SANDIA REPORT

SAND2020-2828
Unlimited Release
Printed March 2020

Conservative Estimation of Tail Probabilities from Limited Sample Data

Charles F. Jekel, Vicente J. Romero

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.

Sandia National Laboratories

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

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@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2020-2828
Unlimited Release
Printed March 2020

Conservative Estimation of Tail Probabilities from Limited Sample Data

Charles Jekel
cjekel@sandia.gov

Vicente Romero
vjromer@sandia.gov

Abstract

Several sparse-sample uncertainty quantification (UQ) methods are compared for conservative but not overly conservative estimation of small tail probabilities involving responses that lay beyond specified thresholds in the tails of probability distributions. Sixteen very differently shaped distributions (or probability density functions, PDFs) and tail probability magnitudes ranging from 10^{-5} to 10^{-1} are considered in order for the study to be relevant to a wide range of risk analysis and quantification of margins and uncertainty (QMU) problems. The emphasis of the study is on limited data regimes ranging from $N = 2$ to 20 samples, reflective of most experimental and some expensive computational situations. Relatively simple sparse-sample UQ methods tested for this regime involve statistical tolerance interval "Equivalent Normal" and related "Ensemble of Normals" and "Superdistribution" (SD) approaches. (The independently derived SD is effectively equivalent to the Bayesian posterior predictive distribution given the assumptions of the derivation.) The performance of the methods was generally improved for $N \geq 5$ samples with a generalized Jackknife resampling technique, which determines a tail probability estimate by averaging estimates from smaller sub-samples. Several quantitative metrics for method conservatism and accuracy of tail probability estimation are used to assess and rank the methods' performance over many random trials for each test PDF and probability magnitude. A variant of Bootstrap resampling was also tried, but did not significantly improve tail probability estimates in most cases. Detailed results are presented from over 100-million tests over the above factors that provide useful granular information on which methods or combination of methods perform best in various areas of the factor space.

Contents

Appendix

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 samples come from. 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 model were perfect in every other way, use of the model with sparse samples of the random-data inputs would likely underestimate the (strength, displacement, etc.) 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.

Previous work by Romero et al. [1] and [2] investigated a class of simple and effective uncertainty quantification (UQ) methods for dealing with sparse samples of random variables. The conservatively biased UQ methods include statistical tolerance intervals (TIs) and their "Equivalent Normal" distributions and related "Ensemble of Normals" (EON) and "Superdistribution" (SD) approaches. These methods are summarized in Section 2 of this report. They can be used to reliably bound useful characteristics of random quantities (for engineering purposes) with very few samples *without trying to infer or model the PDF from which the samples come*. This is key, as it is viewed that trying to estimate the true PDF from sparse sample data (even with estimated uncertainty from mathematically or statistically elaborate methods) is fraught with difficulty and may not be a productive strategy. The methods apply to limited samples from either experimental or simulation data.

Ref. [1] quantifies the performance of these methods at bounding the central 95% of response (between the 2.5 and 97.5 percentiles) on 74 analytical and empirical distributions. Tolerance intervals of 95%coverage/90%confidence and 95%coverage/ 95%confidence were generally found to be more reliably conservative and accurate (in terms of not being overly conservative) than the EON and SD approaches according to several quantitative performance metrics when tested for $N = 2$ to 10 samples on four diverse analytic PDFs (Normal, Log-Normal, Weibull, and 5 degree-of-freedom t). TIs also out-performed EON and SD on 70 empirical PDFs at sample sizes tried of $N = 2$, 4, 10, and 20.

Ref. [3] briefly reviews other UQ approaches, including bootstrapping methods and Bayesian or other type parameter estimation for proposed parametric and non-parametric PDF modeling approaches. Limited testing of many of these in [4] and recommendations in the literature indicate these other methods are generally not suitable for central 95% response estimation with very sparse data (e.g. $N \leq 10$). However, more extensive testing (as on the said 144 PDFs and $N = 2$ to 20 samples) would be necessary to more definitively assess the suitability of these other UQ methods.

Ref. [1] also characterizes the performance of the TI, EON, and SD methods when used to bound one-tail probabilities of magnitude 10^{-4} on 12 analytical and empirical distributions. $N = 2$,

4, 10, and 20 samples were tried. Appropriate UQ treatments for small tail probabilities and sparse sample data are important for assessing performance and safety margins in design, risk, and reliability analysis. Among the TI, EON, and SD methods, SD was found to consistently perform best (high reliability of conservative estimation without being overly conservative) according to several quantitative metrics applied over 10,000 random trials for each test PDF.

The present work substantially expands Ref. [1]'s investigation of tail probability estimation methods and their conservatism and accuracy performance. The 12 analytical and empirical distributions are revisited and four new analytic distributions are added to the study. Figure 1 shows the 16 very differently shaped distributions. Each PDF involves test problems with tail probability magnitudes of 10^{-1}, 10^{-2}, 10^{-3},10^{-4}, and 10^{-5} in order for the study to be widely relevant to a large range of risk analysis and quantification of margins and uncertainty (QMU) problems. The emphasis in the present study is on sparse-data regimes ranging from $N = 2$ to 20 samples, reflective of most experimental and some expensive computational situations.

In addition to characterizing the conservatism and accuracy performance of TI, EON, and SD methods on this expanded test matrix, the present work combines these methods with generalized Jackknife resampling techniques which determine a tail probability estimate by averaging estimates from smaller sub-samples. Appropriate levels of sub-sampling and averaging are suggested from numerical investigation over a subset of the large test matrix. Jackknifing with the suggested sub-sampling is shown to generally improve the performance of TI, EON, and SD methods substantially. The SD and SD-Jackknife methods are found to perform better on average than the other methods, and regimes of best performance of SD vs. SD-Jackknife are identified.

The present work also establishes a much simpler and less computationally expensive method than used in [1] and [2] for calculating tail probabilities with the SD approach. Convergence of EON and SD tail probability estimates are also demonstrated (in an advance beyond [1]) and the numbers of Normal PDF realizations needed for effective use of the methods are identified.

In a late development, the authors have also empirically confirmed the correctness of a reviewer comment (on a follow-on paper to [5] and to this SAND report) that the Superdistribution is essentially the same as a Bayesian posterior predictive distribution (PPD) based on assumptions consistent with those that underlie the SD's construction. The Bayesian PPD has an analytic form of a non-standard t-distribution, see e.g. [6]. A brief discussion is also given in [7]. It is somewhat easier to calculate tail probabilities with the analytic form than with the SD approach (though this is not difficult either). The more important implication is that it is useful and reassuring to know that this equivalency exists. It corroborates the SD method. It is also valuable to see that the SD can be constructed from relatively simple frequentist concepts and sampling approaches, and the result is equivalent to the Bayesian PPD that has substantial mathematical-statistical development behind it and which ultimately yields a simple analytic expression for the distribution.

A variant of Bootstrap resampling with the TI, EON, and SD methods is also studied in this report. It is found that at least this particular variant of Bootstrapping does not improve the performance of TI, EON, and SD methods appreciably in most cases.

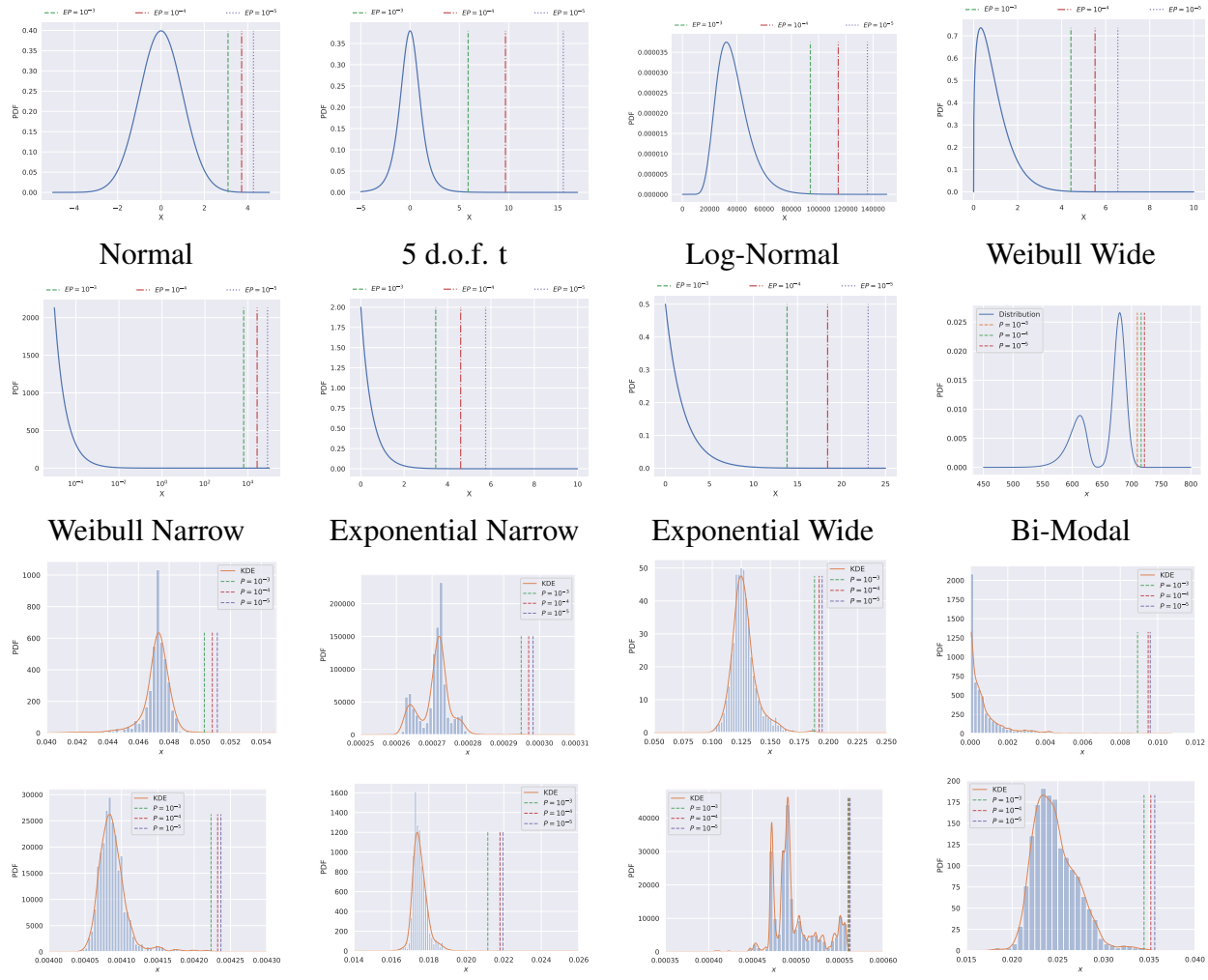


Figure 1: Collection of the 16 distributions analyzed and the corresponding right tail thresholds that integrate to 10^{-3}, 10^{-4}, and 10^{-5}. Note: kernel density fits to the empirical histograms are used in the investigations in this report.

2 A Class of Simple and Effective UQ Methods for Conservative Estimation of Tail Probabilities

Most of the tail probabilities we consider in this report are right-tail probabilities corresponding to integration of the applicable PDF to the right of a specified threshold or limit value. Appendix D. presents an illustrative example of a tail probability calculation. Since this is the probability that random-variable response governed by the PDF exceeds the specified threshold value, the term exceedance probability (EP) is alternatively used in this report for these types of tail probabilities.

2.1 Tolerance-Interval Equivalent Normal (TI-EN)

Tolerance Intervals (TIs) are a simple way to approximately account for the epistemic sampling uncertainty introduced from finite samples of a random variable. 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. However, extensive testing on 143 mildly to highly non-Normal distributions (including multi-modal and highly non-symmetric) in [1] and [3] shows that TIs perform with reasonably (usefully) high reliability or confidence on many highly non-normal distributions.

For instance, 95/90 TIs using $N = 4$ random samples per trial successfully bounded the true central 95% of response in 75% or more of 10,000 trials per PDF for $\approx 76\%$ of the PDFs (56 of 74). Slightly better 95/90 TI performance occurred for 70 other empirical PDFs tried in [3]. From application of 95/95 TIs to a median-representative PDF in [1] it is projected that reliability rates will increase by about 10% on average if 95/95 TIs are used instead. These success rates are reasonably high for $N = 4$ samples even though most of the 144 PDFs are highly non-Normal and information on their shapes is not needed or used by the TI methods. TI reliability levels slowly decline with increasing number of samples [1]. This makes 95/90 TIs of marginal use for $N > 6$ (where $< 70\%$ average reliability exists over the 144 PDFs), and 95/95 TIs are similarly marginal for $N > 8$.

As Figure 2 illustrates, a TI is constructed by first calculating the mean $\tilde{\mu}$ and standard deviation $\tilde{\sigma}$ of a sample. The tolerance interval is centered about the calculated sample mean. The bounds are determined by multiplying the sample standard deviation by a factor f,

$$L = \tilde{\mu} - f\tilde{\sigma} \tag{2.1}$$

$$U = \tilde{\mu} + f\tilde{\sigma} \tag{2.2}$$

where L and U represent the lower and upper bound of a X%/Y% TI. The factor f depends on the desired coverage, confidence, and the N number of samples. There are tables to look up f in [8] and [9]. A table is provided in Appendix C which provides f for various 95% coverage TI, as well

as a Python function to calculate f provided a coverage, confidence, and the number of samples using the equations in [10] and [11].

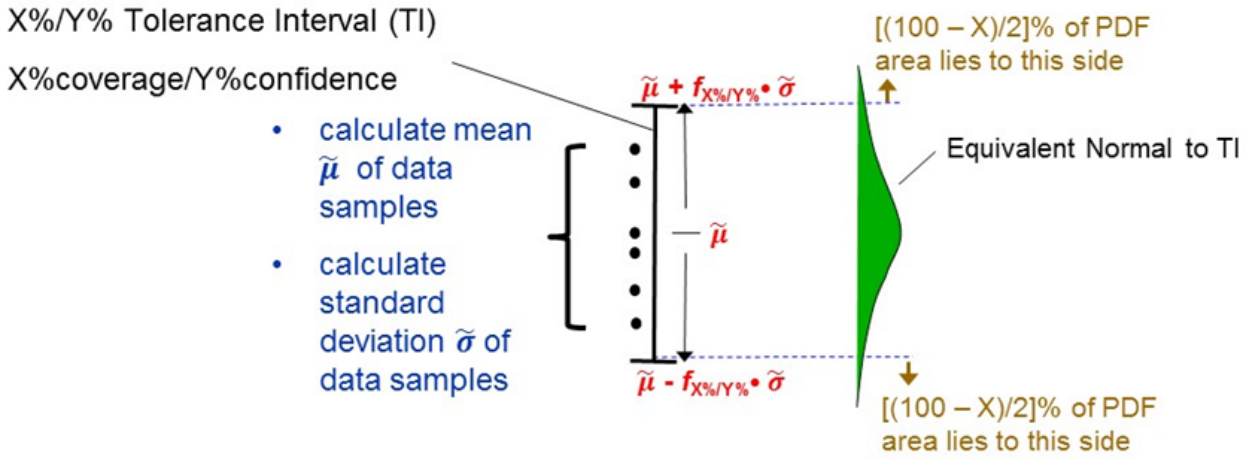


Figure 2: Construction of a tolerance interval and its "Equivalent Normal" distribution (from [1]).

A TI "Equivalent Normal" (TI-EN) distribution (see Figure 2) can be determined by finding an equivalent normal standard deviation σ_{EN}. If 95% coverage is used, then

$$\sigma_{\text{EN}} = \frac{f\tilde{\sigma}}{1.96} \tag{2.3}$$

and the resulting TI-EN is a Normal distribution of mean $\tilde{\mu}$ and variance of σ_{EN}^2.

Alternatively, it was determined during the course of the present work that equations in [10] and [11] can be used to determine σ_{EN} as a function of only the confidence level and sample standard deviation. Effectively we arrive at a factor k which can be multiplied by the sample standard deviation to determine the equivalent normal standard deviation as

$$\sigma_{\text{EN}} = k\tilde{\sigma}. \tag{2.4}$$

Then k can be determined from

$$k = \sqrt{1+N^{-1}}\sqrt{\frac{N-1}{\chi^2_{N-1;\alpha}}}\sqrt{1+\frac{N-3-\chi^2_{N-1;\alpha}}{2(N+1)^2}} \tag{2.5}$$

where N is the number of samples, $\chi^2_{N-1;\alpha}$ is the percentage point function of a Chi-Squared distribution with $N-1$ degrees of freedom evaluated at α, where $1-\alpha$ represents the stated confidence level. For example, a 95% confidence level results in $\alpha = 0.05$. A TI-EN only depends on the sample mean, the sample standard deviation, a desired confidence level, and the number of samples, N. A Python function to calculate k, and a table of calculated values, are included in Appendix C. Because TI-ENs are constructed particular to a specified confidence level like 90%, we sometimes associate them with the confidence level, e.g., "TI-EN90".

Reasoning in [1] predicted that tail probabilities calculated with TI-EN90s based on very few samples of the true PDFs should likely be conservative estimators of tail probabilities, i.e., over-estimate them. Studies in [1] with $N = 2$, 3, and 4 sample TI-ENs found over-estimation of true 10^{-4} tail probabilities in $> 70\%$ of trials for 10 of 12 PDFs from Normal to highly non-Normal. However, the estimates were often overly conservative. Significantly more testing is presented in Section 3 of the current report.

2.2 Ensemble of Normals (EON)

The reasoning and methods for obtaining the Ensemble of Normals (EON) from sparse samples are defined in Section 2 of [1] and illustrated in Figure 3. First, the sample mean and standard deviation are calculated from a given sparse data set. Then a set of candidate means and standard deviations are randomly generated using Student's T and Chi-Squared distributions as illustrated. The T and Chi-Square distributions govern the probabilistic uncertainty of the sample mean and standard deviation when a Standard Normal distribution is sampled. These are used in a reverse sense to generate realizations of what the true mean and standard deviation (and thus true Normal distribution) might be, from a sample mean and sample standard deviation. The EON is the collection of Normal distributions determined from the candidate means and standard deviations. This methodology and the further post-processing explained below are applied to non-Normal distributions with often surprising success in conservative tail probability estimation as discussed later.

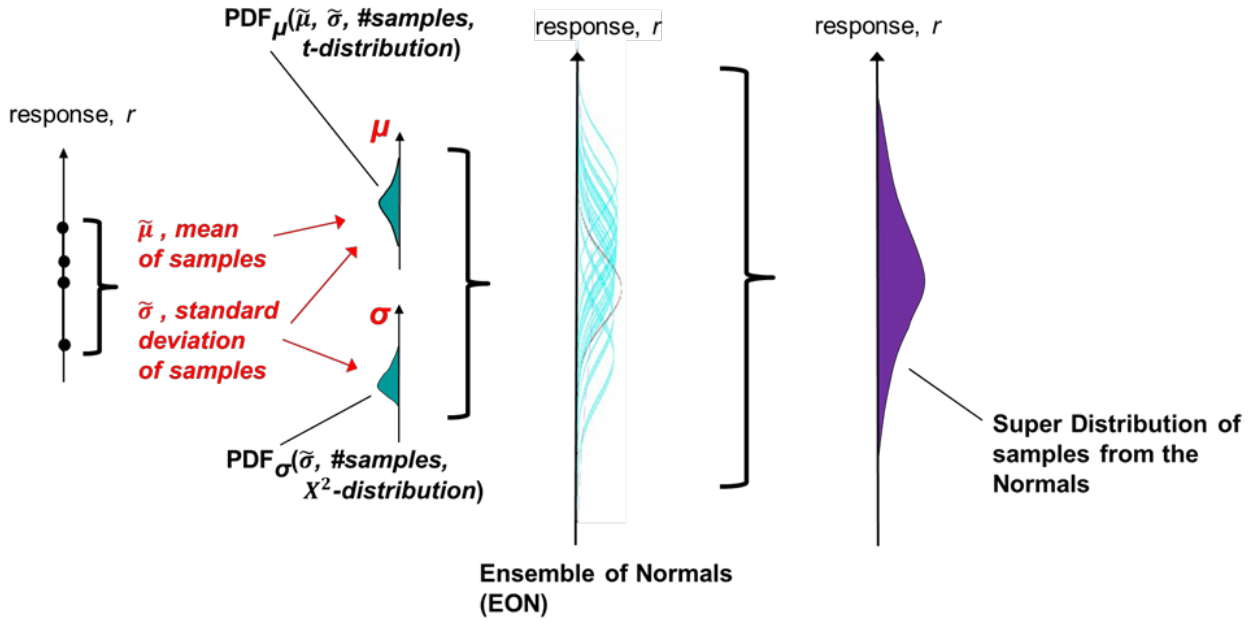


Figure 3: Depiction of process for obtaining an Ensemble of Normals (EON) and associated Superdistribution from a set of response samples. Figure from [1].

If N samples are drawn from a Normal distribution $\mathcal{N}(\mu, \sigma)$ 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.

$$\mu_i = \tilde{\mu} + \frac{T_i \tilde{\sigma}}{\sqrt{N}} \tag{2.6}$$

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 [12] that the central $(1-\alpha)\%$ range of an asymptotically large set or distribution of candidate means $[\mu_i]$, from the distribution's $\alpha/2$ to $(1-\alpha/2)$ quantiles, is a $(1-\alpha)\%$ "confidence interval" (CI) that will contain the true mean μ exactly $(1-\alpha)\%$ of the time. That is, so-produced CIs will successfully contain the true mean in $(1-\alpha)\%$ of a very large number of trials, where each trial X involves drawing N random samples from the said Normal distribution $\mathscr{N}(\mu,\sigma)$ 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.

$$\sigma_i = \tilde{\sigma}\sqrt{(N-1)/\chi_i^2} \tag{2.7}$$

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 [12] that the range between the $\alpha/2$ to the $(1-\alpha/2)$ quantiles of the asymptotically large distribution or set of candidates $[\sigma_i]$ constitutes a $(1-\alpha)\%$ confidence interval that will contain the true standard deviation σ exactly $(1-\alpha)\%$ 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 4) among which the true distribution $\mathscr{N}(\mu,\sigma)$ may exist. In practice it is usually not essential that the true Normal PDF $\mathscr{N}(\mu,\sigma)$ 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;

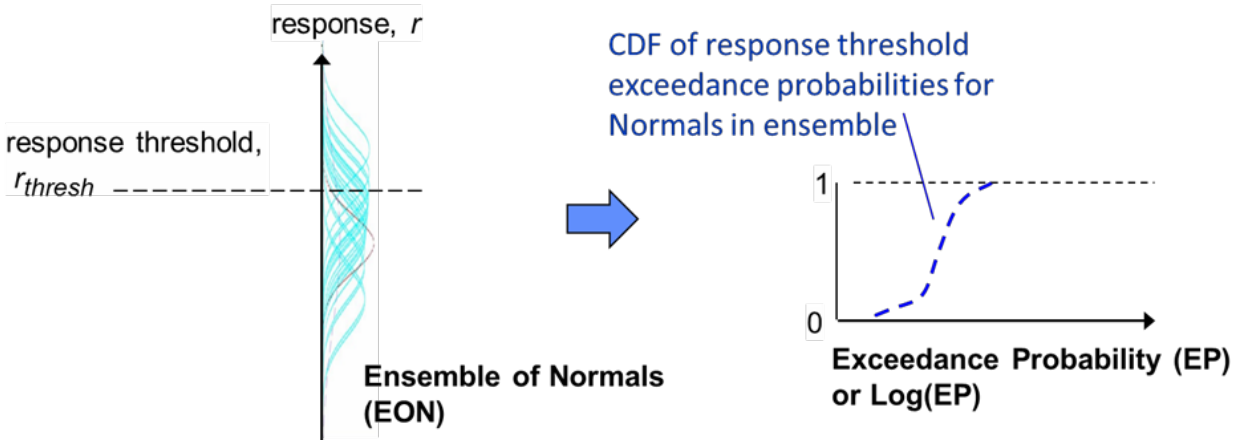


Figure 4: Cumulative density function (CDF) of Exceedance Probabilities calculated from the Ensemble of Normals and a specified threshold level of response.

and on the number of generated candidate Normals. The odds in many practical cases are relatively high as established in [1] 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 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 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 3 and 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.

TIs and EONs are related as follows. The EON approach can provide uncertainty information on PDF percentiles inferred from the sample data, as opposed to just conservative point estimates that TIs or their equivalent Normals provide. For example, consider the 97.5 percentile of response for a Normal distribution. If sparse sample data from the Normal distribution is used to construct an EON, the 97.5 percentile of each Normal in the EON can be identified, and the set of values comprises a distribution within which the true 97.5 percentile of the Normal distribution originally sampled lies. It turns out that the 0.9 quantile of the distribution of possible 97.5 percentile values from the EON coincides with the top of a 95/90 TI (very closely per empirical testing in [2]). Likewise, the lower end of the TI would coincide with the 0.1 quantile of the distribution of possible 2.5 percentiles of the Normals in the EON.

Likewise, given a response threshold, a distribution of EP estimates exists from the EON as

depicted in Figure 4. The question arises: What percentile of this distribution would be most suitable in general for reliably conservative but not overly conservative estimates of tail probabilities? Estimates using the 50th percentile (EON50 estimates) and the 90th percentile (EON90 estimates) were examined in [1] for 12 PDFs and $N = 2$, 4, 10, 20 and 10^{-4} tail probabilities. The EON50 estimates were not reliably conservative enough. The EON90 estimates were usually slightly less reliably conservative than TI-EN90 estimates, with the caveat that the EON results in [1] may not have been sufficiently converged as explained next.

A total of L Normal distributions from L candidate means and candidate standard deviations comprise an EON. The present work uses $L = 10^4$ as a reasonable number of distributions to predict an EP, as established in the convergence study in Appendix E.1. A couple example EONs are presented in Appendix F, which shows overlays of all 10^4 distributions in an EON.

The study in [1] used $L = 100$ normal distributions per EON. Results there are qualitatively similar to results in the present report wherever the same problem is worked, but results in the present report are much better converged (effectively fully) so are the only results to be considered quantitatively correct for the EON method. Furthermore, the testing in the present report (Section 3) involves significantly more test problems involving a broader set of PDFs and tail probability magnitudes, and higher granularity in terms of numbers of samples N. Only EON90 estimates are considered in the present report.

2.3 Superdistribution (SD)

The Superdistribution (SD) method first starts by constructing L number of distributions as an Ensemble of Normals (EON) as described in [1] and shown in Figure 3. The previous SD method described in [1] randomly sampled each distribution in the EON M number of times (e.g. 100,000 times), and then all the $L \times M$ samples were binned to form an empirical SD. Per the description below, a desired exceedance probability of a SD can be determined very computationally efficiently without having to generate, store, and bin the $L \times M$ samples that form an empirical SD.

Note that the PDF and therefore CDF are known for each Normal distribution in an EON. If we were to randomly sample each distribution in the EON, then for each distribution there would be some number of samples $<$ a posed response value x. As M grows to infinity, the proportion of random samples $< x$ (out of the total number of randoms samples M) will define the value of that CDF at x. This will be true for each of the L distributions in the EON. Averaging the L CDF values at x is therefore equivalent to adding all the L distributions' samples that are $\leq x$ and dividing by the grand total number ($L \times M$) of samples. This ratio also defines the value of the SD CDF at x.

Hence, the CDF of the Superdistribution can be defined as

$$\mathrm{SD}_{\mathrm{CDF}}(x) = \frac{\sum_{i=1}^{L} \mathrm{F}_i(x)}{L} \tag{2.8}$$

where F_i is the CDF of the i^{th} distribution in the EON, which is evaluated at x. This process is illustrated in Figure 5, which shows that averaging the individual Normal CDF values at $x = 1.5$ yields

the Superdistribution CDF value at $x = 1.5$. Evaluating the CDF value of a Normal distribution at any input value x is a standard function call in most software packages, even Excel, so this method of SD CDF calculation is fast and convenient. The SD left-tail probability integrated to the left of a specified threshold x in the left tail of the SD is equal to its CDF evaluated at x (Eqn. 2.8). The SD right-tail probability for a specified threshold x in the SD's right tail is given by $1 - \text{SD}_{\text{CDF}}(x)$. Spot checks in the present work confirmed the sameness of tail probabilities calculated the new way using Eqn. 2.8 and the prior method from [1] that constructs the full empirical SD.

EP estimates with SDs in [1] for 12 PDFs and $N = 2$, 4, 10, 20 and 10^{-4} tail probabilities were on average significantly more accurate than estimates with TI-EN90, EON50, and EON90 (according to accuracy metrics used also in the present report that penalize for errors of both over-estimation and under-estimation). For most of the 12 PDFs tried, SD reliability in conservatively bounding the 10^{-4} tail probabilities declined very quickly with sample size N. The results are summarized as follows.

- $N = 2$ samples: all 12 PDFs $> 90\%$ reliability of attaining a conservative EP estimate
- $N = 3$ samples: all 12 PDFs $> 70\%$ reliability and smaller avg. over-estimation errors
- $N = 4$ samples: 10 of 12 PDFs $> 70\%$ reliability and even smaller average errors.

These results are subject to non-convergence error because the study in [1] constructed each Superdistribution from only $L = 100$ normal distributions. Nonetheless, the results are qualitatively similar to results in the present report wherever the same problem is worked. The results in Sections 3 and 4 are much better converged (effectively fully) so are the only results to be considered quantitatively correct for the SD method. The testing in the present report also involves significantly more test problems as mentioned previously.

2.4 Relative Variances of Sparse-sample Method Distributions

Figure 6 plots the standard deviation magnitudes of a TI-EN90, SD, and a histogram of standard deviations from $L = 5000$ Normals of an EON from a sparse-sample problem studied in [2]. The plotted sample standard deviation from the problem's $N = 4$ raw data samples is $\tilde{\sigma} = 0.0056$. The SD has a standard deviation $\sigma_{\text{SD}} = 0.0105$. 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 TI-EN90 is $\sigma_{\text{TI-EN90}} = 0.0142$. This coincides with the 92nd percentile of the histogram and is about 150% larger than the data standard deviation. The TI-EN90 standard deviation is about 35% larger than the SD standard deviation. In general, a TI-EN90 distribution is characteristically broader than its counterpart SD distribution, and both are substantially broader than a Normal distribution fit to the raw data, which would have the sample standard deviation $\tilde{\sigma} = 0.0056$ in this case.

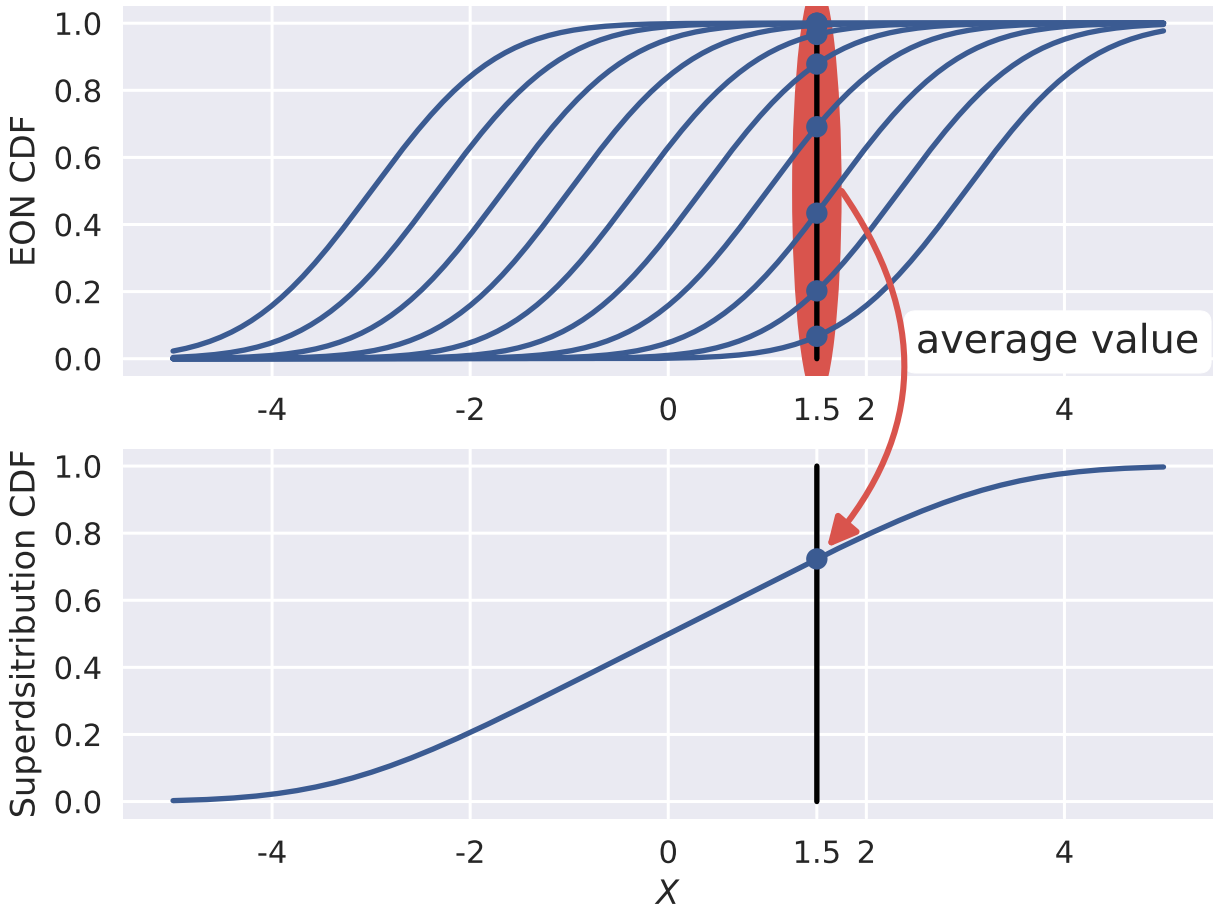


Figure 5: Example of 10 Normal distributions in an EON. The Superdistribution CDF value at $x = 1.5$ is calculated by averaging the CDF values from each Normal distribution in the EON at $x = 1.5$.

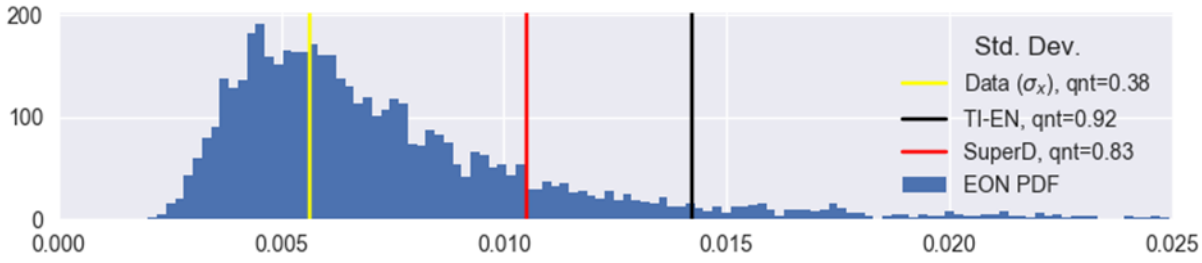


Figure 6: Magnitudes of standard deviations from N=4 data samples and corresponding distributions from sparse-sample methods applied to a problem in [2].

3 Performance of Bounding One-Tail Probabilities on 16 Diverse PDF Shapes

The goal of this study is to quantify the accuracy and reliability of EP predictions from sparse samples. First, a specified number N of random samples were generated from a given statistical distribution. For each sample set, an EP is predicted from each of the UQ methods for a PDF tail value (threshold) corresponding to a true EP of $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$, or 10^{-5}. (Figure 1 shows the threshold levels for $10^{-3} to 10^{-5}$ EP magnitudes.) Each estimated EP was compared to the true EP. This process was repeated 10,000 times to provide insight on the overall performance of each method for the PDF, number of samples, and true EP magnitude. The number of samples considered ranged from $N = 2$ to $N = 20$ for each of the 16 PDFs and EP magnitudes. Subsections 3.1 to 3.9 focus on evaluation of results for $10^{-3} to 10^{-5}$ EP magnitudes. Performance trends over this range prompted a follow-on study for EP magnitudes $10^{-1} and 10^{-2}$ in subsection 3.10.

To illustrate the procedure, a few example EP predictions are shown in Figure 7 for combinations of $N = 2, 10, 20$ and the various true EPs on the Standard Normal distribution. Empirical CDFs of the EON method are shown in all of these cases. A small kink can be seen in the empirical CDFs of the EONs for small sample sizes ($N = 4$). The reason for this kink is explained in Appendix F.

Only the 90th percentile values on the EON CDFs are used in the rest of this report, as a reasonable choice identified in prior studies [1]. The EON90 values are in the mix of the other methods' EP estimates in Figure 7. EON90 EPs are unconservative in three of the nine cases. EON95 EPs would be unconservative in only one of the cases ($N = 10$, EP$= 10^{-5}$), but would be among the most conservative methods in the other eight cases—akin to the TI-EN95 results—which are unnecessarily conservative/inaccurate over the 16 PDFs as will be established later. All methods have cases of both overshooting and undershooting the true EP in these examples.

To quantitatively differentiate the methods in terms of accuracy and conservatism performance, a performance metric from [1] is used:

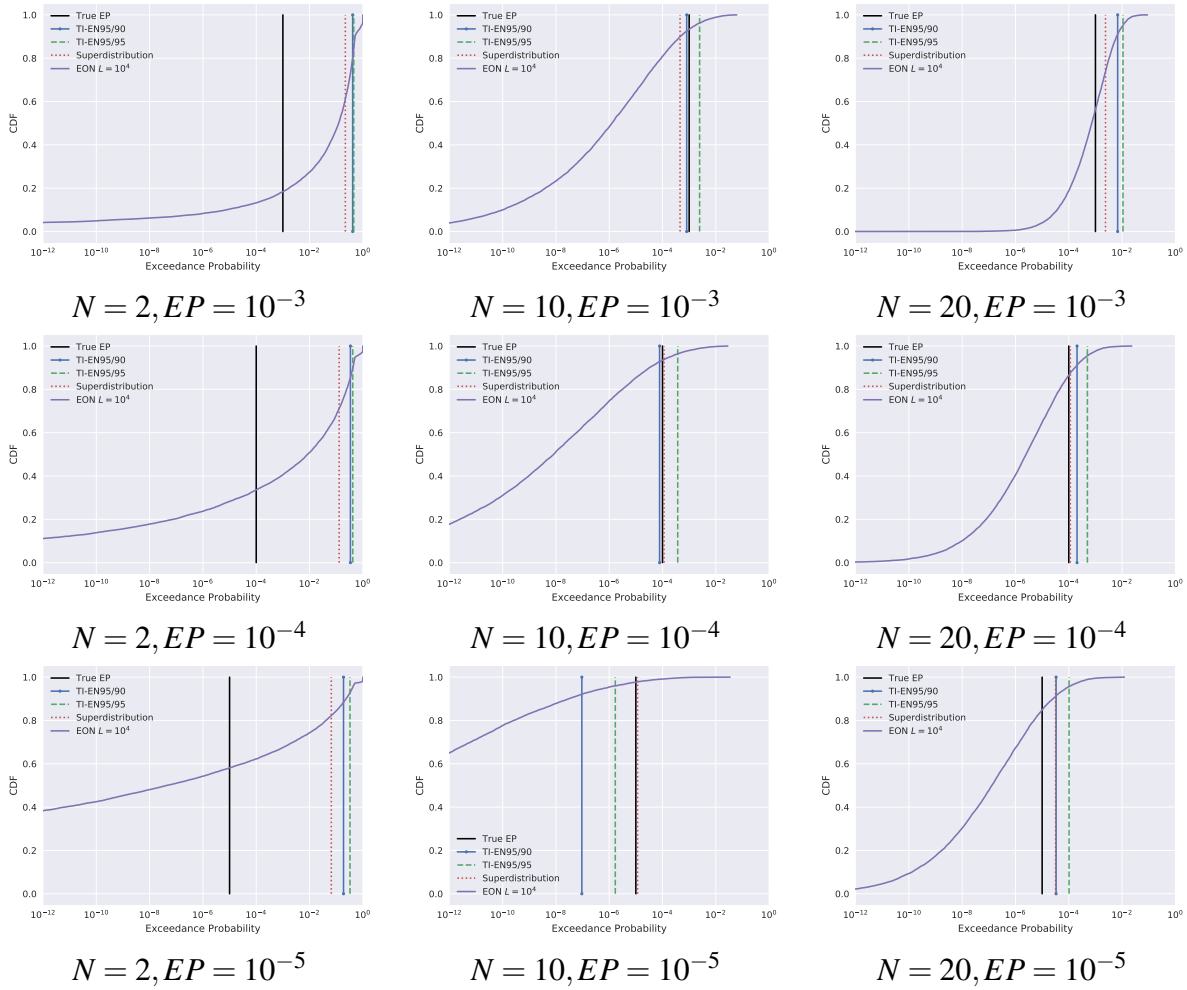
$$\text{EPmetric} = \left[\sum^{N+} \Delta\log + \sum^{N-} |\Delta\log| \right] / N^{+} \tag{3.1}$$

where

$$\Delta\log = \log_{10}(EP_{\text{estimated}}) - \log_{10}(EP_{\text{true}}) \tag{3.2}$$

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. 3.1, the greater the proportion of overshoot errors contributing to that sum (and so the smaller the proportion of contributed undershoot errors), the better the method performance value.

It is important to consider the worst possible EPmetric values for predicting an exceedance probability of EP$= 10^{-4}$. The worst possible overestimate would be an EP$= 1.0$, which would result in a $\Delta\log = 4.0$. The worst possible underestimate would be an EP$= 0.0$, which would result in a $\Delta\log = -\infty$ because $\log_{10}(0.0) = -\infty$. It isn't practical to deal with $-\infty$, so the smallest



$N = 2, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$

$N = 2, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$

$N = 2, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 7: Comparison of different Exceedance Probability predictions from the tested methods for all combinations of $N = 2, 10, 20$ and EP$= 10^{-3}, 10^{-4}, 10^{-5}$ from random samples of the Standard Normal distribution.

number to represent any computed values of EP$= 0.0$ was chosen to be EP$= 10^{-305.5}$. With these approximations in place, the worst possible underestimate of a true EP$= 10^{04}$ (e.g. an estimated probability of 0.0) would yield a $\Delta\log = -309.5$.

A possible alternative performance metric was investigated in Appendix G. The alternative metric did not use a log in the numerator. As a result when predicting extreme one-tail probabilities, there would become a much larger possible bound for an overestimated EP than an underestimated EP. It happened that the method with the lowest performance metric values, also had the lowest alternative metric values. For the rest of our investigations we choose to use the original metric because it's scaling naturally awards conservative over-estimation, as opposed to favoring unconservative under-estimation that the alternative metric naturally rewards by its scaling. In practice an underestimated EP is much worse than a overestimated (conservative) EP, and thus prudent risk mitigation leads us to go forward with the original performance metric.

If undershoot errors are given a $10X$ magnitude amplification 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

$$\text{EPmetric10x} = [\sum^{N+} \Delta\log + 10\sum^{N-} |\Delta\log|]/N^{+}. \tag{3.3}$$

For the same set of overshoot and undershoot errors from a given set of trials, the numerator value in Eqn. 3.3 yields a higher/worse metric value with penalized undershoot errors vs. the non-penalized metric.

In practical applications, it is much safer to overestimate EP than it is to underestimate. Because of this, an additional reliability measure is considered. The reliability measure is defined as

$$\text{Reliability} = \frac{N^{+}}{\text{the total number of trials}} \tag{3.4}$$

which is the proportion of conservative estimates of EP. For this study, 10,000 replicate trials were performed for each case.

The EPmetric represents the combined error/accuracy and conservatism tendency associated with each method, while the Reliability measure represents the confidence that a given estimate will be conservative. The ideal UQ method for any given number of samples would have the lowest EPmetric and the highest Reliability. However, there is a natural trade off between accuracy and conservative predictions which are competing dynamics for the most part.

3.1 Standard Normal Distribution

The PDF and thresholds that integrate to EP=$10^{-3}, 10^{-4}, 10^{-5}$ on the Standard Normal distribution is illustrated in Figure 8. The EPmetric, EPmetric 10x undershoot penalty, and reliability of each method at predicting the EP from sparse samples of the Standard Normal distribution can be seen in Figure 9. The EPmetric of the TI-EN90, TI-EN95, and EON90 are consistently grouped together. The Superdistribution (SD) stands out with the lowest EPmetric for all number of samples and EPs $\leq 10^{-3}$. In most cases the EON90 had the worst EPmetric for small number of samples, but improved over the TI-EN90 and TI-EN95 at higher number of samples. In general for all methods the EPmetric decreases as the number of samples increases, reflecting more accurate estimates on average. Much of the same can be said for the EPmetric 10x undershoot penalty. It appears that the EPmetric 10x value uniformly improves for all methods as the number of samples increases, with the exception of the SD at EPs $\leq 10^{-3}$. Nonetheless, SD has the lowest or is indistinguishable from the lowest EPmetric and EPmetric 10x values over the full $N = 2$ to 20 range for EPs $\leq 10^{-3}$. For EPs $\leq 10^{-2}$ the best method is a bit more complicated, as the trends in EP= 10^{-2} are complete different than the other trends as noticed by the decreasing reliability of each method and higher EPmetric values.

The *Best TI-EN* curve in the figures represents the lowest EPmetric values from TI-EN for various confidence level settings. The curve does not reflect a practical method because the best

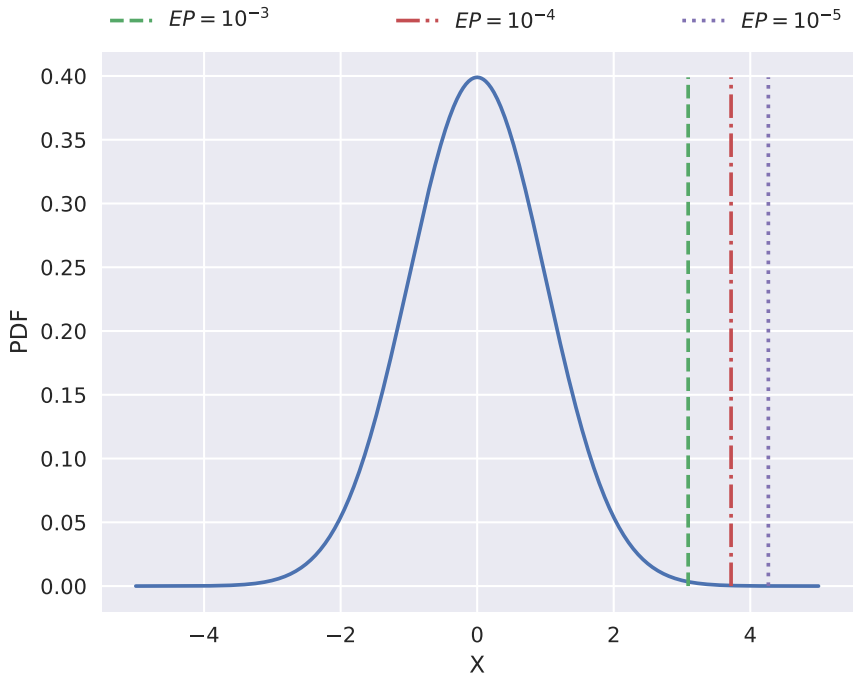


Figure 8: Standard Normal distribution.

confidence level for this metric is dependent upon the definition of the EPmetric and varies with PDF shape and EP magnitude (both are unknown in real problems). However, it does represent the curve from the best possible TI-EN obtainable for the Standard Normal PDF. Figure 10 shows which confidence levels were used in the *Best TI-EN* curve for EP=10^{-4}, where initially the lowest EPmetric value results from the highest confidence level (99.99%). As the number of samples is increased, the *Best TI-EN* confidence level drops to 75% - 80%. Figure 11 shows how the confidence levels affect the EPmetric results for any given number of samples. We can see the trend is much different for $N = 2$, than it is for higher number of samples. At $N = 2$ the best metric came from the highest confidence level. This isn't true with $N = 4$ and up, as the optimal confidence level was somewhere in between 95% and 73% depending on the number of samples. We can see that the best possible TI-EN for the lowest EPmetric doesn't produce EPmetric values that were better than with the Superdistribution. SD also yielded a higher reliability than the *Best TI-EN*.

It may appear odd that the TI-EN with 99.99% confidence had the lowest EPmetric performance values for $N = 2$. Essentially, with $N = 2$ there is limited potential magnitude of performance hits for overestimates, and much larger potential magnitude of performance hits for underestimates. For a 10^{-4} true EP, the difference of logs for an overestimate, e.g. 10^{-1}, gives a value of 3. Thus it's worthwhile to note that the Δlog can at most be 4 for an overestimate. For underestimates, the Δlog can theoretically be unlimited, however with finite math is approximated to be at most a value of -310. With $N = 2$ the Normals in the EON vary so widely that, for a non-100% confidence level, a substantial number of the 10^4 Normals will underestimate. Appendix F shows an example EON with $N = 2$ and $N = 10$, where there are a considerable number of

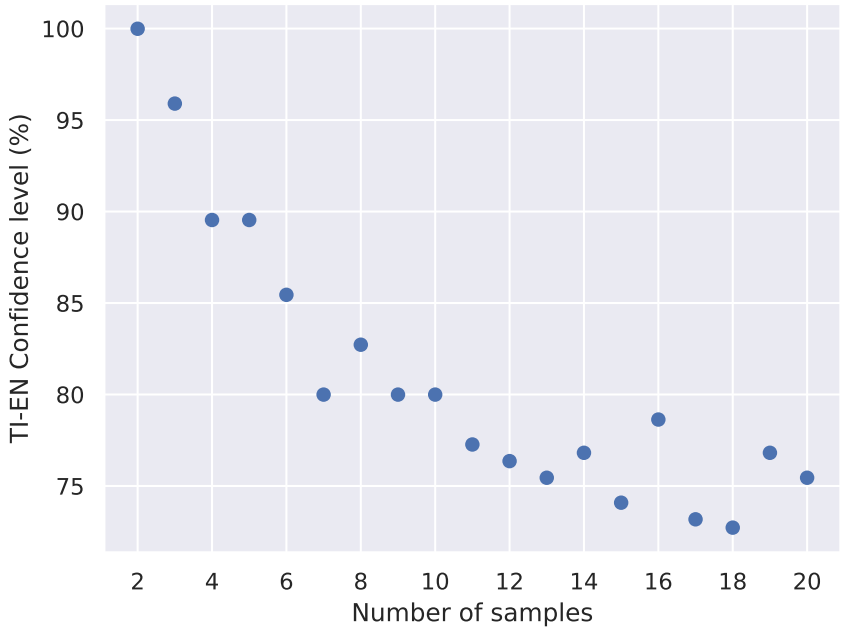


Figure 10: The confidence level which resulted in the *Best TI-EN* EPmetric value for a Normal PDF and EP=10^{-4}.

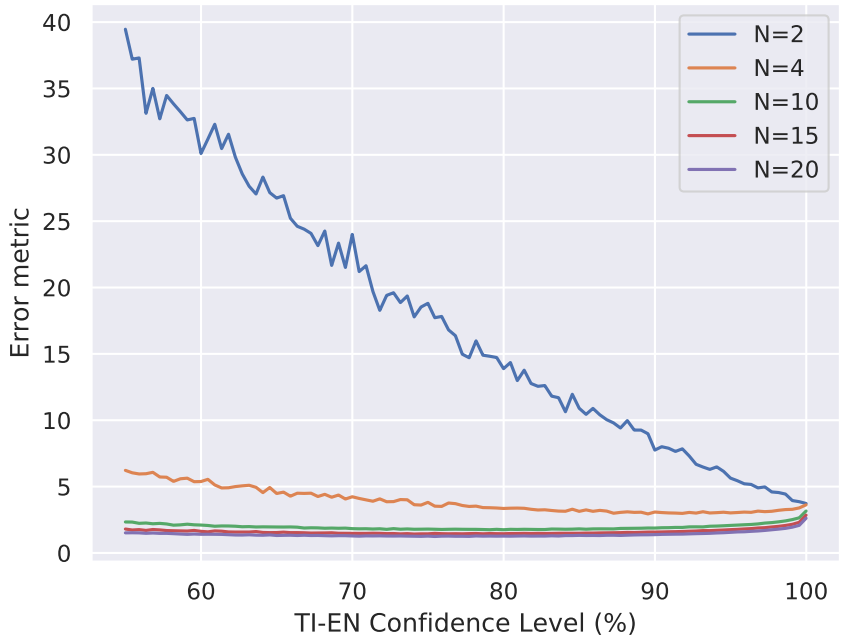


Figure 11: EPmetric value for various TI-EN confidence levels. Different curves represent the different number of samples for EP=10^{-4} and a Normal PDF.

Normal distributions which underestimate the true EP. These severe underestimates will have large Δlog like -10, -50, -100, -200, -300, which overpower the limited (max = 4) penalties of the

overestimates. The lower the TI confidence level, the more overpowering occurs. The least such overpowering occurs for the limiting confidence value of 99.99%, so the metric value is minimum at this confidence level (for $N = 2$).

When N increases, the variability between TI-EN distributions decreases, which reduces the magnitudes of underestimates (on average), which reduces the log differences to continually smaller average as N increases. No longer common are the large $\Delta\log$ values that occurred when $N = 2$. In fact, the $N = 4$ results already show a balance occurs between overestimate and underestimate downsides, at a confidence level of $\approx 89\%$ for EP$= 10^{-4}$. The balance point moves further as N increases, to a value of ≈ 76 % optimal confidence for N=20.

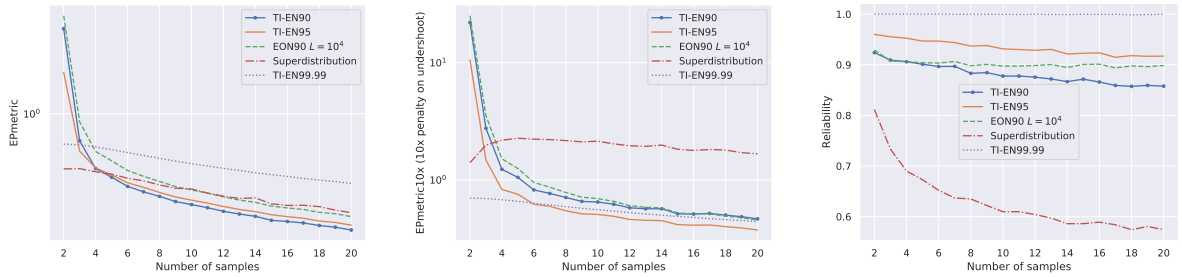
The reliability of the curves indicates what percentage of EP estimates were conservative. In general the TI-EN methods were consistent at producing a reliability near their advertised confidence level, essentially independent of the number of samples (expected if a Normal PDF). The exception occurs for EP$= 10^{-2}$, which shows that the reliability of the methods decreased as the number of samples increased. The TI-EN90 was about 90% reliable, while the TI-EN95 was about 95% reliable. The EON90 also had similar reliability for all number of samples, which was around 86%. Only the SD method had a strong reliability trend vs. the number of samples. For $N = 2$, the SD had a 99% reliability that decreases considerably as the N number of samples is increased.

The decay of SD reliability depends upon the predicted EP magnitude. For EP=10^{-3}, reliability decayed the most, to 78% at $N = 20$, while for EP=10^{-5} the reliability only decayed to 93% at $N = 20$. For a given N, SD was more reliable at predicting a smaller EP than a larger EP. The reliability levels were usefully high for SD being the most accurate of the methods.

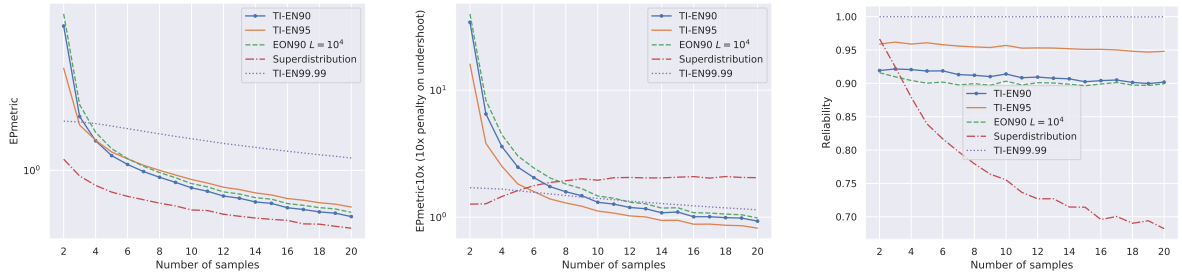
The study of TI performance using optimal confidence values is an interesting hypothetical exercise to see whether TI-ENs can have a better metric value than SDs under fully controlled optimal conditions for TI-ENs. They don't for the Standard Normal PDFs at EP=$10^{-3}, 10^{-4}, 10^{-5}$. Moreover, it is not fruitful to further the investigation of optimal TI-ENs to other EP magnitudes or PDF types because the optimal confidence values vary as a function of these two factors and N. Even if we could map out the function, it would not be possible to use the function in application circumstances because we will not know the true PDF or EP. Therefore, in the remainder of the document, we restrict to what is possible under general application conditions and assess the performance of 95/90, 95/95, and 95/99.99 TI-ENs (the latter being a limiting case of reliability = 1).

Figure 12 shows distributions of EP estimation errors for 10,000 trials of each method with $N =$ 4, 10, and 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. All extremely poor underestimates of EP resulting in $\Delta\log(EP) < -10$ were pooled collectively at a value of $\Delta\log(EP) = -10$. This results in the bump occurring at -10 for low number of samples ($N = 4$), which quickly goes away as the number of samples increases. Although the distribution of EON 90% is closest to the TI-EN90 distribution, there are significant differences; the relationship between the two methods' EPs are different as clearly seen by the previous reliability figures. An interesting observation is that it appears that EP$= 10^{-2}$ had the largest number of bad EP predictions for $N = 4$ based on the large bump occurring at a $\Delta\log = -10$.

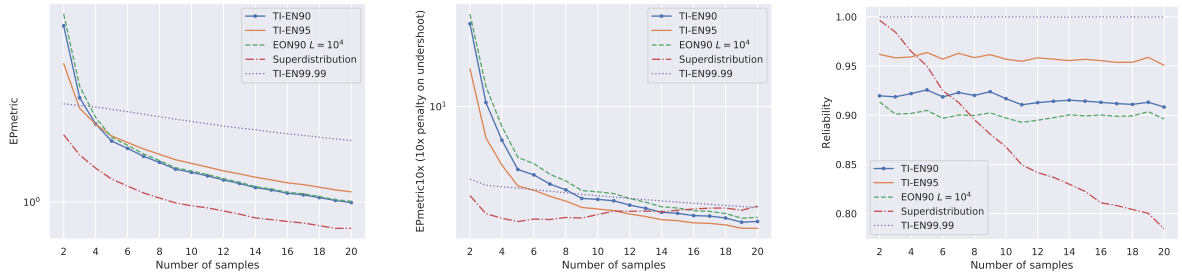
One thing to notice in the distribution of EP predictions is that the large majority of $N = 4$ results are all over-estimates of EP, highly concentrated about + 2 to 4 orders of magnitude error. Superdistribution out-performs the other methods on average over all samples sizes and EP levels; its error distribution always has a mean and peak closest to zero and often appears to be among the most compact in terms of its spread of values. The TI-EN99.99 results are always the most compact and have means farthest to the right; the method gives the largest overestimates on average and these are clustered within an order of magnitude of each EP's maximum possible overestimate. For all methods, the average error gets better (moves closer to zero) as the sample size increases. 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.



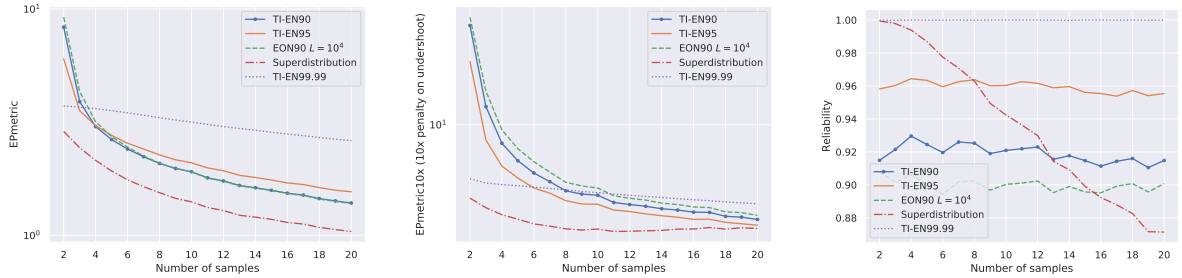
EPmetric, EP= 10^{-1} EPmetric10x, EP= 10^{-1} Reliability, EP= 10^{-1}



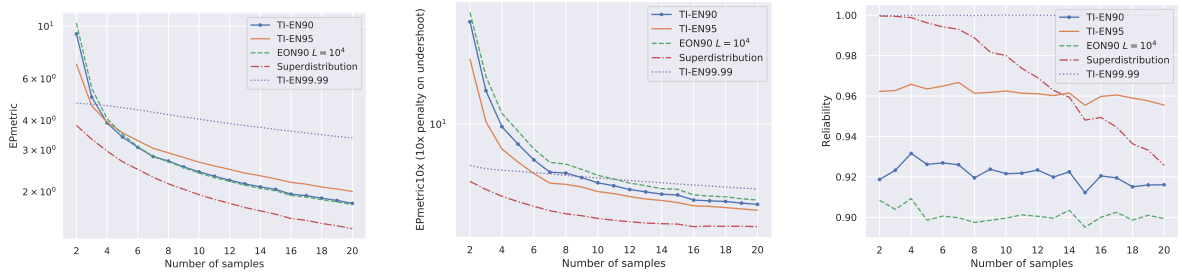
EPmetric, EP= 10^{-2} EPmetric10x, EP= 10^{-2} Reliability, EP= 10^{-2}



EPmetric, EP= 10^{-3} EPmetric10x, EP= 10^{-3} Reliability, EP= 10^{-3}



EPmetric, EP= 10^{-4} EPmetric10x, EP= 10^{-4} Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5} EPmetric10x, EP= 10^{-5} Reliability, EP= 10^{-5}

Figure 9: Results of predicting exceedance probability from the Standard Normal distribution. The EPmetric, EPmetric10x, and Reliability are shown for EP= 10^{-1} through 10^{-5}.

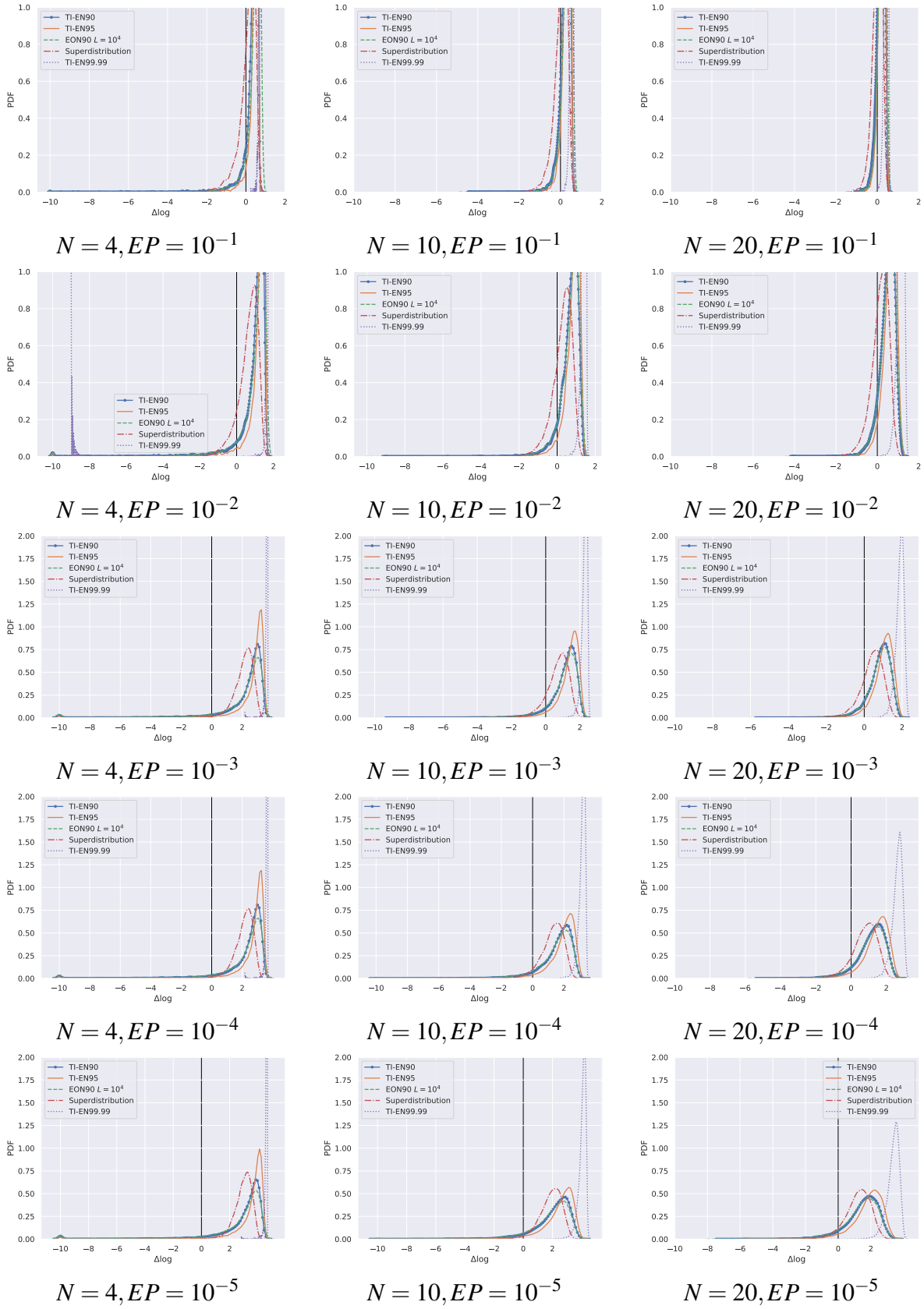


Figure 12: PDFs from 10,000 runs of predicting EPs from the Standard Normal Distribution. Positive Δlog values indicate an overestimate of the true exceedance probability, while negative underestimate.

3.2 Log-Normal distribution

Figure 13 shows the Log-Normal distribution previously studied in [1]. The performance metrics and reliability of each method at conservatively predicting the EP from sparse samples of the distribution are shown in Figure 15. For $N = 2$ to 20 and for all EP magnitudes, the EON90 method universally had worse EP and EP10x metric values with lower reliability than the TI-EN90 method, which universally performed worse than the TI-EN 95 method. The SD method had the lowest metric values (best performance) for $N \leq 4$ to 8 samples, depending on the EP level and the metric type (EP or EP10X). The TI-EN99.99 method had the lowest/best metric values for $N \geq 5$ to 9 samples. As explored next, the optimal choice of method and number of samples also depends on acceptable reliability levels and the methods' reliability performance per the right column of plots in Figure 15.

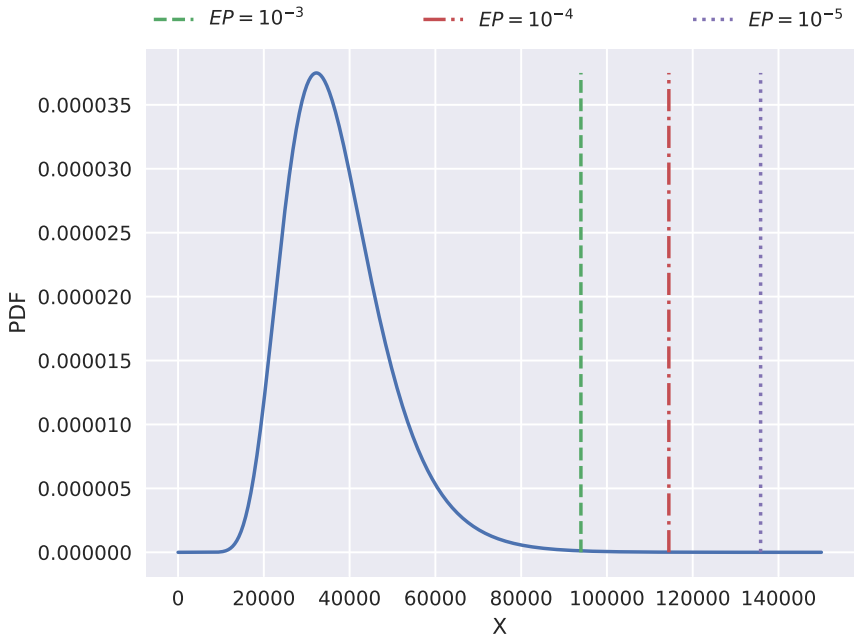


Figure 13: Log-Normal distribution.

The reliability of conservatively bounding the true EP for all methods decreased with number of samples for the Log-Normal distribution. SD and TI-EN99.99 reliabilities start at a highest possible reliability of 100% at $N = 2$ samples (for EP$\leq 10^{-3}$). Reliability of all methods drops as the number of samples increases, falling more quickly for smaller EP magnitudes than for larger magnitudes. SD reliability drops more quickly than for the other methods, falling to a reliability of about 50% at $N = 7$ to 9 samples, depending on EP magnitude. The other methods besides SD and TI-EN99.99 start with reliabilities that are significantly less than 1 at $N = 2$ samples, but fall at a slower rate than for SD (yet reach 50% sooner, at $N = 4$ to 8 samples depending on EP magnitude). For all EP levels, TI-EN99.99 reliability falls much more slowly than for the other methods, so retains the highest reliability over the full range $N = 2$ to 20 samples.

However, a practical downside of TI-EN99.99's high reliability is excessive conservatism. For

example, for $N = 4$ samples the TI-EN99.99 standard deviation multiplier is 27.1 from Table C.2 in Appendix C. The SD method multiplier is far less, approximately 1.9 or $1/14$ the size[1]. The result of the standard deviation multiplier can be seen in Figure 14, which shows the resulting distributions from the TI-EN90, TI-EN99.99, and SD. The distribution that results from the TI-EN99.99 is very flat when compared to the other distributions. The TI-EN99.99 distribution based on $N = 4$ samples will achieve $\approx 100\%$ reliability of conservative tail probability estimation for true EPs 10^3, 10^4, 10^5, but the $\approx 14\times$ wider distribution if used in design or risk analysis would lead to much more conservative and costly outcomes compared to use of the much narrower SD, which has usefully high reliabilities of $\approx 84\%$, 95%, 98% for EPs 10^3, 10^4, 10^5. Hence, even though more reliable at conservative tail probability estimation, the excessive conservatism of the TI-EN99.99 method results in significantly higher/worse EP and EP10X metric values than SD with 4 samples.

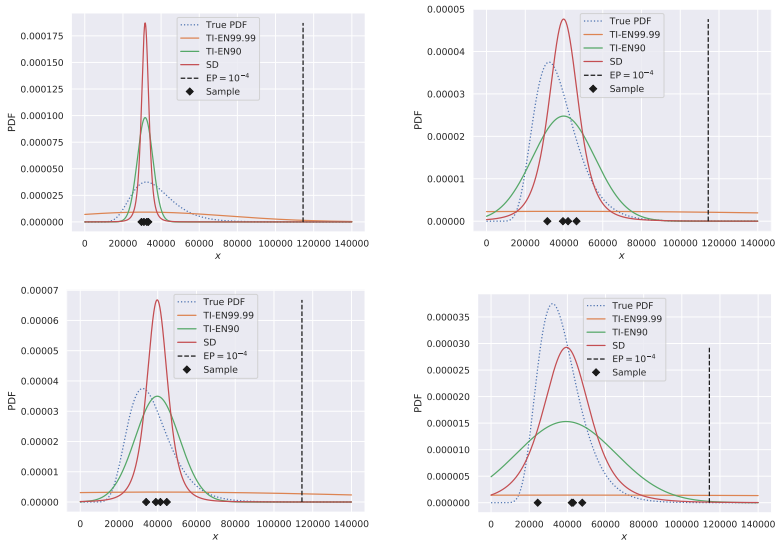


Figure 14: Four examples of the resulting distributions from TI-EN90, TI-EN99.99, and SD methods for $N = 4$ EP=10^{-4}. The true PDF is the Log-Normal distribution.

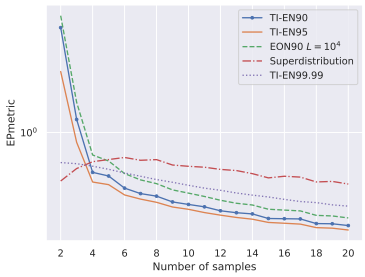
In fact, SD has usefully high reliability of >80% for up to $N = 4$, 5, 6 respectively for EP= $10^3, 10^4, 10^5$. For this regime of $> 80\%$ reliability, SD is the most efficient method in terms of the number of samples involved and the combined reliability and accuracy performance per the EP and EP10X metrics. SD has the lowest metric values overall, e.g. EP10X minima at $N = 3$, 4, 5 respectively for EP= 10^3, 10^4, 10^5. The corresponding reliabilities are all $\geq 95\%$. This is a win-win situation because of the very high reliability and the low sampling cost involved ($N = 3$, 4, 5). For EP= 10^{-3} the TI-EN99.99 method has a EP10X metric at $N = 18$ - 20 samples that reaches a minimum almost as low as the SD minimum. But even if the combined reliability and accuracy performance optimum was as good as the SD optimum, SD would still be far preferable because much fewer (potentially very expensive) samples (3) are needed to achieve the performance optimum. For EP$\leq 10^{-2}$, the SD did not appear to have better performance metrics than

[1] ≈ 1.9 is obtained by multiplying the 90% confidence value for $N = 4$ in Table C.2 ($= 2.54$) by the approximately applicable ratio $(\sigma_{SD}/\sigma_{TI\text{-}EN90}) = 0.0105/0.0142$ obtained from Figure 6. The latter proceeds from an approximately Normal instead of Log-Normal distribution.

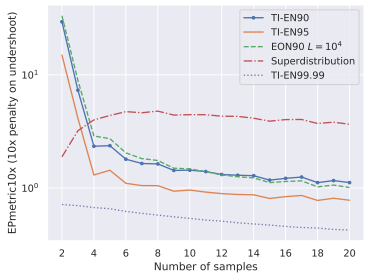
a given TI-EN. However, the best performance values had different TI-EN confidence levels for EP$\leq 10^{-2}$.

Another interesting differentiation is that for any number of samples between $N = 2$ to 20 studied (and if you ignore the EP$= 10^{-1}$ results), the reliability of SD increases as the true EP becomes smaller. The other UQ methods' reliabilities decrease as the true EP becomes smaller. These dynamics are important to note because of implications on choice of method if the order of magnitude of the EP to be estimated is approximately known and is outside the range characterized here of 10^{-5} to 10^{-2}.

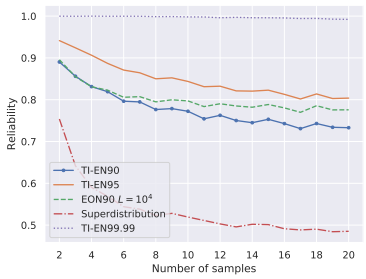
Figure 16 shows distributions of EP estimation errors for 10,000 trials of each method with $N = 4$, 10, 20 random samples per trial. All extremely poor underestimates with errors where $\Delta\log(EP)$ is less than -10 were pooled collectively to be equal to -10. This results in a large bump occurring at -10 in the figure. Unlike with the Standard Normal distribution, the number of these extremely poor estimates increased as N increases. It appears in all of these cases that the SD has the most accurate PDF, being the closest to $\Delta\log = 0$ while having the fewest number of extremely poor underestimates. Note than in cases where $N = 20$, it appears that the TI-EN99.99 was the only method that was conservative at providing such estimates, but was generally at least one order of magnitude off from the true EP.



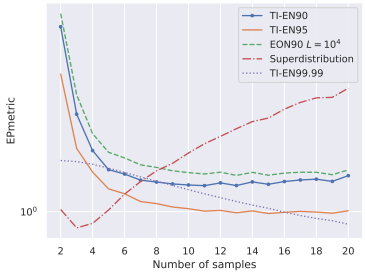
EPmetric, EP= 10^{-1}



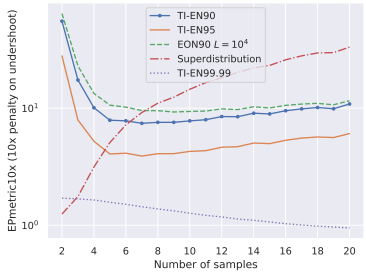
EPmetric10x, EP= 10^{-1}



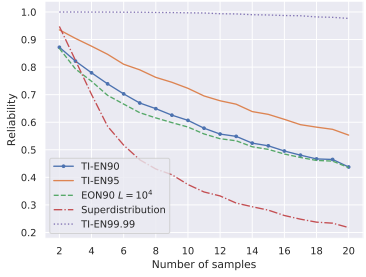
Reliability, EP= 10^{-1}



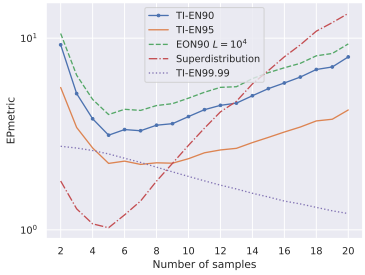
EPmetric, EP= 10^{-2}



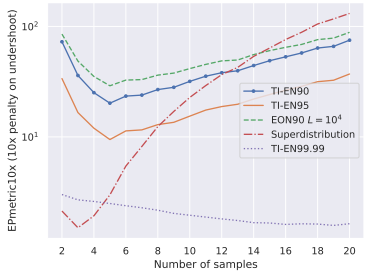
EPmetric10x, EP= 10^{-2}



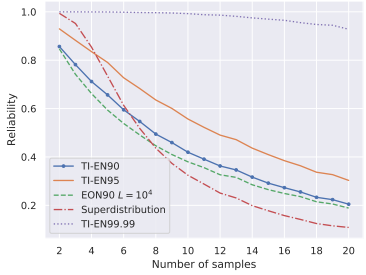
Reliability, EP= 10^{-2}



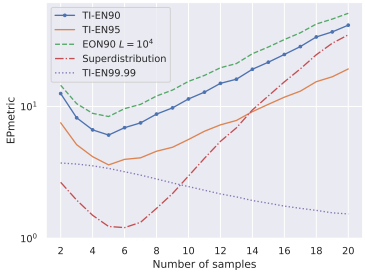
EPmetric, EP= 10^{-3}



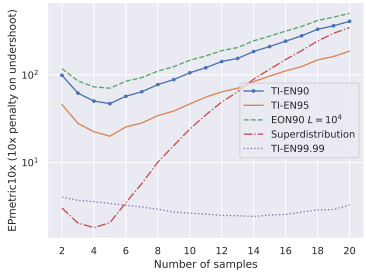
EPmetric10x, EP= 10^{-3}



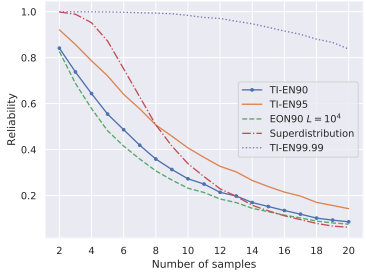
Reliability, EP= 10^{-3}



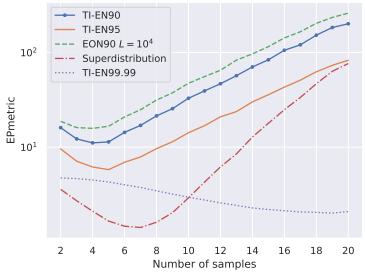
EPmetric, EP= 10^{-4}



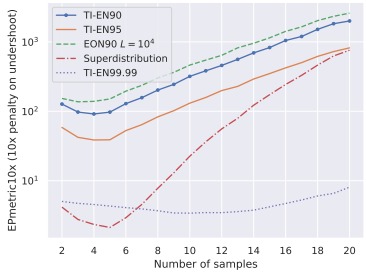
EPmetric10x, EP= 10^{-4}



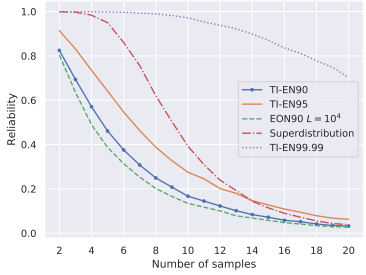
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

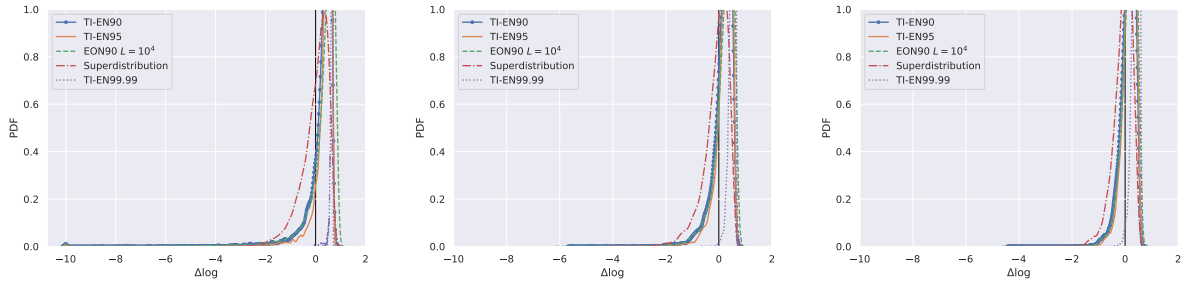


EPmetric10x, EP= 10^{-5}

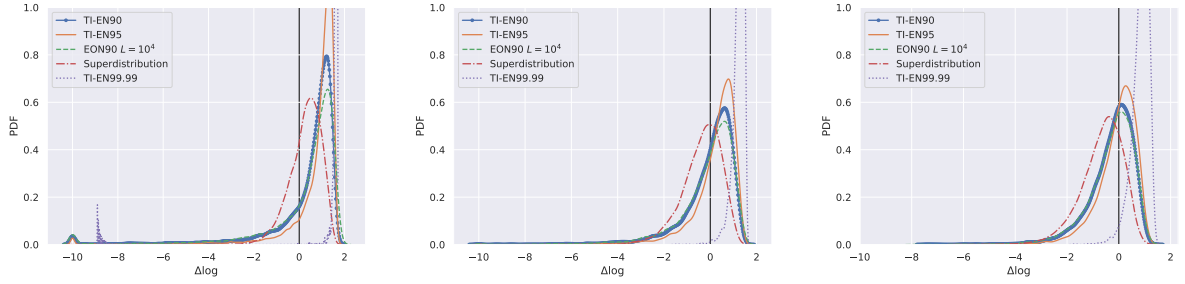


Reliability, EP= 10^{-5}

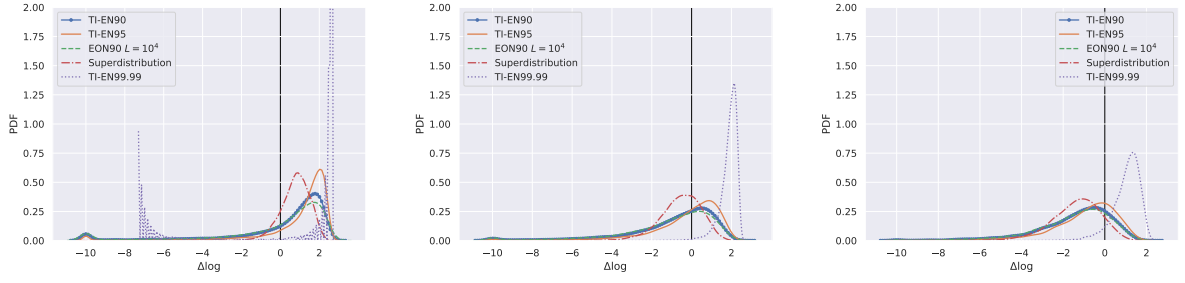
Figure 15: Results for the Log-Normal distribution.



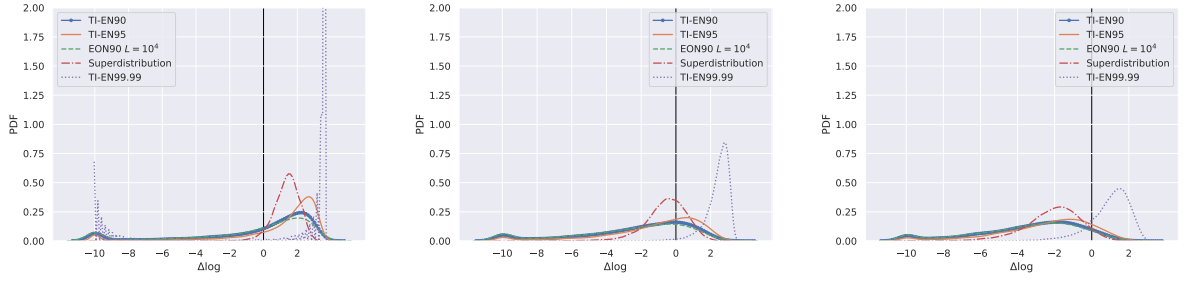
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



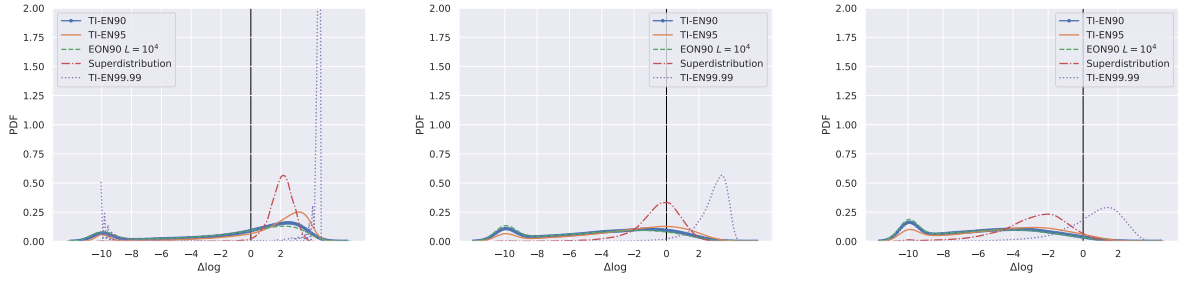
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 16: Distribution of results for the Log-Normal distribution.

3.3 Student's t and Eight Other-distributions with performance characteristics like Log-Normal

Figure 17 shows the 5 degree-of-freedom Student's t (5 d-o-f t) distribution previously studied in [1]. The performance metrics and reliability of each method at conservatively predicting EPs from sparse samples of the t-distribution are shown in Figure 18. Figure 19 shows distributions of EP estimation errors for 10,000 trials of each method with $N = 4$, 10, 20 random samples per trial. The SD was the most accurate method, having the lowest EPmetric in each true exceedance probability case ranging from EP$= 10^{-1}$ to EP$= 10^{-5}$.

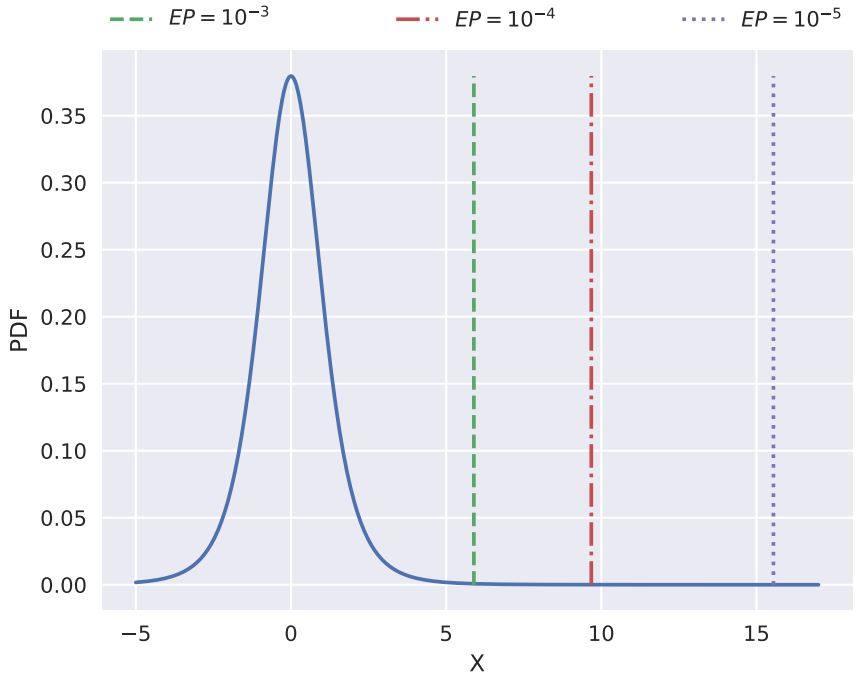


Figure 17: Student's t-distribution.

The 5 d-o-f t and the eight other distributions all yield similar performance characteristics as the Log-Normal distribution when the sparse-sample methods are applied. The eight other distributions are listed after the t-distribution in Table 1 of section 3.9, and Appendix H contains their corresponding performance metrics. SD and TI-EN/99.99 reliabilities start at a highest possible value of 100% at $N = 2$ samples for EP levels $\leq 10^{-3}$ and all 10 PDFs. The other methods start with reliabilities that are significantly less than 1 at $N = 2$ samples. The $N = 4$ plots in Figure 19 reflect SD's high reliability as TI-EN99.99, while not having the excessive conservatism of TI-EN/99.99. The SD method had the lowest EPmetric values for $N = 2$ and all EP levels, implying that with 2 samples the SD had the best combination of accuracy + reliability. SD's superior performance extends to several more samples than 2, depending on EP magnitude and PDFs type as examined next.

Reliability of all methods decreases as the number of samples increases, falling more quickly for smaller EP magnitudes than for larger magnitudes (for all 10 PDFs). TI-EN99.99 reliability

falls much more slowly than for any of the other methods, so retains the highest reliability over the full range $N = 2$ to 20 samples, for all EP levels and all 10 PDFs. SD reliability always decreases the quickest of any of the other methods. Reliabilities of the EON90, TI-EN90, and TI-EN95 methods start significantly lower than for SD and TI-EN99.99 at $N = 2$ samples, but decrease at a slower rate than for SD. This often leads (especially for 10^{-3} and 10^{-4} EP levels, for all PDFs) to crossing over to higher reliabilities than SD at larger numbers of samples (≈ 4 to 12 depending on PDF type). This contributes to EP and EP10X combined performance metric values that are usually lower/better than the SD values at higher numbers of samples. However, TI-EN99.99 metric values are lowest/best of any methods at high numbers of samples. This is largely because TI-EN99.99 has substantially higher reliability than any of the other methods at larger numbers of samples (next paragraph), for all PDF types and EP levels.

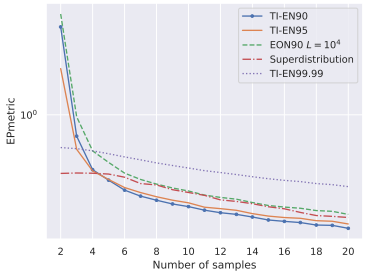
For all 10 distributions: for $N = 2$ to 20 and all EP magnitudes, the EON90 method universally had higher/worse EP and EP10X metric values and lower reliability than the TI-EN90 method, which universally performed worse than the TI-EN/95 method. For all 10 distributions and both EP and EP10X metrics, the best performing method was SD or TI-EN99.99, depending on the number of samples. SD dominated for smaller numbers of samples (typically for less than 4 to 8 samples depending on the EP level and the EP or EP10X metric type), while TI-EN99.99 always overtook SD to have the lowest/best metric values of all methods at higher numbers of samples, typically > 5 to 9. Nonetheless, SD always (for all 10 PDFs) had the absolute minimum/best EP and EP10X values of any method over the full range of samples $N = 2$ to 20 (except in a few possible instances too close to call by eye, so any SD non-dominance is insignificant in these few possible cases). Furthermore, in these possible cases the TI-EN99.99 minima occur at much higher numbers of samples; considerably more expensive for the $a \approx$same performance as optimum SD.

Thus, the combined reliability and accuracy performance metrics EP and EP10X indicate that SD is the optimal choice of methods for the said 10 PDFs and range of EP magnitudes and sample numbers. However, a check must be made to ensure that SD reliability levels are acceptably high at the optimal SD conditions (numbers of samples). Table 1 lists the optimal numbers of SD samples according to the EP and EP10X metrics, for all 16 PDFs tested in this report. As explained in section 3.9, the optimal numbers of SD samples according to the EP10X metric achieve a reasonable and practical objective of $> 80\%$ conservative tail probability estimation for EP magnitudes 10^3, 10^4, 10^5 for 15 of the 16 distributions studied, including all 10 discussed in this subsection (3.3).

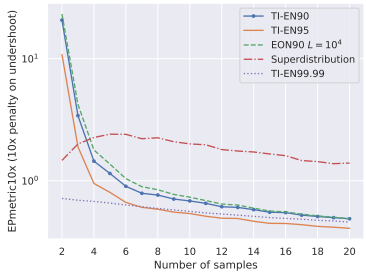
The optimal numbers of SD samples according to the EP metric are typically one to several counts bigger than those identified by the EP10X metric. This is understandable given the reduced (vs. EP10X) weighting of undershoot error magnitudes of non-conservative EP estimates. This deweighting rewards lower overall errors at the tradeoff of less penalty for non-conservative errors. Thus, the penalty for non-reliability is lessened with the EP metric (vs. the EP10X) metric, and this is consistent with the optima shift to higher numbers of samples. Reliability goes down with higher numbers of samples as has been established. Indeed, the EP metric's optimal numbers often do not meet the objective of $> 80\%$ reliability of conservative tail probability estimation.

Another thing to note is that for any number of samples between the $N = 2$ to 20 studied, the reliability of SD increases as the true EP becomes smaller. The other methods' reliabilities *decrease* as the true EP becomes smaller. These dynamics are important to be aware of because of impli-

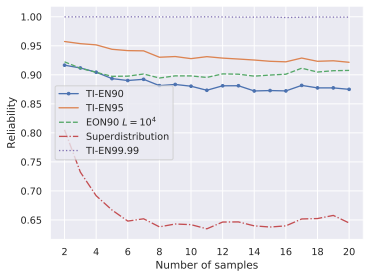
cations on choice of method if the order of magnitude of the EP to be estimated is approximately known and is outside the range characterized here of 10^{-5} to 10^{-3}.



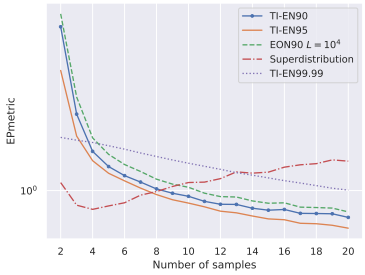
EPmetric, EP= 10^{-1}



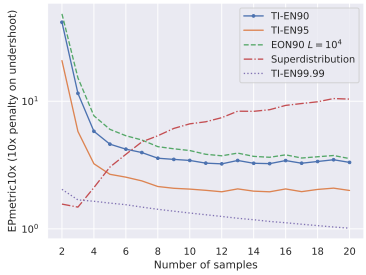
EPmetric10x, EP= 10^{-1}



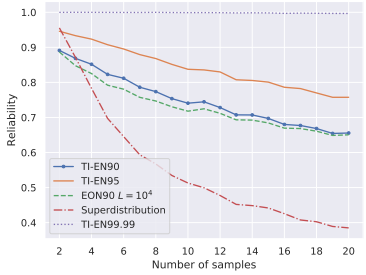
Reliability, EP= 10^{-1}



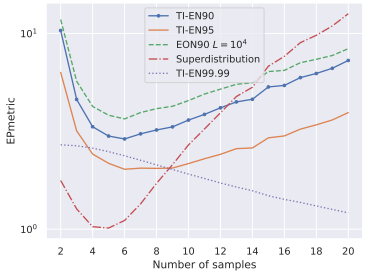
EPmetric, EP= 10^{-2}



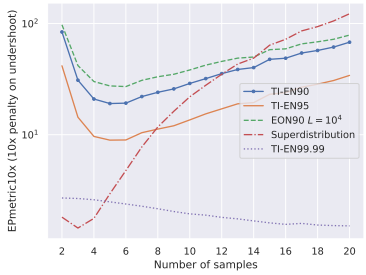
EPmetric10x, EP= 10^{-2}



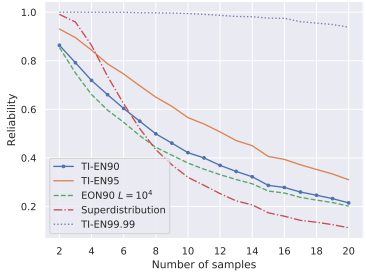
Reliability, EP= 10^{-2}



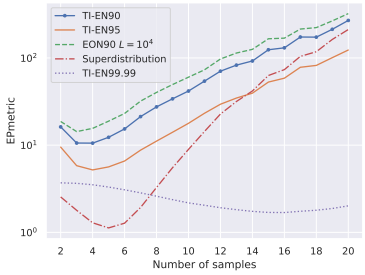
EPmetric, EP= 10^{-3}



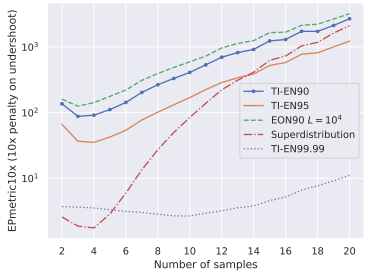
EPmetric10x, EP= 10^{-3}



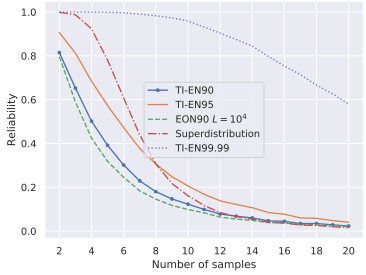
Reliability, EP= 10^{-3}



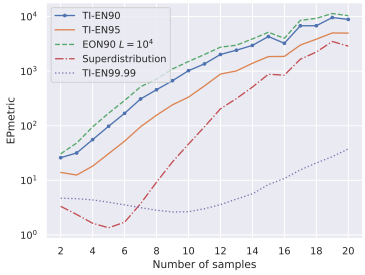
EPmetric, EP= 10^{-4}



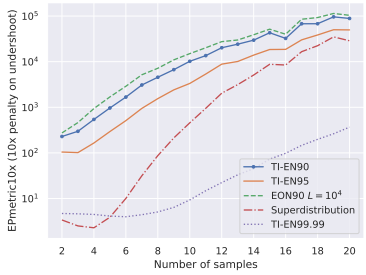
EPmetric10x, EP= 10^{-4}



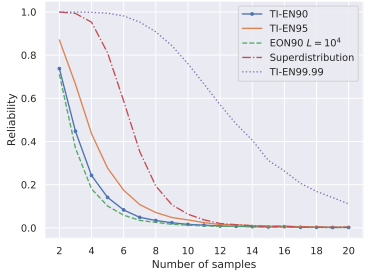
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

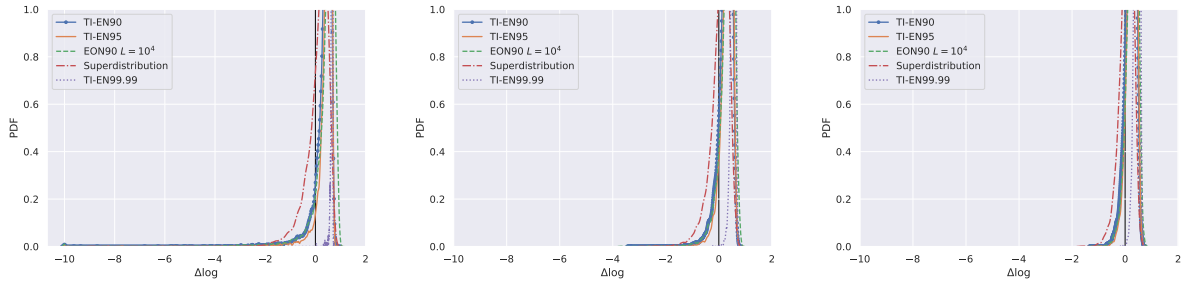


EPmetric10x, EP= 10^{-5}

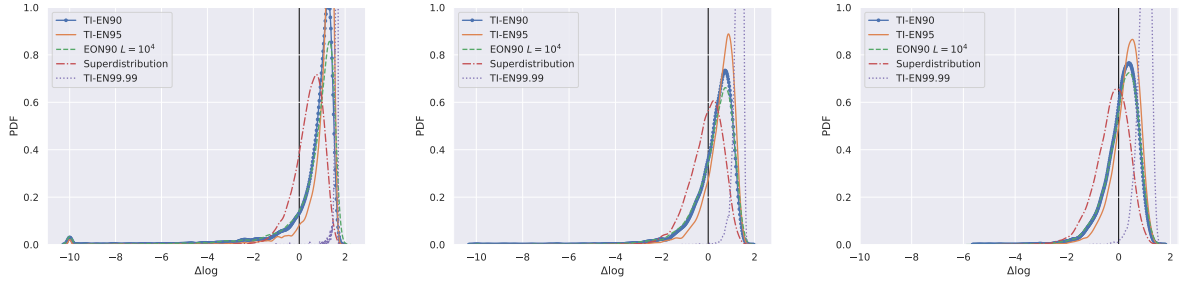


Reliability, EP= 10^{-5}

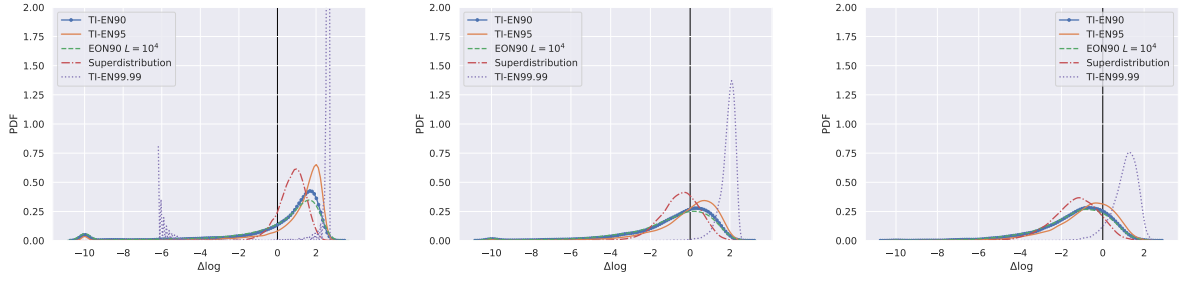
Figure 18: Results for the Student's t-distribution.



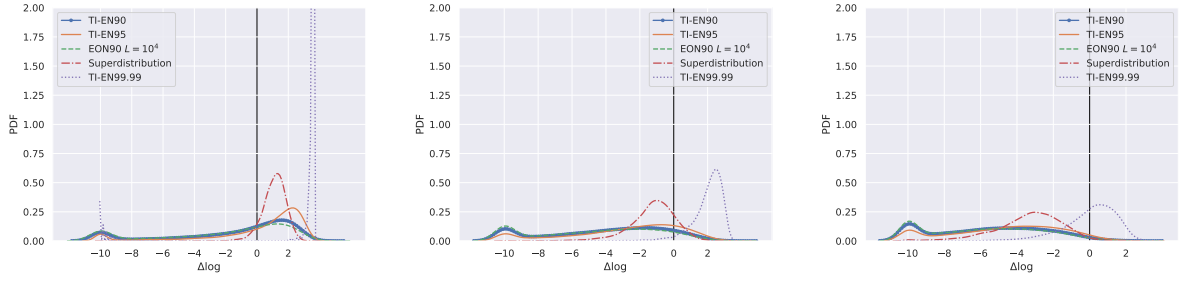
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



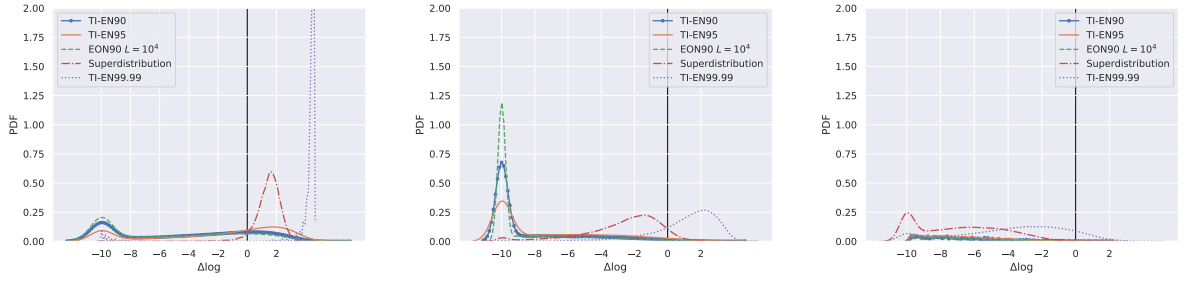
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 19: Distribution of results for the Student's t-distribution.

3.4 Tensile EQPS Weld Max Global 1.0

Figure 20 shows the Tensile Tensile EQPS Weld Max Global 1.0 empirical distribution taken from [3] and studied previously for tail probabilities in [1]. As described in Appendix B, the empirical PDFs sampled in [1] and in the present study are kernel density fits to 1000 samples of each output response quantity of interest from a solid mechanics computational simulation. The distribution looks approximately like how a significantly left-skewed Normal distribution would look with rightward extended tail.

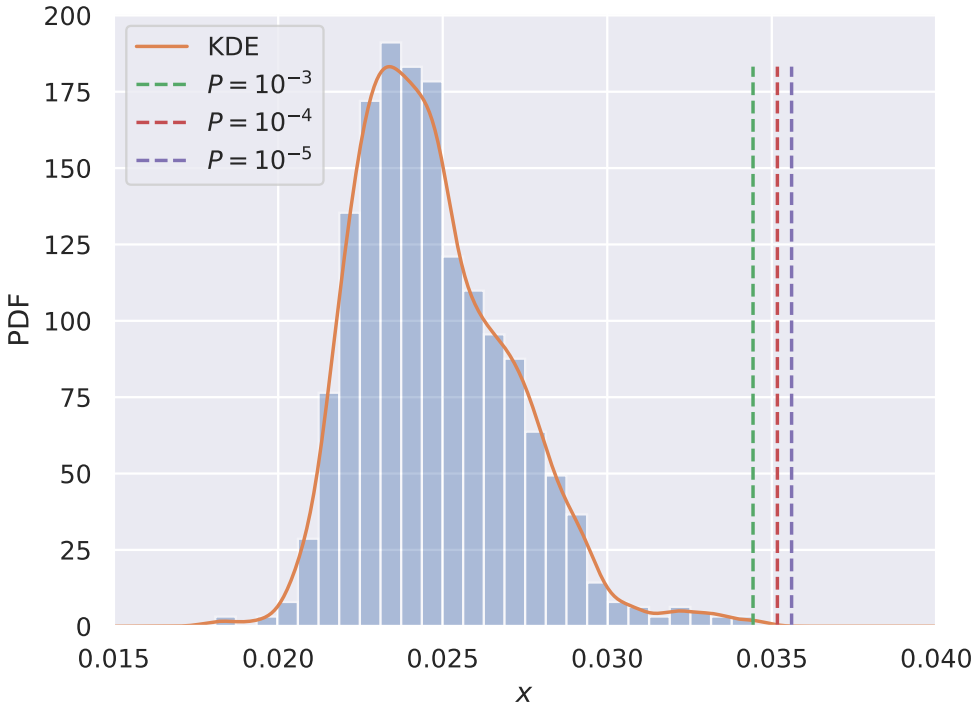


Figure 20: Histogram and KDE for Tensile EQPS Weld Max Global 1.0.

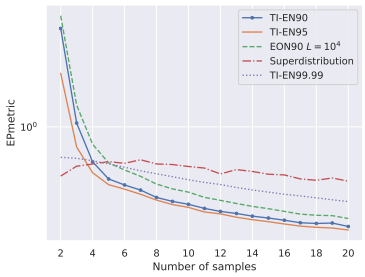
Results are shown in Figures 21 and 22. For the present distribution (Tensile EQPS Weld Max Global 1.0), the reliability curves for all methods are in-character with the reliability curves for the 10 PDFs, except for the TI-EN99.99 reliability curves. These do not slope downward essentially monotonically with increasing numbers of samples like occurs for the 10 PDFs discussed in section 3.3, but instead remain relatively flat-resembling TI-EN99.99 reliability for the Normal distribution.

For some of the metric performance curves, the present distribution yields characteristically different results from those of the 10 PDFs. The 10 PDFs' performance curves in most instances decline from $N = 2$ values as samples are added (indicating improving accuracy and conservatism performance) until a minimum point is reached, beyond which metric values rise (performance worsens) with further samples. Such trough-shaped performance curves, with minima occurring somewhere in the range between 3 to 20 samples, occurs universally (for all EP magnitudes and EP and EP10X metrics) for EON90, TI-EN90, and TI-EN95 methods. The trough behavior occurs

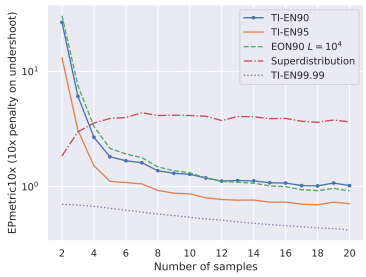
for the SD method in most instances, with the major exception at EP= 10^{-1} which exhibited the opposite performance behavior described by having a minima occur at $N = 2$. The exceptions exhibit increasing metric values (worsening performance) immediately from $N = 2$ values as samples are added, with no initial decline that would give trough shaped performance. Trough shaped performance occurs for the majority of TI-EN99.99 results for the 10 PDFs.

The performance metric curves for the present distribution are only trough-shaped for about half the cases (considering all methods, the three EP magnitudes, and the two performance metrics). Specifically, the SD and TI-EN99.99 results are mostly in-character with the previously discussed performance curves of the other PDFs. One exception is that there are cases where the SD curve remained lower (indicating better performance) than the TI-EN99.99 over the full range of samples from $N = 2$ to 20 for EP= 10^{-4} and EP= 10^{-5}. Another exception is that the SD's EP metric curve for EP= 10^{-5} continually declines (performance continually improves) from the $N = 2$ value as samples are added, instead of exhibiting trough shaped initial improvement but then deterioration of performance. A third exception is that TI-EN99.99 performance with the EP metric does not dominate EON90, TI-EN90, and TI-EN95 performance over the full range of samples from $N = 2$ to 20 (whereas TI-EN/99.99 dominance only occurred for the other PDFs at some very minor range of samples). The three behavioral exceptions described here resemble the behaviors of SD and TI-EN performance on Normal distributions better than the other distributions. This is perhaps not surprising, as the empirical Tensile EQPS Weld Max Global 1.0 distribution appeared like a left-skewed Normal distribution.

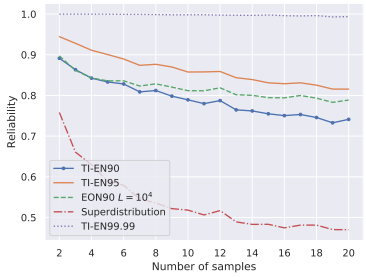
The EON90, TI-EN90, and TI-EN95 methods' performance curve shapes for both EP and EP10X metrics are trough-shaped for EP= 10^{-3} in-character with the 10 PDFs' results. However, for EP = 10^{-4} and 10^{-5} the EP and EP10X metric curves change to those resembling shapes for the Normal distribution: either always declining (performance improves) as samples are added, or initially declining then plateauing to remain essentially flat as samples are added. For this distribution, the performance of SD dominates that of the other methods, like it does for the Normal and other 10 distributions examined thus far.



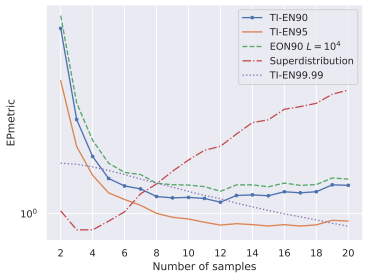
EPmetric, EP= 10^{-1}



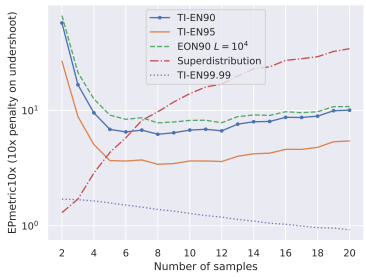
EPmetric10x, EP= 10^{-1}



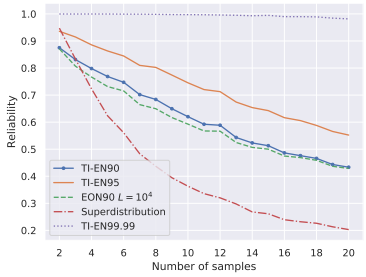
Reliability, EP= 10^{-1}



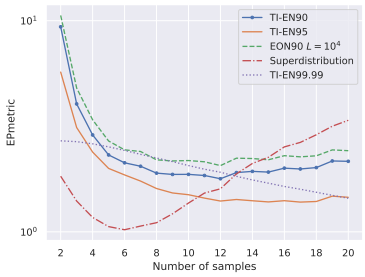
EPmetric, EP= 10^{-2}



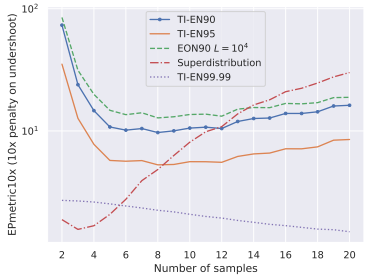
EPmetric10x, EP= 10^{-2}



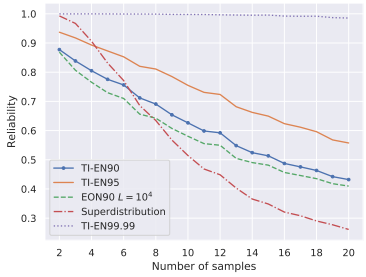
Reliability, EP= 10^{-2}



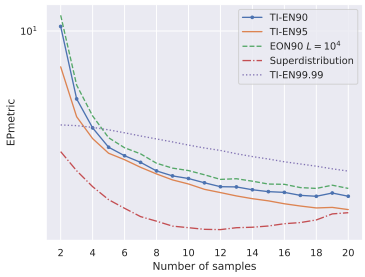
EPmetric, EP= 10^{-3}



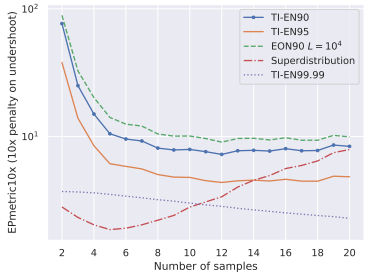
EPmetric10x, EP= 10^{-3}



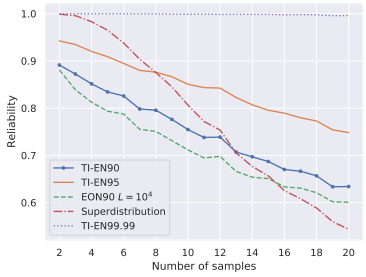
Reliability, EP= 10^{-3}



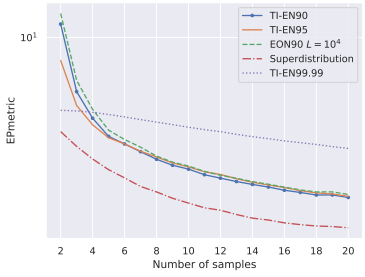
EPmetric, EP= 10^{-4}



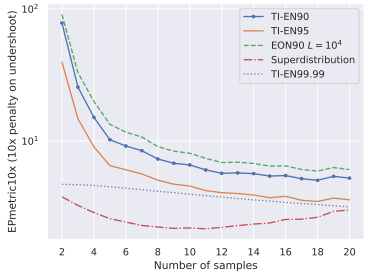
EPmetric10x, EP= 10^{-4}



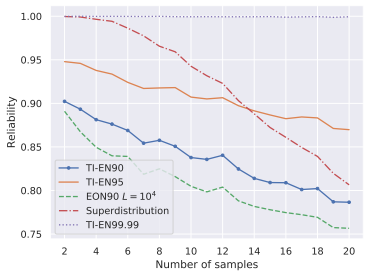
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

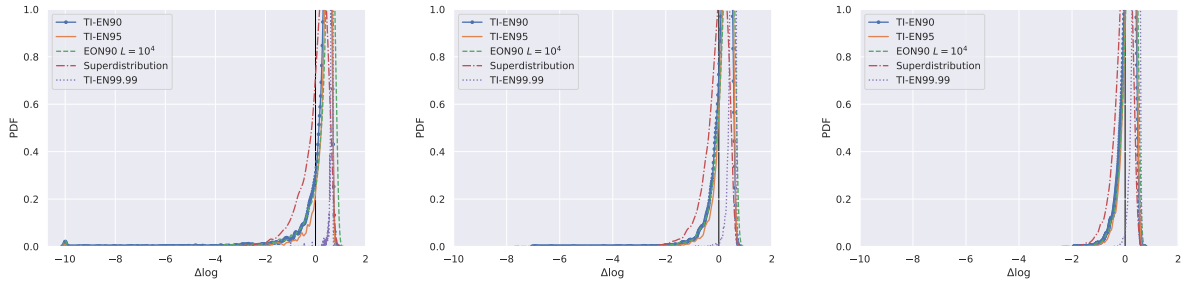


EPmetric10x, EP= 10^{-5}

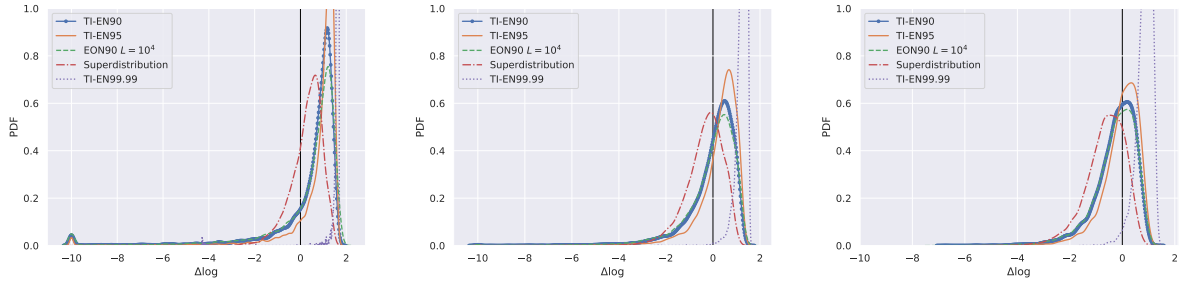


Reliability, EP= 10^{-5}

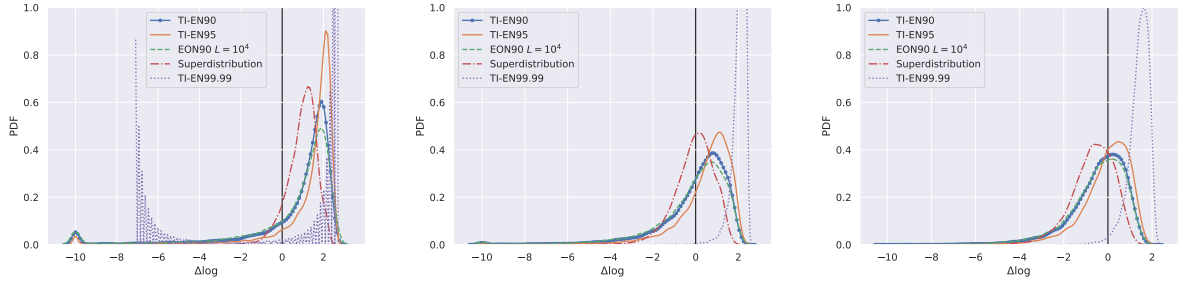
Figure 21: Results for the empirical Tensile EQPS Weld Max Global 1.0 distribution.



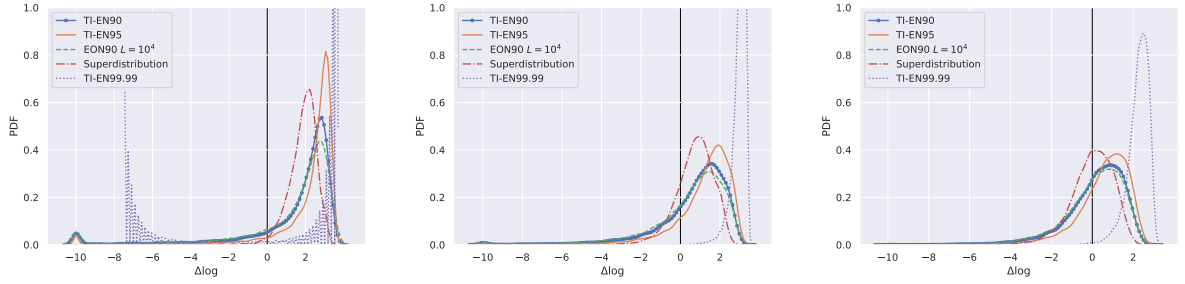
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



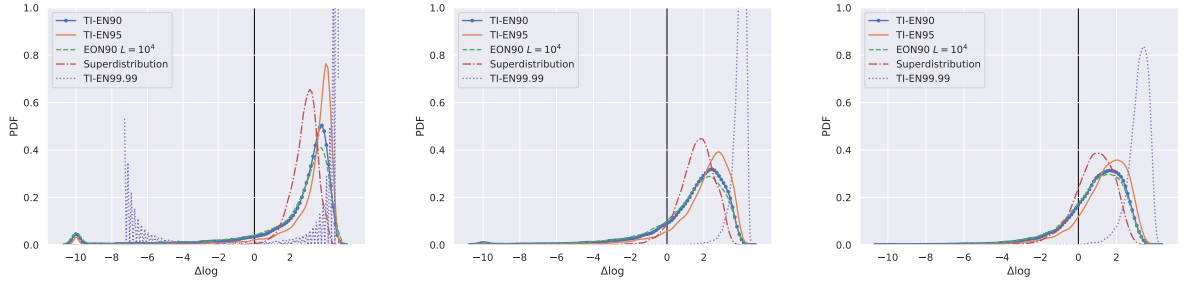
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 22: Distribution of results for the empirical Tensile EQPS Weld Max Global 1.0 distribution.

3.5 EQPS Can Top Element 0.5

Figure 23 shows the EQPS Can Top Element 0.5 empirical distribution from [3] and [1]. The distribution is approximately symmetric.

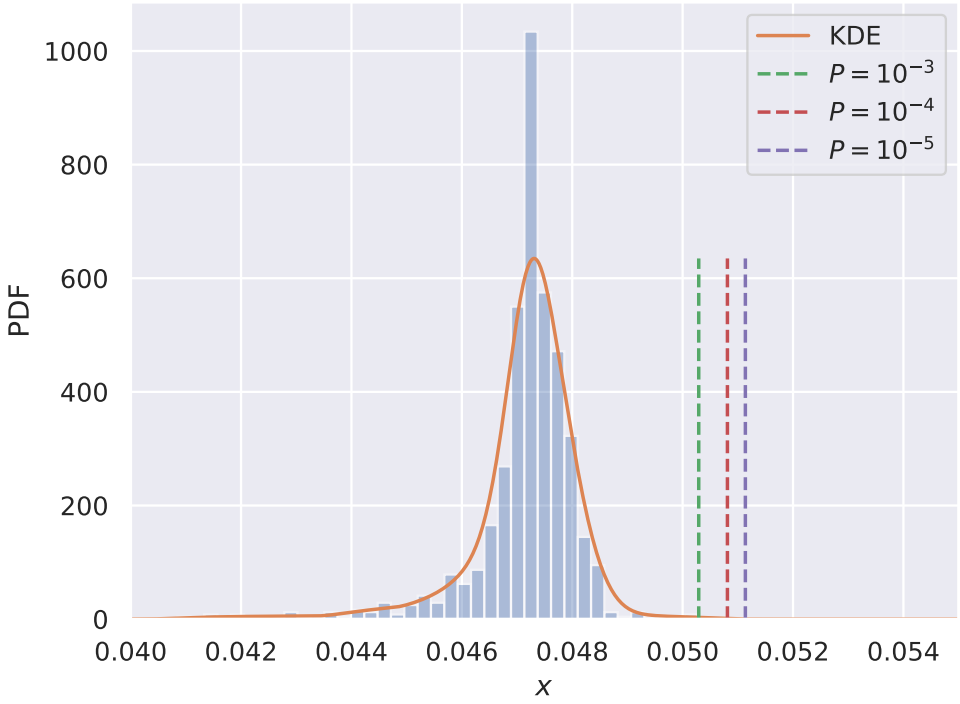


Figure 23: Histogram and KDE for EQPS Can Top Element 0.5.

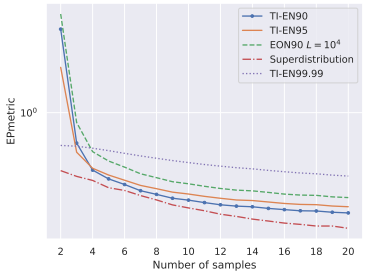
Results are shown in Figures 23 and 24. Most of the observations made in the prior section apply to the present distribution as well. The reliability curves for the SD, EON90, TI-EN90, and TI-EN95 methods for the present PDF are in-character with the other downward sloping reliability curves typical for the non-normal distributions. TI-EN99.99 reliability curves for the present PDF resemble the flat reliability curves obtained for the PDF in the immediately prior subsection and for the Normal distribution.

One departure from the prior section is that trough shaped performance curves occur in all cases for SD, whereas this was not true in one of the six cases in the prior subsection. Another departure is that the SD curves do not always remain lower/better than the TI-EN99.99 curves over the full range from $N = 2$ to 20 samples. Crossing of their performance curves occurs in three of the six cases, so SD is not completely dominant, but still has better optimum performance than TI-EN99.99 optima in all cases for all EP levels. These departures make SD and TI-EN99.99 performance curves very much in-character with their curves for the 10 PDFs.

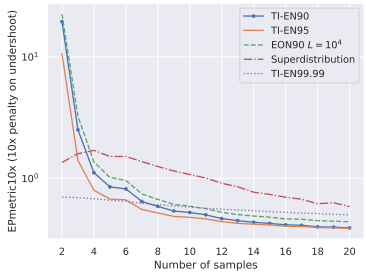
The EON90, TI-EN90, and TI-EN95 methods have performance curves that are even more characteristically different from the curves for the 10 PDFs than the prior subsection's curves are. That is, the performance curves continually fall (performance improves) with added samples all

the way to $N = 20$ in all six cases instead of three as in the prior section.

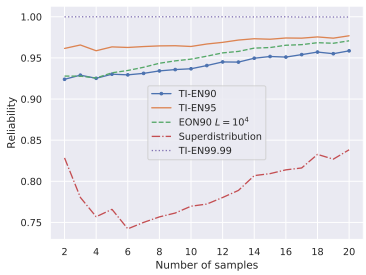
The most important overall finding to this point for the purposes of this report is that SD performance dominates the performance of the other methods for the present distribution and for the other 12 previously discussed.



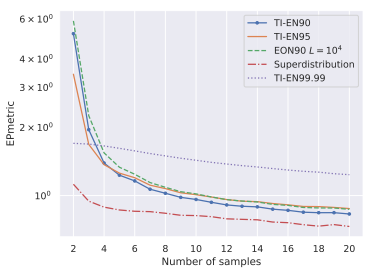
EPmetric, EP= 10^{-1}



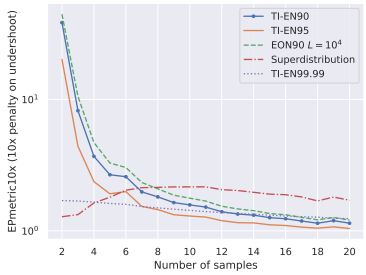
EPmetric10x, EP= 10^{-1}



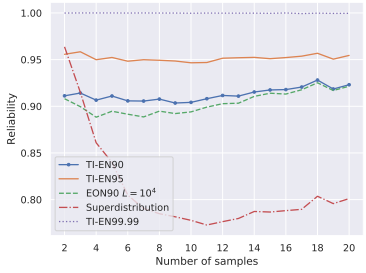
Reliability, EP= 10^{-1}



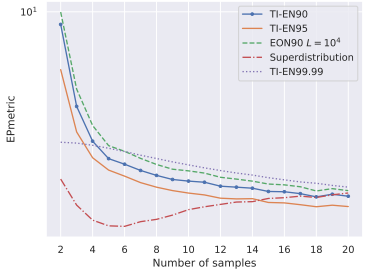
EPmetric, EP= 10^{-2}



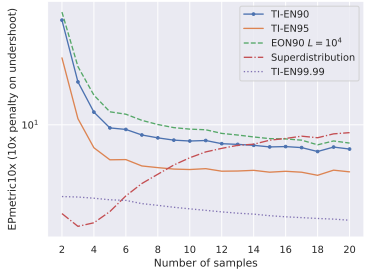
EPmetric10x, EP= 10^{-2}



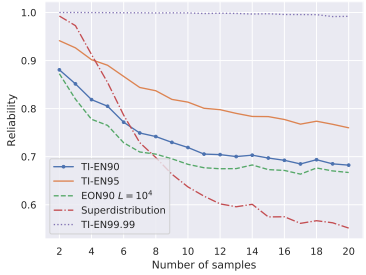
Reliability, EP= 10^{-2}



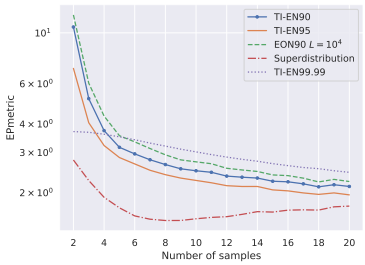
EPmetric, EP= 10^{-3}



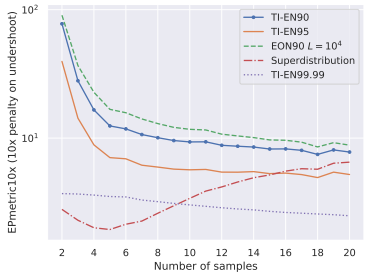
EPmetric10x, EP= 10^{-3}



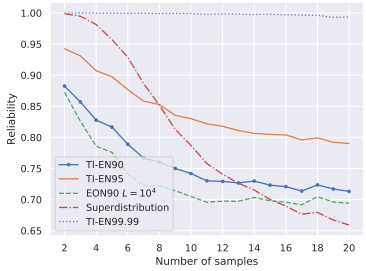
Reliability, EP= 10^{-3}



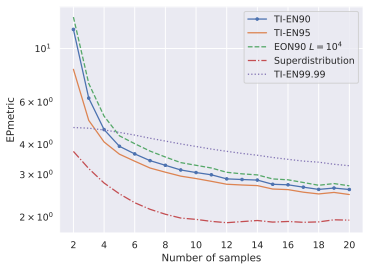
EPmetric, EP= 10^{-4}



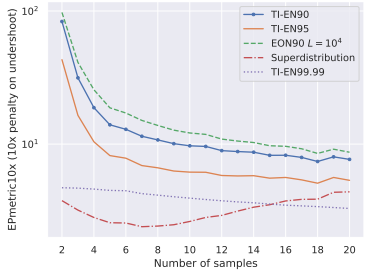
EPmetric10x, EP= 10^{-4}



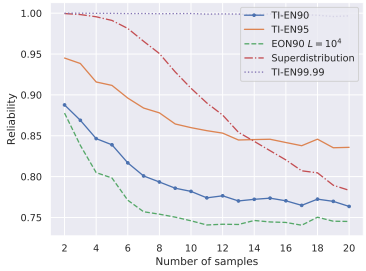
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

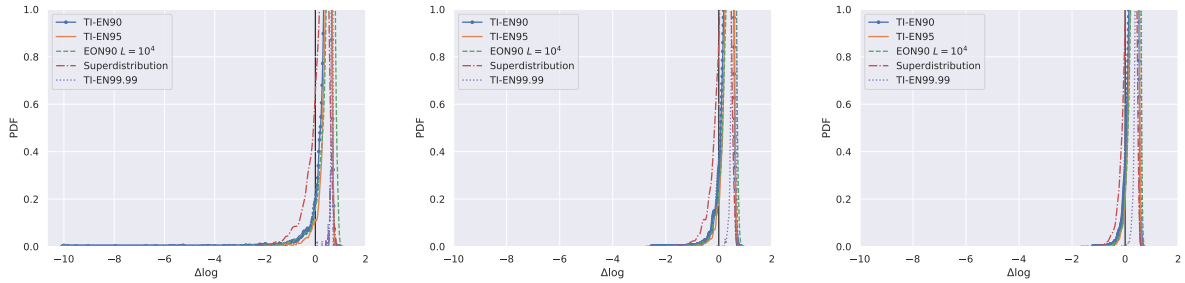


EPmetric10x, EP= 10^{-5}

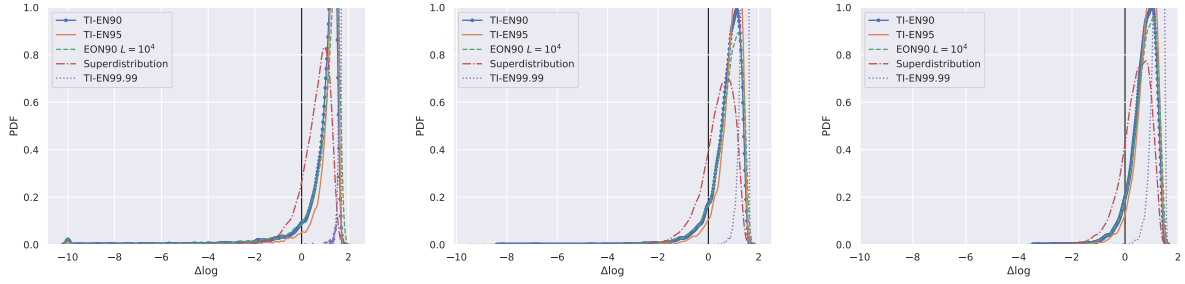


Reliability, EP= 10^{-5}

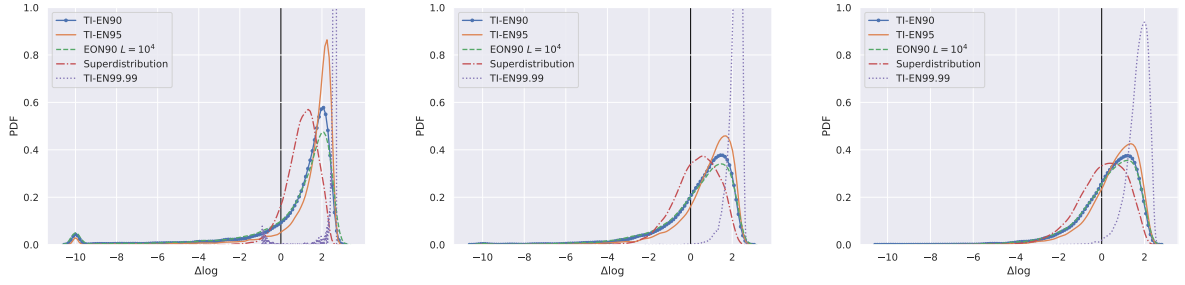
Figure 24: Results for the empirical EQPS Can Top Element 0.5 distribution.



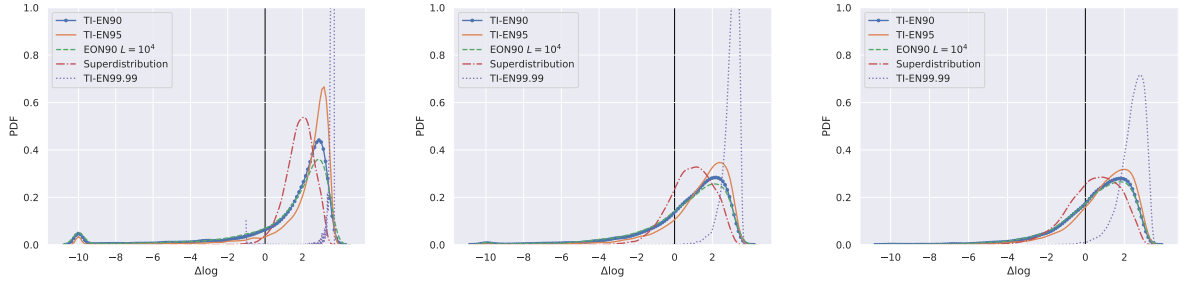
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



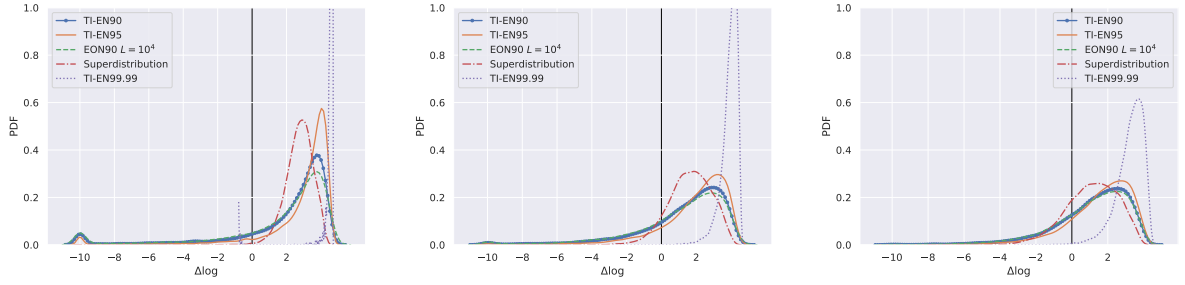
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 25: Distribution of results for the empirical EQPS Can Top Element 0.5 distribution.

3.6 Tensile EQPS Lid Buckle Element 0.25

Figure 26 shows the Tensile EQPS Lid Buckle Element 0.25 empirical distribution from [3] and [1]. It is strongly multi-modal, with eight distinct humps. This would appear to be a very challenging case for any tail-probability estimation method.

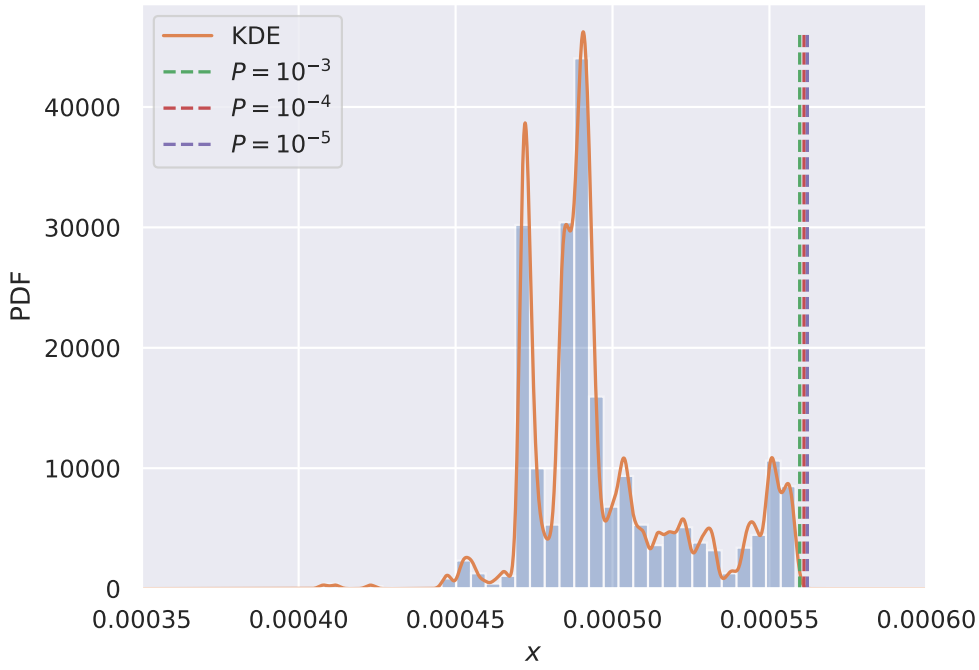


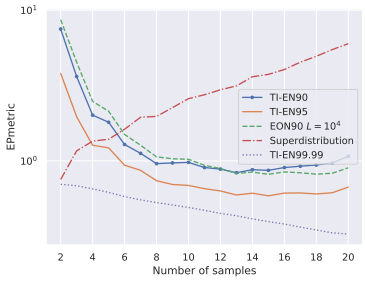
Figure 26: Histogram and KDE for Tensile EQPS Lid Buckle Element 0.25.

Results are shown in Figures 27 and 28. The EP and EP10X performance curve shapes and relative positioning for EON90, TI-EN90, TI-EN95, and TI-EN99.99 are much like for the distribution immediately prior. All the curves monotonically slope downward with increasing samples, indicating improving accuracy-conservatism performance. This is behavior that resembles that for a Normal distribution. Also reminiscent of Normal PDF performance, the SD curves all effectively slope downward (three monotonically and three that have local minima at low numbers of samples but global minima at $N = 19$ or 20). SD again dominates TI-EN99.99; over the entire range $N = 2$ to 20 for five cases, with exception to the EP$= 10^{-1}$ level in which the TI-EN99.99 dominated SD. For EP$\leq 10^{-2}$, the SD optima were superior to TI-EN99.99 performance anywhere within the range. SD dominates EON90, TI-EN90, and TI-EN95 over the entire range, usually substantially, whereas these three methods often dominate TI-EN99.99 over large portions of the range. Again, there is exception for the EP$= 10^{-1}$ level in which the TI-EN99.99 bested all other methods. The results on this empirical distribution appear to be similar for the Normal distribution and the two empirical PDFs in the two immediately prior subsections.

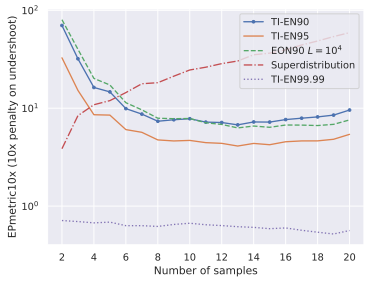
What most stands out as different with the present distribution are the reliability curve behaviors. The TI-EN99.99 reliability curves resemble the flat reliability curves obtained for the Normal distribution and the two empirical PDFs in the two immediately prior subsections. However, the

reliability curves for the SD, EON90, TI-EN90, and TI-EN95 methods are like nothing yet encountered with the other 13 PDFs examined so far. These methods' reliability curves for the present PDF either increase noisily from $N = 2$ values as samples are added, or decline at first and then rise. In either case, the increasing instead of decreasing reliability at the larger numbers of samples is an unfamiliar (to this point) yet beneficial trend. The result is that reliabilities are relatively high for all methods over the entire range $N = 2$ to 20; above 80%, 87%, and 89% respectively for EPs 10^{-3}, 10^{-4}, 10^{-5}. However, reliabilities were significantly lower for EP= 10^{-1} in which methods other than TI-EN99.99 would drop to less than 50% at 20 samples.

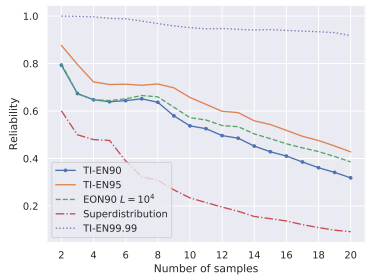
Again, however, the most important overall finding for the present distribution is that SD clearly dominates the other methods (like for the other 13 distributions previously discussed).



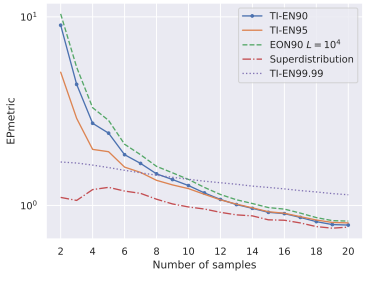
EPmetric, EP= 10^{-1}



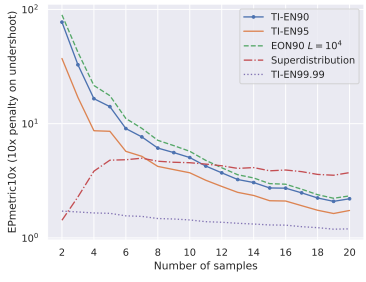
EPmetric10x, EP= 10^{-1}



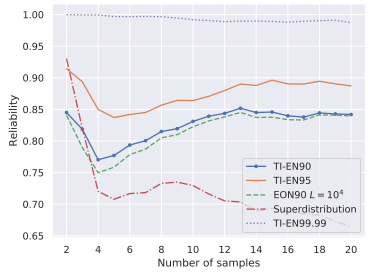
Reliability, EP= 10^{-1}



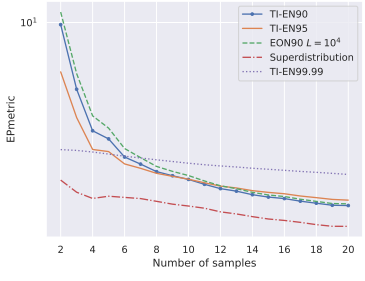
EPmetric, EP= 10^{-2}



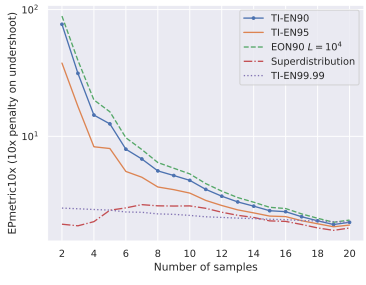
EPmetric10x, EP= 10^{-2}



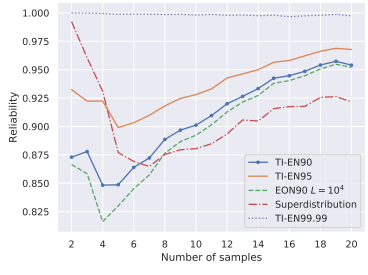
Reliability, EP= 10^{-2}



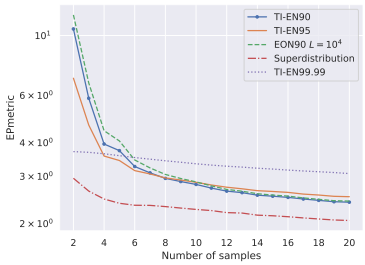
EPmetric, EP= 10^{-3}



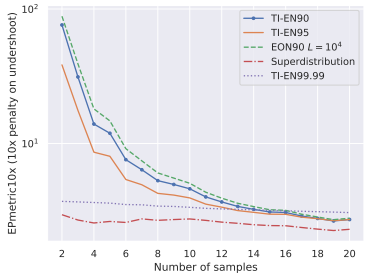
EPmetric10x, EP= 10^{-3}



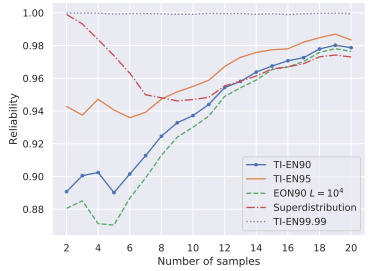
Reliability, EP= 10^{-3}



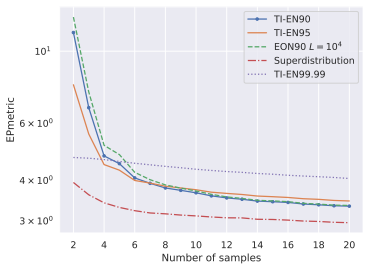
EPmetric, EP= 10^{-4}



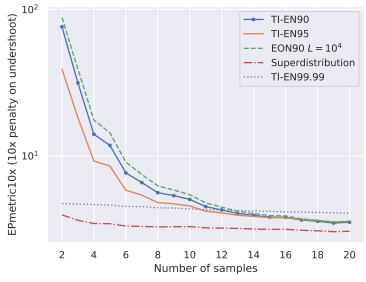
EPmetric10x, EP= 10^{-4}



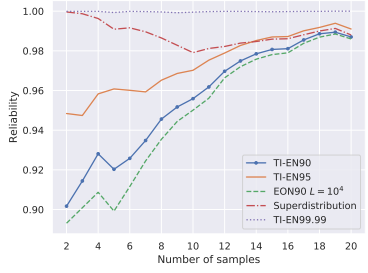
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

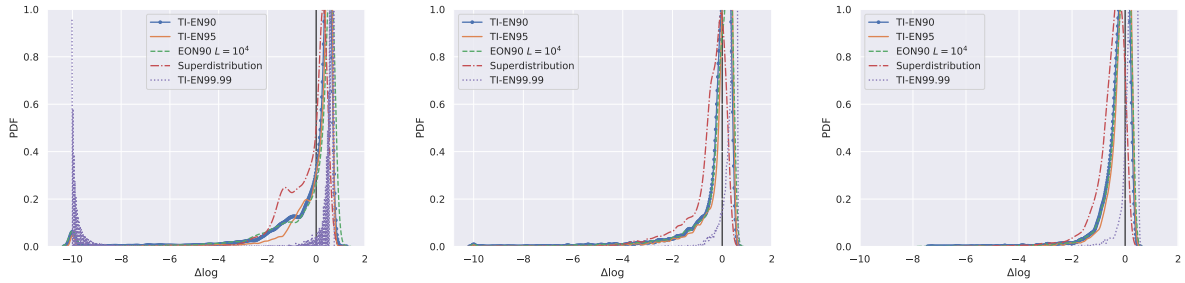


EPmetric10x, EP= 10^{-5}

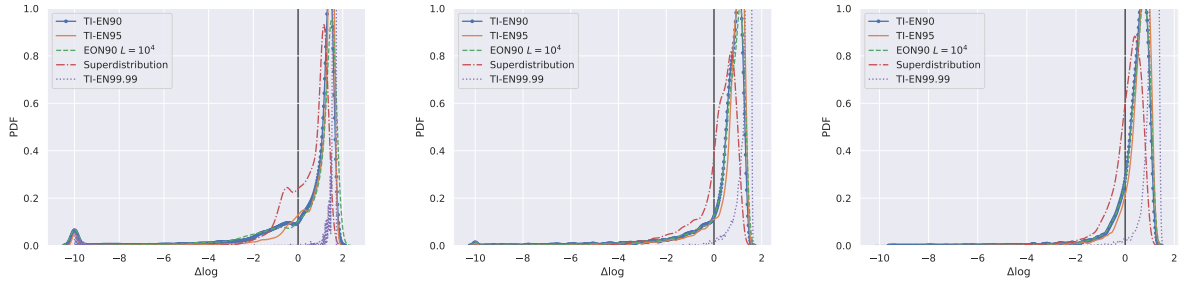


Reliability, EP= 10^{-5}

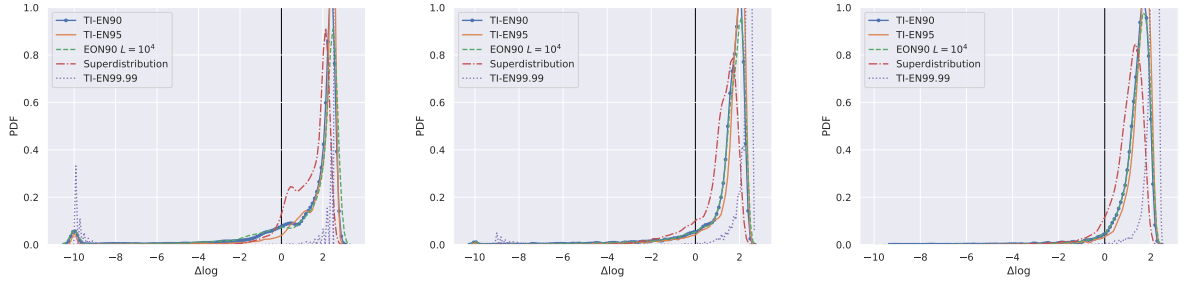
Figure 27: Results for the empirical Tensile EQPS Lid Buckle Element 0.25 distribution.



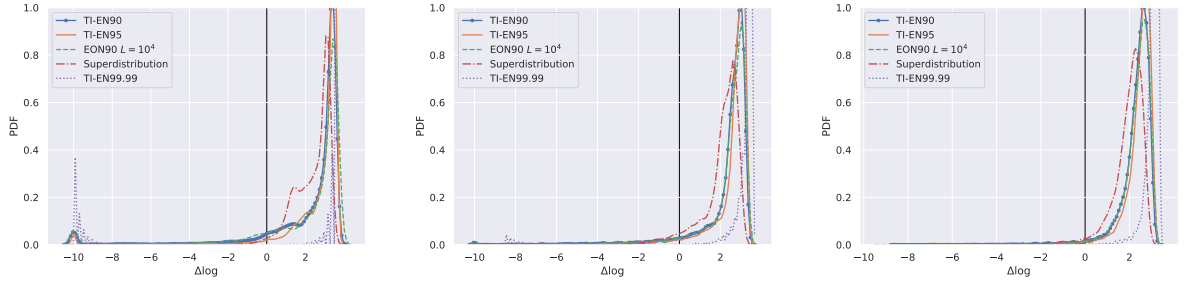
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



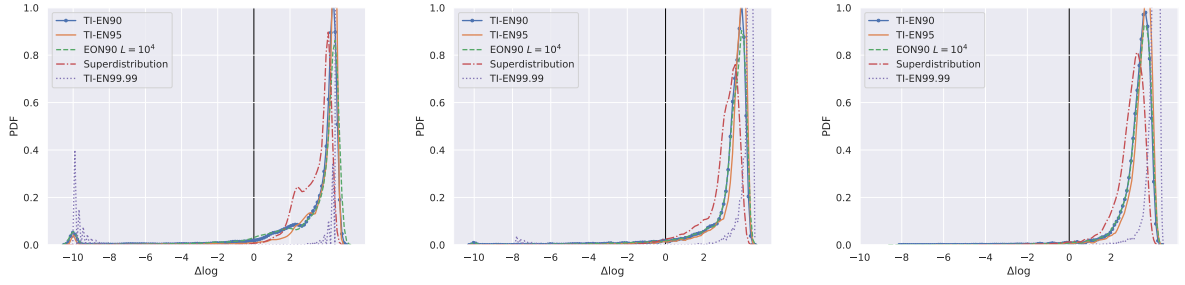
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 28: Distribution of results for the empirical Tensile EQPS Lid Buckle Element 0.25 distribution.

3.7 Bi-modal Log-Gamma Normal distribution

This distribution is assembled with a left Log-Gamma and a right Normal distribution. Figure 13 shows the distribution, and further information about the distribution is available in Appendix A.6. It is strongly multi-modal, with two distinct humps and a much longer left tail off the left wider-shorter hump than the quickly dropping right tail of the right thinner-higher hump. Again, this would appear to be a very challenging PDF for any tail-probability estimation method, but it turned out to be the easiest for right-tail probabilities. The story is very different for left-tail probabilities of 10^{-3} through 10^{-6} magnitudes for this multi-model PDF, as Section **??** of this report explores.

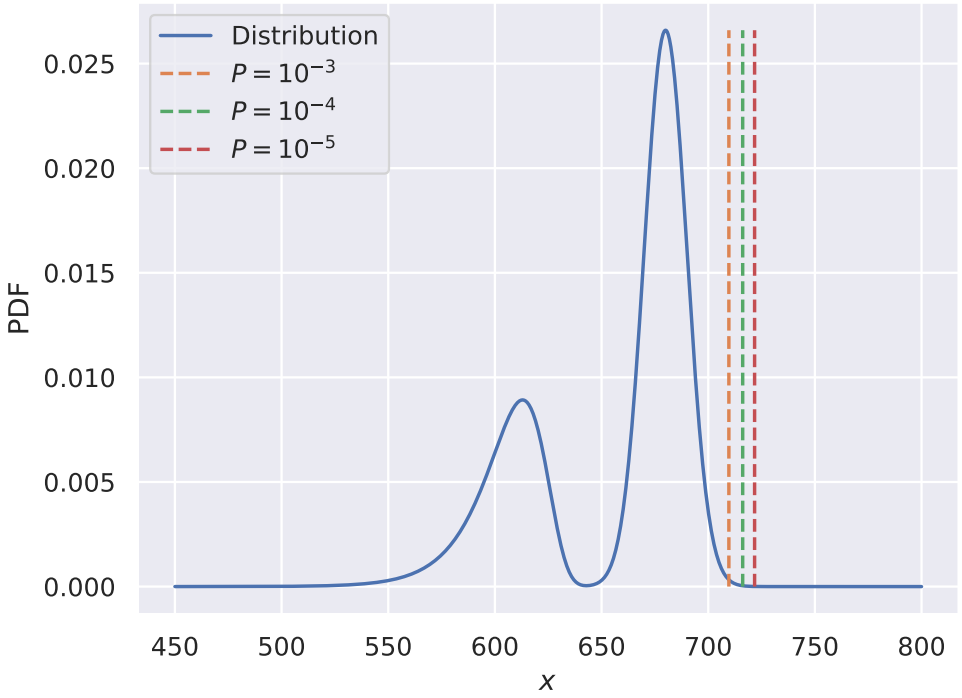
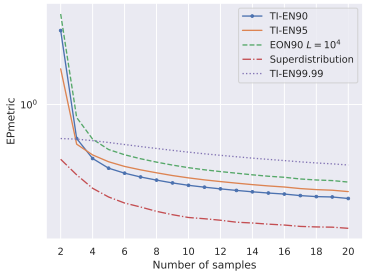


Figure 29: Bi-modal Log-Gamma Normal distribution with thresholds that have right tail EP of $P = 10^{-3}, 10^{-4}, 10^{-5}$.

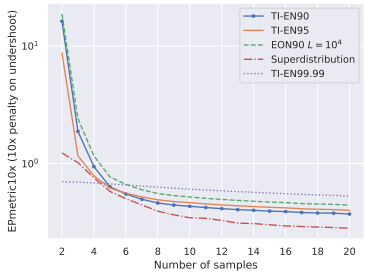
Results for the present distribution are shown in Figures 30 and 31. The reliability curves are even higher and better than for the multi-modal empirical distribution in the immediately previous subsection. All results are above 75% reliability over the entire range $N = 2$ to 20. The reliabilities were significantly higher for EP$\leq 10^{-3}$, in which the lowest reliability for any method was about 93%, and the SD reliability curves never fall below 99% over the range. The EN99.99 are even better, essentially 1 over the full range for any EP level. The other methods' curves rise quicker for the smaller EP and monotonically from their values at $N = 2$ as samples are added. For EP$\leq 10^{-3}$, all of these methods rise above 99% reliability after about $N = 6$ samples.

The EP and EP10x performance curve shapes and relative positioning are much like for the multi-modal empirical distribution in the immediately previous subsection. The performance curve shapes are even more like those for the Normal distribution. This is perhaps not surprising because

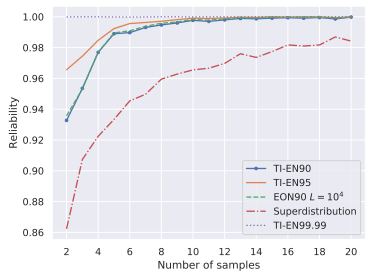
the right hump and tail of the PDF are formed from a Normal distribution. The performance curves decrease monotonically (performance improves monotonically) with added samples. The curves fall rapidly (performance improves rapidly) at very low samples, with much more moderate improvement as medium and larger numbers of samples are encountered. Again, SD dominates all other methods (and over the entire range $N = 2$ to 20 for this particular PDF like for the Normal PDF). TI-EN99.99 performs clearly worst of any of the methods at moderate and high numbers of samples.



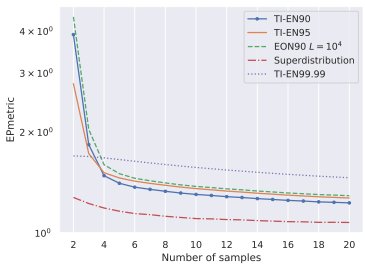
EPmetric, EP= 10^{-1}



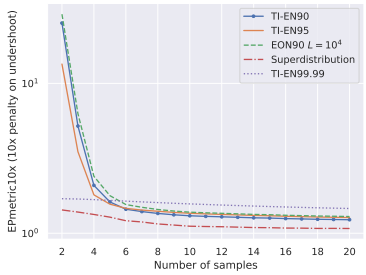
EPmetric10x, EP= 10^{-1}



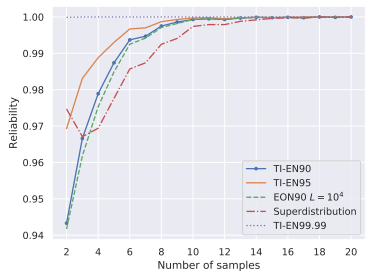
Reliability, EP= 10^{-1}



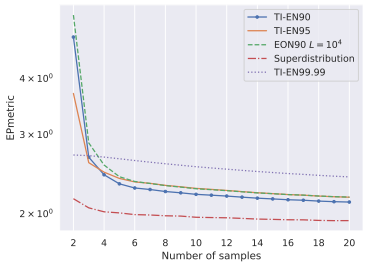
EPmetric, EP= 10^{-2}



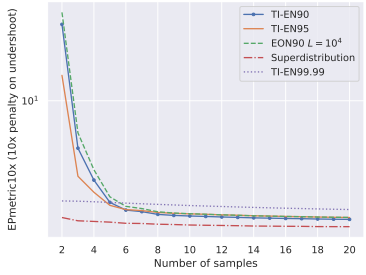
EPmetric10x, EP= 10^{-2}



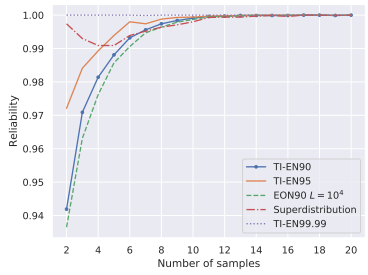
Reliability, EP= 10^{-2}



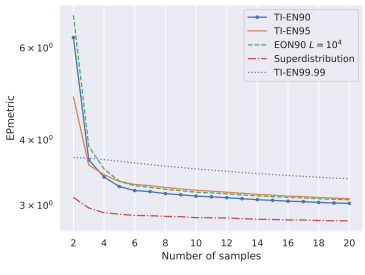
EPmetric, EP= 10^{-3}



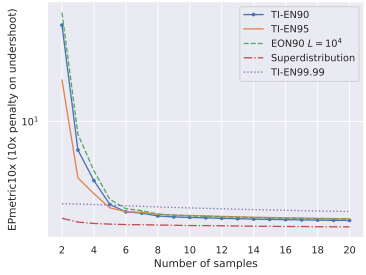
EPmetric10x, EP= 10^{-3}



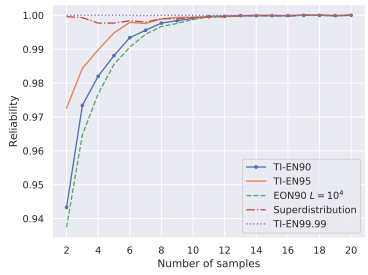
Reliability, EP= 10^{-3}



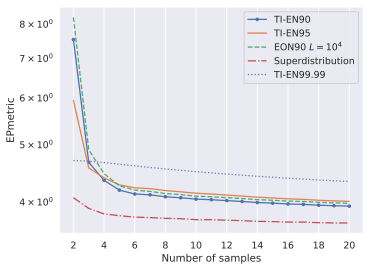
EPmetric, EP= 10^{-4}



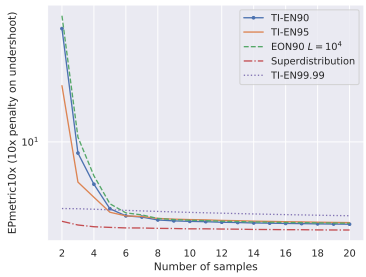
EPmetric10x, EP= 10^{-4}



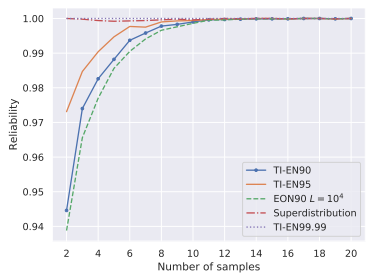
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

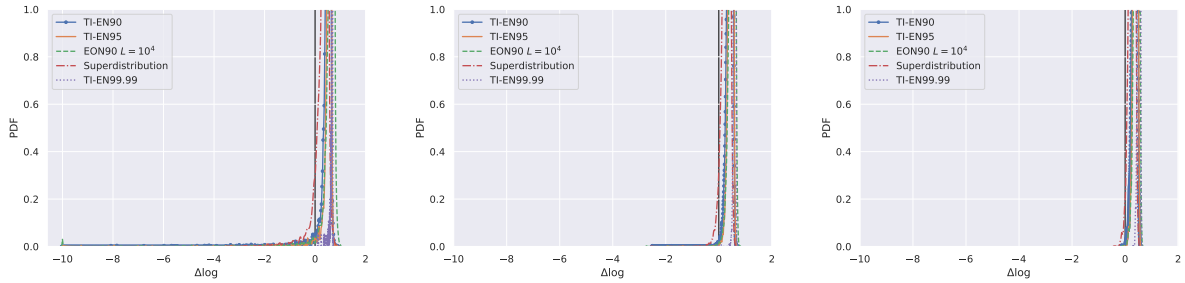


EPmetric10x, EP= 10^{-5}

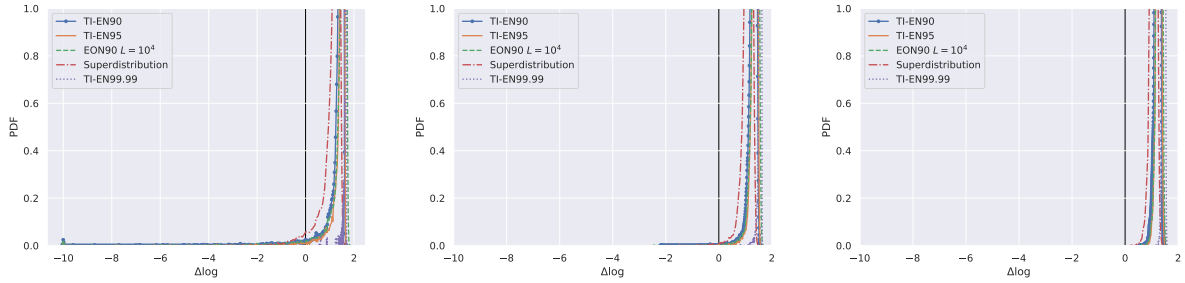


Reliability, EP= 10^{-5}

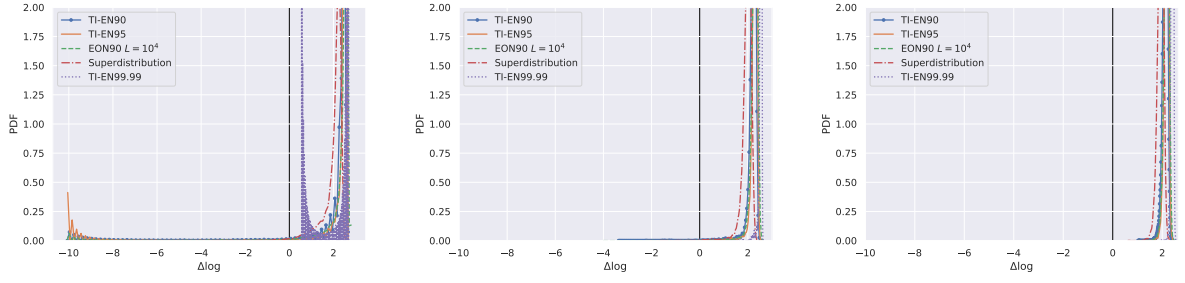
Figure 30: Results for the bi-modal Log-Gamma Normal distribution.



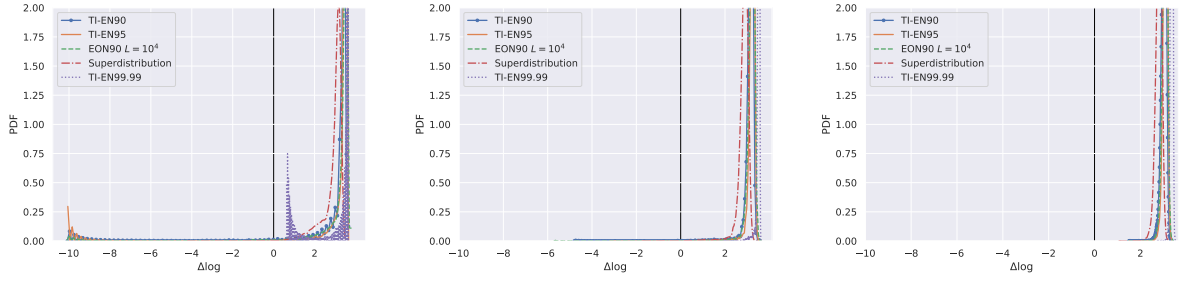
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



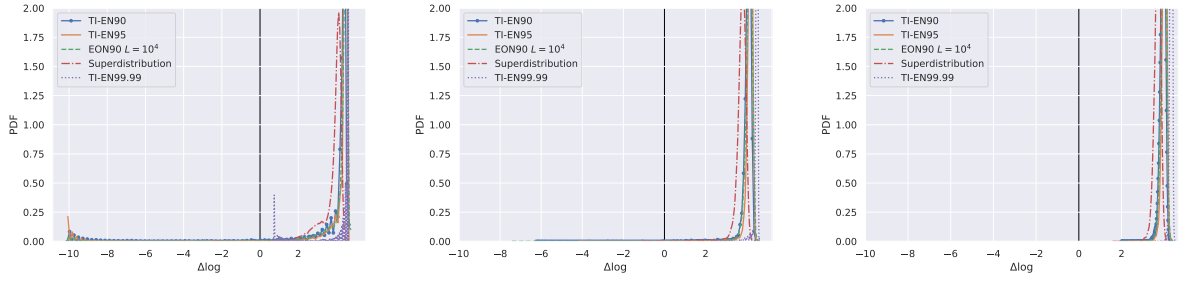
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$

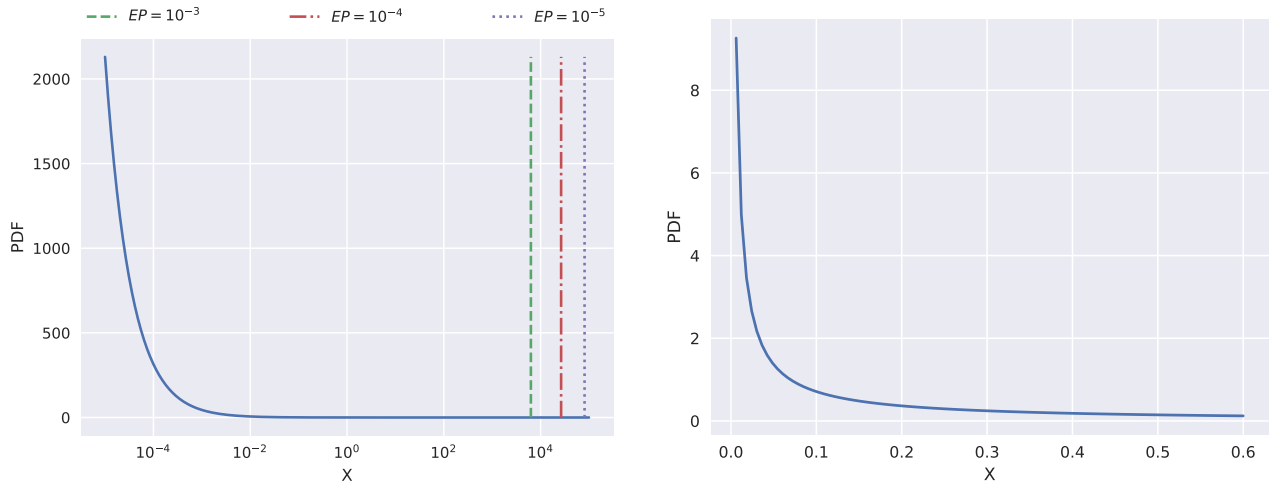


$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 31: Distribution of results for the bi-modal Log-Gamma Normal distribution.

3.8 Weibull Narrow distribution

The Weibull Narrow distribution is shown in Figure A.2, and can be described as a strictly positive distribution where an EP= 10^{-3} is several orders of magnitude away from the mode. This distribution was easily the most difficult for our sparse-sample tail-probability estimation methods. Results are shown in Figures 33 and 34. Note that the noise in the distribution plots is briefly explained in Appendix J. All of the methods were completely unsuccessful to conservatively estimate EPs. TI-EN99.99 performed the best of all methods. It had the best reliability curves over the full range $N = 2$ to 20 and also the best EP and EP10 metric values over the full range. However, even TI-EN99.99 did not perform acceptably for this PDF. Its highest reliabilities of 60%, 50%, 40% respectively for 10^{-3}, 10^{-4}, 10^{-5} EPs occurred at $N = 2$ samples, despite having a k factor that multiplied the raw sample standard deviation's by nearly 10,000. Reliabilities quickly declined to 13%, 5%, and 3% at $N = 4$ samples, trailing off to nearly zero at 6 to 20 samples depending on EP magnitude. Again, the other methods performed even worse. The results of the TI-EN99.99 were significantly better fro the larger EPs, where the TI-EN99.99 had a near 100% reliability for the entire range of samples with an EP= 10^{-1}.

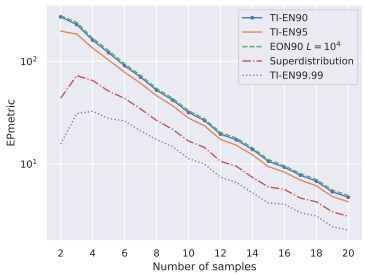


EP locations on log scale

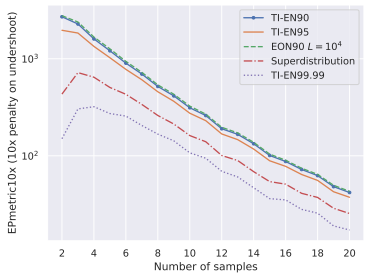
Close up linear scale

Figure 32: Weibull Narrow distribution.

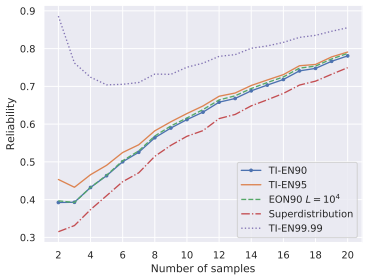
With reliability quickly plunging for all methods as the number of samples increases, the performance metric curves for all methods quickly rise ≈monotonically (with some noise) from their values at $N = 2$ samples. This is a completely different trend than for any of the other PDFs tried (when consider EP$\leq 10^{-3}$). For the other PDFs, performance results improved at least initially with added samples beyond $N = 2$, instead of immediately getting worse. This Weibull Narrow distribution is one example where our methods completely break down.



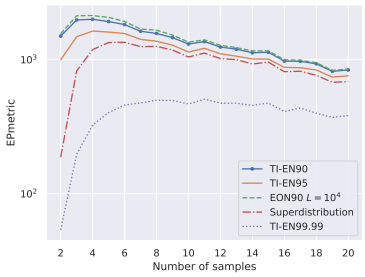
EPmetric, EP= 10^{-1}



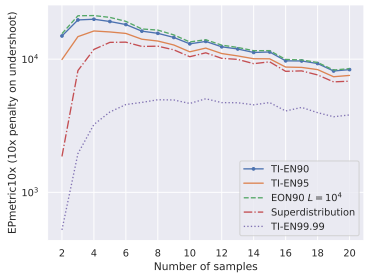
EPmetric10x, EP= 10^{-1}



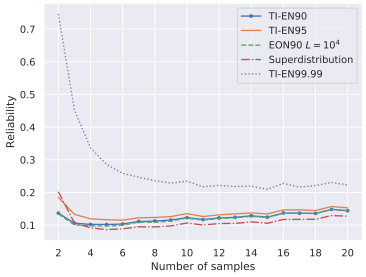
Reliability, EP= 10^{-1}



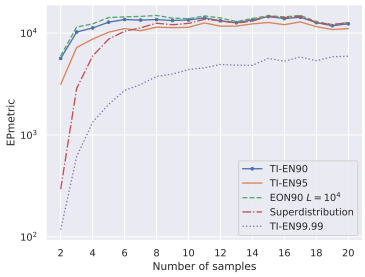
EPmetric, EP= 10^{-2}



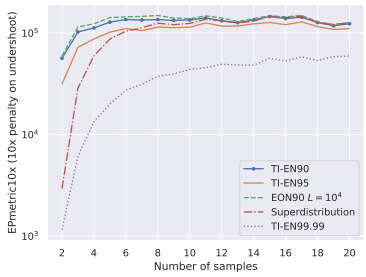
EPmetric10x, EP= 10^{-2}



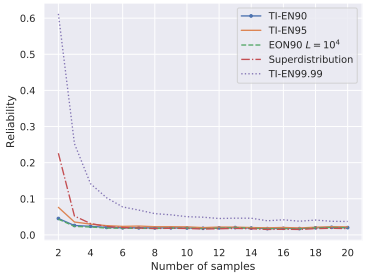
Reliability, EP= 10^{-2}



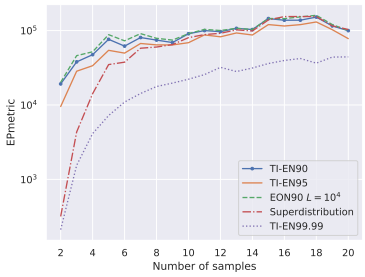
EPmetric, EP= 10^{-3}



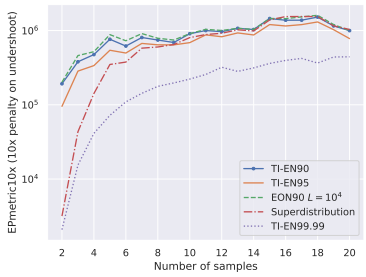
EPmetric10x, EP= 10^{-3}



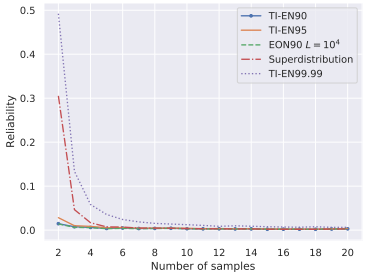
Reliability, EP= 10^{-3}



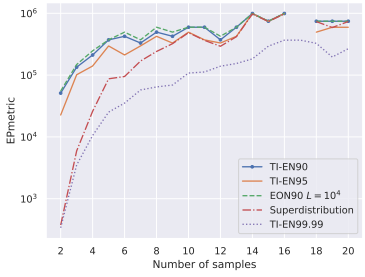
EPmetric, EP= 10^{-4}



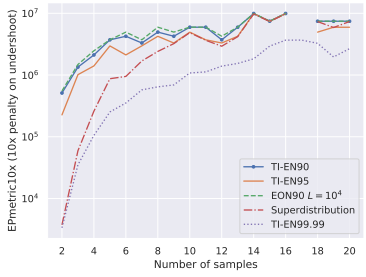
EPmetric10x, EP= 10^{-4}



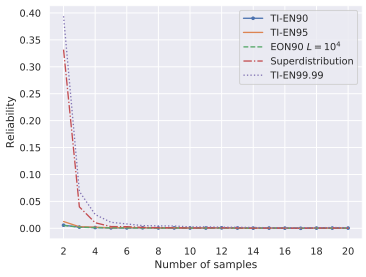
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

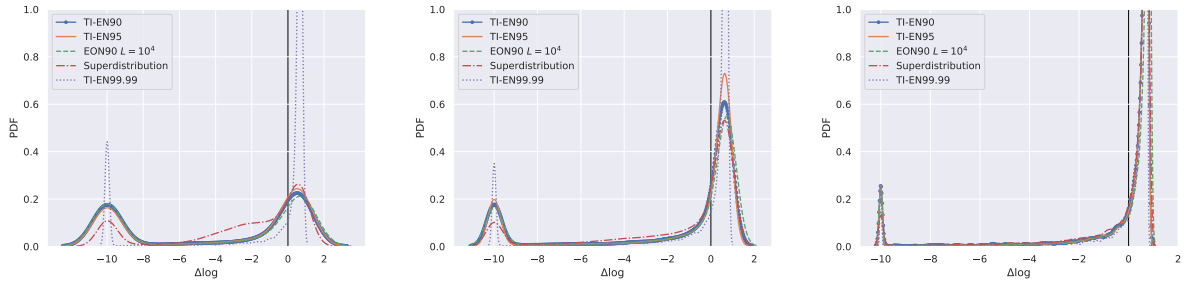


EPmetric10x, EP= 10^{-5}

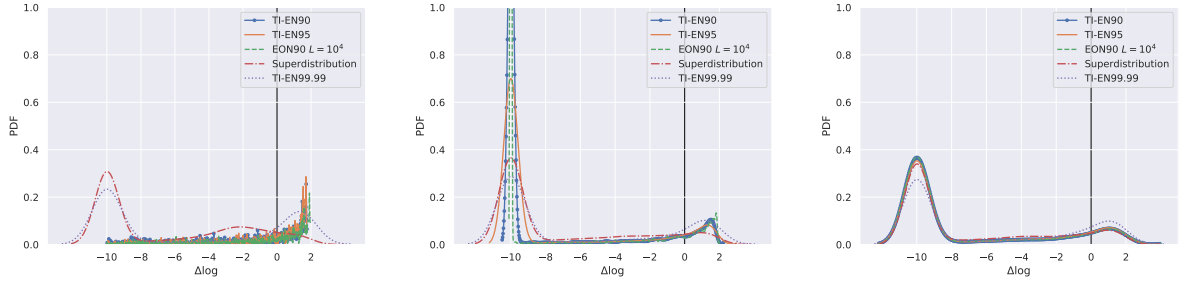


Reliability, EP= 10^{-5}

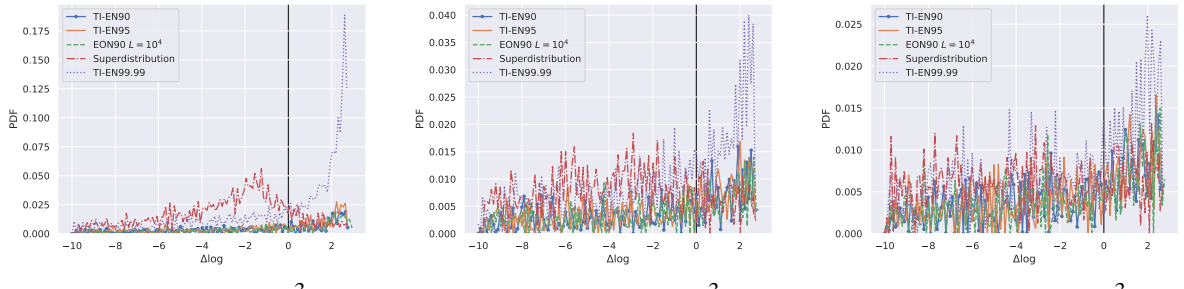
Figure 33: Results for the Weibull Narrow distribution.



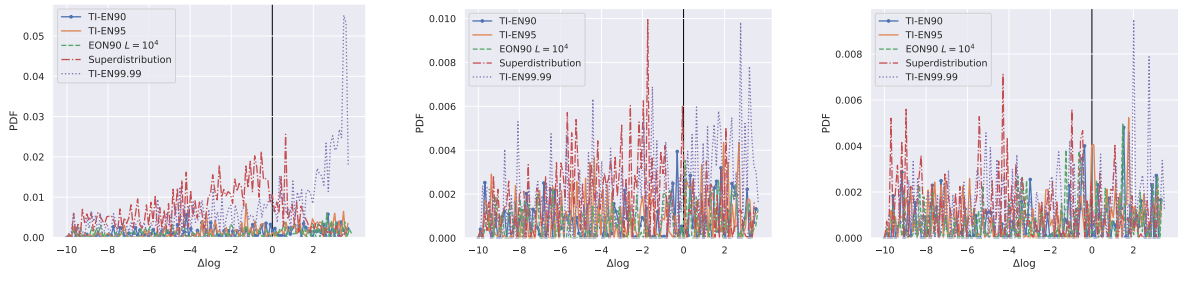
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



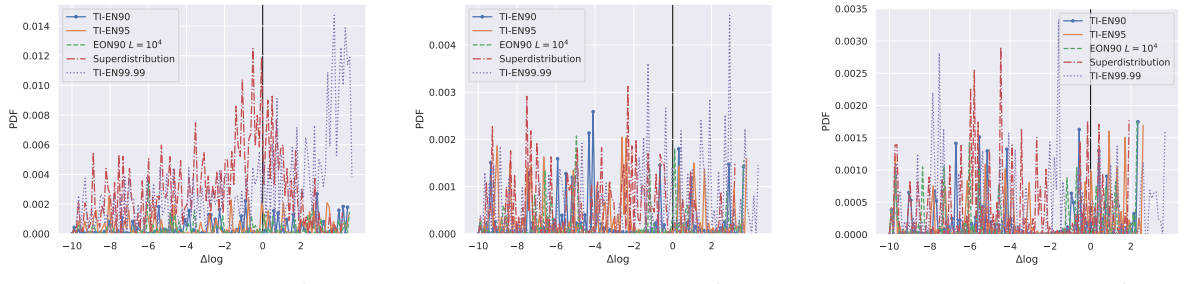
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure 34: Distribution of results for the Weibull Narrow distribution.

3.9 Summary Discussion of Results for Exceedance Probabilities $10^{-3}, 10^{-4}, 10^{-5}$

This subsection summarizes the extensive study just described comparing Tolerance-Interval Equivalent Normal (TI-EN), Ensemble of Normal (EON), and Superdistribution (SD) sparse-sample methods at predicting one-tail exceedance probabilities of magnitudes $10^{-3}, 10^{-4}, 10^{-5}$. The next subsection summarizes and discusses results for EPs 10^{-1} and 10^{-2}. The summaries are with respect to the 16 diverse distributions and sample numbers ranging from $N = 2$ to 20.

For EPs $10^{-5} - 10^{-3}$ the EON90 method performed the worst, often yielding the lowest reliability of conservative estimates while never having the best combined accuracy and reliability according to the EP and EP10X metrics. The TI-EN90 method almost always performed somewhat better than EON90, while the TI-EN95 was usually even better. The TI-EN99.99 method usually performed better yet according to the EP10X metric, but not according to the EP metric-which often indicated TI-EN99.99 was the worst performer. This bifurcation arises because the TI-EN99.99 method always had the highest reliability of conservative estimates (across all PDFs, numbers of samples, and EP magnitudes) and the EP10X metric rewards reliability of being conservative relatively more than small error (accuracy) of the conservative estimates. Hence, TI-EN99.99 performs relatively better under the EP10X metric, but in many cases greatly overestimates the true EP. This drawback is reflected more with the EP metric.

SD outperformed TI-EN99.99 (and the other methods) according to both metrics on 15 of the 16 test distributions. Even though TI-EN99.99 is always most reliable, SD is sufficiently reliable to outperform TI-EN99.99 on the strength of SDs better accuracy on 15 of the 16 distributions for lower numbers of samples (up to 3 to $\approx$10 depending on EP magnitude, metric, and PDF). For the majority of PDFs, SD is eventually outperformed by one or another of TI-EN99.99 and the other methods at higher numbers of samples (beyond $\approx$5 to $\approx$10). This occurs because SD's reliability typically declines faster than the other methods' as samples are added. This often causes SD performance to be eclipsed by one or more other methods as samples are added beyond SD's optimal performance points. However, SD's optimal performance at these points (numbers of samples) is never bested by the other methods for any number of samples, according to both metrics. (Again, this is for 15 of the PDFs. The 16th PDF always has TI-EN99.99 as the best method, but not even TI-EN99.99 performs acceptably on this PDF). Thus, SD had better performance and at lower numbers of samples (less expensive) than the other methods investigated, for 15 of the 16 PDFs. Further summary insights into some sample numbers and their reliabilities with SD follow.

- $N = 2, P \leq 10^{-3}$: 15 of 16 PDFs, reliability $> 95\%$
- $N = 3, P \leq 10^{-3}$: 14 of 16 PDFs, reliability $> 80\%$
- $N = 4, P \leq 10^{-4}$: 14 of 16 PDFs, reliability $> 80\%$
- $N = 4, P \leq 10^{-5}$: 15 of 16 PDFs, reliability $> 85\%$
- $N = 5, P \leq 10^{-5}$: 14 of 16 PDFs, reliability $> 80\%$

Table 1 provides more granular insight into SD performance on the individual PDFs. This information can be used to make decisions about how to best use the SD method if the PDF shape and EP magnitude are approximately known. For now we consider only the data columns for EPs $10^{-5} - 10^{-3}$ and note that the entries and footnotes in the 10^{-1} and 10^{-2} columns reflect a breaking-down of SD performance relative to the TI-EN methods. The substance of the footnotes is developed in the next subsection. Hypothetically, if the PDF shape is known for a quantity being sparsely sampled, then the corresponding row of the table can be used to decide on an optimal or suitable number of samples to use with the SD method. (In section 4 we consider use of the SD method with various resampling approaches, where the best number of samples differs from what we cite here for SD alone.) Within a given row of Table 1, if one further hypothetically knows in advance the approximate order of magnitude of the true EP, then the SD optimal number of samples can be identified per the table. These optima give the best combination of accuracy and conservatism per the EP and EP10X performance metrics.

EP10X optimal sample numbers are typically different (one to several counts lower) than EP metric optimal numbers, for a given PDF and EP magnitude. This is consistent with the EP10X metric's somewhat greater emphasis on reliability over accuracy and vice-versa for the EP metric as explained previously. The lower optimal sample numbers for the EP10X metric engender higher SD reliability rates than do the EP metric optimal sample numbers. In fact, the EP10X metric optimal sample numbers yield $\geq 80\%$ reliability for 15 of the 16 PDFs and the three EP magnitudes $10^{-5} - 10^{-3}$, while the EP metric optimal sample numbers often do not. This is found by comparing the metrics' optimal sample numbers for a given PDF and EP magnitude against the sample numbers in the $\geq 80\%$ reliability columns of the table. These columns give the maximum number of samples allowable for 80% or higher reliability of the SD method. The typically smaller sample numbers in the EP10X columns often yield substantially higher than 80% reliability.

Table 1: Performance summary for Superdistribution or other cited best performing methods. For the reliability column, the number reported is the largest number of samples that resulted in at least 80% reliability for the SD method. For the EP10X and EP10 performance metric columns, the number of samples is reported that resulted in SD's best performance.

	80% SD Reliability, $N \leq$					SD optimal N per EPMetric10X					SD optimal N per EPMetric				
EP	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Normal	2	6	15	20	20	TI[1]	2[A]	4	8	8	4	20	20	20	20
Log-Normal	-	3	4	5	6	TI[2]	2[B]	3	4	5	2	3	5	6	7
5 d.o.f. t	2	3	4	5	5	TI[1]	3[B]	3	4	4	2	4	5	5	5
Weibull Wide	-	2	4	4	6	TI[2]	2[B]	3	4	4	2	3	4	5	6
Exponential Narrow	-	2	3	4	5	TI[2]	2[C]	2	3	3	TI[1]	2	3	4	5
Exponential Wide	-	2	3	4	5	TI[2]	2[C]	2	3	3	TI[1]	2	3	4	5
TP Weld Element 0.75	-	2	2	3	4	TI[2a]	2[C]	2	2	3	TI[1]	2	3	3	4
Tensile EQPS Can Top 0.5	-	3	3	4	5	TI[2]	2[B]	3	3	4	2	3	3	4	5
TP Lid Buckle 1.0	-	3	3	5	8	TI[2]	2[B]	3	4	5	2	3	4	6	8
TP Lid Buckle 0.25	2	10	4	6	9	TI[1]	2[A]	3	4	4	20	20	5	7	10
TP Weld Max Global 0.25	-	2	3	4	6	TI[2]	2[C]	2	3	4	2	2	3	7	8
T EQPS Weld Max Global 1.0	-	3	5	10	20	TI[2]	2[B]	3	5	9	2	3	6	12	20
EQPS Can Top Element 0.5	2	6	5	9	19	TI[1]	2[A]	3	5	7	20	20	5	8	12
T EQPS Lid Buckle Element 0.25	-	3	all	all	all	TI[2]	2[B]	3	4	8	TI[1]	20	20	20	20
Bi-modal	all	all	all	all	all	20[3]	20	20	20	20	20	20	20	20	20
Weibull Narrow	-	-	-	-	-	TI-EN99.99 better[4]					TI-EN99.99 better[4]				

[1] Behaves like Normal PDF in that TI-EN99.99 is best for $N = 2$ to 5; TI-EN95 slightly better/best for $N \geq 6$ with continual improvement with added samples so optimum performance is at $N = 20$ with TI-EN95, which is $> 90\%$ reliably conservative over full range $N = 2$ to 20. If ≥ 6 samples available, use TI-EN95 with as many samples as affordable.

[2] Behaves like Log-Normal PDF in that TI-EN99.99 is best at any N studied (2 to 20); mostly steady improvement with added samples but effectively plateaus at $N = 19$ or 20, $> 90\%$ reliably conservative over full range $N = 2$ to 20.

[2a] Same as Footnote 2 except TI-EN99.99 optimum performance occurs at $N = 6$.

[3] TI-EN99.99 performs best for $N = 2, 3, 4$. SD dominates all methods for $N \geq 5$ with $> 93\%$ reliability and EP10X metric values that are lower/better than TI-EN99.99's best (at $N = 4$); continual improvement with added samples so optimal SD occurs at $N = 20$.

[4] TI-EN99.99 performed better than SD for any N (2 to 20) but reliability is $> 85\%$ only for 2 or 20 samples.

[A] SD is best for $N \leq 4$ where $> 85\%$ reliably conservative; TI-EN99.99 is best or nearly so for $N = 5, 6$ and is $\approx 100\%$ reliably conservative, but performance is inferior to optimum SD with $N = 2$; TI-EN95 is best for $N > 6$ and improves steadily or nearly so with added samples while retaining $\approx 95\%$ reliability, so has optimal performance with $N = 18$ to 20. For $N \geq 10$ samples TI-EN-95 performance surpasses optimal SD with $N = 2$.

[B] SD best for $N = 2$ where $> 90\%$ reliably conservative; TI-EN99.99 best or nearly so for $N = 3$ and clearly best for $N \geq 4$ where $\geq 90\%$ reliably conservative and performance improves steadily or nearly so with added samples to optimal performance with $N = 18$ to 20. For $N \geq 10$ to 12 samples TI-EN99.99 surpasses optimal SD with $N = 2$ or 3 samples.

[C] SD is best for $N = 2$, or at least non-dominated by other methods, and is $> 85\%$ reliably conservative at $N = 2$; for $N \geq 3$ TI-EN99.99 is best but performance does not get significantly better than SD with 2 samples and eventually TI-EN99.99 performance declines with added samples. Therefore, no compelling reason to use anything but SD with economical 2 samples for these PDFs and 10^{-2} EP magnitude.

An important trend to note from the plots in Sections 3.1 to 3.8 and Appendix H is that, for any number of samples between $N = 2$ to 20 studied, the reliability of SD increases as the true EP becomes smaller (considering the EP range $10^{-5} - 10^{-3}$). The other methods' reliabilities *decrease* as the true EP becomes smaller. This is a promising trend regarding SD's continued conservatism (and the other methods' trend toward non-conservatism) when estimating EPs of magnitude $< 10^{-5}$. However, these risk trends reverse for EP magnitudes $> 10^{-3}$. This exposes a potential weakness of SD for EP estimation in this regime. To investigate this, studies with all methods were conducted for EPs 10^{-1} and 10^{-2}. The results are plotted along with the $10^{-3}, 10^{-4}, 10^{-5}$ results in sections 3.1 to 3.8 and Appendix H. The results are discussed next. The method performance trends and preferences for EP magnitudes 10^{-1} and 10^{-2} depart significantly from those for EPs $10^{-3}, 10^{-4}, 10^{-5}$.

3.10 Investigation and Discussion of Methods Performance for Exceedance Probabilities 10^{-1} and 10^{-2}

To get a sense of the various methods' performance vs. number of samples and EP magnitudes 10^{-1} and 10^{-2}, we concentrate on the results for the Normal and Log-Normal distributions. Results for most of the other PDFs are qualitatively similar to those of either the Normal or Log-Normal distributions.

We first examine the reliability and performance curves in Figure 9 for the Normal distribution. For EP $= 10^{-2}$, SD reliability starts at about 96% at $N = 2$ samples but then drops precipitously as samples are added. TI-EN95 reliability also starts at about 96% at $N = 2$ samples but does not drop appreciably as samples are added. TI-EN99.99 reliability remains at about 100% over the entire range $N = 2$ to 20. Considering the EP10X performance metric curves, SD's curve starts lowest/best at $N = 2$ and has a minimum there, then initially rises with added samples then hits a plateau as more samples are added. This is a departure from the SD curves for EPs $10^{-3}, 10^{-4}$, and 10^{-5} which all show initial performance improvement as samples are added (so have minima at more than $N = 2$ samples). SD is the best performer for $N \leq 4$ where $> 85\%$ reliably conservative. TI-EN99.99 is best or non-dominated for $N = 5, 6$ and is $\approx 100\%$ reliably conservative, but perfor-

mance is inferior to optimum SD with $N = 2$. TI-EN95 is best for $N > 6$ and keeps improving with more samples and retains a high reliability of about 95%. For $N \geq 10$, TI-EN95's EP10X metric values are lower/better than the lowest/best value for SD (at $N = 2$). Therefore, SD's optimal performance (with $N = 2$ samples) can be improved upon with TI-EN95 if ≥ 10 samples are available. This is noted by the caveat asterisk [A] on the entry in Table 1 for SD's optimal performance point ($N = 2$ samples) in the EP $= 10^{-2}$ column. The other PDFs that share this same asterisk in the table have qualitatively and quantitatively similar results as the Normal PDF per the explanation beneath the table.

For EP $= 10^{-1}$, SD reliability starts at about 81% at $N = 2$ samples but then drops precipitously. TI-EN95 reliability starts substantially higher at about 96% at $N = 2$ and drops very slowly to about 92% at $N = 20$. TI-EN99.99 reliability remains at about 100% over the entire range $N = 2$ to 20. Of the EP10X performance curves, TI-EN99.99 starts lowest/best at $N = 2$ and continually improves (metric value drops) as samples are added. TI-EN95's performance also improves continually and even faster, eclipsing TI-EN99.99 performance for $N \geq 6$. Optimal TI-EN95 performance occurs at $N = 20$. SD is never competitive over the tested range of 2 to 20 samples. These observations are summarized in Footnote 1 of Table 1. The other PDFs that share this same footnote in the table have qualitatively and quantitatively similar results as the Normal PDF as described in the footnote.

Figure 15 has the results for the Log-Normal distribution. For EP $= 10^{-2}$, SD reliability starts at about 95% at $N = 2$ samples but then drops precipitously as samples are added. TI-EN95 reliability also starts at about 95% at $N = 2$ samples but drops much more slowly to about 56% at $N = 20$. TI-EN99.99 reliability remains at about 100% over the entire range $N = 2$ to 20. Considering the EP10X performance metric curves, SD's curve starts lowest/best at $N = 2$ and rises quickly with added samples. Again, this is a departure from the SD curves for EPs $10^{-3}, 10^{-4}$, and 10^{-5} which all show initial performance improvement as samples are added (so have minima at more than $N = 2$ samples). SD is the best performer for $N = 2$ where $> 80\%$ reliably conservative. TI-EN99.99 is the best or at least non-dominated for $N = 3$ and is clearly best for $N \geq 4$ and keeps improving with more samples and retains a high reliability of nearly 100%. For $N \geq 10$ to 12, TI-EN99.99's EP10X metric values are lower/better than the lowest/best value for SD (at $N = 2$). Therefore, SD's optimal performance (with $N = 2$ samples) can be improved upon with TI-EN99.99 if ≥ 10 to 12 samples are available. This is denoted by the caveat double asterisks [B] on the entry in Table 1 for SD's optimal performance point ($N = 2$ samples) in the EP $= 10^{-2}$ column. The other PDFs that share this double asterisk in the table have qualitatively and quantitatively similar results as the Log-Normal PDF per the explanation beneath the table.

For EP $= 10^{-1}$, SD reliability starts at about 75% at $N = 2$ samples but then drops precipitously. TI-EN95 reliability starts substantially higher at about 94% at $N = 2$ and drops very slowly to about 80% at $N = 20$. TI-EN99.99 reliability remains at about 100% over the entire range $N = 2$ to 20. Of the EP10X performance curves, TI-EN99.99 is always handily lowest/best over the full range from 2 to 20 samples and continually improves (metric value drops) as samples are added. Optimal TI-EN99.99 performance is at $N = 20$. These observations are summarized in Footnote [2] of Table 1. The other PDFs with footnotes [2] and [2a] in the table have qualitatively and quantitatively similar results as the Log-Normal PDF as described in the footnotes.

The 5 d.o.f. t-distribution has results that resemble those of the Normal distribution for EP $= 10^{-1}$, and those of the Log-Normal distribution for EP $= 10^{-2}$. Four other mixed-behavior PDFs exist that resemble Log-Normal behavior for EP $= 10^{-1}$ (Footnote [2] or [2a]) but for EP $= 10^{-2}$ have results signified by caveat [C] beneath the table. These four distributions have a long tail to the side where the EPs are calculated, but no tail or a very short tail to the other side.

Assessing over all the values in the Table's EP10X metric column for EP $= 10^{-2}$ indicates that SD's optimal performance occurs at: $N = 2$ for 13 of the 16 PDFs; $N = 3$ for 1 PDF; and $N = 10$ is the knee in the curve (economical effective optimum) for another. At these optimal points SD performs better than the other methods. The 16th PDF, Weibull Narrow, has TI-EN99.99 as the best method for $N = 2$ to 20 (and for any of the EP magnitudes tested). In the generic situation when PDF shape is not known, it is not knowable what value of N is optimal for SD. But the choice SD with economical $N = 2$ samples would yield optimum SD performance for 13 of the 16 PDFs, and for 15 of the PDFs would yield $> 85\%$ reliability and be the best or at least non-dominated method with $N = 2$ samples. The asterisks in the column signify that for 10 of the 15 PDFs, SD performance with 2 samples can be surpassed by TI-EN95 in 3 cases or TI-EN99.99 in 7 cases if substantially more samples are available/affordable ($N \geq 10$ to 12). In these circumstances reliability is $\geq 90\%$ with the TI-EN methods. *However, for a generic strategy when PDF shape is not known, an economical strategy of 2 samples with SD may be the best approach for* 10^{-2} *EP magnitude. This yielded* $> 85\%$ *reliability for 15 of the 16 PDFs.* The TI-EN95 or TI-EN99.99 methods with $N \geq 12$ as generic strategies would be significantly more expensive and only offer better reliability and accuracy in the best scenario (for half of the 16 PDFs with a generic choice TI-EN99.99), but would underperform SD with 2 samples in the other half of cases.

- SD: $N = 2, P \leq 10^{-2}$: 15 of 16 PDFs, reliability $> 85\%$

For the EP $= 10^{-1}$ column in the table, the EP10X metric indicates that TI-EN99.99 is the best performing method for $N = 2, 3, 4$ and has reliability $\geq 90\%$ over this range for 15 of the 16 PDFs (Weibull Narrow being the outlier). The trend is toward improving performance with increasing samples from 2 to 4, for all PDFs except Weibull Narrow, which has the opposite trend. With added samples beyond $N = 4$, TI-EN99.99 performance continually improves or remains essentially flat for all PDFs, and reliability remains $> 90\%$ for 15 of the 16 PDFs. For 5 of the PDFs, methods other than TI-EN99.99 (associated with footnotes 1 and 3) improve faster with added samples and ultimately perform better than TI-EN99.99 for $N \geq 5$ or 6. For the other 11 PDFs, TI-EN99.99 is the best performer over the entire range $N = 2$ to 20. *A good generic strategy for* 10^{-1} *EP magnitude problems with unknown PDF shape would be to use TI-EN99.99 with as many samples as affordable, up to* $N = 20$ *studied here. This would yield best performance on 11 of the PDFs tried, and best on the other 5 PDFs if the affordable N happens to be* ≤ 4*, with* $> 90\%$ *reliability for 15 of the 16 PDFs with any number of samples* $N = 2$ *to 20.* SD under-performs for 10^{-1} EP as follows.

- SD: $N = 2, P \leq 10^{-1}$: 5 of 16 PDFs, reliability $> 80\%$

3.11 Forward

When very few samples are involved, the SD method generally outperforms the other sparse-sample methods for EP magnitudes 10^{-2} and smaller.

Figure 35 shows how the reliability of the SD method (averaged over the 16 distributions) varies as a function of the number of samples and the EP magnitude. For a given EP magnitude, the average reliability decreases by $\approx 50\%$ as the number of samples increases from 2 to 10. Average reliability generally varies less sensitively with EP magnitude (for a given number of samples) over the range investigated from 10^{-5} to 10^{-1}, but reliability declines significantly and in an accelerating fashion in going from 10^{-2} to 10^{-1} when very few samples (2 or 3) are involved. In this regime, TI-EN 99.99 is preferred to SD. However, the selection of which sparse-sample method to use in a given circumstance is best done in association with the material in next section, where resampling approaches are tested and some are found to significantly improve reliability and accuracy (for the same total number of samples) beyond using the sparse-sample methods alone.

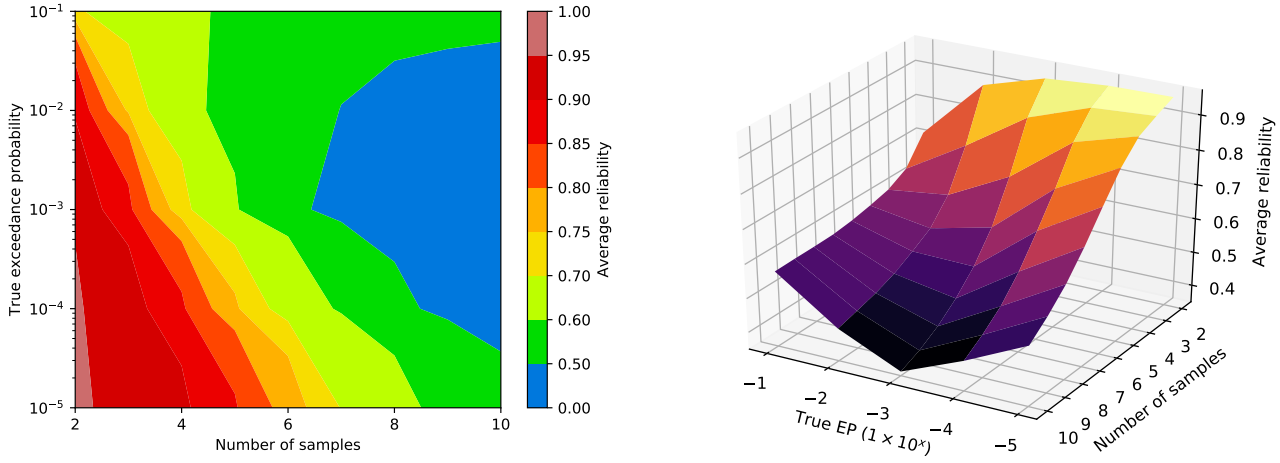


Figure 35: Average reliability of Superdistribution method over the 16 distributions studied, as a function of *EP* magnitude and number of samples.

4 Resampling to Improve SD Performance on Tail Probability Predictions with More Samples than Optimal for SD Alone

The previous section demonstrated that an exceedance probability (EP) can be estimated using the limited-data UQ methods on a variety of analytical and empirical distributions. The Superdistribution method was usually the most accurate and reliably conservative for the fewest samples. However, for SD and the other methods tested, accuracy and reliability deteriorated as the number of samples increased, or an optimum performance was reached beyond which performance deteriorated as more samples were added. This is a very counterintuitive result, but explained by how the sparse-sample UQ methods work. This section explores improving the accuracy and reliability of EP predictions by the SD method using Bootstrapping and Jackknifing resampling techniques which use higher sample sizes than the optimal SD results presented in Table 1.

4.1 Bootstrapping

Statistical resampling methods have been used to reduce the bias (or error) in an estimated statistic [13]. One of the most popular resampling methods today is Bootstrapping, where a statistic is estimated by first calculating the statistic on many different sample combinations created by sampling the original set with replacement. This results in many different estimates and provides insight into the uncertainty regarding the possible true value of the statistic. A best-estimate EP can be found by averaging the individual EP estimates. Bootstrapping was used by Picheny et al. [14] to more conservatively estimate a 99% tail probability from 100 samples. In this work, we are concerned with a much smaller probability and far fewer samples.

The most basic form of Bootstrapping is case resampling. For N number of samples, there are

$$\binom{2N-1}{N} = \frac{(2N-1)!}{N!(N-1)!} \tag{4.1}$$

total number of combinations with replacement. For example, the combinations with replacement for a sample of $N = 3$ with values $[1,\ 2,\ 7]$ are the following:

$$\begin{array}{ccccc} [1,\ 1,\ 1] & [1,\ 1,\ 2] & [1,\ 1,\ 7] & [1,\ 2,\ 2] & [1,\ 2,\ 7] \\ [1,\ 7,\ 7] & [2,\ 2,\ 2] & [2,\ 2,\ 7] & [2,\ 7,\ 7] & [7,\ 7,\ 7]. \end{array}$$

An EP would be calculated (using SD or other UQ method) on each of the above sets, then the average of the EPs would represent the Bootstrapped EP prediction. This type of Bootstrapping is referred to as *exact case* Bootstrap resampling [15].

As N grows larger, computing the EP on all of the combinations with replacement becomes computationally infeasible. With $N = 20$, there are over 68 billion combinations with replacement. In these infeasible cases, the EP can be predicted on a large number of the possible combinations until the mean of the predicted EPs converges.

A generally more popular variant of exact Bootstrapping would perform 27 combinations instead of the 10 shown above. This variant would express the total number of combinations as N^N. This variant might perform better than the variant studied in this report, because it adds more weight to a resample with larger variation (and thus larger resampled standard deviation). Some initial exploration with the more popular variant may make it into [7], which is under revision at the time of publication of this SAND report.

4.2 Generalized Jackknife

Jackknifing is another popular resampling method that predates Bootstrapping. The Jackknife was originally created by Quenouille [16], and the term Jackknifing was coined by Tukey [17]. Jackknifing involves estimating a statistic from N data points by calculating the statistic on all of the $(N-1)$ subsample combinations (without replacement). The resulting Jackknife statistic is simply the average of the subsample statistics. Jackknifing, like Bootstrapping, has been shown to reduce the bias in the estimated statistic. A generalized Jackknife was proposed by Schucany et al. which averages the estimated statistics on combinations of $(N-m)$ samples[18]. Typically m is a small number [19], and in many cases becomes one [18].

We next propose and test a method that is an extreme instance of the generalized Jackknife, such that sparse subsample combinations are considered. We define r as the subsample size, where

$$r = N - m \tag{4.2}$$

and N is the number of samples, m is the generalized Jackknife parameter. While typically m is a small number in most generalized Jackknife applications, in this proposed method m will be a number near N, resulting in r being a small number. This study considered subsample sizes of $r = 2$ through 10. The Jackknife estimated EP is the average EP estimate from the various combinations of subsamples.

Two distinctions can be made between Jackknifing and Bootstrapping. First Bootstrapping considers the combinations with replacement, while Jackknifing combinations draw from the original sample without replacement when creating a subsample. Second, Jackknifing considers subsamples while Bootstrapping considers new samples that are the same size as the original sample set.

The combinations in Jackknife sampling are typically expressed as N choose r or "NCr". For any given NCr, there are

$$\binom{N}{r} = \frac{N(N-1)\cdots(N-r+1)}{r(r-1)\cdots 1} = \frac{N!}{r!(N-r)!} \tag{4.3}$$

total number of possible combinations. This is a Pascal's triangle relationship with respect to r for any number N. An example of how the total number of combinations varies with respect to r is shown in Figure 36 for $N = 10$. In this case, the largest possible number of combinations occurs when $r = 5$. For applications when N is large, it may be impractical to compute and average all of the subsample combinations. In these cases it is recommended to take a large number of the possible combinations until the mean of the statistic estimates converges.

A tail probability can be estimated using this proposed Jackknife method along with any given sparse-sample UQ method. A tail probability is estimated using the following procedure:

1. Pick an appropriate value of r. (This is still being studied. A suggestion is given later.)
2. Consider the NCr combinations of subsamples from a sample of a random variable with N total number of data points.
3. Randomly take one combination and calculate the EP using the UQ method for a subsample size of r.
4. Repeat 3 until all possible combinations have been exhausted, or the mean of the predicted tail probabilities has converged.
5. Report the mean of the EP estimates.

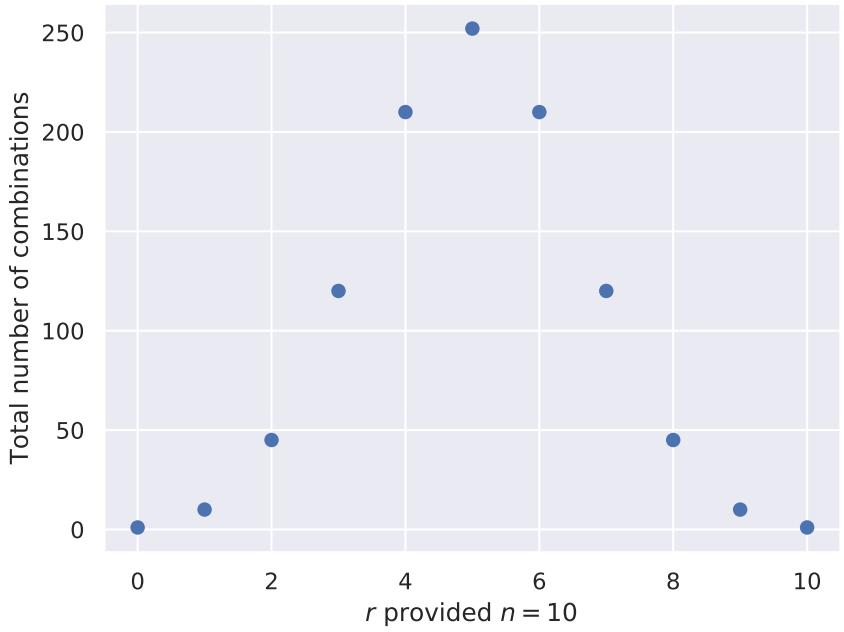


Figure 36: Example of how the total number of possible combinations in NCr follows Pascal's triangle for $N = 10$.

For example consider a sample of $[1,\ 2,\ 7,\ 5,\ 3]$. A $5C3$ Jackknife would consider the following combinations:

$$[1,\ 2,\ 7] \quad [1,\ 2,\ 5] \quad [1,\ 2,\ 3] \quad [1,\ 7,\ 5] \quad [1,\ 7,\ 3]$$
$$[1,\ 5,\ 3] \quad [2,\ 7,\ 5] \quad [2,\ 7,\ 3] \quad [2,\ 5,\ 3] \quad [7,\ 5,\ 3].$$

This generalized Jackknife method requires a subsample size r to be selected. This selection's effect on the method performance will be investigated in the next section. It is also desirable to

explore a parameter-free Jackknife variation. In cases where N is small, it would be feasible to explore all of the complete subsample combinations. For a sample size of N, the EP would be predicted on all of the $NC(N-1), NC(N-2), \cdots, NC2$ combinations. The final EP would result from the average of all Jackknife combinations. In total there are

$$\sum_{k=2}^{N-1} \binom{N}{k} = 2^N - (N+2) \tag{4.4}$$

possible Complete Jackknife subsample combinations to consider. All may not be feasible to compute when N is large. This method is referred to as the Complete Jackknife technique.

If we were to consider the set $[1,2,7,5]$, the Complete Jackknife consists of the following combinations:

$$\begin{aligned} 4C3 &: [1,\ 2,\ 7] \quad [1,\ 2,\ 5] \quad [1,\ 7,\ 5] \quad [2,\ 7,\ 5] \\ 4C2 &: [1,\ 2] \quad [1,\ 7] \quad [1,\ 5] \quad [2,\ 7] \quad [2,\ 5] \quad [7,\ 5]. \end{aligned}$$

4.3 Results of SD with Bootstrapping and Jackknifing on 10^{-4} Tail Probabilities

This section investigates whether Bootstrapping and Jackknifing resampling techniques can be used to improve SD predictions of tail probabilities with limited samples. The techniques are applied to several of the distributions from Section 3 that represent the easiest to the most difficult for EP estimation with SD (and the other methods). The reliability and EP and EP10X performance metrics are used to characterize conservatism and accuracy over 10,000 random trials of the SD+resampling methods for any given number of samples N. We present but don't discuss the EP10X metric results. They tend to over-emphasize reliability compared to accuracy so are a less sensitive measure of performance when reliabilities are all quite high as is the case among the most competitive methods in the following.

Improvements to EP estimates are investigated using the Exact Case Bootstrapping method, along with $NC2$, $NC3$, $NC4$, $NC5$ Jackknifing, and the Complete Jackknife method. The number of samples for the Exact Case Bootstrapping ranged from $N = 3$ to $N = 8$, keeping to low numbers due to computational cost. The generalized Jackknifing methods were also exact, meaning that every possible combination was computed rather than selecting combinations until the mean demonstrated convergence. The NCr methods used up to $N = 14$, and the Complete Jackknife used up to $N = 11$. While all of these resampling techniques do not go up to the full range of $N = 20$ samples, conclusions can be drawn about the trend of their performance when compared to using just SD alone. We consider a true EP of 10^{-4} for each distribution.

Only the results for SD and TI-EN95 with resampling are presented and analyzed in the next two subsections. The results for Bootstrapping and Jackknife resampling with the TI-EN90 and EON methods can be found in Appendix I, where similar trends are observed, but not as good absolute performance. [5] presents side-by-side quantitative comparisons of resampling performance with SD and TI-EN95 methods.

4.3.1 Results on most difficult analytic distributions

The said techniques are applied to the analytical distributions from Section 3 that were the most difficult to predict tail probabilities for. These are Student's t-distribution, the Weibull Narrow distribution, and the Exponential Wide distribution. We keep to analytical distributions here so that other researchers can try their methods on them to compare against our results.

4.3.1.1 Student's t-distribution

The Student's 5 d-o-f t-distribution results for SD are shown in Figure 37. The Exact Bootstrap does not appear to offer reliable improvement in either accuracy or reliability. Reliability is sometimes significantly better and sometimes significantly worse with SD alone than with the Exact Bootstrap when considering the same number of samples. The EPmetric indicates that combined accuracy + reliability performance is sometimes better and sometimes worse between the SD and Exact Bootstrap methods. Thus, no overriding trends above the performance variability differentiate SD-Bootstrap from SD alone.

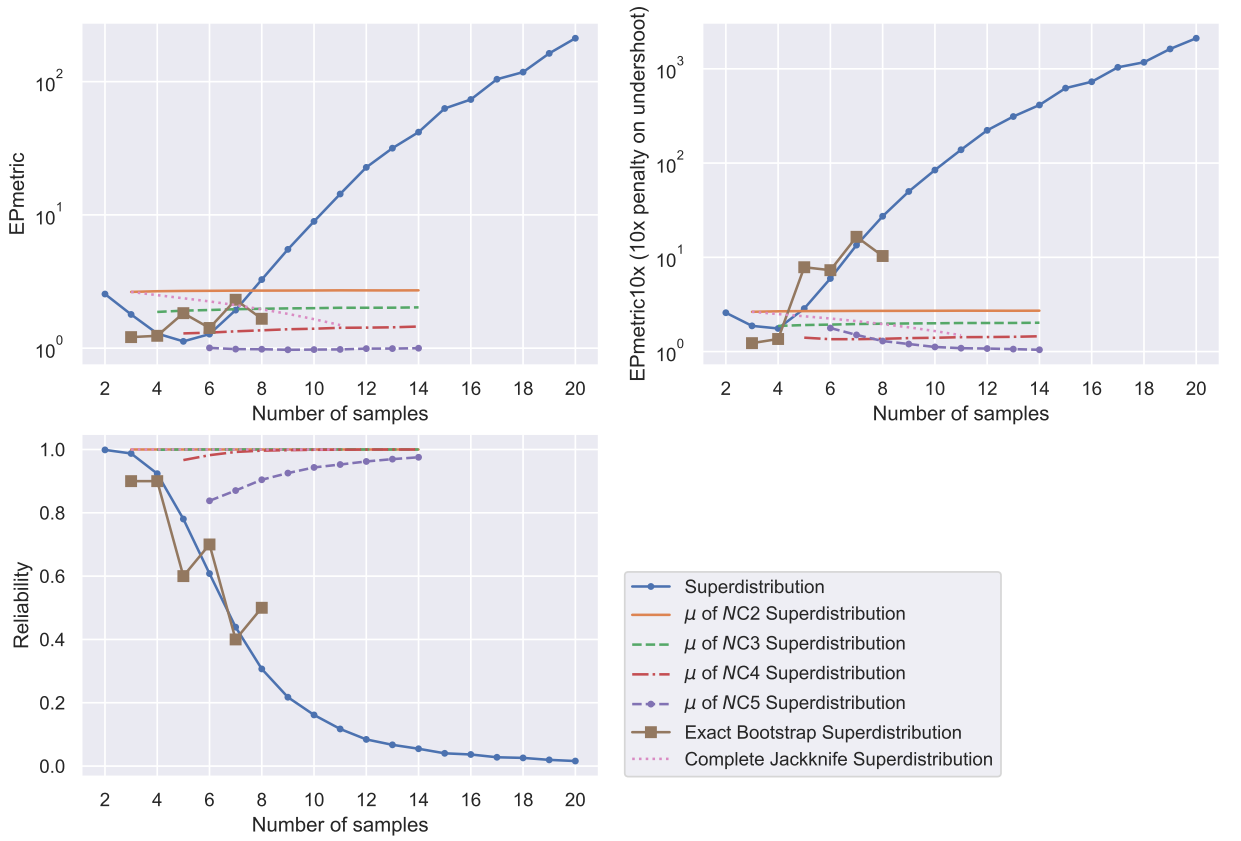


Figure 37: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on a 5 d-o-f Student's t-distribution for $EP = 10^{-4}$.

However, generalized Jackknifing with SD shows promise to offer improved reliability and combined accuracy + reliability performance. Considering reliability, the SD Jackknife (SDJ)

method always improved on SD's reliability for the same number of total samples. For instance, with $N = 5$ the SDJ methods had a reliability $> 95\%$, where the reliability of SD alone was $< 80\%$.

For combined accuracy + reliability according to the EPmetric, SDJ with certain *NCr* parameters significantly improved on the results of SD alone for the same number of total samples. In particular, optimal SD accuracy + reliability by the EPmetric occurs with $N_{\text{SDopt}} = 5$ samples for this t-distribution. Adding more samples degrades EPmetric performance if SD is used without resampling. However, performance improves with added samples if used in the SD-Jackknife method with NC5, which corresponds to *NCr* where $r = N_{\text{SDopt}} = 5$. Lower values of r ($r < N_{\text{SDopt}}$) result in SD-Jackknife performance that never (for any number of samples N) surpasses that of optimal SD with 5 samples. The *NC*5 results have the lowest EPmetric value of any of the methods over the range $N \geq 6$ where *NC*5 is possible. Over this range, *NC*5 SDJ also had considerably better reliability than SD alone. *NC*5 reliability is 83% for $N = 6$ and increases as N increases, while the SD reliability is just 60% at $N = 6$ and quickly deteriorates as N increases.

Optimum SDJ performance occurs within one or two samples beyond the optimum number for SD alone, $N = 6$ or 7 total samples in this case. Further samples add expense but result in improved reliability. However, additional samples do not appreciably improve overall performance (by the EPmetric and also if one considers cost).

The performance of the Complete SDJ *does* continue to improve as samples are added. However, the EPmetric improvement starts from a relatively high/bad value at $N = 3$ samples. The improvement trend indicates about 14 or more total samples are required to achieve the same level of EPmetric accuracy + reliability performance as *NC*5 SDJ with $N = 6$ total samples. However, complete SDJ always has higher reliability (≈ 1) than the optimum SDJ (*NC*5), though the latter has desirably high reliability of > 0.84 for all N tried.

The results of the TI-EN95 method with and without resampling applied to Student's t-distribution are shown in Figure 38. Many of the trends observed with SD and SD+resampling apply here as well. Starting from $N = 2$ samples, the TI-EN95 results first improve with added samples then deteriorate with more samples (EP metric first dips then climbs), and reliability always declines with added samples.

TI-EN95 has significantly worse EPmetric performance than SD for $N \leq 14$. The lower-performing TI-EN95 is improved by the resampling techniques proportionately more than they improve SD, both in terms of reliability and EPmetric reliability+accuracy. Indeed, all the *NCr* results demonstrate higher accuracy and reliability than when using TI-EN95 alone, for the same total number of samples.

The *NCr* EPmetric results appear to plateau as the number of samples increases, while the Complete Jackknife results continue to improve and may be best for $N \geq 14$. Like with SD-Bootstrapping results, TI-EN95-Bootstrapping offered only marginal improvements to the EPmetric and reliability when compared to using TI-EN95 alone.

Although the TI-EN95 results improve significantly with Jackknifing, the results are not better in an absolute sense than the SD-Jackknifing results. If we consider the best results using the

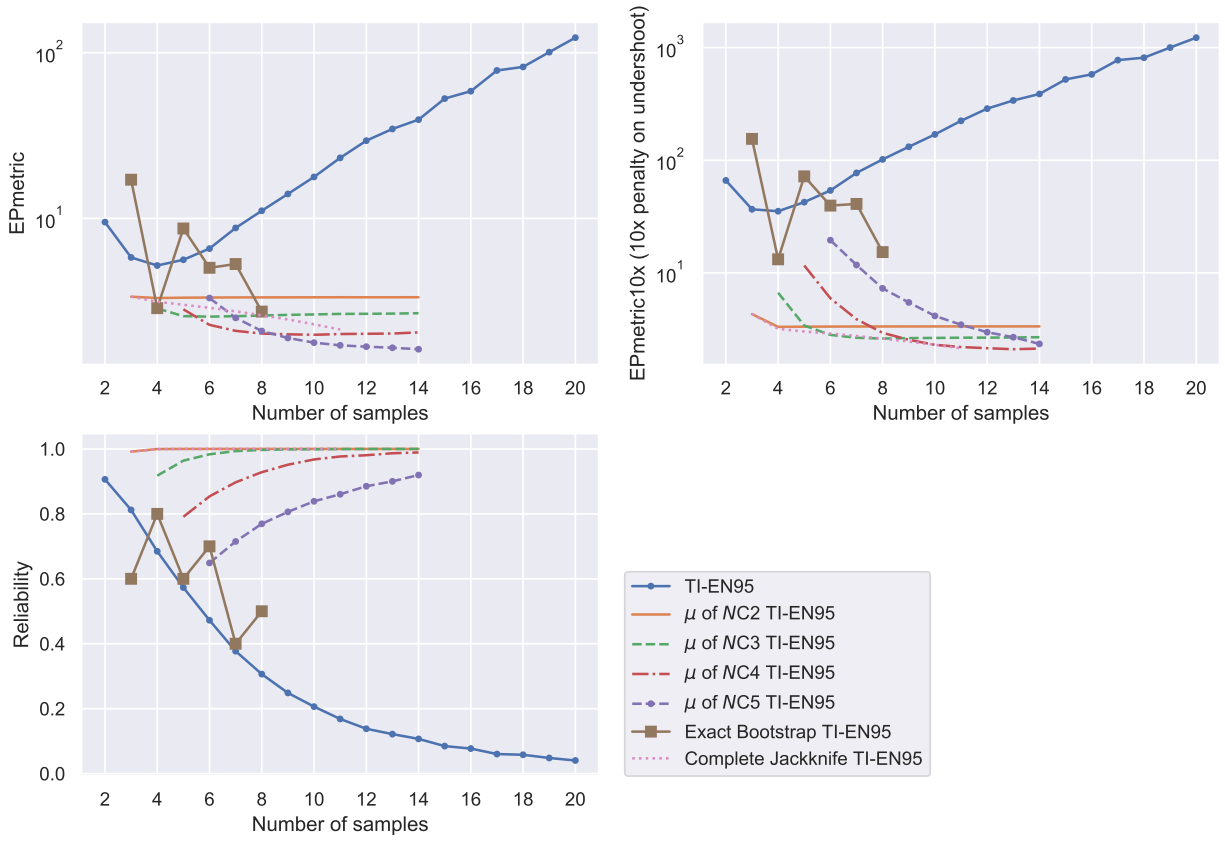


Figure 38: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on Student's t-distribution for $EP = 10^{-4}$.

resampling techniques, the $NC5$ SDJ results had both a lower EPmetric and higher reliability than the TI-EN95J $NC5$ results for all N tried. This is also true for optimal $NC5$ SDJ vs. optimal TI-EN95J NCr where $r = N_{\text{TIEN95opt}} = 4$.

4.3.1.2 Exponential Wide distribution

The SD results applied to the Exponential Wide distribution are shown in Figure 39. Results are qualitatively very similar to those for the 5 d-o-f t-distribution. There is little difference between the SD results without resampling and those with Bootstrap resampling.

Again, SDJ always had higher reliability than SD alone, for the same number of total samples. Concerning EPmetric reliability + accuracy performance, when the number of samples is increased beyond the SD-only optimum $N_{\text{SDopt}} = 4$ for this distribution, the EPmetric performance quickly degrades. Conversely, performance quickly improves when SDJ resampling is used with NCr where $r = N_{\text{SDopt}} = 4$. $NC5$ results also exhibit better EPmetric performance than optimal SD alone, but do not attain a reasonable 80% reliability level until a relatively costly $N = 10$ samples. For half the cost ($N = 5$ samples, which is the lowest number possible for use of $NC4$ SDJ), $5C4$

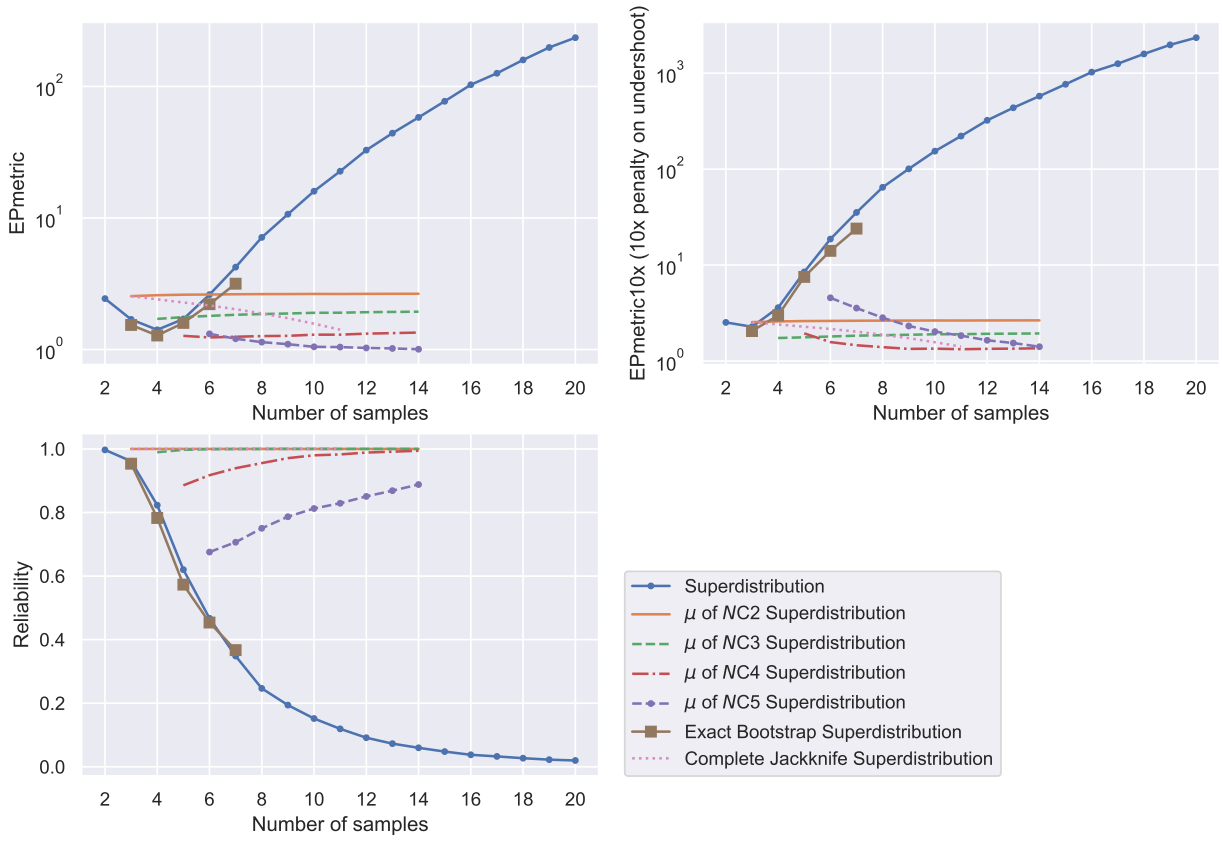


Figure 39: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-4}$.

reliability is about 88%. For SD alone, reliability is only about 60% for 5 samples.

In the other direction with $r < N_{\text{SDopt}}$, *NC*2 and *NC*3 EPmetric results are inferior to *NC*4 and *NC*5 results over the applicable range of *N* studied (up to *N*=14). *NC*2 and *NC*3 never (for any total number of samples studied) have better EP metric performance than optimal SD with 4 samples. However, their reliability is always significantly higher than the 78% reliability of optimal SD.

The optimal *NCr* SD-Jackknife method (*NC*4) exhibits an EPmetric performance optimum at $N = 6$, a few samples beyond the optimum number for SD alone. Further samples add expense, with no improvements–even degradation in–EPmetric performance. However, further samples do improve the already high reliability ($\approx$0.92 at $N = 6$, $\approx$0.99 at $N = 12$).

On the other hand, EPmetric performance of Complete SDJ (CSDJ) *does* keep improving as samples are added. CSDJ required about 12 samples to achieve the same EPmetric level of accuracy + reliability performance as *NC*4 SDJ with the optimal 6 samples. However, *NC*4 SDJ only has desirably high reliability of > 0.8 for $N \geq 10$ samples, while CSDJ has extremely high reliability (≈ 1) for any applicable number of samples $N > 2$.

Results for the TI-EN95 method with and without resampling are shown in Figure 48. Some of

the trends observed with SD apply here as well. Starting from $N = 2$ samples the TI-EN95 reliability always declines with added samples. EPmetric results do not first improve with added samples like they do for SD with this distribution; TI-EN95 EPmetric performance deteriorates immediately with added samples. The *NCr* and Exact Bootstrap EPmetric results appear to asymptote toward different plateaus as the number of samples increases. The exception is that the Complete Jackknife results start best and continue to improve with added samples.

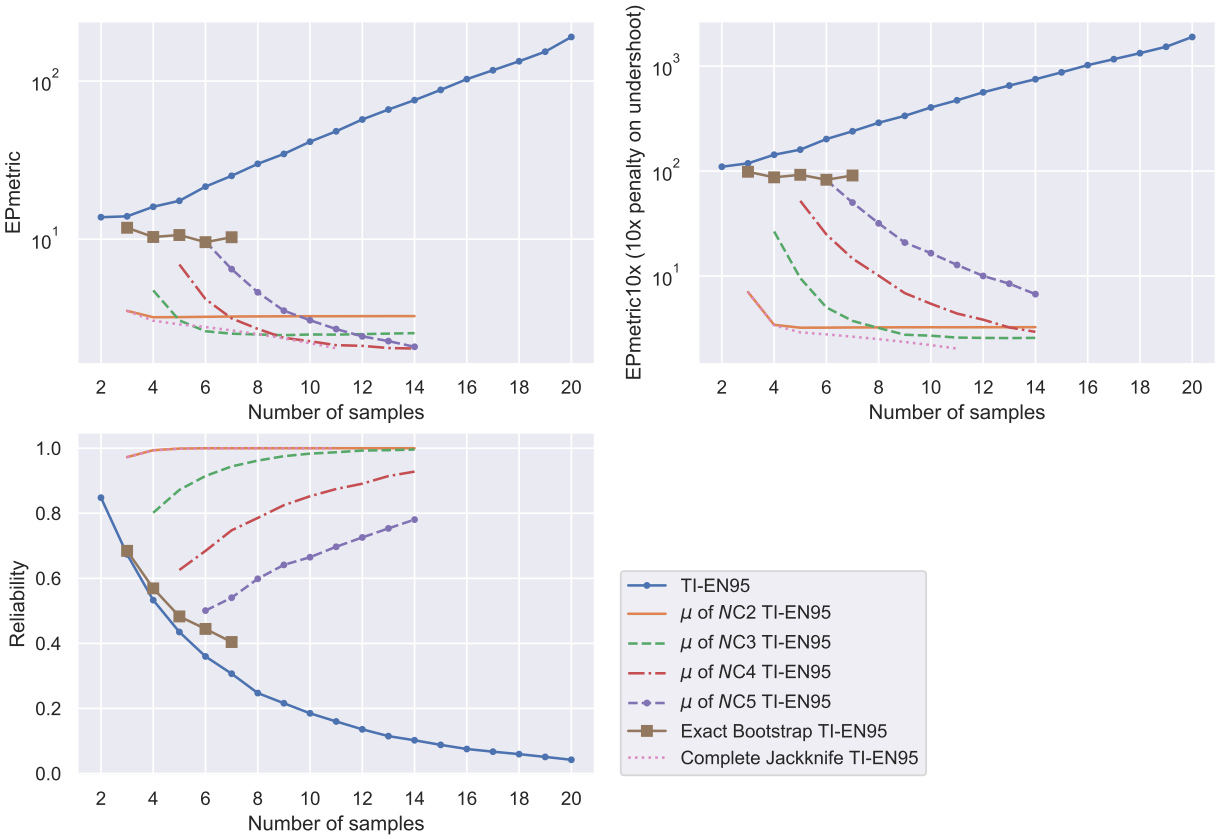


Figure 40: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-4}$.

TI-EN95 has significantly worse EPmetric performance than SD for $N \leq 16$. The lower-performing TI-EN95 method is improved by the resampling techniques proportionately even more than they improve SD-and even more so for this distribution than for the Student-t distribution, both in terms of reliability and EPmetric reliability+accuracy. Indeed, all the resampling results demonstrate higher accuracy and reliability than when using TI-EN95 alone, for the same total number of samples.

Unlike with Bootstrapping for the prior three cases (SD-B and TI-EN95-B on the t-distribution and SD-B on the Exponential Wide distribution), Bootstrapping significantly improved TI-EN95 reliability and EPmetric values compared to using TI-EN95 alone. However, Jackknife resampling, and Complete Jackknife in particular, performed better for a given number of samples. Although resampling improves TI-EN95's results proportionately more than it improves SD's, in absolute

terms SD-Jackknife's accuracy+reliability EPmetric results are best (compare TI-EN95 Complete Jackknife against SDJ *NC*5 at any *N*).

4.3.1.3 Weibull Narrow distribution

The SD results on the Weibull Narrow distribution are shown in Figure 41. Of 16 distributions studied in Section 3, the Weibull Narrow distribution[2] was the most difficult to predict tail probabilities for, given very limited data and the sparse-sample UQ methods tried. This is most evident with the low reliability levels $< 30\%$ for SD alone and $< 3\%$ TI-EN95 alone (see Figure 42). These low reliability levels are the backdrop against which any improvements from resampling are characterized.

SD related results are discussed first. Again, Exact Bootstrapping with SD produced results that were not meaningfully different from using SD alone. Reliability and accuracy are roughly the same with and without bootstrapping for any given number of samples. Also like the previous distributions, SD with Jackknifing always had higher reliability than SD alone for the same number of samples. For instance, the lowest number of samples that the SDJ method can be applied with had a reliability of about 37% whereas the SD-only reliability was about 5% for $N = 3$. This involves the optimal SDJ *NCr* variant, where $r = N_{\text{SDopt}} = 2$ for the EPmetric and this distribution. This variant approached a reasonable reliability of about 73% with $N = 14$ (the maximum *N* investigated). It is projected from the trend in the plot that a useful 80% reliability will occur with about $N = 17$ samples. This optimal variant has a strong trend of increasing reliability with increasing numbers of samples. This is in contrast to the other *NCr* SDJ variants or any of the other resampling methods, which do not yield strong reliability growth with added samples, and do not have reasonable reliability at any number of samples tried.

Concerning EPmetric reliability + accuracy performance, when the number of samples is increased beyond the SD-only optimum $N_{\text{SDopt}} = 2$ for this distribution, the performance quickly degrades. Conversely, performance quickly improves when SDJ resampling is used. Reliability and reliability + accuracy performance of *NC*5 to *NC*2 families of SDJ are consecutively better than SD over the full range $N = 3$ to 14 investigated, with the optimal *NC*2 variant performing vastly better over this range than the other *NCr* variants and than SD alone. Optimal SD alone ($N = 2$ samples) has better EPmetric performance than non-optimal *NC*3, *NC*4, and *NC*5 SDJ methods with substantially more samples (except for *NC*5 with ¿11 samples).

Reliability and reliability + accuracy performance of the Complete SDJ lies between that of the *NC*2 and *NC*3 methods. Like the *NC*2 results, the Complete SDJ combined performance is better (over the full range investigated, $N = 3$ to 14) than the optimal SD-only results (at $N = 2$). Unlike for the *NCr* SDJ methods, reliability of CSDJ declines as the number of total samples increases. This is the apparent cause for CSDJ's accuracy + reliability EPmetric reaching an optimum at about $N = 8$ samples, and then worsening with more samples. In contrast, the EPmetric for the

[2]This distribution was included because of the difficulty to accurately estimate an upper tail EP. The distribution is strictly positive, has a mean of 48, and a standard deviation of 760.4.

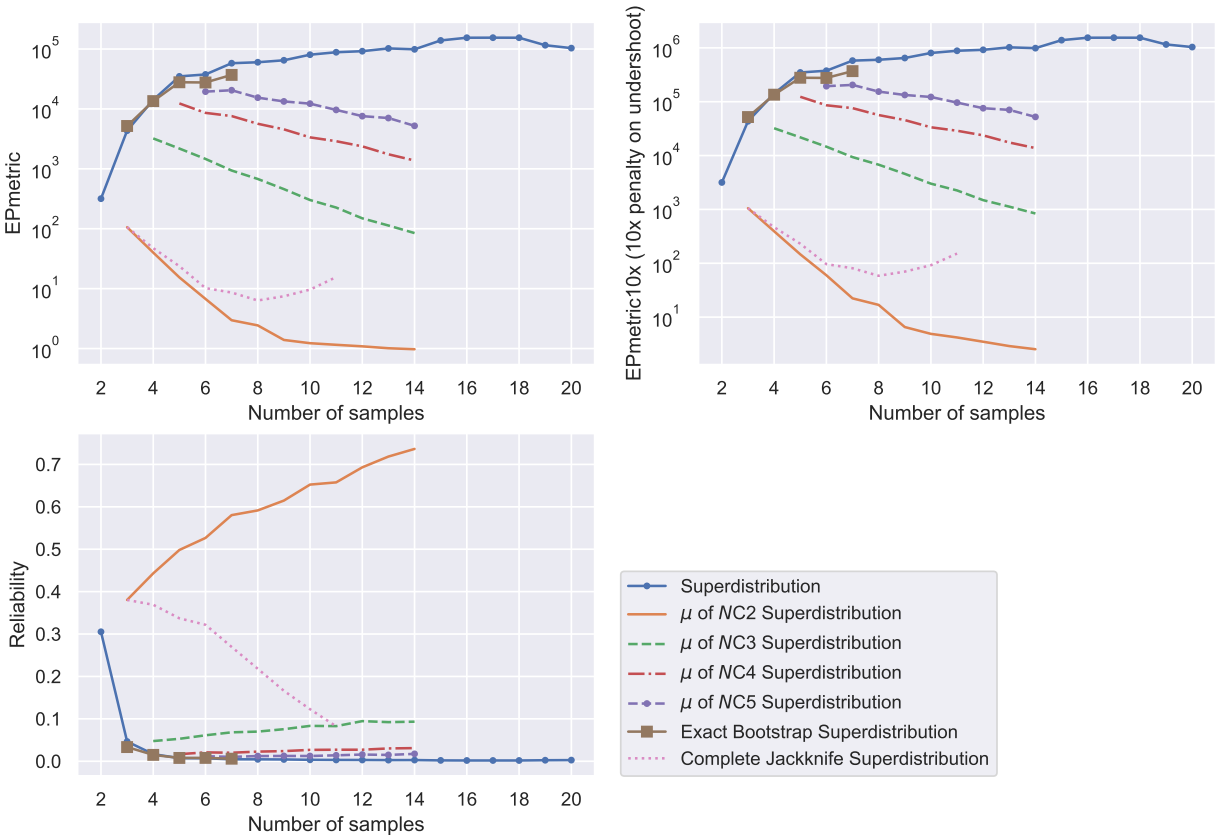


Figure 41: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the Weibull Narrow distribution for $EP = 10^{-4}$.

NCr methods all continually improve with increasing samples.

Results for TI-EN95 on the Weibull Narrow distribution are shown in Figure 42. The trends observed with SD-only apply here as well for TI-EN95-only. Starting from $N = 2$ samples the TI-EN95 reliability initially precipitously declines with added samples and asymptote to near-zero reliability for $N \geq 3$ like they do for SD with this distribution. Also, EPmetric performance immediately declines with added samples like for SD. However, in absolute terms the SD results are much better than the TI-EN95 results, for any given number of samples.

All the resampling methods demonstrate substantially better results than when using TI-EN95 alone (for the same number of samples). Moreover, the lower-performing TI-EN95 method (vs. SD) is improved by the resampling techniques proportionately more than SD is improved–both in terms of reliability and EPmetric reliability+accuracy. For instance, TI-EN95 with Exact Bootstrapping shows noticeable improvement of reliability and EPmetric results with increasing samples, whereas no appreciable improvement occurs for SD with Bootstrapping. Complete Jackknifing (CJ) also has a much more evident positive effect on TI-EN95 than on SD. TI-EN95-CJ reliability starts off better than TI-EN95-alone at $N = 3$ samples (the lowest allowable for CJ), and the TI-EN95-CJ increases with more samples. SD-CJ reliability also starts off better than SD-alone at $N = 3$ samples, but the SD-CJ reliability decreases with more samples. These different trends

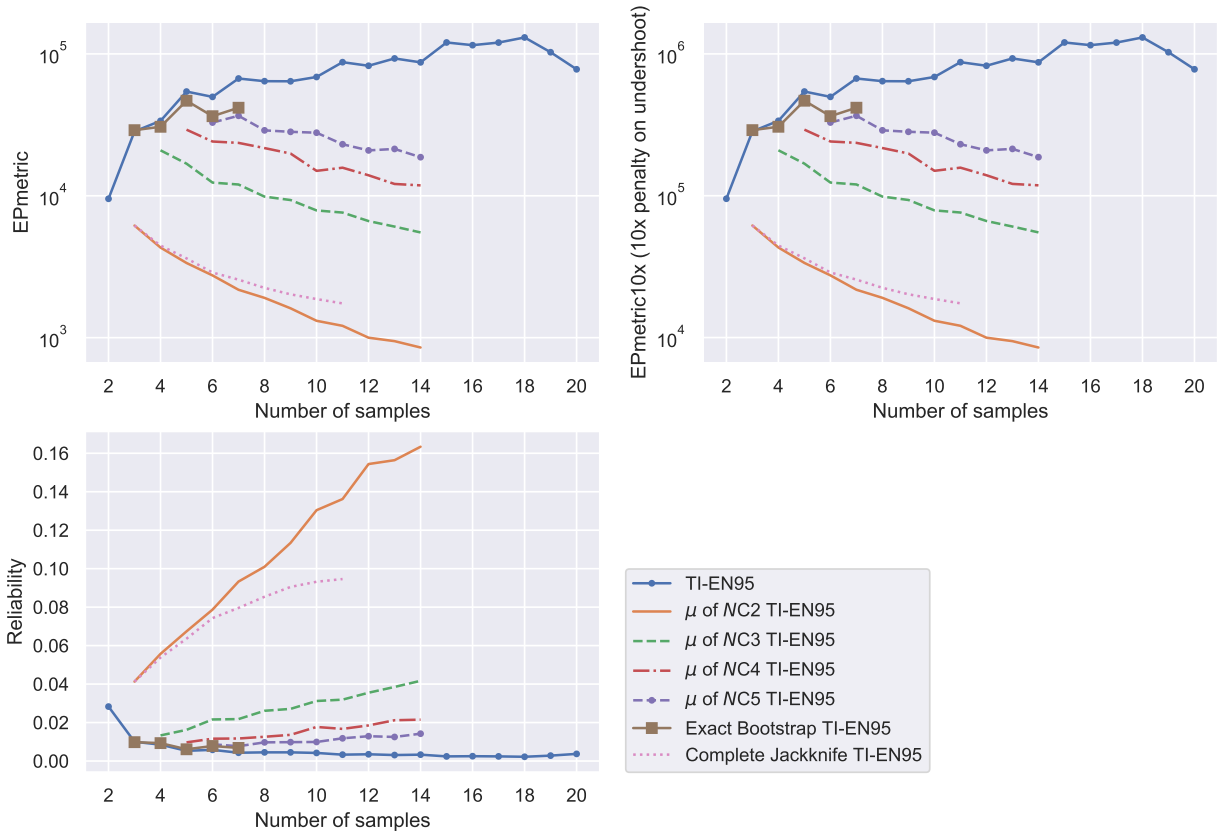


Figure 42: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the Weibull Narrow distribution for $EP = 10^{-4}$.

show up in TI-EN95-CJ EPmetric performance continually improving with added samples, while SD-CJ first improves then declines with added samples. Even so, the SD-B and SD-CJ results are in absolute terms much better than the respective TI-EN95-B and TI-EN95-CJ results for a given number of samples over the range investigated (except for CJ reliability at $N = 11$).

The best TI-EN95 resampling method for both reliability and EPmetric accuracy+reliability is TI-EN95-Jackknifing with optimal NCr subsample size $r = N_{\text{TIEN95opt}} = 2$ for this distribution. The TI-EN95-CJ results are next best, then the $NC3$, $NC4$, and $NC5$ TI-EN95-J results in that order. The rankings in the prior two sentences are also applicable to SD by replacing 'TI-EN95' by 'SD'. In absolute terms, the SD results are much better than the corresponding TI-EN95 results, for any given number of samples. This was the case for the prior two distributions as well.

4.3.2 Results on the easiest distributions

Using generalized Jackknifing with SD was shown to improve the accuracy + reliability (beyond SD alone) for predicting tail probabilities on the most difficult distributions if suitable values are used for the parameters of the Jackknife-SD method. This demonstrated that tail probability esti-

mation could improve as the number of samples is increased beyond what is optimal for SD alone. Here we consider how resampling affects SD performance on some of the distributions that SD performed best on in Section 3. These are the Standard Normal, bi-modal Log-Gamma Normal distribution, and the empirical distribution from the Tensile EQPS Lid Buckle Element 0.25 data.

The results for SD and TI-EN95 with and without resampling are presented in Figures 43-47. Results for the EON and TI-EN90 sparse-sample UQ methods with resampling are shown in Appendix I.

4.3.2.1 Standard Normal Distribution

Generalized *NCr* Jackknifing with SD was shown thus far to be useful for improving the EPmetric accuracy + reliability (beyond what was optimal for SD alone) when predicting tail probabilities on the previous difficult non-normal distributions. This has involved finding a suitable value for the *r* subsample size. However, it is not clear whether the *NCr* Jackknifing would improve the results on the Normal distribution using the TI-EN95 and SD methods. The reason is that the Normal distribution would be the ideal case to use the TI-EN95 or SD methods without Jackknifing, since the TI-EN95 and SD methods were developed considering the behavior of Normal distributions (and thus perform much better on the Normal distribution).

The SD results on the Standard Normal distribution are presented in Figure 43. There are some notable differences between the results on the Standard Normal distribution and the previous distributions. Most notably the EPmetric values for combined accuracy + reliability performance of SD-alone continually decrease/improve with added samples over the full range 2 to 20 studied. The *NCr* SD-Jackknife EPmetric values are all significantly worse than using SD alone, and get increasingly worse as samples increase. However, like with the previous distributions the reliability of the *NCr* methods was always higher than SD alone, for the same number of samples. In fact, all *NCr* reliabilities were about 100%. This higher reliability comes with an unwelcome tradeoff of lower accuracy (worse accuracy + reliability EPmetric value while reliability remains nearly perfect), because the reliabilities are already sufficiently high with SD alone ($> 87\%$ over the entire range from 2 to 20 samples). Note that the relatively poor EPmetric performance for *NCr* SDJ may be because the lowest EPmetric occurred at the highest sample sizes studied, *N*=19 and 20; no small-sample N_{SDopt} exists for this distribution. The order of performance of the *NC*2, *NC*3, *NC*4, *NC*5 methods correlates with the closer their subsample size *r* is toward 20.

Complete SD-Jackknife *does* improve in combined accuracy + reliability performance per the EPmetric as samples are added. Recall that the performance of SD-alone increased as the number of samples increased, so it is not unexpected that averaging-in these better estimates as total *N* increased would improve the Complete SD-Jackknife results. CSDJ also has reliabilities about 100% for the 3 to 11 total samples tried with this method. It's improving trend doesn't appear to reach the best *NCr* method on this distribution (*NC*5) until $N = 12$ samples, where it's EPmetric value is about the same as *NC*5's. It appears that 14 samples are required before the CSDJ trend reaches a better/lower EPmetric value than *NC*5's lowest/best value which occurs at *N*=6. In fact, for any *N* in the range 3 to 11 total samples investigated for Complete SD-Jackknife, an *NCr*

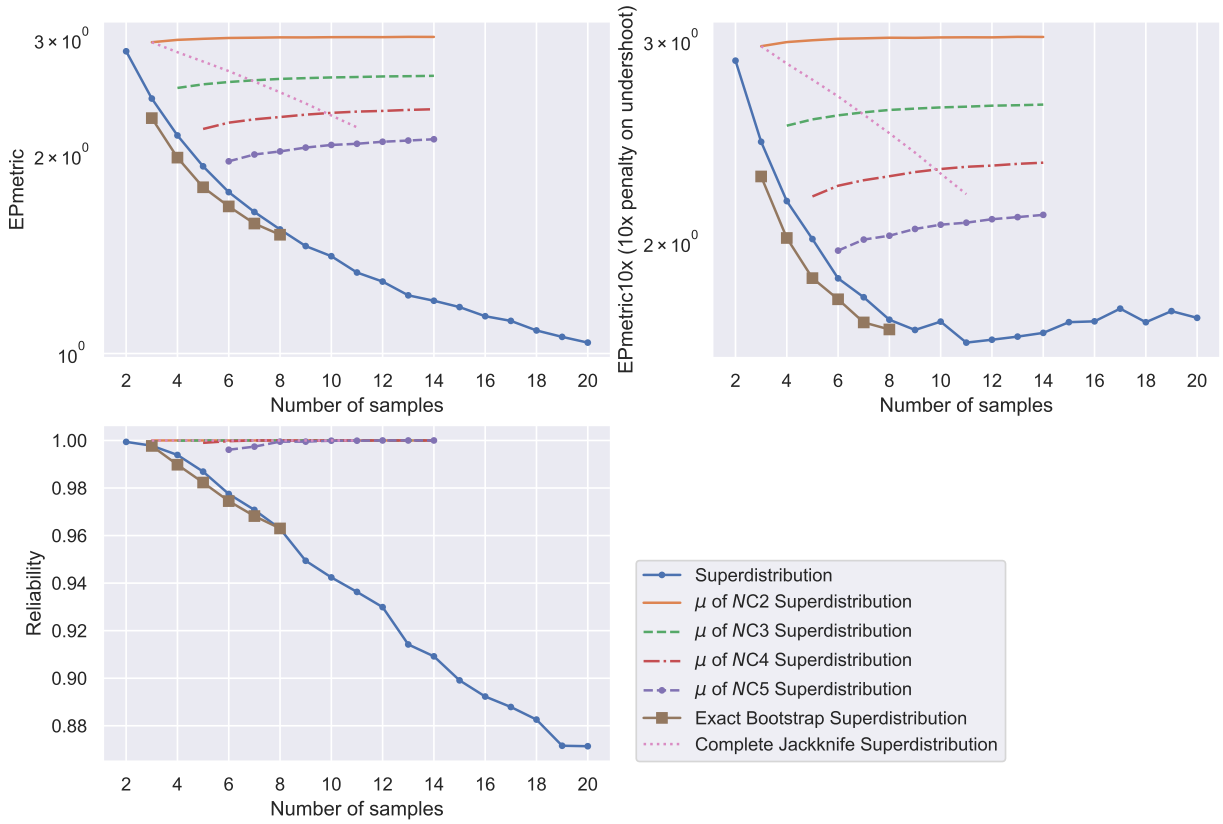


Figure 43: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the standard Normal distribution for $EP = 10^{-4}$.

SD-Jackknife method can be pointed-to that is more cost effective (or equally so at $N = 3$).

The Standard Normal results for TI-EN95 are shown in Figure **??**. The reliability of TI-EN95 was consistently around 95% (as expected from the 95% confidence level and the fact that the distribution is Normal), while the reliability of SD decreased from 100% to 87% as the samples increased from 2 to 20. Like for SD, TI-EN95's accuracy + reliability EPmetric continually improved with added samples. The lowest EPmetric value occurred at the highest sample size studied, N=20; no small-sample $N_{\text{TIEN95opt}}$ exists for this distribution. Overall, SD had better EPmetric performance than TI-EN95 for a given number of samples over the range 2 to 20 studied.

For both TI-EN95 and SD, reliability improved with Complete and *NCr* Jackknifing at the cost of worse combined accuracy + reliability EPmetric values. Reliability and combined reliability + accuracy results with SD Jackknifing are better than or as good as results with TI-EN95 Jackknifing. Unlike the behavior of the previous distributions, the EPmetric performance of TI-EN95 and SD methods improved with bootstrapping with little trade-off in reliability. In fact, TI-EN95 reliability improved slightly with bootstrapping.

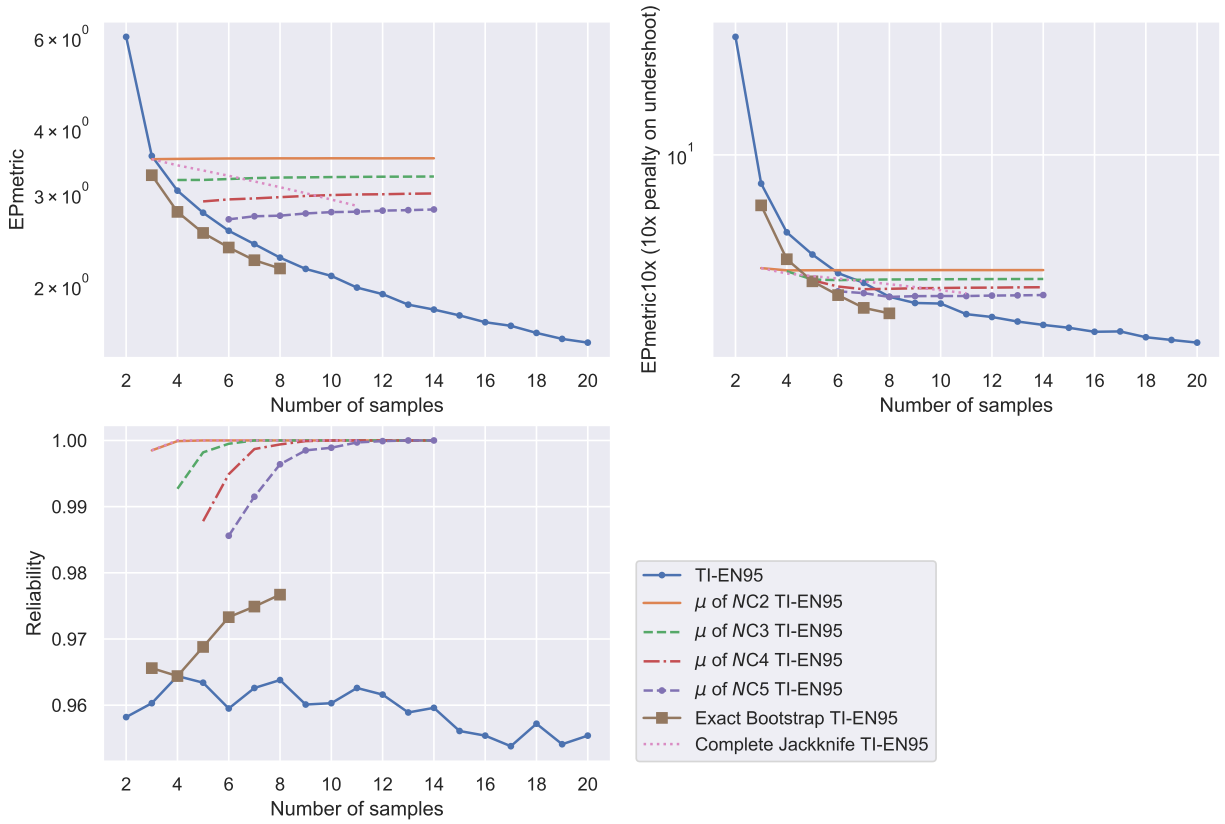


Figure 44: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the standard Normal distribution for $EP = 10^{-4}$.

4.3.2.2 Bi-modal Log-Gamma Normal and Tensile EQPS Lid Buckle Element 0.25

The results on the Bi-modal Log-Gamma Normal distribution are presented in Figures 45 and 46. The results on the empirical Tensile EQPS Lid Buckle Element 0.25 distribution are presented in Figures 47 and 48. These two distributions were some of the best performs for SD in Section 3. The overall trends established in 4.3.2.1 on the Standard Normal distribution appear to hold for the Bi-modal Log-Gamma Normal and empirical Tensile EQPS Lid Buckle Element 0.25 distributional, where Jackknife resampling appears to increase the EPmetric values (which is worse), but also increased the reliability (which is better) of SD alone.

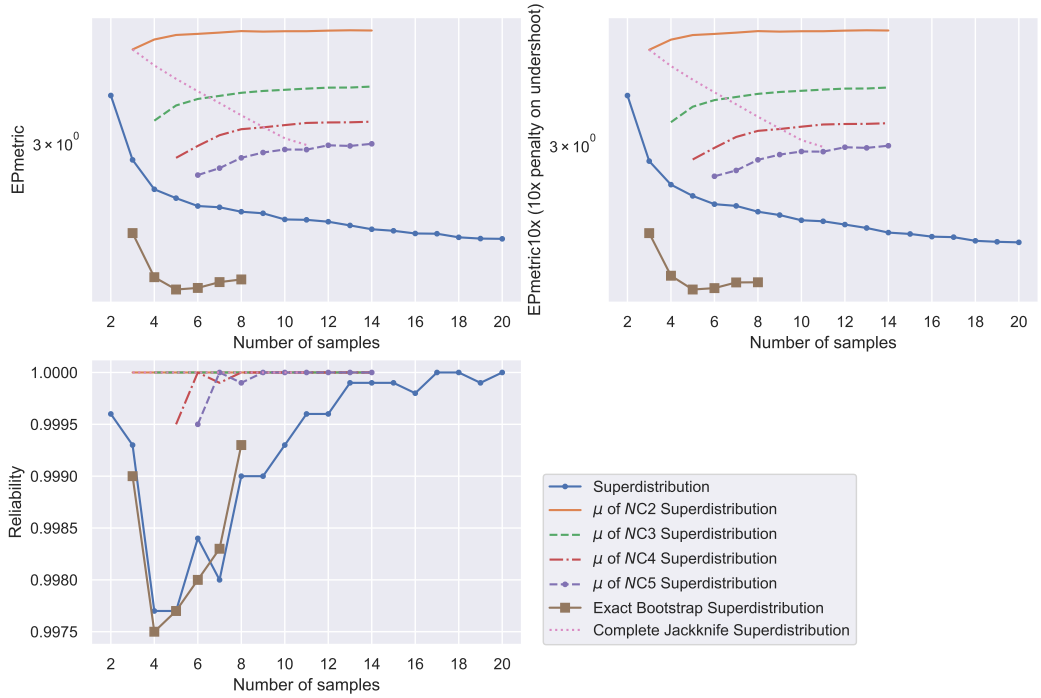


Figure 45: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the bi-modal Log-Gamma Normal distribution for $EP = 10^{-4}$.

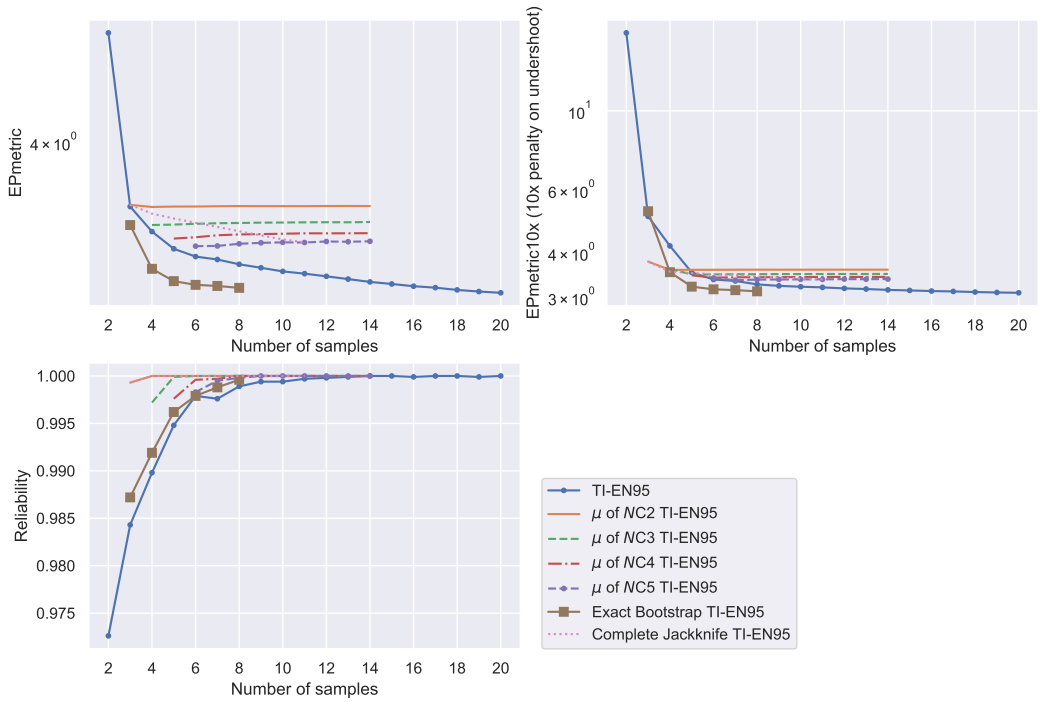


Figure 46: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the bi-modal Log-Gamma Normal distribution for $EP = 10^{-4}$.

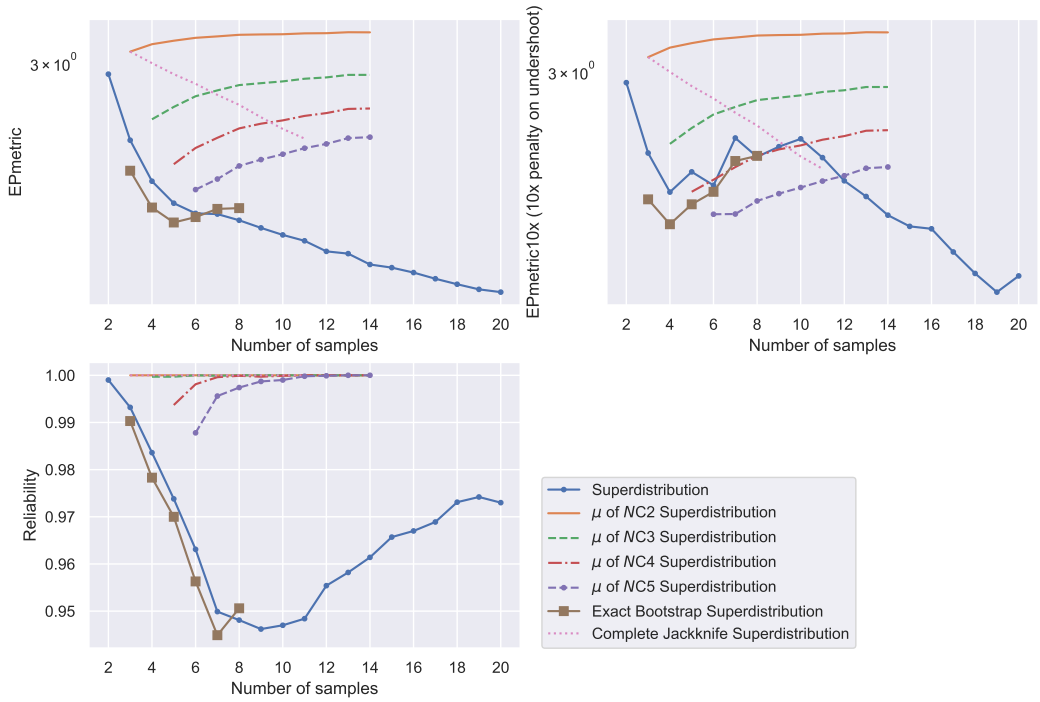


Figure 47: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the empirical Tensile EQPS Lid Buckle Element 0.25 distribution for $EP = 10^{-4}$.

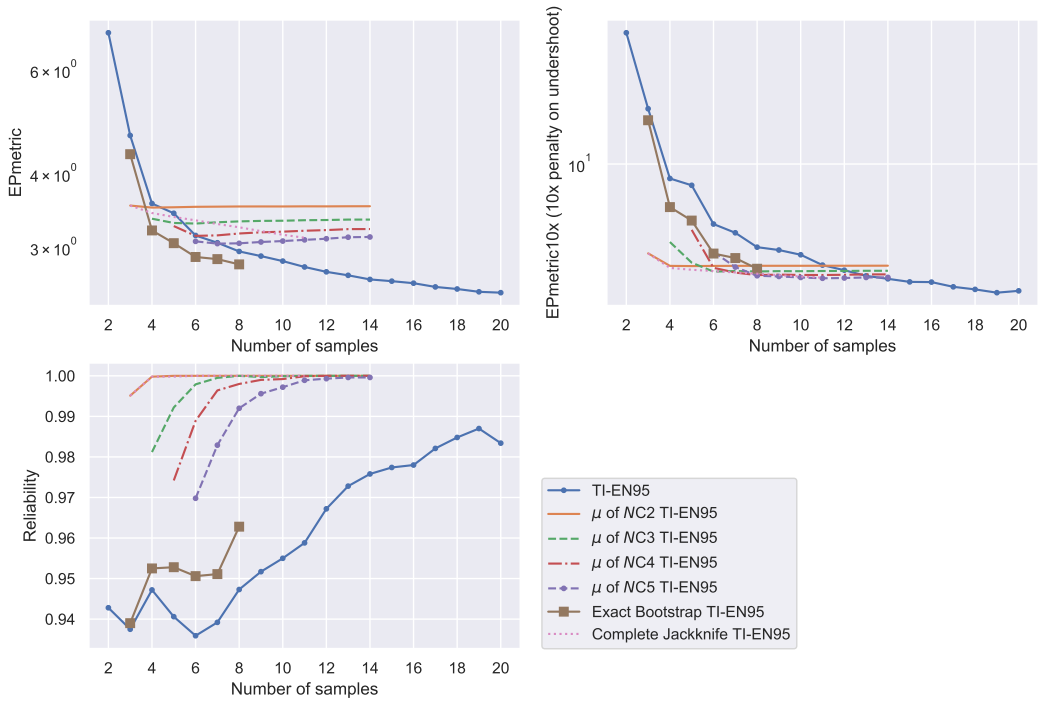


Figure 48: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the empirical Tensile EQPS Lid Buckle Element 0.25 distribution for $EP = 10^{-4}$.

4.4 Discussion and added context

Jackknifing (both Complete and *NCr* versions) was demonstrated to universally increase the reliability of attaining conservative tail-probability estimates (for any given number of samples and for all four distributions) compared to using SD or TI-EN95 methods alone. In the best cases, Jackknife resampling improved both reliability and accuracy of the EP estimates. In the worse cases it improved reliability at the cost of worsened accuracy (over-conservatism). This was true for resampling with both the SD and TI-EN95 sparse-sample UQ methods. Employing the methods with Bootstrap resampling reduced their reliability or did not increase their reliability as much as Jackknifing did, but occasionally gave better combined EPmetric reliability + accuracy than Jackknifing for the same number of total samples (but only rarely better than optimal *NCr* Jackknifing).

For the distributions and tail-probability magnitude of 10^{-4} studied above, SD-alone usually performed better than TI-EN95-alone in terms of reliability and combined reliability + accuracy EPmetric performance for a given number of samples. The optimal/best SD result was always better than the optimal/best TI-EN95 result according to the EPmetric. In broader terms, for 16 distributions including the four in this section and for EP magnitudes 10^{-1} to 10^{-5}, it is found in Section 3 that SD normally performs better than TI-EN methods with 90%, 95%, and 99.99% confidence settings, and one other sparse-sample UQ method tried.

A similar dynamic is found to carry over to SD and TIEN-95 when used with resampling in Section 3. On the four distributions and 10^{-4} EP magnitude studied above, SD with resampling normally had better reliability and EPmetric reliability + accuracy than TI-EN95 with resampling even though resampling improved the lower-performing TIEN-95 method proportionately more than it improved the SD method.

Distribution	$N_{\text{SDopt}} = r_{\text{SDopt}}$	SD reliability $N = N_{\text{SDopt}}$	SDJ reliability (NCr) $r = N_{\text{SDopt}}$ $N = N_{\text{SDopt}} + 1$	SDJ reliability (NCr) $r = N_{\text{SDopt}}$ $N = N_{\text{SDopt}} + 2$	SDJ reliability (NCr) $r = N_{\text{SDopt}} + 1$ $N = N_{\text{SDopt}} + 2$
5 DOF Student-t	5	0.78	0.84 (6C5)	0.87 (7C5)	No results for 7C6
Exponential Wide	4	0.82	0.89 (5C4)	0.92 (6C4)	0.68 (6C5)
Weibull Narrow	2	0.31	0.38 (3C2)	0.44 (4C2)	0.05 (4C3)
Standard Normal	4 see*	0.99	1.00 (5C4)	1.00 (6C4)	1.00 (6C5)

Table 2: Reliability results for the four distributions and SD NCr Jackknifing with various subsample sizes r. *It appears that $N_{\text{SDopt}} = \infty$ for the standard normal distribution, however 4 was chosen as a small sample stand-in.

For most of the 16 distributions studied in Section 3, SD and TI-EN95 have optimum sample sizes N_{SDopt} and $N_{\text{TIEN95opt}}$ where EPmetric values of combined accuracy + reliability performance are lowest/best for each method. Performance worsens with added samples beyond the optimal number unless resampling is used. The most cost effective resampling method in this situation is found to be *NCr* Jackknifing with a "rule of thumb" optimal (or approximately so) subsample size $r = N_{\text{SDopt}}$ or $N_{\text{TIEN95opt}}$ for the corresponding method. See the associated reliability improvement with one added sample from the third to the fourth column in Table 2 for the better-performing SD variant of methods. (Note that the Normal distribution does not necessarily break this "rule". It simply has EPmetric optima at an asymptotically large number of samples, so the rule is not practically implementable or testable for the Normal distribution.)

It is also observed that for the better-performing SDJ variant, the minimum number of samples needed for the rule-of-thumb NCr_{SDopt} Jackknifing is all that is needed for many foreseeable engineering purposes. Additional samples are not highly cost effective. They do not improve the EPmetric value significantly and sometimes even worsen it (except for appreciable improvement for the Weibull Narrow distribution, which is presumed to be an outlier based on outlier results from SD-only and TI-EN-only methods applied to this and 15 other distributions in Section 3, including a Weibull with different parameters values). Reliabilities are already reasonably high with the minimum number of NCr_{SDopt} Jackknifing samples (except for the pathological Weibull Narrow distribution). Reliabilities do improve with more samples, but maybe not be enough to justify the added sampling expense-especially with expensive experiments or simulations. The 4^{th} and 5^{th} columns of Table 2 present results on this point.

Note also that using a *NCr* subsample size $r > r_{\text{opt}}$ should be avoided because this can significantly decrease the reliability of *NCr* Jackknifing. This is exemplified in the last column of Table 2 for the Exponential and Weibull distributions. For the same total number of samples underlying the results in columns 5 and 6, the reliabilities in the last column (6) are substantially lower than those in column 5. In fact, the results in column 6 are substantially lower than those in columns 3 and 4 with fewer total samples. Column 3 shows that even SD alone without Jackknifing and with two fewer samples (1/3 to 1/2 fewer) yields better reliabilities than the column 6 example of *NCr* Jackknifing with subsample size $r > r_{\text{SDopt}}$. The much lower reliabilities in the latter cases also show up in worse combined reliability + accuracy EPmetric values in the plots in Figures 8 and 10. Similar dynamics exist for TI-EN95 variants of the methods.

A different type of deleterious effect occurs when the subsample size is $r < r_{\text{SDopt}}$. In this case, *NCr* SD Jackknifing attains very high reliabilities of ≈ 1 for any number of samples studied here (2 to 20) for the three distributions for which $r < r_{\text{SDopt}}$ is possible. (This excludes the Weibull Narrow distribution.) Essentially perfect reliability is a sign of over-conservatism. This is reflected in combined accuracy + reliability EPmetric values that are inferior, over a broad range of samples, to performance obtained with SD alone or NCr_{SDopt} SD Jackknifing in the regime of small numbers of samples (3 to 6) relevant for expensive tests or simulations. The plots in Figures 6, 8, and 12 show this.

In general, SD or TI-EN95 with *NCr* Jackknifing and non-optimal subsample size can perform less well in terms of combined EPmetric performance than other resampling methods or SD or TI-EN95 alone, for a given number of total samples. Reliability alone is not hurt by Jackknife re-

sampling; SD and TI-EN95 reliabilities are, for a given number of total samples, always improved by any of the Jackknifing methods, including non-optimal *NCr* Jackknifing.[3] But the improved reliability usually comes at the price of worse accuracy in the form of over-conservatism. Therefore, getting the most out of *NCr* Jackknifing requires knowledge of the optimal subsample size for a given distribution and EP magnitude. This is mapped-out in Section 3 for 16 diverse distributions including the distributions in this section and for EP magnitudes 10^{-1} to 10^{-5} (for SD and several other sparse-sample UQ methods). The vast majority of optimum subsample sizes are in the range $N = 2$ to 6 for the best method, SD.

Unfortunately, in real tail-probability estimation problems the exact distribution shape and order of EP magnitude are not known precisely, if at all. Even if the information were available, it may be that the affordable [experimental] sampling budget does not allow the optimum subsample size to be reached for optimal *NCr* Jackknifing.

Some of these *NCr* Jackknifing difficulties can be eased by using Complete Jackknifing. It is a mostly parameter-free resampling technique (with a qualification in the last sentence of this paragraph). This averages all possible *NCr* Jackknife results obtained from all possible r subsample sizes given a total number of samples N. This makes the method more robust to lack of knowledge of the particular optimum subsample size r_{opt} in a problem. The *NCr* Jackknifing EP estimates with $r < r_{\text{opt}}$ will contribute conservative bias relative to the SD or TI-EN95-only estimate and even relative to the (usually) reliable/conservative NCr_{opt} estimate (e.g., in column 4 of Table 2 for SD). On the other hand, *NCr* Jackknifing EP estimates using $r > r_{\text{opt}}$ will contribute non-conservative bias. The latter wins out when Complete SDJ is applied to the pathological Weibull Narrow distribution (see Figure 10) as the total number of samples increases and larger over-sized subsamples $r > r_{\text{opt}}$ are admitted. For the other three distributions, no similar ill effects are apparent up to the highest number of samples tried (11). Indeed, the reliabilities are ≈ 1 over the CSDJ range of samples investigated (3 to 11) and the EPmetric accuracy + reliability trends look like they will eclipse NCr_{opt} Jackknifing at slightly higher numbers of samples, 12 to 14. Nonetheless, in most cases there will be an upper limit to the number of samples that can be used before the described ill effects occur with Complete Jackknifing.

Going forward, we concentrate on SD and SD with resampling because of their broad advantage over TI-EN95 and other sparse-data methods tried to date. Given what we have learned thus far, we would minimize the chances of the said ill effects for any problem with unknown distribution and EP magnitude by limiting the total number of samples used with Complete Jackknifing SD to, say, ≤ 7. This would limit EP under-estimation risk to what is anticipated to be a very low level. (See also the results and discussion below for initial data on CSDJ performance for and 10^{-2} EP magnitudes.) Risk will be even smaller if fewer than 7 samples are used or are affordable. For the distributions we have applied CSDJ analysis to, reliabilities are ≈ 1 for the three non-pathological distributions in this section and two others in Section 3, for an EP magnitude 10^{-4}. This occurs over the range 3 to 11 samples tried.

[3] This statement does not conflict with the discussion around Table 2. Reliability of *NCr* Jackknifing with $r > r_{\text{opt}}$, so $N > (r_{\text{opt}} + 1)$, can yield lower reliability than SD or TI-EN95 alone when allowed a different/lower number of samples $\leq r_{\text{opt}}$.

This conservative strategy is driven by not having analyzed the many other cases mentioned above, and also the realistic ambiguity in any given application problem of not knowing the distribution shape or EP order of magnitude. The downside of this lack of knowledge and therefore the conservatively biased strategy is that over-conservatism likely prevails. Indeed, Complete Jackknifing yielded worse combined reliability + accuracy EPmetric values than Jackknifing over the range of samples tried with it (2 to 11) in all cases studied in this section but the one in Fig. 48. This conservatism could manifest, for example, in an EP estimate of 10^{-3} while it is really several orders smaller, like 10^{-6}. Quantification over the broader data base of 16 diverse distributions and 10^{-1} to 10^{-5} EP magnitudes in Section 3 would help determine the reliability/risk vs. accuracy/-conservatism of using various numbers of samples with CSDJ. It would therefore help identify the best number of samples to use for an appropriate risk-reward balance in a given project.

Bootstrap resampling performed sometimes marginally better and sometimes marginally worse than SD or TI-EN95 without resampling, both in terms of reliability and combined reliability + accuracy by the EP metric. Bootstrap resampling did not yield substantial reliability gains like Jackknifing did for a given total number of samples.

A final consideration in this section is how the resampling methods' performance varies with EP magnitude. For SD-alone, a general trend in Section 3 across the 16 PDFs is that reliability and reliability + accuracy by the EPmetric get worse as EP magnitude increases, for any fixed number of samples. To get an indication of how this trend affects SD when coupled with resampling, we assess this on 10^{-1} and 10^{-2} EPs for the challenging Exponential Wide distribution. We also evaluate TI-EN95 with resampling for these conditions because TI-EN95 shows general improving trends as EP magnitude increases, such that TI-EN95 performance generally eclipses SDs for EP= 10^{-1}. The question is whether TI-EN95 + resampling is also preferable to SD + resampling for EP= 10^{-1} and possibly $= 10^{-2}$.

The SD and TI-EN95 results for EP=10^{-1} are shown in Figures 49 and 50. Note that if resampling is not used, TI-EN95's reliability is always significantly better than SD's for the same number of samples. SD reliability is never above 77% and quickly drops to 0.5 at $N = 4$ samples and worse thereafter, while TI-EN95's reliability is a desirably high ≥ 0.8 up to $N = 4$ samples and never drops below 70% even at 20 samples. SD's EP metric performance is better than TI-EN95's for $N \leq 3$ samples, but this may be mute for many foreseeable engineering purposes because of SD's undesirable reliability level even with $N \leq 3$ samples.

Bootstrapping reduces TI-EN95's reliability considerably, while improving it's reliability + accuracy EPmetric performance slightly. Bootstrapping also reduces SD's reliability, and doesn't help it's EPmetric performance or even makes it worse, depending on the number of samples.

Jackknife resampling significantly improves both methods' reliabilities and EP metric performance. We concentrate on Complete Jackknifing results applicable for the most common case where both distribution and EP magnitude are unknown. CJ with SD yields somewhat better EP metric performance than CJ with TI-EN95 for $N < 8$, but this is mute in many cases because of SD's undesirable reliability level that has a maximum of ≈ 0.76 at $N = 2$ and decreases with more samples. In contrast, TI-EN95-CJ reliability is > 0.92 for all samples tried (3 to 11). TI-EN95-CJ is therefore deemed the generally preferable method for EP magnitude 10^{-1}. This is consistent

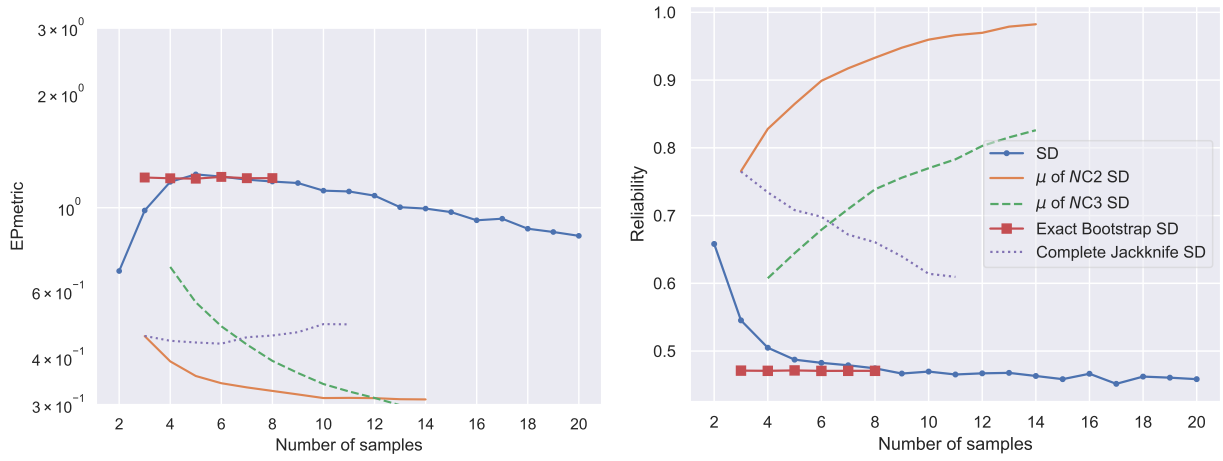


Figure 49: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-1}$.

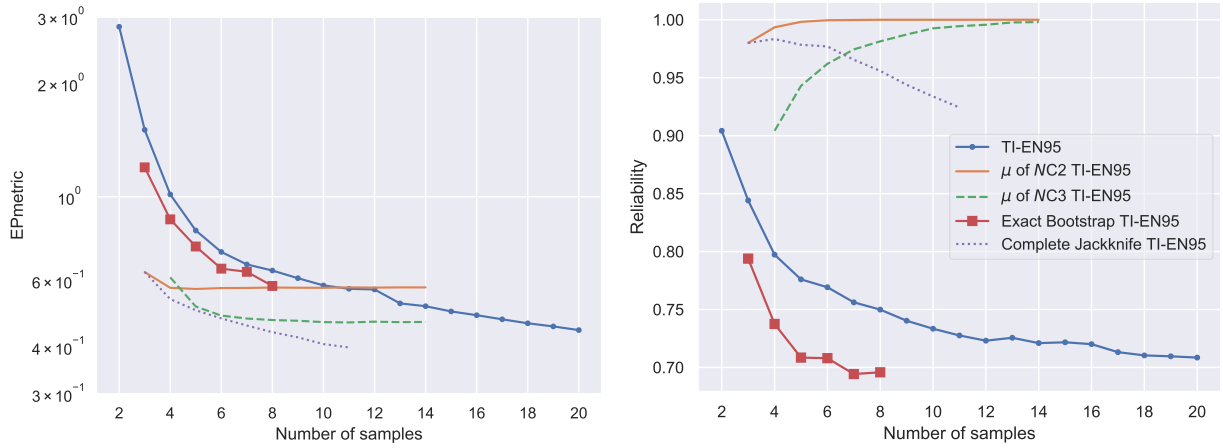


Figure 50: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-1}$.

with the no-resampling case where TI-EN95 is found generally preferable to SD for EP 10^{-1} over the 16 PDFs tested in Section 3.

The EP= 10^{-2} results for SD and TI-EN95 are shown in Figures 51 and 52. If resampling is not used, TI-EN95's reliability is significantly better than SD's except at $N = 2$, where SD reliability is slightly higher, at 90%. It then drops quickly to 65% at $N = 3$ samples and worse thereafter. TI-EN95's reliability is also undesirable (77%) for $N = 3$ samples and gets worse with more samples. Thus, TI-EN95's reliability on the same distribution gets worse with a decrease in EP magnitude from 10^{-1} to 10^{-2}, while SD's reliability gets better (although still not acceptably high with this challenging distribution except for $N = 2$). These trends reflect the general improving (SD) or declining (TI-EN95) reliability trends with decreasing EP magnitude generally found for the 16 PDFs in Section 3.

Bootstrapping does not have much of an impact on TI-EN95 reliability, but helps it's reliability + accuracy EPmetric performance significantly. Bootstrapping generally increases SD's reliability substantially, which also improves its EPmetric performance substantiallymore than TI-EN95's is improved.

Jackknife resampling improves both methods' reliabilities and EP metric performance, and significantly more than Bootstrapping did. Complete Jackknife with SD yields better EPmetric performance than CJ with TI-EN95 for all sample numbers tried (3 to 11). SD CJ has desirably high reliability $> 80\%$ for $N \leq 8$. TI-EN95 CJ has higher reliability ($> 95\%$) over this range of N, but also has lower reliability + accuracy performance than SD CJ over this range. It seems that both SD CJ and TI-EN95 CJ are competitive in the present circumstances. This is reminiscent of the situation with SD-only vs. TI-EN95-only comparisons in Section 3 for the 16 PDFs and EP$= 10^{-2}$. SD was usually deemed preferable, but TI-EN95 was competitive or better for a few PDFs.

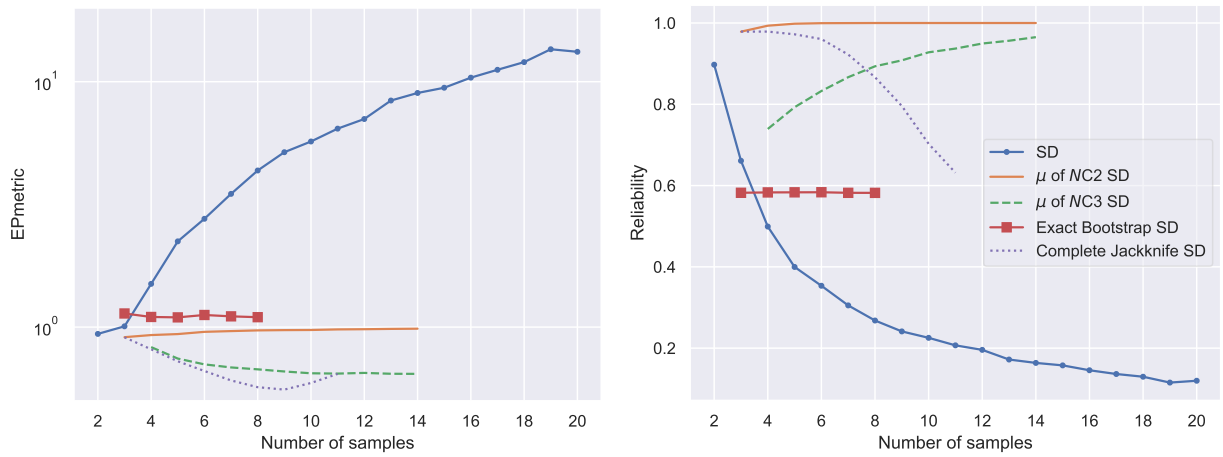


Figure 51: The EPmetric, EPmetric10x, and reliability for the Superdistribution with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-2}$.

In many real situations we can use engineering judgment and prior experience and knowledge with designs or systems approximately resembling the current one in order to reasonably dismiss order 10^{-1} exceedance probability levels as very remote possibilities. (In real engineering situations, risk must be managed but cannot be entirely eliminated if one wants to survive in a cost-constrained and competitive environment, or in order not to ridiculously over-design such that "the plane won't be able to fly because it is too heavy.") In situations where EPs as high as order 10^{-1} are not considered realistic possibilities, one might use 10^{-2} as the lower reasonable limit for determining the general viability and use of sparse-sampling approaches. We take this perspective in providing the following recommendations.

The studies in Section 3 and the current section indicate that SD with Complete Jackknifing appears to be the most robust and accurate/efficient estimation method for reliably conservative tail probability estimates when unknown distribution shapes and EP magnitudes $\leq 10^{-2}$ are involved. Reliabilities of $\geq 90\%$ can normally be expected with CSDJ when ≤ 7 samples are used. As samples are added beyond the minimum 3 for Complete Jackknifing, reliability can decrease for

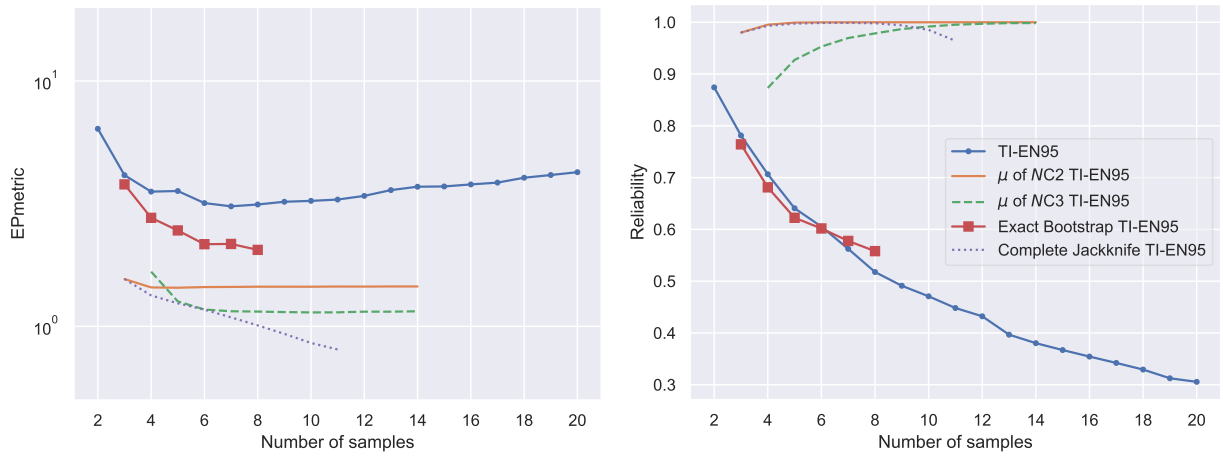


Figure 52: The EPmetric, EPmetric10x, and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-2}$.

the more difficult distributions, but will normally remain above 0.9 with ≤ 7 samples. The benefit of using more samples than the minimum 3 is that the conservative estimation bias that achieves the high reliabilities decreases acceptably with added samples (for N total ≤ 7), such that the combined accuracy + reliability of CSDJ according to the EPmetric gets better as samples are added. This reduces egregious over-conservatism that might otherwise exist. If only 2 samples are available or affordable, SD-alone would be used. This gives reliabilities $> 80\%$ for $EPs \leq 10^{-2}$ for 15 of the 16 PDFs studied in Section 3 (the spoiler being the Weibull Narrow distribution).

Further quantification of Complete and NCr Jackknifing performance with SD and TI-EN over the broader matrix of 16 diverse distributions and 10^{-1} to 10^{-5} EP magnitudes in Section 3 would help more granularly quantify the reliability/risk vs. accuracy/conservatism of using various numbers of samples with the methods. It would therefore help identify the best method and number of samples to use for an appropriate risk-reward objective in a given project.

Finally, both Bootstrapping and Jackknifing can be used to construct a distribution of EP estimates (that are averaged for use in this work). This is perhaps one of the most important features of resampling techniques that was not investigated in this work. Only the average was used in this study, but it may be desirable to work with the distribution of EP estimates in the most risk averse applications. Instead of using the mean of the EP estimates, a more conservative EP estimate may come from the 90^{th} percentile or other high percentile of the distribution of EP estimates. This may be particularly useful if the distribution is known or suspected to be unlike any distribution for which the resampling method's performance has been characterized a-priori.

4.5 Conclusion

This work demonstrated that resampling can be used to improve the performance of the SD and TI-EN95 sparse-sample UQ methods in terms of reliability of attaining a conservative EP estimate,

and accuracy of the estimate. There is generally an optimal number of samples N_{opt} beyond which the performance of the SD and TI-EN95 methods worsen. However, when the methods are paired with resampling techniques, the performance can improve substantially. While Bootstrapping didn't offer much improvement, the *NCr* Jackknife with optimal subsample size $r = r_{\text{opt}} = N_{\text{SDopt}}$ paired with SD showed the most improvement. (SD is broadly found to perform better than TI-EN95 and other sparse-sample UQ methods investigated, whether the methods are used alone or with resampling.)

However, SD *NCr* Jackknifing without the optimal subsample size can perform less well than SD alone for a given number of samples-in terms of combined reliability + accuracy per our EP-metric, although reliability alone always improves with Jackknifing. Because combined reliability + accuracy performance suffers with non-optimal subsample size, and optimal subsample size varies with distribution shape and the EP magnitude involved, and these are normally unknown in real problems, a more robust approach was sought. Complete *NCr* Jackknifing is less parameter-dependent and it yielded higher reliability than non-optimal *NCr* Jackknifing and better combined reliability + accuracy when the *NCr* subsample size is highly non-optimal, as could be for a real problem.

An upper constraint on the total number of samples that can be beneficially used with the Complete Jackknife exists for the reasons explained in the previous section. We have given preliminary guidance on what is anticipated to be a very conservative "safe" upper limit for use of CSDJ. Useful future work would involve refining this guidance by applying CSDJ and analyzing its performance on our full test suite of 16 diverse distributions and 10^{-1} to 10^{-5} EP magnitudes in Section 3.

5 Summary and Conclusions

This work examined the performance of various sparse-sample UQ methods for conservatively but not overly-conservative bounding tail probabilities of magnitudes 10^{-1} to 10^{-5} on a test matrix of 16 diverse distribution shapes. Except for tail probabilities of magnitude 10^{-1}, SD normally performs better with optimal numbers of samples than do TI-EN methods with 90%, 95%, and 99.99% confidence settings and the Ensemble of Normals 90th percentile method when these other methods are used with their optimal numbers of samples. The optimal numbers of samples for SD were often less than for the other methods, making SD typically the most economical method as well.

A similar dynamic carries over when these methods are combined with resampling methods when tested on six of the distributions and 10^{-4} EP magnitude. SD with resampling normally had better reliability and EPmetric reliability + accuracy than TI-EN and EON methods with resampling for the same number of samples, even though resampling improved the lower-performing methods proportionately more than it improved SD.

Resampling often improved performance of SD and the other methods in terms of reliability of attaining a conservative EP estimate, and accuracy of the estimate. There is typically an optimal number of samples N_{opt} beyond which the performance of each sparse-sample method worsens. However, when the methods are paired with resampling techniques, the performance can improve substantially. While the exact case variant of Bootstrapping we studied didn't offer much improvement, the *NCr* Jackknife with optimal subsample size $r = r_{\text{opt}} = N_{\text{SDopt}}$ paired with SD showed the best performance for the tested EP magnitude 10^{-4}. (This conclusion may not extend down to EP magnitude 10^{-1}, as the TI-EN methods paired with resampling may be best in that EP regime, as they are in this regime without resampling.)

However, SD *NCr* Jackknifing without the optimal subsample size can perform worse than SD alone for a given number of samples, although reliability alone always improves with Jackknifing. Because combined reliability + accuracy performance suffers with non-optimal subsample size, and optimal subsample size varies with distribution shape and the EP magnitude involved, and these are normally unknown in real problems, a more robust approach was sought. Complete *NCr* Jackknifing is less parameter-dependent, which yielded higher reliability and better combined reliability + accuracy than non-optimal *NCr* Jackknifing. Complete Jackknifing may be an attractive method for real problems where the optimal sub-sample size is unknown.

An upper constraint on the total number of samples that can be beneficially used with the Complete Jackknife exists for the reasons explained in the previous sections. In the Section 4.5 we give preliminary guidance on what is anticipated to be a very conservative "safe" upper limit for use of CSDJ (Complete Superdistribution Jackknifing). For EPs $\leq 10^{-2}$ we cautiously project that 2 to 7 samples with SD and Complete Jackknifing will yield $> 90\%$ reliability of conservative EP estimation for 15 of the 16 diverse distributions in our test suite. The more samples that can be afforded in this range, the more efficient the estimate will be (shaving off conservatism) without reducing reliability below 90%. For handling EPs between 10^{-2} and EPs $\leq 10^{-1}$ we anticipate that a TI-EN method with Complete Jackknifing would perform best. The best TI-EN confidence

level to use would need to be determined from empirical investigation. Useful future work would involve confirming and/or refining this guidance by applying the methods and quantifying their performance on our full test suite of 16 diverse distributions and 10^{-1} to 10^{-5} EP magnitudes.

In some follow-on work [7] just before publication of this SAND report, a reviewer observed that the Superdistribution is essentially the same as a Bayesian posterior predictive distribution (PPD) based on assumptions consistent with those that underlie the SD's construction. The Bayesian PPD has an analytic form of a non-standard t-distribution, see e.g. [6]. A brief discussion is also given in [7]. It is somewhat easier to calculate tail probabilities with the analytic form than with the SD approach (though this is not difficult either). The more important implication is that it is useful and reassuring to know that this equivalency exists (we have confirmed it empirically). This corroborates the SD method. It is also valuable to see that the SD can be constructed from relatively simple frequentist concepts and sampling approaches, and the result is equivalent to the Bayesian PPD that has substantial mathematical-statistical development behind it and which ultimately yields a simple analytic expression for the distribution.

The reviewer also suggested a variant of the Bayesian PPD approach that was anticipated to replace the effect of Complete Jackknife resampling and also be simpler. Initial investigations explained in [7] found significant merit to this approach for some distributions, numbers of samples, and EP magnitudes. However, the proposed method did not perform as well as CJK resampling in other cases. The reader may want to contact the authors of this SAND report for access to the follow-on work [7].

The reviewer also suggested a different variant of Bootstrapping than the exact case variant we used. The reviewer's variant is briefly discussed in Section 4.1. This is generally a more popular variant of Bootstrapping, so we agree it would be important to try. An initial investigation may make into ref. [7]. Please contact the authors if interested.

References

[1] Vicente Romero, Matthew Bonney, Benjamin Schroeder, and V. Gregory Weirs. Evaluation of a class of simple and effective uncertainty methods for sparse samples of random variables and functions. Technical Report SAND2017-12349, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), November 2017.

[2] Vicente J. Romero and V. Gregory Weirs. A class of simple and effective UQ methods for sparse replicate data applied to the cantilever beam end-to-end UQ problem. AIAA Non-Deterministic Approaches Conference, January 8-12, 2018. doi: 10.2514/6.2018-1665. URL https://arc.aiaa.org/doi/abs/10.2514/6.2018-1665.

[3] Vicente J. Romero, Benjamin B. Schroeder, James F. Dempsey, Nicole L. Breivik, George E. Orient, Bonnie R. Antoun, John R. Lewis, and Justin G. Winokur. Simple effective conservative treatment of uncertainty from sparse samples of random variables and functions. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 4(4):041006–041006–17, April 2018. ISSN 2332-9017. doi: 10.1115/1.4039558. URL http://dx.doi.org/10.1115/1.4039558.

[4] Vicente Romero, Laura Swiler, Angel Urbina, and Josh Mullins. A comparison of methods for representing sparsely sampled random quantities. Technical Report SAND2013-4561, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), September 2013.

[5] Charles F. Jekel and Vicente Romero. Bootstrapping and Jackknife Resampling to Improve Sparse-Sample UQ Methods for Tail Probability Estimation. ASME 2019 Verification and Validation Symposium, Las Vegas, Nevada, USA, May 15-17, 2019. doi: 10.1115/VVS2019-5127. URL https://doi.org/10.1115/VVS2019-5127.

[6] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, chapter 3 introduction to multiparameter models. Chapman and Hall/CRC, 3rd edition, 2013.

[7] Charles F. Jekel and Vicente Romero. Conservative and efficient tail probability estimation from sparse sample data. Sandia National Laboratories document in revision for submission to ASME J. Verification Validation and Uncertainty Quantification.

[8] Gerald J. Hahn and William Q. Meeker. *Statistical Intervals: A Guide for Practitioners*, pages 288–352. Wiley-Blackwell, 2011. ISBN 9780470316771. doi: 10.1002/9780470316771.app1. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316771.app1.

[9] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John Wiley & Sons, 2010.

[10] W. G. Howe. Two-sided tolerance limits for normal populationssome improvements. *Journal of the American Statistical Association*, 64(326):610–620, 1969. doi: 10.1080/01621459.1969.10500999. URL https://doi.org/10.1080/01621459.1969.10500999.

[11] Derek Young. tolerance: An r package for estimating tolerance intervals. *Journal of Statistical Software, Articles*, 36(5):1–39, 2010. ISSN 1548-7660. doi: 10.18637/jss.v036.i05. URL https://www.jstatsoft.org/v036/i05.

[12] I. Miller and J.E. Freund. *Probability and Statistics for Engineers. Premtoce-Hall.* Inc, 1985.

[13] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1295, 12 1986. doi: 10.1214/aos/1176350142. URL https://doi.org/10.1214/aos/1176350142.

[14] Victor Picheny, Nam Ho Kim, and Raphael T. Haftka. Application of bootstrap method in conservative estimation of reliability with limited samples. *Structural and Multidisciplinary Optimization*, 41(2):205–217, Mar 2010. ISSN 1615-1488. doi: 10.1007/s00158-009-0419-8. URL https://doi.org/10.1007/s00158-009-0419-8.

[15] Joanna Kisielinska. The exact bootstrap method shown on the example of the mean and variance estimation. *Computational Statistics*, 28(3):1061–1077, Jun 2013. ISSN 1613-9658. doi: 10.1007/s00180-012-0350-0. URL https://doi.org/10.1007/s00180-012-0350-0.

[16] M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3-4):353–360, 1956. doi: 10.1093/biomet/43.3-4.353. URL http://dx.doi.org/10.1093/biomet/43.3-4.353.

[17] John Tukey. Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614, 1958.

[18] W. R. Schucany, H. L. Gray, and D. B. Owen. On bias reduction in estimation. *Journal of the American Statistical Association*, 66(335):524–533, 1971. doi: 10.1080/01621459.1971.10482296. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1971.10482296.

[19] M. C. Jones and P. J. Foster. Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, 3(1):81–94, 1993. doi: 10.1080/10485259308832573. URL https://doi.org/10.1080/10485259308832573.

[20] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, Dec 1997. ISSN 1573-2916. doi: 10.1023/A:1008202821328. URL https://doi.org/10.1023/A:1008202821328.

A Definition of analytical distributions

This section describes the analytical distributions investigated in this report. Each subsection contains the analytical equation of the PDF for each distribution. Figures will show the PDF of each distribution with the locations of exceedance probability (EP) for $P = 10^{-3}, 10^{-4}, 10^{-5}$. The Python code used to create the frozen distribution object assumes the following import of the SciPy library with minimum version of 1.0.0.

```
from scipy import stats
```

A.1 Standard Normal distribution

The PDF of the Standard Normal distribution is defined as

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \tag{A.1}$$

for a random variable x. The PDF and EP locations are shown in Figure 8 for the Standard Normal distribution . The distribution can be called in Python using the following code.

```
my_dist = stats.norm()
```

A.2 Student's t-distribution

The PDF of the Student's t-distribution is defined as

$$f(x;d) = \frac{\gamma((d+1)/2)}{\gamma(d/2)(1+\frac{x^2}{d})^{(d+1)/2}\sqrt{d\pi}} \tag{A.2}$$

for a random variable x where d represents the degrees of freedom. The PDF and EP locations are shown in Figure 17 for the 5 degree of freedom Student's t-distribution. The same 5 degrees of freedom distribution was used in [1], which is created using the following Python code.

```
my_dist = stats.t(df=5)
```

A.3 Log-Normal distribution

The PDF of the Log-Normal distribution is defined as

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(x)-\mu)^2}{2\sigma^2}\right) \tag{A.3}$$

for a random variable x, and given mean μ and standard deviation σ. The same Log-Normal distribution parameters used in [1] were used in this study with $\mu = 10.48, \sigma = 0.314$. The PDF and EP locations are shown in Figure 13 for this Log-Normal distribution. The Python code to create the distribution is found below.

```
import numpy as np
my_dist = stats.lognorm(s=0.314, scale=np.exp(10.48))
```

A.4 Weibull Wide distribution

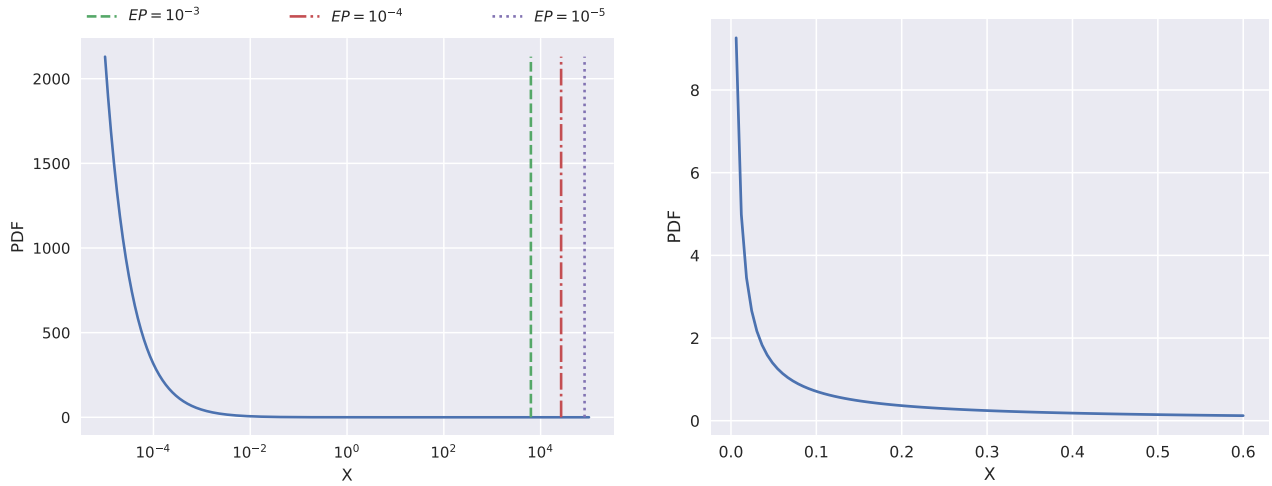
The Weibull Wide distribution is defined as

$$f(x;\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{(\alpha-1)} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \tag{A.4}$$

which has two parameters α and β. The first Weibull Wide distribution used in this report was the same used in [1] with $\alpha = 1.3$ and $\beta = 1.0$, and the PDF can be seen in Figure H.11. A second Weibull Wide distribution was also used, and was referred to as the Weibull Narrow distribution. This Weibull Narrow distribution used $\alpha = 0.2$ and $\beta = 0.4$, and the PDF can be seen in Figure A.2. The Python code to freeze these two distributions follows.

```
my_weibull = stats.weibull_min(c=1.3)

my_weibull_narrow = stats.weibull_min(c=0.2, scale=0.4)
```



EP locations on log scale

Close up linear scale

Figure A.2: Weibull Narrow distribution.

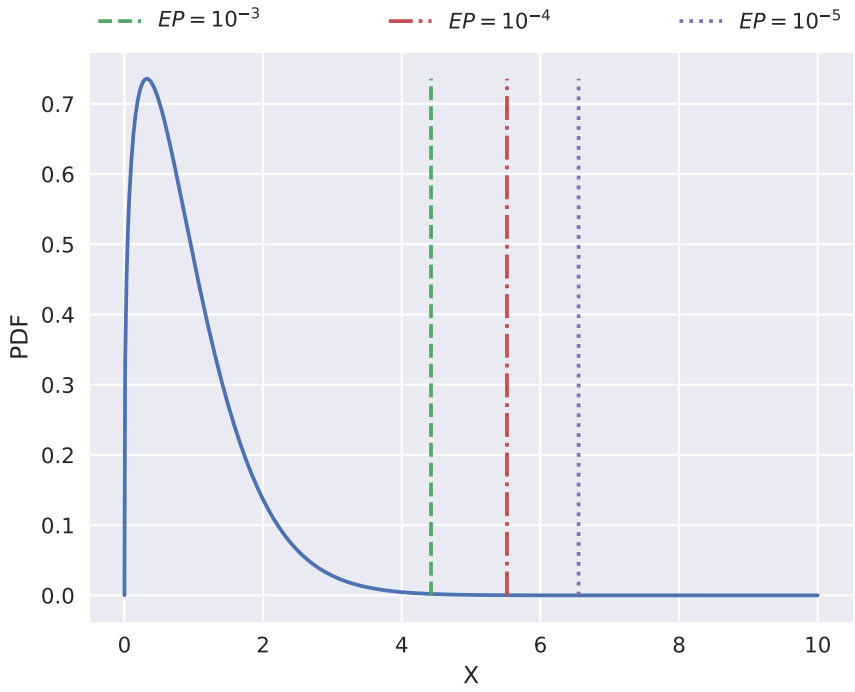


Figure A.1: Weibull Wide distribution.

A.5 Exponential distribution

The PDF of an Exponential Narrow distribution is defined as

$$f(x;\lambda) = \lambda \exp(-\lambda x) \tag{A.5}$$

for a random variable x and a single parameter λ. The first Exponential Narrow distribution used $\lambda = 2.0$, and the PDF can be seen in Figure H.14. An additional Exponential distribution was used that had a wider tail with $\lambda = 0.5$, and is refereed to as the Exponential Wide distribution. The PDF of the Exponential Wide distribution is seen in Figure H.15. The Python code to create the frozen Exponential distributions follows.

```
my_expon = stats.expon(scale=1.0/2.0)

my_expon_wide = stats.expon(scale=1.0/0.5)
```

A.6 Bi-modal Log-Gamma Normal distribution

The bi-modal distribution consists of a left Log-Gamma distribution with a right Normal distribution, and is shown in Figure 29. The Probability Density Function (PDF) of this distribution is defined as

$$f(x;c,l_1,s_1,l_2,s_2) = \frac{3}{2}\left[\frac{\exp\left(c(x-l_1)-\exp(x-l_1)\right)}{2\gamma(c)s_1} + \frac{\exp(-(x-l_2)^2/2)}{\sqrt{2\pi}s_2}\right] \tag{A.6}$$

for parameters c, l_1, s_1, l_2, s_2. A python implementation is shown below, which includes the chosen distribution's parameters.

```
class loggammanorm(stats.rv_continuous):
    def _pdf(self, x, c1, loc1, scale1, loc2, scale2):
        p1 = stats.loggamma.pdf(x, c=c1, loc=loc1, scale=scale1)
        p2 = stats.norm.pdf(x, loc=loc2, scale=scale2)
        return (0.5*p1 + (p2)) / (1.5)
    def _cdf(self, x, c1, loc1, scale1, loc2, scale2):
        p1 = stats.loggamma.cdf(x, c=c1, loc=loc1, scale=scale1)
        p2 = stats.norm.cdf(x, loc=loc2, scale=scale2)
        return (0.5*p1 + (p2)) / (1.5)
# create distribution named mylgn
mylgn = loggammanorm(shapes='c1, loc1, scale1, loc2, scale2')
# freeze mylgn distribution as my_dist with the parameters
my_dist = mylgn(c1=0.85, loc1=615, scale1=12.5, loc2=680, scale2=10)
```

B Empirical distributions

The same eight quantities covered in [1] were investigated in this report. The data originates from various quantities of interest on a can crush solid mechanics computational model that was considered in [3]. Each quantity of interest has 1,000 samples of a random variable. Kernel Density Estimation (KDE) was used to find an empirical distribution to the data. Optimal KDE bandwidths were selected by maximizing the likelihood of a 10-fold cross validation using Differential Evolution [20]. A python example to perform this type of optimization to find optimal KDE bandwidth is shown in Listing 1. The KDE was then fit using the optimal bandwidth to the 1,000 samples. The CDF of the KDE was used to find the locations where the EP integrates to $10^{-3}, 10^{-4}, 10^{-5}$. Sets of 10,000 random samples were then generated from each KDE for samples sizes ranging from $N = 2, \cdots, 20$. The resulting KDEs and EP locations are shown in the lower half of Figure 1.

Listing 1: Python implementation to maximize the 10-fold cross validation log-likelihood to find optimal KDE bandwidth.

```
import numpy as np
from scipy.optimize import differential_evolution
from sklearn.neighbors import KernelDensity
from sklearn.model_selection import cross_val_score
def my_fun(x):
    # fits a KDE for the specified bandwith
    kde_skl = KernelDensity(bandwidth=x[0])
    # returns the 10-fold CV log-likelihood score
    scores = cross_val_score(kde_skl, rv[:, np.newaxis], cv=10)
    # we want to maximize this score, but our optimization
    # algorithm minimizes thus return the negative of the
    # log-likelihood score
    return -np.sum(scores)
# set the np random seed for reproducibility
np.random.seed(121)
# my random variable as 1d array
rv = np.random.normal(size=100)
bounds = np.zeros((1, 2))
bounds[0, 0] = 1e-3
bounds[0, 1] = 1e3
# find the optimal bandwidth that maximizes the 10-fold CV
# log-likelihood score
res = differential_evolution(my_fun, bounds=bounds, polish=True,
                             maxiter=1000, tol=0.0001)
print('Optimal 10fold CV bandwidth:', res.x)
```

C Tolerance Interval f and k factors

The f factor for Tolerance Intervals (TI) can be calculated according to the formulas in [10, 11]. A Python function is provided in Listing 2, which uses the formulas to calculate the f factor for a user input coverage, confidence, and number of samples. For ease of future use, a set of high precision f factors are presented in Table C.1 for TI 95/85, 95/90, 95/95, 95/99, and 95/99.99, with number of samples ranging from $N = 2$ to $N = 40$.

Listing 2: Python function to calculate TI f factor for arbitrary confidence and coverage

```
import numpy as np
from scipy import stats
def calc_two_sided_f(P, alpha, n):
    """
    calculate the two sided tolerance interval factor from
    Howe 1969 as defined by Young 2010
    Input:
    P (float) = Coverage, 95% coverage would input P=0.95
    alpha (float) = Confidence level, 90% confidence would input alpha=0.9
    n (integer, float) = number of samples
    Output:
    f (float) = the two sided confidence interval factor
    Example:
    # Calculate the f factor for 95% coverage, 90% confidence,
    # N=10 samples
    f = calc_two_sided_f(0.95, 0.9, 10)
    """
    alpha = 1.0 - alpha
    mychi2 = stats.chi2.ppf(alpha, df=n-1)
    my_zp = stats.norm.ppf((1+P)/2.0)
    u = my_zp*np.sqrt(1.0 + (1.0/n))
    v = np.sqrt((n-1)/mychi2)
    w = np.sqrt(1.0 + ((n-3-mychi2)/(2*(n+1)**2)))
    f = u*v*w
    return f
```

A Python function is provided in Listing 3, which uses the formulas to calculate the k factor for TI-EN. The user must input the desired confidence and number of samples. For ease of future use, a set of high precision k factors are presented in Table C.2 for TI-EN with confidence levels of 85%, 90%, 95%, 99%, and 99.99%.

Listing 3: Python function to calculate TI-EN k factor for arbitrary confidence and coverage

```
import numpy as np
from scipy import stats
def calc_sigma_en_k(alpha, n):
    """
    calculate k which is used to find sigma_en
    Input:
    alpha (float) = Confidence level, 90% confidence would input alpha=0.9
    n (integer, float) = number of samples
    Output:
    k (float) = sigma_en factor
    Example:
    # Calculate the k factor for 90% confidence,
    # N=10 samples
    f = calc_two_sided_f(0.9, 10)
    """
    alpha = 1.0 - alpha
    mychi2 = stats.chi2.ppf(alpha, df=n-1)
    u = np.sqrt(1.0 + (1.0/n))
    v = np.sqrt((n-1)/mychi2)
    w = np.sqrt(1.0 + ((n-3-mychi2)/(2*(n+1)**2)))
    k = u*v*w
    return k
```

Number of samples	Coverage % / Confidence % 95/85	95/90	95/95	95/99	95/99.99
N =2	12.3222709	18.5557462	37.1977962	186.126675	18613.2404
N =3	5.58532625	6.94934488	9.97676878	22.5678781	226.310781
N =4	4.25795412	4.98558037	6.43997271	11.2990840	53.0788410
N =5	3.68951956	4.19064285	5.13892747	7.97051653	25.8187092
N =6	3.36962859	3.75672711	4.46461776	6.43705831	16.7571922
N =7	3.16252718	3.48117995	4.05027812	5.56264748	12.5461262
N =8	3.01650564	3.28947941	3.76853861	4.99820335	10.1870177
N =9	2.90746319	3.14772314	3.56368823	4.60327959	8.70026350
N =10	2.82259966	3.03822418	3.40749474	4.31096272	7.68497765
N =11	2.75446276	2.95082635	3.28410738	4.08543992	6.95036254
N =12	2.69840727	2.87926838	3.18392853	3.90584282	6.39521693
N =13	2.65138063	2.81947266	3.10079947	3.75919639	5.96130408
N =14	2.61129063	2.76866533	3.03058209	3.63701031	5.61291496
N =15	2.57665312	2.72489119	2.97039001	3.53349462	5.32701294
N =16	2.54638471	2.68673075	2.91814659	3.44456406	5.08810644
N =17	2.51967515	2.65312744	2.87231826	3.36725264	4.88541138
N =18	2.49590580	2.62327788	2.83174698	3.29935402	4.71119837
N =19	2.47459574	2.59655980	2.79554126	3.23919226	4.55978736
N =20	2.45536514	2.57248338	2.76300321	3.18547096	4.42691320
N =21	2.43790988	2.55065753	2.73357836	3.13717114	4.30931230
N =22	2.42198336	2.53076601	2.70682033	3.09348050	4.20444564
N =23	2.40738347	2.51255019	2.68236557	3.05374331	4.11030869
N =24	2.39394285	2.49579642	2.65991482	3.01742420	4.02529795
N =25	2.38152168	2.48032654	2.63921942	2.98408163	3.94811559
N =26	2.37000219	2.46599076	2.62007094	2.95334806	3.87769987
N =27	2.35928443	2.45266217	2.60229329	2.92491499	3.81317386
N =28	2.34928301	2.44023250	2.58573665	2.89852147	3.75380684
N =29	2.33992447	2.42860877	2.57027269	2.87394521	3.69898503
N =30	2.33114531	2.41771066	2.55579079	2.85099564	3.64818906
N =31	2.32289027	2.40746842	2.54219510	2.82950840	3.60097645
N =32	2.31511107	2.39782115	2.52940210	2.80934094	3.55696788
N =33	2.30776533	2.38871542	2.51733868	2.79036899	3.51583625
N =34	2.30081564	2.38010417	2.50594052	2.77248370	3.47729799
N =35	2.29422890	2.37194576	2.49515083	2.75558932	3.44110600
N =36	2.28797567	2.36420320	2.48491924	2.73960122	3.40704398
N =37	2.28202969	2.35684353	2.47520093	2.72444434	3.37492175
N =38	2.27636746	2.34983726	2.46595584	2.71005188	3.34457138
N =39	2.27096785	2.34315791	2.45714810	2.69636413	3.31584403
N =40	2.26581186	2.33678169	2.44874545	2.68332759	3.28860732

Table C.1: Table of high precision f TI factors for 95% coverage and confidence levels between 85% and 99.99%.

Number of samples	Confidence % 85	90	95	99	99.99
$N=2$	6.28698844	9.46739141	18.9788161	94.9643330	9496.72577
$N=3$	2.84970861	3.54564927	5.09028169	11.5144351	115.466806
$N=4$	2.17246549	2.54371019	3.28576074	5.76494474	27.0815389
$N=5$	1.88244253	2.13812238	2.62194995	4.06666479	13.1730529
$N=6$	1.71922985	1.91673273	2.27790806	3.28427377	8.54974497
$N=7$	1.61356393	1.77614485	2.06650640	2.83813760	6.40120246
$N=8$	1.53906177	1.67833666	1.92275911	2.55015061	5.19755355
$N=9$	1.48342685	1.60601071	1.81824169	2.34865520	4.43899152
$N=10$	1.44012833	1.55014286	1.73854967	2.19951119	3.92097901
$N=11$	1.40536397	1.50555131	1.67559578	2.08444642	3.54616850
$N=12$	1.37676370	1.46904147	1.62448318	1.99281357	3.26292574
$N=13$	1.35277008	1.43853289	1.58206962	1.91799258	3.04153756
$N=14$	1.33231562	1.41261031	1.54624376	1.85565160	2.86378475
$N=15$	1.31464310	1.39027616	1.51553296	1.80283651	2.71791369
$N=16$	1.29919974	1.37080618	1.48887766	1.75746294	2.59602037
$N=17$	1.28557217	1.35366132	1.46549543	1.71801761	2.49260263
$N=18$	1.27344473	1.33843168	1.44479541	1.68337482	2.40371680
$N=19$	1.26257205	1.32479975	1.42632277	1.65267948	2.32646487
$N=20$	1.25276034	1.31251564	1.40972142	1.62527015	2.25867069
$N=21$	1.24385443	1.30137980	1.39470846	1.60062693	2.19866913
$N=22$	1.23572850	1.29123088	1.38105616	1.57833538	2.14516475
$N=23$	1.22827944	1.28193692	1.36857901	1.55806093	2.09713480
$N=24$	1.22142186	1.27338892	1.35712434	1.53953043	2.05376118
$N=25$	1.21508441	1.26549598	1.34656526	1.52251860	2.01438170
$N=26$	1.20920701	1.25818167	1.33679545	1.50683792	1.97845465
$N=27$	1.20373867	1.25138125	1.32772506	1.49233099	1.94553262
$N=28$	1.19863580	1.24503946	1.31927764	1.47886466	1.91524276
$N=29$	1.19386095	1.23910888	1.31138771	1.46632552	1.88727194
$N=30$	1.18938171	1.23354852	1.30399885	1.45461634	1.86135515
$N=31$	1.18516987	1.22832279	1.29706215	1.44365326	1.83726664
$N=32$	1.18120082	1.22340062	1.29053499	1.43336355	1.81481288
$N=33$	1.17745293	1.21875475	1.28438007	1.42368381	1.79382697
$N=34$	1.17390710	1.21436118	1.27856457	1.41455849	1.77416423
$N=35$	1.17054646	1.21019864	1.27305953	1.40593875	1.75569859
$N=36$	1.16735598	1.20624829	1.26783924	1.39778141	1.73831969
$N=37$	1.16432226	1.20249328	1.26288082	1.39004817	1.72193049
$N=38$	1.16143331	1.19891859	1.25816385	1.38270494	1.70644532
$N=39$	1.15867836	1.19551070	1.25367003	1.37572126	1.69178825
$N=40$	1.15604770	1.19225746	1.24938288	1.36906984	1.67789171

Table C.2: Table of high precision k TI-EN factors for confidence levels between 85% and 99.99%.

D Exceedance probability

Provided a continuous random variable X, the probability that the random variable will exceed a value x of interest is defined as

$$P(X > x) = S(x) \tag{D.1}$$

where $S(x)$ is the survival function evaluated at the value x of interest. The value of the survival function at a value x is also referred to in the report as the exceedance probability (EP) associated with the response level x. This can be thought of as the area under the probability density function (PDF) of X integrated from x to ∞ (Figure D.3).

Alternatively, the survival function is simply

$$S(x) = 1 - F(x) \tag{D.2}$$

where $F(x)$ is the cumulative distribution function (CDF) of X evaluated at x.

For example, let X come from the Standard Normal distribution. The probability that X exceeds 1.645 is calculated as

$$P(X > 1.645) = S(1.645) \tag{D.3}$$
$$= 0.05 \tag{D.4}$$

which results in a 5% chance. The area under the PDF for this example is illustrated in Figure D.3. The shaded area is equal to 0.05.

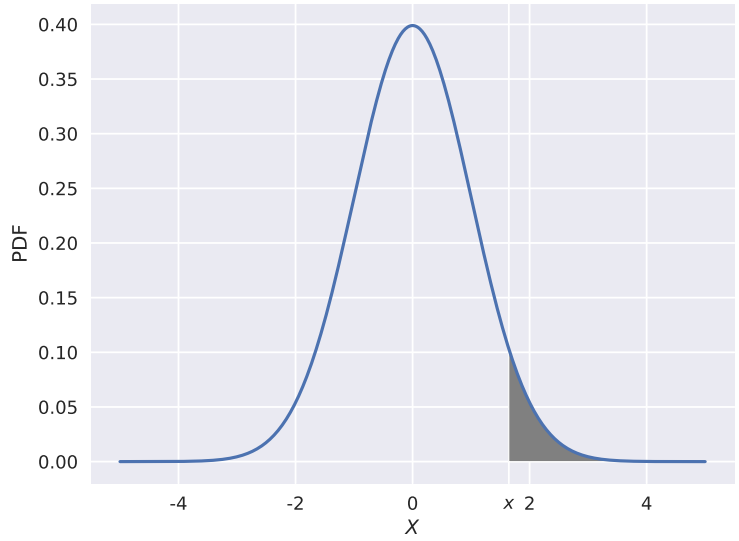


Figure D.3: The probability that a random variable X exceeds x is the same as the area under the probability density function (PDF) from x to ∞. In the example here when $x = 1.645$ and X follows the Standard Normal distribution, the shaded area is equal to 0.05.

E Demonstrated convergence for EON and SD

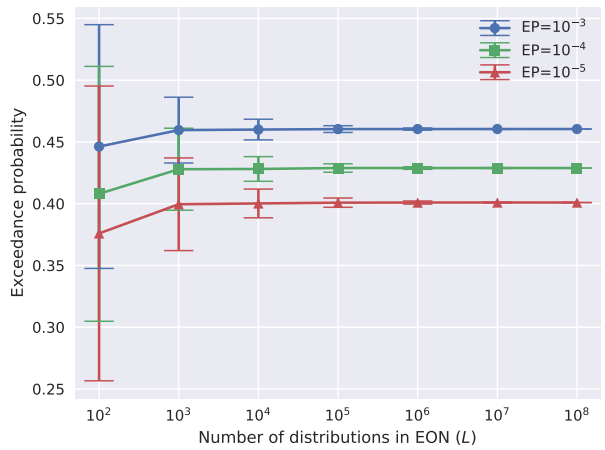
This section describes the convergence behavior for both the Ensemble of Normals (EON) and the SD with respect to the L number of Normal distributions. Means and standard deviations are randomly sampled to populate an EON. As the number of distributions increases, the variability due to the random sampling decreases. An unconverged EON will create noticeably different EP predictions for both the EON90 and the SD.

E.1 Ensemble of Normals one tail EP convergence

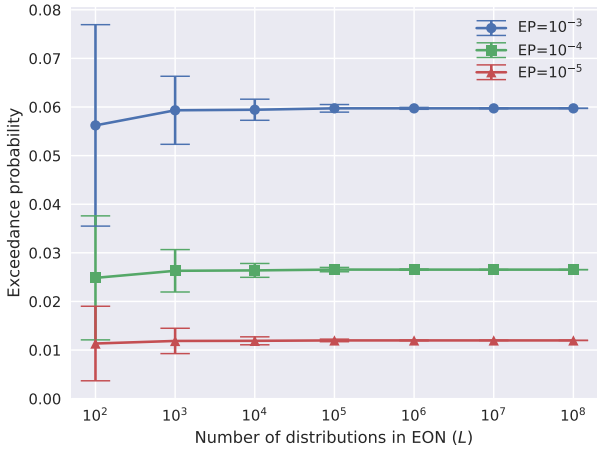
A convergence study was performed on the EON90 method to determine what is an appropriate L number of distributions in the EON. A random sample was taken from a normal distribution, and an EON90 was used to estimate the one-tail EP. Refer to Appendix D for more information on exceedance probability.

The true exceedance probabilities were determined from the Standard Normal distribution to occur at $x = 3.09, 3.72, 4.26$ for exceedance probabilities of $P = 10^{-3}, 10^{-4}, 10^{-5}$. At random, a set of sparse samples were picked from the Standard Normal distribution. The number of data points considered were $N = 2, 10, 20$. The same set of sparse samples were used throughout the study, such that the variation in predicting the exceedance probability was solely due to the L number of distributions. The exceedance probability was then predicted 100 times from different EON.

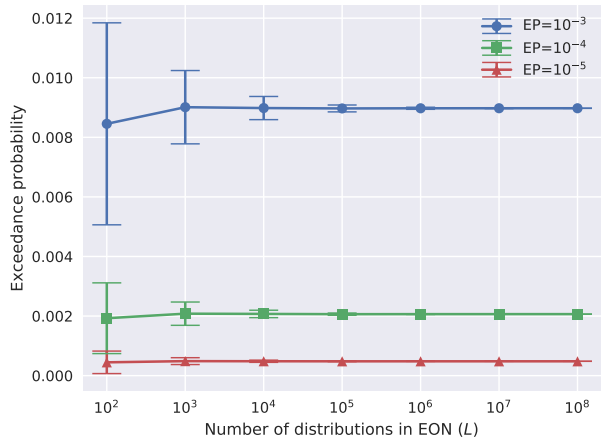
The results of the convergence study are presented in Figure E.4. The data points in the figures were determined by the mean of 100 replicates, while the error bars show ± 1.96 standard deviations from the mean. It appears that with $L = 10^4$ distributions in an EON, that an EP can be consistently predicted with minimum variability. In all cases the mean of $L = 10^3$ appeared to be very similar to the mean of $L = 10^8$, demonstrating early convergence in the mean results. Additionally, it's worthwhile to note that the variability appears to completely diminish with $L = 10^6$ for all cases.



(a) $N = 2$



(b) $N = 10$



(c) $N = 20$

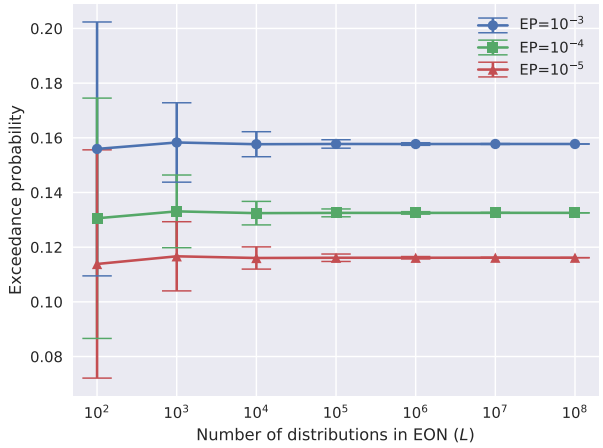
Figure E.4: EON90 convergence results at predicting EP. Each data point represents the mean EP prediction from 100 different sets of EON distributions. The errors bars represent ± 1.96 standard deviations from the mean.

E.2 Superdistribution one tail EP convergence

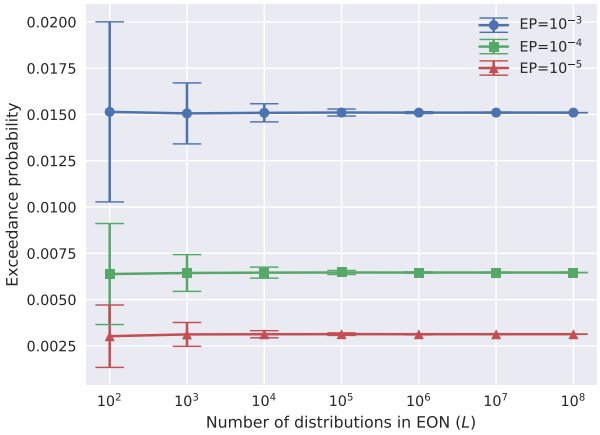
A convergence study was performed on this new Superdistribution method to determine what is an appropriate L number of distributions in the EON to construct a Superdistribution. A random sample was taken from a normal distribution, and a Superdistribution was used to estimate the one-tail exceedance probability. Refer to Appendix D for more information on exceedance probability.

The convergence study follows the same set up as the EON convergence study, and the results of the convergence study are presented in Figure E.5. The data points in the figures were determined by the mean of 100 replicates, while the error bars show ± 1.96 standard deviations from the mean. It appears that with $L = 10^4$ distributions in an EON, that a Superdistribution can reasonably predict the exceedance probability. The variability appears to completely diminish with $L = 10^7$ for all cases. Additionally, the error bars appear to decrease for all cases as you increase the N number of

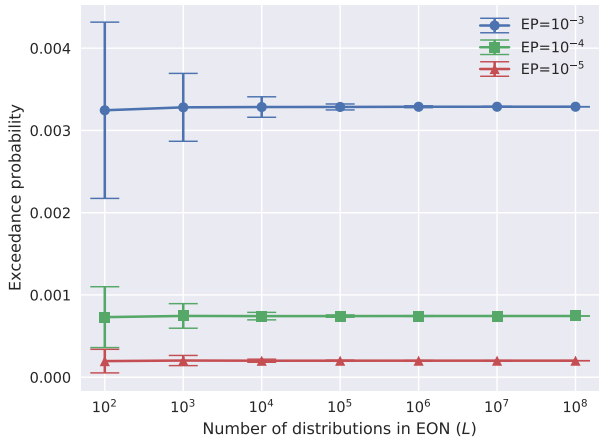
data points.



(a) $N = 2$



(b) $N = 10$

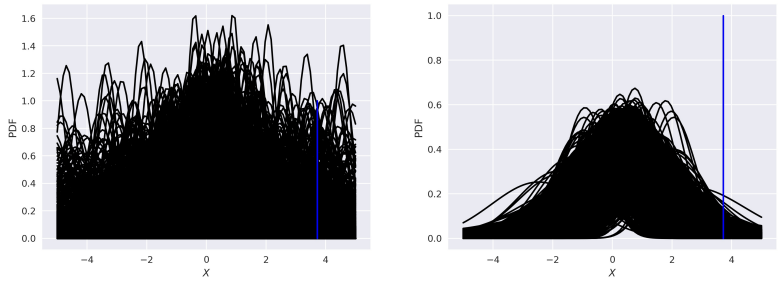


(c) $N = 20$

Figure E.5: Convergence study for Superdistribution method. Data points show the mean value, while error bars show ± 1.96 standard deviation from 100 replicate runs.

F Why the kink in CDFs of EON90 for sparse samples?

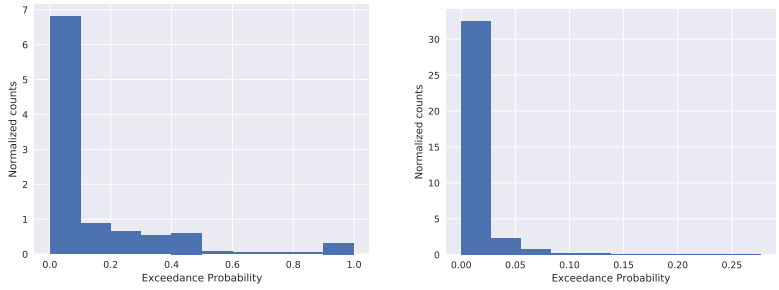
Figure F.6 shows an EON with $L = 10^4$ distributions for predicting an EP of 10^{-4} on the Standard Normal distribution. The response value of interest that yields an EP of 10^{-4} is indicated with he blue line. While the majority of distributions' means sit to the left of the of the true of the line of interest, there are a select few distributions whose means sit entirely to the right. The distributions to the left are likely to predict small EP values, while the distributions on the right are likely to predict large (near one) EP values. A histogram can be constructed from the 10^4 EP predictions and is shown in Figure F.7. There are a lot of distributions predicting a small EP, and an unfortunate number of distributions predicting a large EP for $N = 2$. However, with $N = 10$ there were no distributions predicting an EP greater than $P = 0.3$.



(a) $N = 2$ (b) $N = 10$

Figure F.6: EON distributions for predicting an EP of 10^{-4} on the Standard Normal distribution. The blue line indicates the region where to the right is the true EP on the Standard Normal distribution.

As you increase the N number of samples, the number of distributions in the EON that lay to the right of the true EP diminishes. The relationship with the N number of samples is shown in both figures, where the amount of EP predictions that yields 1.0 is non-existent with $N = 10$ samples. Thus the kink in the CDFs results from the diversity of the EON with sparse samples, with having a separate mode predicting an EP of 1.0.



(a) $N = 2$ (b) $N = 10$

Figure F.7: Normalized histogram of EP predictions in an EON for predicting an EP of 10^{-4} on the Standard Normal distribution.

G On an alternative EPmetric

As discussed, the EPmetric's worse possible value of an underestimated EP could theoretically be $-\infty$, while the worst possible value of an overestimated EP is bounded. It is practical to punish underestimates more than conservative overestimates, but an alternative error metric was investigated which had the opposite effect. Consider

$$\Delta = EP_{\text{estimated}} - EP_{\text{true}} \tag{G.1}$$

and

$$\text{EPdiff} = \left[\sum^{N+} \Delta + \sum^{N-} |\Delta|\right] / N^{+} \tag{G.2}$$

which arrives at a different metric named EPdiff. EPdiff has a bounded worse possible underestimate error, but with a worst possible overestimate of one minus your true EP. In the context of tail probability estimation, this means that EPdiff will have a larger possible overestimation error than underestimation error (unlike EPmetric). In the cases covered in this study, EPdiff would penalize overestimated EPs more than underestimated EPs (again unlike EPmetric).

The results using EPdiff for EP$= 10^{-4}$ on the Standard Normal distribution can be seen in Figure G.8. The Superdistribution appeared to have the lowest EPdiff for all number of samples. While the EPmetric results show the Superdistribution was fairly ahead of the Best TI-EN (Figure 9), the EPdiff results show that the Superdistribution is just slightly ahead of the Best TI-EN. The ordering of the other methods was not consistent when comparing the results with EPmetric (Figure 9) to the results of the alternative EPdiff value (Figure G.8). However, the Superdistribution had the best metric values over the entire range of N =2 to 20 range.

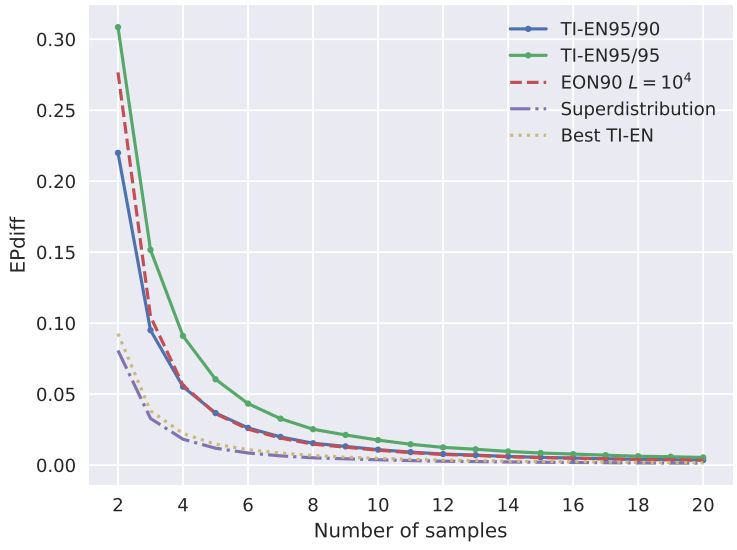


Figure G.8: Results of predicting EP$= 10^{-4}$ from the Standard Normal distribution using EPdiff.

A consequence of using EPdiff instead of EPmetric is that the new Best TI-EN confidence levels were at the lowest end of the confidence interval, because this prevents conservative overestimates. The EPmetric had a tendency to prefer TI-EN with higher confidence intervals. This is

evident by comparing the best TI-EN for EPdiff in Figure G.9, to the best TI-EN for EPmetric in Figure 10. The relationship between the TI-EN confidence level and the EPdiff value can be seen in Figure G.10.

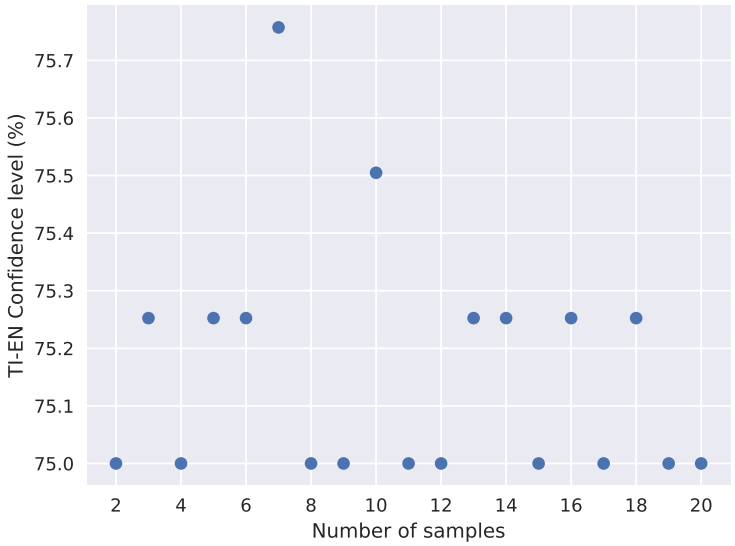


Figure G.9: The confidence level which resulted in the best TI-EN EPdiff value for EP=10^{-4} from the Standard Normal distribution.

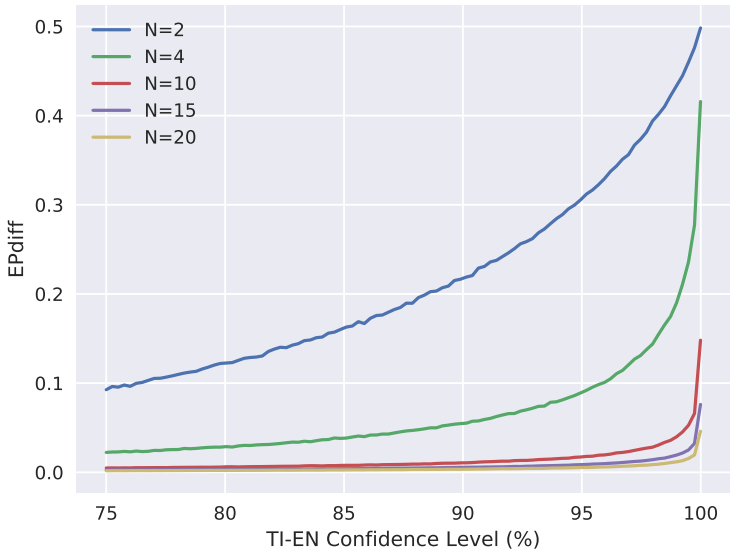


Figure G.10: EPdiff value for various TI-EN confidence levels. Different curves represent the different number of samples for EP=10^{-4} on the standard Normal distribution.

If you wanted to balance EPmetric (or EPdiff) such that the worst possible underestimate was equal to the worst possible overestimate, you could divide each side of the numerator by their appropriate maximum.

H Eight Other-distributions with performance characteristics like Log-Normal

H.1 Weibull Wide distribution

The Weibull Wide distribution used in this report was the same used in [1], and the PDF can be seen in Figure H.11. The performance metric results are found in Figures H.12 and H.13.

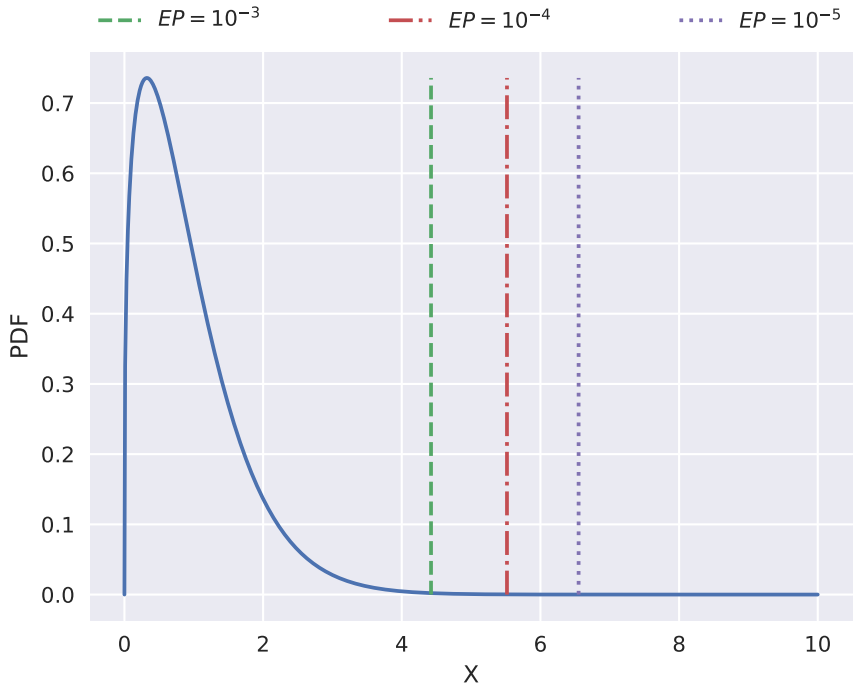
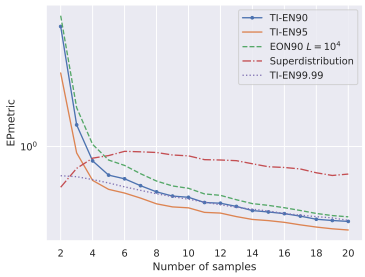
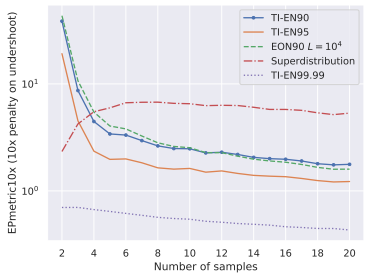


Figure H.11: Weibull Wide distribution.

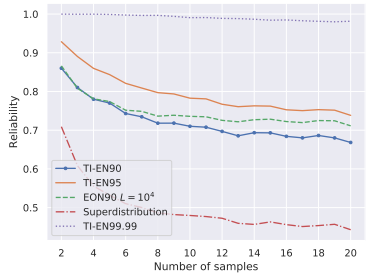
The Weibull Wide distribution has similar trends to the other non-normal distributions covered so far, with the SD have the best EPmetric values for $N \leq 8$ samples in all cases. Additionally, the reliability of all to the Sparse UQ methods decreased as you increased the N.



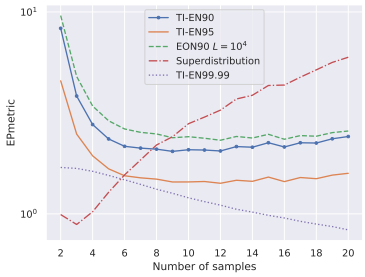
EPmetric, EP= 10^{-1}



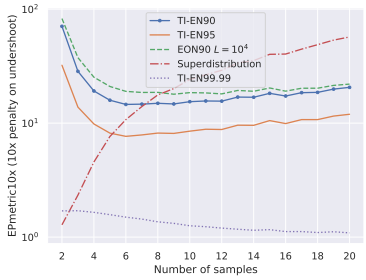
EPmetric10x, EP= 10^{-1}



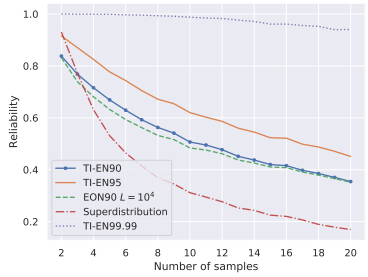
Reliability, EP= 10^{-1}



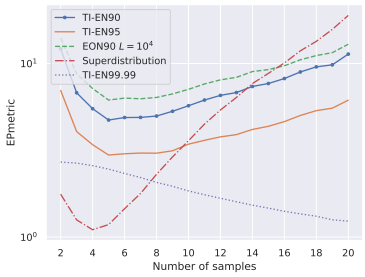
EPmetric, EP= 10^{-2}



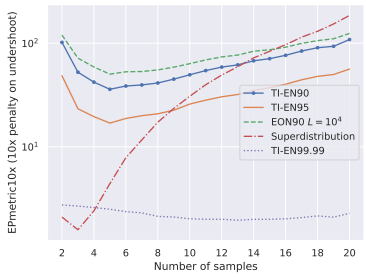
EPmetric10x, EP= 10^{-2}



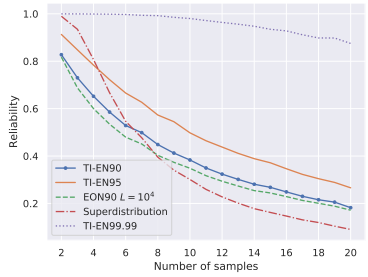
Reliability, EP= 10^{-2}



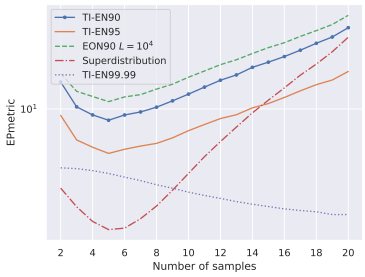
EPmetric, EP= 10^{-3}



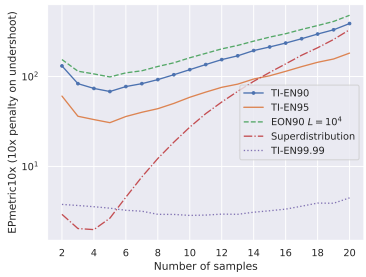
EPmetric10x, EP= 10^{-3}



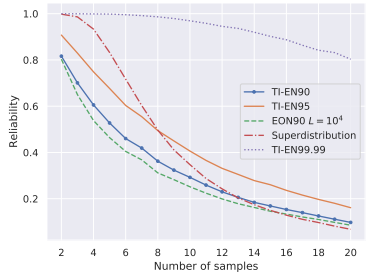
Reliability, EP= 10^{-3}



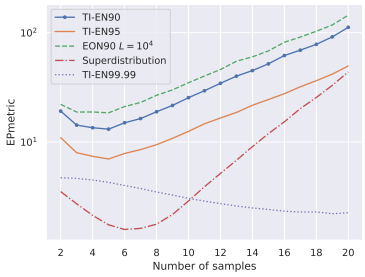
EPmetric, EP= 10^{-4}



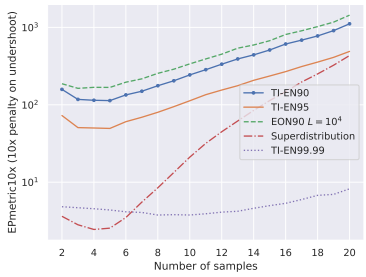
EPmetric10x, EP= 10^{-4}



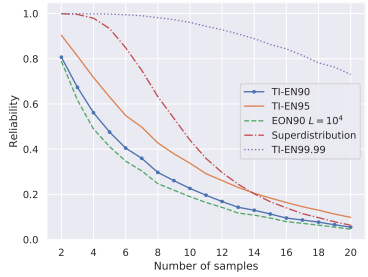
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

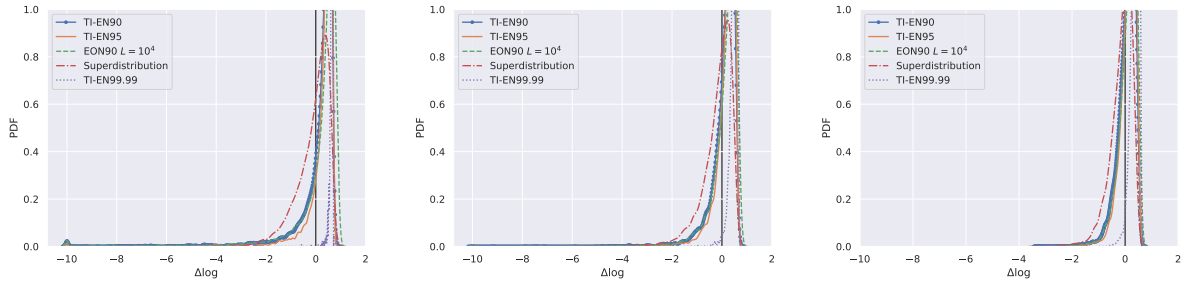


EPmetric10x, EP= 10^{-5}

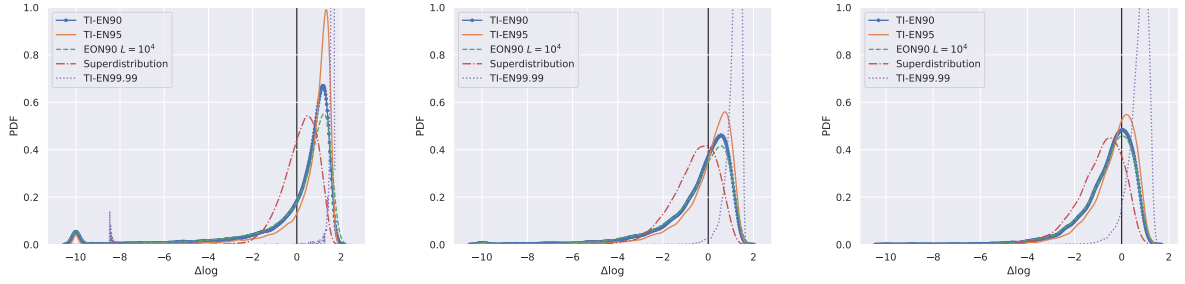


Reliability, EP= 10^{-5}

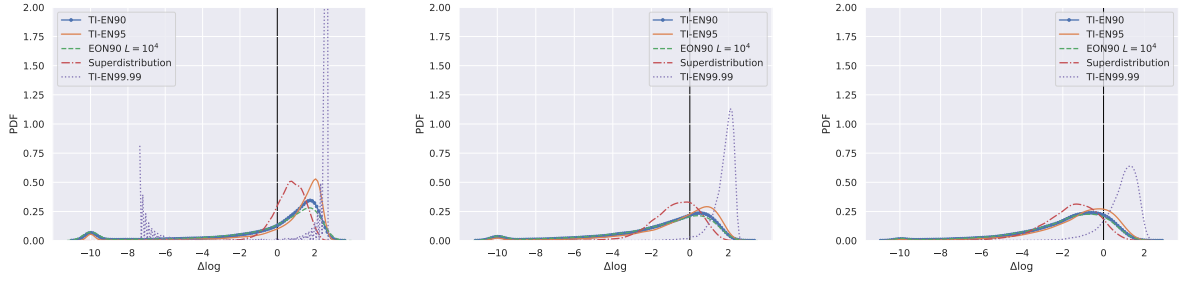
Figure H.12: Results for the Weibull Wide distribution.



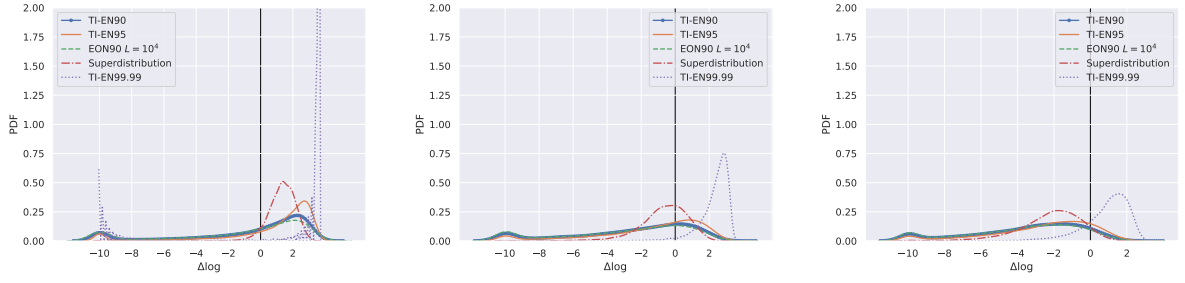
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



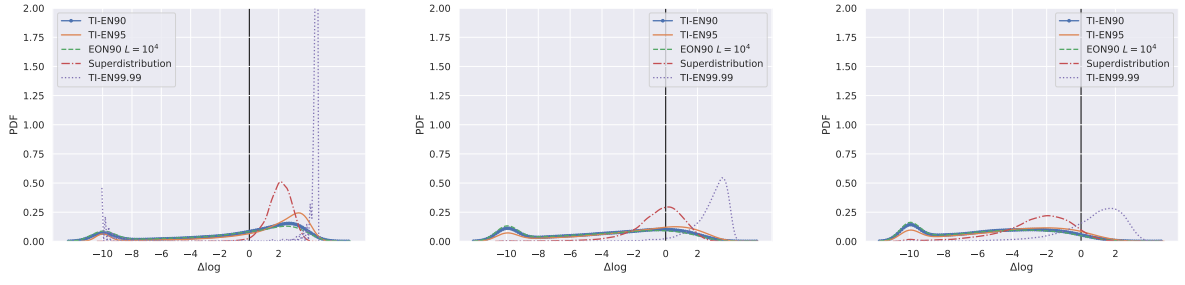
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.13: Distribution of results for the Weibull Wide distribution.

H.2 Exponential Narrow distribution

The PDF for the Exponential Narrow distribution is shown in Figure H.14, and the PDF for the Exponential Wide distribution is shown in Figure H.15. The results of the Exponential Narrow distribution are found in Figures H.16 and H.17. The results of a Exponential Wide distribution are found in Figures H.18 and H.19. The difference between the two Exponential distributions is simply just the scale.

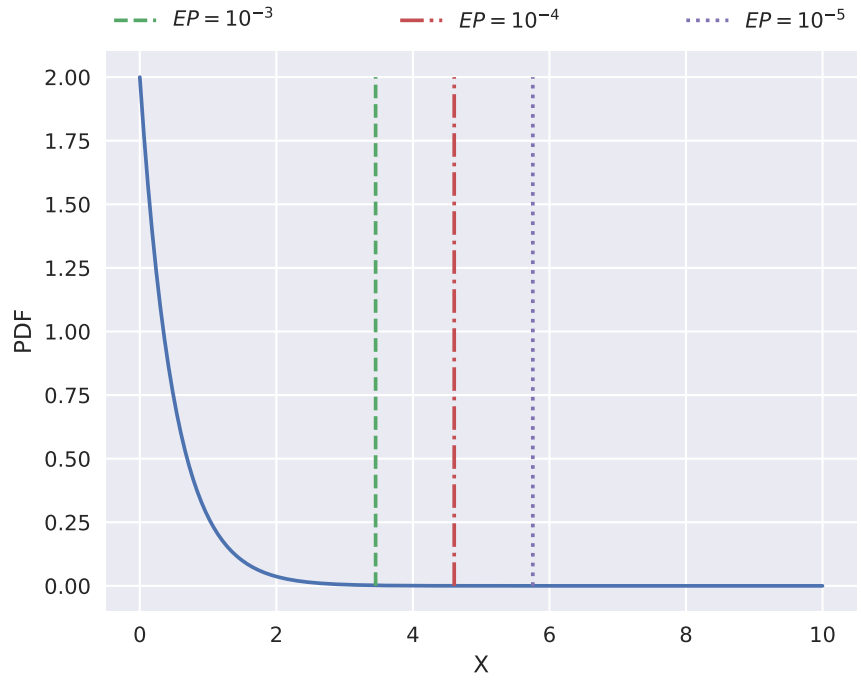


Figure H.14: Exponential Narrow distribution.

The results of the two Exponential distributions are nearly identical for both accuracy and reliability. This leads to the idea that sparse UQ methods may not be sensitive to the scaling for at least this type of distribution. The SD again appears to have the best EPmetric values and reliability at low number of samples from $N \leq 4$.

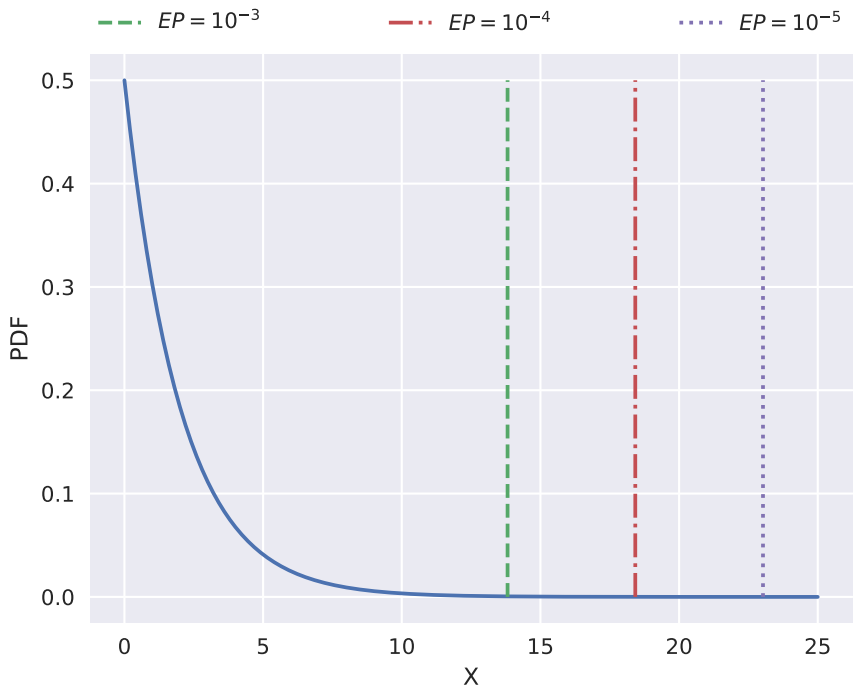
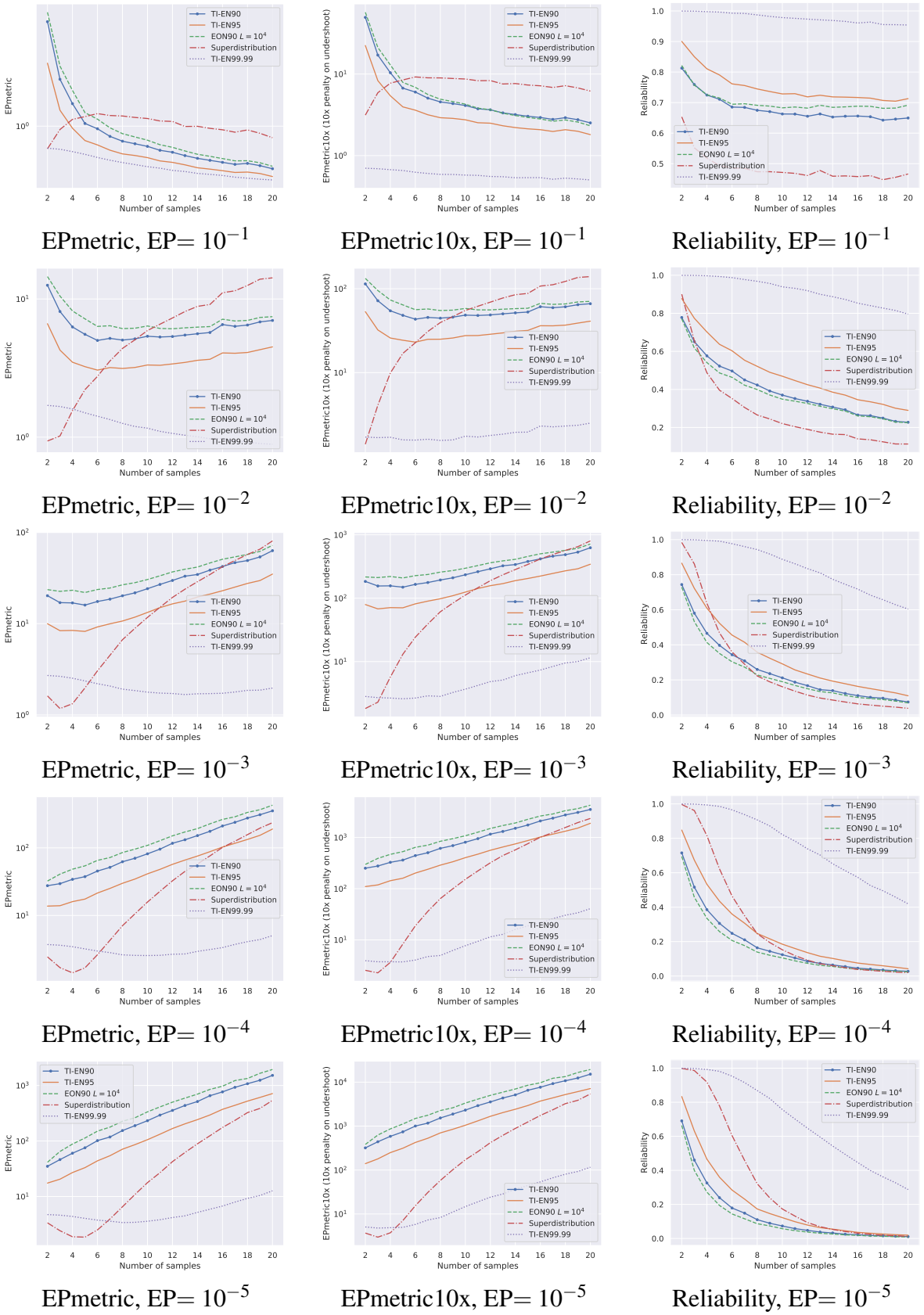


Figure H.15: Exponential Wide distribution.



EPmetric, EP= 10^{-1} | EPmetric10x, EP= 10^{-1} | Reliability, EP= 10^{-1}

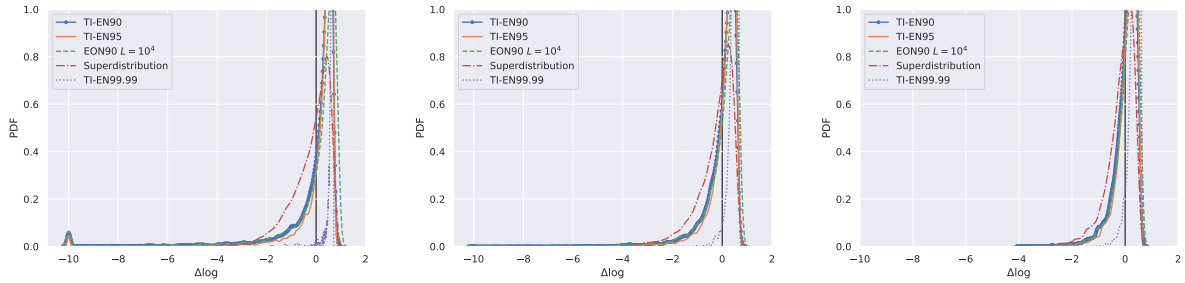
EPmetric, EP= 10^{-2} | EPmetric10x, EP= 10^{-2} | Reliability, EP= 10^{-2}

EPmetric, EP= 10^{-3} | EPmetric10x, EP= 10^{-3} | Reliability, EP= 10^{-3}

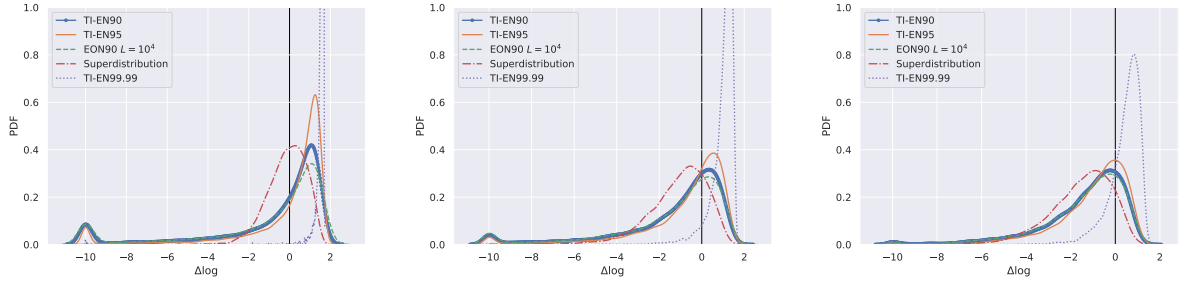
EPmetric, EP= 10^{-4} | EPmetric10x, EP= 10^{-4} | Reliability, EP= 10^{-4}

EPmetric, EP= 10^{-5} | EPmetric10x, EP= 10^{-5} | Reliability, EP= 10^{-5}

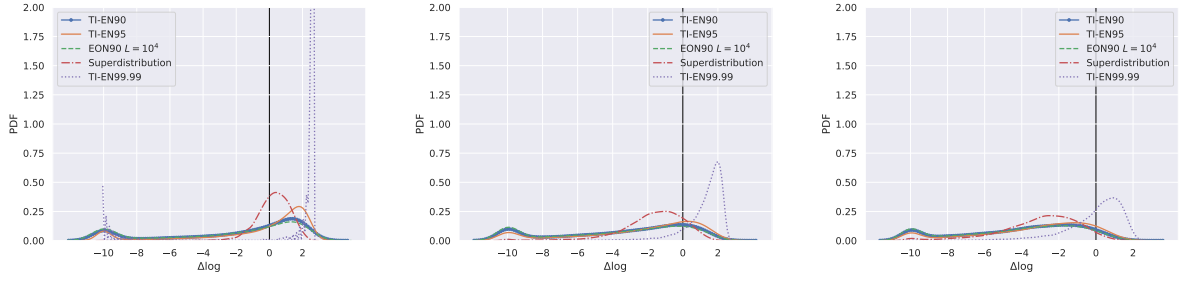
Figure H.16: Results for the Exponential Narrow distribution.



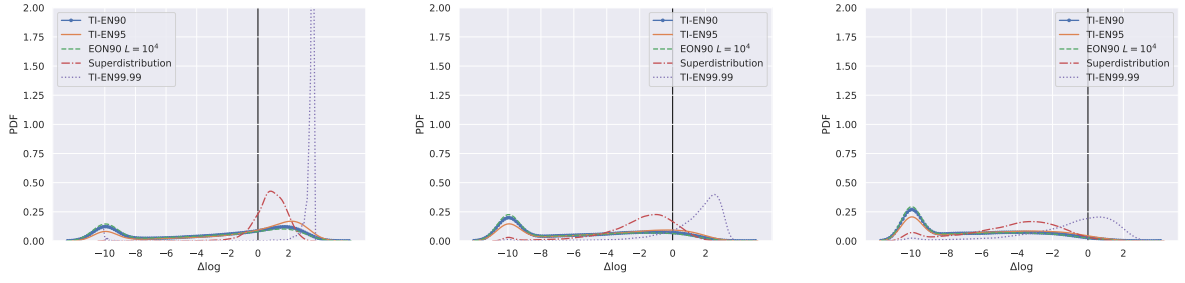
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



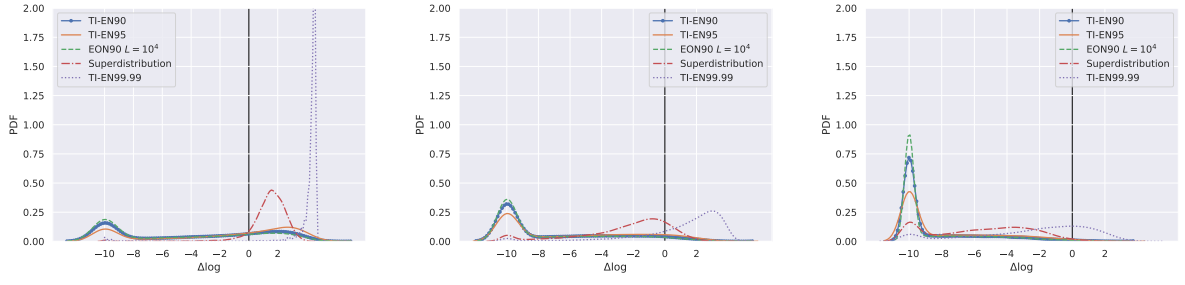
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$

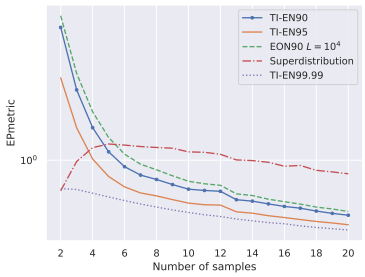


$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$

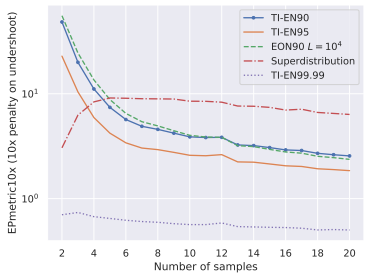


$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

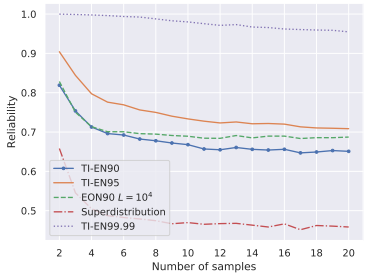
Figure H.17: Distribution of results for the Exponential Narrow distribution.



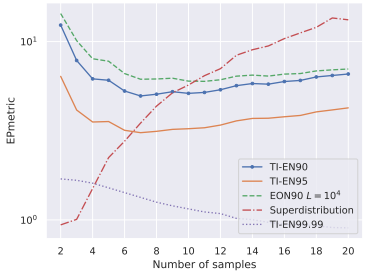
EPmetric, EP= 10^{-1}



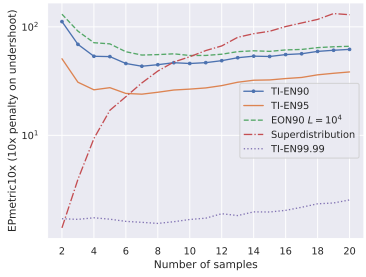
EPmetric10x, EP= 10^{-1}



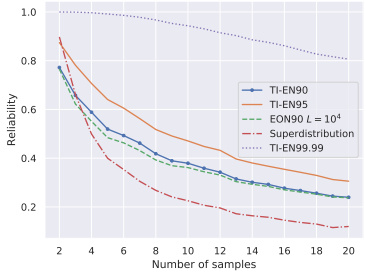
Reliability, EP= 10^{-1}



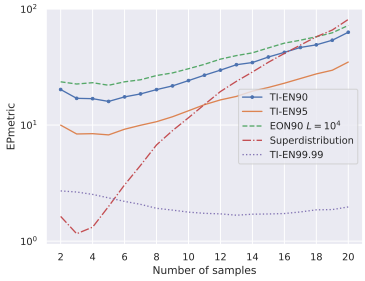
EPmetric, EP= 10^{-2}



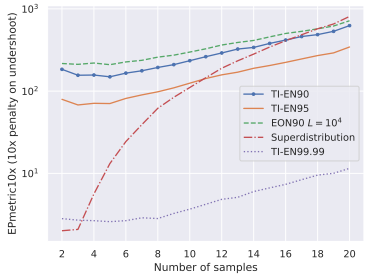
EPmetric10x, EP= 10^{-2}



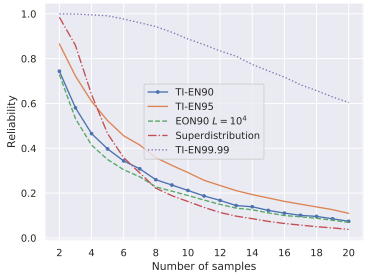
Reliability, EP= 10^{-2}



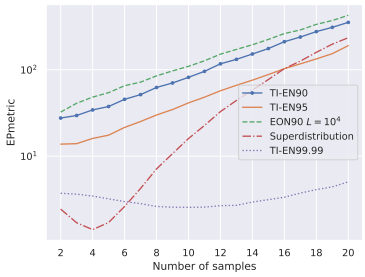
EPmetric, EP= 10^{-3}



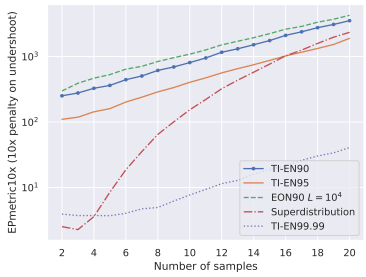
EPmetric10x, EP= 10^{-3}



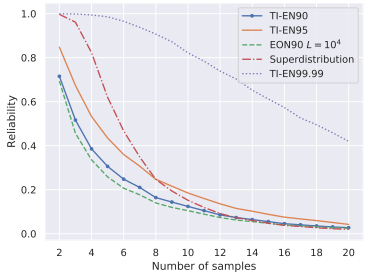
Reliability, EP= 10^{-3}



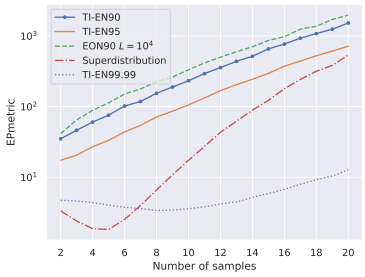
EPmetric, EP= 10^{-4}



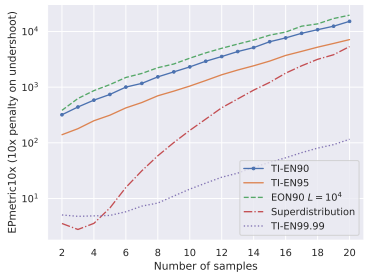
EPmetric10x, EP= 10^{-4}



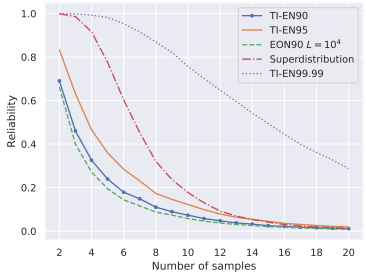
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

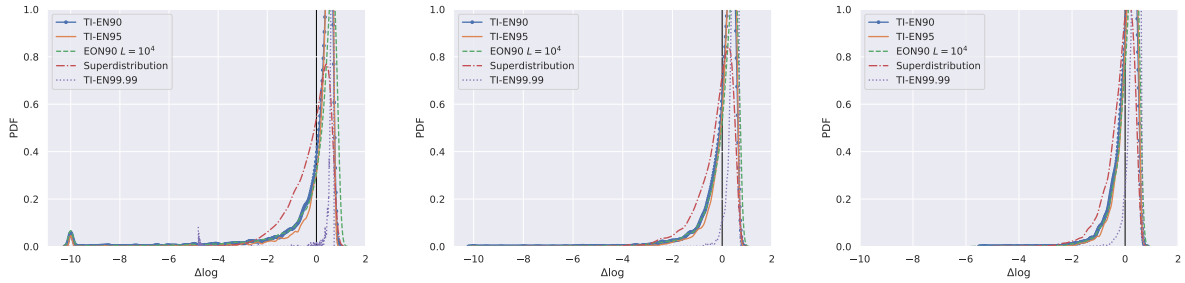


EPmetric10x, EP= 10^{-5}

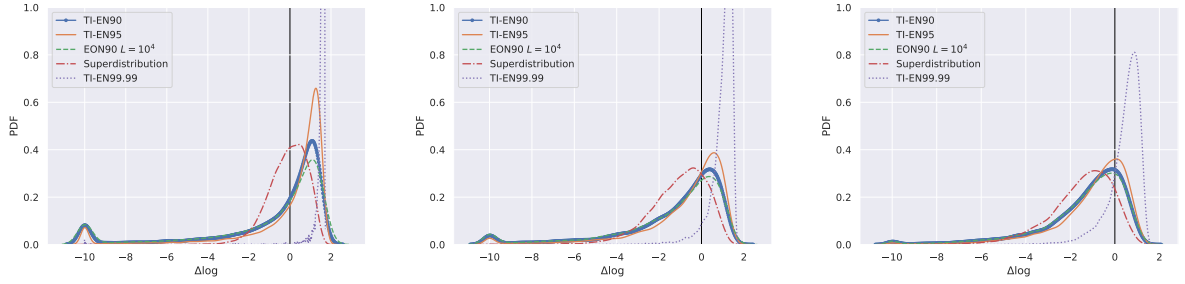


Reliability, EP= 10^{-5}

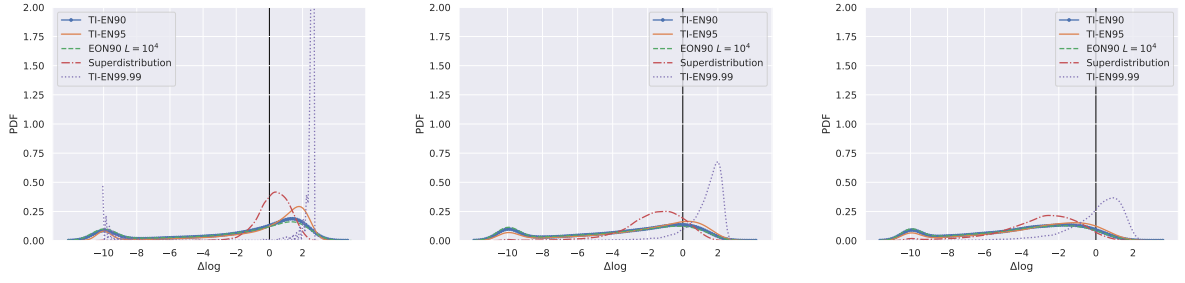
Figure H.18: Results for the Exponential Wide distribution.



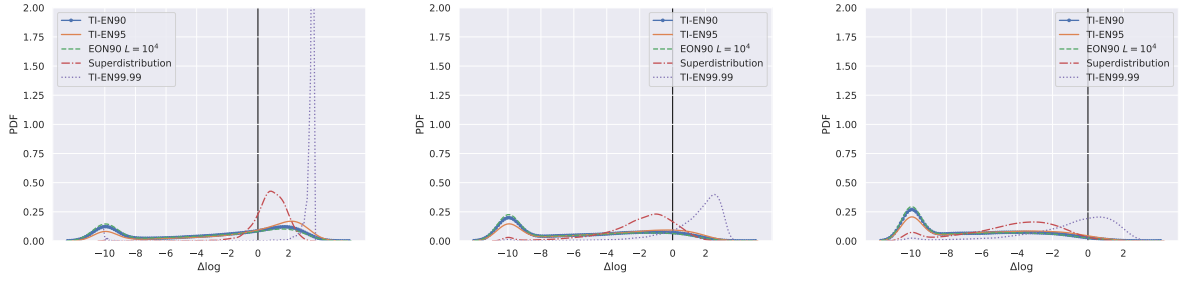
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



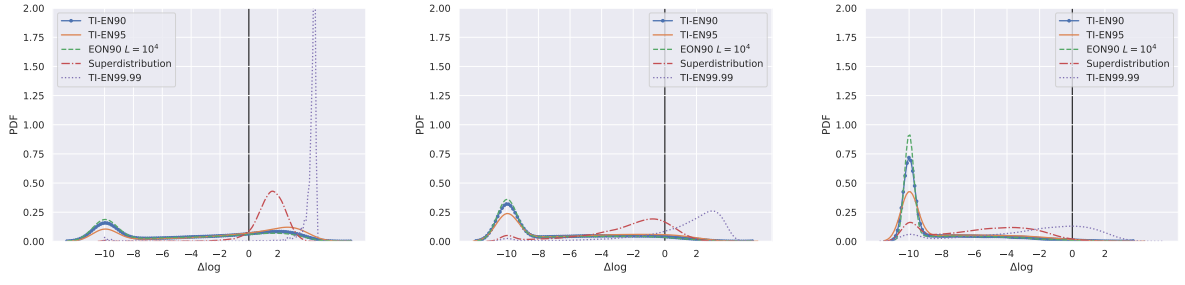
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.19: Distribution of results for the Exponential Wide distribution.

H.3 Tearing Parameter Weld Element 0.75

A histogram and KDE of the Tearing Parameter Weld Element 0.75 data is shown in Figure H.20. The results are shown in Figures H.21 and H.22.

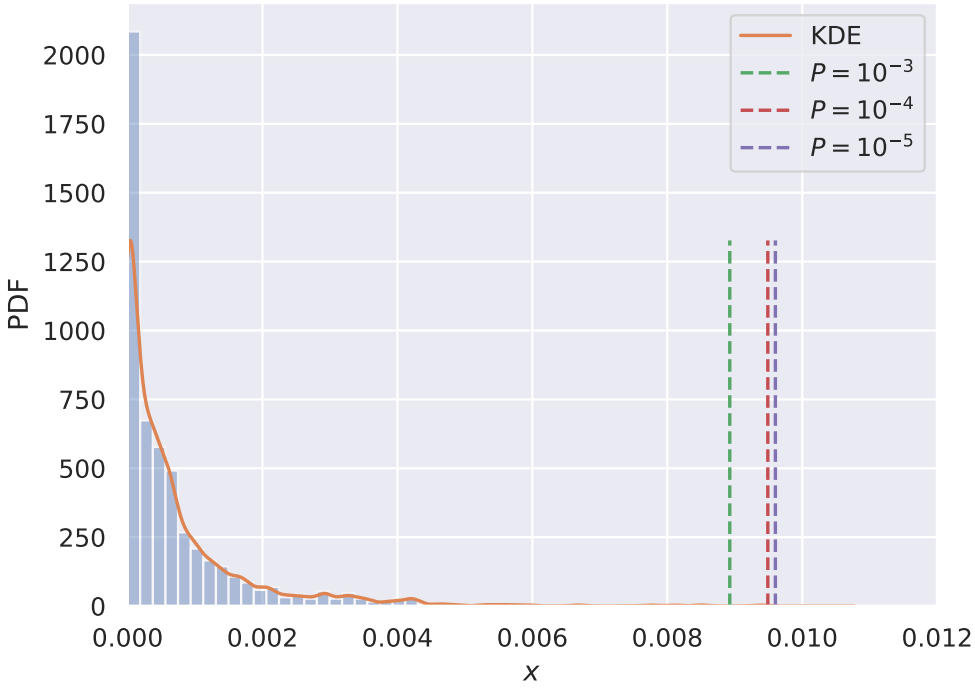
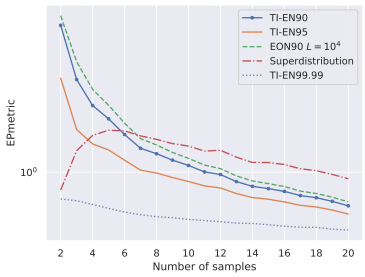
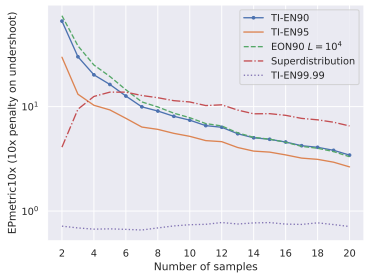


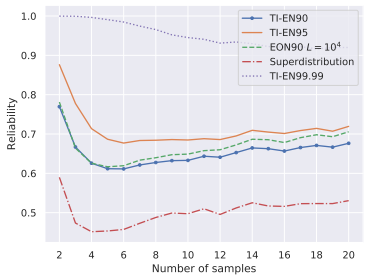
Figure H.20: Histogram and KDE for Tearing Parameter Weld Element 0.75.



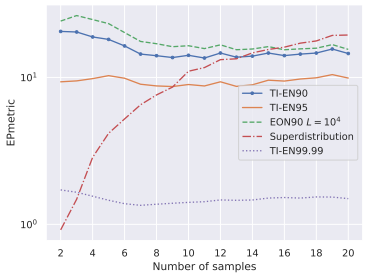
EPmetric, EP= 10^{-1}



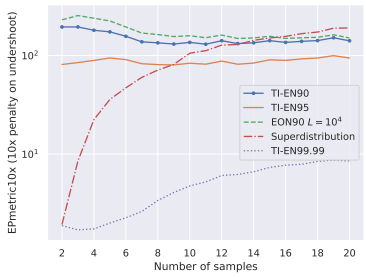
EPmetric10x, EP= 10^{-1}



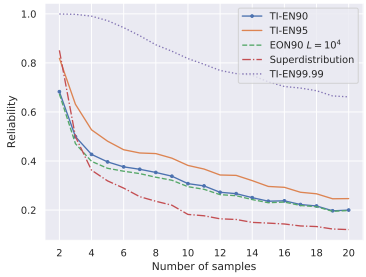
Reliability, EP= 10^{-1}



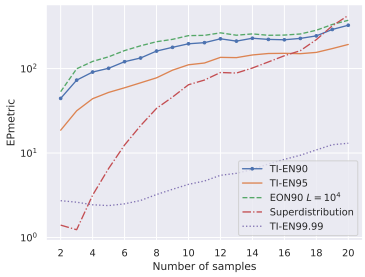
EPmetric, EP= 10^{-2}



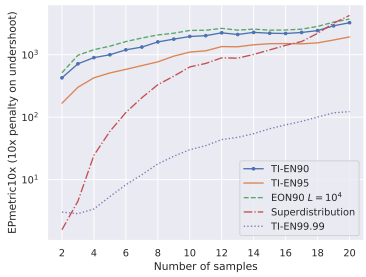
EPmetric10x, EP= 10^{-2}



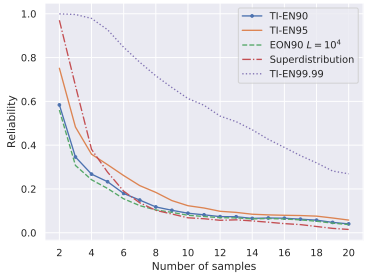
Reliability, EP= 10^{-2}



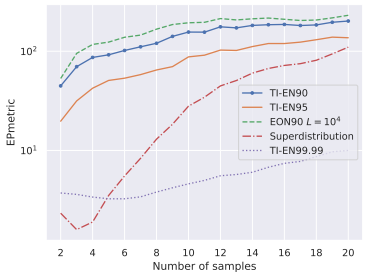
EPmetric, EP= 10^{-3}



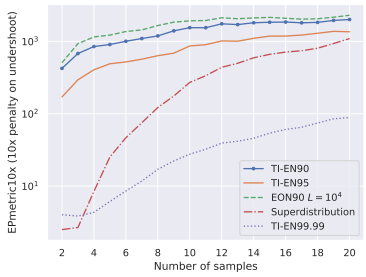
EPmetric10x, EP= 10^{-3}



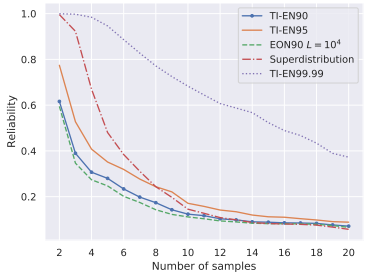
Reliability, EP= 10^{-3}



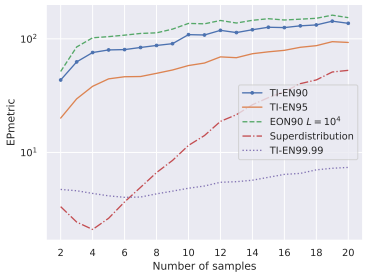
EPmetric, EP= 10^{-4}



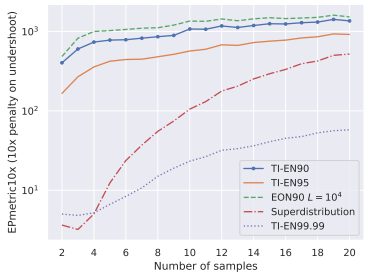
EPmetric10x, EP= 10^{-4}



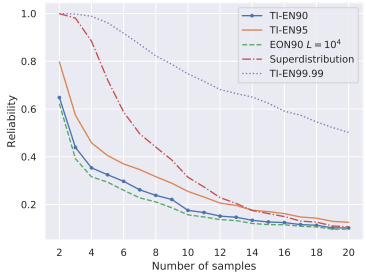
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

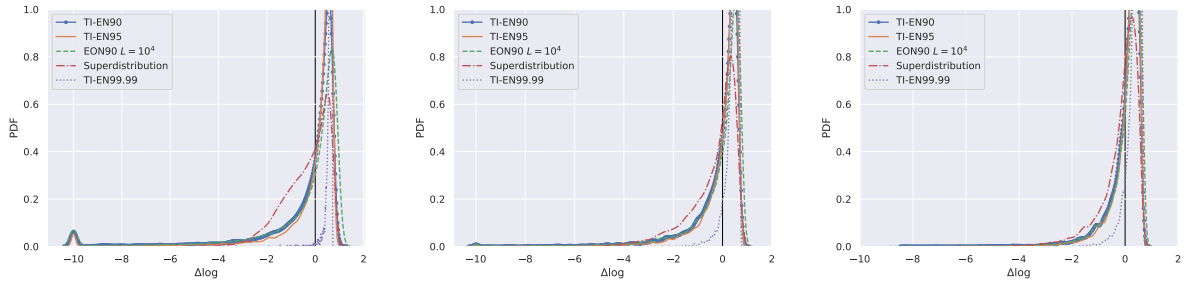


EPmetric10x, EP= 10^{-5}

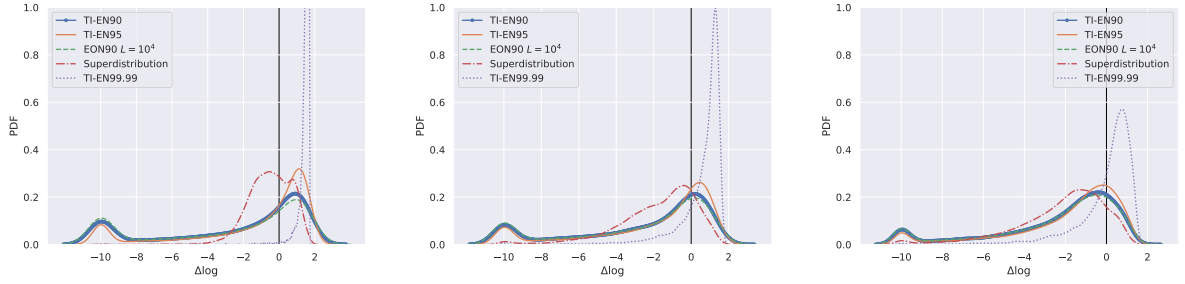


Reliability, EP= 10^{-5}

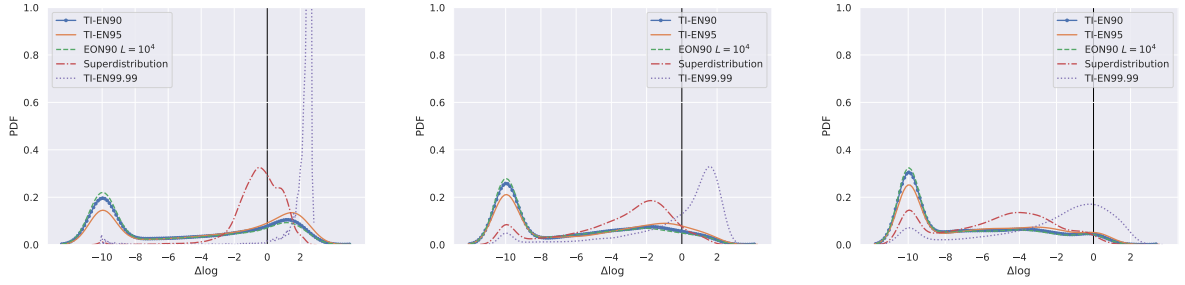
Figure H.21: Results for the empirical Tearing Parameter Weld Element 0.75 distribution.



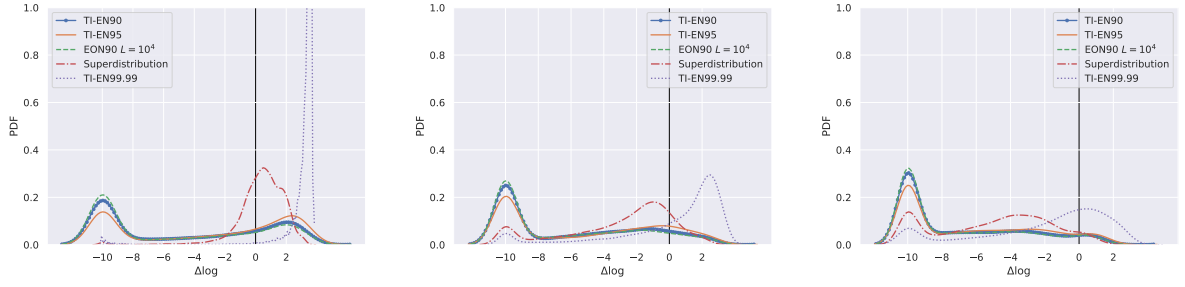
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



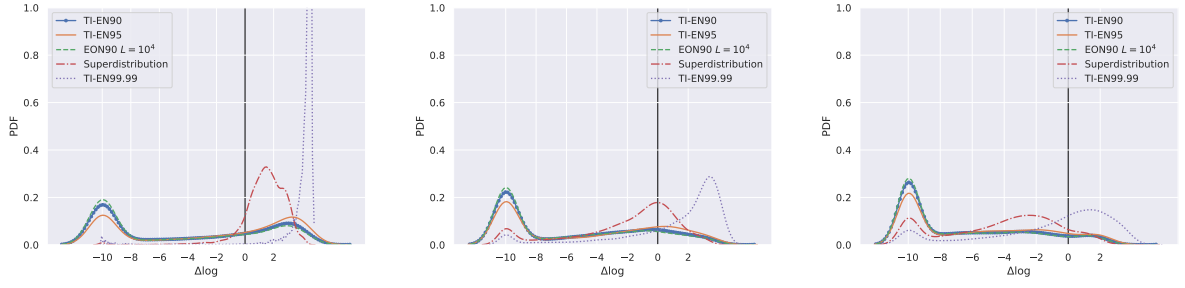
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.22: Distribution of results for the empirical Tearing Parameter Weld Element 0.75 distribution.

H.4 Tensile EQPS Can Top Element 0.5

A histogram and KDE of the Tensile EQPS Can Top Element 0.5 data is shown in Figure H.23. The results are shown in Figures H.24 and H.25.

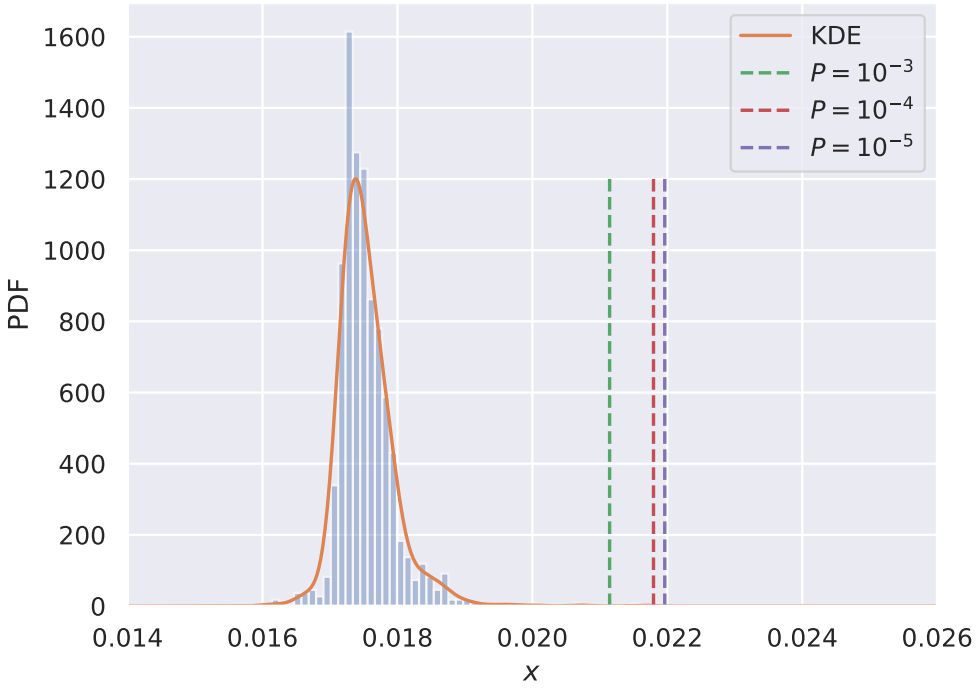
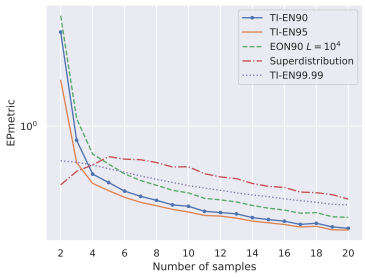
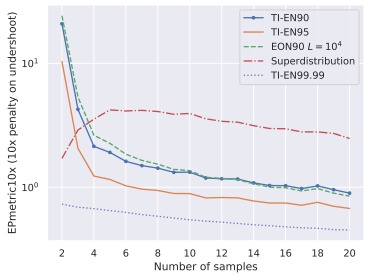


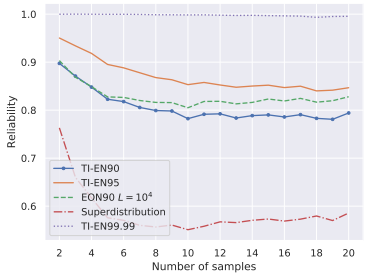
Figure H.23: Histogram and KDE for Tensile EQPS Can Top Element 0.5.



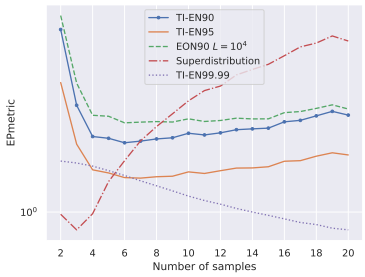
EPmetric, EP= 10^{-1}



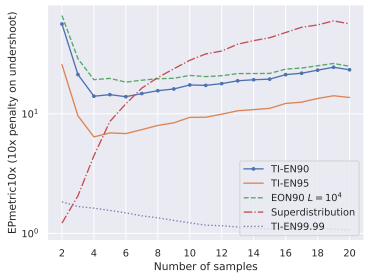
EPmetric10x, EP= 10^{-1}



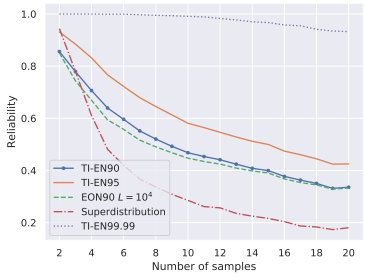
Reliability, EP= 10^{-1}



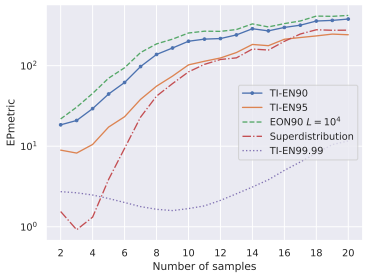
EPmetric, EP= 10^{-2}



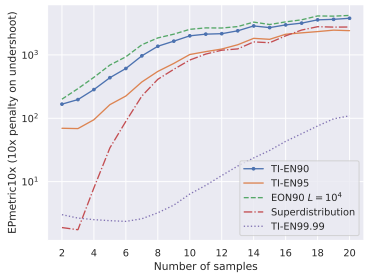
EPmetric10x, EP= 10^{-2}



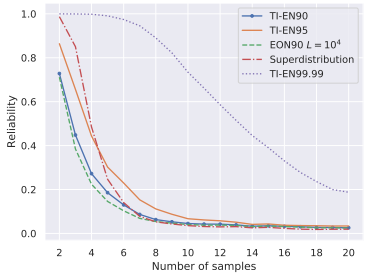
Reliability, EP= 10^{-2}



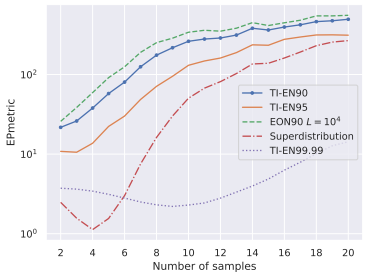
EPmetric, EP= 10^{-3}



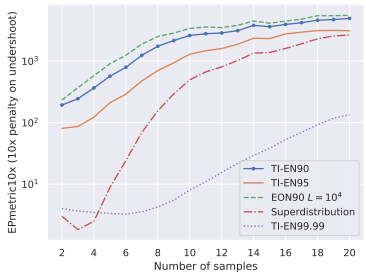
EPmetric10x, EP= 10^{-3}



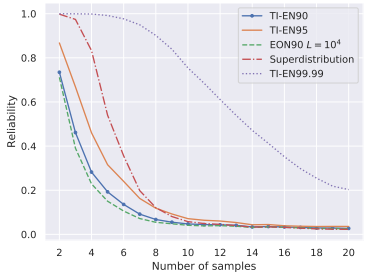
Reliability, EP= 10^{-3}



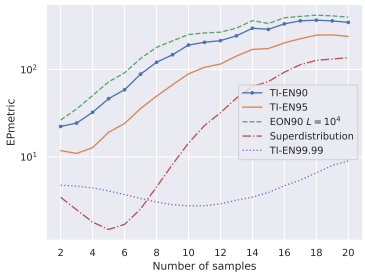
EPmetric, EP= 10^{-4}



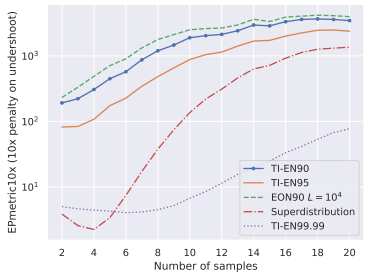
EPmetric10x, EP= 10^{-4}



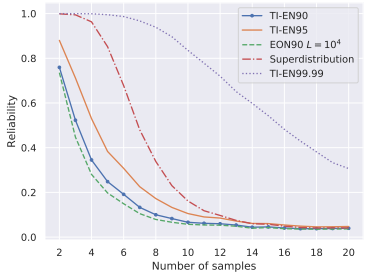
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

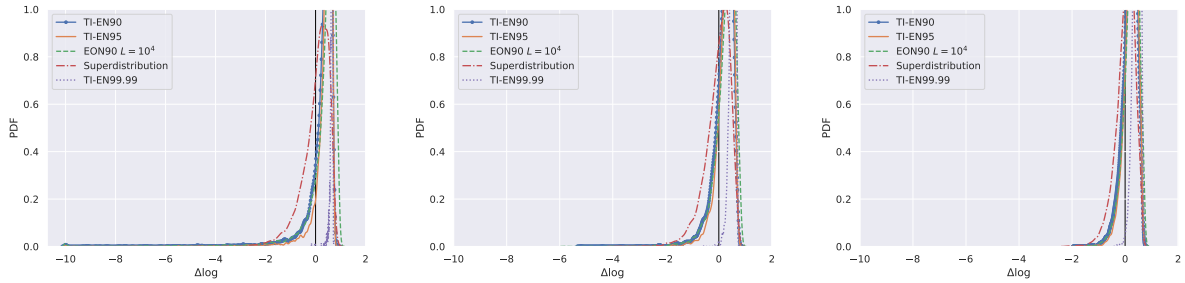


EPmetric10x, EP= 10^{-5}

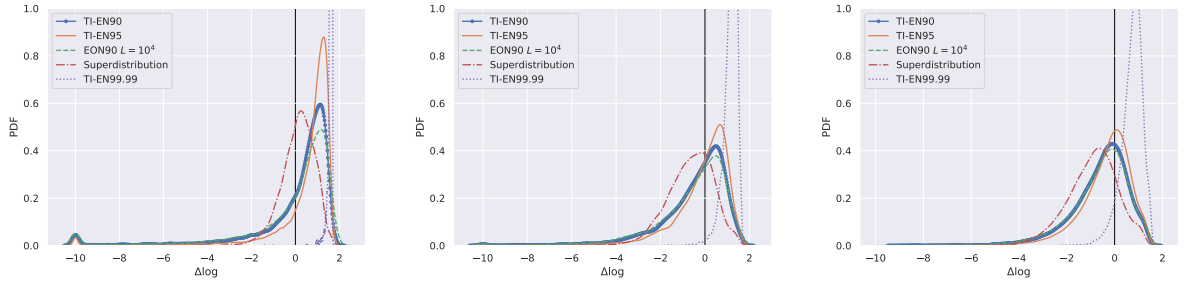


Reliability, EP= 10^{-5}

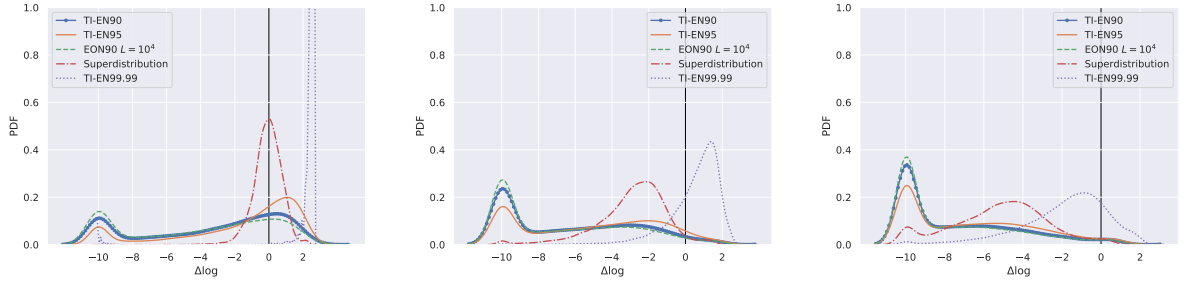
Figure H.24: Results for the empirical Tensile EQPS Can Top Element 0.5 distribution.



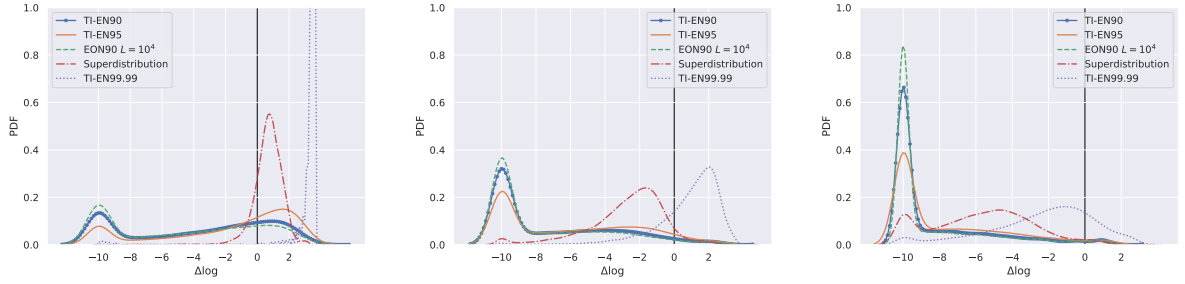
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



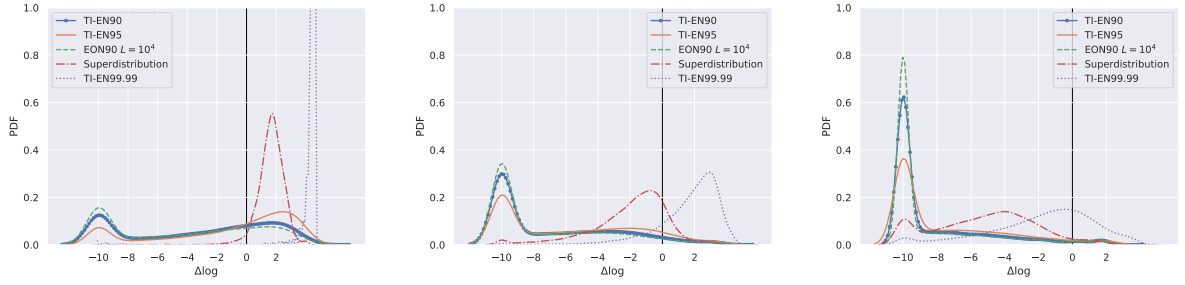
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.25: Distribution of results for the empirical Tensile EQPS Can Top Element 0.5 distribution.

H.5 Tearing Parameter Lid Buckle Element 1.0

A histogram and KDE of the Tearing Parameter Lid Buckle Element 1.0 data is shown in Figure H.26. The results are shown in Figures H.27 and H.28.

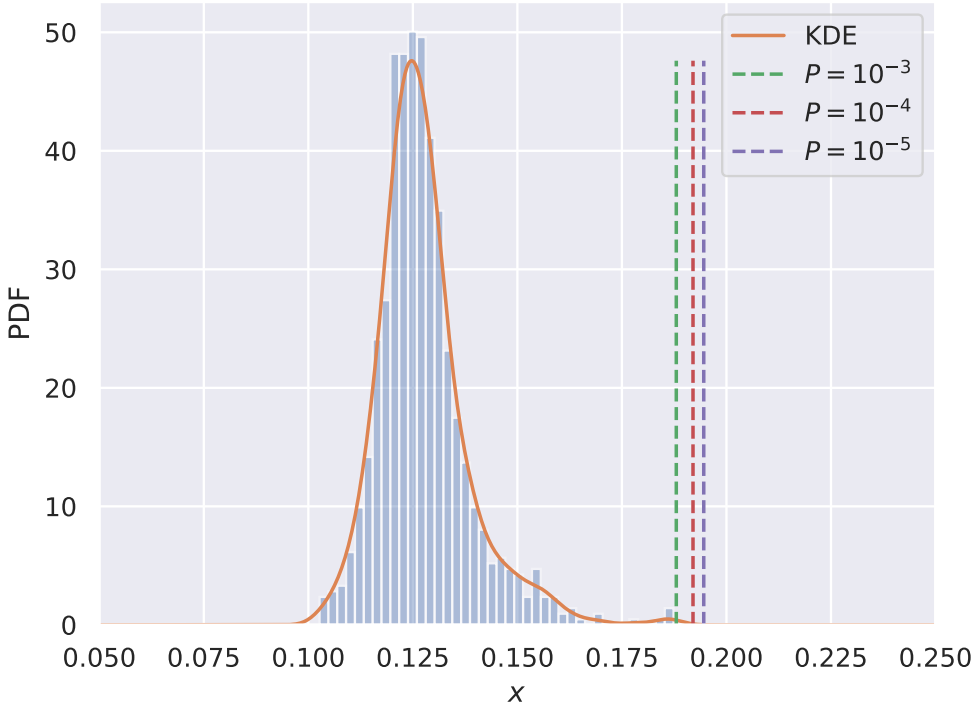
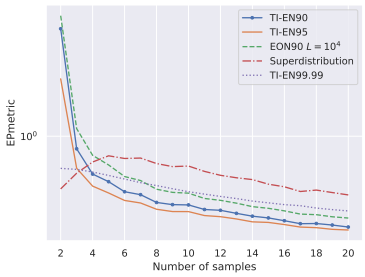
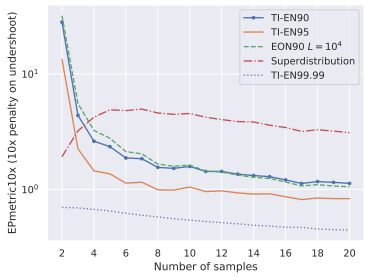


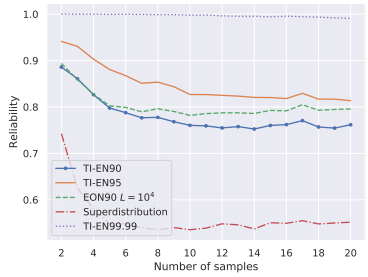
Figure H.26: Histogram and KDE for Tearing Parameter Lid Buckle Element 1.0.



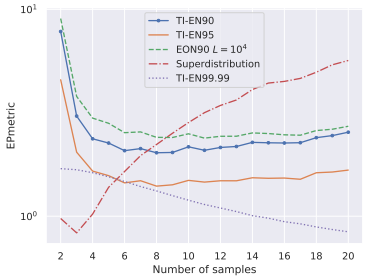
EPmetric, EP= 10^{-1}



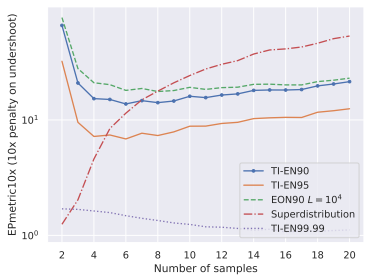
EPmetric10x, EP= 10^{-1}



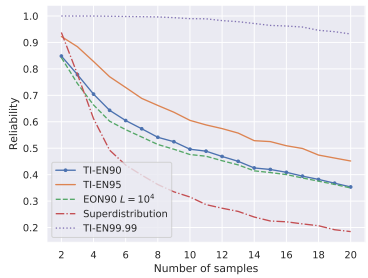
Reliability, EP= 10^{-1}



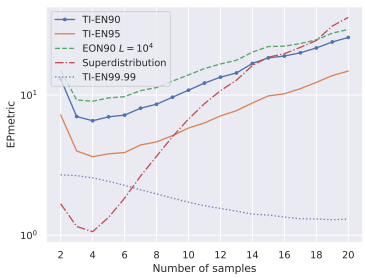
EPmetric, EP= 10^{-2}



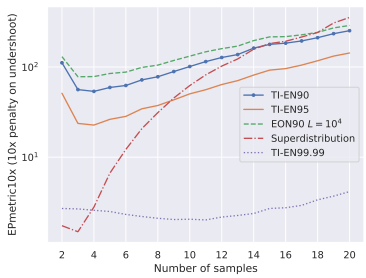
EPmetric10x, EP= 10^{-2}



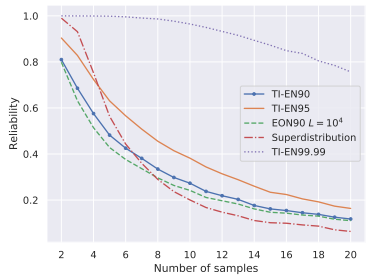
Reliability, EP= 10^{-2}



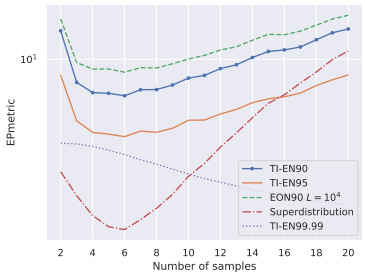
EPmetric, EP= 10^{-3}



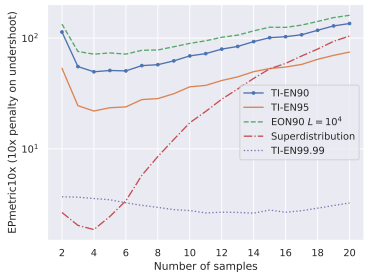
EPmetric10x, EP= 10^{-3}



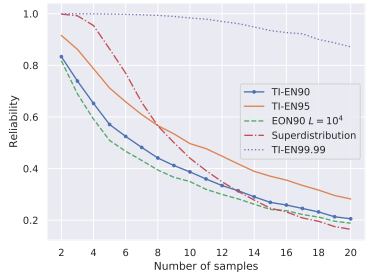
Reliability, EP= 10^{-3}



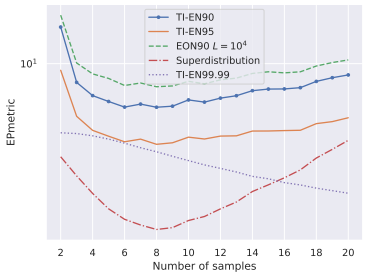
EPmetric, EP= 10^{-4}



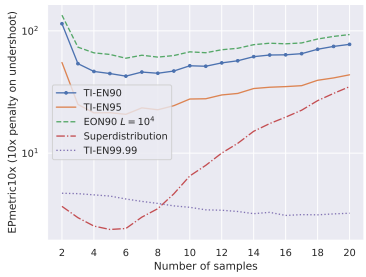
EPmetric10x, EP= 10^{-4}



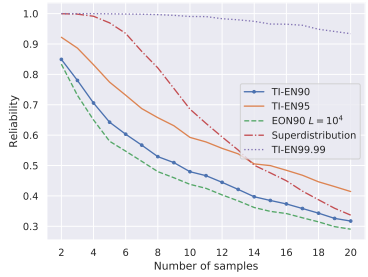
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

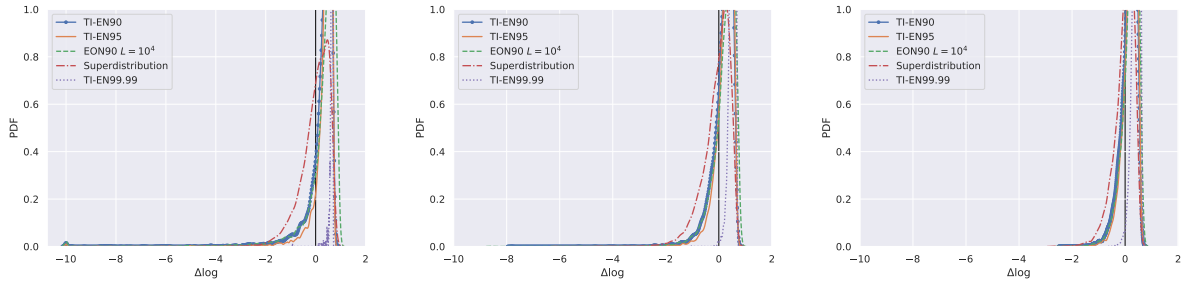


EPmetric10x, EP= 10^{-5}

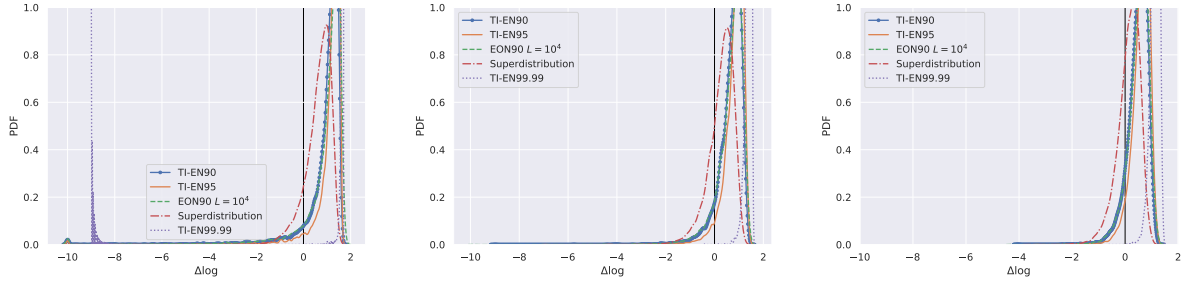


Reliability, EP= 10^{-5}

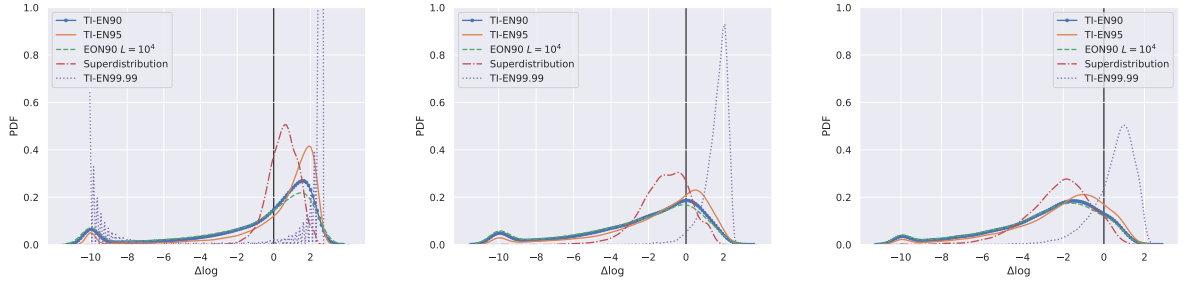
Figure H.27: Results for the empirical Tearing Parameter Lid Buckle Element 1.0 distribution.



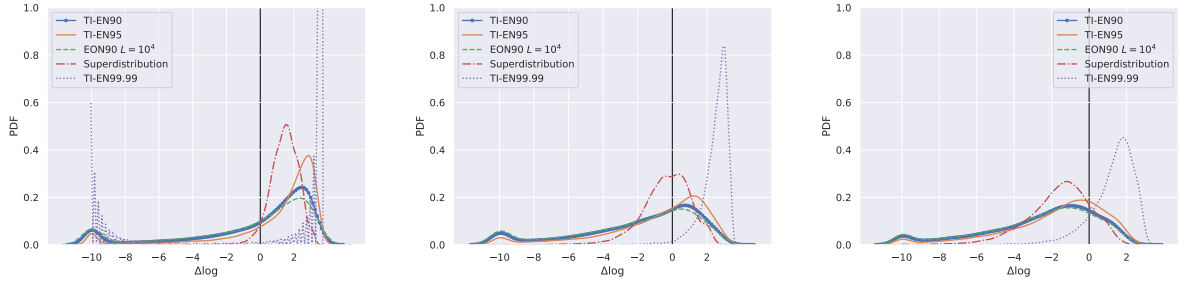
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



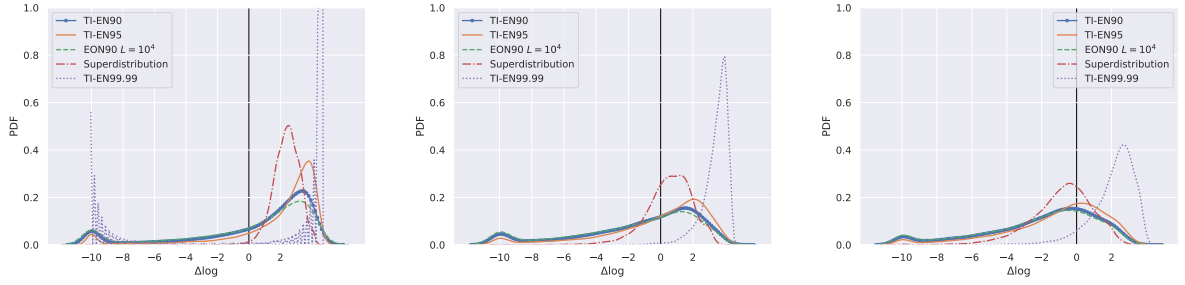
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.28: Distribution of results for the empirical Tearing Parameter Lid Buckle Element 1.0 distribution.

H.6 Tearing Parameter Lid Buckle Element 0.25

A histogram and KDE of the Tearing Parameter Lid Buckle Element 0.25 data is shown in Figure H.29. The results are shown in Figures H.30 and H.31.

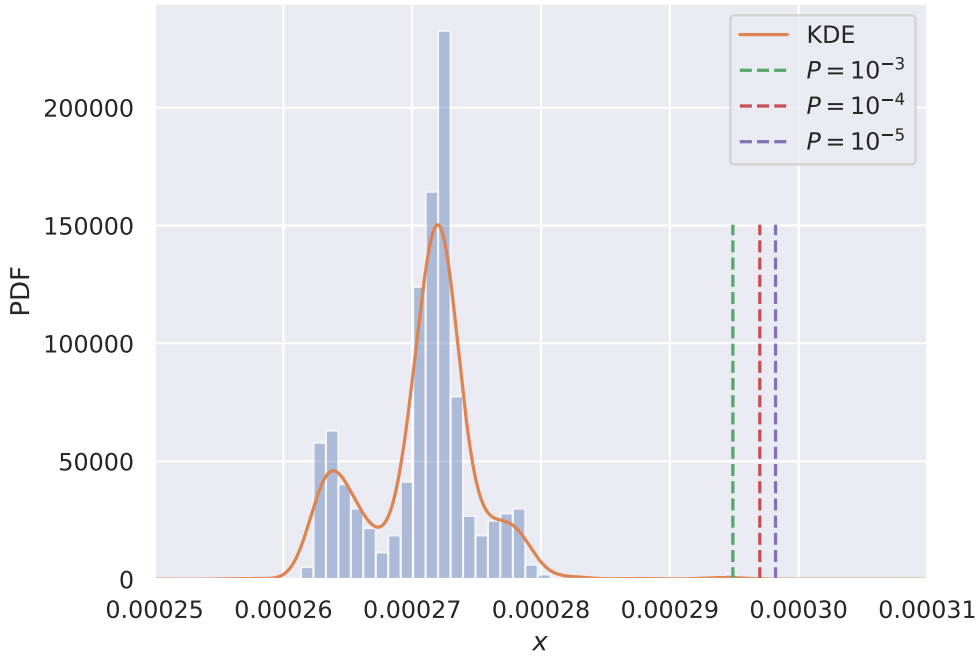
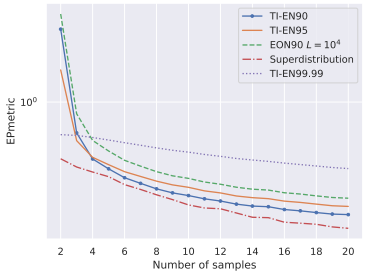
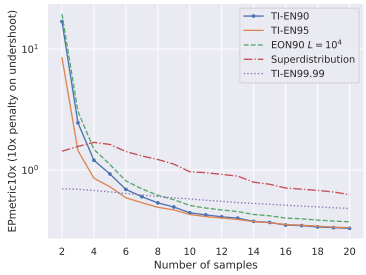


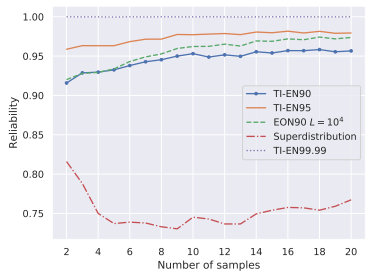
Figure H.29: Histogram and KDE for Tearing Parameter Lid Buckle Element 0.25.



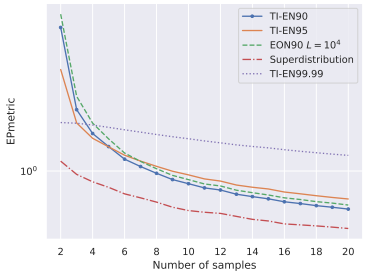
EPmetric, EP= 10^{-1}



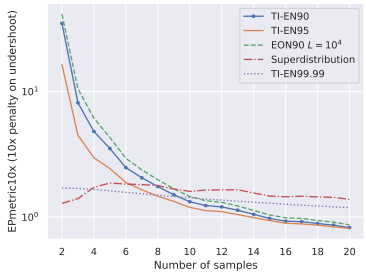
EPmetric10x, EP= 10^{-1}



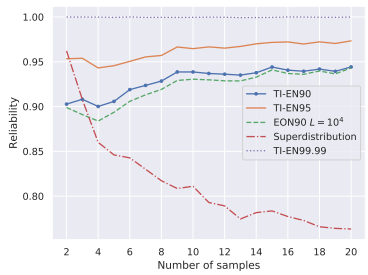
Reliability, EP= 10^{-1}



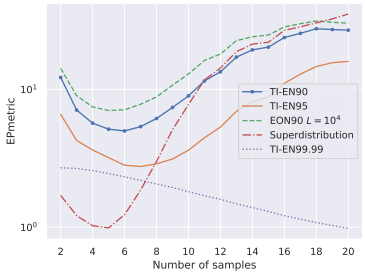
EPmetric, EP= 10^{-2}



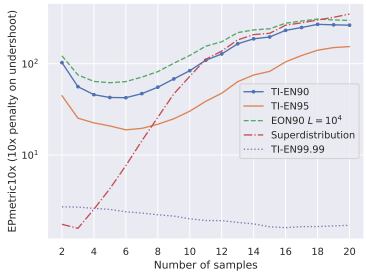
EPmetric10x, EP= 10^{-2}



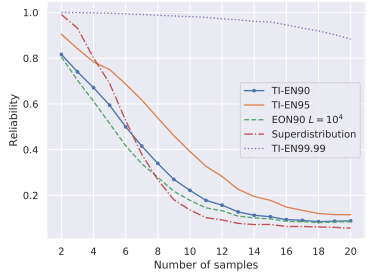
Reliability, EP= 10^{-2}



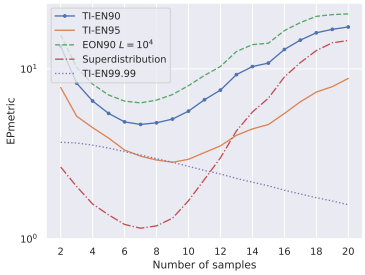
EPmetric, EP= 10^{-3}



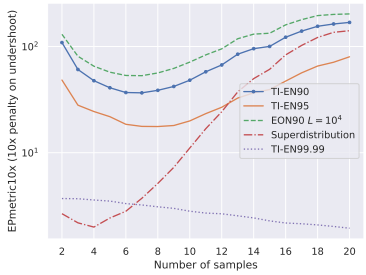
EPmetric10x, EP= 10^{-3}



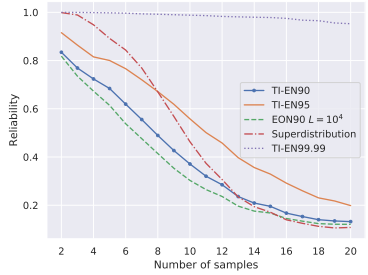
Reliability, EP= 10^{-3}



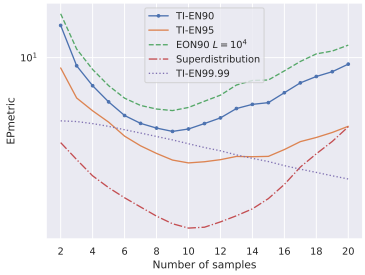
EPmetric, EP= 10^{-4}



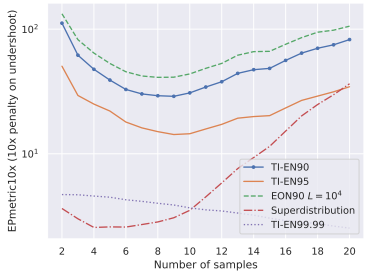
EPmetric10x, EP= 10^{-4}



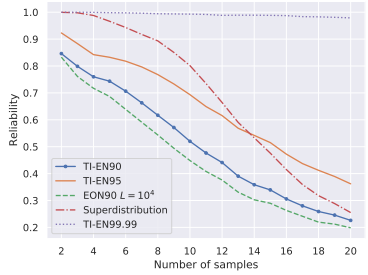
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

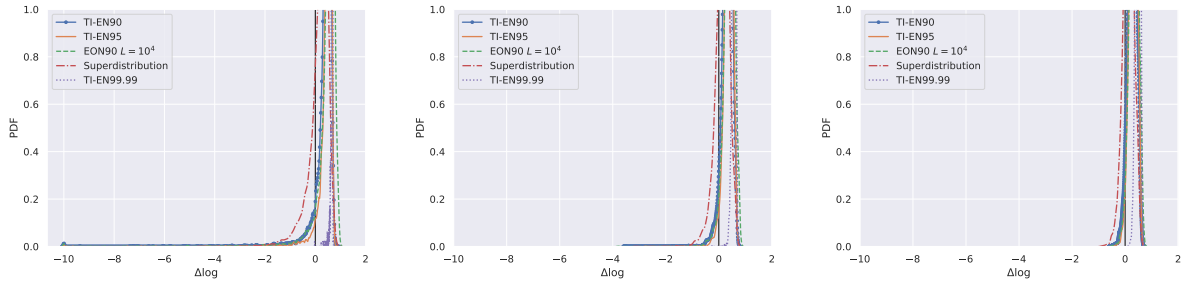


EPmetric10x, EP= 10^{-5}

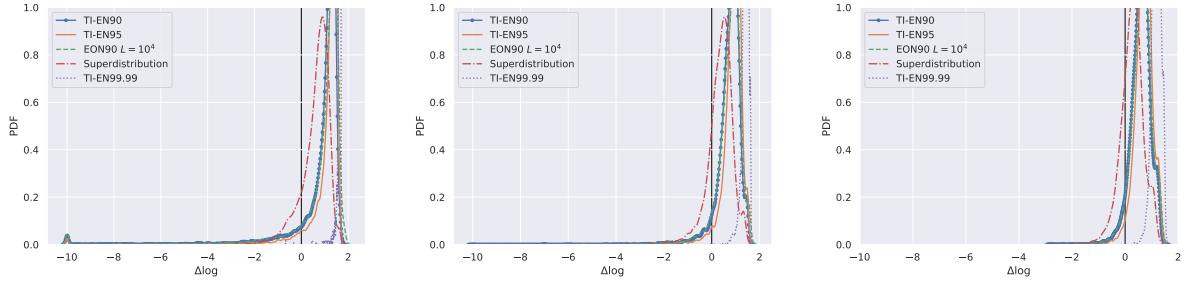


Reliability, EP= 10^{-5}

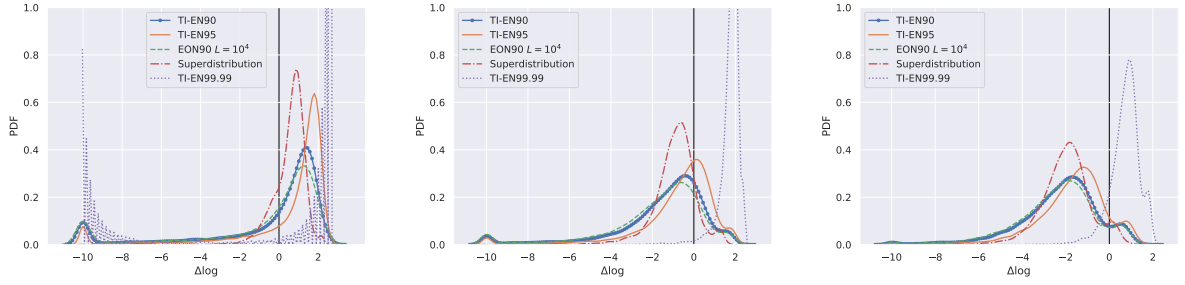
Figure H.30: Results for the empirical Tearing Parameter Lid Buckle Element 0.25 distribution.



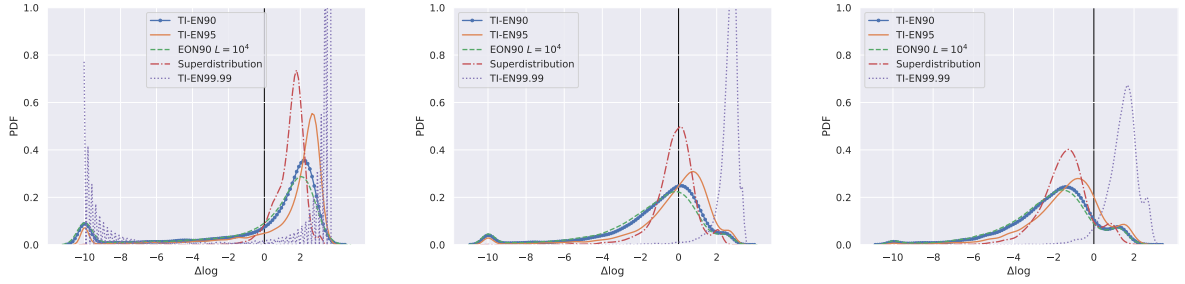
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



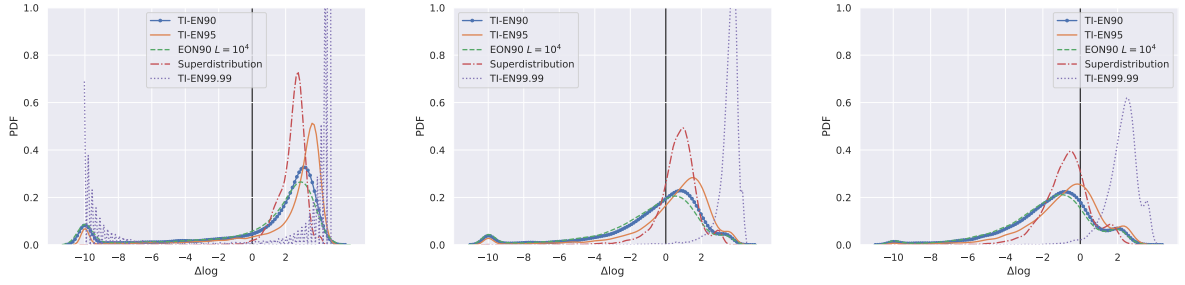
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.31: Distribution of results for the empirical Tearing Parameter Lid Buckle Element 0.25 distribution.

H.7 Tearing Parameter Weld Max Global 0.25

A histogram and KDE of the Tearing Parameter Weld Max Global 0.25 data is shown in Figure H.32. The results are shown in Figures H.33 and H.34.

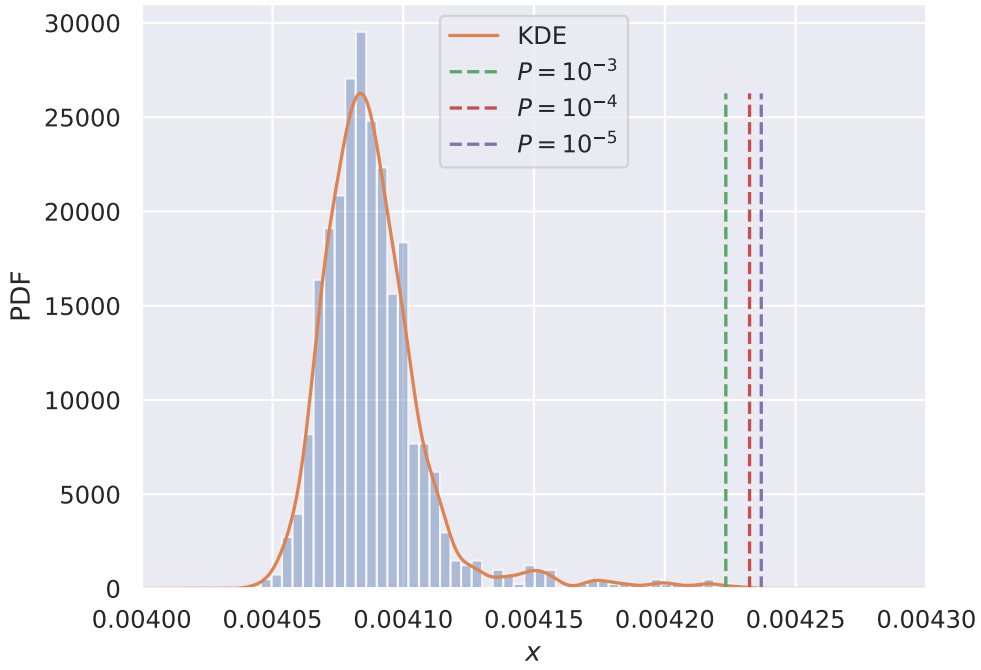
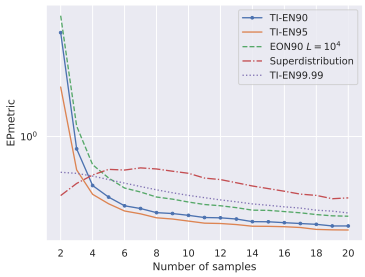
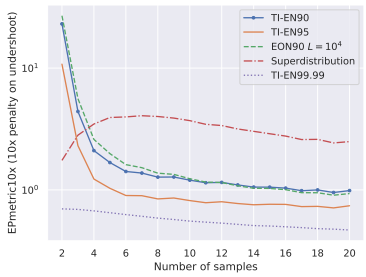


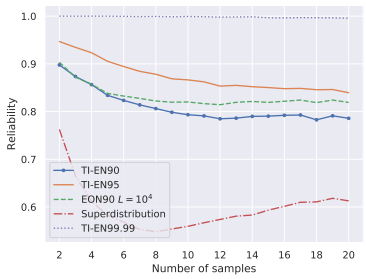
Figure H.32: Histogram and KDE for Tearing Parameter Weld Max Global 0.25.



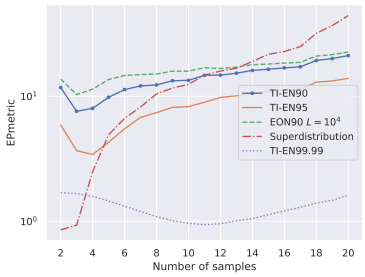
EPmetric, EP= 10^{-1}



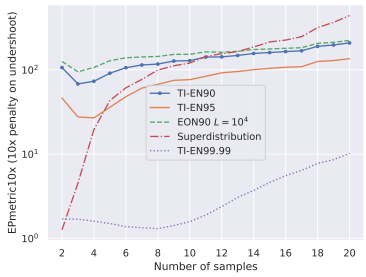
EPmetric10x, EP= 10^{-1}



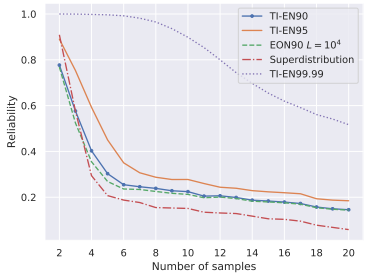
Reliability, EP= 10^{-1}



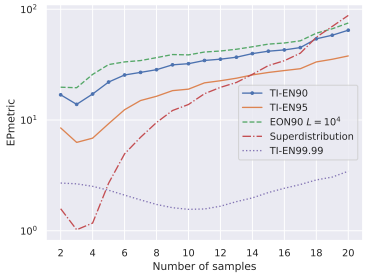
EPmetric, EP= 10^{-2}



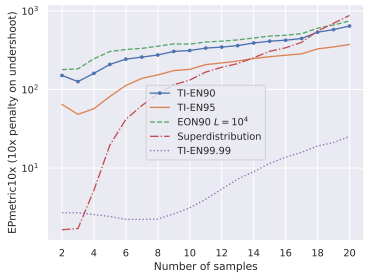
EPmetric10x, EP= 10^{-2}



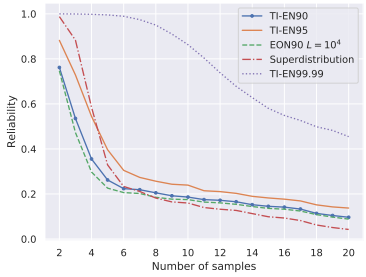
Reliability, EP= 10^{-2}



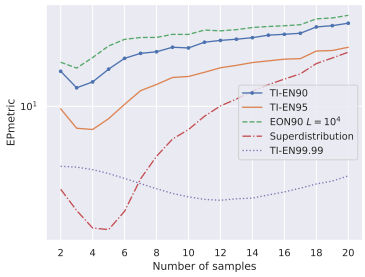
EPmetric, EP= 10^{-3}



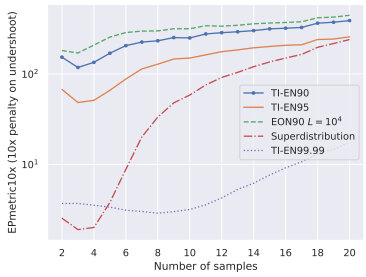
EPmetric10x, EP= 10^{-3}



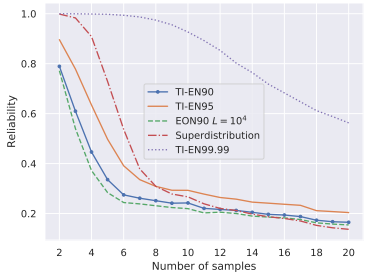
Reliability, EP= 10^{-3}



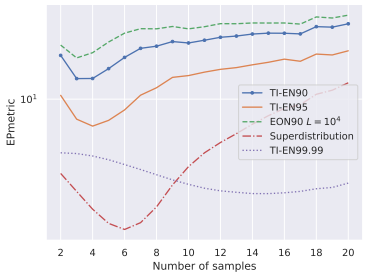
EPmetric, EP= 10^{-4}



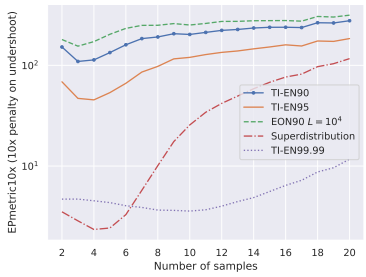
EPmetric10x, EP= 10^{-4}



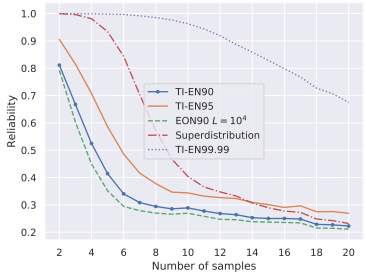
Reliability, EP= 10^{-4}



EPmetric, EP= 10^{-5}

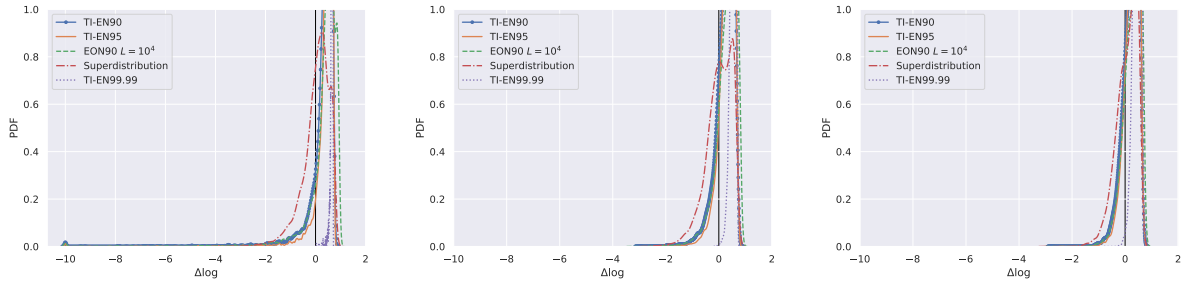


EPmetric10x, EP= 10^{-5}

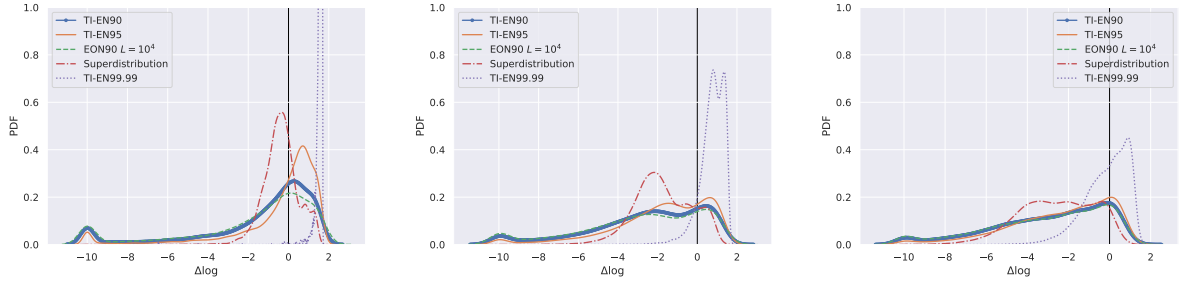


Reliability, EP= 10^{-5}

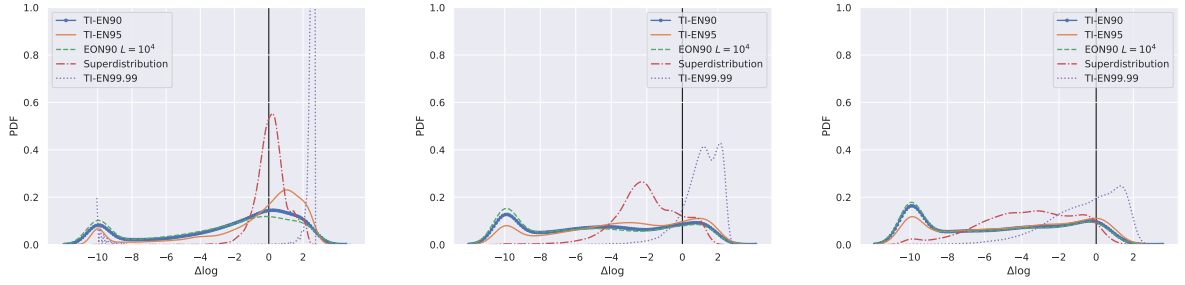
Figure H.33: Results for the empirical Tearing Parameter Weld Max Global 0.25 distribution.



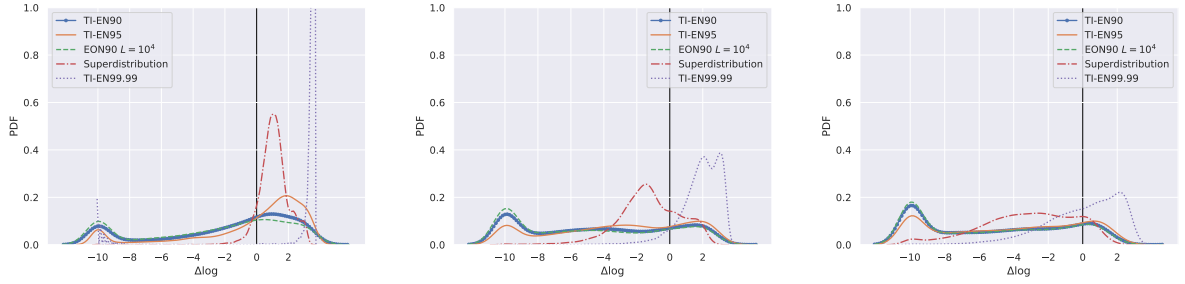
$N = 4, EP = 10^{-1}$ $N = 10, EP = 10^{-1}$ $N = 20, EP = 10^{-1}$



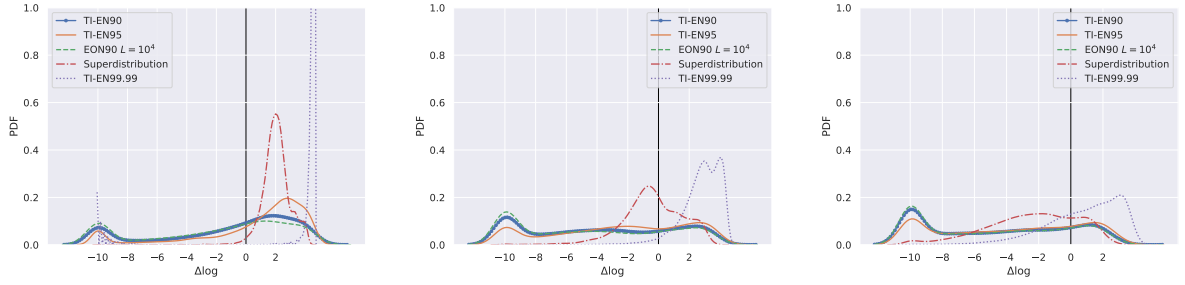
$N = 4, EP = 10^{-2}$ $N = 10, EP = 10^{-2}$ $N = 20, EP = 10^{-2}$



$N = 4, EP = 10^{-3}$ $N = 10, EP = 10^{-3}$ $N = 20, EP = 10^{-3}$



$N = 4, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$



$N = 4, EP = 10^{-5}$ $N = 10, EP = 10^{-5}$ $N = 20, EP = 10^{-5}$

Figure H.34: Distribution of results for the empirical Tearing Parameter Weld Max Global 0.25 distribution.

I TI-EN90 and EON resampling results

The results of using the resampling techniques described in Section 4 for the TI-EN90 and EON90 are shown here. The trends observed on the Superdistribution are similar to the results in this section. In general the *NCr* technique is promising for improving the estimated EP with larger number of samples. The *NCr* always resulted in more conservatively estimated EPs. It did not appear that Bootstrapping was successful at improving the EP predictions, as Bootstrapping resulted in significantly more variability in the results. Additionally, the Bootstrapping results were very near to the results without resampling.

I.1 Student's t-distribution

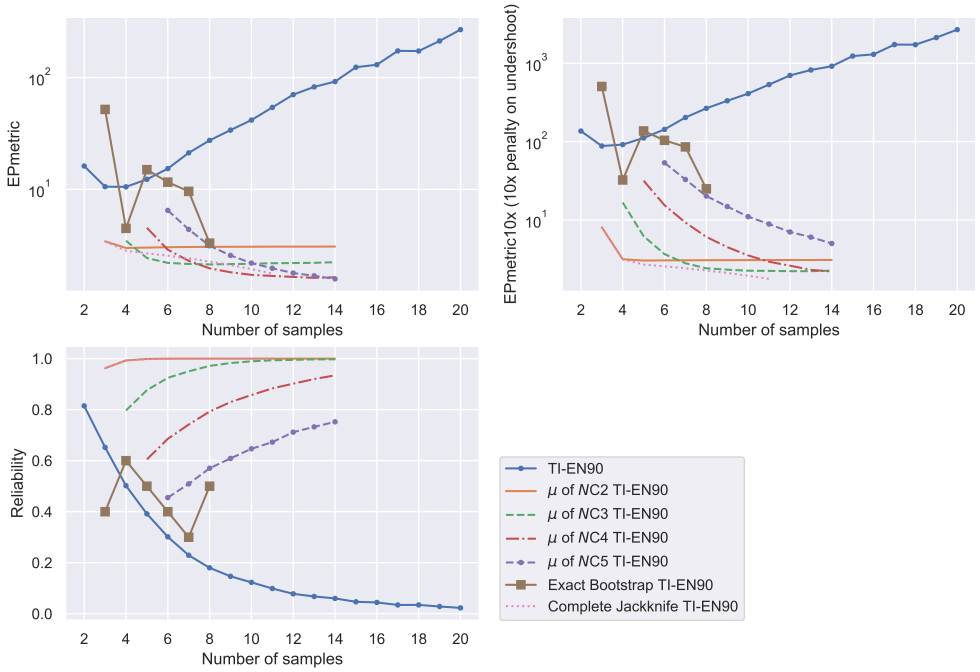


Figure I.35: The EPmetric, EPmetric10x, and reliability for the TI-EN90 with Bootstrapping and Jackknifing on Student's t-distribution for $EP = 10^{-4}$.

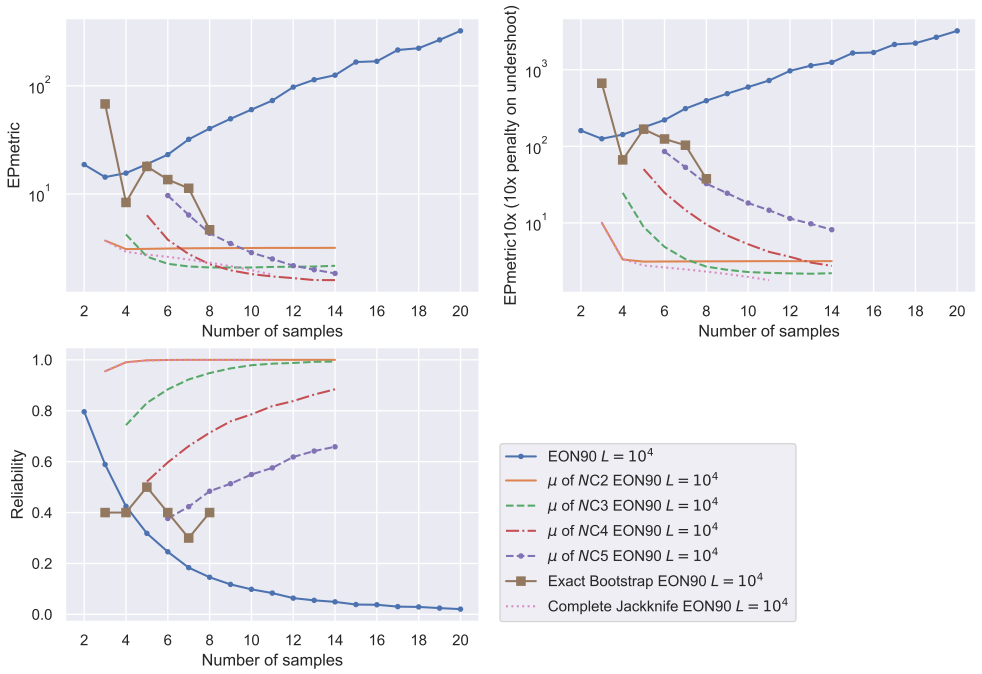


Figure I.36: The EPmetric, EPmetric10x, and reliability for the EON 90 with Bootstrapping and Jackknifing on Student's t-distribution for $EP = 10^{-4}$.

I.2 Weibull Narrow distribution

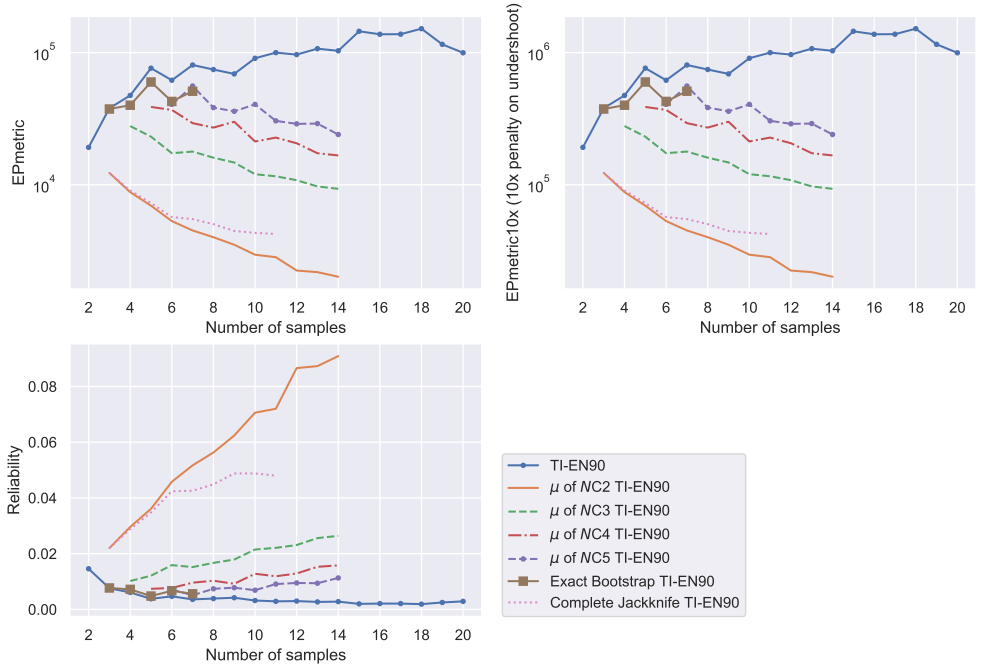


Figure I.37: The EPmetric, EPmetric10x, and reliability for the TI-EN90 with Bootstrapping and Jackknifing on the Weibull Narrow distribution for $EP = 10^{-4}$.

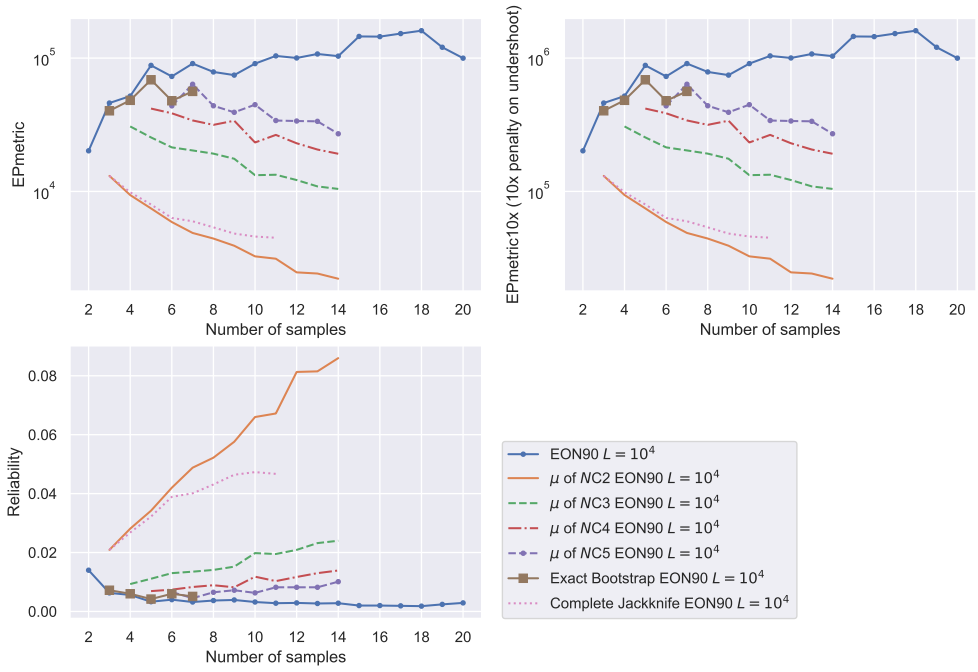


Figure I.38: The EPmetric, EPmetric10x, and reliability for the EON 90 with Bootstrapping and Jackknifing on the Weibull Narrow distribution for $EP = 10^{-4}$.

I.3 Exponential Wide distribution

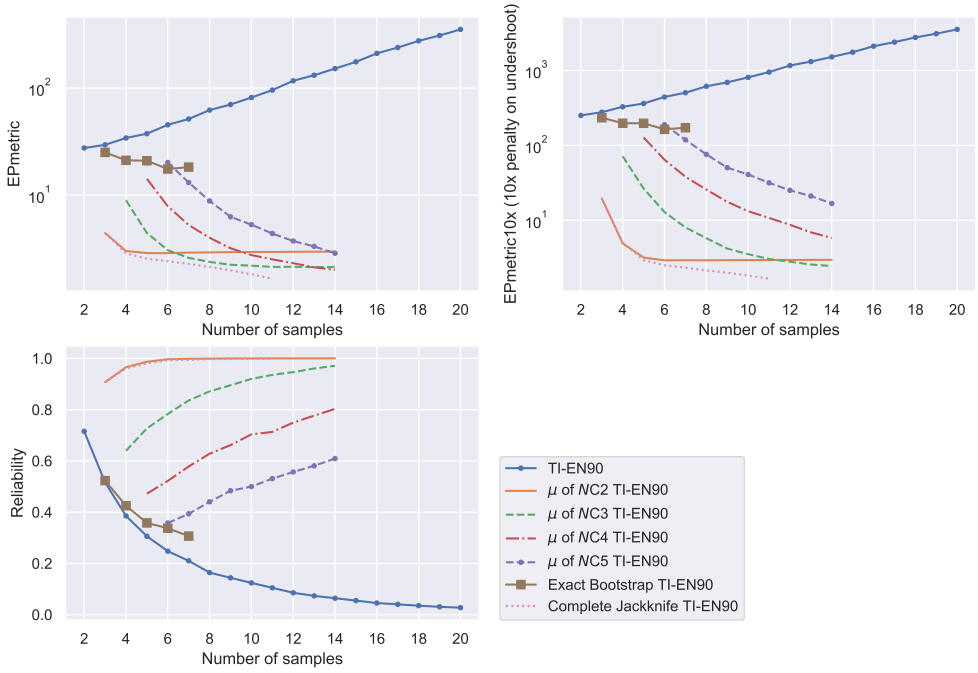


Figure I.39: The EPmetric, EPmetric10x, and reliability for the TI-EN90 with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-4}$.

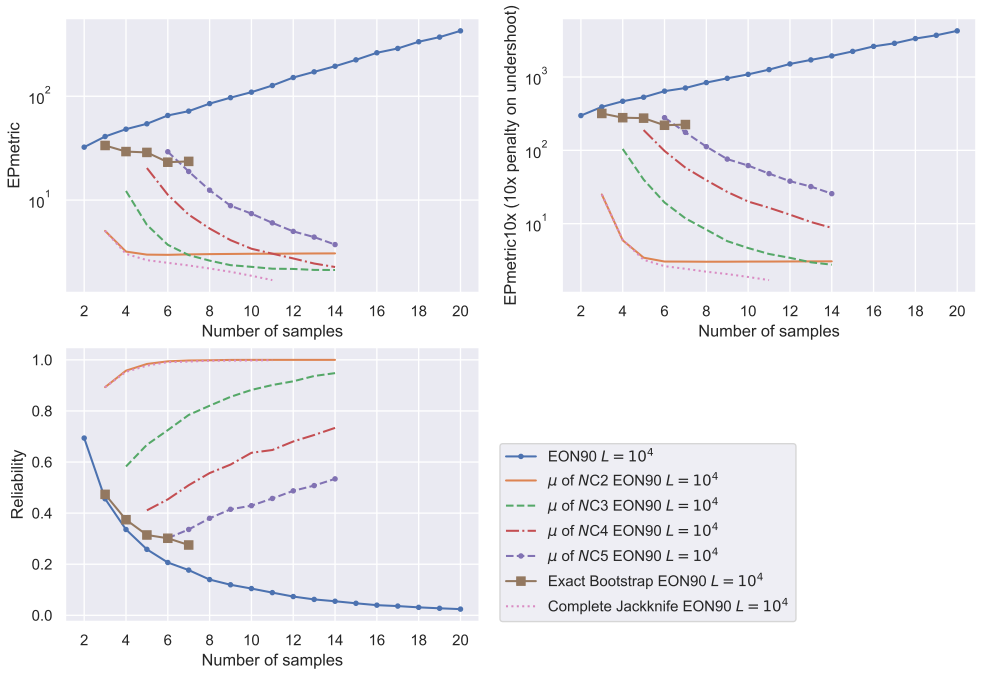


Figure I.40: The EPmetric, EPmetric10x, and reliability for the EON 90 with Bootstrapping and Jackknifing on the Exponential Wide distribution for $EP = 10^{-4}$.

I.4 Standard Normal distribution

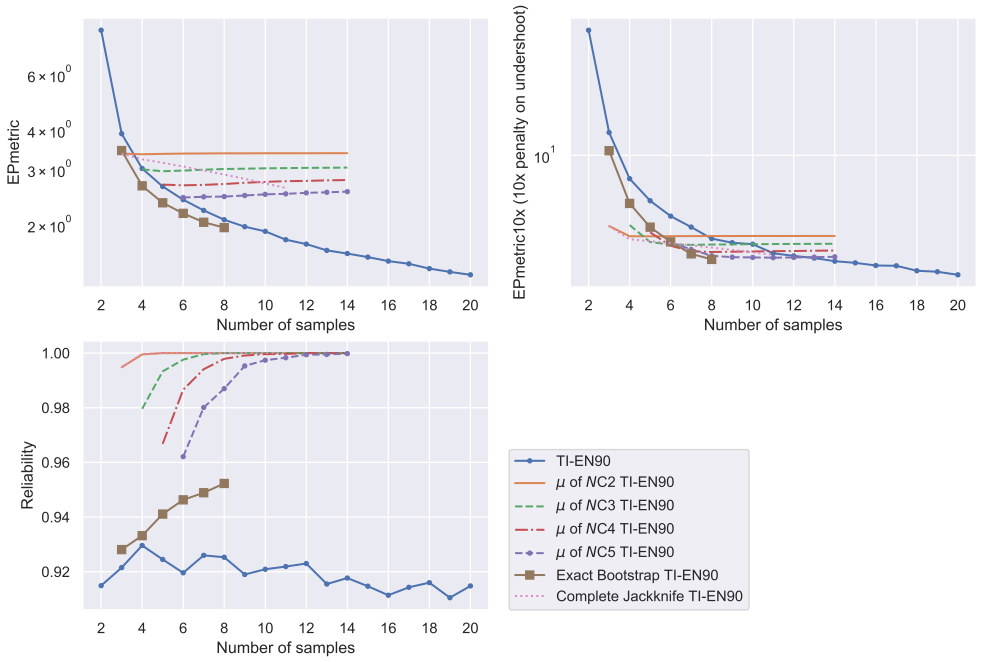


Figure I.41: The EPmetric, EPmetric10x, and reliability for the TI-EN90 with Bootstrapping and Jackknifing on the standard Normal distribution for $EP = 10^{-4}$.

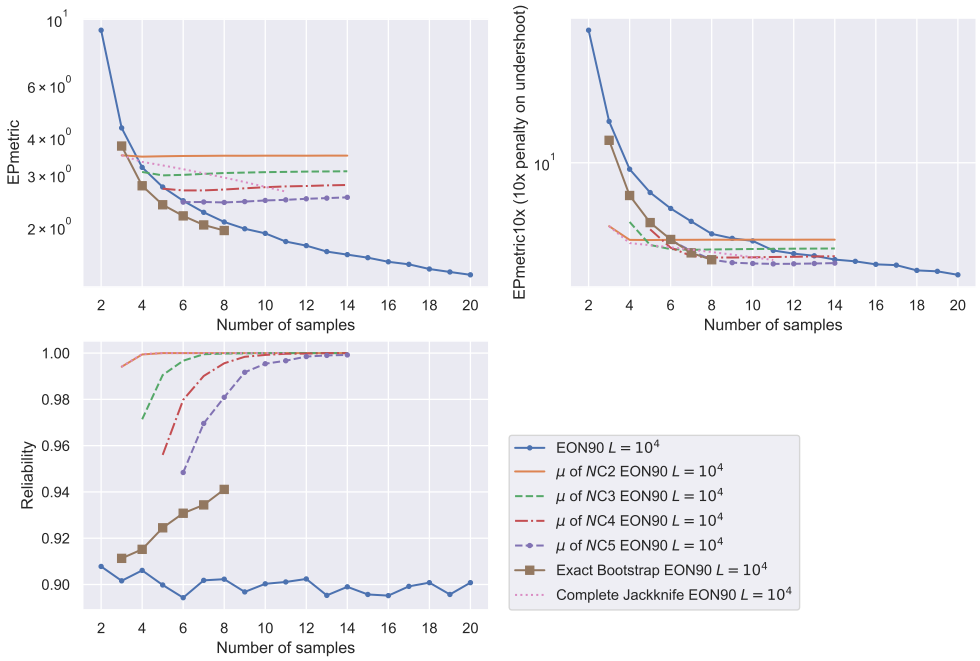


Figure I.42: The EPmetric, EPmetric10x, and reliability for the EON 90 with Bootstrapping and Jackknifing on the standard Normal distribution for $EP = 10^{-4}$.

I.5 Bi-modal Log-Gamma Normal distribution

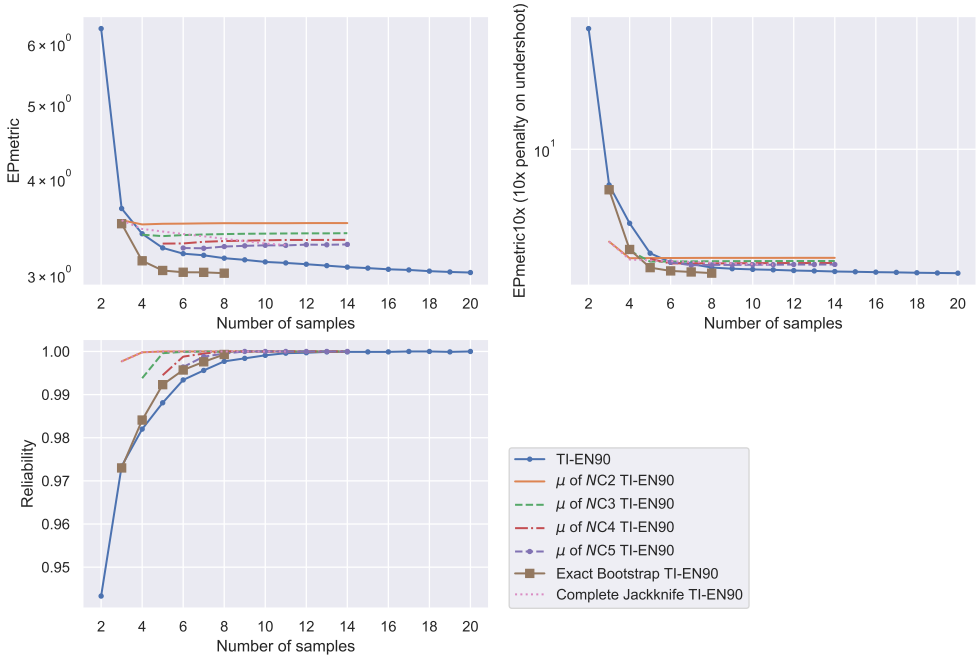


Figure I.43: The EPmetric, EPmetric10x, and reliability for the TI-EN90 with Bootstrapping and Jackknifing on the bi-modal Log-Gamma Normal distribution for $EP = 10^{-4}$.

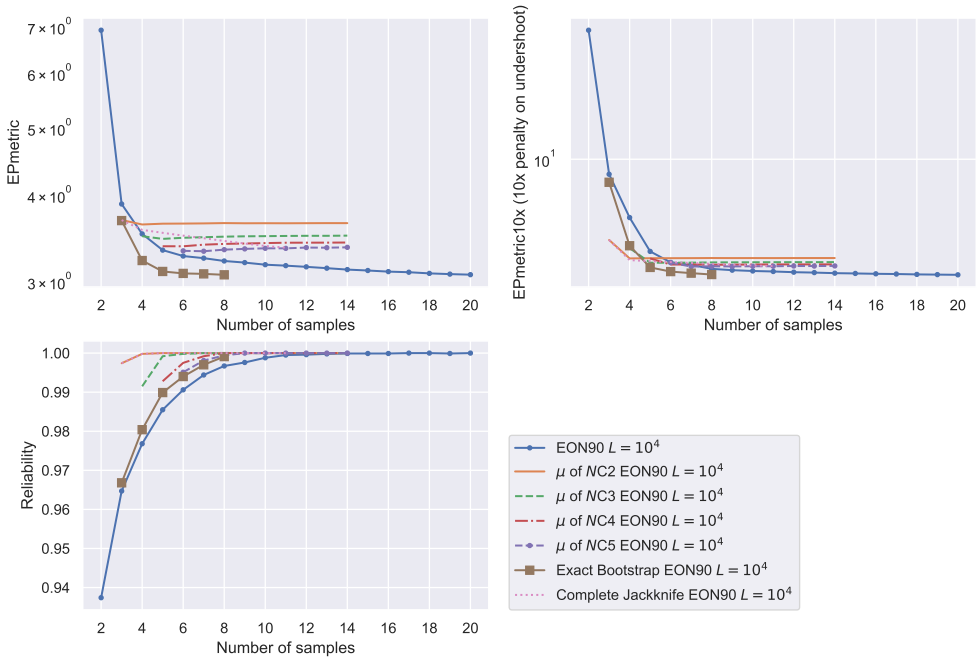


Figure I.44: The EPmetric, EPmetric10x, and reliability for the EON 90 with Bootstrapping and Jackknifing on the bi-modal Log-Gamma Normal distribution for $EP = 10^{-4}$.

I.6 Empirical Tensile EQPS Lid Buckle Element 0.25

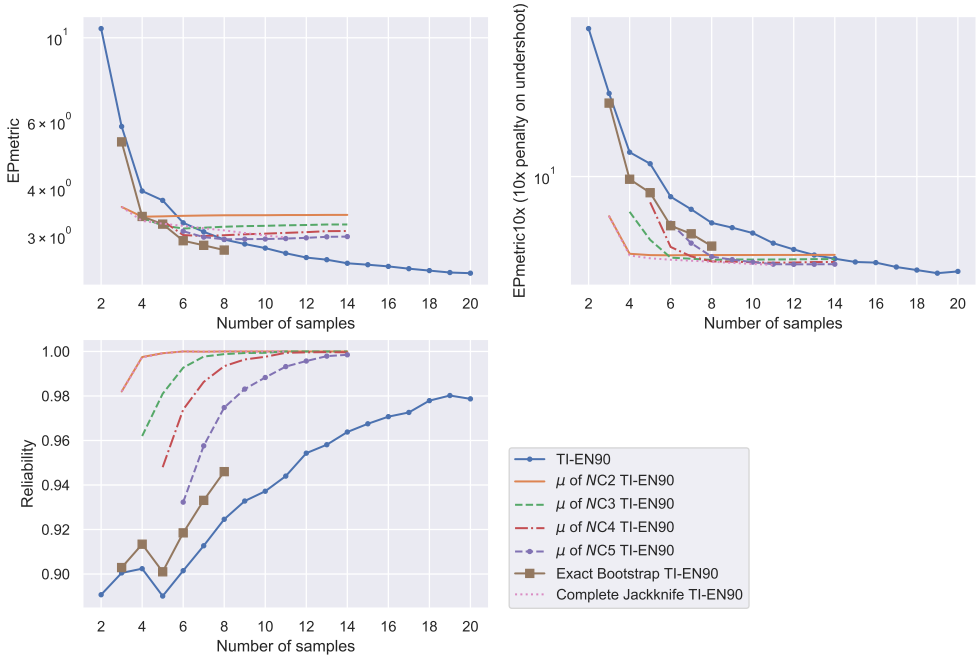


Figure I.45: The EPmetric, EPmetric10x, and reliability for the TI-EN90 with Bootstrapping and Jackknifing on the empirical Tensile EQPS Lid Buckle Element 0.25 distribution for $EP = 10^{-4}$.

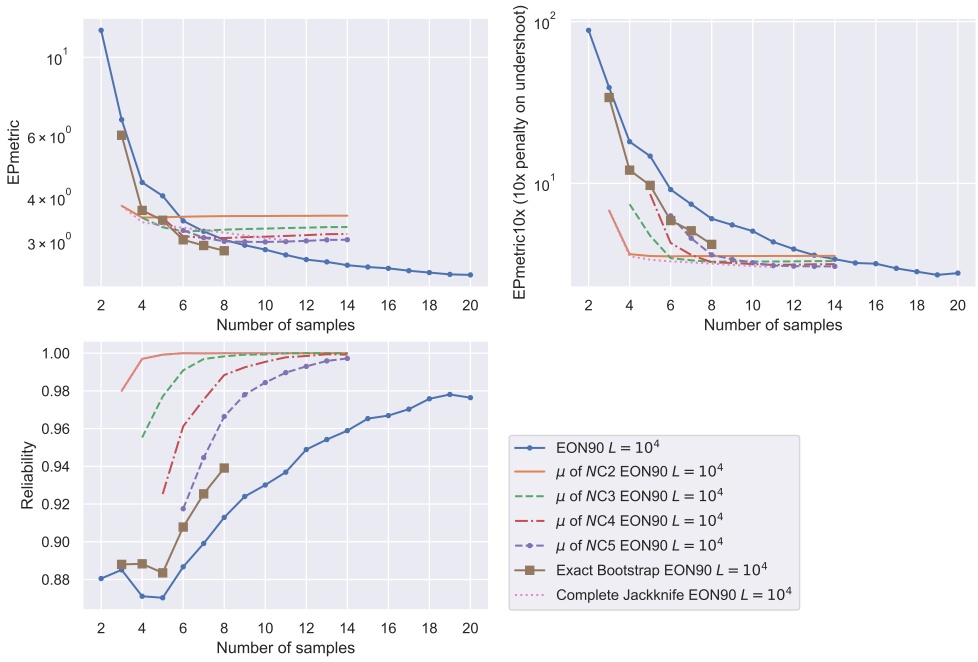


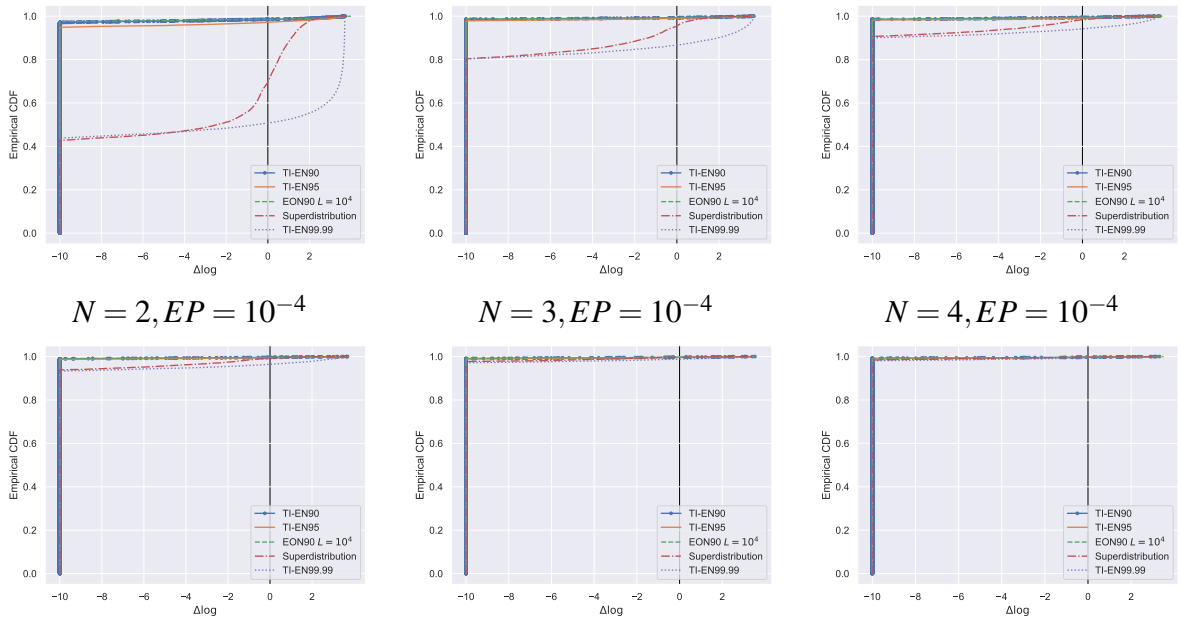
Figure I.46: The EPmetric, EPmetric10x, and reliability for the EON 90 with Bootstrapping and Jackknifing on the empirical Tensile EQPS Lid Buckle Element 0.25 distribution for $EP = 10^{-4}$.

J Artifacts in empirical PDFs constructed from 10,000 trials

Several artifacts appear in the empirical PDFs of the Δlog results in Section 3. This is particular evident in Figure 34 from the Weibull Narrow distribution, where some of the distributions appear to be random noise. This is an artifact of how the distributions were plotted, in which there are a series of unusual delta functions in the PDF.

A better way to plot the Δlog results may have been with empirical CDFs instead. Figure J.47 shows a few empirical CDFs from the 10,000 trials on the Weibull Narrow distribution for a $EP = 10^{-4}$. The empirical CDFs appear smooth, while the empirical PDFs of Figure 34 appeared like random noise.

There are a few notable advantages of plotting empirical CDFs instead of empirical PDFs of the results. One advantage of an empirical CDF is that it is easy to read the reliability (percentages of conservative estimates) from the empirical CDFs. The reliability would be the percentage of Δlog values that had a positive value. The other advantage is that the x and y axis of empirical CDFs can be easily constrained for all plots, while PDFs may required different scaling on the y axis in order to visualize the result. The disadvantage of using CDFs instead of PDFs is that it is much harder to identify the distribution of the random trials from the CDF.



$N = 2, EP = 10^{-4}$ $N = 3, EP = 10^{-4}$ $N = 4, EP = 10^{-4}$

$N = 5, EP = 10^{-4}$ $N = 10, EP = 10^{-4}$ $N = 20, EP = 10^{-4}$

Figure J.47: Empirical CDFs of results for the Weibull Narrow distribution.

DISTRIBUTION:

1	MS	0153	E. Blansett, 2832	(electronic copy)
1	MS	0525	T.P. Swiler, 2643	(electronic copy)
1	MS	0825	M. Pilch, 1544	(electronic copy)
1	MS	0828	A. Urbina, 1544	(electronic copy)
1	MS	0828	B. Carnes, 1544	(electronic copy)
1	MS	0828	K.J. Dowding, 1544	(electronic copy)
1	MS	0828	A. Eckert, 1544	(electronic copy)
1	MS	0828	B. Freno, 1544	(electronic copy)
1	MS	0828	K.W. Irick, 1544	(electronic copy)
1	MS	0828	C. Jekel, 1544	(electronic copy)
1	MS	0828	S. Kieweg, 1544	(electronic copy)
1	MS	0828	A. Krueger, 1544	(electronic copy)
1	MS	0828	B. Lance, 1544	(electronic copy)
1	MS	0828	J. Mullins, 1544	(electronic copy)
1	MS	0828	G.E. Orient, 1544	(electronic copy)
1	MS	0828	J.R. Red-Horse, 1544	(electronic copy)
1	MS	0828	V.J. Romero, 1544	(electronic copy)
1	MS	0828	B. Schroeder, 1544	(electronic copy)
1	MS	0828	J. Winokur, 1544	(electronic copy)
1	MS	0829	A. Doser, 6673	(electronic copy)
1	MS	0829	B.M. Rutherford, 6673	(electronic copy)
1	MS	0829	L. Wilson, 6673	(electronic copy)
1	MS	0829	A. Zang, 6673	(electronic copy)

1	MS	0845	W.R. Witkowski, 1540	(electronic copy)
1	MS	1168	B.S. Paskaleva, 1356	(electronic copy)
1	MS	1168	S. Wix, 1356	(electronic copy)
1	MS	1173	A. Mar, 5499	(electronic copy)
1	MS	1177	H. Thornquist, 1355	(electronic copy)
1	MS	1179	J.P. Castro, 1341	(electronic copy)
1	MS	1318	G. Geraci, 1463	(electronic copy)
1	MS	1318	J. Jakeman, 1463	(electronic copy)
1	MS	1318	L.P. Swiler, 1463	(electronic copy)
1	MS	1320	V.G. Weirs, 1446	(electronic copy)
1	MS	1323	W.J. Rider, 1446	(electronic copy)
1	MS	9006	P.D. Hough, 8754	(electronic copy)
1	MS	9042	C. Lam, 8715	(electronic copy)
1	MS	9042	S.M. Nelson, 8752	(electronic copy)
1	MS	0899	Technical Library, 9536	(electronic copy)


Sandia National Laboratories