})} - k_2^{(\text{cal})} \\ \cdots \\ k_N^{(\text{exp})} - k_N^{(\text{cal})} \end{bmatrix} = B \left( \vec{k}^{(\text{exp})} - \vec{k}^{(\text{cal})} \right).$$

This equation emulates the TSURFER final expression for $\vec{\delta}_m$ by lumping all the mathematical operators into one matrix operator B, operating on a vector describing the discrepancies between the measured and reference calculated $k_{\text{eff}}$ values for the $N$ experi-

ments. Combing this equation with Equation (4), one can update the epistemic PDF for the application $k_{eff}$ value $k_t^{(\text{app})}$, which is also expected to be a normal distribution, with mean value given by

$$k_m^{(\text{app})} = k_{\text{app}}^{(\text{cal})} + \vec{\theta}_{\text{app}}^{T} \vec{\delta}_m + \varepsilon = k_{\text{app}}^{(\text{cal})} + \vec{\theta}_{\text{app}}^{T} B \left( \vec{k}^{(\text{exp})} - \vec{k}^{(\text{cal})} \right).$$

Rewriting this equation with the substitution $\vec{\lambda} = \vec{\theta}_{\text{app}}^{T} B$ results in

$$k_m^{(\text{app})} - k_{\text{app}}^{(\text{cal})} = \sum_{i=1}^{N} \lambda_i \left( k_i^{(\text{exp})} - k_i^{(\text{cal})} \right).$$

Since the cross-sections PDF in Equation (6) is normal, and $k_t^{(\text{app})}$ is assumed to depend linearly on $\vec{\delta}_t$ per Equation (4), the resulting PDF for $k_t^{(\text{app})}$ will also be normal, with mean value and variance that are analytically determined. In doing so, the standard deviation of the consolidated PDF is an analytic function of the true experimental standard deviation and the prior standard deviation, implying that it represents the true (not an estimate) standard deviation of the consolidated PDF for $k_t^{(\text{app})}$. This greatly simplifies the calculation of the tolerance limit as compared to the general case when both the mean and standard deviations are not known and have to be estimated from the consolidation analysis (see Appendix A.7). Rewriting the equation above in the form of deviations (i.e., biases) from the prior calculated values results in

$$\Delta k^{(\text{app})} = \sum_{i=1}^{N} \lambda_i \Delta k_i^{(\text{exp})} = \vec{\lambda}^{T} \Delta \vec{k}^{(\text{exp})}. \tag{7}$$

This equation describes a linear relationship between the initial experimental biases for the $N$ experiments and the TSURFER calculated bias for the application, representing the change in the location of the mean $k_{eff}$ value for the application. The goal is to establish a tolerance interval for the application bias. Since the standard deviations for both the experimental measurements and reference calculated values are known, the true standard deviation for the application bias can be calculated as

$$\sigma_{\Delta k^{(\text{app})}} = \sqrt{\sum_{i=1}^{N} \lambda_i^2 \left( \sigma_t^{(\text{exp}_i)2} + \sigma_t^{(\text{exp}_i,\text{cal})2} \right)}.$$

This equation strongly emulates the derivation of the sample standard deviation using the sample mean, as discussed at the beginning of Appendix A.7, specifically in Equation (A23). Unlike Equation (A23), the weights are not unity. By analogy, the $\Delta k_t^{(\text{app})}$ may be thought of as a feature derived from the $N$ PDFs representing the $N$ experiments, each contributing a single sample. The standard deviation of the estimated quantity $\Delta k_t^{(\text{app})}$ may be calculated by first projecting the samples along the unit vector:

$$\vec{e}_N = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_N \end{bmatrix}^{T} / \sqrt{\sum_{i=1}^{N} \lambda_i^2},$$

and calculating the residual $\mathrm{P}_e \Delta \vec{k}^{(\mathrm{exp})}$ and the estimated standard deviation as

$$s_{\Delta k^{(\mathrm{app})}} = \sqrt{\frac{\left\| \mathrm{P}_e \Delta \vec{k}^{(\mathrm{exp})} \right\|_2^2}{N-1}}. \tag{8}$$

Since the true standard deviation of the bias can be calculated analytically, the above quantity may simply be used for verification purposes by checking its likelihood from the $\chi^2$-distribution. Specifically, per the discussion in the Appendix A.7, the following quantity $(N-1)s_{\Delta k^{(\mathrm{app})}}^2 / \sigma_{\Delta k^{(\mathrm{app})}}^2$ should have a $\chi^2$-distribution with $N-1$ degrees of freedom. If the calculated value is significantly different from the analytically calculated value, then it serves as a measure of the adequacy of the assumptions given in Equation (5). If the standard deviation of the experiment is assumed to be unknown, then the above quantity provides an approach for estimating a realistic estimate of the experimental procedure standard deviation. However, this is outside the scope of the current work.

As noted in Appendix A, the determination of the tolerance interval for normally distributed quantities is largely dependent on what the unknowns are, i.e., the mean and/or the standard deviation. When both the mean and standard deviation are unknown, then the treatment in Appendix A.7 must be used, which requires the use of the noncentral $t$-distribution. When only the mean is unknown, then the tolerance interval can be evaluated directly using the normal distribution (see Appendix A.5). Finally, when only the standard deviation is unknown, the $\chi^2$-distribution must be used (see Appendix A.6). Recall that TSURFER analysis assumes that the standard deviations of the experimental measurement and the calculated $k_{eff}$ value are both known. These assumptions, coupled with the linearity assumption in Equations (4) and (5), imply that the standard deviation of the bias is also known: that is, it can be analytically calculated. Furthermore, since only a single measurement is available from each experiment, the implication is that the mean value is unknown, so the analysis in Appendix A.5 is the most suitable for estimating the tolerance interval for the application bias.

As shown in that discussion, equivalent results may be obtained with either the epistemic or aleatory treatment, simply because the true standard deviation of the parameter is assumed to be known. However, if the measurement procedure results in non-normally distributed uncertainties, or if it lacks a credible estimate of the standard deviation, then the assumption of a known standard deviation is no longer applicable, so the methods in Appendix A.7 must be used to determine the tolerance limits. Finally, when the reported standard deviation for the measurement is based on a conservative analysis, then the resulting tolerance intervals will be conservative, as well. Therefore, the explicit treatment of measurement uncertainties would be warranted only when the tolerance intervals for the associated applications biases are deemed unacceptably high.

Based on the TSURFER assumptions, a tolerance limit is given by (see Appendix A.5):

$$\Delta k_{\mathrm{K-Tolerance}}^{(\mathrm{app})} = \left( \sum_{i=1}^{N} \lambda_i \Delta k_i^{(\mathrm{exp})} \right) + K_{95}\left(1 + \frac{1}{\sqrt{N}}\right)\sqrt{\sum_{i=1}^{N} \lambda_i (\sigma_t^{(\mathrm{exp}_i)2} + \sigma_t^{(\mathrm{exp}_i, \, \mathrm{cal})2})}, \tag{9}$$

where $K_{95}$ is the coverage parameter calculated from a normal distribution, and $K_{95/95} = K_{95}(1 + \frac{1}{\sqrt{N}})$ is the corrected value to compensate for the uncertainty in the estimate mean bias. If a 95%/95% tolerance limit is required, then a value of $K_{95} = 1.65$ should be used.

## 3. Numerical Verification Results

This section introduces a number of numerical tests to validate the calculated tolerance interval. The objective is to employ a random sampling approach to calculate the distribution of biases, considering the uncertainties in the experimental measurements and the

code's prior calculations. The resulting distributions are numerically evaluated to calculate the upper threshold limit for the tolerance interval per the discussion in Appendix A. The distributions are compared to the values analytically calculated by Equation (9). These results are generated using a TSURFER model containing 17 critical experiments and 1 application. The experiments are all from the LEU-COMP-THERM-008 benchmark experiments set, denoted hereafter, LCT-008 [35]. The LCT-008 experiments are critical lattices with low-enriched (2~4 wt.%) $UO_2$ fuel rods and perturbing rods in a cylindrical, borated water tank. Examples of the experiments are depicted in a cross-sectional view in Figure 2, in which the fuel rods with aluminum alloy are presented in blue dots and the perturbing rods are shown in yellow surrounded by borated water, which is shown in green. Different critical experiments in the LCT-008 set include various perturbing rods materials and arrangements, or the boron concentrations are varied. For instance, Figure 2a,d presents two LCT-008 experiments without perturbing rods, whereas experiment LCT-008-016 contains a borated water slab in the center, of 1158-ppm boron, whereas the boron concentration in experiment LCT-008-001 was 1511 ppm.
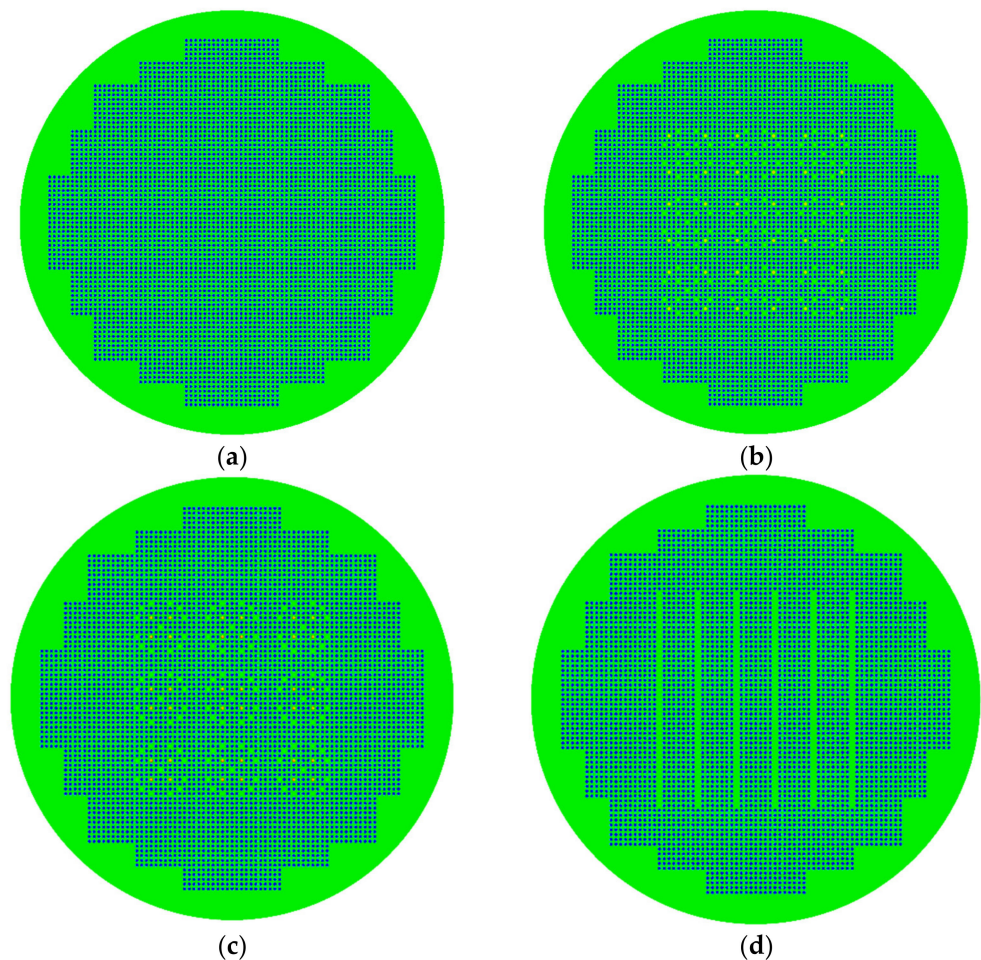


**Figure 2.** Layout of Critical Experiments in LCT-008 Benchmark Set. (**a**) LCT-008-001; (**b**) LCT-008-005; (**c**) LCT-008-012; (**d**) LCT-008-016.

Figure 2b,c show the layout of two experiments with different perturbing rod materials, in which both experiments have 3 × 3 central arrays of water holes and perturbing rods using Pyrex in LCT-008-005, whereas alumina is used in LCT-008-012. The application is an accident-tolerant fuel (ATF) assembly which employs a 10 × 10 BWR-dominant lattice of GE14 design, using $UO_2$ fuel pins with FeCrAl cladding. The layout of the application model is shown in Figure 3, with 92 fuel pins and two water rods in the center, surrounded

by coolant water in a channel box. There should be 78 $UO_2$ fuel pins with enrichments ranging from 2~4%, and 14 rods with gadolinium.
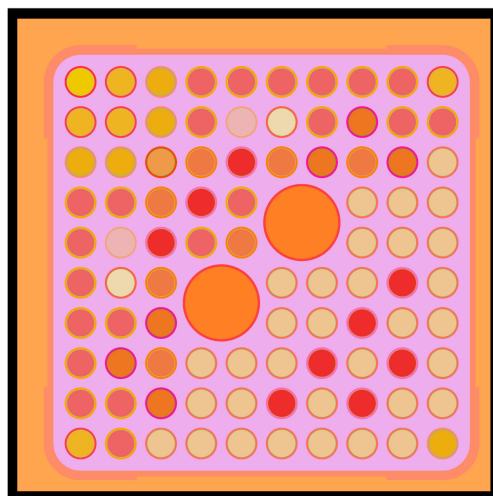


**Figure 3.** Layout of Application ATF Assembly.

Table 1 lists the prior values for $k_{eff}$ and the measured values, as well as the adjusted values resulting from the TSURFER application for all experiments and the application. Note that the application does not have a measured value. The adjusted values post-TSURFER represent the mean values of the PDFs generated by TSURFER. Sample results of the prior and post-TSURFER PDFs for a number of experiments and the application are shown in Figure 2.

**Table 1.** TSURFER $k_{eff}$ Values.

| Benchmark Experiment | Code-Prior Value | Measured Value | Adjusted by TSURFER |
|---|---|---|---|
| LEU-COMP-THERM-008-001 | 0.999830 | 1.00000 | 0.999694 |
| LEU-COMP-THERM-008-002 | 1.000356 | 1.00000 | 1.000293 |
| LEU-COMP-THERM-008-003 | 1.001015 | 1.00000 | 1.000947 |
| LEU-COMP-THERM-008-004 | 1.000146 | 1.00000 | 1.000116 |
| LEU-COMP-THERM-008-005 | 0.999817 | 1.00000 | 0.999773 |
| LEU-COMP-THERM-008-006 | 1.000374 | 1.00000 | 1.000380 |
| LEU-COMP-THERM-008-007 | 0.999652 | 1.00000 | 0.999657 |
| LEU-COMP-THERM-008-008 | 0.998910 | 1.00000 | 0.999157 |
| LEU-COMP-THERM-008-009 | 0.999103 | 1.00000 | 0.999334 |
| LEU-COMP-THERM-008-010 | 0.999998 | 1.00000 | 0.999937 |
| LEU-COMP-THERM-008-011 | 1.000675 | 1.00000 | 1.000614 |
| LEU-COMP-THERM-008-012 | 1.000216 | 1.00000 | 1.000148 |
| LEU-COMP-THERM-008-013 | 1.000363 | 1.00000 | 1.000292 |
| LEU-COMP-THERM-008-014 | 1.000220 | 1.00000 | 1.000115 |
| LEU-COMP-THERM-008-015 | 1.000051 | 1.00000 | 0.999962 |
| LEU-COMP-THERM-008-016 | 1.000088 | 1.00000 | 1.000087 |
| LEU-COMP-THERM-008-017 | 0.999352 | 1.00000 | 0.999487 |
| B-GE14DOM-hf.v40-FeCrAl-1 | 0.933281 | - | 0.932013 |

In the discussion below, prior PDFs refer to the PDFs consolidated by TSURFER, including the code-based PDF, centered around the calculated value, and the measured PDF, centered around the measured value. For example, Figure 4 shows the code-based prior PDF (in red), the measured PDF in black, and the consolidated PDF in blue for LEU-COMP-THERM-008-008, as denoted by "Fused-Post".
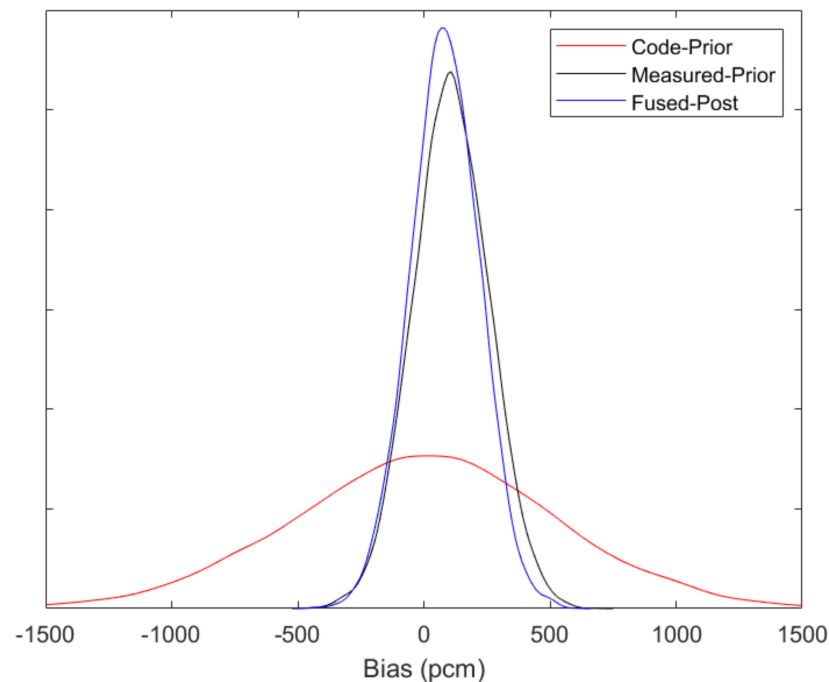


**Figure 4.** Prior and consolidated PDFs for the representative experimental model.

The *x*-axis is given in units of pcm, subtracting all $k_{eff}$ values from the reference code-calculated prior values for the respective experiments. Notice the measured-prior PDF (in black) has a mean value slightly above zero, implying that the code underestimates the measured value, which is assumed to be 1.0 for all experiments. The spread of the measured-prior PDF is consistent with the measurement uncertainty, which is assumed to be 150 pcm for all experiments. The TSURFER-consolidated PDF (also referred to as *post-PDF*) has a narrower width than the code-prior PDF because the measurement uncertainty is much smaller than the code-calculated prior uncertainty (on the order of 500 pcm).

Figure 5 shows a similar plot but for the application, which does not have a measured-prior PDF. Although the mean value of the consolidated PDF did not change by much, the standard deviation has experienced a drop from 670 pcm as calculated from code-prior PDF (in red) down to 388 pcm (in blue). The drop is not as significant as in the case with the experiments. This is because there is a low similarity between the application and the experiments. The similarity index is a scalar quantity used by TSURFER to determine the relevance of experiments to the given application; its use is not required by the Bayesian methodology.

As mentioned in the discussion above, TSURFER can calculate all its quantities of interest, including the experimental biases and the application bias and their standard deviations, via an analytical procedure, based on two reasons: (a) the model relating the variations in $k_{eff}$ to the cross sections is assumed to be linear, and (b) the standard deviations of the code-calculated $k_{eff}$ and measurements are assumed to be known. The calculated standard deviations represent the basis for tolerance-limit evaluation, so the next series of numerical experiments is designed to estimate the mean value of the bias and the standard deviation using a sampling-based (i.e., aleatory) approach to test the adequacy of the estimated quantities for tolerance limits estimation.
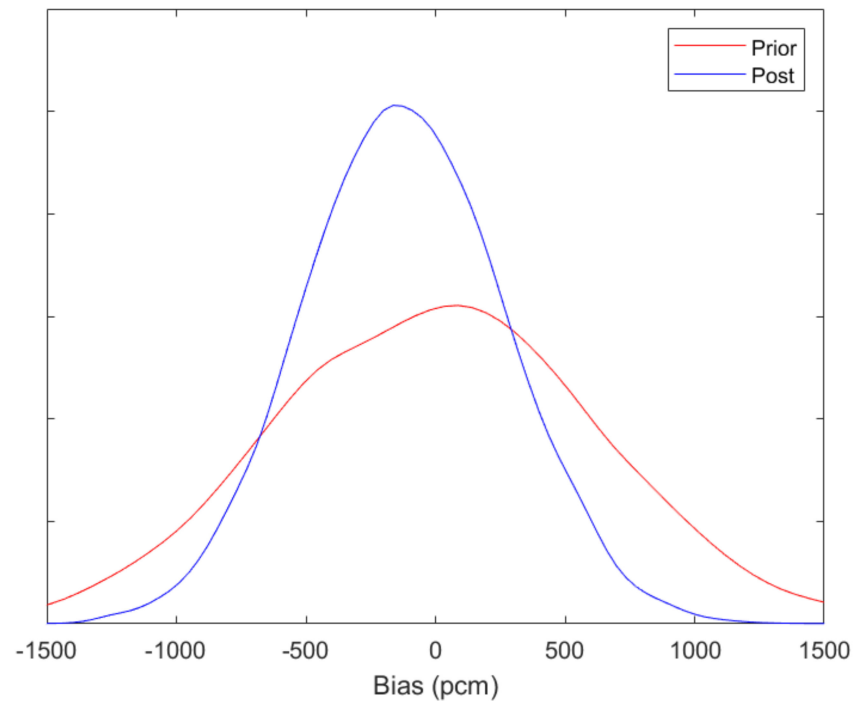
**Figure 5.** Application of prior- and post-TSURFER PDFs.

Equation (7) represents the key equation employed by TSURFER to calculate the application bias as a function of the discrepancy between measured and code-calculated $k_{eff}$ values for all available experiments. This equation calculates an estimator for the mean value of the TSURFER bias, so a sampling-based approach is employed, emulating the discussion in the Appendix A.5. To achieve this, $N$ random samples are generated for the discrepancy term $\Delta \vec{k}_j^{(\exp)}$ by generating a random sample from the measured-prior PDF—Equation (2)—and a random sample from the code-prior PDF—Equation (3)—and averaging the $N$ biases calculated from Equation (7), as follows:

$$\Delta k^{(\mathrm{app})} = \frac{1}{N} \sum_{i=1}^{N} \vec{\lambda}^T \, \Delta \vec{k}_j^{(\exp)} .$$

This process is repeated 10,000 times, producing 10,000 estimates of the mean, each based on $N$ samples of the discrepancy vector. Figure 6 shows the sampling distribution for the mean using different values of $N$. The blue PDF corresponds to $N = 1$, which is equivalent to the PDF generated by TSURFER, shown in red in Figure 5. The standard deviations for the various values of $N$ are shown in the legend, thus confirming the $1/\sqrt{N}$ behavior for mean value estimation. Figure 7 shows the plots of the quantity $(N-1)s^2_{\Delta k^{(\mathrm{app})}}/\sigma^2_{\Delta k^{(\mathrm{app})}}$, which are found to follow a $\chi^2$ distribution, with 16 degrees of freedom, as expected from Equation (8).
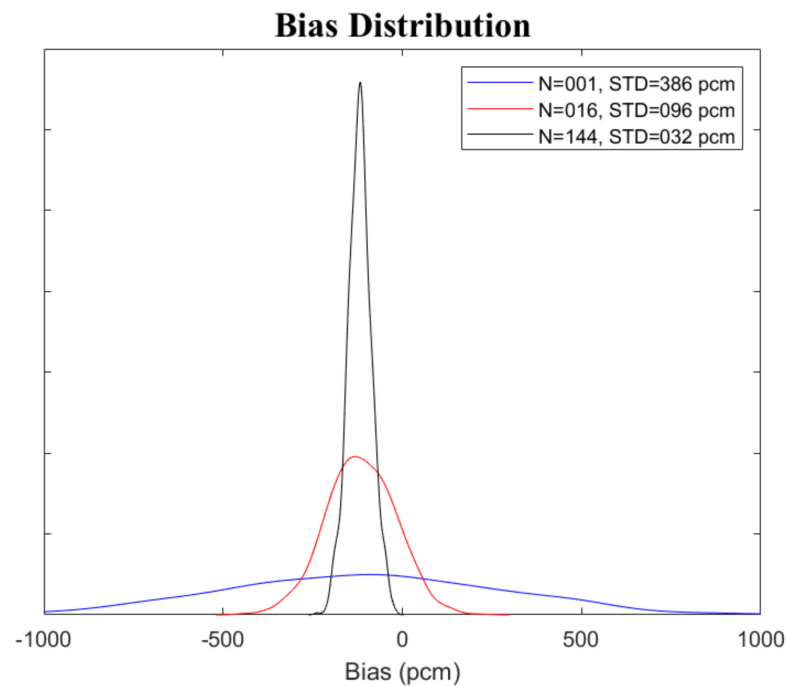
**Bias Distribution**



**Figure 6.** Mean estimation of application bias.

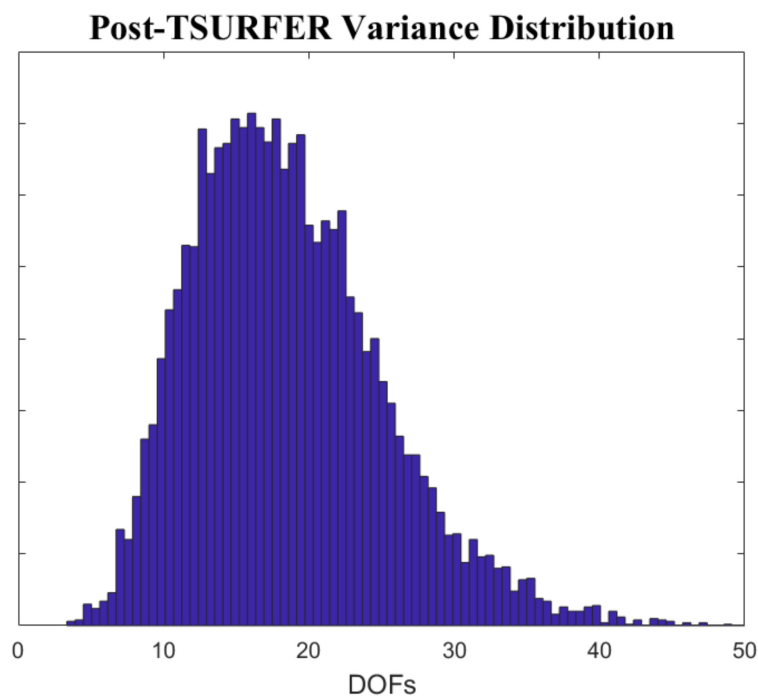**Post-TSURFER Variance Distribution**



**Figure 7.** DOFs of the variance of experiments' residuals.

To calculate the *K*-coverage parameter, the process described in the Appendix A.7 is used, (1) assuming that neither the mean value nor the standard deviation is known, (2) using a fixed number of samples, and (3) estimating the sample mean and standard deviation. This produces Figure 8, which is similar to Figure A7. The red line passes approximately through the center of the cloud corresponding to the $K_{95} = 1.64$, which is calculated from the normal distribution. A higher value $K_{95/95}$ is selected to ensure that 95% of the values are above the line, which ensures that the upper threshold of the tolerance limit is covering 95% of the bias values. Similar plots can be generated for different values of *N*, recording the corresponding $K_{95/95}$ value. Figure 9 shows the result

of this experiment, with a sharp initial decline in the ratio $K_{95/95}/K_{95}$ for a smaller value of $N$. In reality, the $K$-coverage should approach 1.0 for an infinite number of samples, which is taken to be 10,000 in this numerical experiment. Assuming the mean value is based on 17 samples, one per experiment, $K_{95/95}/K_{95}$ =1.54.
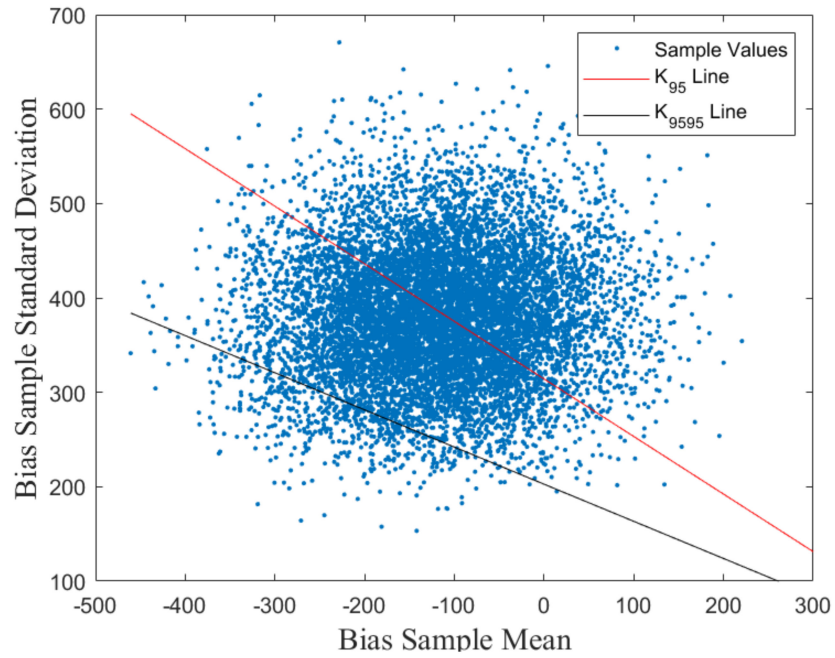


**Figure 8.** Tolerance limit upper threshold determination.
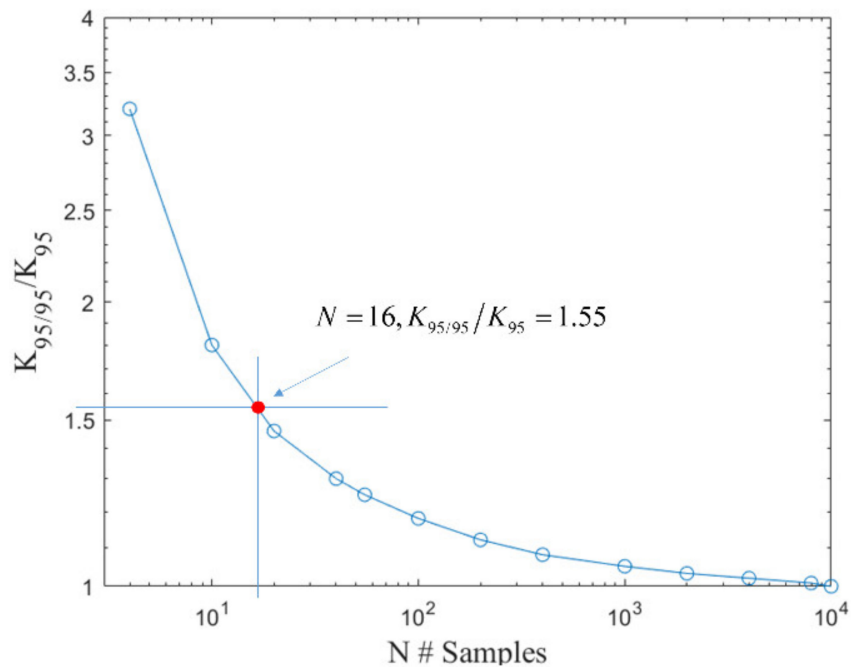


**Figure 9.** *K*-Coverage determination with unknown mean and standard deviation.

Finally, given that TSURFER bias PDF is analytically available as shown in Figure 10, the true mean and standard deviation are known, which is expected to give a lower upper threshold value than the one calculated above. The upper thresholds per Equation (9), with and without the correction factor $1 + 1/\sqrt{N}$, are shown in Figure 10. The area below the non-corrected value $K_{95}$ is found to be 94.4%, and the corrected value $K_{95/95}$ covers

98% of the area. It is not surprising that the two values are close, differing only by a factor of $1.24 = 1 + 1/\sqrt{17}$.
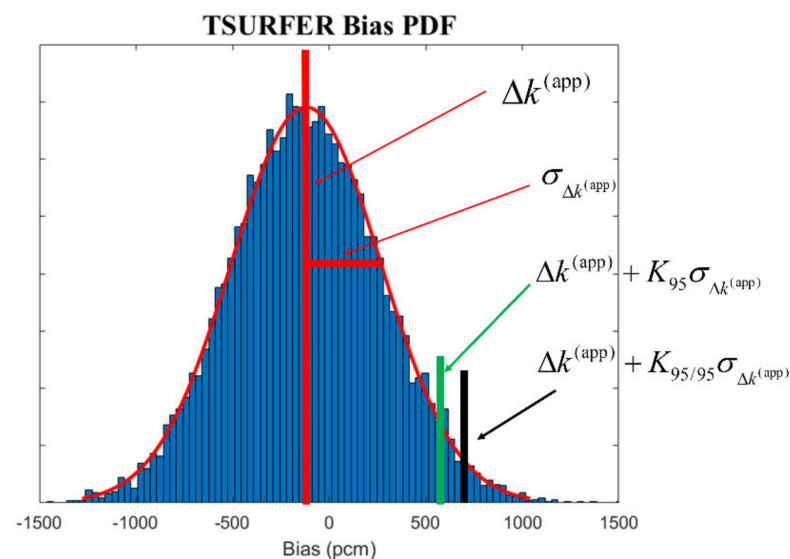
**TSURFER Bias PDF**



**Figure 10.** Application bias tolerance interval.

## 4. Conclusions

In support of using TSURFER for criticality safety applications, this work has focused on the development of a defendable methodology to calculate tolerance limit for TSURFER-calculated bias. TSURFER currently calculates both the bias and its uncertainty, but it lacks the ability to determine the probability that the bias may exceed a certain upper threshold value, which is needed to support the calculations of upper subcritical limits. In response to this need, the mathematical basis for tolerance limit calculation has been developed in a manner consistent with TSURFER Bayesian-based methodology. Deviating markedly from other mathematically heavy presentations, the current work adopts a pedagogical presentation style intended to target practitioners to support the end-users of TSURFER. In this spirit, the work focuses on providing an intuitive understanding of the meaning of tolerance limits and their calculational methods. Basic mathematical equations supported by graphical aids are employed to illustrate the meanings of the terms and methods developed. Finally, numerical experiments using realistic critical benchmark experiments are used to exemplify application of the TSURFER code. Beyond TSURFER, this paper is expected to benefit a wide range of model-validation and uncertainty-quantification activities, for which the goal is to characterize and consolidate confidence from multiple knowledge sources. The work presents general description for the $C^3$ process, which may be adapted to various end-user applications, such as the assimilation of results from experiments and calculations, from codes with different levels of fidelity, and so on.

**Author Contributions:** Conceptualization, H.A.-K.; methodology, H.A.-K.; formal analysis, H.A.-K.; investigation, H.A.-K., U.M. and W.M.; data curation, H.A.-K.; writing—original draft preparation, H.A.-K., D.H. and U.M.; writing—review and editing, U.M., W.M. and W.W.; supervision, H.A.-K. All authors have read and agreed to the published version of the manuscript.

## Appendix A. Basic Statistical Definitions and Operations

Brief introductions to the frequentist and Bayesian viewpoints are made, followed by the concept of sampling and inference analyses, which represent the fundamental algorithmic building blocks for the frequentist and Bayesian methods. The following discussion applies these analyses to estimate the tolerance interval, starting with idealized scenarios, and gradually building up to the general case. The discussion in this section is meant to be generic to motivate the value, interpretation, and calculation methods for tolerance limits. The intended application is presented in Section 2 for the TSURFER methodology.

*Appendix A.1. Frequency vs. Bayesian Probability*

Statistical description of the outcome of an experiment seeking to determine the value of an unknown parameter is essential under two scenarios: (a) when the analyst cannot make a deterministic statement about the value of the parameter, or (b) when the parameter itself is inherently random, so it does not assume a single value. An example of the first scenario is when the parameter has a fixed single value like a physical constant, but the experimental procedure introduces unavoidable sources of uncertainties, thus resulting in seemingly random outcomes for the measured parameter value. Mathematically, this is described as follows:

$$y_i = z + \varepsilon_i, \ i = 1, 2, \ldots, N, \tag{A1}$$

where $y_i$ is the $i$th recorded or inferred measurement of the parameter's unknown true value $z$. The measurements are contaminated by a random error $\varepsilon_i$ resulting from a number of uncontrolled experimental conditions. The estimate of the true value $z$ may be improved by repeating the same experiment a number of times $N$ or by conducting $N$ other experiments. This type of inference is referred to as *Bayesian inference*, and the parameter $z$ is referred to as an *epistemic parameter*.

In the second scenario, the parameter is inherently random, as dictated by the physics, such as the number of counts in a Geiger counting experiment; hence, its measured value is expected to exhibit random behavior, which may be described mathematically as follows:

$$y_i = z_i + \varepsilon_i, \ i = 1, 2, \ldots, N. \tag{A2}$$

Unlike the previous equation, the true parameter value $z_i$ is inherently random, because it changes its value every time it is recorded. This value is referred to as an *aleatory parameter*. The recorded values $z_i$ may be split into two terms emulating Equation (A1), as follows:

$$z_i = \mu_t + \delta_i, \ i = 1, 2, \ldots, N. \tag{A3}$$

whereas Equations (A1) and (A3) are mathematically similar, their interpretation is distinctly different. Equation (A1) asserts that $z$, the true value of the parameter, is single-valued, but its inference is obfuscated by the uncontrolled random experimental errors. If a high precision experiment is employed, then the random term in Equation (A1) will be significantly diminished, ultimately approaching zero in the limit of perfect and/or many

repeated measurements. In Equation (A3), the split into two terms $\mu_t$ and $\delta_i$ is artificial. It represents a mathematically convenient way to describe an aleatory parameter in terms of a constant term $\mu_t$, representing the mean parameter value, and a random term with zero mean $\delta_i$, representing deviations from $\mu_t$. Note that the random term $\delta_i$ will not vanish if perfect measurements are possible. Therefore, $\mu_t$ may be thought of as a mathematical feature that compactly describes the distribution of the random values assumed by the aleatory parameter. Other mathematical features may also be derived, such as the standard deviation $\sigma_t$ of the random term $\delta_i$, and the tolerance interval containing a certain portion of the random values, which is defined in the next subsection.

Note that the extracted features, $\mu_t$ and $\sigma_t$, have fixed single values, despite the aleatory nature of the parameter they describe. The implication is that one may resort to Bayesian or epistemic methods to infer the features of an aleatory parameter. Conversely, frequentist or aleatory methods may be used to describe an epistemic parameter. For example, the variable $y$ in Equation (A1) may be regarded as an aleatory parameter, since it is contaminated by uncontrolled random errors. If the PDF of the aleatory parameter can be captured, then its mean value may be used to infer the unknown epistemic parameter value $z$. Examples of the interaction between aleatory and epistemic methods are shown below.

Figure A1 plots two PDFs that describe the true values assumed by typical epistemic and aleatory parameters. The epistemic parameter is described by a delta function (in black) centered around the true value—a mathematical abstraction of a PDF with zero spread. The other distribution describes an aleatory parameter which assumes a range of values resulting from its inherent randomness. The two types of uncertainties are relevant to the TSURFER methodology. The variable $y$ in Equation (A1) may represent the measured $k_{eff}$ value from a number of experiments in which the measurement process is contaminated by numerous sources of aleatory and epistemic uncertainties. The experimentalist typically lumps these uncertainties together in the form of a single source, and the standard deviation of the source may be reported with each measurement. The variable $z$ may represent the cross sections for which the true mean values are unknown.

Cross sections are an example of aleatory parameters that are treated as epistemic in the TSURFER methodology. In theory, cross sections characterize the probability of interaction between a nucleus and a neutron, which is an inherently random event, implying aleatory treatment. However, the cross sections evaluation procedures [36] (a combination of differential and integral measurements and analytical methods) result in a mean value estimate (dotted blue line) that is significantly different from the true mean value of their aleatory spread (dotted black line), thereby justifying their treatment as epistemic.
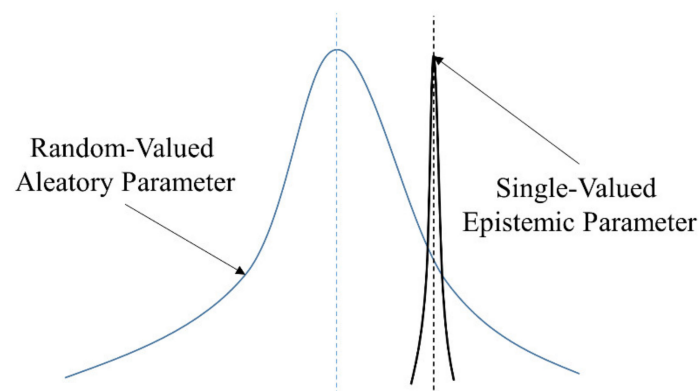


**Figure A1.** PDF Examples for True Values of Epistemic and Aleatory Parameters.

*Appendix A.2. Statistical Sampling and Inference Analyses*

Like deterministic calculations, two analyses can be defined in statistics: the sampling and the inference analyses, representing the equivalent of deterministic forward and inverse analyses, respectively. In the sampling analysis, the true PDF is known, and the goal is to generate samples that are consistent with that PDF, often to be employed

in subsequent downstream calculations such as sampling the cross sections from their prior PDFs to calculate the corresponding PDF of $k_{eff}$ for a target application or critical experiment. Consistency can be measured in a number of ways, such as by comparing features (e.g., mean, standard deviation, tolerance intervals) that are extracted from the true PDF and the available samples. In the inference analysis, $N$ samples of the parameter, or functions thereof (e.g., samples of experimental $k_{eff}$ which are related to the target application $k_{eff}$ and/or the nuclear data) are presented with the goal of determining the true PDF that generated the samples. In many practical applications, the objective is to determine features like the PDF tolerance levels rather than defining the entire PDF. This objective represents the key focus of this work.

An important byproduct of sampling analysis is the ability to determine the probability that a sample is drawn from a certain range of values within the distribution. This is a key requirement for the inference analysis when a single measurement is available, as illustrated in the next subsection. Statements such as the following are typical: Prob $(z < z_H) = 0.9$, where $z$ is an aleatory parameter, and $z_H$ is some fixed upper limit. This statement asserts that there is $p = 90\%$ chance that a single sample has a value lower than $z_H$. This statement, while probabilistically posed, has the following deterministic interpretation: *if the entire population's values are known (i.e., only achieved with infinite number of samples), then exactly p fraction of these values will be below $z_H$.* Even though this statement has no uncertainty, it is unrealizable in practice. With a finite number of samples, this $p$ fraction is no longer an exact deterministic number. This can be demonstrated by drawing $N$ samples and recording the fraction of values that are below $z_H$, and then repeating the experiment by drawing $N$ new samples, and so forth. In each experiment, a different fraction $p$ is calculated, representing the fraction of the $N$ values that are below $z_H$. The recorded values of $p$ are expected to have a distribution of their own, as shown in the three subplots on the left in Figure A2 for different values of $N$. Note that the distribution of $p$ values intuitively shrinks as $N$ increases, ultimately converging to a delta function centered around the exact value of $p$, which can be calculated from the $z$ distribution for the given threshold value $z_H$; for example, in a normal distribution with a zero mean and a unit standard deviation, $p = 95\%$ for $z_H = 1.65$. These are the numerical values employed in Figure A2. The plots show that when $N$ is finite, there is always a non-zero probability to calculate values for $p$ that are less than the true value of 95%. For example, in the case of $N = 1000$, the distribution of $p$ values is approximately centered around the 95% value, suggesting that there is 50% chance that the threshold value of $z_H = 1.65$ will contain at least 95% of the distribution. There is an equal 50% chance that coverage of less than 95% would be obtained. Clearly, if $N$ continues to increase, then a 50% chance that the calculated $p$ value will be less than 95% will continue to exist. However, the range of values is shrinking closer to the 95% value.

If $z_H$ is multiplied by a number greater than 1.0, then the probability of covering at least 95% of the population will increase above 50%, as demonstrated on the right three subplots of Figure A2, which employ a threshold value of $z_H = 2.0$. These graphs suggest that when $N > 1000$, there will be a near-zero (in reality a very small number) chance that the calculated $p$ value will be less than 95%, thus, providing much more confidence than 50%. Hence, the confidence in the estimated coverage can be increased by simply increasing the size of the tolerance interval. This illustrates the basic concept behind tolerance interval estimation techniques. Details on how this can be achieved for various distributions are presented below.
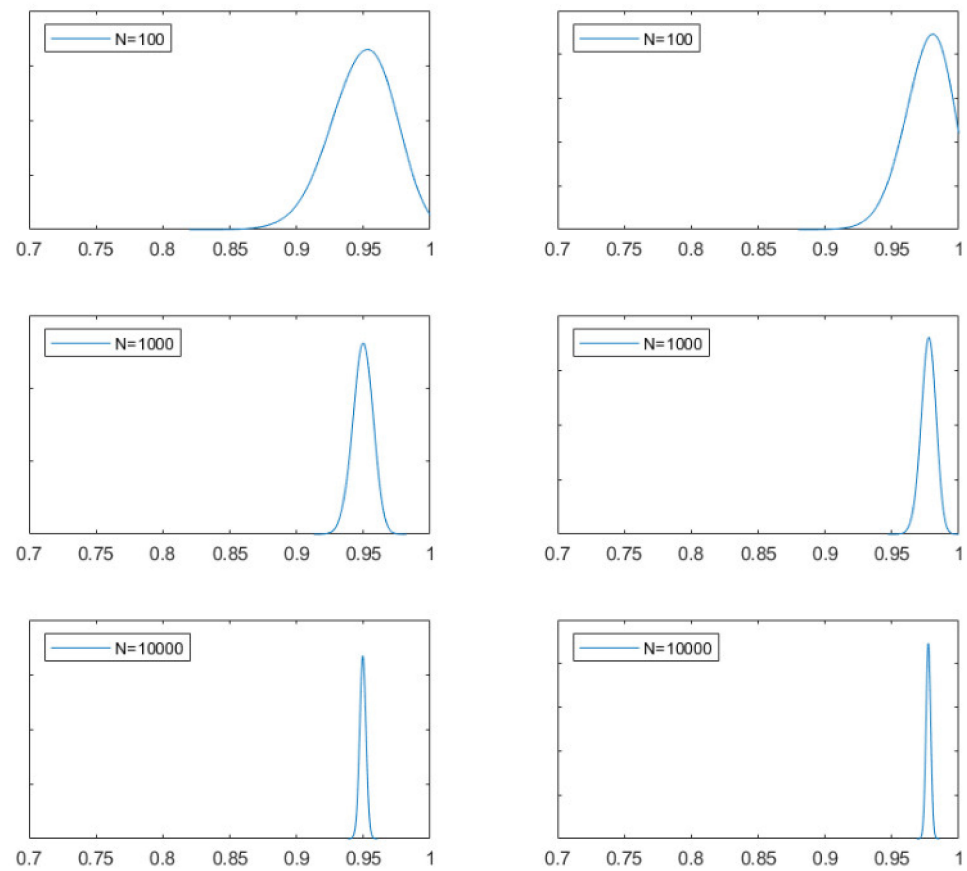
**Figure A2.** Tolerance Interval $\alpha$-Coverage with Finite Samples.

With the epistemic uncertainty, the sampling analysis is trivial, since all samples will assume the same value. In reality, the true PDF is unknown, so the samples are generated based on a prior PDF, representing the best knowledge about the parameter. In the inference analysis, the goal is to improve the estimate of the true PDF in order to approach the shape of the delta function. Thus, upon the introduction of new measurements, a successful inference analysis for epistemic uncertainty will generate PDFs that are successively shrinking their spread and ultimately converging to the true value in the form of a delta function. With aleatory uncertainty, the sampling analysis is tasked with generating samples with different values to emulate the aleatory nature of the parameter. Although this is a straightforward process, the additional step of generating a cumulative density function (CDF) with values between 0 and 1 is required. Details on this step are not relevant to the current discussion. The goal of the inference analysis is to use the available samples to recover the true PDF. In most practical problems, it is much more convenient to infer features extracted from the PDF rather than the full PDF. For example, with a normal distribution, it suffices to know the mean and standard deviation to fully describe the distribution. Other features are also valuable, such as the *K*-coverage parameter, which defines an interval, which is often centered around the mean and stretched by the standard deviation for normal PDFs, according to the following:

$$\text{Prob}\,(z \;<\; z_H) = \; \alpha \text{where } z_H = \mu_t + K\sigma_t, \tag{A4}$$

where $\mu_t$ and $\sigma_t$ are the true mean and standard deviation of the distribution. The *K*-coverage parameter defines an interval which covers a portion $\alpha$ of the area under the PDF. Note that $\alpha$ has units of probability or fraction, and *K* is dimensionless, such that $K\sigma_t$ has the units of the original parameter *z*. Both are referred to as the *coverage parameters*, with *K* determining the size of the interval, and $\alpha$ determining the associated portion covered by

that interval. Tables may be found relating $K$ and $\alpha$ for many key distributions, including normal distribution, $\chi^2$-distribution, $t$-distribution, and non-central $t$-distribution.

In aleatory analysis, the $K$-coverage provides information on the interval, which is expected to contain the most variations of the parameter. This is valuable in manufacturing processes that involve assurances on product quality. In epistemic analysis, the $K$-coverage informs the interval that is most likely, with confidence $\alpha$, to contain the real value of the parameter. This information is key to quantifying risk as measured by the consequences of the parameter being outside this interval. This interval is referred to as the *tolerance interval*, and it represents the goal of this work, where the $z$ parameter refers to the bias calculated by TSURFER and the objective is to create an upper limit $z_H$ to cover $\alpha = 95\%$ of the distribution with 95% level of confidence. As mentioned earlier, if one employs infinite number of samples $N$ to calculate the tolerance limit, the confidence would be 100% that $z_H$ covers 95% of the distribution. Therefore, the goal is to find a mathematical approach to estimate a multiplier $K$ that brings the confidence up to a user-defined value.

*Appendix A.3. Interaction between Aleatory and Epistemic Methods*

This section discusses the key enabling mathematical principle that allows the analyst to seamlessly transition between the aleatory and epistemic analysis methods/interpretation. The true PDF that generated the samples in Equations (A1) and (A2) can be described as follows, assuming a normal PDF, which is one of the most commonly used PDFs in statistics,

$$p(z) \propto e^{-\frac{(z-\mu_t)^2}{2\sigma_t^2}}, \tag{A5}$$

where the proportionality implies that a constant is needed to normalize the area under the PDF to 1, (the exact definition of which is omitted since it does not benefit the discussion). In aleatory settings, this PDF describes the distribution of all the random values $z$ assumed by the aleatory parameter—the mean value of which is $\mu_t$, and the standard deviation is $\sigma_t$—typically referred to as the *sampling distribution*. A basic statistical result states that if $N$ measurements of $z$ are recorded, then the following can be shown [2]:

$$\mu_t = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} z_i, \text{ and } \sigma_t^2 = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} (z_i - \mu_t)^2, \tag{A6}$$

and furthermore, the integral,

$$\int_{z_L}^{z_H} p(z) dz = p, \tag{A7}$$

measures the frequency (or fraction of times) with which the parameter $z$ assumes values between the two limits $z_L$ and $z_H$. This statement implies that the parameter is expected to assume values outside this interval with a probability of $1 - p$. This interpretation of Equation (A5) represents the basis of the sampling analysis. As mentioned above, if $N$ is indeed infinity, then the above statements in Equations (A6) and (A7) produce exact values for these quantities.

In the inference analysis, the objective is to build an estimate of the PDF in Equation (A5) using a finite number of samples. More importantly, if additional samples become available, then the estimated PDF must be updated and converged to the true PDF given in Equation (A5). If only a single measurement is available, then the integral in Equation (A7) allows for the assertion that there is a probability $p$ that the single sample belongs to the interval $(z_L, z_H)$. As discussed below, this provides an aleatory-based approach for estimating confidence intervals for quantities such as the true mean and standard deviation of an aleatory parameter.

In the epistemic setting, this same PDF may be written as follows:

$$p(\mu_t) \propto e^{-\frac{(z-\mu_t)^2}{2\sigma_t^2}}. \tag{A8}$$

This PDF has the same form as that in Equation (A5), but the $z$ value is assumed to be single-valued and known, and $\mu_t$ is the unknown. This describes a scenario in which a single sample has been measured, and in the objective is to infer the value of the unknown parameter $\mu_t$. In statistical literature [5], this PDF is often denoted by: $p(\mu_t|z)$ and is referred to as the *distribution conditioned on the measurements*, which is referred to herein as the *inference distribution* to distinguish it from the sampling distribution. Unlike the aleatory case, this PDF does not imply that $\mu_t$ assumes random values; instead, it measures the analyst's degree-of-belief about the true value given the single measurement $z$. For example, if the first sample $z_1$ is very far from the true mean $\mu_t$, then the corresponding PDF will be centered around $z_1$, implying that $\mu_t$ is likely to be close to $z_1$. Although this may seem to be incorrect, it is the best guess possible, based on the single measurement $z_1$. Mathematically, the interval in Equation (A7) will be used to measure the analyst's confidence that the true value lies between $z_L$ and $z_H$. This probability $p$ hedges against the analyst's lack of complete knowledge about the true value of $\mu_t$ by stating that there is a $1-p$ chance that the true value might lie outside the given interval. With very few samples and high measurement uncertainty, the probability $1-p$ is expected to be high, but it will gradually approach zero via the consolidation approach (explained in the following section) as more measurements are accumulated.

*Appendix A.4. Consolidation of PDFs*

A key mathematical tool used in the inference analysis consolidates the PDFs obtained from multiple independent sources/different experiments. This mathematical approach allows for the unique differences between epistemic and aleatory uncertainties. For the aleatory case, the following two PDFs can be consolidated, based on $N_1$ and $N_2$, respectively, as independently recorded samples of an aleatory parameter $z$:

$$p_1(z) \propto e^{-\frac{(z-\mu_{t,1})^2}{2\sigma_{t,1}^2}} \quad \text{and} \quad p_2(z) \propto e^{-\frac{(z-\mu_{t,2})^2}{2\sigma_{t,2}^2}}.$$

The first (second) PDF implies that, based on the available $N_1$ ($N_2$) samples, the mean value of the samples is $\mu_{t,1}(\mu_{t,2})$ and the standard deviation of the samples is given by $\sigma_{t,1}(\sigma_{t,2})$. As these PDFs have the meaning of frequency, they can be consolidated according to the following rule:

$$p(z) = \frac{N_1 p_1(z) + N_2 p_2(z)}{N_1 + N_2}. \tag{A9}$$

This straightforward averaging approach is consistent with the frequency interpretation of aleatory uncertainty. It is equivalent to combining all samples from the two experiments into one experiment and recalculating the frequencies of the various assumed parameter values. This consolidation approach ensures that if the two PDFs are exactly the same, then the resulting PDF will assume the same shape, which is intuitively sound.

For the epistemic case, consider the following two PDFs:

$$p_1(\mu_t) \propto e^{-\frac{(z_1-\mu_t)^2}{2\sigma_{t,1}^2}} \quad \text{and} \quad p_2(\mu_t) \propto e^{-\frac{(z_2-\mu_t)^2}{2\sigma_{t,2}^2}}.$$

Each of these two PDFs provides a measure of the degree-of-belief about the true value of the mean value $\mu_t$ based on a single measurement. In practical applications, $z_1$ may represent the best estimate from prior knowledge, which is the reference value for $k_{eff}$ for a given model of a critical benchmark experiment, and $z_2$ is the corresponding real measurement for $k_{eff}$. For a critical benchmark model describing a thermal flux spectrum, typical

values for $\sigma_{t,1}$ are on the order of 500–700 pcm, and a typical value for $\sigma_{t,2}$, representing the measurement uncertainty, is on the order of 150 pcm. Bayesian consolidation of these two PDFs gives [37]

$$p(\mu_t) \propto p_1(\mu_t) p_2(\mu_t). \tag{A10}$$

This consolidation approach recognizes that with additional knowledge, the consolidated PDF must converge to the true delta-function–looking PDF, since the parameter has a single value rather than a range of values, as is the case with an aleatory parameter. Multiplying the PDFs achieves this goal. (A simple social example is provided here for illustration. Consider asking a random person about which road to take to reach a certain destination, with the options being limited to Road A or Road B. The first individual recommends Road B and claims to be 80% confident. If another individual provides the same information independently of the first individual, the overall consolidated confidence that Road B is the correct way should increase. One can show that the confidence increases to $(0.8 \times 0.8)/(0.8 \times 0.8 + 0.2 \times 0.2) = 0.94$. However, if the individuals provide contradicting information (e.g., if the second individual recommends Road A with 80% confidence), then the consolidated knowledge should assign equal confidence to both ways, for 50% each, thus representing the state of complete ignorance. This can be achieved via PDF multiplication. Averaging the PDF would not result in increased confidence when consistent knowledge is gleaned from multiple sources). This consolidation approach ensures that if similar PDFs are consolidated, then the resulting PDF will have a smaller spread, indicating higher confidence in the true value of the parameter. This approach is depicted in Figure A3.
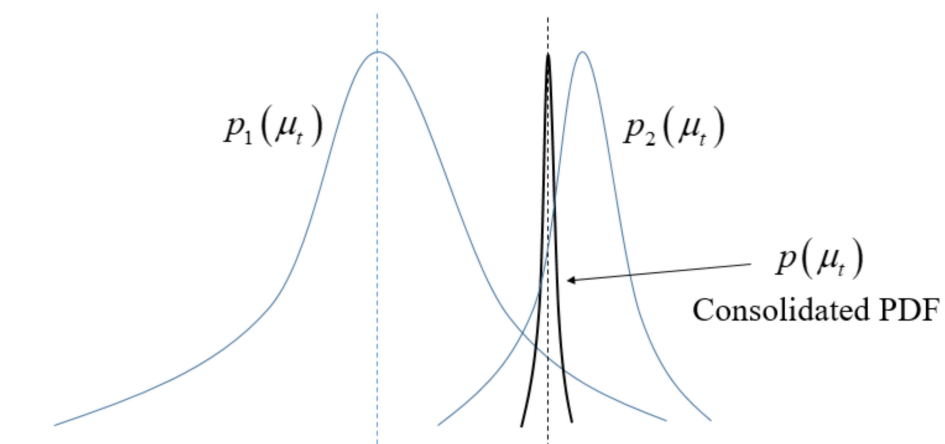


**Figure A3.** Bayesian Consolidation of Two Prior PDFs.

*Appendix A.5. Aleatory and Epistemic Mean Value Estimation*

One interesting observation is that while $z$ may be an aleatory parameter, its derived features, $\mu_t$, $\sigma_t$, and $K$, are epistemic because their true values are single-valued. The implication is that one could transition between the two viewpoints as dictated by the problem, as demonstrated in this and the next few subsections. This subsection starts with the simplest inference problem, in which the true value $\mu_t$ is unknown; however, $\sigma_t$ is known, and the PDF is known to be normal, which defines the relationship between $K$ and $p$, leaving $\mu_t$ to be the only unknown. This problem may be used to describe both epistemic and aleatory uncertainties per Equations (A1) and (A3), respectively. For illustration, assume that one is interested in estimating the true value of an epistemic parameter per Equation (A3). The Bayesian approach employs a sample $z_i$ to calculate an epistemic PDF for $\mu_t$ of the form:

$$p_i(\mu_t | z_i) \propto e^{-\frac{(\mu_t - z_i)^2}{2\sigma_t^2}}, \ i = 1, \ldots, N.$$

According to Bayes, the best-estimate epistemic PDF that consolidates these $N$ PDFs (assuming no prior knowledge) is given by

$$p(\mu_t) \propto \prod_{i=1}^{N} p_i(\mu_t) \propto \prod_{i=1}^{N} e^{-\frac{(\mu_t - z_i)^2}{2\sigma_t^2}} \propto e^{-\frac{(\mu_t - \bar{z})^2}{2\sigma_\mu^2}}, \tag{A11}$$

where it can be shown that [6]

$$z_m = \frac{\sum_{i=1}^{N} z_i}{N} \text{ and } \sigma_\mu = \frac{\sigma_t}{\sqrt{N}}. \tag{A12}$$

The resulting PDF is normal and is centered around the mean value of the samples $z_m$, and its standard deviation is $1/\sqrt{N}$ smaller than $\sigma_t$, the true standard deviation of $z$. As $N$ goes to infinity, the resulting PDF converges to a delta function centered around the true mean value $\mu_t$. The implication here is that while the parameter $z$ is aleatory, the inference analysis has adopted an epistemic setting to extract a feature of the aleatory parameter PDF: that is, its mean value $\mu_t$.

The same inference problem can be solved using an aleatory setting. This requires defining an estimator, which is a function of the samples that can be used to estimate the unknown quantity of interest, or the true mean $\mu_t$. The idea of using an estimator is inspired by the features definition in Equation (A12). This follows because it is desirable to use an estimator that converges to the true value with an infinite number of samples. If the number of samples is finite, then these very definitions are expected to produce random values, thus allowing for an aleatory treatment. (The same approach for an estimator is discussed in relation to Figure A2). The following estimator $z_m$ of the mean is thus intuitively reasonable:

$$z_m = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{A13}$$

Note that $z_m$ is a new aleatory parameter; it can be thought of as the mathematical average of $N$ aleatory parameters, all having the same PDF as the aleatory parameter $z$. To generate a sample for $z_m$, $N$ samples must be generated for the parameter $z$, and their average must be calculated (this basic approach is used by TSURFER whereby $N$ experiments are employed to estimate the bias, with each experiment producing a single sample measurement. See the discussion in Section 2.2 regarding Equation [7]). Assuming the true PDF of $z$ is known, the true PDF of $z_m$ can be predicted as [1] (this can be verified numerically by creating many samples for $z_m$, each of which is generated using $N$ random samples of $z$, and finally, building a histogram for the samples for $z_m$. This process can be repeated using different values for $N$. As $N$ increases, the histogram for $z_m$ will shrink at a rate proportional to $1/\sqrt{N}$):

$$p(z_m) \propto e^{-\frac{(z_m - \mu_t)^2}{2\sigma_\mu^2}}. \tag{A14}$$

Interestingly, this PDF has the same mean value as the true mean value of $z$, which is $\mu_t$, and its standard deviation is related to that of $z$ by $\sigma_\mu = \sigma_t/\sqrt{N}$. The implication is that if this experiment can be run many times, with each run producing a sample for $z_m$, then a distribution will be generated that is centered around the true mean value $\mu_t$ and a standard deviation $\sigma_\mu$. An example is shown in Figure A4 for a normal distribution, with $\mu_t = 1$ and $\sigma_t = 2$, for various values of $N$, with $N = 1$ representing the original distribution.

However, this approach is not a practical solution, since it requires many executions of the experiment. If this approach were feasible, then a PDF could have been built for the original parameter $z$ directly rather than employing an estimator $z_m$. To overcome this challenge, recall the approach employed above to establish confidence with a single measurement $z_m$. Since the goal is to estimate the fixed value of $\mu_t$, an epistemic treatment

per Equation (A8) is possible. Therefore, the following equation can be used to characterize confidence about $\mu_t$:

$$p(\mu_t|z_m) \propto e^{-\frac{(z_m - \mu_t)^2}{2\sigma_\mu^2}} \tag{A15}$$

This is the same PDF as the true PDF of the estimator $z_m$ in Equation (A14), but now as a function of $\mu_t$, with $z_m$ serving as a fixed single-valued measurement from the experiment. Per Equation (A7), confidence $p$ can be established that the true value lies within the interval $(z_m - K\sigma_\mu, z_m + K\sigma_\mu)$.
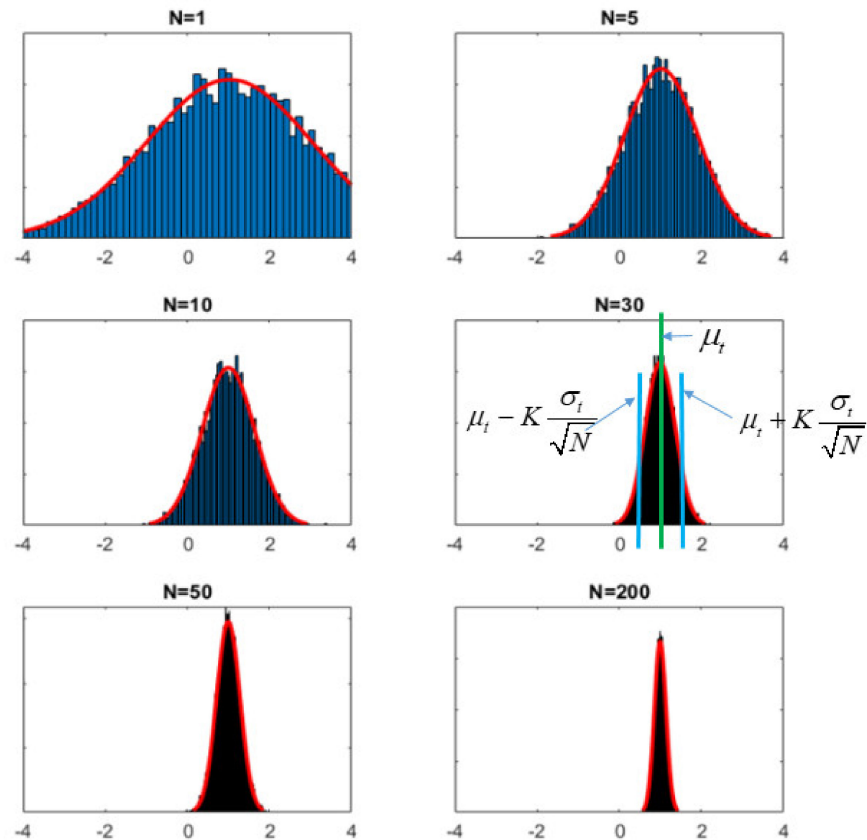


**Figure A4.** Distribution of the samples' mean with known standard deviation.

A discussion of an aleatory approach to characterize confidence using the concept of confidence interval follows. The objective is to find an interval that contains the true mean $\mu_t$. As discussed above, such intervals are typically described using a single $K$-coverage parameter. As depicted in Figure A4 (subplot with $N = 30$), there is a $p$ chance that the estimated value $z_m$, which represents a single sample drawn from the distribution in Equation (A11), will lie in the interval $(\mu_t - K\sigma_\mu, \mu_t + K\sigma_\mu)$ (represented by the two blue lines surrounding the mean value indicating by the green line), i.e.,

$$\text{Prob}\left(\mu_t - K\sigma_\mu < z_m < \mu_t + K\sigma_\mu\right) = p. \tag{A16}$$

This statement can also be interpreted as follows: there is $p$ chance that the estimator's value will be within no more than $K\sigma_t$ units away from the true mean $\mu_t$. Using numerical values of $p = 0.95$, $K = 2$, $N = 100$, $\mu_t = 1$, $\sigma_t = 1.5$, and $z_m = 1.2$, there is 95% chance that $|\mu_t - 1.2| < 0.3$, i.e., the true mean is over- or under-estimated by at most 0.3 units from the value 1.2. The interval [0.9, 1.5] is called a 95% confidence interval for $\mu_t$ and it allows the above equation to be rewritten as follows:

$$\text{Prob}\left(z_m - K\sigma_\mu < \mu_t < z_m + K\sigma_\mu\right) = p \tag{A17}$$

These two statements are equivalent simply because the true standard deviation of the estimator $\sigma_\mu$ is known. The first statement shows that $z_m$ cannot be away from the true mean by more than $K\sigma_\mu$ with $p$ confidence, and the second statement indicates that the true mean cannot be away from the single sample $z_m$ by more than $K\sigma_\mu$ with the same confidence $p$. The second statement can be readily calculated using the Bayesian PDF by integrating over $\mu_t$ between the two limits, $z_m - K\sigma_\mu$ and $z_m + K\sigma_\mu$. If interpreted in an aleatory sense, this statement hedges only once, emphasizing the fact that if the experiment is repeated $M$ times, then $(1 - p) M$ of these intervals will fail to contain the true mean $\mu_t$. It is interesting to note, here, that the true mean $\mu_t$ is an epistemic parameter, whereas the confidence interval concept is aleatory. If the confidence interval is interpreted in a Bayesian sense (it is called a *credible interval*), then based on the single measurement, the analyst thinks that the true value may be outside the interval with confidence $1 - p$, thus measuring a lack of confidence in the single measurement as $1 - p$.

This discussion highlights the mechanics of an inference analysis that started with an aleatory approach to form an aleatory estimator, per Equation (A13), and its associated PDF per Equation (A14), which then switched to an epistemic approach per Equation (A15) to characterize confidence using a single measurement. Note that the aleatory approach required the selection of an estimator, a suitable functional form to estimate the quantity of interest with the given samples. Clearly this decision can be difficult for general quantities of interest, as shown below for quantities such as the $K$-coverage parameter, in which both the true mean and the standard deviation are unknown. Thus, a clear advantage for the Bayesian approach is that it did not require the use of estimators. It only required interpretation of the original PDF as a measure of confidence for the unknown quantity given a single measurement in the sense of Equation (A8). With additional measurements, a straightforward approach employing multiplication of PDFs, as in Equation (A11), resulted in a PDF that automatically improves the estimate of the mean. Interestingly, it could be argued that the Bayesian approach provides insight on what estimator function should be used for aleatory analysis, simply because its definition emerges naturally following the successive multiplication of the $N$ PDFs from the $N$ samples.

As explained above, the mean value represents one possible feature that may be extracted from the PDF. Other important features include an interval or a range of values that covers a user-defined portion of the distribution. This was defined earlier as the $K$-coverage parameter, which covers an $\alpha$ portion of the distribution. This interval is referred to as the *tolerance interval* in both frequentist and Bayesian statistics. As an example, using the sampling distribution in Equation (A5), it can be shown that the following interval $(-\infty, \mu_t + 1.96\sigma_t)$ contains 97.5% of the population of random values for an aleatory parameter, or a 97.5% degree-of-belief for the true value of an epistemic parameter. As discussed above, if infinite samples are available to estimate $\mu_t$, then the upper limit of this interval would be exact. Employing a simple estimator $z_m^H = z_m + K\sigma_\mu$ (with $K = 1.96$) for the true upper threshold $z_t^H$ of this tolerance interval, the sampling distribution for $z_m^H$ is given by

$$p(z_m^H) = C e^{-\frac{(z_m + K\sigma_t - z_t^H)^2}{2\sigma_\mu^2}}. \tag{A18}$$

This distribution is simply a shifted version of the distribution in Equation (A14) since the true standard deviation $\sigma_t$ is known. When using a single estimate of $z_m$, it is clear that there is a 50% chance that the estimated $z_m$ will be higher than the true mean $\mu_t$, causing the estimated upper threshold $z_m + 1.96\sigma_t$ to be higher than the true upper threshold $\mu_t + 1.96\sigma_t$. The implication is that there is 50% chance that the estimated tolerance interval $(-\infty, z_m^H) = (-\infty, z_m + 1.96\sigma_\mu)$—which is based on a single estimate of $z_m$—will contain less than 97.5% of the population of random values. To increase the confidence above 50%, the value of $K$ can be increased above 1.96 to make up for the under-estimated $z_m$ value: a higher value of $K$ is needed to ensure the following:

$$z_m + K_m\sigma_t > \mu_t + 1.96\sigma_t,$$

re-arranging as follows:

$$K_m > 1.96 + \frac{\mu_t - z_m}{\sigma_t} \rightarrow K_m > 1.96 + \frac{1}{\sqrt{N}} \frac{\mu_t - z_m}{\sigma_\mu}. \tag{A19}$$

This equation shows that if it can be ensured with 100% confidence that $K$ remains above the noted value, then the estimated tolerance interval will contain at least 97.5% of the population of $z$ values. This is only possible if $K$ is infinite, because a single sample of $z_m$ can potentially have very small values. In practice, it suffices to ensure that $z_m$ is only a few standard deviations away from the mean, which is expected to occur with a very high probability. For example, per Equation (A14), there is only 2.5% chance that $z_m < \mu_t - 1.96\sigma_\mu$. Inserting that in Equation (A19), it is straightforward to see that with $(100-2.5)\% = 97.5\%$ confidence, the following $K = 1.96(1 + 1/\sqrt{N})$ value ensures that at least 97.5% of the $z$ values will be contained in the interval upper-limited by $z_m + K\sigma_t$. This is referred to as an *upper tolerance limit* with 97.5% probability and 97.5% confidence. If $N$ is infinity, then $K$ approaches the minimum value required to contain 97.5% of the population with 100% confidence. As $N$ decreases, $K$ must increase to make up for the uncertainty in $z_m$, and this reduces the confidence below 100%, because it is unrealistic to select $K$ to be infinity.

In the previous example, both the aleatory and epistemic treatments have led to similar results. It is thus instructive to determine when the two approaches would be different. The key difference lies in the consolidation of prior knowledge. To understand this difference, assume that the confidence interval has been established for the true mean based on $N_1$ measurements of $z$. These results can be rendered using the estimator in Equation (A13) in the form of a PDF in Equation (A14). When additional measurements are made using $N_2$ samples, a new PDF may be independently constructed. The Bayesian approach only requires access to the prior PDF that was generated with $N_1$ samples to consolidate it with the new PDF. However, the aleatory consolidation requires not only the prior PDF, but also the number of samples used to generate it, as well as the functional form of the estimator, to ensure consistency among both sets of samples $N_1$ and $N_2$. Effectively, using the aleatory approach is equivalent to conducting a third virtual experiment that combines all available samples and employs a unified estimator. In practical settings, when knowledge is obtained from multiple sources, the details on the inference process (e.g., the exact functional form of the estimator and the number of samples) are often not well-documented. Results are often communicated in a minimal manner as the confidence interval and/or the associated PDF, for example. Therefore, it becomes difficult to justify consolidating knowledge from multiple sources as shown in Equation (A9). To overcome this problem, approximate methods [16,38,39] have been developed to consolidate knowledge for the tolerance interval from multiple sources. This problem is nonexistent with the Bayesian consolidation approach, as reflected in Equation (A10). Therefore, Bayesian methods have the advantage of consolidating knowledge from multiple sources, precluding the need to track the details of the inference analysis from each source.

*Appendix A.6. Standard Deviation Inference*

This subsection extends the inference analysis to characterize confidence in the estimation of the true standard deviation $\sigma_t$ using $N$ samples of the aleatory parameter $z$. The PDF shape is assumed to be normal, and only $\sigma_t$ is assumed to be unknown. The following section addresses the more general case in which both the mean and the standard deviation are unknown. Employing an estimator of the standard deviation, emulating the true value obtained with infinite number of samples per Equation (A6),

$$s_m = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (z_i - \mu_t)^2}. \tag{A20}$$

As shown above, the estimator $s_m$ is expected to be an aleatory parameter since it is based on a finite number of samples. An analytical expression can be derived for the true PDF of $s_m$, which is generated numerically for different values of $N$ with $\sigma_t = 2$, in Figure A5. It can be shown that the distribution of $Ns_m^2\sigma_t^2$ is given by a $\chi^2$-distribution with $N$ degrees of freedom [2]. With increasing $N$, the distribution approaches a normal distribution, and the corresponding distributions for $s_m$ (i.e., dividing the $\chi^2$-distribution by $N$ and taking the square root) ultimately converge to a delta function centered around the true value of the standard deviation. (As mentioned above, while the $\chi^2$-distribution can be analytically derived, it can be numerically verified by running an experiment in which $N$ samples are generated, from which an estimate of the variance is calculated; the process is repeated many times, building a histogram for the recorded variance samples. If all the recorded variances are divided by the real standard deviation, then the normalized form of the $\chi^2$-distribution is obtained. The value of numerical verification of the estimator distribution is stressed, because when the true PDF is not normal, then the analytical results relating the confidence to the K-coverage parameter can no longer be developed, so numerical methods must suffice. This provides a clear, intuitive method for developing the relationship between $p$ and $K$, as shown in the subsequent subsections).
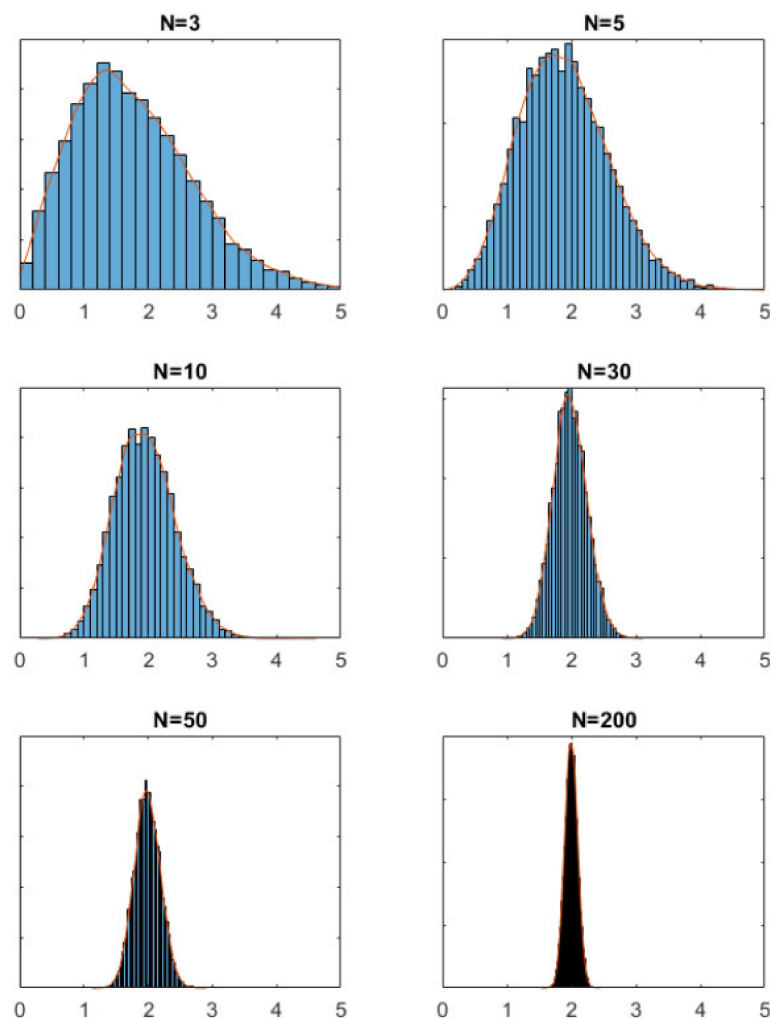


**Figure A5.** Distribution of the samples' standard deviation.

It can be shown that the shape of this distribution is unaffected by the true value, so it is possible to describe the PDF (it is also common to work directly with the variance rather than the standard deviation, since the $\chi^2$-distribution is a function of the variance. However, this subtlety is bypassed here for the sake of using simpler notations). $p_N(s_m/\sigma_t)$

in terms of a normalized variable $s_m/\sigma_t$. As in the previous analysis, this PDF allows for calculation of the confidence $p$ that a single sample, such as $s_m'$, will lie within a given interval. As the goal is often to estimate an upper bound on the true standard deviation $\sigma_t$, the following integral can be used to describe the confidence in the estimated value $s_m'$:

$$\text{Prob}\,(s_m' > \sigma_t) = \text{Prob}\,(\frac{s_m'}{\sigma_t} > 1) = p = \int_1^\infty p_N\left(\frac{s_m}{\sigma_t}\right) d\frac{s_m}{\sigma_t}.$$

This integral can be readily calculated since the form of $p_N(s_m/\sigma_t)$ is analytically known. This integral is intuitive, because it asserts that all estimates $s_m'$ greater than $\sigma_t$ contribute to the probability $p$. The implication is there is 1-$p$ chance that the true value $\sigma_t$ will exceed the estimated value $s_m'$, which is given by:

$$1 - p = \int_0^1 p_N\left(\frac{s_m}{\sigma_t}\right) d\frac{s_m}{\sigma_t} \tag{A21}$$

In practical applications, with $N$ being relatively small, the value of $1 - p$ might be too high to be satisfactory. Two useful mathematical approaches may be used to reduce the probability while not requiring a significant increase in $N$. The first approach involves the use of a multiplier, which serves as a conservative approach against the under-estimation of the standard deviation. This is achieved by multiplying the estimated $s_m'$ by a fixed multiplier $K$ and converting the integral above to the following:

$$\text{Prob}\,(Ks_m' > \sigma_t) = \text{Prob}\,(\frac{s_m'}{\sigma_t} > \frac{1}{K}) \to 1 - p = \int_0^{1/K} p_N\left(\frac{s_m}{\sigma_t}\right) d\frac{s_m}{\sigma_t}.$$

This shrinks the upper limit of the integration by a factor $K$, hence increasing the confidence $p$ that the new limit $s_m'' = Ks_m'$ is above the true standard deviation $\sigma_t$. This practice is very common among engineers, in which an additional multiplier is used to obtain a more conservative upper-bound on uncertain quantities of interest.

The second approach is to generate two estimates of the standard deviation, $s_{m,1}'$ and $s_{m,2}'$, and take the maximum $s_m'' = \max(s_{m,1}', s_{m,1}')$ to represent the best estimate of the standard deviation. It is easy to show that

$$1 - p = \left(\int_0^1 p_N\left(\frac{s_m}{\sigma_t}\right) d\frac{s_m}{\sigma_t}\right)^2.$$

This equation implies that $s_m''$ will be lower than the true standard deviation $\sigma_t$, only if both estimates are below the limit, which results in squaring the probability since the two estimates are independently drawn form $p_N(s_m/\sigma_t)$. This approach is known as *order statistics* [40,41], an example of which is Wilks statistics [14,15]. Combining the two approaches shown above using a $K$ multiplier and the maximum of $n$ independent estimates of standard deviation results in much higher confidence $p$:

$$p = 1 - \left(\int_0^{1/K} p_N\left(\frac{s_m}{\sigma_t}\right) d\frac{s_m}{\sigma_t}\right)^n. \tag{A22}$$

If the integral in Equation (A21) is initially equal to 0.22 for $N = 3$, implying an initial confidence of $p = 0.78$, then $K = 1.25$ and $n = 2$ increase the confidence to 0.98. These numbers may be interpreted as follows. First, $nN = 6$ samples of the aleatory parameter $z$ must be drawn and divided into two batches, $n = 2$, with each batch containing $N = 3$ samples. From each batch, a single estimate of the standard deviation is calculated per

Equation (A20), $s'_{m,1}$ and $s'_{m,2}$, and the maximum of the two is multiplied by $K = 1.25$ to form the best estimate $s''_m$ for the standard deviation $\sigma_t$. This estimate is expected to be higher than the true value with $p = 98\%$ confidence. The implication is that if this experiment is repeated $M$ times, then the estimated values will be lower than the true value in $0.02M$ of the repeated times, or $1 - p = 0.02$.

The previous discussion provides a method to estimate the true standard deviation of the PDF of an aleatory parameter. The confidence in the analysis is described using a probability measure $p$. The language suggested a single hedging, as before. However, these results can be used to introduce the concept of double hedging presented earlier. Recall that $\sigma_t$ is a fixed feature extracted from the true PDF of $z$. It may be interpreted as the square root of the average squared distance from the mean. In this interpretation, it is only necessary to hedge once, as in $p\,(s_m > \sigma_t) = 0.98$. The double hedging arises if $s_m$ is interpreted as a $K$-coverage interval with $K = 1$, or $(\mu_t - \sigma_t,\ \mu_t + \sigma_t)$, thus representing the interval that captures approximately 68% of the normally distributed random values:

$$\text{with confidence } p, \ \text{Prob}\,(\mu_t - s_m\ <\ z\ <\ \mu_t + s_m) = \alpha$$

this interval is referred to as a *tolerance interval* with $K = 1$. By definition, a tolerance interval is sought to capture a certain portion $\alpha$ of the population. If the distribution is known, then there is a one-to-one relationship between the $K$-coverage parameter and its corresponding $\alpha$. For example, in a normal distribution, a $K = 1$ corresponds to $\alpha = 0.683$, $K = 2.0$ to $\alpha = 0.955$, $K = 3.0$ to $\alpha = 0.997$ (Values rounded to the third significant figure), and so on. When the true standard deviation is known, these statements are exact, and their confidence is 100%. The hedging here refers only to the aleatory nature of future samples. However, when only a single estimate of the standard deviation is available, then it is possible that with probability $1 - p$, $s_m$ is less than $\sigma_t$, thus implying that the coverage $\alpha$ will be less than its true value. Using the following two equations,

$$\alpha(s_m) = \int\limits_{\mu_t - s_m}^{\mu_t + s_m} p(z)dz = \int\limits_{\mu_t - s_m}^{\mu_t + s_m} Ce^{-\frac{(z-\mu_t)^2}{2\sigma_t^2}}\,dz, \text{ and } p\left(\frac{s_m}{\sigma_t}\right) = 1 - \int\limits_{0}^{s_m/\sigma_t} p_N\left(\frac{s'_m}{\sigma_t}\right) d\frac{s'_m}{\sigma_t},$$

one can calculate the exact coverage $\alpha$ corresponding to different estimates $s_m$ using the true parameter PDF, as well as the associated confidence $p$ using the $\chi^2$-distribution. This information can be translated into economical or safety metrics that can be used to determine the appropriate values for the required coverage and the confidence $p$.

If a Bayesian approach is used, then the same PDF, $p_N(s_m/\sigma_t)$, can be used to characterize confidence in the true value $\sigma_t$ by treating $s_m$ as a fixed value and $\sigma_t$ as the unknown, as shown in the previous subsection. The PDFs corresponding to different estimates can be consolidated together via multiplication [6]:

$$p(\sigma_t) \propto \prod_{i=1}^{n} p_{N,i}(s_{m,i}/\sigma_t).$$

*Appendix A.7. Simultaneous Inference of Mean and Standard Deviation*

This subsection discusses the more general case in which both the mean and standard deviations are to be inferred from the samples. As before, the first step is to select the estimators for the mean and standard deviations:

$$z_m = \frac{1}{N}\sum_{i=1}^{N} z_i \text{ and } s_m = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(z_i - z_m)^2}.$$

Note the two changes in the definition of $s_m$: (1) it employs the mean value estimator $z_m$ since the true mean $\mu_t$ is unknown, and (2) it divides by $N$-$1$ rather than $N$. This is

explained in terms of the available number of degrees of freedom (DOFs), a key concept in statistical inference. If there are $N$ random samples of the parameter $z$, then they may be regarded as a vector in an $N$ dimensional space, as denoted by $\vec{z} = [z_1 \ z_2 \dots \ z_N]^T$. A coordinate system is required to describe the components, with the samples representing the components in the standard coordinate system. Clearly, different coordinate systems will result in different components obtained via

$$\xi_i = \vec{u}_i^{\,T} \vec{z},$$

where $\xi_i$ is the $i$th component in a different coordinate system, as described by $N$ orthonormal vectors, $\vec{u}_i|_{i=1}^N$. It is important to ensure that the statistical properties of the samples, as calculated by the selected estimators, remain invariant to the choice of the coordinate system. Invariance implies that, if an estimator of the standard deviation is calculated, then it should have the same value, regardless of the coordinate system employed. As shown below, when the estimated mean $z_m$ is subtracted from all the samples, it can be shown that there is a coordinate system that will always have one of its $N$ components zeroed out. This implies that the true number of DOFs has been diminished by one. To show this, recall that the sample mean is given by

$$z_m = \frac{1}{N}\sum_{i=1}^{N} z_i \rightarrow z_m = \frac{1}{N}\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_N \end{bmatrix} = \frac{1}{\sqrt{N}}\vec{e}_N^{\,T}\vec{z}. \tag{A23}$$

This equation shows that the sample mean is the projection (weighted by $1/\sqrt{N}$) of the random vector $\vec{z}$ onto the unit vector $\vec{e}_N = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T / \sqrt{N}$. The residual vector is thus given by

$$\begin{bmatrix} z_1 - z_m \\ z_2 - z_m \\ \dots \\ z_N - z_m \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_N \end{bmatrix} - z_m \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_N \end{bmatrix} - \left( \vec{e}_N^{\,T} \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_N \end{bmatrix} \right)\vec{e}_N = \vec{z} - \left(\vec{e}_N^{\,T}\vec{z}\right)\vec{e}_N = \left(\mathbf{I} - \vec{e}_N\vec{e}_N^{\,T}\right)\vec{z} = \mathbf{P}_e \vec{z},$$

where $\mathbf{P}_e$ is an orthogonal projector designed to remove the component along the vector $\vec{e}_N$. This implies that, while the residual vector has $N$ components, it only has $N-1$ random DOFs, with the component along the direction $\vec{e}_N$ zeroed out.

As the estimated standard deviation should be invariant to linear transformation, the estimator function should recognize that the transformed random vector, while still random, has only $N-1$ random components. Therefore, the mean may be thought of as a mathematical feature extracted from the random vector component along the direction $\vec{e}_N$, and the standard deviation may be considered as another feature extracted from the remaining $N-1$ components, which are independent of the component along the direction $\vec{e}_N$. As expected, this ensures that the two estimators are independent, as shown in the scatter plot provided in Figure A6, with the $x$-axis representing the sample mean values, and the $y$-axis representing the sample standard deviations.

The distribution of the $s_m$ values is discussed first. As described in Appendix A.6, the true PDF of $(N-1)s_m^2\sigma_t^2$ is the $\chi^2$-distribution can be shown analytically with $N-1$ DOFs. This is not surprising, because the removal of the sample mean reduces the DOFs by 1. It also ensures that the distribution of $s_m$ is independent of the distribution of $z_m$. This implies that the distribution of $s_m$ will be the same, regardless of the value of $z_m$, even if the true value $\mu_t$ is unknown. Hence, the discussion in Appendix A.6 still applies when estimating the true standard deviation, even when the true mean value is not known.

Therefore, one can calculate the confidence that the estimated standard deviation bounds the true standard deviation as follows:

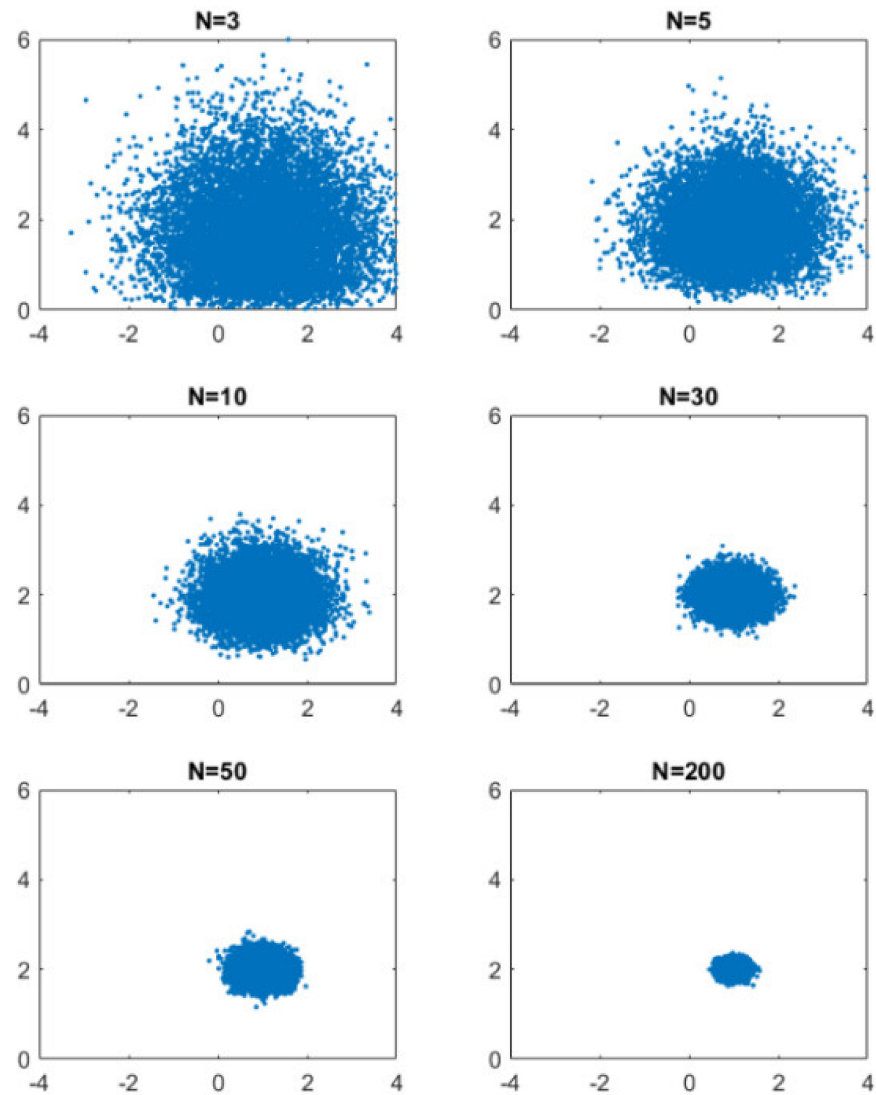$$\text{Prob}\left(Ks'_m > \sigma_t\right) = p$$



**Figure A6.** Joint distribution of samples' means and standard deviations.

If the mean is known, then this formula can be used to establish a tolerance interval with $K$-coverage and confidence $p$, as shown in the previous section. As discussed above, if the true standard deviation is known, the confidence becomes $p = 1.0$. However, if neither the true mean nor the standard deviation is unknown, then different samples will result in different sizes of the tolerance interval, thereby resulting in different coverage $\alpha$. That is, for every sample of the mean and standard deviation, a different $K$-coverage parameter would be required to have the same $\alpha$ coverage. Unlike the case presented in Appendix A.5, in which $K$ must compensate for uncertainties in the mean value only, in this case, $K$ must compensate for uncertainties in both the estimated mean and the standard deviation.

Figure A7 illustrates this scenario, in which samples of the mean and standard deviation are shown for the case of a normal distribution with $\mu_t = 1$, $\sigma_t = 2$, as indicated by a red point. All values are calculated based on a small number of samples: $N = 10$. The purple line traces the values that satisfy $z_m + Ks_m = z_t^H$ with $K = 1.65$, namely, $z_m + 1.65s_m = z_t^H = 4.3$, which corresponds to a coverage of 95%. All samples above the line will result in tolerance intervals with upper threshold values that are larger than the value $z_t^H$, which is the mini-

mum required to have 95% coverage. All values below the line will result in lower upper thresholds and hence coverage that is lower than 95%. Consider for example the black point (with $z_m = 0.25$ and $s_m = 1.5$), which requires a *K* value of 2.7 to reach a 95% coverage. Thus, the goal of the tolerance interval estimation reduces to finding the minimum value of *K* to ensure, with 95% probability, that the estimated mean and standard deviation will produce an upper threshold value greater than $z_t^H$.
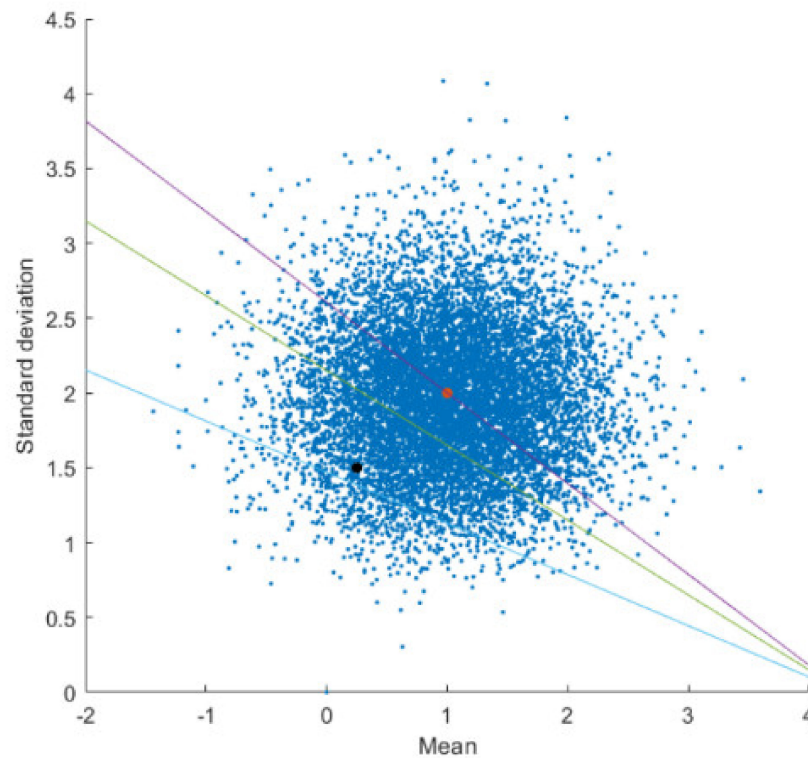


**Figure A7.** Joint distribution of mean values and standard deviation.

First, consider the purple line that passes through the true mean $\mu_t$ and the true standard deviation $\sigma_t$. The points above that line are numerically calculated to be 46% of the total, which is not exactly equal to 50%, because the distribution of standard deviation values is not symmetric with low values of *N*. Above $N = 50$, the distribution of the standard deviation approaches the symmetric normal shape. Next, consider the green line that traces the points satisfying the equation $z_m + 2.0s_m = z_t^H$, increasing the value of *K*. The number of points above this line is increased to approximately 88%, meaning that the confidence associated with the upper threshold value of the tolerance interval increased from 46 to 88% by a moderate increase in the *K* value. This implies that, by increasing the value of *K*, one can have higher confidence that a single estimate of the sample mean and standard deviation can produce an upper threshold for the 95% tolerance interval. This exercise can be repeated by gradually increasing the *K* value until 95% of the points are above the line, which is found to be 2.9 for this example. Fortunately, the joint distribution of $z_m$ and $s_m$ can be analytically derived, allowing one to find the distribution of *K* values, which can be integrated numerically to obtain the minimum value required to reach a probability of 95%. This distribution is called the *non-central t-distribution*, which is discussed below [1,2].

Using the two estimators for mean and standard deviation, define the following estimator:

$$\eta = \frac{z_m - \mu_t}{s_m / \sqrt{N}}. \tag{A24}$$

This estimator is inspired by the quantity appearing in the exponent of Equation (A14), by replacing $\sigma_t$ by $s_m$. This estimator recognizes that the true value for the standard

deviation is unknown, so it is reasonable to try the sample standard deviation. This quantity has a *t*-distribution [42], which looks very similar to the normal distribution, with the distinction of only practical relevance for small values of *N*. For low values of *N*, the *t*-distribution appears to be wider than the normal distribution, with heavier tails. The *K*-coverage parameter for a given portion of the distribution is expected to be larger than that calculated from the normal distribution. Repeating the previous analysis from Appendix A.5,

$$z_m + K_m s_m > \mu_t + K\sigma_t \rightarrow K_m > \frac{K + \frac{1}{\sqrt{N}}\left(\frac{\mu_t - z_m}{\sigma_t/\sqrt{N}}\right)}{\left(\frac{s_m}{\sigma_t}\right)},$$

leading to the following minimum value of the estimator $K_m$:

$$K_m = \frac{K + \frac{1}{\sqrt{N}}\left(\frac{\mu_t - z_m}{\sigma_t/\sqrt{N}}\right)}{\left(\frac{s_m}{\sigma_t}\right)},$$

which ensures with 100% probability that the tolerance limit $(-\infty, z_m + K_m s_m)$ covers at least $\alpha\%$ of the population, where $\alpha$ represents the exact coverage obtained with the true parameters $\mu_t$, $\sigma_t$ and $K$. The distribution of the $K_m$ values is called the *non-central t-distribution* [43], as shown in

Figure 8 for the case of $\mu_t = 1.0$, $\sigma_t = 2.0$, and $K = 1.65$, corresponding to 95% coverage, shown as a red vertical line. As demonstrated above, if the objective is to ensure 95% coverage with a given probability $p\%$, then the $K$ value must be increased until the area under the curve is at least $p\%$.
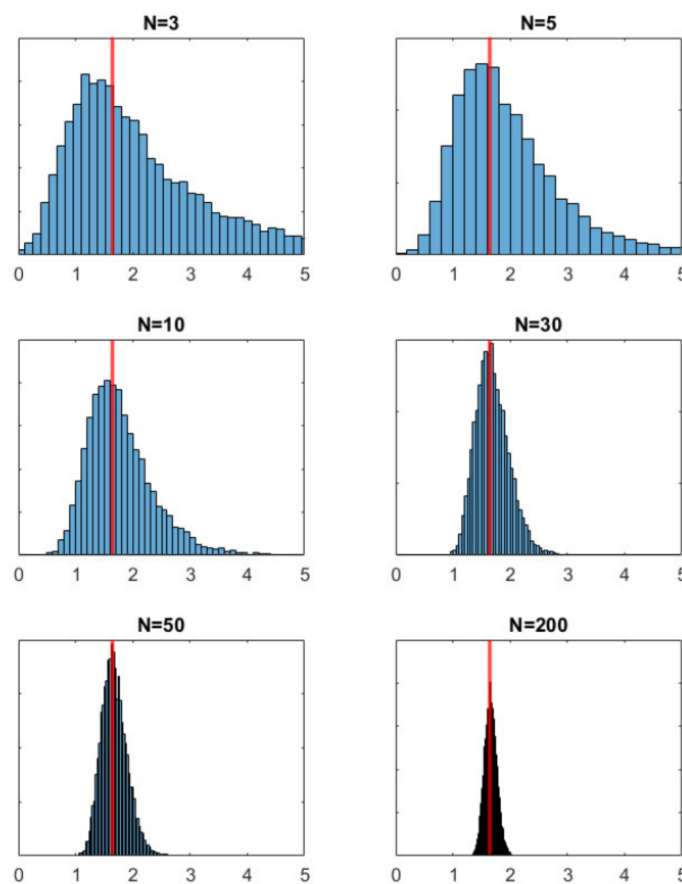


**Figure A8.** Noncentral t-distribution.

## References

1.  Fisher, R.A. *Contributions to Mathematical Statistics*; Shewhart, W.A., Ed.; J. Wiley and Sons, Inc.: New York, NY, USA, 1950.
2.  Hotel, P.G. *Introduction to Mathematical Statistics*, 4th ed.; J. Wiley and Sons, Inc.: New York, NY, USA, 1971.
3.  Jaynes, E.T. Confidence Intervals vs. Bayesian Intervals. In *Harper and Hooker*; Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science: D. Reidel Publishing Company: Dordrecht, The Netherlands, 1976; Volume II, pp. 175–257.
4.  Samaniego, F.J. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2010.
5.  Stanton, J.M. *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*; The Guilford Press: New York, NY, USA, 2017.
6.  Bolstad, W.M. *Introduction to Bayesian Statistics*; Wiley: Hoboken, NJ, USA, 2017.
7.  Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Science: Berlin/Heidelberg, Germany, 2006.
8.  Wakefield, J. *Bayesian and Frequentist Regression Methods*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2013.
9.  Thornber, H. *An Autoregressive Model: Bayesian Versus Sampling Theory Analysis*; Department of Economics, University of Chicago: Chicago, IL, USA, 1965.
10. Broadhead, B.L.; Williams, M.L.; Wagschal, J.J. Generalized Linear Least-Squares Adjustment, Revisited. *J. ASTM Int.* **2006**, *3*, 1–8. [CrossRef]
11. Burrows, G.L. Statistical Tolerance Limits—What are They? *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1963**, *12*, 2. [CrossRef]
12. Meeker, W.Q.; Hahn, G.J.; Escobar, L.A. *Statistical Intervals: A Guide for Practitioners and Researchers*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2017.
13. OECD. *Overview of Approaches Used to Determine Calculational Bias in Criticality Safety Assessment, State-of-the-Art Report (Part 1)*; Nuclear Energy Agency: Paris, France, 2013.
14. Wilks, S.S. Determination of Sample Sizes for Setting Tolerance Limits. *Ann. Math. Stat.* **1941**, *12*, 91–96. [CrossRef]
15. Wilks, S.S. Statistical Prediction with Special Reference to the Problem of Tolerance Limits. *Ann. Math. Stat.* **1942**, *13*, 400–409. [CrossRef]
16. Wald, A.; Wolfowitz, J. Tolerance Limits for a Normal Distribution. *Ann. Math. Stat.* **1946**, *17*, 2. [CrossRef]
17. Patel, J. Tolerance Limits: A Review. *Commun. Stat.-Theory Methods* **1986**, *15*, 2719–2762. [CrossRef]
18. Commission, U.N.R. Emergency Core Cooling Systems, Revisions to Acceptance Criteria. *Fed. Regist.* **1988**, *53*, 35996–36005.
19. Pal, L.; Makai, M. Remarks on Statistical Aspects of Safety Analysis of Complex Systems. *arXiv* **2003**, arXiv:physics/0308086.
20. Guba, A.M.; Makai, L.P. Statistical Aspects of Best Estimate Method—I. *Reliab. Eng. Syst. Saf.* **2003**, *80*, 217–232. [CrossRef]
21. Glaeser, H. GRS Method for Uncertainty and Sensitivity Evaluation of Code Results and Applications. *Sci. Technol. Nucl. Install.* **2008**, *2008*, 1–7. [CrossRef]
22. Loukusa, H.; Tulkki, V. Determination of Tolerance Limits for Fuel Assembly Fission Gas Release Summary Statistics. *Nucl. Eng. Des.* **2020**, *358*, 110438. [CrossRef]
23. Mishra Datta, D.S.; Rajest, S.S. Quantification of Tolerance Limits of Engineering System Using Uncertainty Modeling for Sustainable Energy. *Int. J. Intell. Netw.* **2020**, *1*, 1–8.
24. Aures, A.; Berner, N.; Pautz, A.; Zwermann, W. Uncertainty and Sensitivity Analysis of PWR Mini-Core Transients in the Presence of Nuclear Data Uncertainty Using Non-Parametric Tolerance Limits. *Ann. Nucl. Energy* **2020**, *137*, 107146. [CrossRef]
25. Porter, N. Wilks' Formula Applied to Computational Tools: A Practical Discussion and Verification. *Ann. Nucl. Energy* **2019**, *133*, 129–137. [CrossRef]
26. Han, S.; Kim, T. Numerical Experiments on Order Statistics Method Based on Wilks' Formula for Best-Estimate Plus Uncertainty Methodology. *J. Environ. Manag.* **2019**, *235*, 28–33. [CrossRef]
27. Mui, T.; Kozlowski, T. Confirmation of Wilks' Method Applied to TRACE Model of Boiling Water Reactor Spray Cooling Experiment. *Ann. Nucl. Energy* **2018**, *117*, 53–59. [CrossRef]
28. Park Zimmer, Z.D.; Mathew, T. Tolerance Limits under Normal Mixtures: Application to the Evaluation of Nuclear Power Plant Safety and to the Assessment of Circular Error Probable. *Comput. Stat. Data Anal.* **2016**, *103*, 304–315. [CrossRef]
29. Zhang, H.; Szilard, R.; Zou, L.; Zhao, H. Comparisons of Wilks' and Monte Carlo Methods in Response to the 10CFR50. 46 (c) Proposed Rulemaking. In Proceedings of the NUTHOS-11, Gyeongju, Korea, 9–13 October 2016; pp. 1–12.
30. Shockling, M. Non-Parametric Order Statistics: Providing Assurance of Nuclear Safety. In Proceedings of the NURETH-16, Chicago, IL, USA, 30 August–4 September 2015; pp. 2477–2490.
31. Hong, I.S.; Oh, D.Y.; Kim, I.G. *Generic Application of Wilks' Tolerance Limit Evaluation Approach to Nuclear Safety*; NEA/CSNI/R(2013)8/PART2; Nuclear Energy Agency of the OECD (NEA): Paris, France, 2013.
32. Martin, R.P.; Nutt, W.T. Perspectives on the Application of Order-Statistics in Best-Estimate Plus Uncertainty Nuclear Safety Analysis. *Nucl. Eng. Des.* **2011**, *241*, 274–284. [CrossRef]
33. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
34. Williams, M.L.; Rearden, B.T. SCALE-6 Sensitivity/Uncertainty Methods and Covariance Data. *Nucl. Data Sheets* **2008**, *109*, 2796–2800. [CrossRef]
35. Briggs, J.B. *International Handbook of Evaluated Reactor Physics Benchmark Experiments*; NEA/NSC/DOC: Paris, France, 2006.

36. Larson, N.M. *Updated Users' Guide for SAMMY: Multilevel R-Matrix Fits to Neutron Data Using Bayes' Equations*; No. ORNL/TM-9179/R4; Oak Ridge National Lab.: Oak Ridge, TN, USA, 1998.
37. Bayes, T. LII. An Essay towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philos. Trans. R. Soc. Lond.* **1763**, *53*, 370–418.
38. Lu, J. Tolerance Interval for Multiple Regression. *J. Farm Econ.* **1960**, *42*, 910–920. [CrossRef]
39. Galambos, J.; Simonelli, I. *Bonferroni-Type Inequalities with Applications*; Springer: New York, NY, USA, 1996.
40. Robbins, H.E. On the Measure of a Random Set. *Ann. Math. Stat.* **1944**, *15*, 70–74. [CrossRef]
41. Arnold, B.C.; Balakrishnan, N.; Nagaraja, H.N. *A First Course in Order Statistics*; SIAM: Philadelphia, PA, USA, 2008.
42. Gosset, W.S. The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1–25.
43. Lenth, R.V. Cumulative Distribution Function of the Non-central t Distribution. *J. R. Stat. Soc. Ser. C* **1989**, *38*, 185–189.