Proceedings of ASME 2019
Verification and Validation Symposium
VVS2019
May 15-17, 2019, Las Vegas, NV, USA

VVS2019-5127

BOOTSTRAPPING AND JACKKNIFE RESAMPLING TO IMPROVE SPARSE-SAMPLE UQ METHODS FOR TAIL PROBABILITY ESTIMATION*

Charles F Jekel
R&D Graduate Intern
Sandia National Laboratories‡
Dept. of Mechanical and Aerospace Engineering
University of Florida, Gainesville, Florida 32611

Vicente Romero†
V&V, UQ, and Credibility Processes Dept.
Sandia National Laboratories‡
Albuquerque, New Mexico 87185
vjromer@sandia.gov

ABSTRACT

Tolerance Interval Equivalent Normal (TI-EN) and Superdistribution (SD) sparse-sample uncertainty quantification (UQ) methods are used for conservative estimation of small tail probabilities. These methods are used to estimate the probability of a response laying beyond a specified threshold with limited data. The study focused on sparse-sample regimes ranging from $N = 2$ to 20 samples, because this is reflective of most experimental and some expensive computational situations. A tail probability magnitude of 10^{-4} was examined on four different distribution shapes, in order to be relevant for quantification of margins and uncertainty (QMU) problems that arise in risk and reliability analyses. In most cases the UQ methods were found to have optimal performance with a small number of samples, beyond which the performance deteriorated as samples were added. Using this observation, a generalized Jackknife resampling technique was developed to average many smaller subsamples. This improved the performance of the SD and TI-EN methods, specifically when a larger than optimal number of samples were available. A Complete Jackknifing technique, which considered all possible sub-sample combinations, was shown to perform better in most cases than an alternative Bootstrap resampling technique.

INTRODUCTION

Estimation of tail probability magnitudes is a challenging task with only a small number of samples and no prior knowledge of the true distribution. Tolerance Interval Equivalent Normal (TI-EN) and Superdistribution (SD) sparse-sample uncertainty quantification (UQ) methods described and investigated in [1–3] and briefly described in the appendices of this paper have recently been studied along with other sparse-sample UQ methods to estimate the probability of a random event or occurrence beyond a specified threshold. This is commonly known as exceedance probability (EP) estimation or tail probability estimation. While it may always be desirable to have additional samples [4], in practice the number of samples is generally limited by some cost restriction. This is especially true for some of the most expensive experiments or largest computer simulations, where only a few test replicates or samples N can be performed, e.g. $N = 2$ to 20 examined in this study.

The prior studies in [1] and [2] found that SD and TI-EN95 (targeted to a 95% confidence level per Appendix A) typically perform best of the sparse-sample UQ methods studied under sparse-data conditions. However, an interesting phenomenon occurs with these methods. It is generally anticipated that additional samples will improve EP estimation. However, as estab-

*Sandia National Laboratories document SAND2019-3256 C. This paper is a work of the United States Government and is not subject to copyright protection in the U.S.

†Address all correspondence to this author.

‡Sandia National Laboratories is a multi-mission 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 Energys National Nuclear Security Administration under contract DE-NA0003525.

lished in [3], there is typically some optimal number of samples for peak accuracy and/or reliability of conservative estimation with the TI-EN and SD methods which are based on the Normal distribution. When additional samples are added beyond the optimal number it may become evident that the distribution is non-normal, and thus the accuracy and reliability of the estimated EP deteriorates.

This is according to a performance measure (see next section) assessed over thousands of random-sample trials which considers both accuracy of the estimate relative to the true EP value and the reliability rate or proportion of conservative estimates that do not under-estimate the true EP. While the reasons are understood for the performance deterioration (which occurs for many of the 16 distributions studied in [1] and [3]), this effect leaves something to be desired when more samples are available than what yields optimal performance with the TI-EN or SD methods and the said distributions.

Using the observation that a small/optimal number of samples yields most accurate and reliable EP estimate in these cases, a generalized Jackknifing resampling technique is studied in this paper in an attempt to improve TI-EN and SD performance for larger-than-optimal sample sizes. This Jackknife resampling technique averages many smaller subsample estimates in an effort to improve the performance for a larger sample size. This is an initial study to investigate whether Jackknifing can be used to improved the performance of the TI-EN and SD sparse-sample UQ methods. Bootstrapping was also investigated as an alternative resampling technique to Jackknifing.

The focus of this study is on sparse-sample regimes ranging from $N = 2$ to 20 samples. A tail probability magnitude of 10^{-4} is examined in order for the study to be relevant to quantification of margins and uncertainty (QMU) problems that arise in risk and reliability analyses. The sparse-sample UQ methods combined with resampling will be analyzed for accuracy and reliability on four distributions in this initial study. These distributions are the Standard Normal, Student's t, Weibull, and Exponential.

METHODS

The goal of this study is to quantify the accuracy and reliability of EP predictions from sparse samples. This is done by evaluating many random trials from known statistical distributions. The process for conducting these random trials follows.

First, a number of N random samples were generated from a given statistical distribution. From this random sample, an EP is estimated beyond a threshold that corresponds to a probability of 10^{-4}. The EP was estimated using the following UQ methods: TI-EN, SD, TI-EN with Jackknifing, SD with Jackknifing, TI-EN with Bootstrapping, and SD with Bootstrapping. A 95% confidence level was used with the TI-EN. The TI-EN and SD methods are summarized in this papers appendix and the Jackknifing and Bootstrapping extensions are explained later in this section.

Each EP estimate was compared to the true EP of 10^{-4}. This process was repeated 10,000 times for each N number of samples, where $N = 2, 3, \cdots, 20$. This was done for reach of the four said distributions.

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

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

$$\Delta\log = \log_{10}(EP_{\text{estimated}}) - \log_{10}(EP_{\text{true}}) \tag{2}$$

where N^{+} and N^{-} are the numbers of overshoot ($>$, conservative) and undershoot ($<$, unconservative) cases respectively. For this study the total number of trials was $N^{+} + N^{-} = 10,000$.

In performance or safety related design and analysis applications it is usually preferred to overestimate EP than to underestimate it. Thus, for a given numerator sum of overshoot and undershoot error magnitudes in Eqn. 1, the greater the number of overshoot errors contributing to that sum, the higher the proportion of conservative errors, indicating better method performance, and the larger the denominator in the performance metric Eqn. 1. The resulting lower overall ratio (metric value) in Eqn. 1 thus correlates with better method performance. This is true in general with our performance metric: lower metric value corresponds to better method performance.

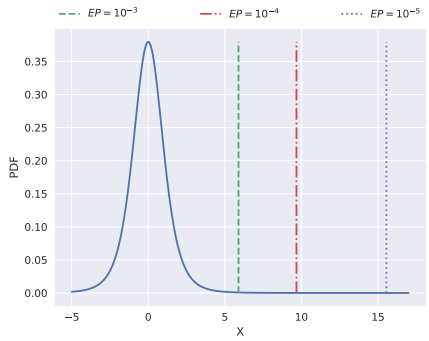

Figure 1. Student's t-distribution with 5 degrees of freedom and locations of EPs of 10^{-3} to 10^{-5}.

The performance metric combines aspects of both EP estimation accuracy and reliability of being conservative. Because there is often a desired or required lowest acceptable level of reliability of obtaining an EP estimate that is conservative, an EP

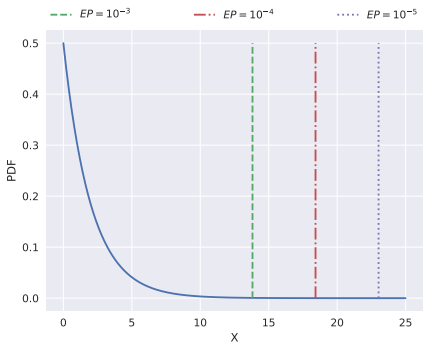

Figure 2. Exponential Wide distribution with $\lambda = 0.5$ and locations of EPs of 10^{-3} to 10^{-5}.

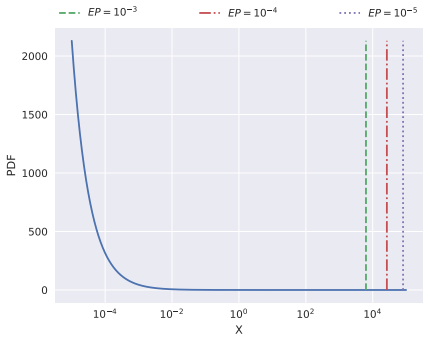

Figure 3. Weibull Narrow distribution with $\alpha = 0.2$ and $\beta = 0.4$, as well as the locations of EPs of 10^{-3} to 10^{-5}.

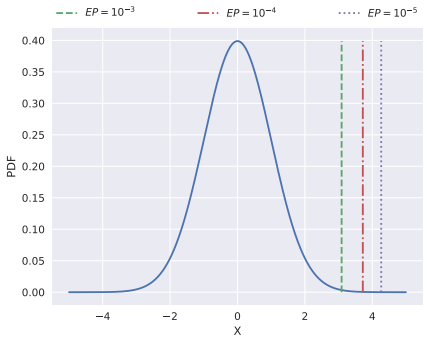

Figure 4. Standard Normal distribution with the locations of EPs of 10^{-3} to 10^{-5}.

estimate that is conservative, a reliability measure alone is also calculated and reported defined as

$$\text{Reliability} = \frac{N^+}{\text{the total number of trials}}. \tag{3}$$

The statistical distributions studied are: 1) Students t-distribution with 5 degrees of freedom, 2) Exponential Wide distribution with $\lambda = 0.5$, 3) Weibull Narrow distribution with $\alpha = 0.2$ and $\beta = 0.4$, 4) Standard Normal distribution. The threshold location for EPs on each distribution is shown in Figures 1-4. The Bootstrap and Jackknife resampling techniques are explained in the next two subsections.

Bootstrapping

Statistical resampling methods have been used to reduce the bias (or error) in an estimated statistic [5]. 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 possible true value of the statistic. An estimated EP can be found by averaging the EP estimates from the combinations with replacement. Bootstrapping was used by Picheny et al. [6] to more conservatively estimate a 99% tail probability from 100 samples. In this paper 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

$$\binom{2N-1}{N} = \frac{(2N-1)!}{N!(N-1)!} \tag{4}$$

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* Bootstrapping resampling [7].

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.

Generalized Jackknife

Jackknifing is another popular resampling method that predates Bootstrapping. The Jackknife was originally created by Quenouille [8], and the term Jackknifing was coined by

Tukey [9]. 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 [10]. Typically m is a small number [11], and in many cases becomes one [10].

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

$$r = N - m \tag{5}$$

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 $r = 2,3,4,5$. 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. Additionally, Jackknifing considers subsamples while Bootstrapping considers new samples that are the same size of the original sample set.

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

$$\binom{N}{r} = \frac{N(N-1)\cdots(N-r+1)}{r(r-1)\cdots 1} = \frac{N!}{r!(N-r)!} \tag{6}$$

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 5 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.

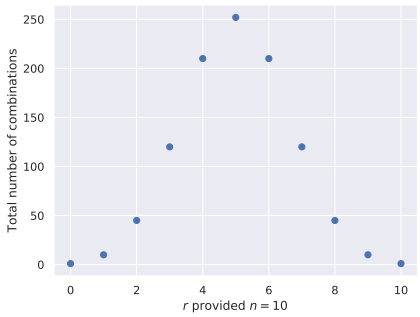

Figure 5. Example of how the total number of possible combinations in NCr follows Pascal's triangle for $N = 10$.

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

[1, 2, 7] [1, 2, 5] [1, 2, 3] [1, 7, 5] [1, 7, 3]
[1, 5, 3] [2, 7, 5] [2, 7, 3] [2, 5, 3] [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

$$\sum_{k=2}^{N-1} \binom{N}{k} = 2^N - (N+2) \tag{7}$$

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:

$$4C3: [1,\ 2,\ 7] \quad [1,\ 2,\ 5] \quad [1,\ 7,\ 5] \quad [2,\ 7,\ 5]$$
$$4C2: [1,\ 2] \quad [1,\ 7] \quad [1,\ 5] \quad [2,\ 7] \quad [2,\ 5] \quad [7,\ 5].$$

RESULTS

This section investigates whether Bootstrapping and Jackknifing can be used to improve the SD and TI-EN95 predictions of tail probabilities with limited samples. The techniques were applied to the four distributions previously defined. The EPmetrics characterize the conservatism and accuracy performance of 10,000 random trials for any given number of samples. The reliability represents the percentage of EP estimates that were conservative from the 10,000 random trials.

Improvements to EP estimates were investigated using the Exact Bootstrapping method, along with *NC*2, *NC*3, *NC*4, *NC*5 Jackknifing, and the Complete Jackknife method. The number of samples for the Exact Bootstrapping ranged from $N = 3$ to $N = 8$ due to the cost limitations. 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 the SD or TI-EN95 UQ methods. As a reminder, the true EP for each distribution was 10^{-4}.

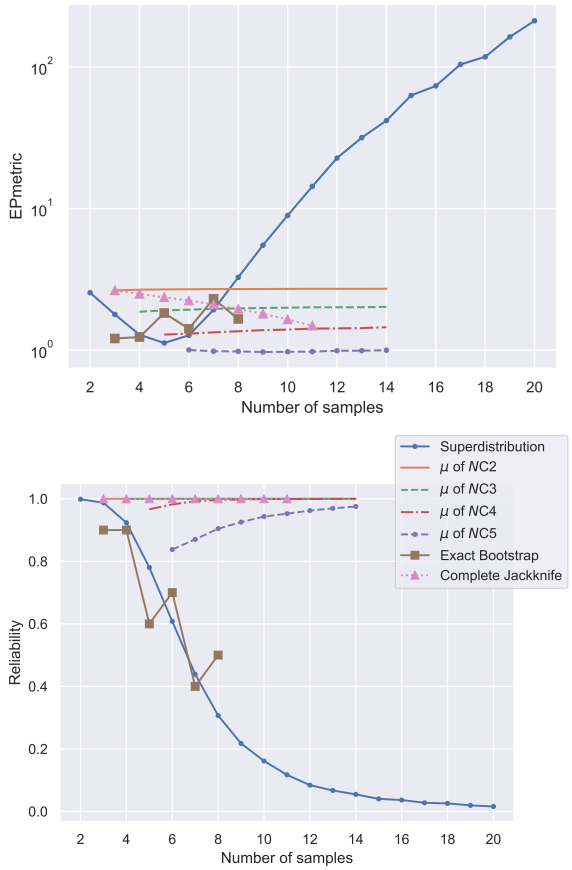

Figure 6. The EPmetric and reliability for the Superdistribution with Bootstrapping and Jackknifing on a 5 d-o-f Student's t-distribution.

Student's t-distribution

The Student's 5 d-o-f t-distribution results for SD with and without resampling are shown in Figure 6. 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.

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, consider that 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 appre-

ciably 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 $N = 2$ 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 $N = 5$ total samples.

The results of the TI-EN95 method with and without resampling applied to Student's t-distribution are shown in Figure 7. 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 resampling techniques, the *NC*5 SDJ results had both a lower EPmetric and higher reliability than the TI-EN95J *NC*5 results for all *N* tried. This is also true for optimal *NC*5 SDJ vs. optimal TI-EN95J *NCr* where $r = N_{\text{TIEN95opt}} = 4$.

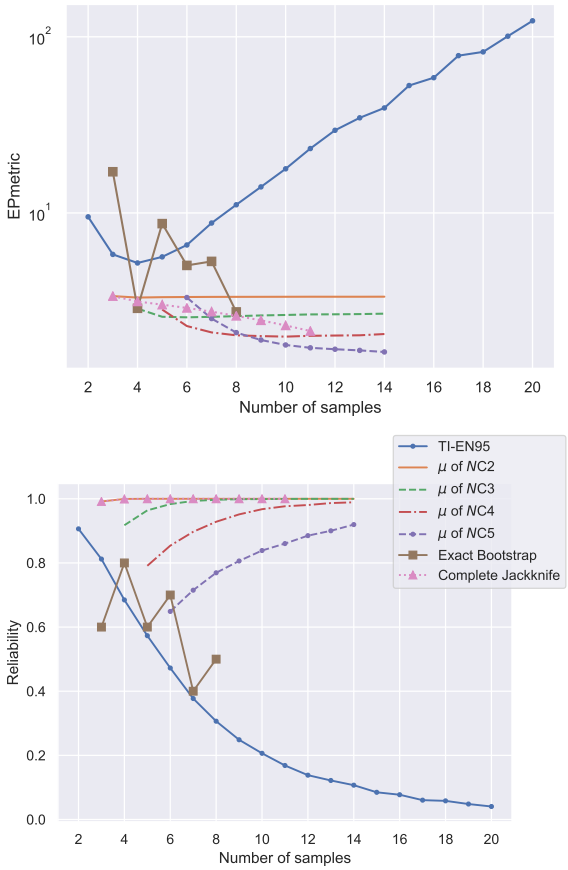

Figure 7. The EPmetric and reliability for the TI-EN95 with Bootstrapping and Jackknifing on Student's t-distribution.

Exponential Wide distribution

The results for SD with and without resampling applied to the Exponential Wide distribution are shown in Figure 8. Results are qualitatively very similar to those for the 5 d-o-f t-distribution. Very little difference exists between the SD results without resampling and those with Bootstrap resampling. Reliability and accuracy are roughly the same for both methods for any given number of samples.

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 $r = N_{\text{SDopt}} = 4$. *NC*5 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 *NC*4 SDJ), 5*C*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 *NC*4 SD-Jackknife method 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 $N = 12$).

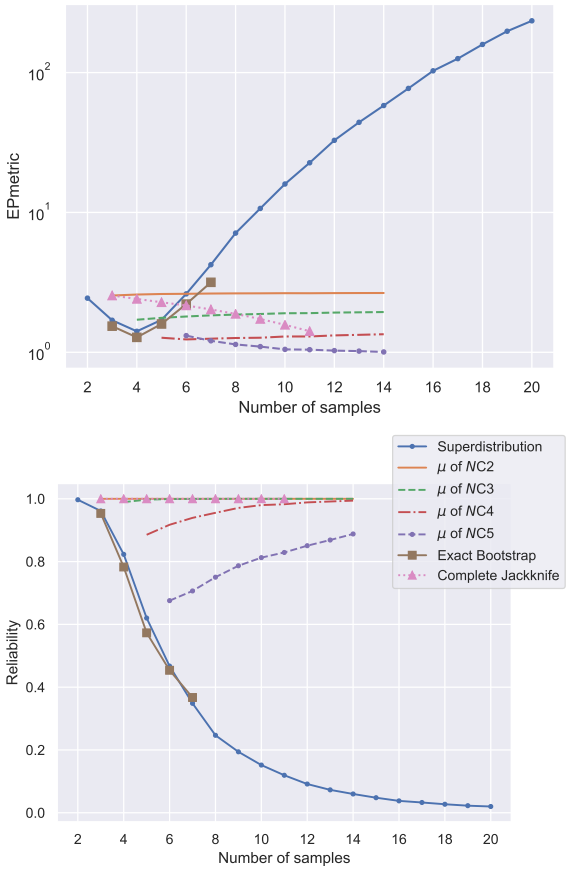

Figure 8. The EPmetric and reliability for the Superdistribution with Bootstrapping and Jackknifing on the Exponential Wide distribution.

On the other hand, performance of the Complete SDJ (CSDJ) *does* keep improving as samples are added. However, the CSDJ was not as effective of using SDJ with optimal subsample size r. CSDJ required about 12 samples to achieve the same EPmetric level of accuracy + reliability performance as $NC4$ SDJ with the optimal 6 samples.

Results for the TI-EN95 method with and without resampling are shown in Figure 9. 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. However, EPmetric results do not first improve with added samples like they do for SD with this distribution; TI-EN95 EPmetric performance always deteriorates with added samples.

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 the present 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.

The NCr and Exact Bootstrap EPmetric results plateau or appear to asymptote toward different plateaus as the number of samples increases, while the Complete Jackknife results start best and continue to improve with added 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 results proportionately more than it improves SDs, in absolute terms SD-Jackknifes accuracy+reliability EPmetric results are best (compare TI-EN95 Complete Jackknife against SDJ $NC5$ at any N).

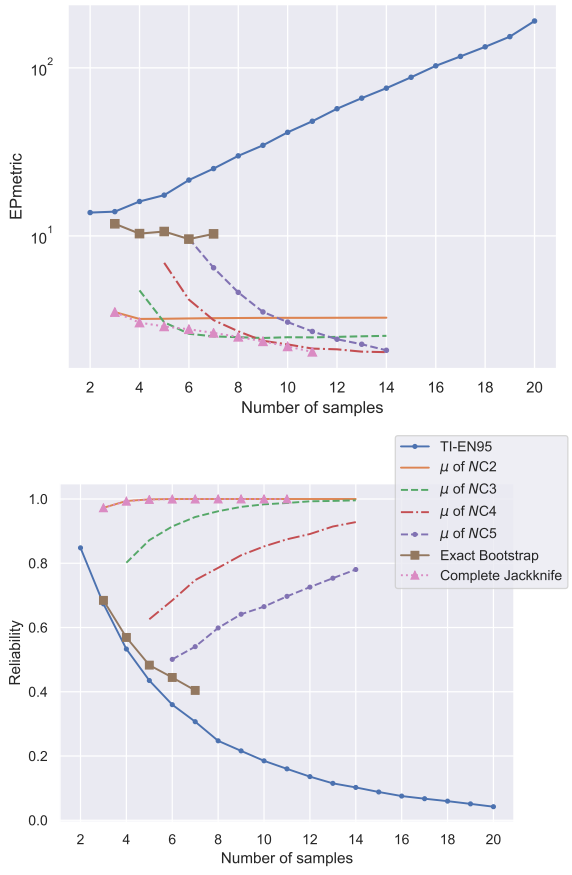

Figure 9. The EPmetric and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the Exponential Wide distribution.

Weibull Narrow distribution

The results of using SD and TI-EN95 with and without resampling on the Weibull Narrow distribution are shown in Figures 10 and 11. Of 16 distributions studied in [3], the Weibull Narrow distribution was the most difficult to predict tail probabilities for, given very limited data and the sparse-sample UQ methods tried. This is most evident in the low reliability levels for SD alone and TI-EN95 alone in Figures 10 and 11. The best SD reliability is 0.3 with $N = 2$ samples, dropping precipitously to about 0.05 with $N = 3$, 0.02 at $N = 4$, and 0.01 for $N = 5$ to 20. TI-EN95 reliability is roughly an order of magnitude worse at any sample size N. This is 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 ($N = 3$) had a reliability of about 42% whereas the SD-only reliability was about 5%. This involves the optimal SDJ *NCr* variant, where $r = N_{\text{SDopt}} = 2$ for the EPmetric and this distribution. This variant has 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 EPmetric performance quickly degrades. Conversely, performance quickly improves when SDJ resampling is used with optimal *NC*2. The *NC*2 EPmetric results in Figure 10 exhibit vastly better combined accuracy + reliability performance than SD (for the same number of total samples) over the full range $N = 3$ to 14 investigated. Larger numbers of samples will result in even greater advantage of *NC*2 SDJ if its trend continues of improving reliability and combined accuracy + reliability with increasing total number of samples.

Reliability and reliability + accuracy performance of *NC*5 to *NC*3 families of SDJ are consecutively better than SD over the full range $N = 3$ to 14 investigated, but are vastly worse than the optimal *NC*2 family over this range (according to EPmetric). 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).

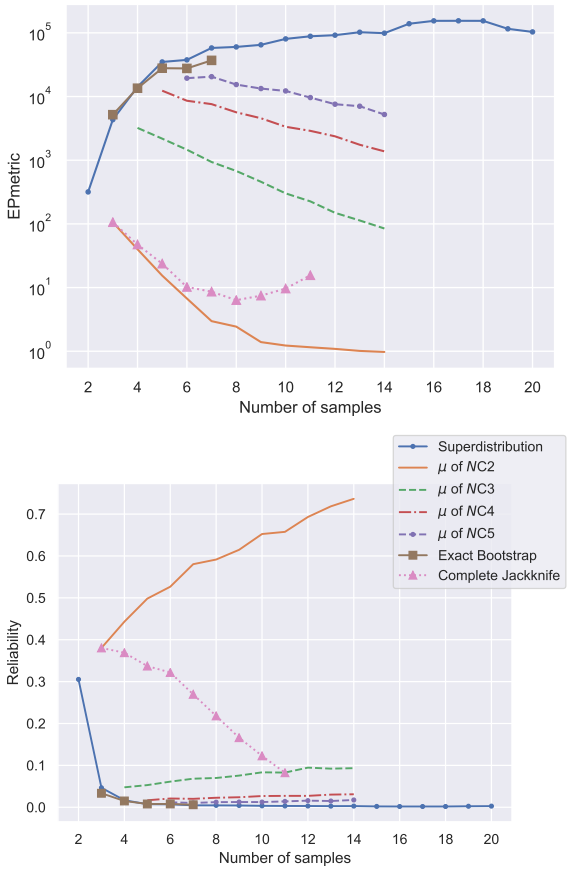

Figure 10. The EPmetric and reliability for the Superdistribution with Bootstrapping and Jackknifing on the Weibull Narrow distribution.

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 *NCr* methods all continually improve with increasing samples.

Results for TI-EN95 with and without resampling are shown in Figure 11. The TI-EN95-only trends observed with SD-only apply here as well. 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 results immediately decline with added samples like they do for SD. (This is also seen for TI-EN95 with the Exponential Wide distribution.) 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 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. The *NC*3, *NC*4, and *NC*5 TI-EN95-J results are progressively worse TI-EN95-CJ results. The rankings in the above three sentences are also applicable to SD by replacing 'TI-EN95' by 'SD'. In absolute terms, the SD results are much better than their corresponding TI-EN95 results, for any given number of samples. This was the case for the prior two distributions as well.

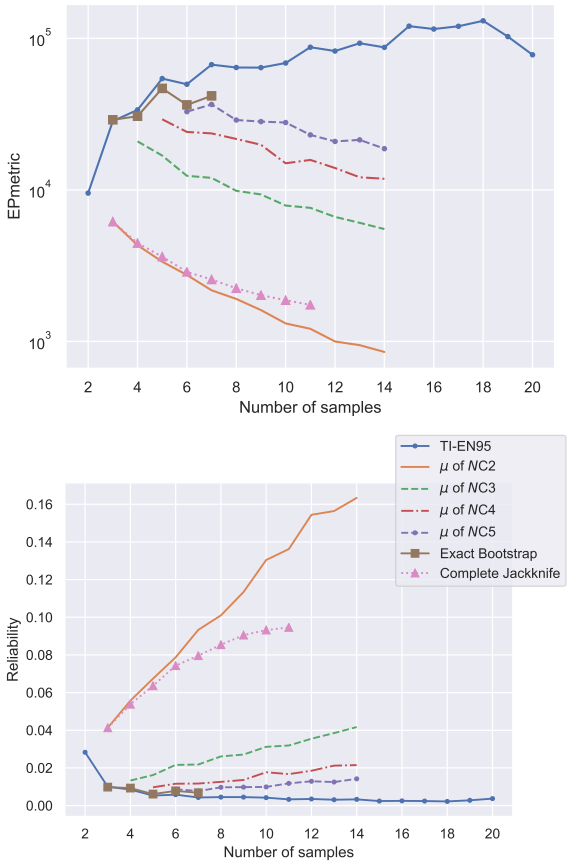

Figure 11. The EPmetric and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the Weibull Narrow distribution.

Standard Normal Distribution

Using generalized Jackknifing with SD was shown 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.

The results for SD with and without resampling on the Standard Normal distribution are presented in Figure 12. 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 1.0. This higher reliability comes with an unwelcome tradeoff of lower accuracy (worse accuracy + reliability EPmetric value while reliability remains 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 performance for *NCr* SDJ may be because the lowest EPmetric occurred at the highest sample sizes studied, *N*=19 and20; 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 larger their subsample size *r* is toward 20.

Complete SD-Jackknife *does* improve in combined accuracy + reliability performance per the EPmetric as samples are

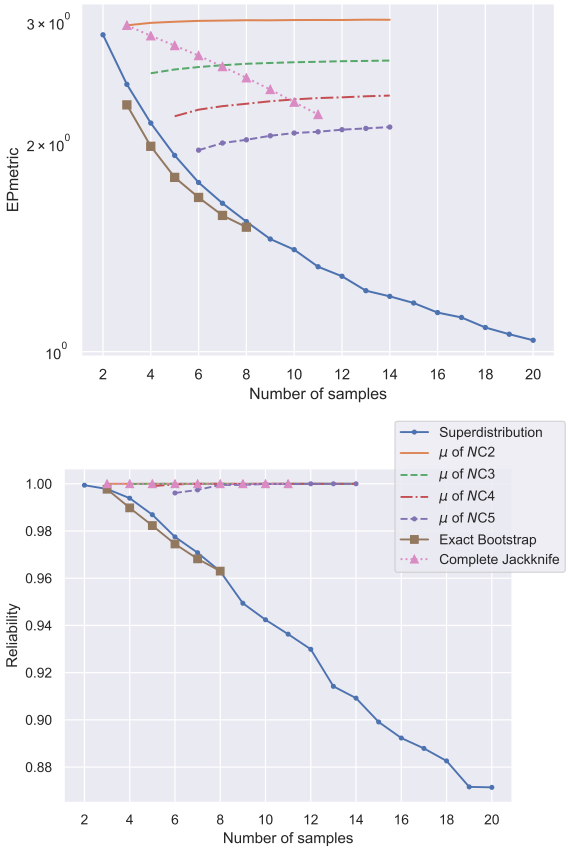

Figure 12. The EPmetric and reliability for the Superdistribution with Bootstrapping and Jackknifing on the standard Normal distribution.

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 would improve the Complete SD-Jackknife results. CSDJ also has reliabilities 1.0 for the 3 to 11 total samples tried with this method. Its improving trend doesnt appear to reach the best *NCr* method on this distribution (*NC*5) until N=12 samples. There its EPmetric value is the about same as *NC*5s. However, it appears that 14 samples are required before the CSDJ trend reaches a better/lower EPmetric value than *NC*5s 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* SD-Jackknife method can be pointed-to that is more cost effective.

The Standard Normal results for TI-EN95 with and without resampling are shown in Figure 13. 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 1 to 0.87 as the samples increased from 2 to 20. Like for SD, TI-EN95s 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.

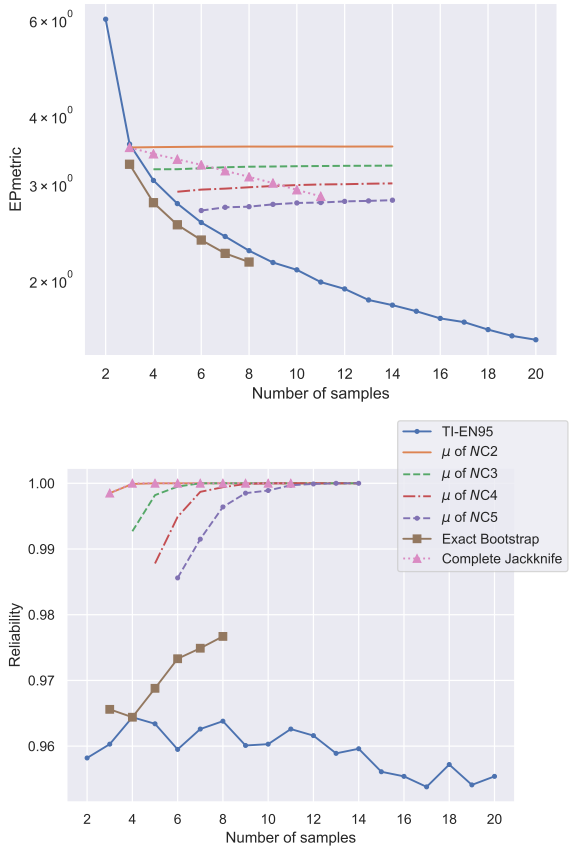

Figure 13. The EPmetric and reliability for the TI-EN95 with Bootstrapping and Jackknifing on the standard Normal distribution.

DISCUSSION

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 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 sometimes reduced their reliability but sometimes also gave better combined EPmetric reliability + accuracy than Jackknife resampling did.

For the distributions and a tail-probability magnitude of 10^{-4} studied here, 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 paper and for EP magnitudes 10^{-1} to 10^{-5}, it is found in [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 [3] and in this paper. On the four distributions and 10^{-4} EP magnitude studied here, 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.

For most of the 16 distributions studied in [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" 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 1 for the better-performing SD variant of methods. This "rule" applies more strictly for SDJ than for the TI-EN95J variant. The latters EPmetric value is, in the case of Figure 7, slightly better when $r = (N_{\text{TIEN95opt}} - 1)$ in *NCr* Jackknifing with the same total number of samples N needed as a minimum for the rule-of-thumb use (see the $N = 5$ cases in Figure 7). Nonetheless, the EPmetric differences are so minor that the rule is deemed to effectively apply. 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 [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$^{\text{th}}$ and 5$^{\text{th}}$ columns of Table 1 present results on this point.

Note also that using a 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 1 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 with 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 resampling; SD and TI-EN95 reliabilities are, for a given number of total samples, always improved by any of the Jackknifing methods, including

Table 1. Reliability results for the four distributions and SD NCr Jackknifing with various subsample sizes r.

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[1]	0.99	1.00 (5C4)	1.00 (6C4)	1.00 (6C5)

non-optimal NCr Jackknifing.[2] 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 [3] for 16 diverse distributions including the four in this paper and for EP magnitudes 10^{-1} to 10^{-5} (for SD and several other sparse-sample UQ methods). 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 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 Jackknife 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 (see column 4 of Table 1 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 best 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.

The optimum subsample sizes mapped-out in [3] (as mentioned above) show that the vast majority are in the range 2 to 6 for SD. (In the following 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 CSDJ to, say, 7. This would limit EP under-estimation risk to what is anticipated to be a reasonably low level. (We cannot be more quantitative without applying CSDJ to our full test suite of 16 diverse distributions and 10^{-1} to 10^{-5} 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 paper and two others in [3], for an EP magnitude 10^{-4}. This occurs over the range 3 to 11 samples tried. For 2 samples, SD-alone would be used. This gives reliabilities of 1 in the five non-pathological cases.

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 over the range of samples tried with it (2 to 11) than NCr_{opt} Jackknifing in all cases studied in this paper but the one in Fig. 9. This conservatism could manifest, for example, in an EP estimate of 10^{-3} while it is really several orders smaller, like 10^{-6}. The indicated quantification over the broader data base of

[2] This statement does not conflict with the discussion around Table 1. 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 with a different/lower number of samples $\leq r_{\text{opt}}$.

16 diverse distributions and 10^{-1} to 10^{-5} EP magnitudes would help quantify 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.

Finally, both Bootstrapping and Jackknifing can be used to construct a distribution of EP estimates that are averaged. This is perhaps one of the most important features of resampling techniques 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 to be unlike any distribution for which the resampling methods performance has been characterized a-priori.

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 EPmetric, 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 [3].

REFERENCES

[1] Romero, V., Bonney, M., Schroeder, B., and Weirs, V. G., 2017. 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.

[2] Romero, V. J., and Weirs, V. G., 2018. "A class of simple and effective uq methods for sparse replicate data applied to the cantilever beam end-to-end uq problem". In 2018 AIAA Non-Deterministic Approaches Conference, p. 1665.

[3] Jekel, C., and Romero, V. Conservative estimation of tail probabilities from limited sample dataexpanded investigation of a class of simple uncertainty methods for sparse data. Technical report in preparation, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States).

[4] Bhachu, K. S., Haftka, R. T., and Kim, N. H., 2016. "Comparison of methods for calculating b-basis crack growth life using limited tests". *AIAA Journal,* ***54***(4), Mar., pp. 1287–1298.

[5] Wu, C. F. J., 1986. "Jackknife, bootstrap and other resampling methods in regression analysis". *Ann. Statist.,* ***14***(4), 12, pp. 1261–1295.

[6] Picheny, V., Kim, N. H., and Haftka, R. T., 2010. "Application of bootstrap method in conservative estimation of reliability with limited samples". *Structural and Multidisciplinary Optimization,* ***41***(2), Mar, pp. 205–217.

[7] Kisielinska, J., 2013. "The exact bootstrap method shown on the example of the mean and variance estimation". *Computational Statistics,* ***28***(3), Jun, pp. 1061–1077.

[8] Quenouille, M. H., 1956. "Notes on bias in estimation". *Biometrika,* ***43***(3-4), pp. 353–360.

[9] Tukey, J., 1958. "Bias and confidence in not quite large samples". *Ann. Math. Statist.,* ***29***, p. 614.

[10] Schucany, W. R., Gray, H. L., and Owen, D. B., 1971. "On bias reduction in estimation". *Journal of the American Statistical Association,* ***66***(335), pp. 524–533.

[11] Jones, M. C., and Foster, P. J., 1993. "Generalized jackknifing and higher order kernels". *Journal of Nonparametric Statistics,* ***3***(1), pp. 81–94.

[12] Hahn, G. J., and Meeker, W. Q., 2011. *Statistical Intervals: A Guide for Practitioners*. Wiley-Blackwell, pp. 288–352.
[13] Montgomery, D. C., and Runger, G. C., 2010. *Applied statistics and probability for engineers*. John Wiley & Sons.
[14] Howe, W. G., 1969. "Two-sided tolerance limits for normal populationssome improvements". *Journal of the American Statistical Association,* ***64***(326), pp. 610–620.
[15] Young, D., 2010. "tolerance: An r package for estimating tolerance intervals". *Journal of Statistical Software, Articles,* ***36***(5), pp. 1–39.

Appendix A: Tolerance-Interval Equivalent Normal

Tolerance Intervals (TIs) are a simple way to 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.

As Figure 14 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 at the sample mean. The bounds are determined by multiplying the sample standard deviation by a factor f,

$$L = \tilde{\mu} - f\tilde{\sigma} \tag{8}$$
$$U = \tilde{\mu} + f\tilde{\sigma} \tag{9}$$

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 [12] and [13]. Equations to calculate f provided a coverage, confidence, and the number of samples can be found in [14] and [15].

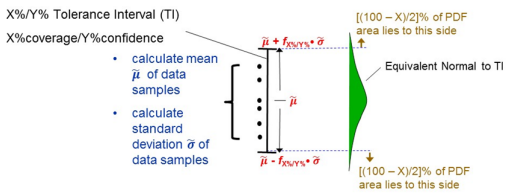

Figure 14. Construction of a tolerance interval and its "Equivalent Normal" distribution (from [1]).

A TI "Equivalent Normal" (TI-EN) distribution (see Figure 14) can be determined by finding an equivalent normal standard deviation σ_{EN} as explained in [1]. If 95% coverage is used, then

$$\sigma_{\text{EN}} = \frac{f\tilde{\sigma}}{1.96} \tag{10}$$

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

Alternatively, equations in [14] and [15] 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

$$\sigma_{\text{EN}} = k\tilde{\sigma}. \tag{11}$$

Then k can be determined from

$$k = \sqrt{1+N^{-1}}\sqrt{\frac{N-1}{\chi^2_{N-1;\alpha}}}\sqrt{1+\frac{N-3-\chi^2_{N-1;\alpha}}{2(N+1)^2}} \tag{12}$$

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 desired 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. Because TI-ENs are constructed particular to a specified confidence level like 95%, we sometimes associate them with the confidence level, e.g., "TI-EN95".

Appendix B: Superdistribution (SD)

The Superdistribution (SD) method first starts by constructing L number of distributions as an Ensemble of Normals (EON) as described in [1]. Each Normal distribution in the EON is defined as $\mathcal{N}(\mu_i, \sigma_i^2)$ for a candidate mean μ_i and candidate standard deviation σ_i. A candidate mean μ_i can be determined from

$$\mu_i = \tilde{\mu} + \frac{T_i\tilde{\sigma}}{\sqrt{N}} \tag{13}$$

where T_i is a random sample from a Student's t-distribution with $N-1$ degrees of freedom. A candidate standard deviation σ_i can be determined from

$$\sigma_i = \tilde{\sigma}\sqrt{\frac{N-1}{\chi^2_i}} \tag{14}$$

where χ_i^2 is a random sample from a $N-1$ degree of freedom Chi-Square distribution. The SD will converge for a large number of L. A convergence study in [3] showed that $L = 10,000$ was sufficient for a converged SD.

Note that the PDF and therefore CDF are known for each Normal distribution in the EON. Hence, the CDF of the SD can be defined as

$$\mathrm{SD}_{\mathrm{CDF}}(x) = \frac{\sum_{i=1}^{L} \mathrm{F}_i(x)}{L} \tag{15}$$

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 15, 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. The SD right-tail probability for a specified threshold x in the SD's right tail is given by $1 - \mathrm{SD}_{\mathrm{CDF}}(x)$.

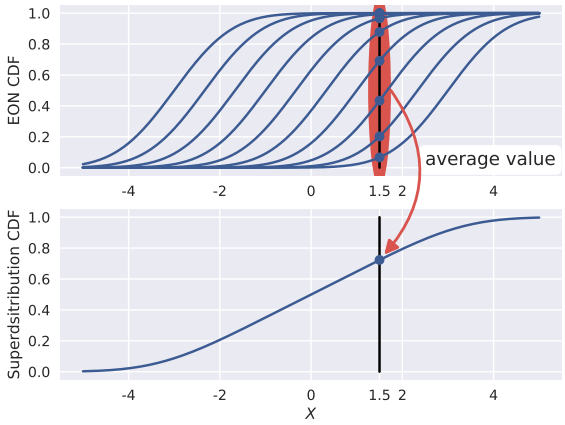

Figure 15. Example of 10 Normal distributions in an EON. The Superdistribution CDF value at $x = 1.5$ is calculated by averaging the CDF values from each Normal distribution in the EON at $x = 1.5$.