SANDIA REPORT
SAND2017-12349
Unlimited Release
Printed November 2017

Evaluation of a Class of Simple and Effective Uncertainty Methods for Sparse Samples of Random Variables and Functions

Vicente Romero, Matthew Bonney, Benjamin Schroeder, V. Gregory Weirs

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods/

SAND2017-12349
Printed November 2017
Unlimited Release

Evaluation of a Class of Simple and Effective Uncertainty Methods for Sparse Samples of Random Variables and Functions

Vicente Romero—V&V, UQ, Credibility Processes Dept. 1544
Matthew Bonney—student intern, Component Science & Mechanics Dept. 1556
Benjamin Schroeder—V&V, UQ, Credibility Processes Dept. 1544
V. Gregory Weirs—Multiphysics Applications Dept. 1446

Sandia National Laboratories
P. O. Box 5800
Albuquerque, New Mexico 87185-MS0828

Abstract

When very few samples of a random quantity are available from a source distribution of unknown shape, it is usually not possible to accurately infer the exact distribution from which the data samples come. Under-estimation of important quantities such as response variance and failure probabilities can result. For many engineering purposes, including design and risk analysis, we attempt to avoid under-estimation with a strategy to conservatively estimate (bound) these types of quantities—without being overly conservative—when only a few samples of a random quantity are available from model predictions or replicate experiments. This report examines a class of related sparse-data uncertainty representation and inference approaches that are relatively simple, inexpensive, and effective. Tradeoffs between the methods' conservatism, reliability, and risk versus number of data samples (cost) are quantified with multi-attribute metrics used to assess method performance for conservative estimation of two representative quantities: central 95% of response; and 10^{-4} probability of exceeding a response threshold in a tail of the distribution. Each method's performance is characterized with 10,000 random trials on a large number of diverse and challenging distributions. The best method and number of samples to use in a given circumstance depends on the uncertainty quantity to be estimated, the PDF character, and the desired reliability of bounding the true value. On the basis of this large data base and study, a strategy is proposed for selecting the method and number of samples for attaining reasonable credibility levels in bounding these types of quantities when sparse samples of random variables or functions are available from experiments or simulations.

TABLE OF CONTENTS

FIGURES

TABLES

1. INTRODUCTION

When very few samples of a random quantity are available from a source distribution or probability density function (PDF) of unknown shape, it is usually not possible to accurately infer the PDF from which the data samples come. Thus, a significant component of epistemic uncertainty exists concerning the source distribution of random or aleatory variability. The likely error that accompanies sparse sampling has a bias toward underestimating the true variability of the source; the variance calculated from just a few samples will usually be less than the variance calculated from a large number of samples, for many common PDF types. This unconservative bias is undesirable for many engineering purposes. If a structure or pressure-vessel model were perfect in every other way, use of the model with sparse samples of the random-data inputs would likely underestimate the (strength or displacement) response variance of the real system. In design and risk analysis one would normally want to avoid such variance underestimation. The calculated mean from sparse samples will also likely have significant error, which also contributes to uncertainty and risk in response estimation.

Therefore, we desire to conservatively estimate (bound) random quantities—without being overly conservative—when only a few samples are available from testing or simulation. This paper considers two representative uncertainty quantities of interest associated with a random quantity being sampled: A) the 2.5 to 97.5 percentile "central 95%" range of the sparsely sampled PDF; and B) 10^{-4} exceedance probability (EP) associated with a tail of the PDF integrated beyond a specified limit. Appropriate UQ treatments for small tail probabilities like this are important for assessing performance and safety margins in design, risk, and reliability analysis using sparse samples of experimental data or model predictions. The quantity in category A is important e.g. for assessing how well model results match or bound a representative or adequate span of the experimental data and its uncertainty for model calibration or validation purposes.

Because accurate description of the aleatory distribution is not possible with relatively sparse samples, the set of related sparse-data UQ methods in this section are used within a strategy of conservative treatment of the aleatory and epistemic uncertainties involved, while at the same time attempting to avoid being overly conservative. Of course, the demarcation between what is viewed as appropriate conservatism versus inappropriate over-conservatism in a particular circumstance is usually not sharply identifiable in absolute terms and is dependent on subjective judgments and often un-crisply defined relationships and objectives regarding risk-benefit-cost tradeoffs. Nonetheless, this paper introduces and explores some of the tradeoffs involved and begins to illustrate methodology toward arriving at "conservative but not overly conservative" treatments of uncertainty in cases of sparse replicate data from random variables and functions.

The sparse-data strategies and conclusions also apply for sparse model simulation results where random-variable inputs are being sampled and the affordable number of model evaluations is very limited to, say, 5 Latin-Hypercube Monte Carlo samples. Similar considerations hold when model inputs are experimental or synthetic samples of random functions, like various stress-strain curves from replicate material tests [Romero et al., 2014, 2015, 2017a], or realizations of random fields like spatial variation of surface roughness or material properties.

Section 2 describes a class of related sparse-data uncertainty representation and inference approaches that are relatively simple, inexpensive, and effective under typical realistic circumstances established in sections 3 and 4. Section 3 investigates method performance for conservatively bounding the central 95% range of a sparsely sampled PDF. Section 4 investigates performance for bounding 10^{-4} tail probabilities. Corresponding performance metrics and analysis are employed to identify the best performing methods and sample sizes N to use for credible bounding estimates for these two uncertainty quantities. Sample sizes of N= 2, 4, 10, 20 were investigated. Different methods are found to be best for each quantity. The number of samples for credible bounding with these best methods is relatively low for both quantities—on the order of four samples. Section 5 summarizes the main points from this large study.

2. A CLASS OF RELATED SPARSE-DATA UQ METHODS: TOLERANCE INTERVALS AND THEIR EQUIVALENT NORMAL DISTRIBUTIONS, ENSEMBLE OF NORMALS AND THEIR SUPERDISTRIBUTIONS

Tolerance Intervals (TIs) and their Equivalent Normal (EN) Distributions

Tolerance Intervals are a simple way to approximately account for the epistemic sampling uncertainty introduced from finite samples of a random variable. When very few samples are available, it is usually not possible to accurately infer the source probability distribution from which the samples came. Instead, a viable strategy is to attempt to be conservative, but not overly conservative, in estimating e.g. the central 95% range of response between the distribution's 2.5 and 97.5 percentiles. This range could be useful for many envisioned purposes, including engineering design and model calibration and validation. The Tolerance Interval method is an easy and economical way to obtain such bounding estimates, as explained next.

TIs are parameterized by two user-prescribed levels: one for the desired "coverage" proportion of a distribution and one for the desired degree of statistical "confidence" in covering or bounding at least that proportion. For instance, a 95%coverage/90%confidence TI (95%/90% TI, 95/90 TI, or 0.95/0.90 TI) prescribes lower and upper values of a range said to have at least 90% odds that it covers or spans 95% of the "true" probability distribution from which the random samples were drawn—if they were drawn from a Normal distribution. Promising findings from TI robustness investigations on a large variety of other PDF types will be discussed later.

As Figure 2-1 illustrates, a X%coverage/Y%confidence TI is constructed by multiplying the calculated standard deviation $\tilde{\sigma}$ of the data samples by an appropriate factor f to create a TI of total length $2f\tilde{\sigma}$. The interval is centered about the calculated mean $\tilde{\mu}$ of the samples, so the interval's top and bottom endpoints are defined by

$$\tilde{\mu} \pm f\tilde{\sigma}. \tag{2.1}$$

The factor f depends on the parameters X, Y, and the number N of samples, and can be obtained from look-up tables (e.g. [Hahn & Meeker, 1991], [Montgomery & Runger, 1994]) or can be calculated from formulas (e.g. [Howe, 1969]) or associated functions available in mathematics and statistics software packages, e.g. [Young, 2010].

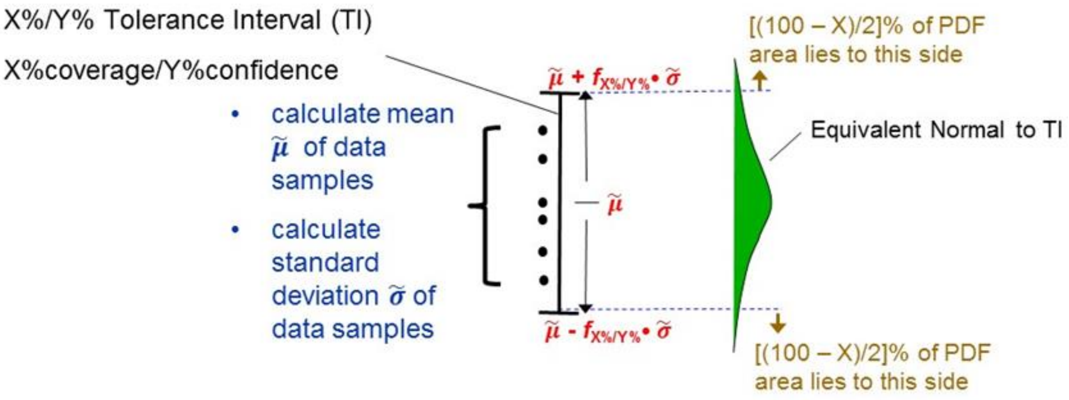

X%/Y% Tolerance Interval (TI)
X%coverage/Y%confidence
• calculate mean $\tilde{\mu}$ of data samples
• calculate standard deviation $\tilde{\sigma}$ of data samples
$\tilde{\mu} + f_{X\%/Y\%} \cdot \tilde{\sigma}$
$\tilde{\mu}$
$\tilde{\mu} - f_{X\%/Y\%} \cdot \tilde{\sigma}$
[(100 – X)/2]% of PDF area lies to this side
Equivalent Normal to TI
[(100 – X)/2]% of PDF area lies to this side

Figure 2-1. Scalar data samples, Tolerance Interval, and Equivalent Normal

Table 2-1 shows factors for constructing 95%/90% TIs. The table and Figure 2-2 reveal that 95/90 TI size decreases quickly with the number of data samples. The rate of uncertainty decrease per added sample has a knee at 4 to 6 samples, with the rate of decrease being fairly small after 8 samples. The tolerance interval has an asymptotic standard deviation multiplier of 1.96 for an infinite number of samples. In this case, the TI corresponds to the exact 95% central percentile range of a Normal PDF with $\tilde{\mu}$ and $\tilde{\sigma}$ exact from ∞ samples.

Although TI construction is based on sparse-sampling theory for Normal populations, investigations by Romero et al. (2013a,b) showed that TIs provide reliably conservative estimates of the combined epistemic and aleatory uncertainties associated with very sparse samples for Normal, uniform, and right-triangular PDFs, and four other distributions resulting from convolving various combinations of these three PDF types. The TI approach is also much easier to use than the four other sparse-data methods investigated.

However, it was also found that 95/90 TIs can egregiously exaggerate the true variability when very few samples are involved. The approach of Pradlwarter and Schuëller (2008) usually exaggerates the true variability significantly less than 95/90 TIs when few samples are available, but has significantly lower success rates of central 95% capture and is somewhat more involved to implement. Its performance remains to be broadly tested and characterized, but other indications from [Winokur et al., 2016] are also favorable. The common practice of simply fitting the random data with a normal distribution was found to have substantial risk of underestimating the true variability of the population being sampled—even if the sampled distribution is Normal.

The TI method has been used in several engineering applications at Sandia National Laboratories (e.g. Romero et al., 2014, 2015, Jamison et al., 2016)[1] but has not yet been verified on highly non-linear application problems and non-symmetric distribution shapes because of the computational expense of the application models and because of other time and resource constraints in those projects. Section 3 involves testing on a highly non-linear solid mechanics problem with 140 non-Normal distribution shapes and analytic Normal, wide-tailed 5 degree of freedom T, highly skewed Log-Normal, and highly-skewed and wide-tailed Weibull distributions.

TI robustness for sparse data has been established on non-symmetric gamma distributions in [Bhachu et al., 2016]. They found that the best performing parametric and non-parametric Bootstrap methods they studied required order 20 samples or more to be competitive with or sometimes surpass the accuracy of the TI methods. Bootstrap methods are also more complicated.

Bayesian methods also have substantial difficulty with sparse data. Recent Sandia research [Romero, Weirs, Schroeder, et al., 2018] has shown that Bayesian approaches for dealing with very sparse data are significantly more complicated and not demonstrably more effective than the UQ methods in this report. However, the simple TI approach loses reliability and efficiency advantages over many of the other methods mentioned when data becomes less sparse. Then more sophisticated data-fitting methods such as in [Pradlwarter & Schuëller, 2008] and the

[1] Applications in calibration and validation of a device structural dynamics model, a solid propellant combustion model, and several radiation-damaged electronics models have also been conducted by the first author with the TI UQ method, but references are not publicly available.

Table 2-1. 95%/90% Tolerance Interval Factors *f* (standard deviation multipliers) vs. # of samples of random quantity. (Selected results computed from formulas found in [Howe, 1969].)

# samples	$f_{0.95/0.90}$
2	18.56
3	6.95
4	4.99
5	4.19
6	3.76
8	3.29
10	3.04
20	2.57
30	2.42
40	2.34
∞	1.96

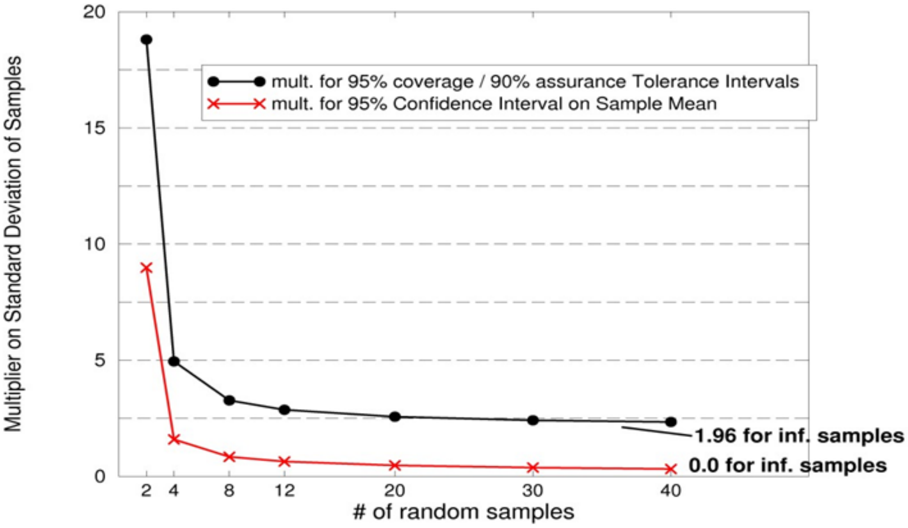


Figure 2-2. Multiplier *f* on calculated standard deviation used to form 0.95/0.90 tolerance interval ranges vs. number of random samples. (Figure reproduced from [Romero et al., 2011], ignore confidence interval curve.)

single- and multi- distribution Bayesian approaches surveyed in [Romero, Schroeder, et al., 2017] may be preferable to the Normal PDF based single- and multi- distribution approaches investigated in the present report.

For other uncertainty representation and analysis purposes, 95/90 and 95/95 “Equivalent Normal” (EN) PDFs are constructed from TIs such that the EN’s 0.025 and 0.975 quantiles coincide with the end points of the TIs. This is portrayed in Figure 2-1. A TI and its Equivalent Normal have the same mean $\tilde{\mu}$ calculated from the data samples. As a Normal distribution, the 0.025 and 0.975 quantiles of the EN occur at 1.96 times the EN’s standard deviation (σ_EN). Equating the TI half-length to 1.96•σ_EN yields the following equation for calculating the Equivalent Normal’s standard deviation.

$$\sigma\_\mathrm{EN} = f_{X\%/Y\%} \bullet \tilde{\sigma}/1.96 \tag{2-2}$$

By their construction, 0.95/0.90 Equivalent Normal PDFs have the same high reliability as 0.95/0.90 TIs that the EN 0.025 and 0.975 quantiles contain the 0.025 and 0.975 quantiles of the true PDFs from which the random samples come (for a large array of PDF types). Furthermore, 95/90 TI-ENs will have even higher reliabilities of capturing extended quantiles like 0.01 and 0.99 of the true PDFs being sampled. Figure 2-3 depicts the basis for this statement. For any number N of data samples, the 99% central range of a 95%/90% TI Equivalent-Normal PDF can be shown to envelope a 99%_coverage/90%_confidence TI determined from TI tables. Therefore, whatever empirical reliability or confidence a 99/90 TI has in bounding the 99% central range of a sampled PDF, a greater reliability of bounding will exist if using the 99% central range of a 95%/90% TI Equivalent Normal.

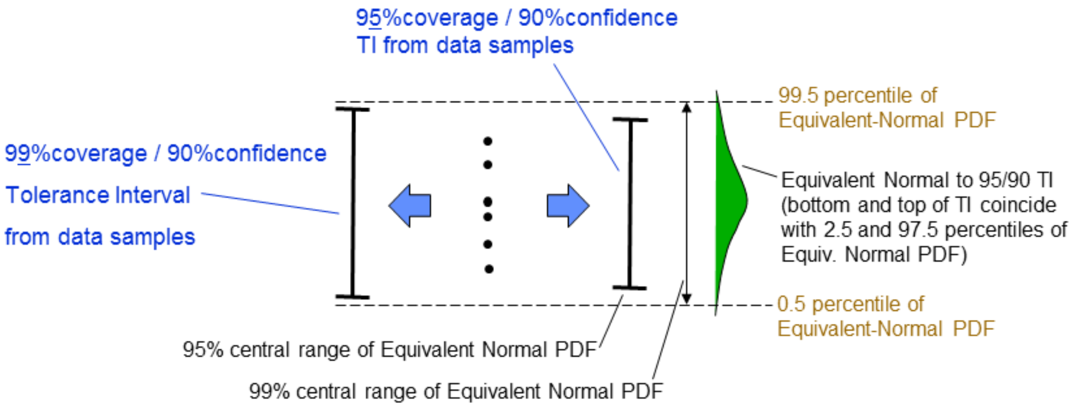

95%coverage / 90%confidence
TI from data samples
99.5 percentile of
Equivalent-Normal PDF
99%coverage / 90%confidence
Tolerance Interval
from data samples
Equivalent Normal to 95/90 TI
(bottom and top of TI coincide
with 2.5 and 97.5 percentiles of
Equiv. Normal PDF)
0.5 percentile of
Equivalent-Normal PDF
95% central range of Equivalent Normal PDF
99% central range of Equivalent Normal PDF

Figure 2-3. Comparison of bounds from a 99/90 TI and the 99% Central Coverage range of a 95/90 TI Equivalent-Normal PDF.

Ensemble of Normals (EON) and associated Superdistribution (SD)

The approaches for dealing with sparse replicate data in this subsection are somewhat more involved than TIs and ENs, but add versatility and are still relatively simple and inexpensive. These methods have a strong relation to Tolerance Intervals as explained in the next subsection. We first discuss the Ensemble of Normal distributions (EON) approach, and then the associated

Superdistribution (SD) approach.

An ensemble of Normal distributions is a useful device for treating sparse replicate data with improved freedom for expressing the uncertainty of probabilities associated with outcomes or ranges of outcomes. EON are constructed as follows.

If N samples are drawn from a Normal distribution *Normal*(μ,σ) where μ and σ are the mean and standard deviation of the distribution, the "sample" mean $\tilde{\mu}$ and the "sample" standard deviation $\tilde{\sigma}$ calculated from the N data samples will usually have error relative to the true mean and standard deviation μ and σ. Distributions of possible values of the true mean and standard deviation can be constructed from the sample mean and standard deviation as follows.

A reasonable candidate μ_i for the value of the true mean can be obtained by drawing a random sample T_i from a Student's T distribution with (N – 1) degrees of freedom (DOF) and using it in the following equation. The T distribution is symmetric about zero and resembles a Normal distribution but has wider tails.

$$\mu_i = \tilde{\mu} + T_i\tilde{\sigma}/\sqrt{N} \tag{2-3}$$

An asymptotically large set of reasonable candidates $[\mu_i]$ is generated from an asymptotically large set $[T_i]$ of samples from an N-1 DOF T-distribution. It can be shown [Miller & Freund, 1985] that the central (1-α)% range of an asymptotically large set or distribution of candidate means $[\mu_i]$, from the distribution's α/2 to (1 - α/2) quantiles, is a (1-α)% "confidence interval" (CI) that will contain the true mean μ exactly (1-α)% of the time. That is, so-produced CIs will successfully contain the true mean in (1-α)% of a very large number of trials, where each trial *X* involves drawing N random samples from the said Normal distribution *Normal*(μ,σ) and using the sample mean and standard deviation $\tilde{\mu}_x$ and $\tilde{\sigma}_x$ to generate a distribution of candidate means $[\mu_i]_x$ and a corresponding CI.

Analogously, a reasonable candidate σ_i for the value of the true standard deviation can be obtained by drawing a random sample χ_i^2 from a (N – 1) DOF Chi-Square distribution and using it in the following equation. The Chi-Square (χ^2) distribution is a non-symmetric distribution that starts at zero and proceeds rightward per the example in Figure A.5.

$$\sigma_i = \tilde{\sigma}\sqrt{(\mathrm{N}-1)/\chi_i^2} \tag{2-4}$$

An asymptotically large set of reasonable candidates $[\sigma_i]$ is generated from an asymptotically large set $[\chi_i^2]$ of samples from an N-1 DOF χ^2-distribution. It can be shown [Miller & Freund, 1985] that the range between the α/2 to the (1 - α/2) quantiles of the asymptotically large distribution or set of candidates $[\sigma_i]$ constitutes a (1-α)% confidence interval that will contain the true standard deviation σ exactly (1-α)% of the time.

Another theoretical result is that the T and χ^2 distributions are independent of each other, so sample means and standard deviations generated are not correlated with each other.

Uncorrelated pairings of samples from the sets $[\mu_i]$ and $[\sigma_i]$ can be used to generate candidate Normal distributions (see Figure 2-4) among which the true distribution *Normal*(μ,σ) may exist.

In practice it is usually not essential that the true Normal PDF *Normal*(μ,σ) lie among the candidate PDFs, as long as a sought analysis quantity like exceedance probability (EP), or a percentile or percentile range from the true distribution, is within or likely bounded by a suitably determined continuous uncertainty band constructed from the generated set of candidate Normals (where 'suitably' is considered later). The odds or reliability of this occurring depend on the sought quantity; on the number of samples N; and on the number of generated candidate Normals. The odds in many practical cases are relatively high as established later with 100 candidate Normals. Even in cases when the distribution drawn-from is highly non-Normal, odds are reasonably high that the EON procedure yields useful bounds on percentiles, percentile ranges, and exceedance probabilities.

To recap, Figure 2-4 and the following steps summarize the EON procedure for data samples from a Normal or non-Normal distribution.

1. Given n_x data samples, compute the sample mean, $\tilde{\mu}_x$, and sample standard deviation, $\tilde{\sigma}_x$.

2. Generate a set of n_r random samples from the T distribution corresponding to $n_x - 1$ degrees of freedom. Refer to this set as $[T_i]$ and the i-th sample in the set as T_i. Similarly generate a set of n_r samples from an $n_x - 1$ DOF χ^2 distribution.

3. Use the samples in $[T_i]$ and $[\chi_i^2]$ and equations 2-3 and 2-4 to generate n_r candidate means $[\mu_i]$ and standard deviations $[\sigma_i]$ for an ensemble of n_r Normal distributions to be used for uncertainty analysis as described later.

An associated "Superdistribution" is shown in Figure 2-4. The SD is obtained by sampling each Normal PDF of the ensemble and accumulating or binning all the samples into a single distribution. A promising more direct method of construction is presently being investigated in [Romero, 2018]. The Superdistribution should be symmetric about its mean (which should be the same as the nominal mean $\tilde{\mu}_x$ of the data set, and of the T-based PDF of possible means inferred from the few samples in the data set).

Figure 2-5 shows the standard deviation magnitudes of a 95/90 TI Equivalent-Normal, a Superdistribution, and a histogram of 5000 standard deviations for n_r = 5000 Normals that would make up an associated Ensemble of Normals. (These results come from [Romero & Weirs, 2018].) The nominal standard deviation from N=4 data samples in this example is $\tilde{\sigma}$ = 0.0056. The histogram of 5000 standard deviations is generated from Eqn. 2-4 scaling of 5000 samples from a Chi-Square distribution with N - 1 = 3 degrees of freedom. The EON derived Superdistribution has a standard deviation σ_{SD} = 0.0105. This value coincides with the 83rd percentile of the histogram and is about 88% larger than the nominal standard deviation from the data samples. The standard deviation of the 0.95/0.90 TI Equivalent Normal is σ_{EN} = 0.0142. This coincides with the 92rd percentile of the histogram and is about 150% larger than the data standard deviation. The TI-EN standard deviation is about 35% larger than the SD standard deviation. In general, a 95/90 TI-EN distribution is characteristically broader than its counterpart SD distribution, and both are substantially broader than a Normal distribution fit to the raw data.

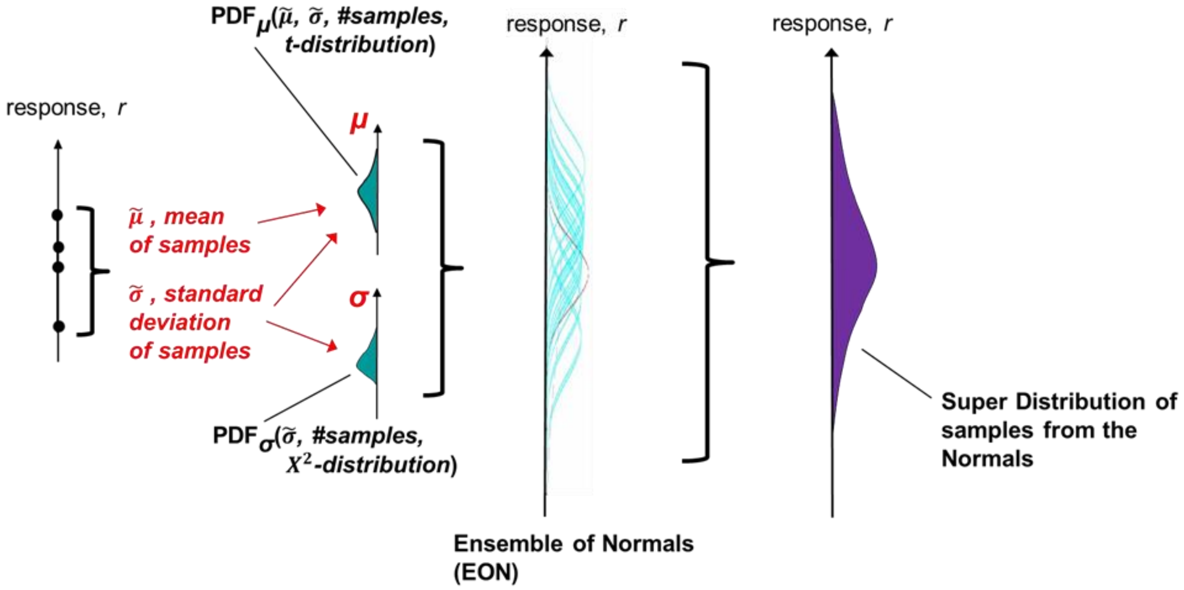

response, r
$\widetilde{\mu}$, mean of samples
$\widetilde{\sigma}$, standard deviation of samples
PDF$_{\mu}$($\widetilde{\mu}$, $\widetilde{\sigma}$, #samples, t-distribution)
μ
σ
PDF$_{\sigma}$($\widetilde{\sigma}$, #samples, X^2-distribution)
response, r
Ensemble of Normals (EON)
response, r
Super Distribution of samples from the Normals

Figure 2-4. Construction of Ensemble of Normals (EON) and their Superdistribution (SD) from scaled T and Chi-Square distributions given the mean and standard deviation of the raw data samples.

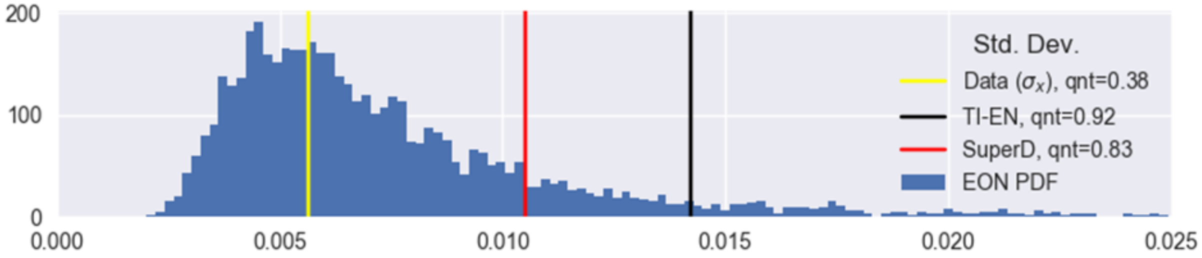


Figure 2-5. Magnitudes of standard deviations from N=4 data samples and corresponding distributions from sparse-data methods (from [Romero & Weirs, 2018]).

<u>*Percentile Estimation Relationship between Ensemble of Normals and Tolerance Intervals*</u>

Figure 2-6 shows how the EON approach can provide uncertainty information on PDF percentiles inferred from the sample data, as opposed to just point estimates that TI Equivalent Normals and Superdistributions provide. The figure shows an example for 2.5 and 97.5 percentiles of response. These percentiles are determined on each of the n_r = 5000 Normals of the EON, and a PDF of the n_r estimates is formed for each response percentile as depicted in the figure. The percentile PDFs are not symmetric, although depicted so in the figure. See e.g.

[Romero & Weirs, 2018] for example shapes of PDFs for the 2.5 and 97.5 percentiles from several N=4 sample sets, along with associated SD and TI-EN distributions.

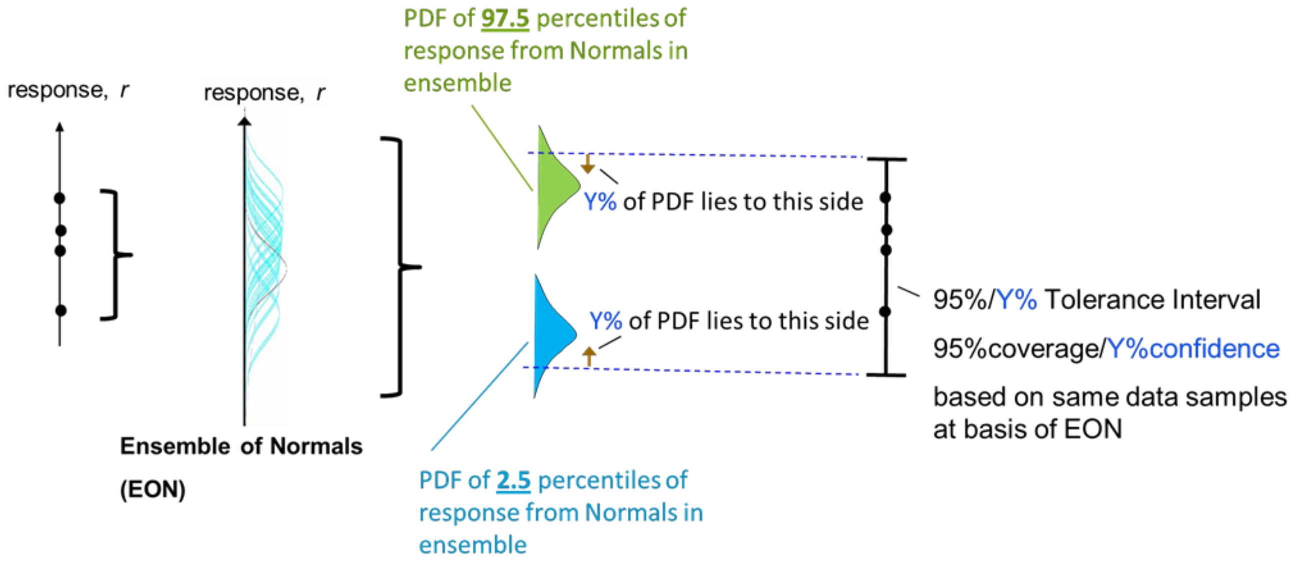

response, r
response, r
PDF of 97.5 percentiles of response from Normals in ensemble
Y% of PDF lies to this side
Y% of PDF lies to this side
95%/Y% Tolerance Interval
95%coverage/Y%confidence
based on same data samples at basis of EON
Ensemble of Normals (EON)
PDF of 2.5 percentiles of response from Normals in ensemble

Figure 2-6. EON-derived uncertainty distributions on inferred 2.5 and 97.5 percentiles of population response (and approximate equivalence of 95%/90% TI end-points and corresponding 90% confidence levels on PDFs of 2.5 and 97.5 percentiles of response).

The figure indicates an approximate equivalence between the end-points of a 95%coverage/Y%confidence TI and corresponding quantiles on PDFs of the 2.5 and 97.5 percentiles of response per the previous paragraph. For example, the lower end of a 95/90 TI approximately coincides with the 0.1 quantile of the uncertainty distribution for the 2.5 percentile, and the upper end of a 95/90 TI coincides with the 0.9 quantile of the PDF for the 97.5 percentile. Accordingly, 90% confidence exists that the 2.5 and 97.5 percentiles of a Normal distribution being sampled will lie within the said upper and lower quantiles. The designation of this range as a '95%/90% EON interval' is then appropriate, where Y% in Figure 2-6 equals 90% in the example here. This also amounts to the statement we've ascribed to 95/90 TIs—that the ends of the TI will, with 90% reliability or confidence, contain the range between 2.5 and 97.5 percentiles of the Normal distribution being sampled. This ascription is not exactly true, per "non-centrality" concepts explained at the beginning of section 3.1, but the approximation is very close. Empirical evidence of the closeness of this relationship is presented in figures 2-7 to 2-9.

For each trial in Figure 2-7, N=2 or N=20 samples are drawn from a Standard Normal distribution and a 95/90 TI and a 95/90 EON interval are constructed from the samples. For most of the 20 trials for N=2 samples, the 95/90 TIs closely coincide with the 95/90 EON intervals. The correspondence is uniformly very close for all 20 trials for N=20 samples.

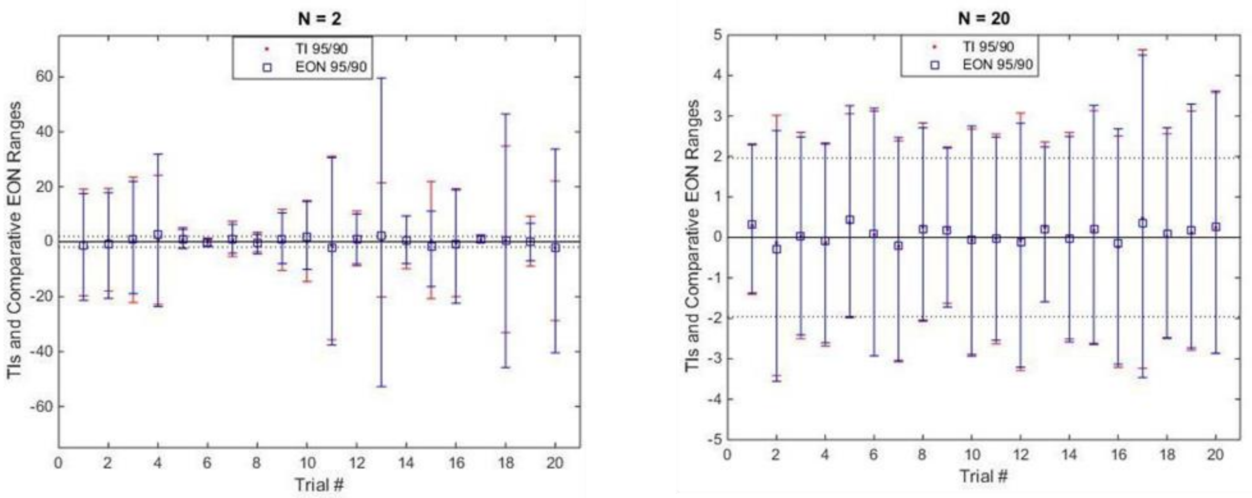


Figure 2-7. Twenty trials of N=2 and N=20 samples drawn at random from a Standard-Normal PDF, where 95/90 TIs constructed from the samples are plotted overlying intervals constructed from lower (0.1) to upper (0.9) quantiles respectively of PDFs of 2.5 and 97.5 percentiles of 100 Normals in an EON (n_r=100). Horizontal dashed lines mark true 2.5 and 97.5 percentiles of Std. Normal PDF. Note the >10X different vertical scales for the N=2 and N=20 plots; the latter shows much smaller deviations from the true percentiles.

Figure 2-8 focuses on the 2.5 percentile of response and plots each UQ method's distribution of 10K results from 10K trials for N=2,4,10,20 samples per trial. Only the 2.5 percentile results are plotted because results for the 97.5 percentile are negative-reflections about a vertical line at abscissa=0. (This was confirmed by plotting them.) The plotted vertical green lines at -1.96 (= -1.96*standard deviation of unity for the Standard-Normal PDF) mark the true 2.5 percentile from the PDF being sampled.

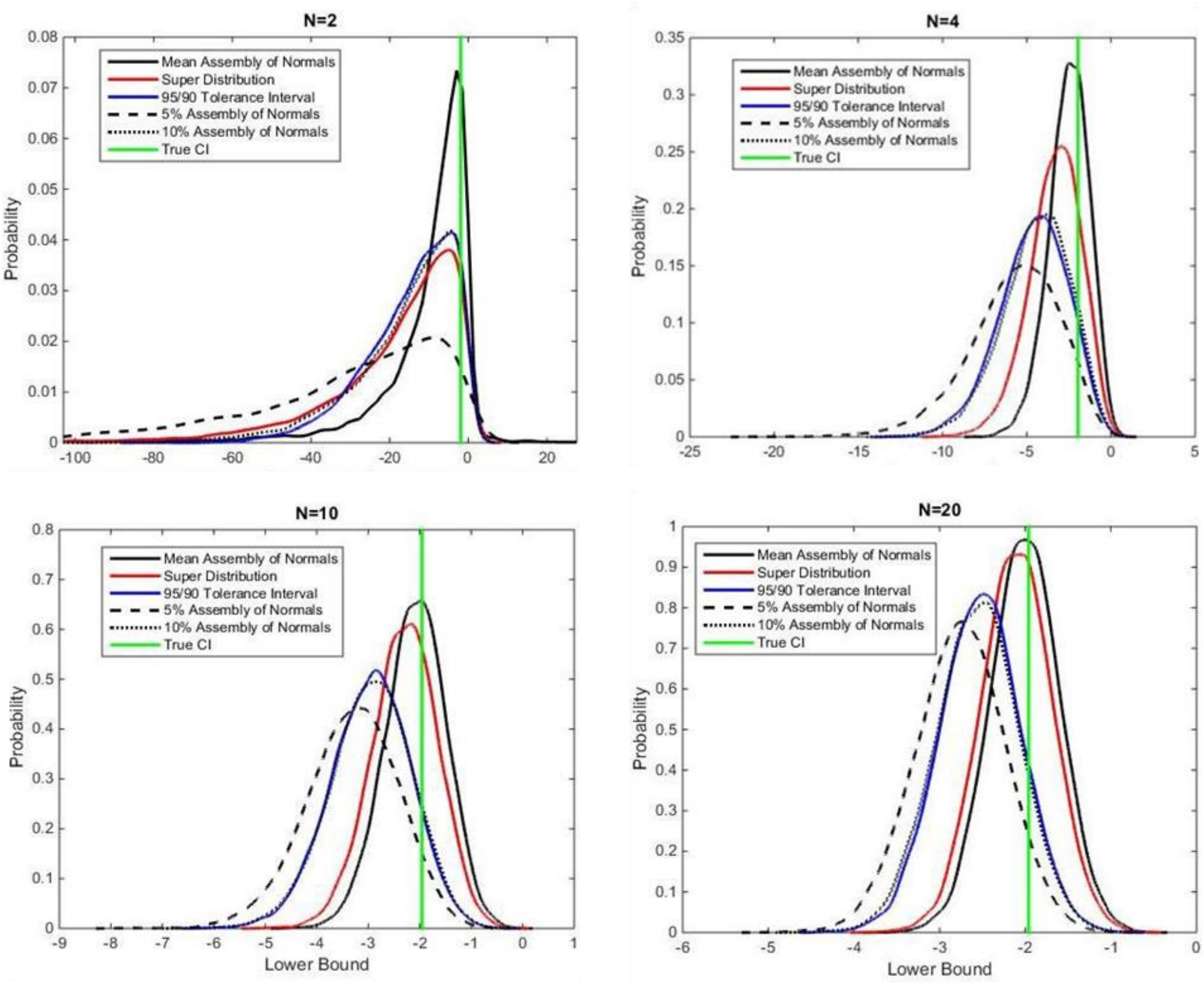


Figure 2-8. Distributions of results extending Figure 2-7 results for the lower (2.5) percentile to 10K trials and N=4,10 and other sparse-data UQ methods per the plot legends.

Consistent with the closeness of the TI and EON results in Figure 2-7, for each of N = 2, 4, 10, 20 in Figure 2-8 the distribution of 95/90 TI results plots essentially on top of the distribution of '10% Assembly of Normals' results. (Here 'Assembly of Normals' (AON) is a synonym for 'Ensemble of Normals' (EON), and a label 'Q% Assembly of Normals' in the plots in Figure 2-8 and Figure 2-9 corresponds to 95/(100-Q) EON results in the terminology introduced earlier.) Figure 2-9 shows a similar correspondence between 95/95 TI results and 5% AON results. Thus, the relationship depicted in Figure 2-6 is empirically seen to be effectively true. This allows the more easily constructed X%/Y% TIs to be stood in for purposes the X%/Y% EON intervals might be used (which are much more difficult to construct). For example, a X%/Y% TI can be used for a Y% confidently conservative bound on an individual percentile *Z* of response, instead of a range between two percentiles, where the X% parameter of the TI is calculated as [2*(Z) - 100]% for e.g. a Z=99 percentile of response (the equation X% = [100 - 2*(Z)]% applies if Z% < 50% and a conservative bound in this case is a bound from below; the estimate is conservative if less than the true percentile value Z%).

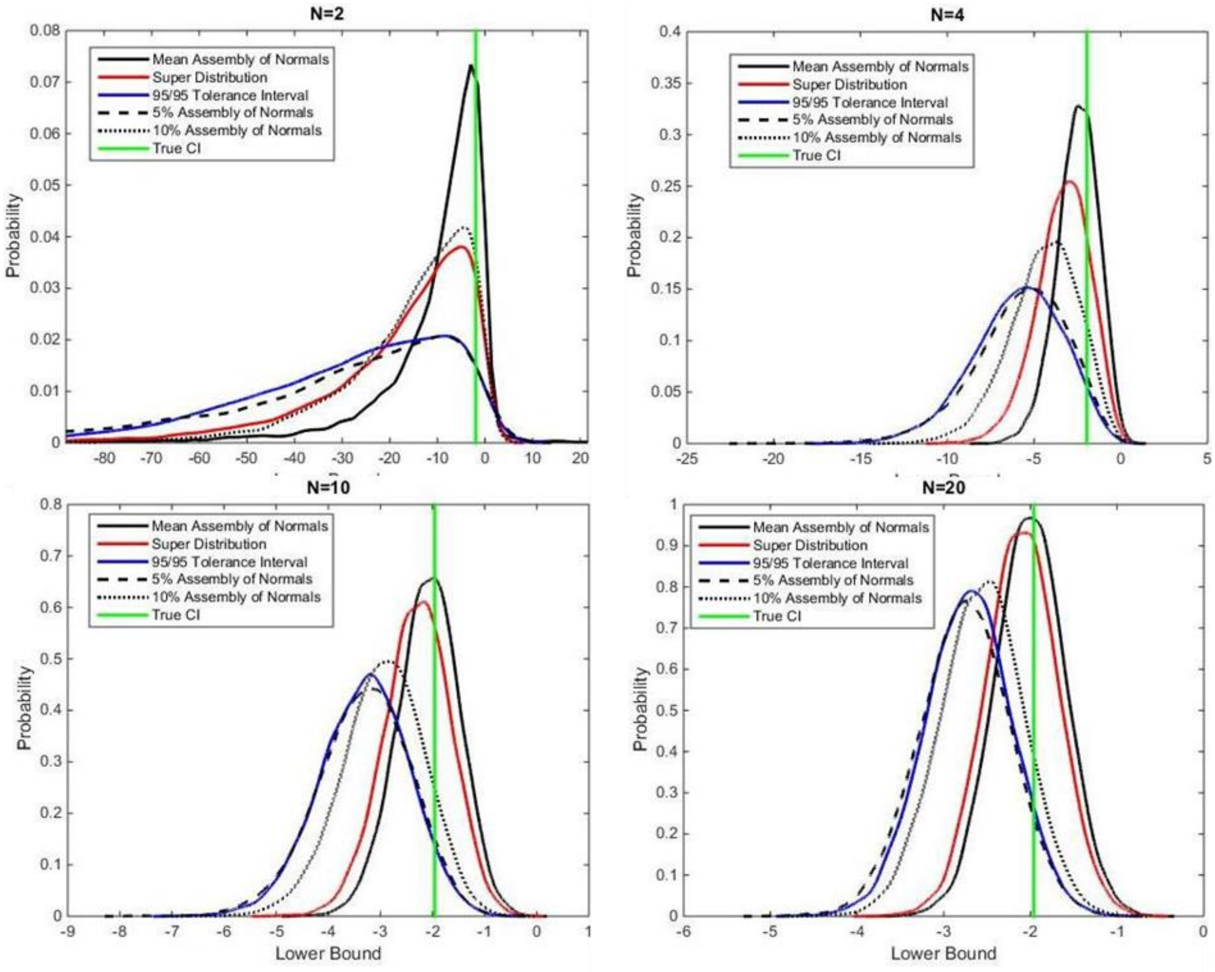


Figure 2-9. Version of Figure 2-8 with 95/90 TI results replaced by 95/95 TI results.

Quantitative analyses of the sparse-data UQ method results will be conducted in the next two sections, but a few more qualitative observations and comparisons are made here. All methods' distributions of results become less wide and more peaked about the true percentile value (they all get more accurate) as the number of samples increases. (The relatively similar sizes of the PDFs in the four plots is deceptive because the plots have significantly increasing ordinate scales and decreasing abscissa scales as N goes from 2 to 20.) The plots show three distinct PDF groups at N=2 samples, four distinct groups at N=4 samples, and only two distinct groups at N=10 and 20 samples. These results accompany the following dynamics as N increases from 2 to 20 samples.

The PDF of Superdistribution results is very different from the PDF of Mean AON results for N=2, but the SD PDF changes considerably as N goes from 2 to 20 and progressively approaches the shape of the Mean AON PDF, which has a more stable shape in a self-similar sense over the range of N. (A Mean AON result in an individual trial is the equivalent of a 95/50 TI or 95/50 EON interval between the 50th percentiles of the upper and lower PDFs in Figure 2-6.) For N=20 samples the SD PDF is quite similar to the Mean AON PDF.

3. METHODS' PERFORMANCE FOR BOUNDING CENTRAL 95% OF RESPONSES

Performance of the sparse-data UQ methods described in Section 2 is characterized in this section for accuracy and efficiency in bounding the central 95% of responses for a diverse and challenging test-bed of analytic and empirical PDF shapes.

3.1. Performance on Four Analytic PDF Shapes

The four analytic PDF shapes used in this subsection are a Normal distribution; a five degree-of-freedom (5 DOF) Student's t distribution with zero mean (has wider tails than a Normal distribution); and two skewed distributions, a Log-Normal and a more highly skewed Weibull, both of which are described in detail in Appendix A.

Normal PDF Results and Development and Use of a Multi-Attribute Performance Metric

On average, 90% of 95/90 TIs would be expected to capture 95% or more of the Normal distribution being sampled. This includes capture of the central 95% range of the PDF between its 0.025 and 0.975 quantiles as one possibility, but also "non-central" quantile ranges such as the 0.01 to 0.96 range of the PDF. TI confidence/reliability/success rates for capturing just the central 95% of response will be smaller than success rates for the less restrictive case of capturing *any* contiguous 95% of the PDF. Capture of only the central 95% of response as the success criterion anticipates the use of TIs for design or safety assessment purposes. In such endeavors we envision objectives such as: "We want high reliability that no more than 2.5% of results lie above the predicted 95/~90 TI we're using to size safety or performance margins with." TIs that bound the central 0.025 to 0.975 range of response can be used for such design objectives, or for objectives that no more than 2.5% of responses lie below the predicted 95/~90 TI, or no more than 2.5% of results lie above the predicted 95/~90 TI and no more that 2.5% of results lie below it. There is less design certainty about what one is establishing with TIs that enclose non-central 95% ranges. Therefore we choose to only count central TI successes, as a truer measure of the reliability of TIs when used for design and safety analysis purposes.

Thus, TI success rates for capturing the central 95% range of a Normal PDF are smaller than the "advertised" 90% for 95/90 TIs. Anecdotally, five or six of the 40 TIs in Figure 2-7 do not span the -1.96 to 1.96 central 95% range of the Standard Normal being sampled. This corresponds to < 90% reliability or success rate for the 95/90 TIs. Table 3-1 gives precise TI reliability rates from 10K trials at various numbers of samples. The 95/90 TI success rate is about 89% for N=2 samples, falling to about 81% for N=20. These results are consistent with previous studies in [Romero et al. 2013a,b]. The 95/95 TI success rate is about 95% for N=2 samples, falling to about 90% for N=20. Apparently, capture of only the central 95% of response (as the success criterion) affects TI capture-success rates increasingly more as the number of samples N increases and TI length drops precipitously (see Table 2-1 and Figure 2-2).

The results in Table 3-1 for all sparse-data UQ methods are plotted in Figure 3-1. As expected, the results for 95/90 TIs and EON90% plot effectively on top of each other, as do the results for 95/95 TIs and EON95%. A few cases in the table show the EON reliability rate is equal to or slightly higher than the corresponding TI reliability rate, but the majority of cases show the TI reliability rate is better. This is reflected by the TI methods' better average reliability rate than the corresponding EON reliability rate (last column in the table). The table data and the plots also show that the TI and EON methods have significantly higher central 95% capture reliability than the Superdistribution and Mean EON (EON50%) methods. EON50% performs least well, with 75% reliability for N=2 samples, degrading to only 35% at N=20 samples. The SD method does not perform quite as poorly, but does considerably less well than 95/90 and 95/95 TI and EON.

Reliability rates decrease with number of samples N at much faster rates for the Superdistribution and EON50% methods than for the 95/90 and 95/95 TI and EON methods. Because the SD and EON50% intervals have the same means or midpoints as the intervals from the 95/90 and 95/95 TI and EON methods, the sizes of the SD and EON50% intervals are apparently smaller than those of the 95/90 and 95/95 TI and EON methods, and decrease in size much faster than the 95/90 scaling behavior depicted in Figure 2-2.

Table 3-1. Empirical Reliabilities of Sparse-Data UQ Methods for capturing the central 95% range of a Normal distribution sampled N times. Results from 10K random trials of each method.

Method	N=2	N=4	N=10	N=20	avg.scor
95/90 TI	89.4%	87.4%	84.5%	81.3%	85.7%
EON 90%	89.1%	85.4%	83.4%	82.4%	85.1%
95/95 TI	94.6%	93.4%	91.6%	89.6%	92.3%
EON 95%	94.6%	92.4%	91.0%	90.1%	92.0%
EON 50%	74.4%	51.9%	39.2%	35.4%	50.2%
Super D.	89.6%	72.4%	54.9%	45.6%	65.6%

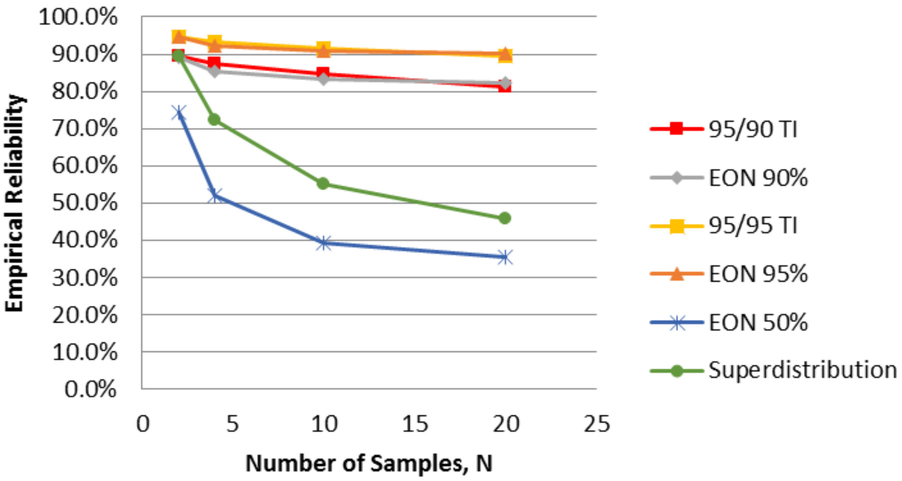


Figure 3-1. Empirical Reliabilities for capturing the central 95%range of a Normal distribution sampled N times (plotted data from Table 3-1).

A broad consideration of engineering objectives indicates that the decrease in coverage success with added samples does not mean that less samples are better. With less samples, the "overshoot" errors of the methods can be very large, with large potential to yield very conservative designs or to significantly under-estimate safety margins as discussed above. These outcomes can lead to unnecessary design and product costs and highly pessimistically skewed performance and safety perceptions of the designs and products. To get a sense of the conservatism vs. risk tradeoff involved, consider the design objective mentioned previously: "We want high reliability that no more than 2.5% of results lie above the predicted 95/90 TI we're using to size safety or performance margins with." If one can accept e.g. 85% expected reliability of meeting this objective (15% risk of not meeting it), instead of say 90% expected reliability (10% risk), then the 95/90 TI curve in Figure 3-1 indicates that 10 samples would suffice. Ten-sample TIs are less than 1/6 the size of 2-sample TIs (see Table 2-1), which have an expected reliability of about 89% in Table 3-1. The 6X difference in the size of the TI designed or analyzed with could translate to very large improvements in design cost, weight, and other objectives, larger indicated performance and safety margins, etc., with a relatively small (5%) increase in risk (from 10% to 15%) that the requirement is not met.

Additionally, in this example it is of far less concern whether the TI bounds the lower 2.5 percentile of response, as the design requirement only expresses a need to bound the upper 97.5 percentile—a "one-sided" upper bounding requirement. Then TI lower-bounding of the 2.5 percentile is not necessary and a somewhat higher TI success rate would be expected with this relaxed criterion for success. Moreover, different UQ methods could have similar success rates in bounding the desired percentiles, but at the same time have significantly different error magnitudes (e.g. the diversity of distribution shapes in Figure 2-8) that favor one method over the other.

In general, a more refined and comprehensive performance measure than just capture success rate is needed to more fully characterize UQ method performance. Winokur et al. (2017) propose and use a multi-attribute weighted performance metric to help quantify tradeoffs between risk, conservatism, and number of samples and how these tradeoffs differ between several proposed UQ approaches for multi-material sparse-data problems. The performance metric is summarized next and a particularization of it is applied. The following development uses 95/90 TIs for illustration.

Figure 3-2 shows TI mismatch errors relative to a reference percentile range that the TI is desired to bound. Mismatches at the upper (u) and lower (ℓ) ends of the reference percentile range are defined as:

$$\epsilon_u = r_u - r_{u\text{-reference}} \qquad (3\text{-}1)$$

$$\epsilon_\ell = r_{\ell\text{-reference}} - r_\ell \qquad (3\text{-}2)$$

where r stands for 'response' of some random quantity.

These equations return a positive value of mismatch error ϵ_ℓ when the TI range extends below or bounds the reference 2.5 percentile at the lower end. In coordinated fashion, the mismatch error ϵ_u at the upper end is measured positive when the TI range extends above or bounds thereference 97.5 percentile. Thus, when both ϵ_ℓ and ϵ_u are positive this is termed a +/+ error case and the TI bounds the reference range from both above and below as desired. Table 3-2 provides designations for the other possible error cases.

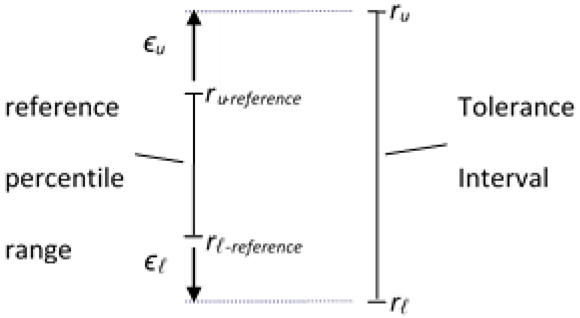
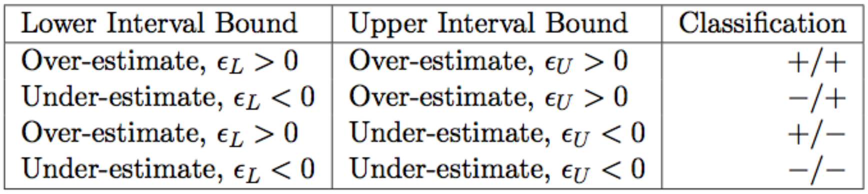


Figure 3-2. Definition of mismatch errors between TI and reference percentile range.

Table 3-2. TI Mismatch Error Classifications



Lower Interval Bound	Upper Interval Bound	Classification
Over-estimate, $\epsilon_L > 0$	Over-estimate, $\epsilon_U > 0$	$+/+$
Under-estimate, $\epsilon_L < 0$	Over-estimate, $\epsilon_U > 0$	$-/+$
Over-estimate, $\epsilon_L > 0$	Under-estimate, $\epsilon_U < 0$	$+/-$
Under-estimate, $\epsilon_L < 0$	Under-estimate, $\epsilon_U < 0$	$-/-$

In the performance scoring above for TI success or reliability rates, only errors in the +/+ category are "valued." High proportions of +/+ errors relative to the other error types were sought. But other error categories may also involve successful TI performance, as in the case of the one-sided bounding requirement discussed above. The Winokur et al. metric incorporates preference weighting degrees of freedom (see below) to express relative desirability of the errors in the four categories in Table 3-2.

Error magnitudes in the various error categories are also very important. For example, while errors in the -/- category may be undesirable in a given situation, if these "undershoot" or "shortfall" errors are very small, then they may be nominally acceptable, and in any case they would be considered preferable to larger -/- errors. Thus, it is important for a performance metric to involve not just the relative proportions of the error types, but also the error magnitudes in the various categories.

For a TI in any error category, its upper and lower mismatch errors have a combined magnitude given by adding their absolute values:

$$|\epsilon| = |\epsilon_\ell| + |\epsilon_u|. \tag{3-3}$$

Among TIs in a given error classification bin like +/+, TIs with larger absolute error sums $|\epsilon|$ are considered worse performers than TIs with smaller average absolute errors. To assess average error magnitudes over a number of TI trials, errors in the various classification types can be grouped, summed, and averaged as follows.

$$|\bar{\epsilon}| = \frac{1}{N_{trials}} \left[\sum^{N^{++}} (\epsilon_u + \epsilon_l)_{++} + \sum^{N^{+-}} (|\epsilon_u| + \epsilon_l)_{+-} + \sum^{N^{-+}} (\epsilon_u + |\epsilon_l|)_{-+} + \sum^{N^{--}} (|\epsilon_u| + |\epsilon_l|)_{--} \right] \qquad \text{(3-4)}$$

Each summation sign in this equation includes all trial TIs in the indicated error category. Absolute value signs are written in this equation only for negatively signed errors in a given error category. A form of this equation that is less vulnerable to coding mistakes takes the absolute value of all error terms/quantities.

In [Romero, Schroeder, et al., 2017] it is established that on average for 70 diverse PDF shapes (also to be discussed in section 3.2 of the present paper) the magnitude of + + errors decreases with the number of samples N, in scale with the TI multiplier in Figure 2-2. This beneficial drop of + + error magnitudes is unfortunately accompanied by the undesired effect of lower + + success rates per Figure 3-1. If the rate of decrease of + + error magnitude is faster than the rate of decrease of the proportion of + + errors, and if these performance attributes are equally weighted, then the overall quality of performance would improve with added samples—as seems most reasonable. The Winokur et al. metric accounts for these two competing performance attributes, and those associated with the other error types, with the following performance metric.

$$metric = \frac{|\bar{\epsilon}|}{w_1\, p_{++} + w_2\, p_{+-} + w_3\, p_{-+} + w_4\, p_{--}} \quad \Bigg| \quad \sum w_i = \sum p_{ij} = 1 \qquad \text{(3-5)}$$

The numerator in this equation comes from Equation 3-4. The proportion p_{++} in the denominator is equal to the success rate of attaining ++ errors in the 10K trials. Similarly, the other proportions p_{ij} in the denominator are the rates or proportions obtained for the other error types in the trials. The multiplier weights w_i in the denominator allow for expression of relative preferences for the various error types. Once the weights (which must sum to one) are prescribed, then the denominator value is determined by the error proportions p_{ij} in a given set of trials whose performance is being scored. The numerator is set by the error magnitudes involved. For a given set of error preferences/weights, if two sets of trials with different sparse-data UQ methods, for example, have the same error proportions but different average error magnitudes, then the denominator will be the same for both cases, but the case with the larger average error will have the larger numerator and therefore the larger metric value. Thus, a *larger* performance metric value coincides with *lower* overall performance.

Next we briefly examine how the metric value is affected by different error preference weightings. See Appendix B in [Winokur et al., 2017] for a more in-depth investigation with numerical examples of metric behavior under different error proportions and preference weightings.

In general, a set of trial results fixes the numerator value and the proportions $p$ij in the denominator. The error preference weights wi also affect the denominator value such that it, and thus the overall metric value, vary with the values of the weights. For instance, if a high proportion of + + errors exist, say p++ = 0.8, and these are highly preference-weighted relative to the other types of errors, say w1 = w++ = 0.8, then the expressed preference for high proportion of + + errors is largely satisfied. The (p++)(w++) term dominates the denominator in this case, and the denominator is larger than if + + errors are not highly preferred/weighted, e.g. by a prescribed weight $w1 = w{+}{+} = 0.2$. The smaller denominator in the latter case, and the unchanged numerator value, mean a larger value of the overall performance metric, so worse performance is indicated, as it should be, than in the former case where $w1 = w{+}{+} = 0.8$ expresses a relatively high preference for + + errors and a high proportion of them exists, $p{+}{+} = 0.8$.

On the other hand, if the same high preference $w_{++} = 0.8$ is specified for + + errors, and in a second set of trials with another method, the incidence of them is much lower (e.g. $p_{++} = 0.2$) than in the paragraph above, then the denominator would likely decrease relative to the $p_{++} = 0.8$ case above. This would, in isolation, tend to increase the metric value in a reflection of mal-satisfaction of the high preference $w_{++} = 0.8$ for + + errors. But the error magnitudes (thus the numerator value) would also generally be different along with the new error proportions p_{ij} and new denominator value in the second case. Therefore, it is not evident a-priori whether the overall metric value would increase or decrease, i.e., whether performance would be indicated better or worse in the second case. The performance metric would have to be calculated to decide this.

We now compare the behavior of the multi-attribute performance metric Eqn. 3-5 to the single-attribute performance measure in Figure 3-1, as the number of samples per trial increases. The limited performance measure in Figure 3-1 credits only + + errors as desirable. This corresponds to an error preference weighting $w_{++} = 1.0$ in Eqn. 3-5, with all other weights = zero. For this case, the denominator of Eqn. 3-5 decreases with added samples as the proportion of + + errors decreases per Figure 3-1. But the magnitude of the numerator's errors appear to decrease at an even faster rate; the performance metric value falls with added samples as the left plot in Figure 3-3 shows for all methods. Results for EON 90% and 95% methods are given in Table B.1 in Appendix B, but are not plotted because the corresponding TIs are much easier to construct and their results are always very close to, and typically better than, the EON results as the table shows. The decline with added samples for all methods indicates improving overall performance under Eqn. 3-5's broad measure of performance.

By this broad performance measure, the 95/90 TI and EON methods are indicated best in Figure 3-3's left plot and Table B.1 by a significant relative margin when averaged over all N=2,4,10,20 cases. The 95/95 methods are indicated worst, in stark contrast to Figure 3-1's ranking of the 95/95 methods as best by the single criterion of capture success while ignoring magnitudes of overshoot and undershoot errors. Table B.1 reveals that EON50% does best at N=4 and SD does best at N=10. On average over all N=2,4,10,20 cases, SD does second best and EON50% does third best.

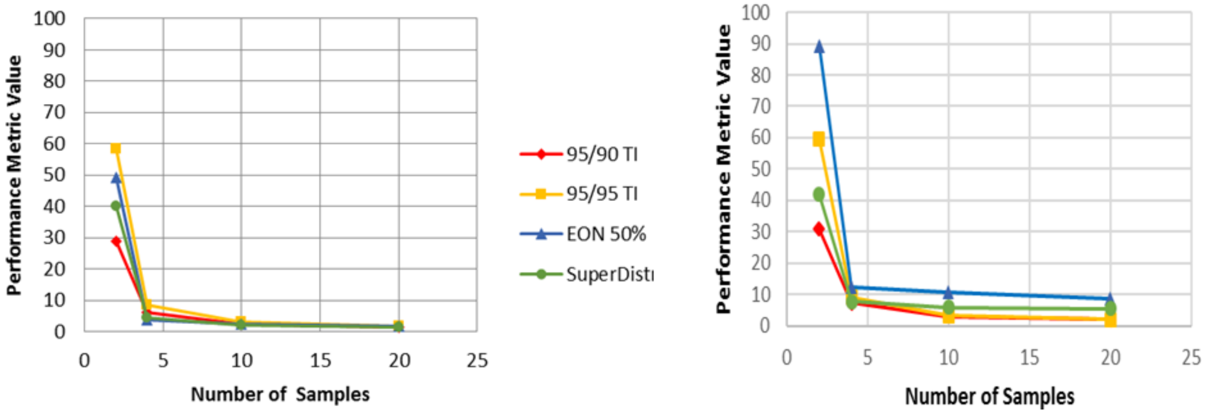


Figure 3-3. Performance Metric results vs. number of samples for Normal distribution said error preferences/weights and unpenalized and 10X penalized shortfall errors as explained in the text. Results for each # of samples are from 10K TI trials with each method.

We next consider adding to the metric a penalty factor that multiplies the magnitudes of the least preferred or least acceptable errors in cases where they are considered to be potentially highly harmful or dangerous. Equation 3-6 adds to Equation 3-5 an illustrative 10X penalty on shortfall/undershoot errors, where these are considered 10 times worse than overshoot errors of the same magnitude.

$$|\bar{\epsilon}| = \frac{1}{N_{trials}} \left[\sum^{N^{++}} (\epsilon_u + \epsilon_l)_{++} + \sum^{N^{+-}} (10|\epsilon_u| + \epsilon_l)_{+-} + \sum^{N^{-+}} (\epsilon_u + 10|\epsilon_l|)_{-+} + \sum^{N^{--}} 10(|\epsilon_u| + |\epsilon_l|)_{--} \right] \quad (3\text{-}6)$$

The penalty factor in Eqn. 3-6 increases the numerator value in the performance metric Equation 3-5 (relative to use of the Eqn. 3-4 non-penalty version of the numerator) while leaving the denominator unchanged. The net effect is to increase the performance metric value, which decreases the perceived quality of method performance. This is logically consistent with having 10X larger effective magnitudes of the least preferred or acceptable errors. Accordingly, the plot at right in Figure 3-3 shows that, for a given number of samples N, all the performance curves have higher/worse values than their counterpart curves in the plot at left with un-penalized shortfall errors (numerical data of the plot at right in Figure 3-3 are found in Appendix B Table B.2). Results for EON 90% and 95% methods are not in the plot because the corresponding TI results are always very close to, and typically better than, the EON results, as Table B.1 shows.

The performance metric indicates improving overall performance with added samples for all methods. The 95/90 TI method is indicated best for N=2,4,10, and 95/95 TI is best at N=20 with 95/90 TI a very close second (see Table B.2). When averaged over all N=2,4,10,20 cases, 95/90 TI does best by a significant relative margin; SD does second best; 95/95 TI does third best (better than SD at large sample sizes N=10,20 but not at small sample sizes N=2,4); and

EON50% always does worst. The latter appears to be a consequence of the 10X penalty on shortfall errors, which are prevalent with the EON50% method given its relatively low capture success rates (Figure 3-1).

All performance curves exhibit substantial knees at N=4 data samples—for both unpenalized and 10X penalized metrics. Beyond four samples, the incremental gains in both performance measures are much smaller per added sample than are the incremental gains up to four samples. It will be seen that this knee is consistent for the many other PDF types studied in this paper.

5 DOF t-distribution Results

A 5 DOF t-distribution with zero mean is now considered. This distribution is symmetric about zero and has wider tails than a Normal distribution having the same standard deviation as the 5DOF t. The EON50% method is not studied here because of its poor relative performance on the Normal PDF.

Table 3-3 and Figure 3-4 present the result for UQ methods applied to the t-distribution. As expected, all method capture success rates fall with added samples and the 95/90 TI and 95/90 EON results are very similar, as are the 95/95 TI and EON results. A few cases in the table show the EON reliability rate is equal to or slightly higher than the corresponding TI reliability rate, but in the majority of cases the TI reliability rate is better. TI has a higher average score over the set of cases N=2,4,10,20, by only by one or two tenths of a percentage point (last column in the table). The Superdistribution method has comparable reliability to the TI and EON methods at N=2 samples, but its reliability decrease with the number of samples N at much faster rate. SD has only 50% reliability at N=20 samples, where it under-performs the TI and EON 95/90 and 95/95 methods respectively by about 30 and 40 percentage points.

Reliabilities for the TI and EON methods at small sample sizes N=2,4 are on the order of one percentage point better for the 5 DOF t than for the Normal PDF, but on the order of two or three percentage points worse at the larger sample sizes N=10,20. On average over all cases N=2,4,10,20, TI and EON reliabilities are one to two percentage points lower for the 5 DOF t distribution. Reliabilities for the SD method are about 1 percentage point better for the 5 DOF t than for the Normal PDF at N=2 samples, increasing to about 5 percentage points better at N=20 samples, for a N=2,4,10,20 average of about 2 percentage points better than for the Normal PDF. Overall, performance is fairly similar on the Normal and the 5 DOF t distributions.

Table 3-3. Empirical Reliabilities of Sparse-Data UQ Methods for capturing the central 95% range of a 5 DOF t-distribution sampled N times. Results from 10K random trials of each method.

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	90.2%	87.3%	81.9%	77.9%	84.3%
EON 90%	90.4%	85.6%	81.6%	79.2%	84.2%
95/95 TI	95.2%	93.7%	89.1%	85.9%	91.0%
EON 95%	95.1%	92.5%	88.5%	87.2%	90.8%
Super D.	90.9%	73.0%	57.2%	50.7%	67.9%

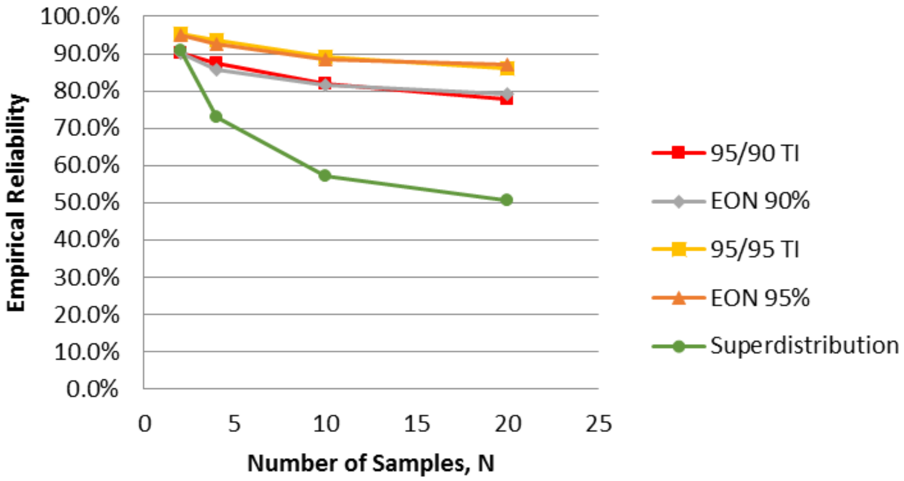


Figure 3-4. Empirical Reliabilities for capturing the central 95% range of a 5 DOF t-distribution sampled N times (plotted data from Table 3-3).

The left plot in Figure 3-5 shows the no-penalty metric results. Metric values decrease with added samples (indicating improving overall performance) for all methods. The EON95% results are not plotted because they are very close to, but not as good on average, as the 95/95 TI results (see last column in Appendix B Table B.3). The 95/95 TI method is the worst performer of all the plotted methods, presumably because of its larger relative overshoot errors. The 95/90 TI method is the next best performer on average over N=2,4,10,20. The SD method performs slightly better—second best on average. EON90% results are the best on average, in a rare outperformance of 95/90 TIs on average.

The right plot in Figure 3-5 shows the 10X penalty metric results. The EON 90% and 95% results are not plotted because the corresponding TI results are very close to, but typically better than, the EON results per Table B.4, in particular the N=2,4,10,20 average metric values in the table's last column. Metric values decrease with added samples (indicating improving overall performance) for all methods except for SD, which increases beyond N=4 samples. Apparently, a prevalence of shortfall errors give SD the lowest capture success rates (Figure 3-4), and the 10X penalty on these errors outweighs the effect of decreasing error magnitudes as samples are added. (Decreasing error magnitudes are evident; they reduce the unpenalized metric's numerator enough to drive the SD metric lower/better with added samples (left plot) despite the metric's declining denominator from declining capture proportion p++ in Figure 3-4.) Per the last column in Table B.4, 95/90 TIs average the best performance by far, with the lesser performing SD and 95/95 TI methods scoring very similar on average.

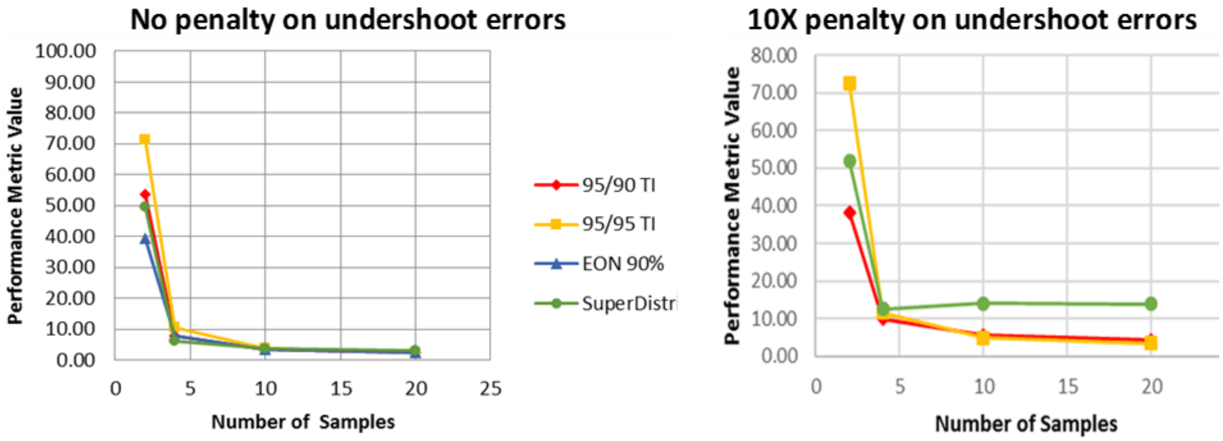


Figure 3-5. Performance Metric results vs. number of samples for 5 DOF t-distribution and unpenalized and 10X penalized shortfall errors. Results for each # of samples are from 10K TI trials with each method.

For this PDF type as well, all performance curves exhibit substantial knees at N=4 data samples for both unpenalized and 10X penalized metrics.

Log-Normal Distribution Results

The Log-Normal (L-N) distribution defined in Appendix A is now considered. The particular L-N distribution used comes from [Bhachu et al., 2016]. The EON50% method is not studied because of its poor relative performance on the Normal PDF.

Table 3-4 and Figure 3-6 present the result for UQ methods applied to the L-N distribution. As expected, all method capture success rates fall with added samples and the 95/90 TI and 95/90 EON results are very similar, as are the 95/95 TI and EON results. Several cases in the table show the EON reliability rate is equal to or slightly higher than the corresponding TI reliability rate. EON 95/90 scores very slightly better than 95/90 TIs on average (last column in Table 3-4), while EON 95/95 scores very slightly worse than 95/95 TIs on average.

The capture success rates for all methods are significantly lower at each sample size N for the log-normal distribution than for the Normal and 5 DOF t distributions. Reliabilities for the L-N distribution are fairly high for all methods (>86%) for N=2 samples, but the drop in reliabilities with added samples is dramatic for all methods and is steepest (very steep) for the Superdistribution method. At N=4 samples, the TI and EON 95/95 methods still have reasonably high reliabilities of about 85%, and the 95/90 versions have reliabilities >70% which is perhaps acceptable for many engineering purposes, but SD reliability is less than 50% which would appear to be unacceptable in many engineering circumstances.[2] A linear interpolation of the 95/95 TI and

[2] Reliability rates of 75% or 85% are often adequate to sufficiently manage risk, especially if conservatism from other sources exists in the analysis or results—such as applied factors of safety, and/or large indicated design, safety, or performance margins from high-quality analysis, and/or when more than one source of uncertainty is present that involves sparse data conservatively treated with the TI method. In the latter circumstance, studies in [Romero, Swiler, et al, 2013] and [Winokur & Romero, 2017] indicate that when more than one dominant or influential uncertainty source is represented conservatively (each with reliabilities of say 70%), then if the conservatively represented uncertainties are combined in linear propagation or aggregation, the individual conservative biases compound to yield substantially greater than 70% conservative bias.

EON trends shows their reliabilities are about 70% at N=7 samples. More samples may present too much risk for this distribution type.[3] In fact, less samples and a different sparse-data method may more optimally achieve 70% assurance in this case. This is discussed next with the help of the performance metric scores.

Table 3-4. Empirical Reliabilities of Sparse-Data UQ Methods for capturing the central 95% range of a Log-Normal distribution sampled N times. Results from 10K random trials of each method.

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	86.2%	73.9%	35.8%	10.3%	51.6%
EON 90%	86.3%	70.6%	35.2%	12.0%	51.0%
95/95 TI	92.9%	86.5%	53.7%	19.5%	63.1%
EON 95%	93.1%	84.0%	53.3%	22.4%	63.2%
Super D.	87.2%	48.0%	7.9%	1.5%	36.1%

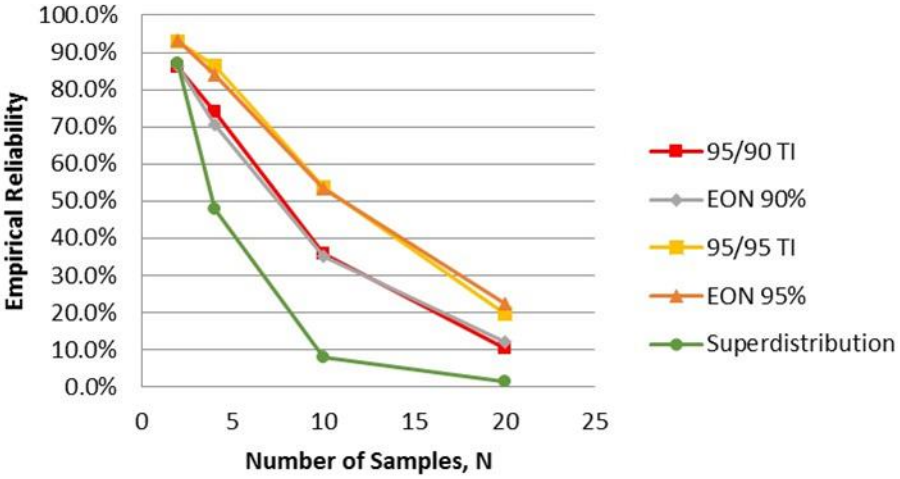


Figure 3-6. Empirical Reliabilities for capturing the central 95% range of a Log-Normal distribution sampled N times (plotted data from Table 3-4).

The left plot in Figure 3-7 shows the no-penalty metric results. The EON results are not plotted because they are very close to, but not as good as, the corresponding TI results on average, even though several individual EON results are better than their the corresponding TI results (see Appendix B Table B.5). Metric values decrease with added samples (indicating improving overall performance) for all methods except SD in going from N=10 to 20. The 95/95 TI method

[3] More accurate/reliable tolerance intervals can be generated for log-normal distributions using a simple transform applied to standardly generated TIs (see e.g. [Young, 2010], [MIL-HDBK-17-1F, 2002]), but we use the L-N case as a bounding circumstance where an unknown distribution shape is being sparsely sampled but is strongly suspected to be less skewed than a log-normal with the parameters here, but is also perhaps not even highly skewed at all, even ~symmetric.

is the worst performer of all the plotted methods except at N=20 where SD is worst. The 95/90 TI method is the best performer except for SD at N=4. In terms of average performance over the set N=2,4,10,20, the 95/90 TI method is best, then SD, then 95/95 TI.

The right plot in Figure 3-7 shows the 10X penalty metric results. The EON 90% and 95% results are not plotted because the corresponding TI results are very close to, but typically better than, the EON results per Table B.6, in particular the N=2,4,10,20 average metric values in the table's last column. Metric values decrease with added samples (indicating improving overall performance) for all methods except for SD, which increases beyond N=4 samples for the same reasons stated for the right plot in for the 5 DOF t-distribution. Per the last column in Table B.6, 95/90 TIs average the best performance by far, then 95/95 TIs followed closely by SDs.

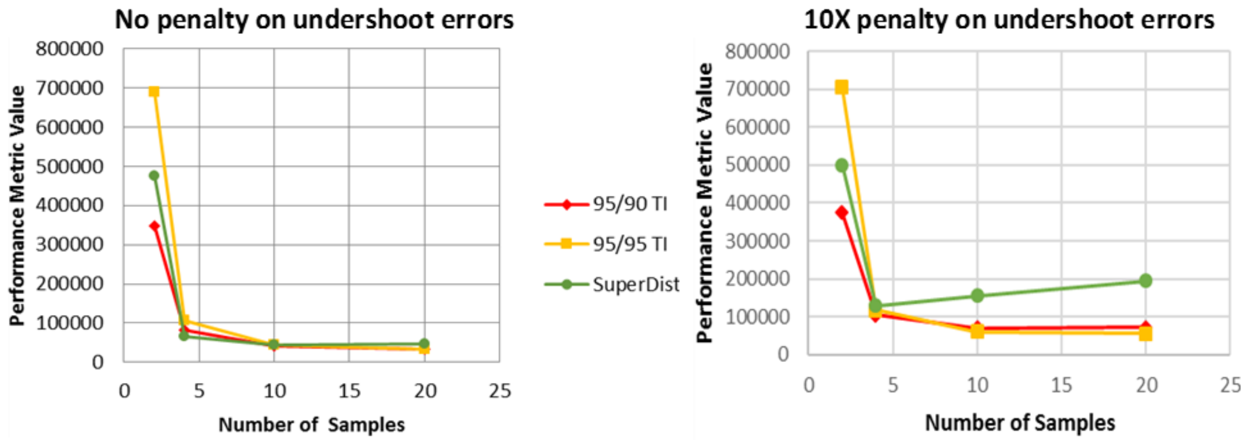


Figure 3-7. Performance Metric results vs. number of samples for Log-Normal distribution and unpenalized and 10X penalized shortfall errors. Results for each # of samples are from 10K TI trials with each method.

For this PDF type like for the others, all performance curves exhibit substantial knees at N=4 data samples for both unpenalized and 10X penalized metrics.

We now continue the discussion started immediately above Table 3-4, with the objective of 70% reliability and in view of the circumstances in footnotes 2 and 3. From Figure 3-6 and Table 3-4, for effectively the same level of non-capture risk one could use N=7 samples with 95/95 TIs, or 4 samples with 95/90 TIs, or 3 samples with SD. Of course, the smaller the number of samples the more attractive from a cost standpoint (e.g. each sample requires a test or an expensive model simulation), but the size/conservatism of the overshoot errors must also be taken into account because this drives design and performance margin perceptions and associated cost, weight, etc. Accordingly, the performance metric results are applied.

Both performance metric plots in Figure 3-7 show that N=4 with 95/90 TIs is preferable to N=3 with SD. The non-penalized metric and left plot indicate that N=4 with 95/90 TIs is preferable to N=7 with 95/95 TIs. The penalized metric and right plot indicate that N=4 with 95/90 TIs is slightly less preferable to N=7 with 95/95 TIs. The magnitude of shortfall errors is emphasized with the penalty metric, which are evidently a larger share of the metric's numerator for the 95/90 TIs than for the 95/95 TIs. One must weigh the potential consequences of the slightly larger potential shortfall errors of the 95/90 method against the use of the 95/95 method with added cost of three more tests or simulations. These considerations are also weighed against the potential cost of larger potential overshoot errors of the 95/95 method which scores lower than 95/90 in the left (unpenalized metric) plot. The best answer depends on the particulars of the situation and cannot be determined here, but at least some constraints, tradeoffs, sensitivities, and a narrowing of UQ method choices, have been determined.

Weibull Distribution Results

The Weibull distribution defined in Appendix A is now considered. Kanwar et al. (2015) indicate that tolerance intervals exist specifically for two-parameter Weibull distributions are available in e.g. [Young, 2010], [MIL-HDBK-17-1F, 2002]. However, the present treatment assumes the specific form of the distribution being sampled is unknown—as is very commonly the case in engineering practice, see e.g. section 3.2. The EON50% method is not studied because of its poor relative performance on the Normal PDF.

Table 3-5 and Figure 3-8 present the result for UQ methods applied to the Weibull distribution. As expected, all method capture success rates fall with added samples and the 95/90 TI and 95/90 EON results are very similar, as are the 95/95 TI and EON results. Several cases in the table show the EON reliability rate is equal to or slightly higher than the corresponding TI reliability rate, but the average TI score is slightly higher than the average score of the corresponding EON over the full set of cases N=2,4,10,20 (last column in the table).

The capture success rates for all methods are even lower (and significantly so at each sample size N) for the Weibull distribution than for the log-normal distribution. Reliabilities for the Weibull distribution are respectably high for all methods (>75%) for N=2 samples, but the drop in reliabilities with added samples is very dramatic for all methods and is again steepest for SD. At N=4 samples, the TI and EON 95/95 methods are in the low to mid 50% range (unacceptable) and all other methods have considerably lower reliabilities, < 35%. The Weibull distribution is significantly more skewed than the log-normal distribution (see plots in Appendix A) and appears to be quite problematic for the sparse-data UQ methods in the present study. For a reasonable reliability, one is limited to 2 or 3 samples and stuck with the potential for very large overshoot errors that could hamper design feasibility or perceptions of whether an existing system has adequate safety or performance margins.

Table 3-5. Empirical Reliabilities of Sparse-Data UQ Methods for capturing the central 95% range of a Weibull distribution sampled N times. Results from 10K random trials of each method.

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	75.4%	34.6%	1.6%	0.0%	27.9%
EON 90%	75.2%	31.9%	1.8%	0.0%	27.2%
95/95 TI	87.3%	56.2%	5.0%	0.1%	37.1%
EON 95%	87.3%	52.6%	5.6%	0.1%	36.4%
Super D.	76.9%	12.0%	0.1%	0.0%	22.2%

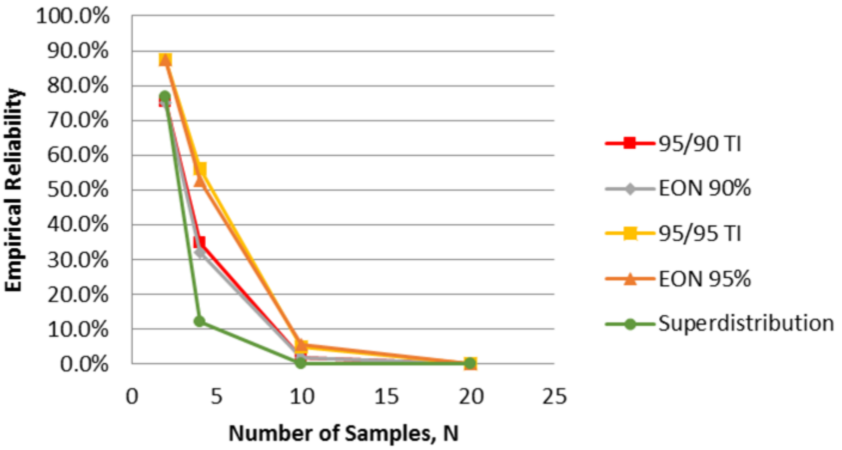


Figure 3-8. Empirical Reliabilities for capturing the central 95% range of a Weibull distribution sampled N times (plotted data from Table 3-5).

The left plot in Figure 3-9 shows the no-penalty metric results. The EON results are not plotted because they are very close to, but not as good as, the corresponding TI results on average, even though several individual EON results are better than their the corresponding TI results (see Appendix B Tables B.7). As with the log-normal distribution, metric values decrease with added samples (indicating improving overall performance) for all methods except SD in going from N=10 to 20. As with the log-normal distribution, the 95/95 TI method is the worst performer of all the plotted methods except at N=20 where SD is worst. The 95/90 TI method performs best at N=2,10, SD is best at N=4, and 95/95 TIs are best at N=20. In terms of average performance over the set N=2,4,10,20, the 95/90 TI method is best, then SD, then 95/95 TI.

The right plot in Figure 3-9 shows the 10X penalty metric results. The EON 90% and 95% results are not plotted because the corresponding TI results are very close to, but typically better than, the EON results per Table B.6, in particular the N=2,4,10,20 average metric values in the table's last column. For all methods, metric values decrease with added samples initially (indicating improving overall performance) but then at some point increase with added samples, presumably for the reasons stated earlier. The increase is most notable for SD, which increases beyond N=4 samples. Per the last column in Table B.8, 95/90 TIs average the best performance by far, then 95/95 TIs, followed by SDs.

The feasibility of all methods is limited to 2 to 3 samples for the Weibull PDF. In this regime, both the penalized and unpenalized performance measures show a clear preference for the 95/90 TI (or EON) method.

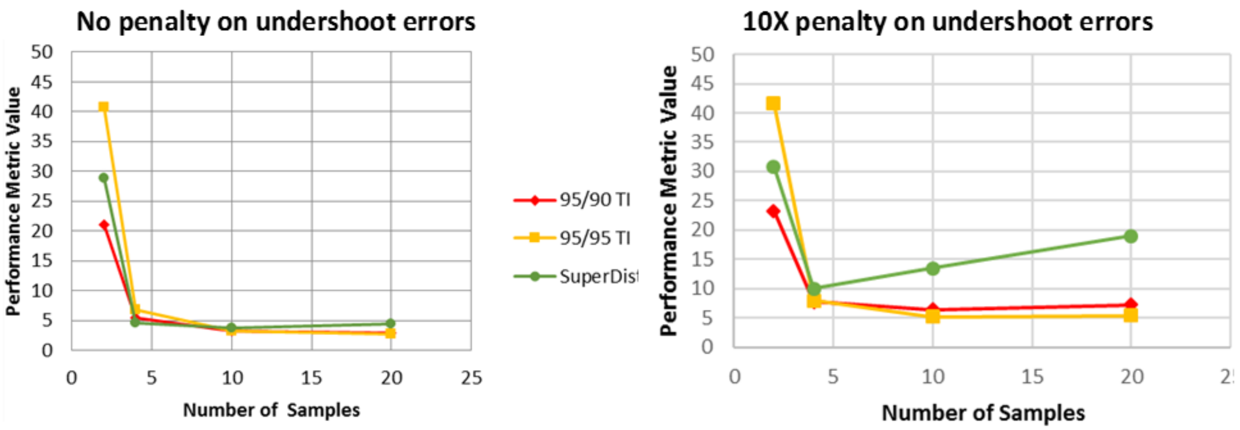


Figure 3-9. Performance Metric results vs. number of samples for Weibull distribution and unpenalized and 10X penalized shortfall errors. Results for each # of samples are from 10K TI trials with each method.

For this PDF type as well, all performance curves exhibit substantial knees at N=4 data samples for both unpenalized and 10X penalized metrics.

3.2. Performance on 70 Empirical Distributions

Romero, Schroeder, et al. (2017) examined the variability of 140 predicted responses when multiple stress-strain curves (reflecting variability from replicate material tests) are propagated through a finite element model of a ductile steel can being slowly crushed. The response quantities of interest (QOIs) include displacements, stresses, strains, and calculated measures of material damage. Each response quantity's behavior varies according to the particular stress-strain curves used for the materials in the model. The desire is to estimate response variabilitywhen only a few stress-strain (ss) curve samples are available from material testing. Like with random variables, propagation of just a few ss curve (random function) samples will usually result in significantly underestimated response variability relative to propagation of a much larger population that adequately samples the presiding random-function source of ss curves. Accordingly, 95/90 Tolerance Intervals were applied to the sparse realizations of response from propagation of small numbers of ss curves through the model, and performance of the TI method was characterized with similar procedures and metrics to those in Section 3.1 above.

The ss curves were synthetically generated to resemble real stress-strain curves and to enable generation of a large population (1000) that could be propagated through the model to form reasonably well resolved response distributions and associated statistics for the 70 QOIs. Ten thousand trials were conducted where, for each trial, N ss curves were drawn at random from the population of 1000, and each drawn curve was used in a can-crush simulation. N responses, one for each ss curve, were used to construct a 95/90 TI and the ends of the TI were compared to the reference 2.5 and 97.5 percentiles of that response resolved (to within small error accounted for as explained next) from the set of 1000 results for that QOI. Figure 3-10 reproduced from [Romero, Schroeder, et al., 2017] shows the 95/90 TI reliability results for 70 QOIs when ss curves corresponding to a temperature of 200C are used in the study.

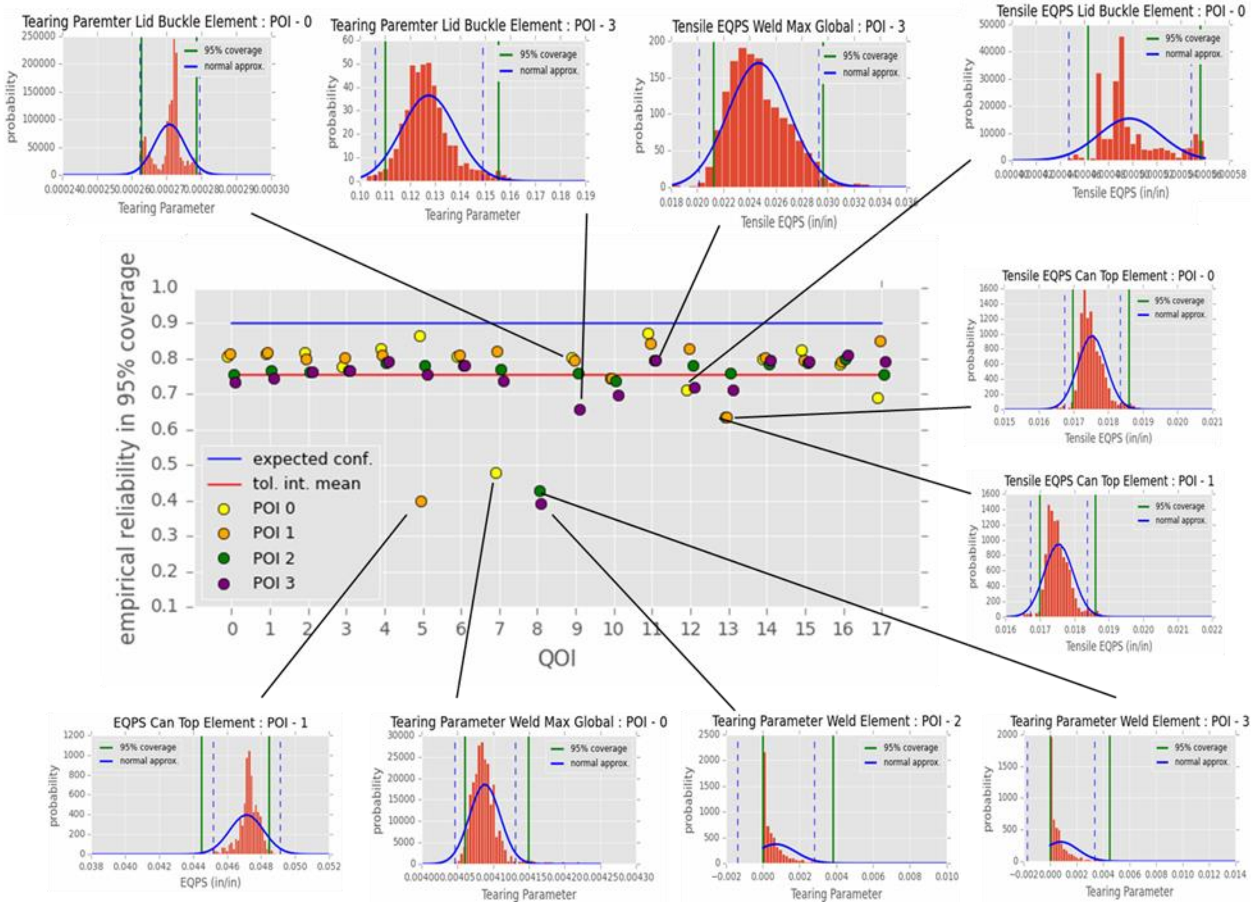


Figure 3-10. 95/90 TI success rates over 10,000 trials, and some of the 70 QOI response histograms, for N=4 stress-strain curves drawn at random per trial. All 70 histograms from this study are shown in [Romero, Schroeder, et al., 2017]

Two green vertical lines on each histogram in the figure mark the 25th and 975th ordered samples of the 1000 samples, providing nominal estimates of the 2.5 and 97.5 percentiles of the asymptotic distributions. Normal distributions using the calculated mean and standard deviation from the 1000 samples in each histogram are also plotted for reference. The 2.5 and 97.5 percentiles of the Normal distributions are signified by vertical blue dashed lines in the plots.

The offsets between the blue and green lines for the 2.5 and 97.5 percentiles correlate well with performance of the TI UQ method: smaller offsets correlate with better performance. In particular, it appears (and stands to reason) that TI success rates suffer when the 2.5 and/or 97.5 percentiles of a population (histogram) lie relatively far outside the corresponding percentiles of a Normal distribution based on the mean and standard deviation of the population.

TI reliability for capturing the central 95% of response between the 2.5 to 97.5 percentiles is 70% to 90% for most (62 of 70 = 89%) of the QOIs, with an average reliability of about 76% over all 70 QOIs. Average reliability was 79% for a second set of 70 QOIs in a similar study in [Romero, Schroeder, et al., 2017] for somewhat characteristically different ss curves corresponding to a temperature of 400C. These success rates are pleasingly high even though most of the 140 response histograms are highly non-Normal and the following factor biases the calculated success rates downward substantially.

Sampling uncertainty exists regarding the true 2.5 and 97.5 percentiles of a given QOI population because these percentiles are calculated from only M=1000 samples of response. Let P* be a finite-sample estimate of the true quantile or proportion P of response that lies beyond some threshold value. If enough samples are taken, then the estimate P* can be said with some percent likelihood or "confidence" to lie within a corresponding "confidence interval" of the true proportion, P. From [Devore, 1982], when the number M of total Monte Carlo (MC) samples meets the condition

$$M \bullet P \geq 5 \qquad (3\text{-}7)$$

then the following formula for 95% confidence intervals (CIs) applies:

$$|P - P^*| \leq 1.96[P(1-P)/M]^{1/2} \, . \qquad (3\text{-}8)$$

The condition (3-7) is met for both the 2.5 and 97.5 percentiles of interest here. Substituting either P=0.025 or P=0.975 into Equation (3-8) yields $|P - P^*| \leq 0.01$. Thus, the 95% CI on the estimated quantile P*=0.025 indicates that the estimate is within ± 0.01 or ± one quantile of the true 0.025 quantile, with 95% confidence. The response QOI value corresponding to the true 2.5 percentile is therefore 95% likely to lie between the QOI values of the calculated 1.5 and 3.5 percentiles. Likewise, the response value corresponding to the true 97.5 percentile is estimated to lie between the QOI values of the calculated 96.5 and 98.5 percentiles. For the results in Figure 3-10 we form a "conservative" or "pessimistic" range denoted by calculated percentiles 1.5 to 98.5 as the reference range to compare the trial 95% TIs against. We could also form a "non-conservative" or "optimistic" reference range for comparison by using the range between the 3.5 and 96.5 percentiles. The pessimistic/conservative choice reduces the estimated reliability rate by about 7 percentage points vs. the optimistic/non-conservative choice when N=4 samples are used, per the plot in Figure 3-11 reproduced from [Romero, Schroeder, et al., 2017]. The plot shows that the calculated performance difference increases to about 16 percentage points at N=10 samples. These pessimistic reporting choices bias the reported success rates to substantially reduced values.

When capture of any 95% range (such as 0.01 to 0.96) of the 70 PDFs are counted as TI successes, the top curve in Figure 3-11 shows the nominal success rates. These are calculated using the nominal 95% coverage value and not the conservative 97% or non-conservative 93% values just discussed above. The curve's nominal reliability rate for N=4 samples is about 85%. This is about six percentage points higher than the 79% rate that is most directly comparable—which is approximately midway between the conservative and non-conservative central 95% curves at N=4 in the figure. The differential increases to about twelve percentage points at N=10. Thus, there is significant added conservatism from counting just the central 95% response range vs. any 95% range in determining TI success rates.

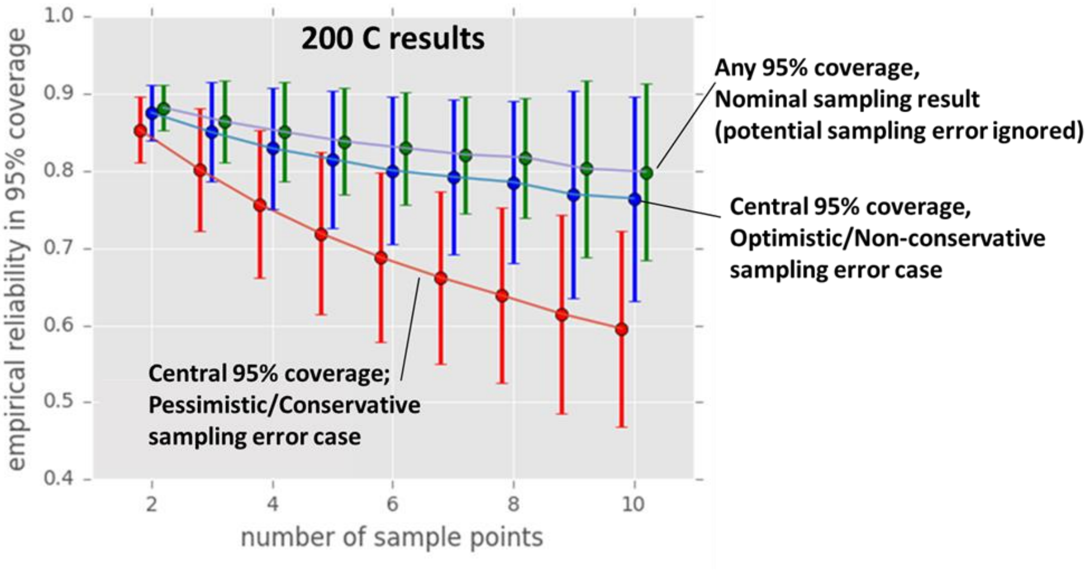


Figure 3-11. Variation of TI method success rates versus the number N of 200C stress-strain curves drawn at random per trial (e.g. N=4 in Figure 3-10), and for conservative and non-conservative 2.5 to 97.5 percentile ranges discussed, and for non-central 95% ranges also counted as successes in the green curve. Dots are averages over the 70 200C QOI cases, uncertainty bars are ± 1 standard deviation of the 70 individual success rates.

Even if the optimistic/non-conservative choices are used to define the reference 95% ranges that the trial TIs are tested against, the reliability rates indicated by the green curve in Figure 3-11 are still slightly below the pessimistic success rates for 0.95/0.90 TI in Figure 3-1 and Figure 3-4 for the Normal and 5 DOF t distributions. So on average, the 0.95/0.90 TIs do less well on the 70 distributions than on the Normal and t distributions. On a more even basis of comparison, the 87.3% reliabilities of 0.95/0.90 TIs for central 95% capture and N=4 in tables 3-1 and 3-3 are compared against the central 95% capture reliabilities shown in Figure 3-10 for the 70 individual PDFs and N=4. Only two of the 70 PDFs show TI success rates at or above 87%. Even if all dots in the figure are moved upward by seven percentage points so they reflect the average benefit in Figure 3-11 of optimistic instead of pessimistic rates of central 95% capture for N=4, the average reliability would go from 76% to 83%, still below 87% reliability rates for the Normal and 5 DOF t distributions. We can therefore conclude that that the majority of the 70 response PDFs in the can-crush study are more difficult than Normal and 5 DOF t distributions for the 95/90 TI method to be successful on.

Whether the optimistic 83% reliability or the pessimistic 76%, the average reliability for the 70 PDFs is slightly higher than the 74% reliability for the log-normal PDF (N=4, 95/90 TI method in Table 3-4). This loosely supports an argument that the "average" PDF in the can-crush study is somewhat easier than a log-normal distribution for the 95/90 TI method to handle well. The two most difficult PDFs for the TI method engender ~40% pessimistic reliabilities (in Figure 3-10) or 47% optimistic reliabilities. These are far better than the Weibull distribution, which engenders 35% reliability for 95/90 TIs and N=4 (Table 3-5). Thus, the Weibull can be said to bound the 70 PDFs in terms of difficulty for the 95/90 TI method.

To get performance metric values for all 70 PDFs on a similar scale, a given trial TI's errors $\epsilon\ell$ and ϵu in Figure 2-1 are normalized by dividing them by the QOI's mean value from its 1000 histogrammed results. Figure 3-12 shows corresponding performance metric results for the 70 PDFs individually and on average over all PDFs. For the average metric curves, performance metric Equations 3-4 to 3-6 also apply for a larger averaging process over the 700K normalized results of all 70 QOIs (10K results per QOI).

The two performance curves far apart from the others in the left and right plots correspond to the two right-most distributions in the bottom row of Figure 3-10. These distributions are highly skewed, with a large build-up of realizations against a physical lower limit of possible values these two quantities can take (a lower possible limit of zero material damage at this point in the can crush event). When taking N samples of these two PDFs for the 10K trials, many trials contained multiple samples of zero response. These many trials yielded very short TIs centered toward zero, which did not capture the PDFs' 97.5 percentiles of response and had relatively large shortfall errors there and large overshoot errors at the zero end of the PDF. Thus the low capture success rates in Figure 3-10 and the poor performance metric results in Figure 3-12.

When dealing with cases like this where distributions of results are expected to butt up against known upper and/or lower response constraints, without knowing specific distribution shapes, restricted distributional forms like exponential or log-normal suitable for these cases should be used for determining tolerance intervals (see e.g. [Young, 2012]).

For the unpenalized metric, performance for these two PDFs increases with added samples until N=7. More samples add essentially no value, causing the metric value to remain essentially flat or increase slightly. In fact, a distinct knee at N=3 samples exists beyond which only very marginal benefit occurs with added samples. For the 10X penalized metric, performance for the two PDFs decreases immediately with added samples beyond N=2. The effect of shortfalls from the true 97.5 percentile are magnified with this metric. This overwhelms any decrease in overshoot errors and/or increase in capture success rates that would appear to be causing the decrease in the no-penalty metric with added samples beyond N=2.

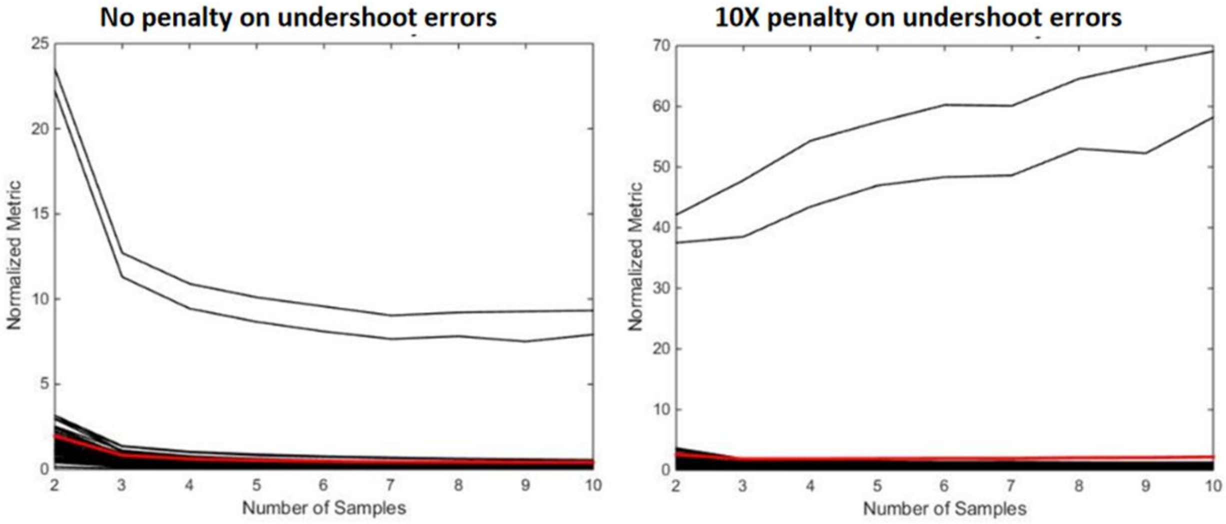


Figure 3-12. Performance Metric results vs. number of samples for 70 can-crush response PDFs and unpenalized and 10X penalized shortfall errors. Black curves are from 10K 95/90 TI trials for each PDF. Red curve is the average of the 70 curves. Note different vertical scales in plots.

Zoomed versions of the plots in Figure 3-12 are shown in the upper left plots in Figure 3-13 and Figure 3-14. These figures compare 95/90 TI performance against 95/95 TI, SD, and EON50% methods. Note that the vertical scales are not the same in all plots in these figures. The plots zoom in on the 68 non-outlier performance curves to better examine them. (The two outlier performance curves exist for all methods but are outside the frames of all zoomed plots in Figure 3-13 and Figure 3-14 except for the lower-left plot in Figure 3-14 where they are partially visible at upper-right in the plot.)

In all plots the red average performance curves for the 70 PDFs are strongly affected (pulled up) by the two outlier curves outside the plot frames. The red average curves are significantly higher than visually estimable average curves of the 68-curve populations in each plot. The two outliers were not included for the performance metric plots in [Romero, Schroeder, et al., 2017], but are included in the present paper to reflect the effects of these more difficult PDF shapes.

In Figure 3-13, the red curve in the 95/90 TI plot decreases continually with added samples (see Table B-9 in Appendix B), as do all 68 individual curves (it appears). The red average curve and all the individual curves have a strong knee at N=3 samples, beyond which improvement is marginal with added samples. The SD results are plotted on the same vertical scale. Individual and average SD and 95/90 TI curves are fairly close in height. Detailed comparisons of the average curves will be made later. The bulk of individual SD curves appear to decrease with added samples, although not all individual curves appear to, and the average curve certainly does not—which appears to be due to large influence from the two outliers. The individual and average SD curves have a strong knee at N=3 samples; improvement with more samples is very marginal or even negative after N=4 to 5 for the two outliers and after N=5 for the red average curve.

The 95/95 TI results are plotted on a different vertical scale. Individual and average curves are higher than for SD and 95/90 TI except for more than about N=7 samples. For this method as well, the individual and average curves have a strong knee at N=3 samples; improvement with more samples is very marginal but does continue to the end of the plot (N=10) for the red average curve (see Appendix B Table B-9) and for the 68 individual curves (it appears).

The 'Mean Ensemble of Normals' plot in Figure 3-12 uses the 95/50 EON (EON50%) method explained earlier, which is approximately a 95/50 TI method per Figure 2-6. Noting the vertical scale in the plot, the EON50% method performs relatively well on the 68 PDFs, though found to significantly underperform the other methods for Normal PDFs so was not tried on the other analytic PDFs in Section 3.1. But the red average curve for all 70 empirical PDFs is only lower/better than the other methods at N=2. The average performance and most of the individual performances do not improve significantly beyond N=2 samples like for the other methods.

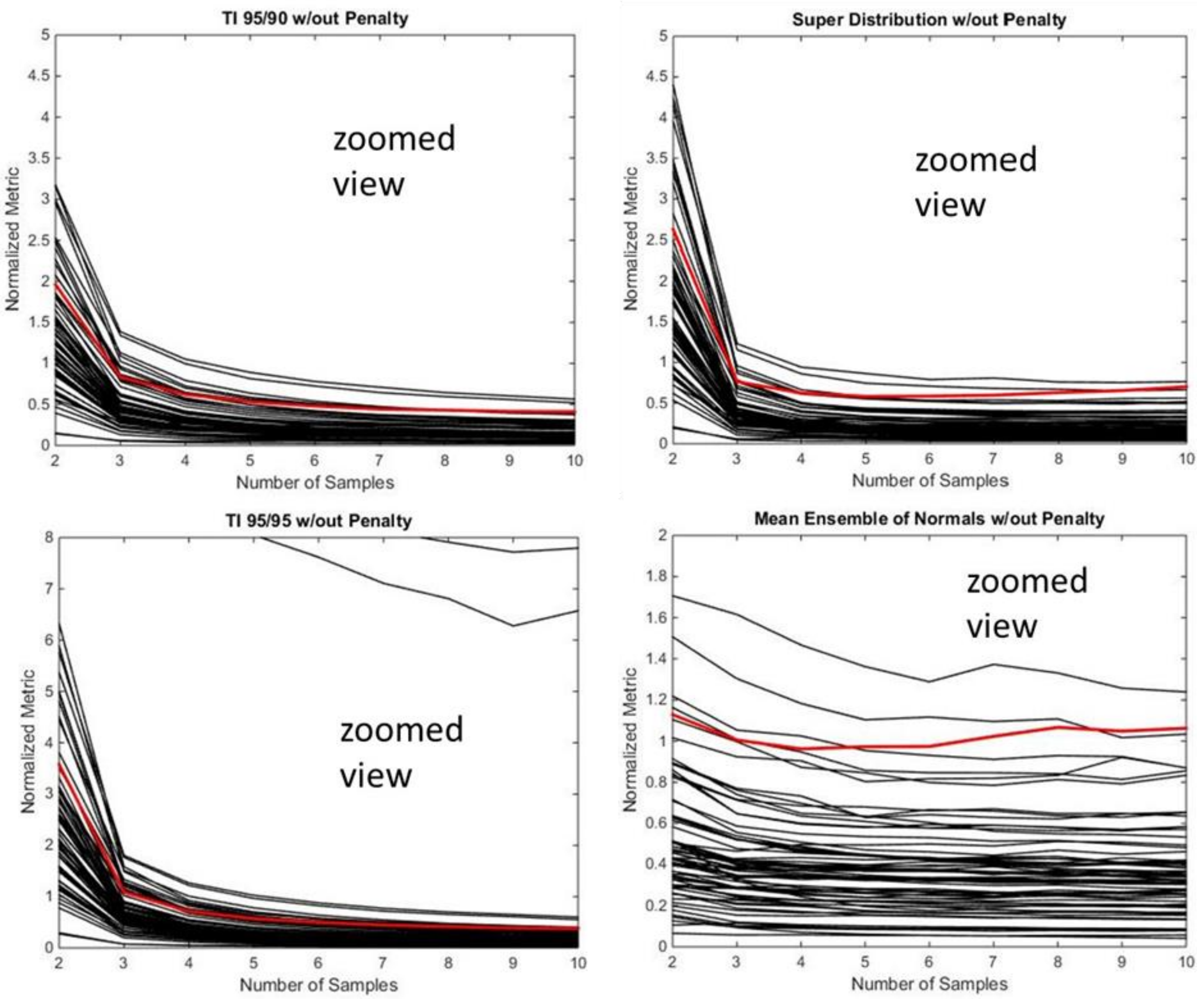


Figure 3-13. Performance Metric results vs. number of samples for 70 can-crush response PDFs and unpenalized shortfall errors. All figures are zoomed versions that relegate all or most of the two outlier curves to lie outside the plot frames in order to show more detail of the other curves. Vertical scales are not the same in all plots.

In Figure 3-14 all the red average curves for all methods have higher 10X penalized metric values at all sample sizes N than their counterpart average curves in Figure 3-13 with unpenalized metrics. The curve populations are also generally higher in metric value in Figure 3-14 than in Figure 3-13 as expected, but not to the same degree as the average curves. The two outlier cases have outsized effects under the 10X penalized metric compared to the unpenalized metric (e.g. compare the >2X different vertical scales of the left and right plots in Figure 3-12). These outsized effects have outsized impact on the red average curves in Figure 3-14, which are noticeably more different from their populations of 68 individual curves than in Figure 3-13 with the unpenalized metric. Indeed, for 95/90 TIs the average curve shows a minimum (best performance) at N=3, then worsens with more samples. But the 68 individual curves appear to generally show continued improved performance with added samples.

A similar story exists for 95/95 TIs except that the average curve shows a minimum (best performance) at N=5, then worsens with more samples. In contrast, the SD method's individual results generally appear to worsen with added samples beyond N=3, as does the average curve.

Furthermore, the SD average curve and the population of curves lie at higher/worse metric values than the 95/90 and 95/95 TI results. In this respect we can use just the average curves to compare the relative performances of the SD, 95/90, and 95/95 TI methods. This goes for the EON50% method as well. The EON50% method performs relatively poorly compared to the other three methods, per the 68 individual curves and the average curve (note the ~2X different vertical scale on the EON50% plot compared to the other three).

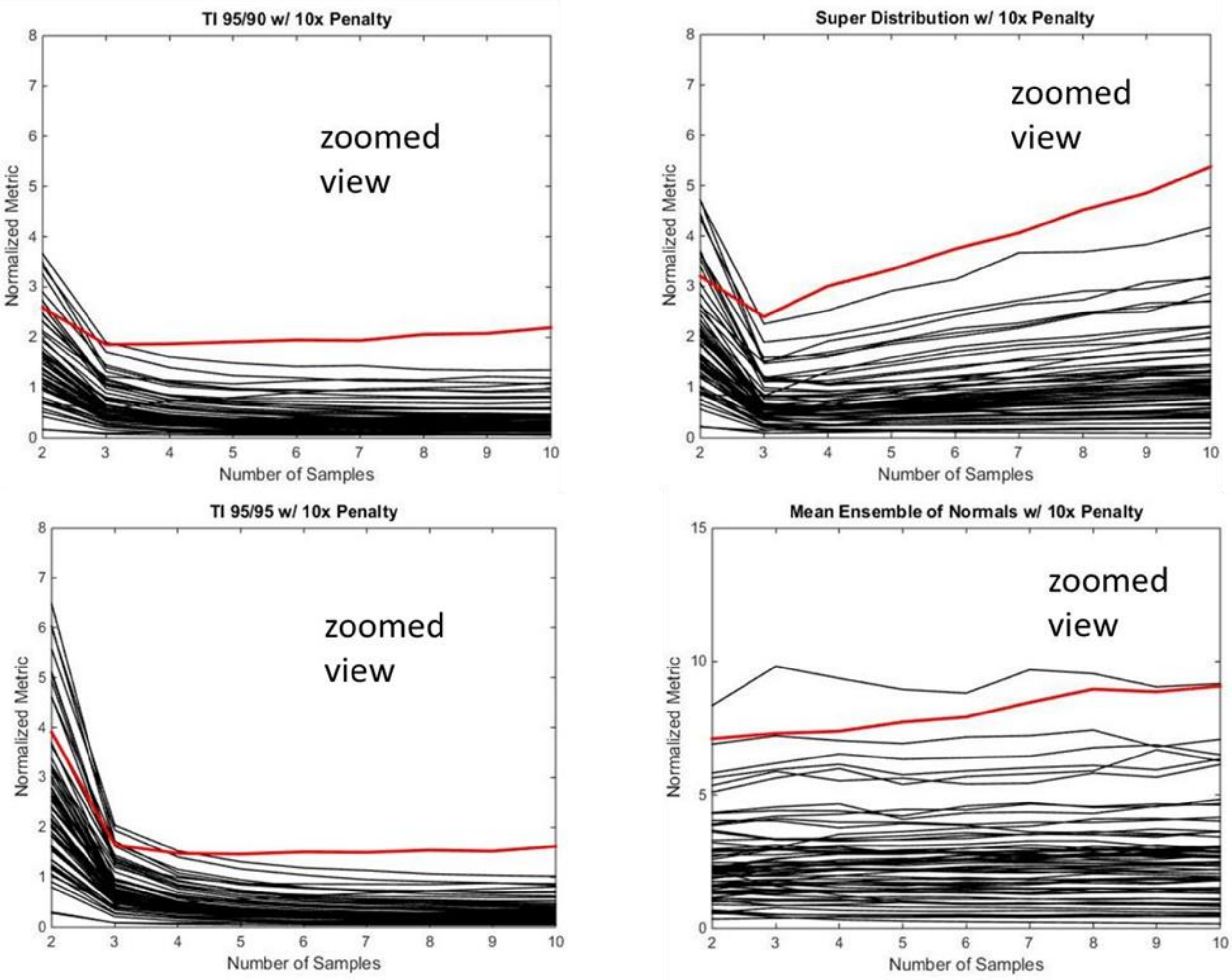


Figure 3-14. Performance Metric results vs. number of samples for 70 can-crush response PDFs and 10X penalized shortfall errors. All figures are zoomed versions that relegate the two outlier curves to lie outside the plot frames in order to show more detail of the other curves. Vertical scales are not the same in all plots.

Figure 3-15 consolidates the four methods' average performance curves from Figure 3-13 and Figure 3-14. The plotted average performance data in Figure 3-15 is given in tables B.9 and B.10 in Appendix B. These consolidated results are discussed in the next subsection.

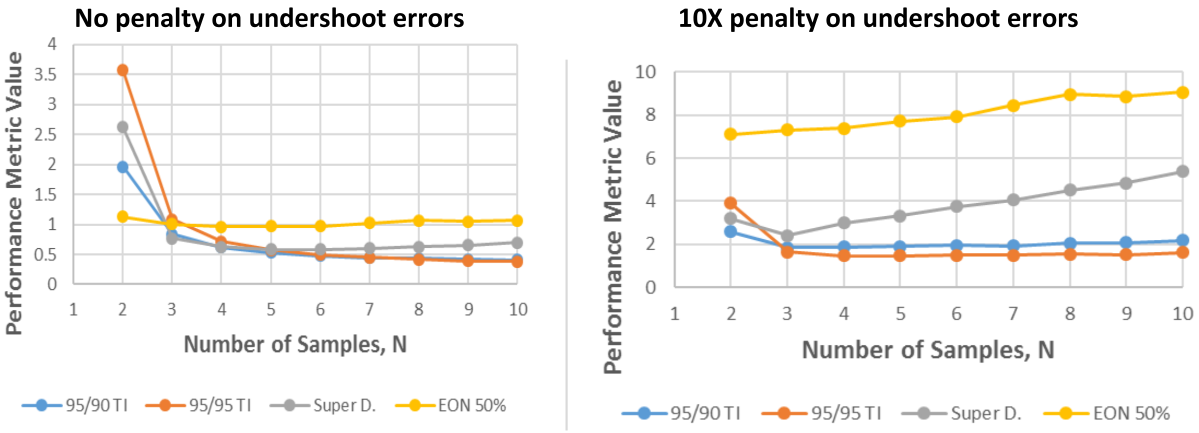


Figure 3-15. Average Performance Metric results vs. number of samples for 70 can-crush response PDFs and unpenalized and 10X penalized shortfall errors. (These are the red curves from Figure 3-13 and Figure 3-14). Note different vertical scales in plots.

3.3. Discussion, Conclusions, and Recommendations for Capture of Central 95% of Response

For the unpenalized performance metric plot at left in Figure 3-15, the EON50% method is dominated by the other methods except at N=2 and 3. At N=3 the EON50% method performs worse than two other methods: 95/90 TIs and SD. At N=2 the EON50% method performs best of all methods. However, this is somewhat misleading because the relatively good performance at N=2 appears to be more a result of large relative overshoot errors of the other methods, rather than desirable performance of the EON50% method. Indeed, even though the method has relatively low non-penalized performance values at N=4,10,20 in Figure 3-3 (left plot), the EON50% method simultaneously has unacceptably low capture success rates in Figure 3-1. The EON50% method produces shorter prediction intervals than the other methods with a given set of samples. The resulting low success rate and more and greater undershoot errors are offset in the metric by the small relative overshoot errors. The worsening capture success rates and shortfall errors as samples are added appear to approximately evenly counteract the decreasing overshoot errors such that overall performance remains flat in the left plot in Figure 3-15. Consistent with these observations, the EON50% method scores far worse than the other methods at all sample sizes in Figure 3-3 and Figure 3-15 using the performance metric with 10X penalized shortfall errors. Hence, the EON50% method is substantially less fit than the other studied methods for general sparse-data problems.

The Superdistribution method is considered next. SD performs about the same as, or significantly worse than, 95/90 TIs considering penalized and unpenalized metrics and all sample sizes in Figure 3-15. A similar observation holds for the analytic PDFs studied. SD performance in Figure 3-15 gets worse on average for the 70 PDFs with more samples than N=5 for the unpenalized metric and N=3 with the penalized metric, whereas 95/90 (and 95/95) TI performance improves or holds approximately flat with added samples (for both versions of performance metric). SD results are also considerably more involved to calculate than TIs. For these reasons, 95/90 TIs are considered preferable to SDs for bounding the central 95% of a generic PDF.

The 95/90 TI method has been argued in the prior two paragraphs to perform better on average than SD and EON50% methods on the 70 PDFs. The same is true for the analytic PDFs studied (all things considered). We now examine the absolute and relative performances of 95/90 and 95/95 TIs to determine which, if either, is preferable.

For N=5 to 10 samples in the left non-penalized metric plot in Figure 3-15, average performance is effectively equivalent for 95/95 and 95/90 methods applied to the 70 PDFs. From this performance metric the 95/95 method's larger overshoot errors and commensurately larger capture reliabilities and smaller/fewer shortfall errors are an approximately equivalent tradeoff with the 95/90 method's smaller overshoot errors but commensurately worse capture reliabilities and shortfall errors. For N=2 to 4 samples, 95/90 TIs score better than the 95/95 TIs. Presumably the outsized overshoot errors of the 95/95 method at very sparse samples overwhelm its other advantages and overall performance is worse than 95/90 TIs according to the unpenalized metric. But when shortfall errors are penalized by 10X, the region of the 95/90 method's advantage reduces to the lower limit of N=2 samples. The right plot in Figure 3-15 shows a slight 95/95 performance advantage for N > 2 samples. The relative average effectivnesses of the 95/90 and 95/95 TI methods on the 70 empirical PDFs are appreciably similar on the analytic PDFs. Whether to use a 95/90 or 95/95 TI in a given instance comes down to the following observations and suggested strategy.

If one is fairly sure that the distribution being sampled is symmetric or approximately so, then for a desired capture reliability such as 90%, the 5 DOF t distribution reliability results in Figure 3-4 can be used to determine the number of samples to attain this reliability (or nearly so for a foreseeably large subset of symmetric PDFs, though this has not been investigated beyond uniform PDFs in [Romero, Swiler, et al., 2013]). Figure 3-4 indicates 90% reliability with N=2 samples for 95/90 TIs or N=8 for 95/95 TIs. The TI multiplier factor for 95/90 TIs is 18.56 in Table 2-1.The multiplier for 95/95 TIs and N=8 is only about 3.7. This gives far smaller TIs to work with in design, model calibration or validation, etc. so avoids unnecessary and costly conservatism in attaining the same 90% capture reliability that 95/90 TIs with N=2 would. The downside, of course, is the cost of six additional samples with the 95/95 method. Ultimately, project resources and circumstances would presumably dictate which route is the most feasible to achieving the desired risk level regarding capture reliability. Also note that benefits continue to accrue with added samples for TI methods and the symmetric PDFs studied, but a sharp knee in benefit vs. sampling cost exists at N=4 samples, beyond which far less marginal benefit accrues with added samples.

If one is not fairly sure that the distribution being sampled is symmetric or approximately so, then the following strategy is suggested. The 140 empirical PDFs give a pessimistic or conservative reliability vs. N curve represented by the red curve in Figure 3-11. Variation of TI method success rates versus the number N of 200C stress-strain curves drawn at random per trial (e.g. N=4 in Figure 3-10), and for conservative and non-conservative 2.5 to 97.5 percentile ranges discussed, and for non-central 95% ranges also counted as successes in the green curve. Dots are averages over the 70 200C QOI cases, uncertainty bars are ± 1 standard deviation of the 70 individual success rates.. (This curve has slightly lower reliabilities vs. N (is conservative) compared to a similar curve shown in [Romero, Schroeder et al., 2017] for a second set of 70 PDFs.) For an example minimum acceptable capture reliability of 70% (see Footnote 2), the lowest number of samples that can be taken is N=6 from the red curve in Figure 3-11. For N=6 the TI multiplier is 3.76 from Table 2-1. This is compared to the multiplier for 95/95 TIs determined as follows.

The 95/90 reliability vs. N relationship in Figure 3-6 for the log-normal PDF conservatively bounds (from below) the average 95/90 TI reliability for the 140 empirical PDFs (conservatively represented by the red curve in Figure 3-11 as just explained). Figure 3-6 indicates 10 or more percentage points higher reliability with the 95/95 method than with the 95/90 method for N ≥ 4 samples. We use this performance differential on the log-normal PDF as an estimator for the average performance differential that would exist between 95/90 and 95/95 TI methods applied to the 140 PDFs. Applying a +10 percentage point shift to Figure 3-11 for N ≥ 4 yields an estimate that N=10 samples would yield ~70% average capture reliability with 95/95 TIs. The associated TI multiplier is 3.4. This is a somewhat smaller/better multiplier than 3.76 with N=6 and 95/90 TIs determined in the preceding paragraph, but involves more tests or model simulations. Again, the best choice between these two options will depend on what is most feasible given project resources and circumstances.

Thus, if one is willing to risk that a PDF being sampled has a central 95% capture difficulty that is equal to or less than the average difficulty of the 140 highly diverse and challenging PDFs in the empirical study (see examples in Figure 3-10), one could figure the most appropriate number of samples and smallest TI multiplier to achieve a desired capture reliability using the red curve in Figure 3-11 for 95/90 TIs and a +10 percentage point shifted version of that curve for 95/95 TIs. The large majority of the 140 PDFs (89%) in the empirical study enjoyed TI capture reliabilities ≥ these average reliability curves. See for example the highly non-normal and even multi-modal PDF shapes in Figure 3-10 that lie near or above the average reliability line. This large study suggests at the very least a high *plausibility* that the methodology described in the paragraph immediately above can be used to figure the most appropriate number of samples and smallest TI multiplier to achieve >70% capture reliability. If one wants to be more conservative, the Log-Normal results in Figure 3-6 or the Weibull results in Figure 3-8 could be used with the following considerations.

The last paragraph in the Log-Normal subsection indicates that >70% reliability is achieved with N=4 samples or less with 95/90 TIs. This is similar to average reliability of the 140 PDFs with N=4 and 95/90 TIs. The L-N PDF is proposed as a convenient surrogate for the 140-PDF average reliability at N=4 samples (established at N=4 for 95/90 TIs and a presumed surrogate for 95/95 TIs as explained two paragraphs above). As established, about 89% of the 140 PDFs have reliabilities of 70% or greater. Thus, capture of the central 95% of the Log-Normal PDF is as difficult or more difficult than for ~90% of the 140 empirical PDFs studied.

For N > 4, the L-N PDF can reasonably be concluded to be more difficult than 90% of the 140 PDFs because the L-N reliability rate drops more quickly vs. N than the average reliability rate does for the 140 PDFs (compare the 95/90 TI curve in Figure 3-6 with the red curve in Figure 3-11).

For N < 4, the L-N PDF is estimated to be more difficult than about 85% of the 140 PDFs. (The average reliability rate for the 140 PDFs and the L-N PDF is the same ~85% at N=2 and an assumption is made that the percentage of the 140 results above and below the average is similar at N=2, 3, and 4).

Thus, the analytic Log-Normal PDF and its 95/90TI and 95/95 TI results in Figure 3-6 are proposed as a convenient conservative estimator of TI reliability if one is willing to risk that the PDF being sampled has a central 95% capture difficulty less than 85% to 90% of the 140 empirical PDFs studied. The substantial test-bed of 140 diverse and challenging PDF shapes makes it *highly plausible*, even *reasonably credible* that the desired capture reliability is achieved if figuring # of samples from the log-normal 95/90 and 95/95 reliability curves when the PDF being sampled is unknown.

Very strong *belief* or *credibility* may require use of the Weibull distribution as a severe case (second most challenging in terms of capture success among the 144 analytic and empirical PDFs studied). As explained in the Weibull subsection, acceptable reliability levels limit the number of allowable samples to N=2 or 3. The 95/90 method performs best in this regime according to the no-penalty and 10X penalty metrics. N=2 samples yields a capture reliability of 75% but gives a very high multiplier of 18.6. A preferential strategy may be to use N=3 samples with 95/95 TIs. This yields a slightly lower capture reliability of 70% but a considerably smaller (less prohibitive) multiplier of 9.9. In any case, use of the Weibull PDF as a severe case is a very risk averse strategy and comes with the downsides of large bias toward conservatism. There is a remarkable irony here of needing to keep the number of samples very low (two or three) in this very risk-averse strategy when little to nothing is known about the PDF being sampled.

4. METHODS' PERFORMANCE FOR BOUNDING 10^{-4} TAIL PROBABILITIES OF PDFS

In this section, related sparse-data UQ methods are characterized for accuracy and efficiency in bounding the exceedance probability (EP) in PDF tails integrated beyond prescribed limits that yield EP = 10^{-4}. A diverse and challenging set of 12 analytic and empirical PDF shapes are employed in the study.

For TI Equivalent Normals (TI-ENs) exemplified in Figure 2-1 and Superdistributions exemplified in Figure 2-4, a single EP is yielded when integrating the PDF above or below a specified threshold value. When dealing with Ensemble of Normals representations of uncertainty exemplified in Figure 2-4, each Normal PDF in the ensemble yields an EP estimate, so a distribution of EP estimates occurs as depicted in

Figure 4-1. Each EON in the study involved L=100 Normal distributions, so each CDF in the study, exemplified at right in the figure, is constructed from 100 EP estimates.

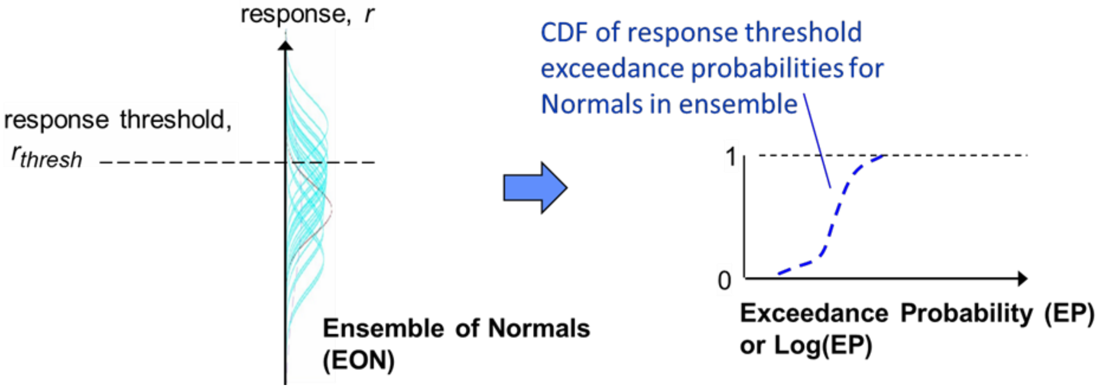


Figure 4-1. CDF of Exceedance Probabilities calculated from the Ensemble of Normals and a specified threshold level of response.

Example PDF Results and Multi-Attribute Performance Metric for Exceedance Probability

One of the test problems for characterizing exceedance probability estimation of the sparse-data methods is shown in Figure 4-2. For our study to be relevant to risk and reliability analysis we characterize estimation accuracy for very small exceedance probabilities (10^{-4}) in the tails of the test PDFs. The empirical histograms from the can-crush problem (specifically the histograms in Figure 3-10) presumably provide a challenging set of distribution shapes to test the UQ methods. Kernel density (KD) fits to selected histograms from Figure 3-10 are determined in Matlab® and these are normalized to PDFs of unit integrated area. The 0.9999 quantiles of the PDFs are determined from 10^{6} samples, which are enough to obtain negligible error for the study's purposes. The QOI response value at the 0.9999 quantile is the critical response value to the right of which the PDF tail integrates to 10^{-4} exceedance probability. Figure 4-2 shows how relatively small the PDF area is that lies to the right of the vertical dashed line and integrates to 10^{-4}

probability. Each PDFs critical response value, so determined, provides the integration limit for EP estimates with the UQ methods.

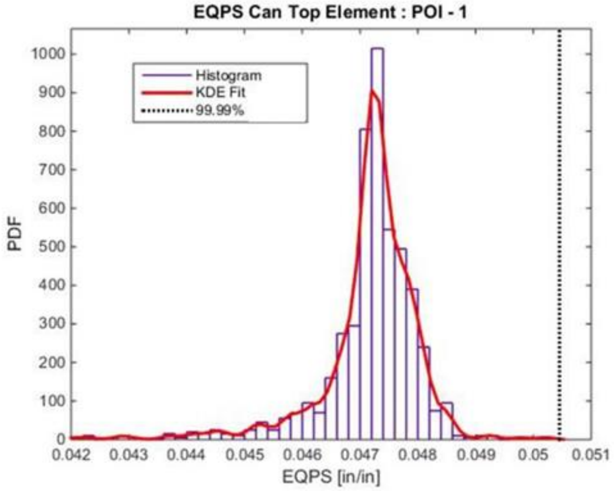


Figure 4-2. Kernel Density fit to a sample empirical histogram of response from can-crush UQ study (this is the QOI at bottom left corner in Figure 3-10). Vertical dashed line marks 0.9999 quantile of a PDF from the Kernel Density fit normalized to a unit integrated area.

Figure 4-3 shows distributions of EP estimation errors for 10,000 trials of each method with N=2,4,10,20 random samples per trial. The plot abscissas quantify estimation errors in terms of the number of orders of magnitude difference from the exact EP of 10^{-4}. Note that the TI 95/90 results lie below the curve labeled 'TI Extended' [Romero & Weirs, 2018], which method was apparently mishandled in the study and is ignored in this paper. The results labeled 'EON 90%' are the 90th percentile result in each trial, where each trial yields a distribution of results per Figure 4-1. Although the distribution of EON 90% results in Figure 4-3 is closest to the TI 95/90 distribution, there are significant differences; the relationship between the two methods' EP results is not directly the same as in Figure 2-6 for percentile results, although the correlation is high.

One thing to notice in Figure 4-3 is that the large majority of N=2 results are all over-estimates of EP, highly concentrated about + 3 to 4 orders of magnitude error. This means the EP estimates were concentrated about values 10^{-1} to 1 for all methods. The peak and average of the SD errors are at about 10^{-1}, while the other distributions' peaks and averages are closer to 1. Thus, the SD method is on average about an order of magnitude more accurate than the other methods for this PDF and N=2 samples. SD also out-performs at the other samples sizes; its error distribution always appears by eye to have a mean and peak closest to zero and to be more compact in terms of its spread of values. For all methods, the average error appears to typically get better as sample size N increases (note that the same horizontal scales but differing vertical scales in the four plots of the figure). But as average error typically decreases with sample size, each error distribution's proportion to the right of zero typically also decreases, indicating typically declining reliability for conservatively bounding the true EP (from above). Table 4-1 confirms this. Each method's empirical reliability is listed vs. sample size N. Reliabilities are the proportion of trials (by counting, not by integrating PDFs of results) in which the EP estimage exceeds the true EP. The reliabilities range from a high of ~98% (SD method) at N=2, to a low of ~67% (SD method) at N=20.

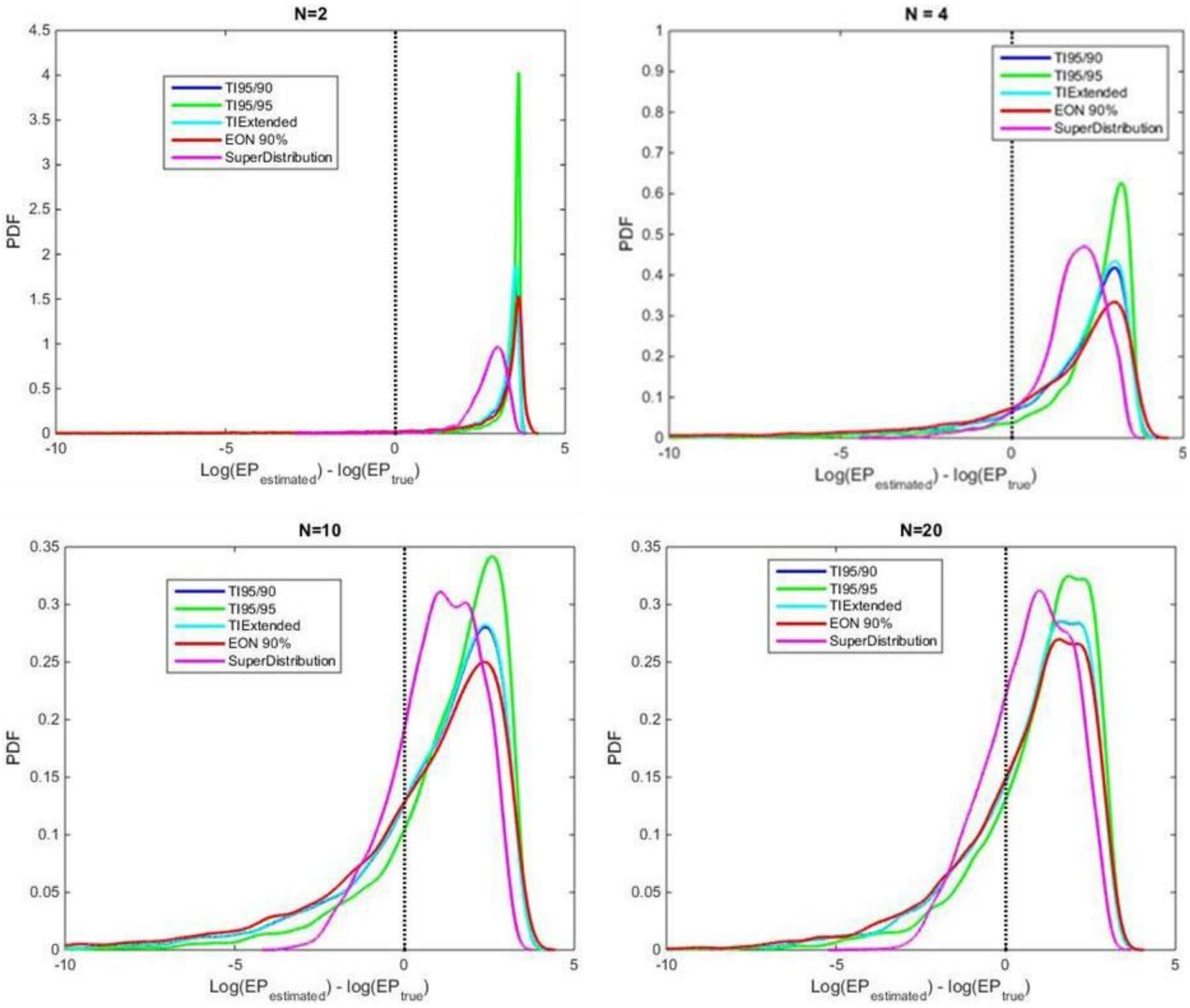


Figure 4-3. Exceedance Probability estimation results in terms of # of orders of magnitude error from exact 10^{-4} for 10K trials with sparse-data UQ methods and N=2, 4, 10, 20 random samples per trial. Note: TI 95/90 results lie below distribution curve labeled 'TI Extended', which method appears to have been inadvertently assigned the 95/90 results in the plotting procedure.

Table 4-1. Empirical Reliabilities of Sparse-Data UQ Methods for Conservative Exceedance Probability estimation (see Figure 4-2)

N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	87.6%	93.8%	85.3%	98.2%
4	80.4%	88.5%	74.8%	90.4%
10	71.9%	80.2%	67.5%	73.6%
20	72.1%	78.4%	69.9%	66.9%

Although reliability of conservatively bounding the actual 10^{-4} EP typically declines with increasing samples, the overshoot and undershoot errors also typically decline for this PDF. This causes generally improving performance metric results as shown in Figure 4-4. The performance metrics for EP estimation are described next and then the metric results in the figure are discussed.

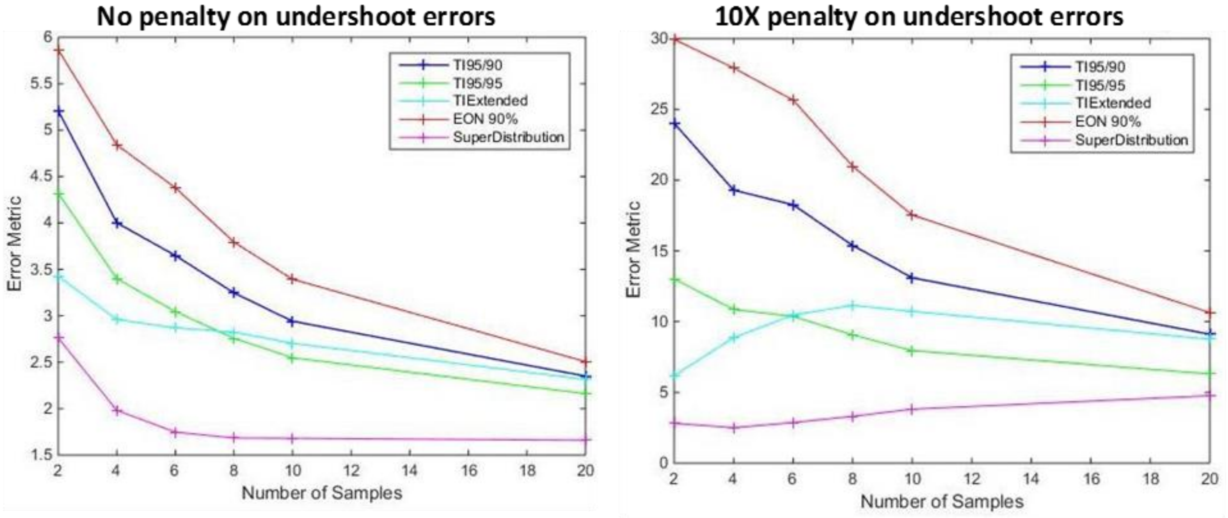


Figure 4-4. EP Performance Metric results vs. number of samples for PDF in Figure 4-2 and unpenalized and 10X penalized undershoot errors as explained in the text. Results for each # of samples are from 10K TI trials with each method. Note: Ignore TI Extended results, which are suspect.

A method performance metric that penalizes overshoot or undershoot errors of the same magnitude equally is

$$\text{EPmetric} = [\textstyle\sum^{N+} \Delta log \ \ + \sum^{N-} |\Delta log| \ ]/N^{+} \tag{4-1}$$

where

$$\Delta log = \text{abscissa in the error PDF plots} = \ log(EP\_estimated) - \ log(EP\_true) \tag{4-2}$$

and N^+ and N^- are the numbers of overshoot and undershoot cases respectively in the total number of trials = $N^+ + N^-$. For a given numerator sum of overshoot and undershoot error magnitudes in Eqn. 4-1, the greater the number of overshoot errors contributing to that sum (and so the smaller the number of contributed undershoot errors), the better the method performance would be because the proportion of conservative bounding cases would increase vs. unconservative/undershoot cases while the total error over all the trials remains the same. This is reflected in a lowering of the metric value because the numerator stays constant but the denominator increases. Again, as with the performance metric for central 95% PDF capture, *smaller metric values imply better method performance*.

If undershoot errors are given a 10X magnitude amplification as in Section 3 to reflect that a (non-conservative) undershoot error is considered much worse than a (conservative) overshoot error of the same magnitude, then the performance metric becomes

$$\text{10X penalty EPmetric } = [\sum^{N+} \Delta log \; + 10\sum^{N-} |\Delta\text{log}| \,]/N^{+}. \qquad (4\text{-}3)$$

For the same set of overshoot and undershoot errors from a given set of trials, the numerator value in Eqn. 4-3 yields a higher/worse metric value with penalized undershoot errors vs. the non-penalized metric Eqn. 4-1.

The non-penalty metric results in the left plot in Figure 4-4 show that performance of all methods improves with added samples. The EON 90% method performs worst at all sample sizes. The 95/90 TI-EN method performs 2nd worst at all sample sizes. The 95/95 TI-EN results are only slightly better. SD results are best by a substantial margin at all sample sizes. Similar trends hold with the 10X penalized metric, but in contrast to the other methods (besides the 'TI Extended' method which is suspect and ignored in the following), SD performance does not continue to improve, and in fact slowly declines, with added samples beyond N=4. Nonetheless, at all sample sizes SD maintains a significant performance margin over even the closest competitor, 95/95 TI- EN.

Another empirical PDF that would presumably provide a challenging distribution shape to test the UQ methods is presented in Figure 4-5.

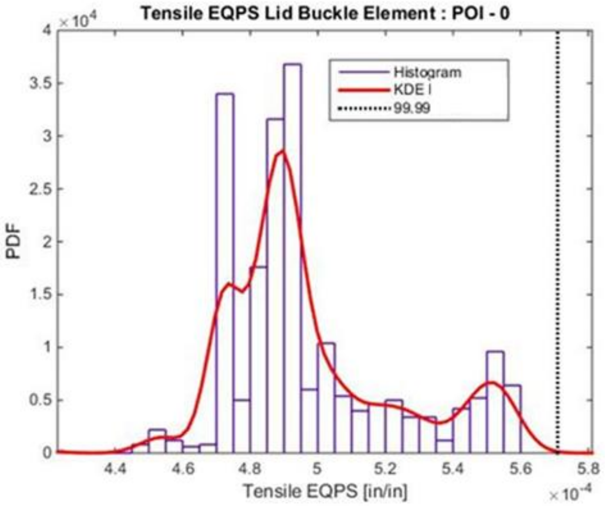


Figure 4-5. Kernel Density fit to a sample empirical histogram of response from can- crush UQ study (this is the QOI at top right corner in Figure 3-10). Vertical dashed line marks 0.9999 quantile of a PDF from the Kernel Density fit normalized to a unit integrated area.

Figure 4-6 shows that SD peak and average errors are noticeably closer to zero than for the other methods, at all sample sizes. But unlike for the first PDF, the SD error distribution is the least compact of all the methods. Furthermore, all methods' error distributions have a substantially reduced trend of shifting leftward and gaining significantly more undershoot errors (more PDF area to the left of zero) as samples are added. Thus, method reliabilities do not appreciably degrade as sample size increases like for the first PDF. In fact, Table 4-2 shows that for all methods, reliabilities with N=20 samples are >95%, close to or higher than with N=2 samples. This might be because this PDF has an extended right tail with substantial prominence (PDF area) relatively close to the EP integration region. This PDF gives a much greater chance than the first one of attaining samples and thus UQ method candidate PDFs with weight/area to the right of the integration limit, and hence exaggerated EP estimates. Consistently, the first PDF's relatively more prominent extended left tail than right tail would appear to work against it in terms of high reliability of attaining conservative EP estimates associated with its right tail which abruptly falls to very small value/weight relatively far from the integration limit.

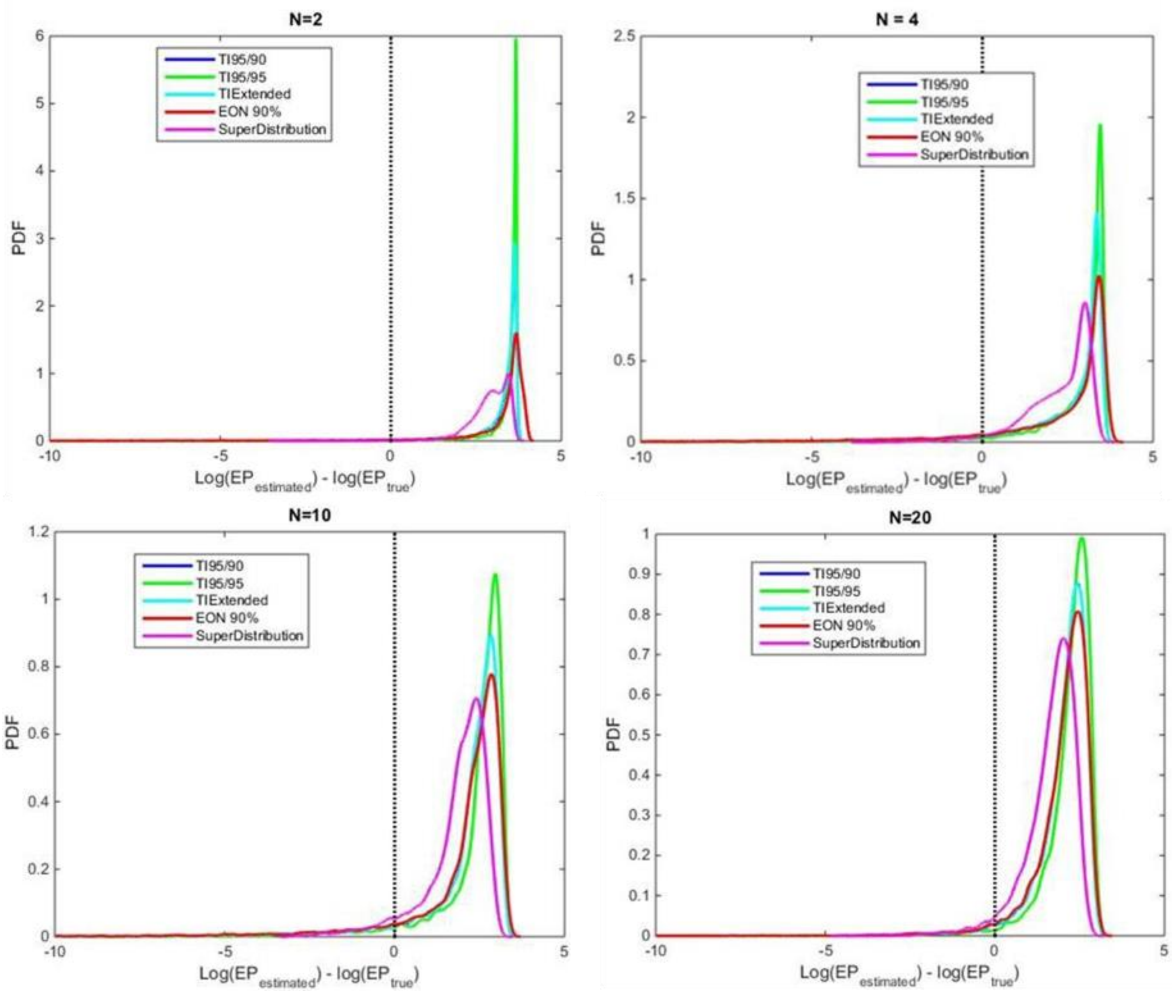


Figure 4-6. Exceedance Probability estimation results in terms of # of orders of magnitude error from exact 10^{-4} for 10K trials with sparse-data UQ methods and N=2, 4, 10, 20 random samples per trial. Note: TI 95/90 results lie below distribution curve labeled 'TI Extended,' which method appears to have been inadvertently assigned the 95/90 results in the plotting procedure.

Table 4-2. Empirical Reliabilities of Sparse-Data UQ Methods for Conservative Exceedance Probability estimation (see Figure 4-5)

N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	90.1%	95.0%	88.8%	98.2%
4	88.0%	93.2%	84.6%	94.6%
10	92.5%	94.7%	91.1%	92.5%
20	97.0%	98.0%	96.6%	95.9%

Figure 4-4 shows generally improving performance metric results (both penalized and unpenalized) for all methods as the number of samples increases. The metric denominator scales with the reliability rate, and since this remains relatively flat over the range N=2 to N=20 for all methods (Table 4-2), the improving performance metric values come from declining overshoot and undershoot error magnitudes. The other finding of note is that for this PDF the SD method appreciably out-performs the other methods at all sample sizes and for both metric variants.

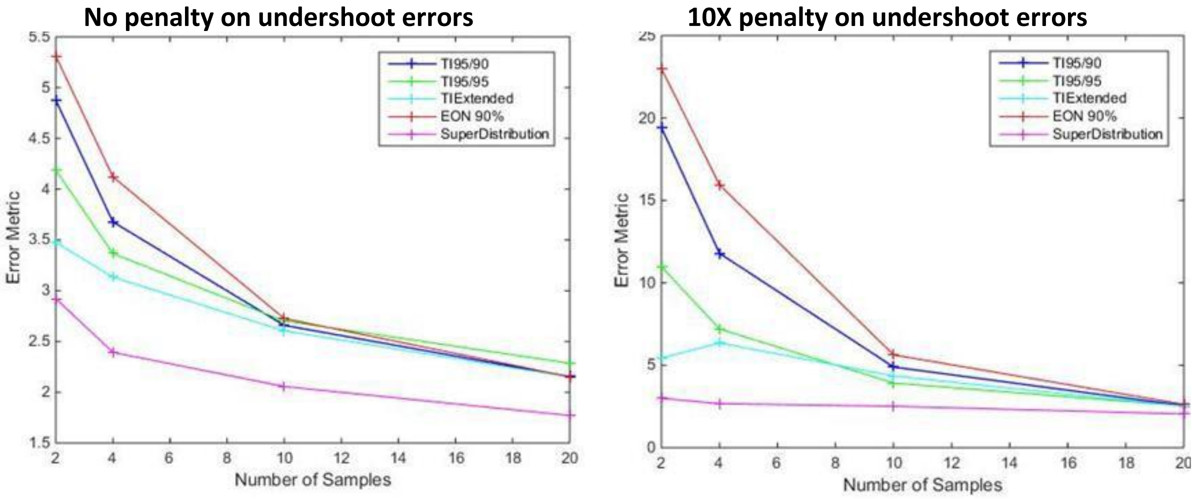


Figure 4-7. EP Performance Metric results vs. number of samples for PDF in Figure 4-5 and unpenalized and 10X penalized undershoot errors. Results for each # of samples are from 10K TI trials with each method. Note: Ignore TI Extended results.

The UQ methods are next tested on an analytic PDF, the standard-normal shown in Figure 4-8.

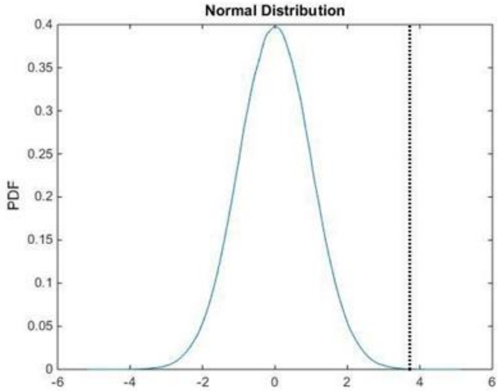


Figure 4-8. Standard-Normal PDF with vertical dashed line at 0.9999 quantile.

Figure 4-9 shows that, for this distribution as well, the SD peak and average errors are discernably closer to zero than for the other methods, at all sample sizes. Table 4-3 reveals that only the SD method reliabilities appreciably change with sample size. The 95/90 TI-EN method has reliabilities slightly greater than 90% at all sample sizes (ranging between 90.9% and 92.4%). The 95/95 TI-EN method has reliabilities close to 95% at all sample sizes (95.4% to 96.2%). The EON 90% method has reliabilities slightly less than 90% at all sample sizes (88.5% to 89.8%). The SD method reliabilities, however, drop significantly with added samples: from 96% for N=2, to 84% for N=20.

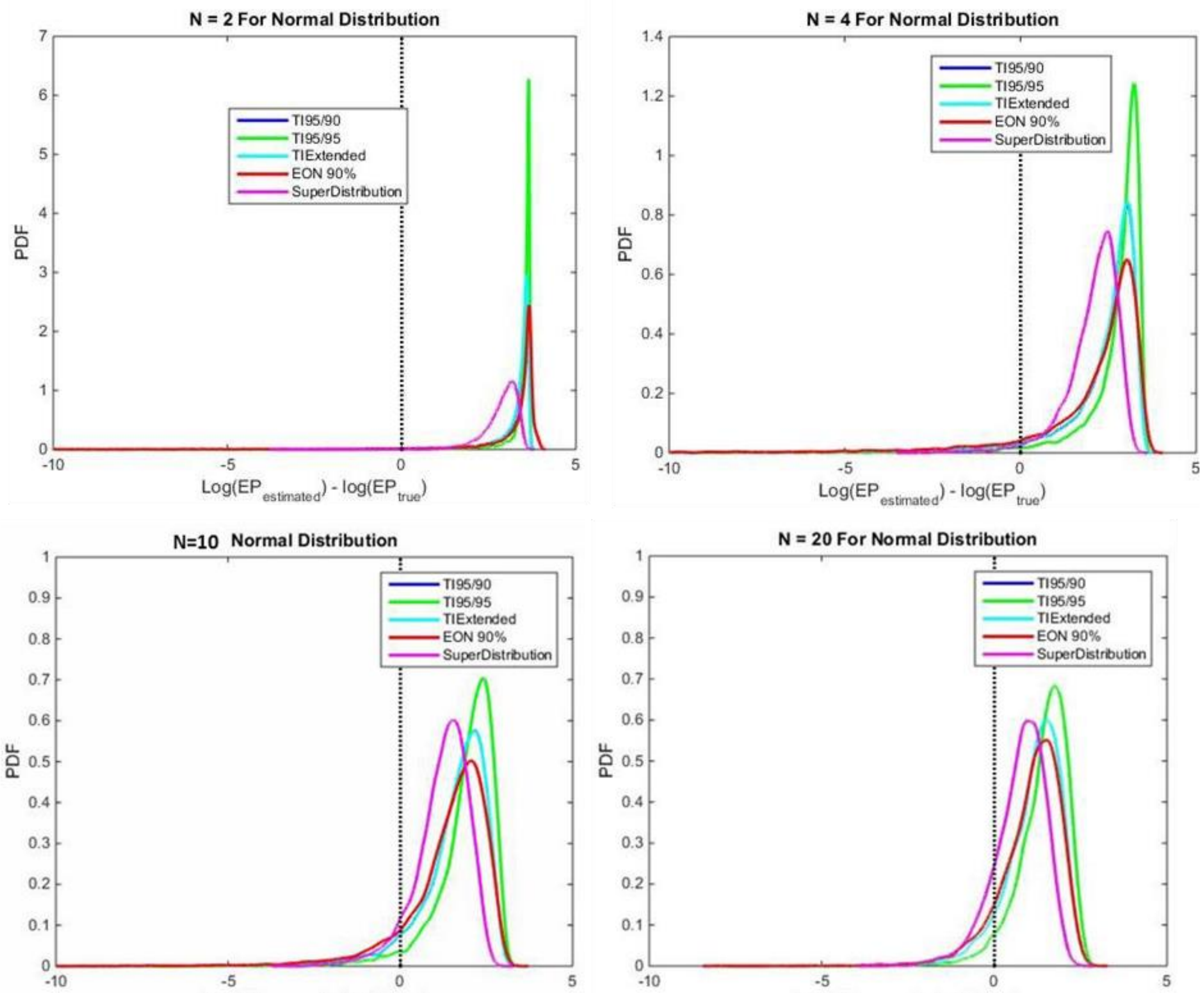


Figure 4-9. Exceedance Probability estimation results in terms of # of orders of magnitude error from exact 10^{-4} for 10K trials with sparse-data UQ methods and N=2, 4, 10, 20 random samples per trial. Note: TI 95/90 results lie below distribution curve labeled 'TI Extended', which method appears to have been inadvertently assigned the 95/90 results in the plotting procedure.

Table 4-3. Empirical Reliabilities of Sparse-Data UQ Methods for Conservative Exceedance Probability estimation (see Figure 4-8)

N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	91.1%	95.6%	89.8%	98.6%
4	91.7%	95.8%	88.7%	96.7%
10	92.4%	96.2%	89.3%	92.2%
20	90.9%	95.4%	88.5%	84.4%

Although SD reliability drops with added samples, its overall performance generally improves if the reductions in magnitudes of overshoot and undershoot errors are accounted for per the performance metrics. Figure 4-10 shows continuous improvement with added samples according to the unpenalized metric, and improvement through at least N=10 samples according to the penalized metric. With either performance metric the SD method again out-performs the other methods at all sample sizes.

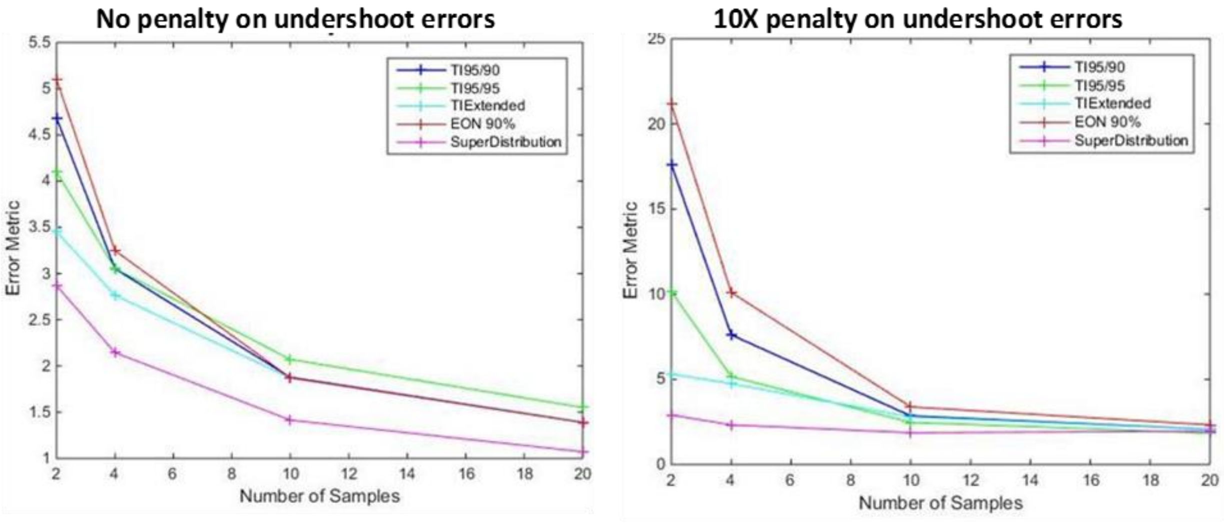


Figure 4-10. EP Performance Metric results vs. number of samples for PDF in Figure 4-8 and unpenalized and 10X penalized undershoot errors. Results for each # of samples are from 10K TI trials with each method. Note: Ignore TI Extended results.

In dramatic contrast, the 5 DOF t-distribution shown in Figure 4-11, which is somewhat shallower than a Normal distribution and has somewhat wider tails, yields UQ method results that are quite different than for a Normal distribution. Figure 4-12 shows considerable leftward shifting and widening of the error distributions of all methods as N increases. The causes all methods' error distributions to have increasingly smaller proportions to the right of zero. SD errors are differentiated by markedly less worsening than for the other methods. Figure 4-12 shows that for 5 DOF t-distribution like for the other three considered above, the SD peak and average errors are discernably closer to zero than for the other methods, at all sample sizes. SD errors are also considerably more compactly distributed than the other methods' errors are.

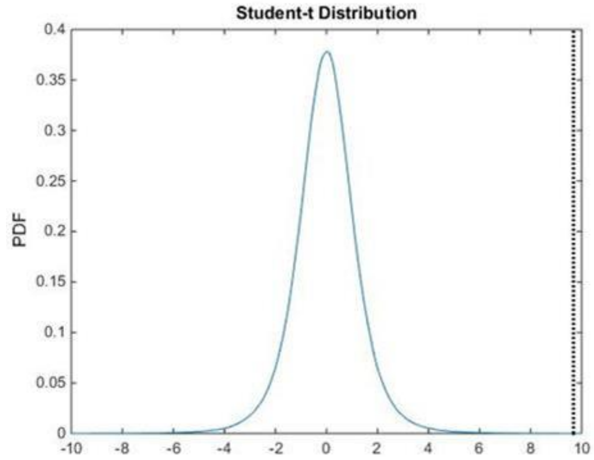


Figure 4-11. 5 Degree-of-Freedom t distribution with vertical dashed line at 0.9999 quantile.

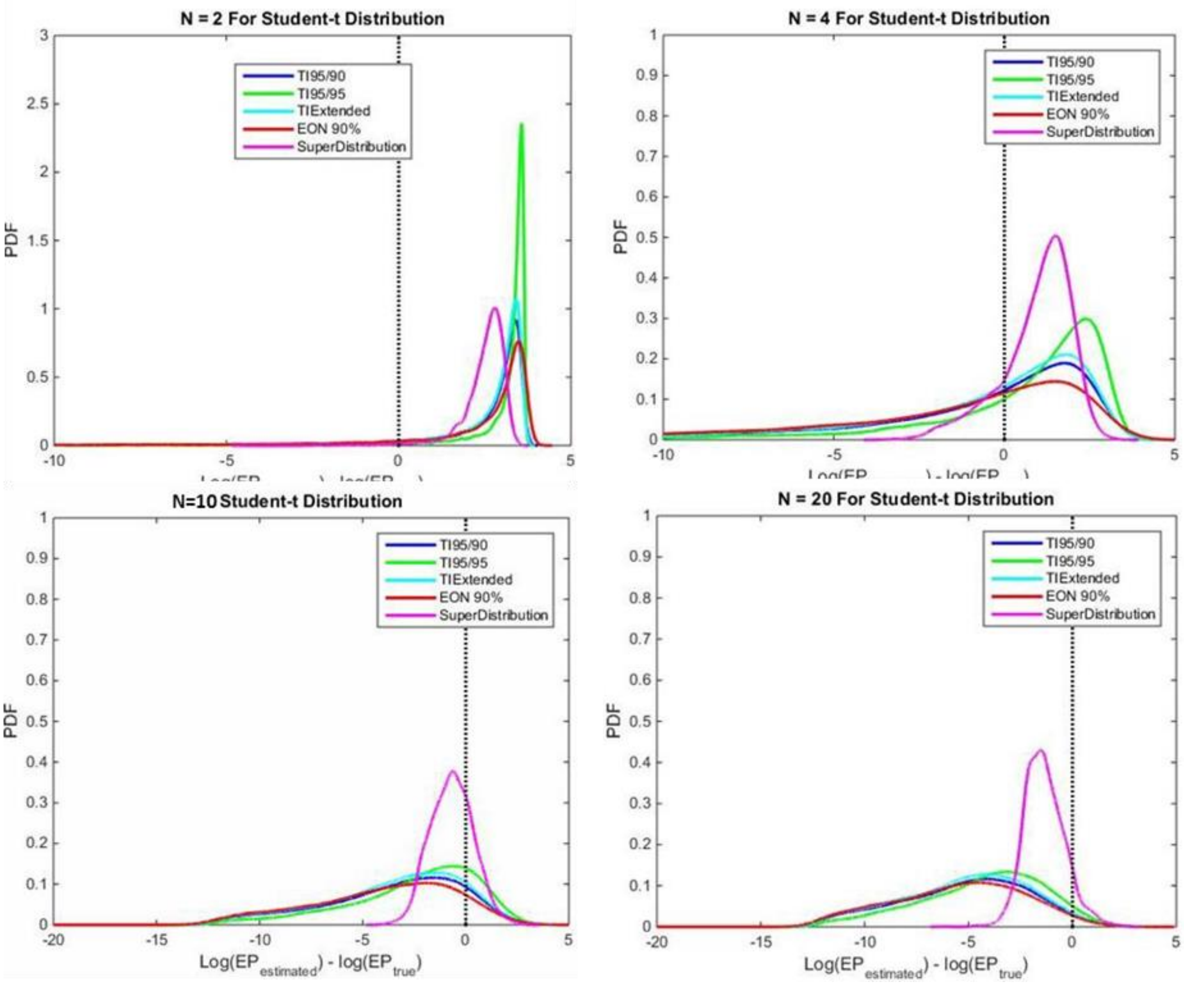


Figure 4-12. Exceedance Probability estimation results in terms of # of orders of magnitude error from exact 10^{-4} for 10K trials with sparse-data UQ methods and N=2, 4, 10, 20 random samples per trial. Note: TI 95/90 results lie below distribution curve labeled 'TI Extended', which method appears to have been inadvertently assigned the 95/90 results in the plotting procedure.

As expected from the declining proportions of errors to the right of zero, all method reliabilities drop precipitously with increasing sample size as Table 4-4 shows. Reliabilities with N=2 samples range from ~78% for EON 90% to ~97% for SD. Reliabilities with N=20 samples fall to very low values ranging from ~2% for SD to ~5% for 95/90 TI-EN. In a departure from the three PDFs investigated above, overall performance *drops* with added samples as shown by increasing performance metric values in Figure 4-13. Associated performance declines are smaller with added samples for SD than for the other methods, and for this PDF like for the others, SD significantly out-performs the other methods at all sample sizes.

Table 4-4. Empirical Reliabilities of Sparse-Data UQ Methods for Conservative Exceedance Probability estimation (see Figure 4-11)

N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	80.9%	90.1%	77.6%	96.6%
4	50.6%	69.0%	41.5%	76.0%
10	11.6%	21.0%	9.1%	14.5%
20	2.7%	4.5%	2.3%	1.9%

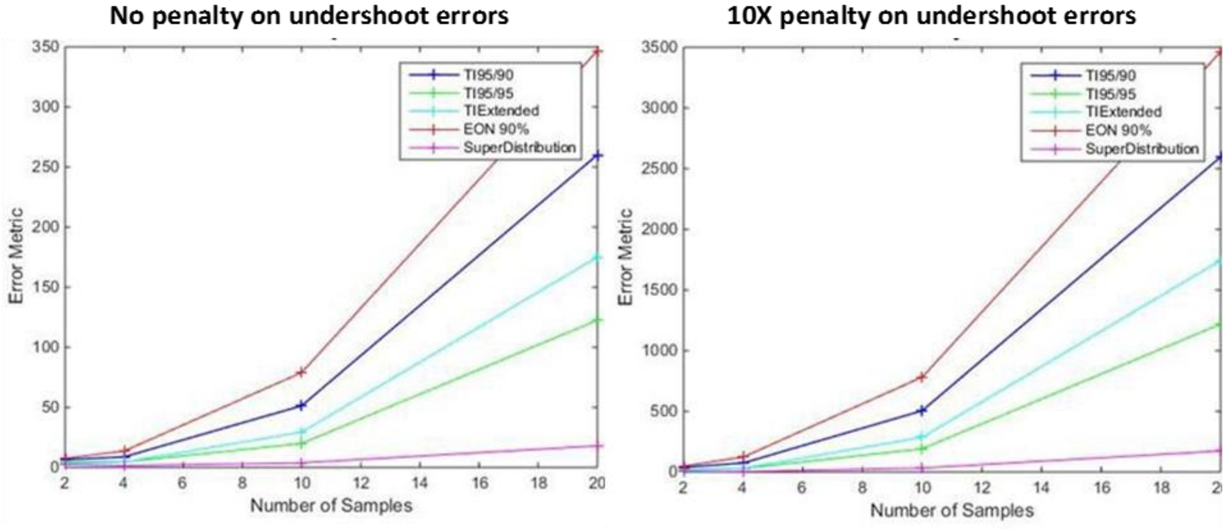


Figure 4-13. EP Performance Metric results vs. number of samples for PDF in Figure 4-11 and unpenalized and 10X penalized undershoot errors. Results for each # of samples are from 10K TI trials with each method. <u>Note</u>: Ignore TI Extended results.

<u>Summary of Other PDF Results and a Proposed Strategy for Selection of Method and # Samples</u>

Appendix C presents the methods' results for log-normal and Weibull analytic distributions and six more empirical PDFs selected from Figure 3-10 thought to be diversely challenging. The PDF for Tensile EQPS Weld Max Global 1.0 in Appendix C behaves roughly like the PDF in Figure 4-2, although looks very different from it. The other analytic and empirical PDFs in Appendix C evoke method performances qualitatively similar to that in Figure 4-12 and Figure 4-13 for the 5 DOF t distribution; performance metric values generally increase and performance generally decrease with added samples. SD is always the best performer for all 12 PDFs, at all sample sizes.

Tables of the number of samples vs. method reliability for conservatively bounding the 10^{-4} EPs are given in Appendix C for the PDFs there. Method reliabilities for the 5 DOF t distribution generally bound (from below) the reliabilities for all the other PDFs above and in Appendix C except for the third PDF in Appendix C (Figure C.3) at N=2,4 and the fourth PDF in Appendix C (Figure C.4) at all samples sizes. Therefore, one might claim reasonably small risk (i.e., *high plausibility* or approximately equivalently, *reasonable credibility*) in using the 5 DOF t distribution's method-specific reliability curves as conservatively biased estimators for reliability vs. sample size relationships in estimating ~10^{-4} magnitude EPs. If one wants higher credibility of attaining conservative results, one could construct for each method a more conservatively biased N vs. reliability relationship by using at each N the lowest reliability value from the tables for the third and fourth PDFs in Appendix C.

A simpler distillation of the information in this section is obtained from focusing on just the SD method, as the best performer for all samples sizes and PDFs studied. The following SD results are the most relevant, assuming a conservative bias of 70% or greater desired reliability.

N=2 samples: all 12 PDFs > 90% relia. (range 93.2 % to 98.6%)

N=3 samples: all 12 PDFs > 70% relia. (range 71.9% to 97.6%) -linear interp. from tables

N=4 samples: 10 of 12 PDFs > 70% relia. (range 72.2% to 96.7%); other two: 51% and 54%

Thus, exceedance probabilities of magnitude ~10^{-4} estimated with SD and N=2 samples could be considered to conjure strong belief or credibility (from 12 of 12 cases tried) that the EP estimate has a high chance (at least 90%) of being conservative, and is most accurately estimated with SD than with the other methods studied. SD with N=3 samples could be considered to conjure strong belief or credibility (from 12 of 12 cases) that the EP estimate is conservative at a medium-high level of reliability (at least 70%). With N=4 samples, a high plausibility or ~equivalently a medium degree of belief or credibility (from 10 of 12 cases) could be considered to exist that the EP estimate is conservative at a medium-high level of reliability (at least 70%).

For EP estimation as well, a striking counter-intuitiveness exists that the number of samples should be kept very low (two or three) for a risk-averse strategy when little to nothing is known about the PDF being sampled.

5. SUMMARY AND CONCLUSIONS

An Ensemble of Normals most fundamentally represents the aleatory and epistemic uncertainties present when statistical inferrences are made from sparse samples drawn from a Normal distribution, although this paper finds substantial applicability to even highly non-normal distributions. Relationships were established between EONs and derivative uncertainty representations depicted in Figure 2-4 and Figure 2-6 (Tolerance Intervals, TI Equivalent Normals, and Superdistributions). The best representation to use depends on the particular UQ purpose. In this paper two UQ purposes were studied: A) bounding the central 95% of PDF response (relevant for e.g. model calibration and validation); and B) bounding a 10^{-4} tail probability of the PDF (representative of risk and reliability analysis purposes).

For bounding a PDF's central 95% of response between its 2.5 and 97.5 percentiles, the suspected relationship depicted in Figure 2-6 between EONs and TIs was found to exist (within a very close approximation) for four symmetric to highly skewed analytic PDFs studied: Normal, five degree-of-freedom t, log-normal, and Weibull. Thus, the 95/90 and 95/95 TI methods are concluded to be simple and economical means to obtain commensurate EON estimates of the central 95% of response. This is anticipated to extend more generally to the central *X*% capture problem with *X*%/90% and *X*%/95% TIs, and to use of these TIs for one-sides bounds on *individual* [100 ± *X*]/2 percentiles of response (aside from two-sided upper and lower bounds on the range between these percentiles).

In general, reliabilities of capturing the central 95% of response are fairly sensitive to the number of samples N. Capture reliabilities decrease as N increases, for all methods. Higher reliability (good) is strongly correlated with larger average over-estimation of the true range of the central 95% of response and associated higher design or risk mitigation costs (bad). Because of these and other conflicting incentives involved, it is important to consider the relationships between capture reliability, # of samples, magnitude of conservatism, and relative desirability of over-estimation vs. under-estimation errors. Multi-attribute scoring metrics were used to combine these factors into overall performance scores as a function of sample size N. The 95/90 and 95/95 TI methods were found to perform better on average than the SD method on the four analytic PDFs and 70 empirical distributions at the tried sample sizes of N=2, 4, 10, and 20. This range of sample sizes and the 74 diverse and challenging distribution shapes (some of which are shown in Figure 3-10) provide a significant test-bed to characterize UQ method performance on.

For the top-performing 95/90 and 95/95 TI methods, overall performance improves greatly in going from N=2 samples to N=3 or 4 (for all four analytic and 70 empirical PDFs), but then only marginally improves or even declines with more samples, as undesirable declines in capture reliability overwhelm beneficial error-magnitude declines. This N=4 performance "sweet spot" between risk-cost-conservatism tradeoffs was also found for 95/90 TIs on 70 other empirical PDFs studied in [Romero, Schroeder, et al., 2017]. Furthermore, findings with a linear test problem in [Winokur & Romero] indicate that nominally 4 experimental realizations and model evaluations per important random variable, function, and/or field source of variability are suitable with 95/90 TIs to attain a reasonable cost-risk-conservatism balance per the considerations in Footnote 2.

About 89% of all the 144 PDFs discussed in this paper have reliabilities of ~75% or greater with 95/90 TIs and N=4 random samples per trial. Related analysis projects that 89% of all 144 PDFs have reliabilities ≥85% with 95/95 TIs and N=4. In the authors' judgment, this large 89%

proportion of the highly diverse and challenging 144 PDFs studied provides a basis for high plausibility ≈ reasonable credibility that reliabilities ≥75% or ≥85% are reasonable expectations for 95/90 TIs or 95/95 TIs respectively and N=4 samples when a distribution of unknown shape is being sampled.

Other choices of N and/or reliability levels and/or credibility levels may be more suitable for the particular cost-risk constraints and objectives in a project. Section 3.3 offers some analysis of the study results to facilitate a strategy for picking the best combination of the # of samples and 95/90 TI or 95/95 TI method to achieve a desired reliability and credibility level (also considering what, if anything, is known about the shape of the PDF being sampled). From the large data base of PDF shapes and results examined, it is proposed that the strategy offers moderate to high credibility options for achieving the desired reliability levels within the specified parameters. Added considerations affect optimal selection of sparse-data uncertainty representations when the uncertainty is to be propagated, as discussed in [Romero, Weirs, Schroeder, et al., 2018].

For the purpose of conservatively bounding a 10^{-4} tail probability of a PDF, the following related sparse-data UQ methods were tested: the Superdistribution method, 95/90 and 95/95 TI-Equivalent Normal methods, and the 90% highest EP estimate from 100 Normals of an ensemble (EON-90%). The methods were tested on the four analytic PDFs and eight empirical distributions chosen for shape diversity and perceived high difficulty for EP estimation. According to performance metrics established in this paper that weigh estimation error magnitudes against reliabilities of bounding the true EP, SD was universally the best performer for all 12 PDFs at all sample sizes.

For all methods, reliabilities decrease as sample size increases, with high sensitivity to N for most of the 12 PDFs. Higher reliability (good) is strongly correlated with larger average over-estimation of the true EP and associated higher design or risk mitigation costs (bad). For most PDFs, reliability declines with sample size faster than estimation error magnitudes, leading to declining overall performance with N for most PDFs. Reliabilities decline with N so quickly that N=4 samples is the maximum that retains reasonable credibility that at least 70% reliability is achieved (when seeking to bound a 10^{-4} magnitude EP for a PDF of unknown shape). It is also concluded that N=3 samples comes with high credibility of attaining >70% reliability, and N=2 samples comes with high credibility of attaining >90% reliability. The basis for these statements follows.

N=2 samples: all 12 PDFs > 90% reliability of attaining a conservative EP estimate (exact=10^{-4})
N=3 samples: all 12 PDFs > 70% reliability
N=4 samples: 10 of 12 PDFs > 70% reliability

For both EP and central 95% estimation there is a striking counter-intuitiveness that the number of samples must be kept very low with these UQ methods to have low risk of under-estimation (non-bounding) of these quantities when little or nothing is known about the PDF being sampled. The large bias toward conservatism that achieves this low risk comes at a cost of increased engineering and product expenses and/or perceptions of smaller design or safety margins. Other UQ approaches surveyed in Section 2 would appear preferable if much is known about the PDF up front and/or the number of samples is more than identified in this report to give reasonably reliable bounding estimates. More research is needed here.

These findings apply to sparse samples of experimental or model simulation scalar results. The latter may come from propagation of sparse realizations of random variable, random function, and/or random field data (see e.g. [Winokur et al., 2017], [Romero, Schroeder, et al., 2017]).

REFERENCES

[1] Bhachu, K.S., R.T. Haftka, N.H. Kim, "Comparison of Methods for Calculating B-Basis Crack Growth Life Using Limited Tests," *AIAA Journal*, Vol. 54, No. 4, April 2016.

[2] Devore, J.L., *Probability & Statistics for Engineering and the Sciences*, Brooks/Cole Publishing Co., Wadsworth, Inc., Belmont, CA., 1982, pp. 99 - 104.

[3] Hahn, G.J., and Meeker, W.Q., *Statistical Intervals—A Guide for Practitioners*, Wiley & Sons, 1991.

[4] Howe, W. G. (1969). "Two-sided Tolerance Limits for Normal Populations - Some Improvements", *J. American Statistical Association*, 64 , pages 610-620.

[5] Miller, I., and Freund, J.E., *Probability and Statistics for Engineers*, 3rd Ed., Prentice-Hall, 1985.

[6] MIL-HDBK-17-1F Composites Materials Handbook," Department of Defense, Vol. 1, Chapt. 8, 2002.

[7] Montgomery, D.C., and Runger, G.C., *Applied Statistics and Probability for Engineers*, Wiley & Sons, 1994.

[8] Pradlwarter, H.J., and G.I. Schuëller, "The use of kernel densities and confidence intervals to cope with insufficient data in validation experiments," *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, Issues 29-32, May 2008, pp. 2550-2560.

[9] Romero, V., B. Rutherford, J. Newcomer, "Some Statistical Procedures to Refine Estimates of Uncertainty when Sparse Data are Available for Model Validation and Calibration," paper AIAA- 2011-1709, 13th AIAA Non-Deterministic Approaches Conference, Denver, CO, April 4-7, 2011.

[10] Romero, V., J. Mullins, L. Swiler, A. Urbina, "A Comparison of Methods for Representing and Aggregating Experimental Uncertainties involving Sparse Data—More Results," *Soc. Automot. Engrs. Int. J. of Materials and Manufacturing*, 6(3):2013, doi:10.4271/2013-01-0946.

[11] Romero, V., L. Swiler, A. Urbina, J. Mullins, "A Comparison of Methods for Representing Sparsely Sampled Random Quantities," Sandia National Laboratories report SAND2013-4561, September 2013.

[12] Romero, V., J.F. Dempsey, B. Antoun, "UQ and V&V Techniques applied to Experiments and Simulations of Heated Pipes Pressurized to Failure," Sandia National Laboratories report SAND2014-3985, May 2014.

[13] Romero, V., A. Black, N. Breivik, G. Orient, J. Suo-Anttila, B. Antoun, A. Dodd, "Advanced UQ and V&V Procedures applied to Thermal-Mechanical Response and Weld Failure in Heated Pressurizing Canisters," Sandia National Laboratories document SAND2015-3005C presented at Soc. Auto. Engrs. 2015 World Congress, April 21-23, 2015, Detroit, MI.

[14] Romero, V., B. Schroeder, J.F. Dempsey, N. Breivik, G. Orient, B. Antoun, J.R. Lewis, J. Winokur, "Simple Effective Conservative Treatment of Uncertainty from Sparse Samples of Random Variables and Functions," Sandia National Laboratories document SAND2017-5177 J accepted for *ASCE-ASME Journal of Uncertainty and Risk in Engineering Systems: Part B. Mechanical Engineering*.

[15] Romero, V.J, and V.G. Weirs, "A Class of Simple and Effective UQ Methods for Sparse Replicate Data applied to the Cantilever Beam End-to-End UQ Problem," Sandia National Laboratories document SAND2017-12365 C, 20th AIAA Non- Deterministic Approaches Conference, AIAA SciTech 2018, Jan. 8-12, Kissimmee, FL.

[16] Romero, V., “Demonstration and Analysis of Discrete-Direct Model Calibration and Uncertainty Propagation involving Aleatory and Epistemic Uncertainties in the Experiments and Models,” Sandia National Laboratories document in preparation (2018).

[17] Romero, V., V.G. Weirs, B. Schroeder, J.R. Lewis, L. Hund, J. Mullins, “Approaches to Experimental Data UQ and QMU for Scalar Data from Stochastically Varying Systems,” draft Sandia National Laboratories document, 2018.

[18] Winokur, J., and V. Romero, “Optimal Design of Computer Experiments for Uncertainty Quantification with Sparse Discrete Sampling,” Sandia National Laboratories document SAND2016-12608 J in revision for *ASCE-ASME Journal of Uncertainty and Risk in Engineering Systems: Part A. Civil Engineering*.

[19] Young, D.S. (2010). ``Tolerance: An R Package for Estimating Tolerance Intervals.” *Journal of Statistical Software*, 36(50), pp. 1-39.

APPENDIX A. DEFINITION OF LOG-NORMAL AND WEIBULL DISTRIBUTIONS USED IN THIS STUDY

The Matlab function calls for the four analytic distribution used in this study are:

```
random('norm',0,1,[N,1]);
random('t',5,[N,1]);
random('wbl',1,1.3,[N,1]);
random('logn',10.48,0.314,[N,1]);
```

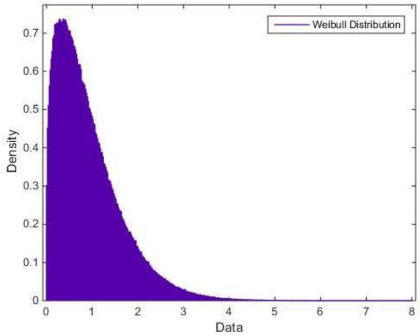
where N is the number of samples.

The normal distribution has a mean of zero and unit variance (a Standard Normal). The Student-t distribution has 5 degrees of freedom with a zero mean.

The Weibull distribution has 2 inputs: scale parameter (a) and shape parameter (b). The scale parameter is set to a=1 and the shape parameter is set to b=1.3. The pdf equation is:

$$f(x|a,b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-\left(\frac{x}{a}\right)^{b}}.$$

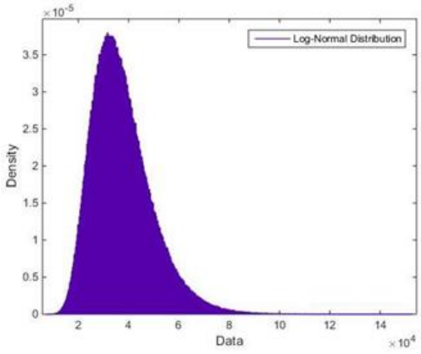
This pdf with the specified input parameters is plotted below.



The log-normal distribution also contains two inputs: normal mean (μ) and normal standard deviation (σ). These do not correspond to the resulting distribution, but the underlying normal distribution the logarithm is taken of. The pdf equation for this log-normal distribution is:

$$f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\{\frac{-(\mathrm{L}-\mathrm{N}(x)-\mu)^{2}}{2\sigma^{2}}\}.$$

This pdf with the specified parameter values is plotted below.



APPENDIX B. MULTI-ATTRIBUTE PERFORMANCE METRIC SCORES FOR UQ METHOD CAPTURE OF CENTRAL 95% OF PDFS

The performance metric values in this appendix use Equation 3-4 for the numerator value in the performance metric Equation 3-5 for tables with titles indicating ‘Non-Penalized’ performance metric results. Tables with titles indicating ‘10X Penalized’ performance metric results use Equation 3-6 for the numerator to apply a 10X penalty to shortfall errors (see Figure 3-2 and Table 3-2).

Normal PDF

Table B-1 and Table B-2 contain the numerical data plotted in Figure 3-3 for a sampled Normal PDF. As expected, in each table the results for 95/90 TIs and EONs are very close to each other, as are the results for 95/95 TIs and EONs. A few cases in the tables have EON performance values that are slightly lower/better than the corresponding TI values, but the majority of cases show TI performs better. This is reflected in each table’s last column which shows better average performance of the TI methods than the corresponding EON methods. Other observations on the performance of these and the SD and EON50% methods are given in the body text discussing Figure 3-3.

Table B-1. Non-Penalized Performance Metric Values, Normal distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	29.03	6.32	2.46	1.52	9.8
EON 90%	32.1	6.14	2.54	1.58	10.6
95/95 TI	58.56	8.68	3.03	1.75	18.0
EON 95%	72.29	8.48	3.07	1.84	21.4
EON 50%	49.12	3.95	2.51	1.82	14.4
Super D.	40.14	4.52	2.09	1.54	12.1

Table B-2. 10X Penalized Performance Metric Values, Normal distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	30.88	7.4	3.14	2.06	10.9
EON 90%	33.97	7.42	3.23	2.07	11.7
95/95 TI	59.45	9.18	3.32	2	18.5
EON 95%	73.17	9.09	3.38	2.07	21.9
EON 50%	89.21	12.25	10.57	8.74	30.2
Super D.	41.86	7.66	5.89	5.52	15.2

5 DOF t-distribution

Table B-3 and Table B-4 contain the numerical data plotted in Figure 3-5 for a sampled 5 DOF t-distribution. As expected, in each table the results for 95/90 TIs and EONs are very close to each other, as are the results for 95/95 TIs and EONs. In Table B.3 for the unpenalized metric shows 95/95 EON performs slightly better than the corresponding TIs for N=4, but the other cases show 95/95 TIs performs better such that its average is better in the table's last column. EON 95/90 results show a rare outperformance of the corresponding TIs on average, although the TIs outperform EON at the larger samples sizes N=10,20. In Table B.4 for the 10X penalized metric, 95/90 and 95/95 TIs do better on average (at N=2,4,10 but not at N=20) than their EON counterparts. Other observations on the performance of these and the SD method are given in the body text discussing Figure 3-5.

Table B-3. Non-Penalized Performance Metric Values, 5 DOF t distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	53.61	7.89	3.48	2.34	16.8
EON 90%	39.32	7.74	3.51	2.37	13.2
95/95 TI	71.43	10.59	3.94	2.47	22.1
EON 95%	88.37	10.42	4.00	2.56	26.3
Super D.	49.54	6.19	3.70	3.02	15.6

Table B-4. 10X Penalized Performance Metric Values, 5 DOF t distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	38.29	9.98	5.48	4.29	14.5
EON 90%	42.04	10.24	5.62	4.16	15.5
95/95 TI	72.64	11.54	4.87	3.46	23.1
EON 95%	89.63	11.55	5.00	3.44	27.4
Super D.	52.03	12.56	14.12	13.90	23.2

Log-Normal Distribution

Table B-5 and Table B-6 contain the numerical data plotted in Figure 3-7for a sampled log-normal distribution. As expected, in each table the results for 95/90 TIs and EONs are very close to each other, as are the results for 95/95 TIs and EONs. In Table B-5 for the unpenalized metric shows EONs performs slightly better than the corresponding TIs in a few cases, but the majority of cases and the averages over all cases N=2,4,10,20 show in the table's last column that TIs perform better on average than their corresponding EONs. In Table B-6 for the 10X penalized metric, 95/90 and 95/95 TIs do better on average (at N=2,4,10 but not at N=20) than their EON counterparts. Other observations on the performance of these and the SD method are given in the body text discussing Figure 3-7.

Table B-5. Non-Penalized Performance Metric Values, Log-Normal distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	347530	81513	41167	33436	125912
EON 90%	382180	79871	41399	33427	134219
95/95 TI	691010	106930	44631	33215	218947
EON 95%	843560	105040	45594	33580	256944
Super D.	475610	65192	44141	45893	157709

Table B-6. 10X Penalized Performance Metric Values, Log-Normal distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	374820	105190	69956	71225	155298
EON 90%	409300	107030	71290	68565	164046
95/95 TI	704220	118410	60387	55659	234669
EON 95%	856570	118420	62122	53755	272717
Super D.	501000	129600	155110	194100	244953

Weibull Distribution

Table B-7 and Table B-8 contain the numerical data plotted in Figure 3-9 for a sampled Weibull distribution. As expected, in each table the results for 95/90 TIs and EONs are very close to each other, as are the results for 95/95 TIs and EONs. Table B-7 for the unpenalized metric shows EONs performs slightly better than the corresponding TIs in a few cases, but the majority of cases and the averages over all cases N=2,4,10,20 show in the table's last column that TIs perform better on average than their corresponding EONs. In Table B-6 for the 10X penalized metric, 95/90 and 95/95 TIs do better on average (at N=2,4,10 but not at N=20) than their EON counterparts. Other observations on the performance of these and the SD method are given in the body text discussing Figure 3-9.

Table B-7. Non-Penalized Performance Metric Values, Weibull distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	21.10	5.45	3.21	2.94	8.18
EON 90%	23.13	5.37	3.23	2.89	8.65
95/95 TI	40.74	6.83	3.28	2.72	13.39
EON 95%	50.61	6.77	3.31	2.72	15.85
Super D.	28.91	4.57	3.69	4.53	10.43

Table B-8. 10X Penalized Performance Metric Values, Weibull distribution sampled N times, results from 10K random trials of each method

Method	N=2	N=4	N=10	N=20	avg.score
95/90 TI	23.3	7.7	6.4	7.2	11.2
EON 90%	25.3	7.9	6.5	6.8	11.6
95/95 TI	41.7	8.0	5.2	5.4	15.1
EON 95%	51.6	8.1	5.3	5.1	17.5
Super D.	30.9	10.0	13.5	19.0	18.3

Average of 70 Empirical Distributions

Table B-9 and Table B-10 contain the numerical data plotted in Figure 3-15 for average results of 70 empirical distributions as explained in section 3.2. Observations on these performance data are given at the beginning of section 3.3.

Table B-9. Non-Penalized Performance Metric Values, average results for 70 empirical distributions sampled N times, results from 10K random trials of each method for each distribution

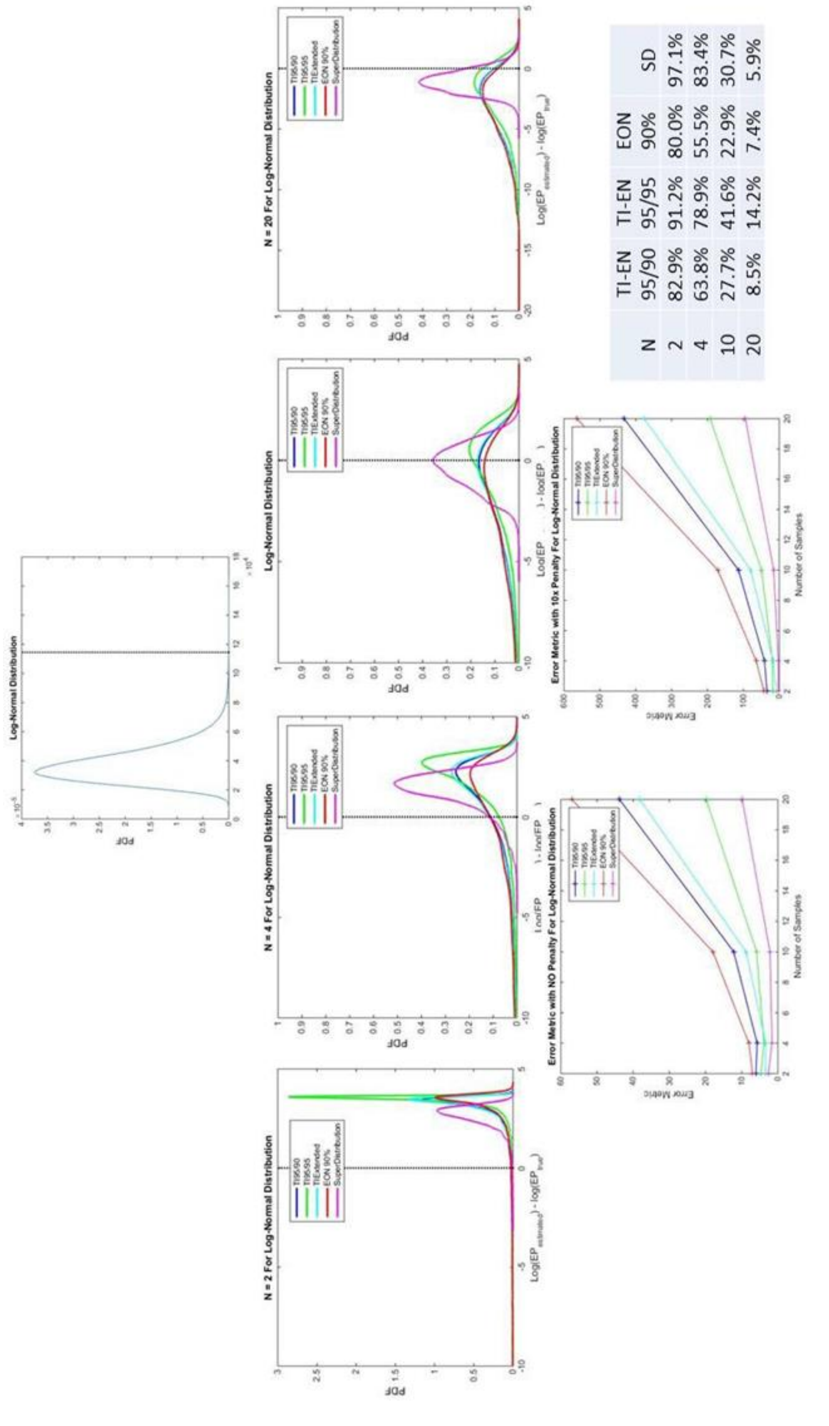
Method	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10	avg. score
95/90 TI	1.96	0.83	0.62	0.53	0.48	0.44	0.43	0.41	0.41	0.68
95/95 TI	3.57	1.09	0.71	0.57	0.50	0.45	0.42	0.39	0.38	0.90
Super D.	2.62	0.76	0.62	0.58	0.59	0.60	0.63	0.65	0.70	0.86
EON50%	1.13	1.01	0.96	0.97	0.97	1.02	1.07	1.05	1.06	1.03

Table B-10. 10X Penalized Performance Metric Values, average results for 70 empirical distributions sampled N times, results from 10K random trials of each method for each distribution

Method	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10	avg.score
95/90 TI	2.59	1.86	1.87	1.91	1.95	1.94	2.06	2.08	2.19	2.05
95/95 TI	3.90	1.65	1.48	1.46	1.51	1.50	1.54	1.52	1.62	1.80
Super D.	3.20	2.40	3.01	3.34	3.75	4.06	4.52	4.85	5.38	3.83
EON50%	7.11	7.30	7.38	7.73	7.92	8.46	8.96	8.86	9.07	8.09

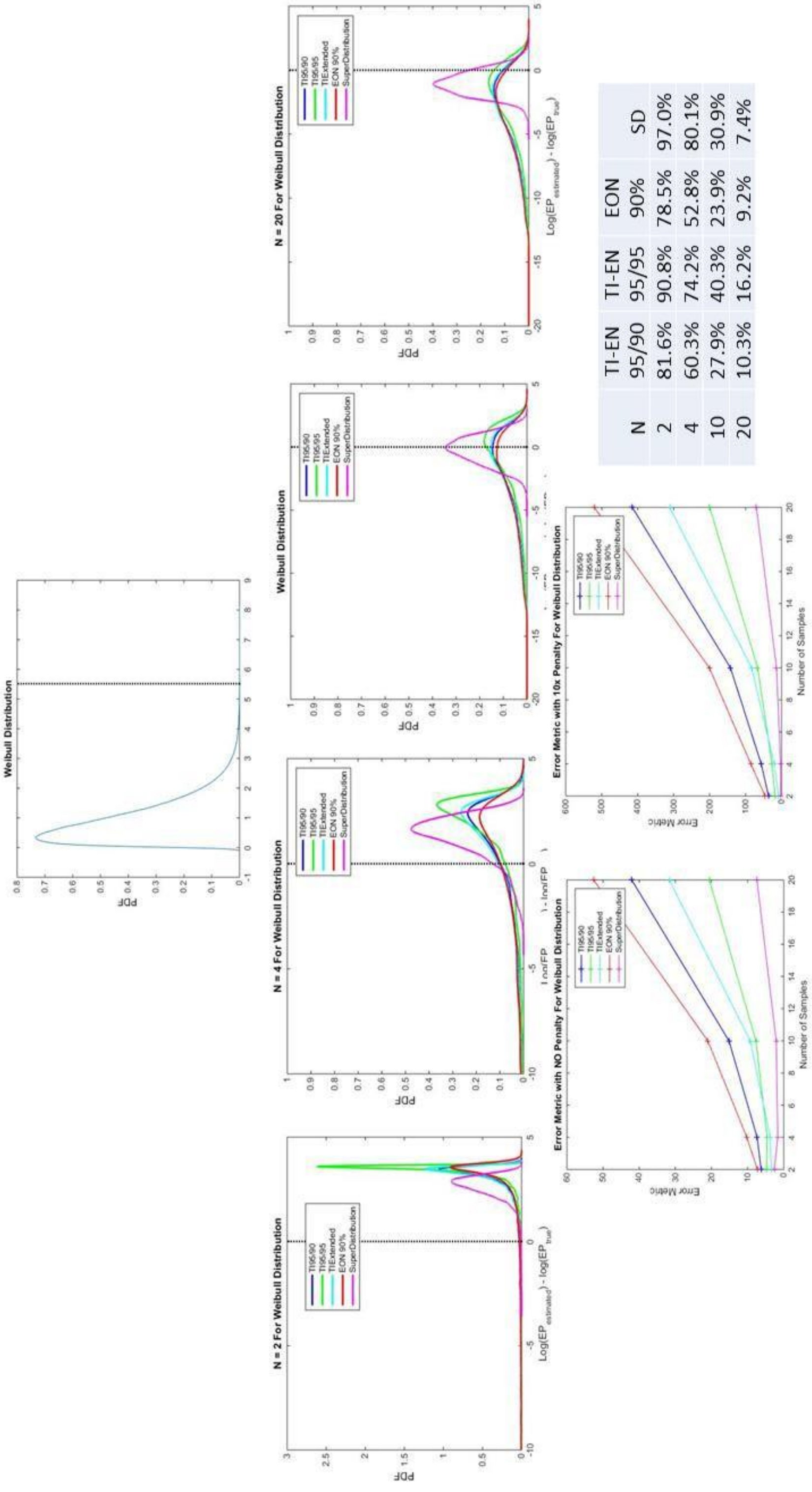
APPENDIX C. EXCEEDANCE PROBABILITY ESTIMATION RESULTS FOR TWO ANALYTIC AND SIX EMPIRICAL PDFS

The performance metric values in this appendix use Equation 4-1 through 4-3.



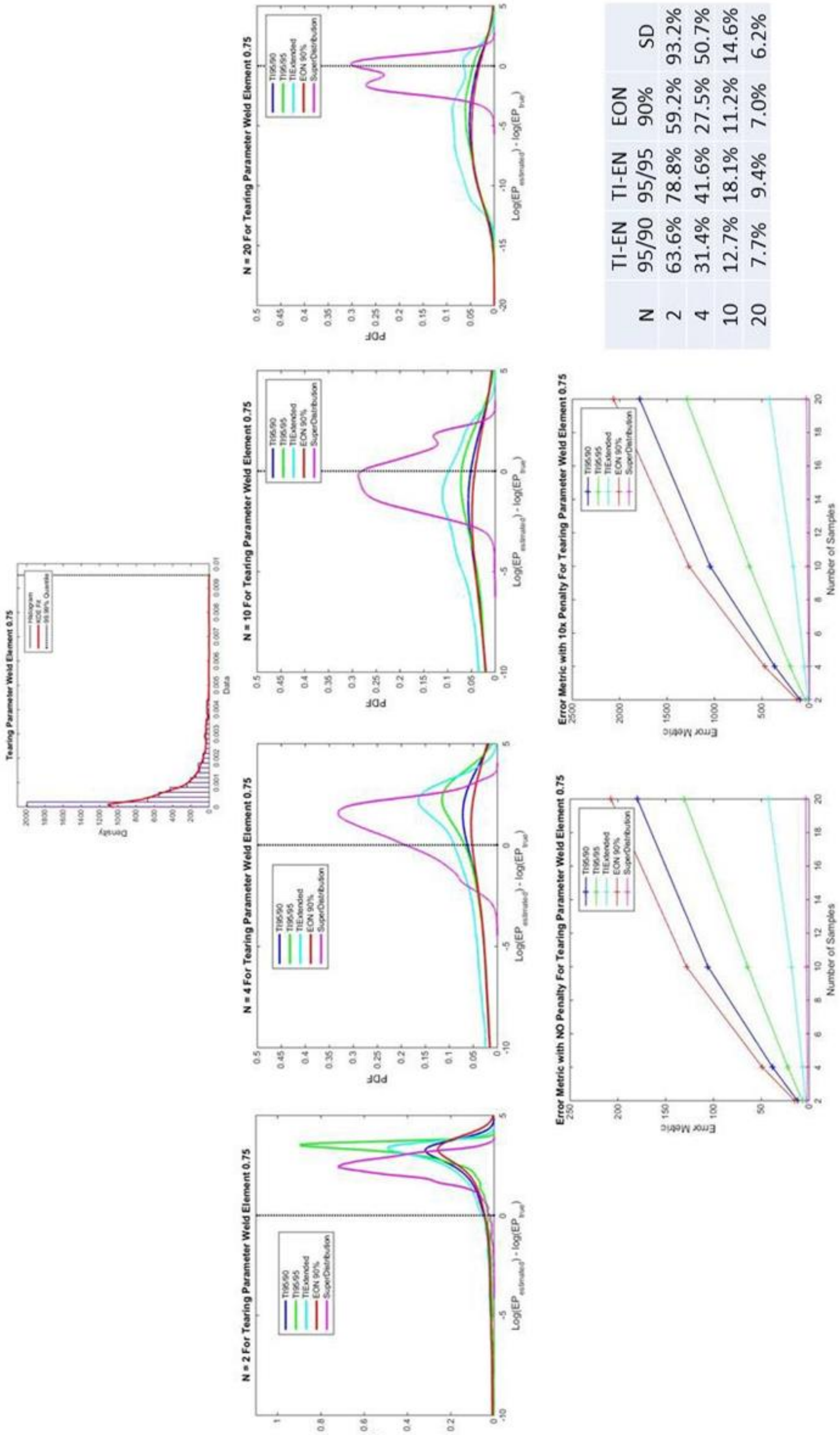
N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	82.9%	91.2%	80.0%	97.1%
4	63.8%	78.9%	55.5%	83.4%
10	27.7%	41.6%	22.9%	30.7%
20	8.5%	14.2%	7.4%	5.9%

Figure C-1. 10^{-4} EP estimation results for Log-Normal PDF defined in Appendix A



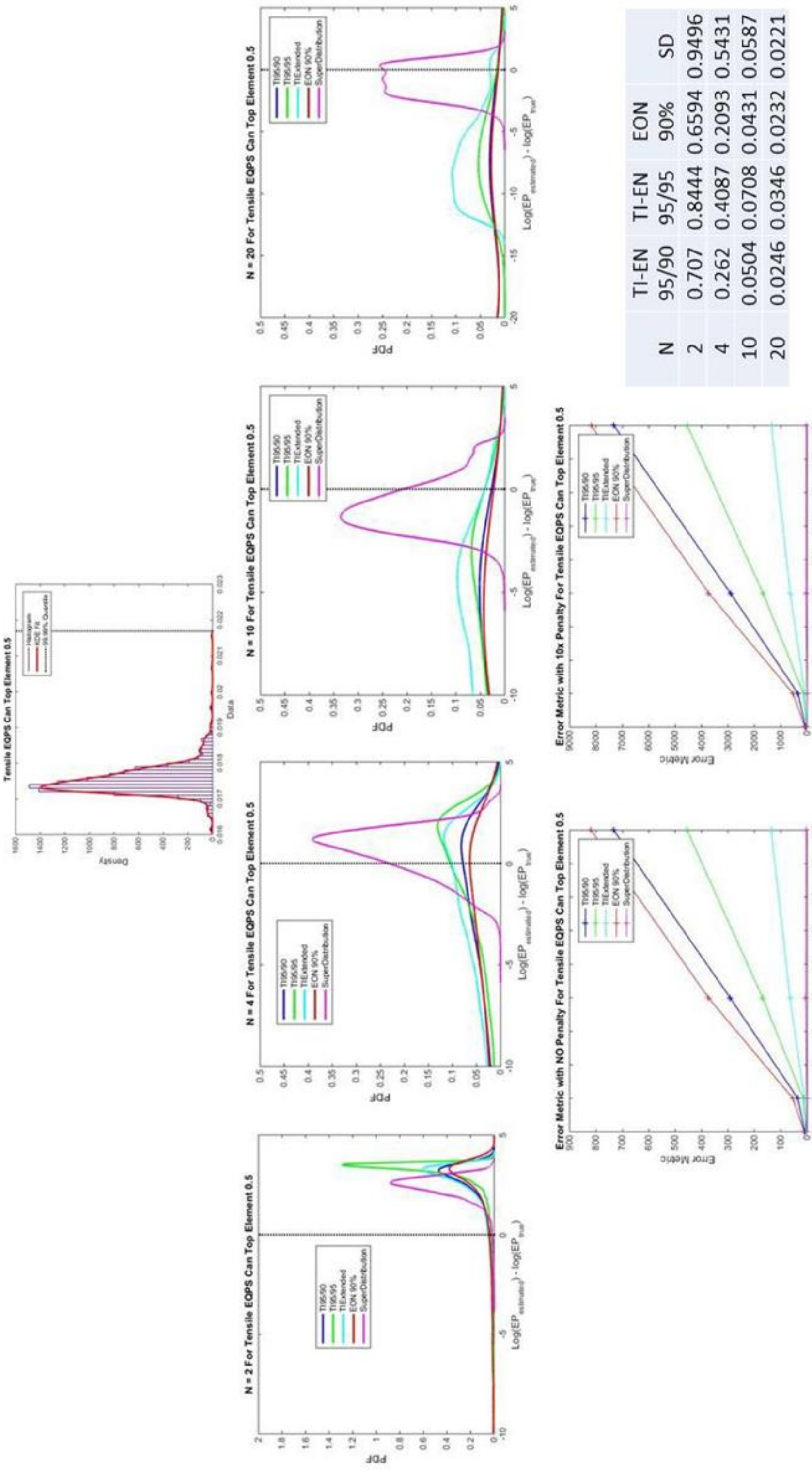
N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	81.6%	90.8%	78.5%	97.0%
4	60.3%	74.2%	52.8%	80.1%
10	27.9%	40.3%	23.9%	30.9%
20	10.3%	16.2%	9.2%	7.4%

Figure C-2. 10^{-4} EP estimation results for Weibull PDF defined in Appendix A



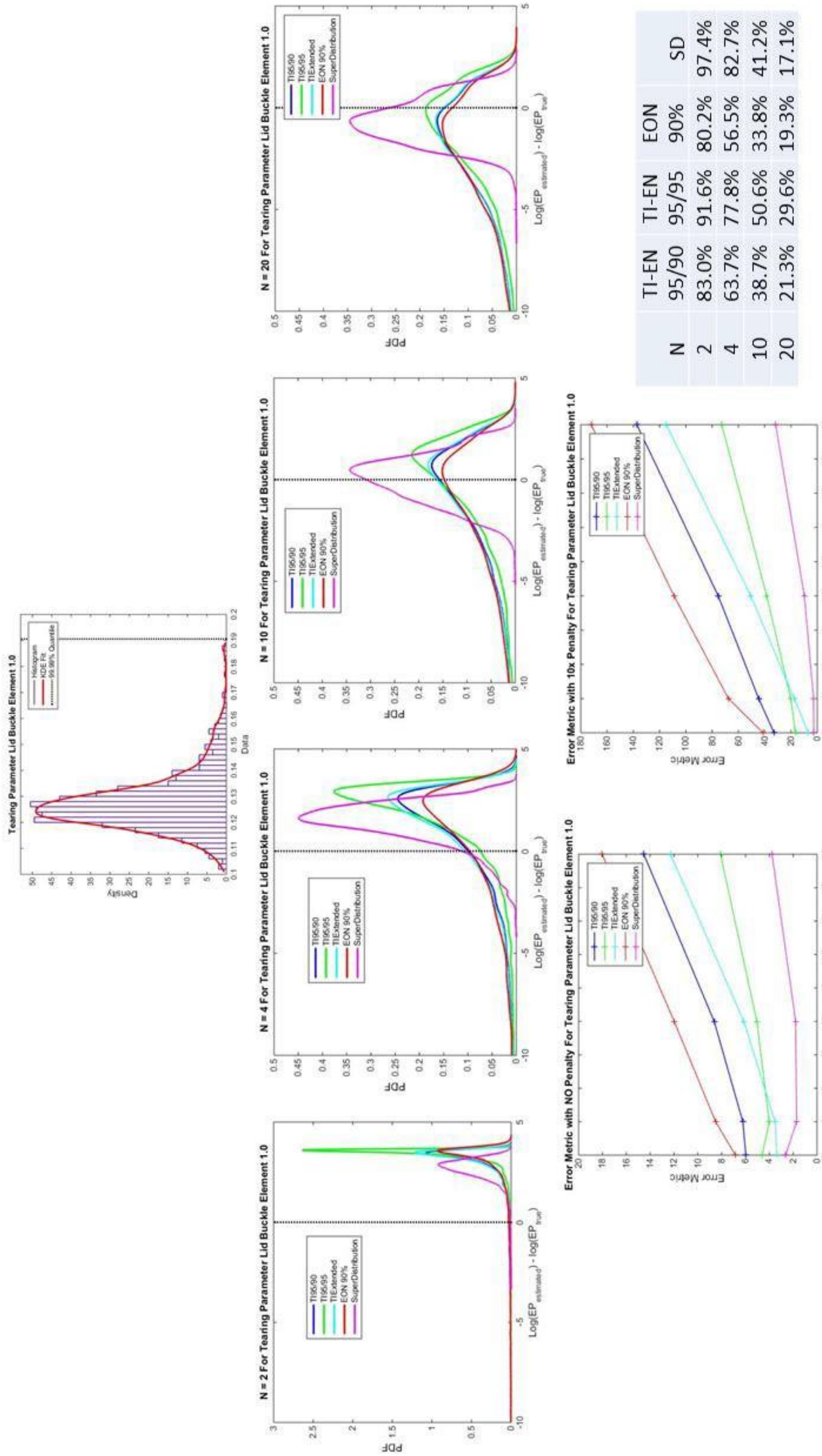
N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	63.6%	78.8%	59.2%	93.2%
4	31.4%	41.6%	27.5%	50.7%
10	12.7%	18.1%	11.2%	14.6%
20	7.7%	9.4%	7.0%	6.2%

Figure C-3. 10^{-4} EP estimation results for Can-Crush empirical PDF for Tearing Parameter Weld Element 0.75



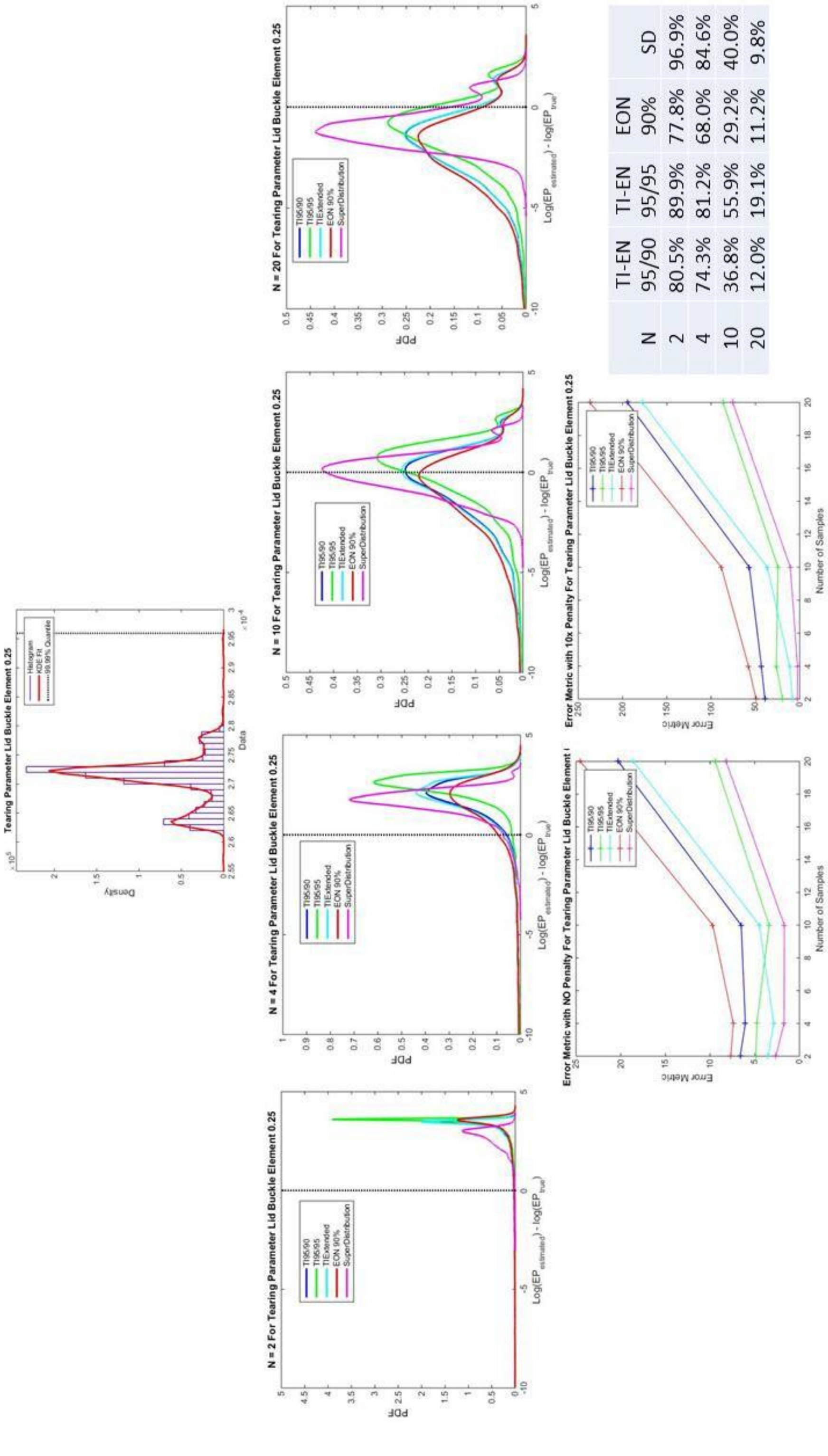
N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	0.707	0.8444	0.6594	0.9496
4	0.262	0.4087	0.2093	0.5431
10	0.0504	0.0708	0.0431	0.0587
20	0.0246	0.0346	0.0232	0.0221

Figure C-4. 10^{-4} EP estimation results for Can-Crush empirical PDF for Tensile EQPS Can Top Element 0.5



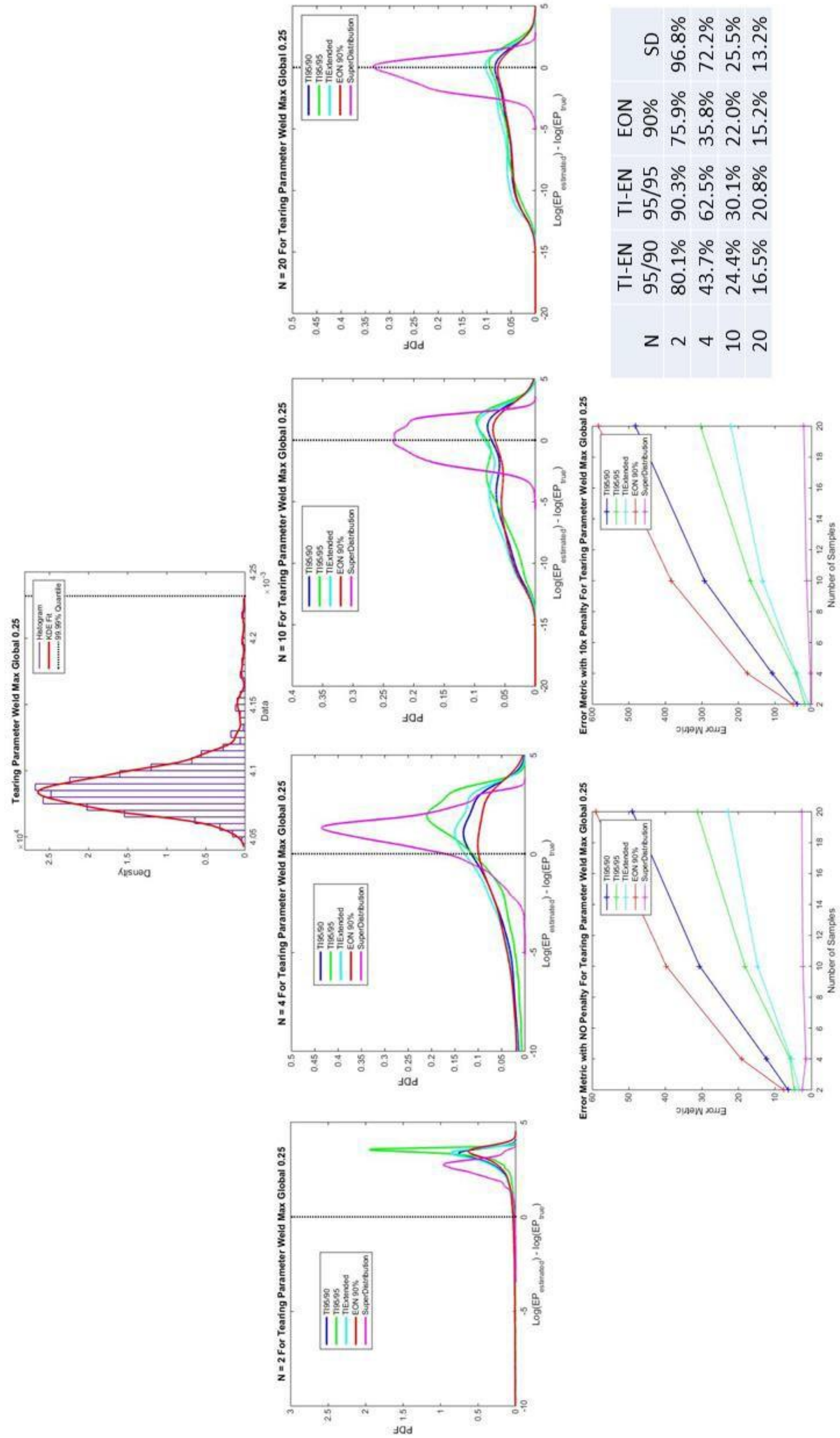
N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	83.0%	91.6%	80.2%	97.4%
4	63.7%	77.8%	56.5%	82.7%
10	38.7%	50.6%	33.8%	41.2%
20	21.3%	29.6%	19.3%	17.1%

Figure C-5. 10^{-4} EP estimation results for Can-Crush empirical PDF for Tearing Parameter Lid Buckle Element 1.0



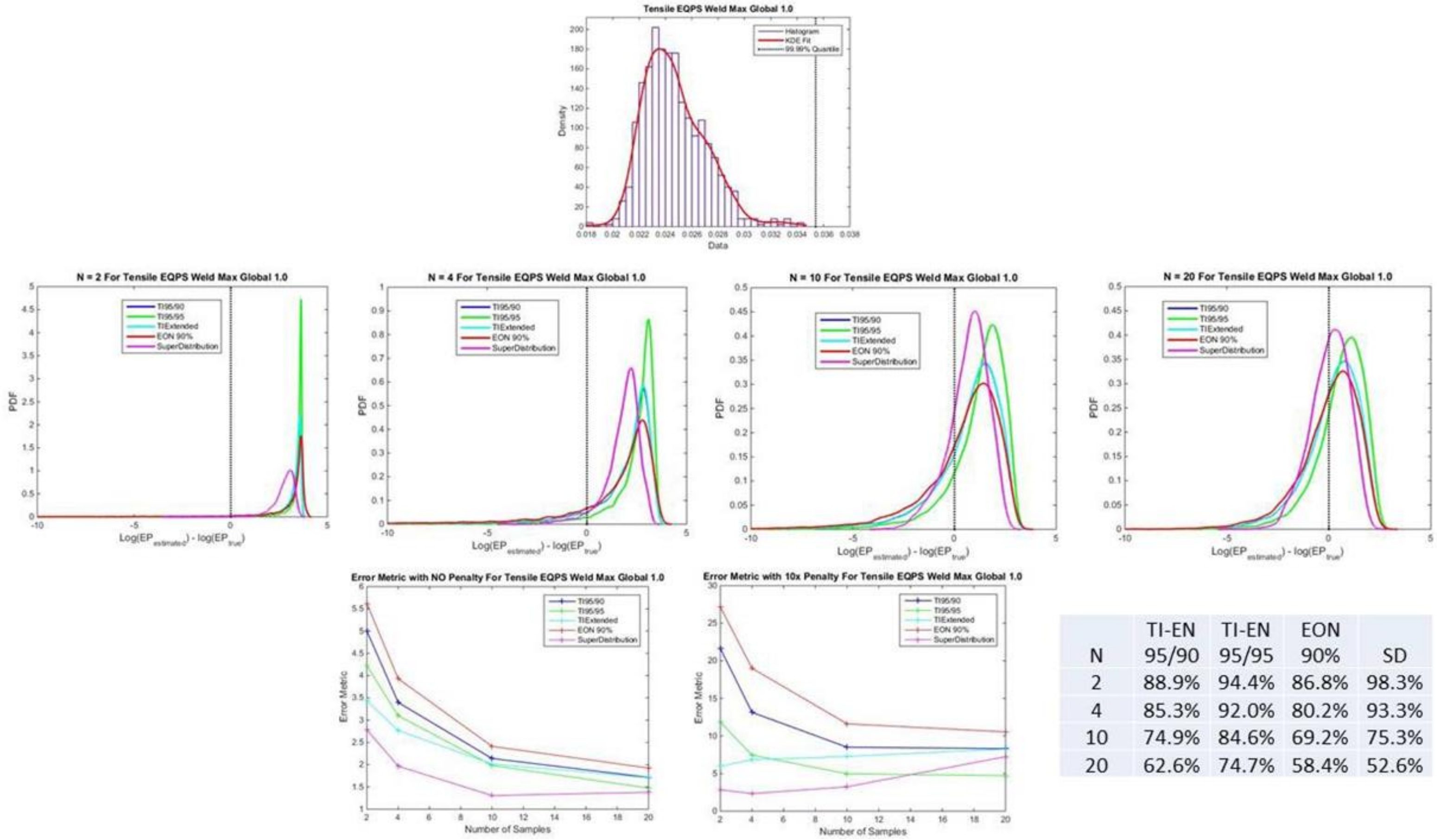
N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	80.5%	89.9%	77.8%	96.9%
4	74.3%	81.2%	68.0%	84.6%
10	36.8%	55.9%	29.2%	40.0%
20	12.0%	19.1%	11.2%	9.8%

Figure C-6. 10^{-4} EP estimation results for Can-Crush empirical PDF for Tearing Parameter Lid Buckle Element 0.25



N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	80.1%	90.3%	75.9%	96.8%
4	43.7%	62.5%	35.8%	72.2%
10	24.4%	30.1%	22.0%	25.5%
20	16.5%	20.8%	15.2%	13.2%

Figure C-7. 10^{-4} EP estimation results for Can-Crush empirical PDF for Tearing Parameter Weld Max Global 0.25



N	TI-EN 95/90	TI-EN 95/95	EON 90%	SD
2	88.9%	94.4%	86.8%	98.3%
4	85.3%	92.0%	80.2%	93.3%
10	74.9%	84.6%	69.2%	75.3%
20	62.6%	74.7%	58.4%	52.6%

Figure C-8. 10^{-4} EP estimation results for Can-Crush empirical PDF for Tensile EQPS Weld Max Global 1.0

DISTRIBUTION

External Distribution

1 Matthew Bonney (electronic copy)
7685 NW16th St
Ankeny, IA 50023

Internal Sandia Distribution

1	MS0348	5352	Ethan Blansett	(electronic copy)
1	MS0812	1544	V.J. Romero	(electronic copy)
1	MS0825	1514	M. Pilch	(electronic copy)
1	MS0828	1544	W.R. Witkowski	(electronic copy)
1	MS0828	1544	A.R. Black	(electronic copy)
1	MS0828	1544	B. Carnes	(electronic copy)
1	MS0828	1544	K.J. Dowding	(electronic copy)
1	MS0828	1544	A.C. Hetzler	(electronic copy)
1	MS0828	1544	G.E. Orient	(electronic copy)
1	MS0828	1544	J.R. Red-Horse	(electronic copy)
1	MS0828	1544	A. Urbina	(electronic copy)
1	MS0828	1544	Ben Schroeder	(electronic copy)
1	MS0828	1544	Josh Mullins	(electronic copy)
1	MS0828	1544	Justin Winokur	(electronic copy)
1	MS0828	1544	Sarah Kieweg	(electronic copy)
1	MS0828	1544	Tom Paez	(electronic copy
1	MS0829	431	B.M. Rutherford	(electronic copy)
1	MS0829	431	J.T. Newcomer	(electronic copy)
1	MS0829	9436	Lauren Hund	(electronic copy)
1	MS0829	9436	Adah Zang	(electronic copy)
1	MS0830	9436	John R. Lewis	(electronic copy)
1	MS0897	1544	K.D. Copps	(electronic copy)
1	MS1138	6923	B.S. Paskaleva	(electronic copy)
1	MS1168	1356	Steven Wix	(electronic copy)
1	MS1173	5443	Alan Mar	(electronic copy)
1	MS1177	1355	J.P. Castro	(electronic copy)
1	MS1318	1440	T.G. Trucano	(electronic copy)
1	MS1318	1441	B.M. Adams	(electronic copy)
1	MS1318	1441	M.S. Eldred	(electronic copy)
1	MS1318	1441	L.P. Swiler	(electronic copy)
1	MS1320	1441	V.G. Weirs	(electronic copy)
1	MS1323	1443	W.J. Rider	(electronic copy)
1	MS9152	8759	Jaideep Ray	(electronic copy)
1	MS9159	8954	P.D. Hough	(electronic copy)
1	MS0899	9536	Technical Library	(electronic copy)

Sandia National Laboratories