Reliability Engineering and System Safety 180 (2018) 201–210

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Distinguishing between model- and data-driven inferences for high reliability statistical predictions

Lauren Hund[*], Benjamin Schroeder, Kellin Rumsey, Gabriel Huerta

Sandia National Laboratories, 1515 Eubank SE, Albuquerque NM 87123, United States

ARTICLE INFO

Keywords:
Extreme value modeling
Model uncertainty
Margin
Risk assessment
Quantification of margins and uncertainties

ABSTRACT

Estimating the tails of probability distributions plays a key role in complex engineering systems where the goal is characterization of low probability, high consequence events. When data are collected using physical experimentation, statistical distributional assumptions are often used to extrapolate tail behavior to assess reliability, introducing risk due to extrapolation from an unvalidated (statistical) model. Existing tools to evaluate statistical model fit, such as probability plots and goodness of fit tests, fail to communicate the risk associated with this extrapolation. In this work, we develop a new statistical model validation metric and relate this metric to engineering-driven model validation metrics. The metric measures how consistent the parametric tail estimates are with a more flexible model that makes weaker assumptions about the distribution tails. An extreme-value based generalized Pareto distribution is used for the more flexible model. Models are updated using a Bayesian inference procedure that defaults to reasonably conservative inferences when data are sparse. Properties of the estimation procedure are evaluated in statistical simulation, and the effectiveness of the proposed metrics relative to the standard-of-practice statistical metrics is illustrated using a pedagogical example related to a real, but proprietary, engineering example.

1. Introduction

In complex engineering systems, it is often of interest to characterize the likelihood of high consequence but rare events. Examples include probabilistic risk assessment (PRA) applied to nuclear power plants and quantification of margins and uncertainties (QMU) applied to nuclear weapons [1]. Both PRA and QMU involve decomposition of the system reliability into events (using, for example, fault trees or reliability block diagrams) and rolling up event-level likelihoods to the system-level. In such analyses, high system reliability at the top (system) level is achieved by specifying even higher reliability for sub-system level events. To characterize performance at the sub-system (e.g., event or component) level in PRA and QMU, continuous performance measures related to sub-system functionality are often modeled using probability distributions. However, with high reliability requirements, characterizing sub-system reliability requires accurately estimating tails of probability distributions [2,3]. Hence, characterizing the tails of probability distributions plays a key role in low probability, high consequence fields like PRA [4–6] and QMU [2].

Probability distributions for continuous measures of sub-system functionality are often estimated by collecting data from experiments. That is, given a set of experimental data, the user selects a model (probability distribution) for the data; then, the parameters of the probability distribution are estimated from the data using statistical methods. Reliability is estimated from the tails of the fitted probability distributions. However, with high reliability requirements, the experimental sample size is often too small to accurately characterize distribution tails without strong modeling assumptions. Consequently, parametric distributional assumptions (e.g., normality) are often used to extrapolate the extreme tail behavior of the underlying probability distribution. The risk associated with extrapolating distribution tail behavior is known to be large [2,4,7–11], because tail estimates are extremely sensitive to the selected statistical model. Tail fitting methods such as statistical extreme value theory, which use a subset of the data most informative of the distribution's tail behavior, are an approach to providing estimates of low probability events [12,13], while avoiding assumptions about the shape of the high probability regions of the distribution. In sparse data situations, alternative methods to traditional statistical analysis can inform reliability. For instance, evidence theory is commonly applied to incorporate epistemic uncertainty with limited information [14–16]. Utilizing auxiliary information such as historical data and expert elicitation in sparse data situations can also improve reliability assessments [17]. However, in practical applications, the analyst must decide when enough

[*] Corresponding author.
E-mail address: lbhund@sandia.gov (L. Hund).

https://doi.org/10.1016/j.ress.2018.07.017
Received 9 February 2018; Received in revised form 10 July 2018; Accepted 12 July 2018

information exists to use statistical approaches instead of these more expert-driven methods.

The epistemic uncertainty associated with probability distribution selection is a known issue in many engineering applications, including PRA and QMU [1,4], as well as lifetime and accelerated testing [18,19], input distribution specification for computational simulation modeling (e.g. [20,21]) and statistical quality control (e.g. [22]). However, existing statistical model fit tools, specifically goodness of fit tests and probability plots, are inadequate for assessing and communicating the validity of model assumptions. Interpretation of probability plots can be subjective [23]; is sensitive to outliers [24]; and, most importantly for tail characterization, highlights the center of the distribution rather than the tails [11]. Distributional goodness of fit tests are even more problematic; goodness of fit tests can detect evidence of lack of model fit, but do not provide evidence that a model is a good fit for the data, even though the end-user is typically aiming to prove the latter [24,25]. Accepting a parametric model just because the model cannot be disproved using data reflects a logical fallacy [26] and “can lead to disastrous results” when making inferences about tail behavior [8]. However, we are unaware of other commonly-used alternatives to goodness of fit tests and probability plots for assessing statistical model adequacy.

The goal of this article is to propose validation metrics for communicating model form risk when considering whether to estimate distribution tails from statistical models to characterize rare event probabilities. To improve upon statistical goodness of fit tests for assessing model fit, we develop metrics that are directly tied to validation of physics-based computational simulations [27–29]. While the precise definition of model validation is debatable, we will define validation as, “the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model” [29] (this goal is often never exactly achievable in practice, and validation consists of analyses to assess validity, albeit imperfectly). In short, validation examines whether a model’s prediction accuracy is good enough for the intended application. As in computational simulation, statistical model validation should show model accuracy for the intended application, *e.g.*, for tail characterization, provide evidence that the model accurately captures the tail behavior. Statistical extreme value theory is concerned with making inferences about extreme observations [30] and thus provides a suite of tools relevant for reliability analysis. Instead of strictly using statistical extreme value methods to make tail extrapolations, we use it to construct validation metrics to communicate risk associated with tail characterization. To complement the validation metrics, we provide a graphical aid that highlights model fit in the tails, such that the proposed tools parallel probability plots and goodness of fit tests. Our approach addresses the limitations of the existing methods by directly explicating the risk being absorbed in model-based extrapolation. These new metrics can help distinguish between inferences that are data-driven, i.e., driven by the observed data values, versus model-driven, i.e., driven by the selected statistical model for the data.

This article is structured as follows. In Section 2, we provide an example problem from QMU for nuclear weapons reliability, highlighting where existing statistical tools can break down for tail estimation. In Section 3, we propose a novel approach for evaluating whether the distributional assumptions underlying the methods are adequate for the intended use, rooting our methods in engineering-driven model validation concepts. Lastly, in Section 4, we illustrate the proposed methods on the QMU example.

2. Motivating example: parametric tolerance intervals for QMU

To define the tail characterization problem, we consider estimation of an extreme percentile of a distribution; in practice, this percentile may correspond to a reliability or safety requirement. To characterize confidence (or, conversely, sampling uncertainty) in percentile estimates, we use statistical tolerance intervals [31]. Statistical tolerance intervals are an appropriate tool for this application, because they map continuous performance measures back to the question of interest, namely how confident are we that the reliability is at least r?

Without loss of generality, we consider percentile estimation for the upper-tail of a probability distribution. Further, we assume all performance measures are independent and identically distributed (*i.i.d.*), a common (but often violated) assumption in practice (see the Discussion for more information). Let $X = \{X_1, X_2, ..., X_n\}$ denote an *i.i.d.* random sample from a population. Given a set of data X, we can estimate the distribution of X, denoted $\mathcal{F}_X$, and then calculate percentiles from this distribution. A percentile of a distribution Q_r is defined as the value of the distribution for which r percent of observations are below this value. Mathematically, Q_r is defined as $P(X < Q_r) = r$.

In practice, the distribution $\mathcal{F}_X$ and percentiles Q_r are estimated from finite samples and contain sampling uncertainty. We denote estimates of $\mathcal{F}_X$ and Q_r as $\widehat{\mathcal{F}}_X$ and $\widehat{Q}_r$. Statistical tolerance intervals can be used to quantify sampling uncertainty in a percentile $\widehat{Q}_r$. Tolerance intervals can be one- or two-sided; we consider only one-sided tolerance intervals in this paper, often referred to as tolerance bounds, because, in practice, we are typically trying to find a bound on a performance measure. Mathematically, an upper one-sided $(r, 1-\alpha)$ tolerance bound $\widehat{Q}_{r,\alpha}$ is defined as:

$$P_{\mathbf{X}}\{P_X(X \leq \widehat{Q}_{r,\alpha}|\mathbf{X}) \leq r\} = 1-\alpha, \tag{1}$$

where r is a percentile, α is the confidence in the estimate of the percentile. Heuristically, a one-sided upper statistical tolerance bound is simply an upper confidence bound on a percentile.

Extrapolation based on a parametric statistical model is commonly used to estimate percentiles and tolerance intervals, i.e. $X \sim \mathcal{F}_X$, where $\mathcal{F}_X$ is an assumed probability model for X. In engineering applications, common choices for $\mathcal{F}_X$ are the normal, Weibull, and lognormal distributions [4,31]. Parameters of the distributions are estimated using statistical methods, such as maximum likelihood estimation or method of moments. After fitting the model, percentiles are estimated based on quantiles of the fitted model $\widehat{\mathcal{F}}_X$.

2.1. Launch safety device application

We consider an example motivated by quantification of margin and uncertainty (QMU), a framework for evaluating system-level nuclear weapon performance (e.g., reliability and safety) by rolling-up component level margin estimates [1,32]. In QMU, a quantitative measure of system performance is compared to a performance requirement. For instance, given a reliability requirement r, the estimated rth percentile of the performance distribution ($\widehat{Q}_r$) must be sufficiently far from the requirement τ, accounting for uncertainty (Fig. 1). Following [32], margin can be defined as the distance between the percentile estimate $\widehat{Q}_r$ and requirement τ; (sampling) uncertainty in the percentile estimate is measured through a tolerance interval $\widehat{Q}_{r,\alpha}$. We assume the requirement τ is fixed and known. As the demand for QMU using experimental data increases, a common question is when enough data exists to reliably estimate a percentile Q_r and a tolerance interval $\widehat{Q}_{r,\alpha}$. More precisely, when can we feel confident that a statistical model $\mathcal{F}_X$ is valid for estimating a percentile and tolerance interval?

As an example, we consider a hypothetical launch safety device on a missile. The launch safety device is an electromechanical switch that operates upon sensing a specific environmental input. The device closes electrical contacts upon receiving the correct input. Hence, the device acts as a safety mechanism by failing to close unless the correct input is received; however, the switch must close reliably within a certain time window for the downstream devices to operate appropriately and achieve system success. Hence, the reliability of this device directly informs the overall system reliability. Suppose the component has a requirement to close within 23.5 s of launch with 99.5% reliability; that