Bootstrapping and Jackknife Resampling to Improve Sparse-Sample UQ Methods for Tail Probability Estimation

Charles Jekel, U. Fla. PhD Candidate and Sandia Intern
Vicente Romero, Sandia National Laboratories
V&V, UQ, and Credibility Processes Dept.

ASME Verification & Validation Symposium
May 15-17, 2019, Las Vegas, NV

\(^1\)Sandia National Laboratories document SAND2019-5422 C
Tail probability estimation with sparse sample data
—Introduction

General Sparse-Sample Tail-Probability Estimation Objective:
◦ Conservatively estimate risk or reliability related frequency-distribution
tail probabilities from very sparse sample data (e.g. experimental data)
◦ Avoid being overly conservative
◦ These competing objectives make this a very challenging problem

Objective of the Present Study and V&V Symposium Paper:
◦ What is the best tail-probability estimation approach
given very sparse sample data?

• Factor Space of a Current Investigation:
  • 2 to 20 data samples
  • Tail probabilities or “exceedance probabilities” (EPs) of $10^{-1}$, $10^{-2}$, $10^{-3}$, $10^{-4}$, $10^{-5}$
  • > 25 variants of Sparse-Sample UQ Methods or combinations of methods
  • 16 diverse and challenging distributions shapes
  • 10,000 random sampling trials per combination of factor levels tried

• A very extensive study (150 pp. Sandia draft report in review)
  • ~1/4 full-factorial investigation, >100-million tests of the methods’ performance

• The paper and presentation give highlights of the methods and interim findings
Sparse-Sample UQ Methods — Building Block 1

A class of relatively simple and effective sparse-sample UQ methods tailored for Normal distributions

Construction of:
- Ensemble of Normals (EON)
- Super Distribution (SD)

Sparse sample data
Study of Sparse-Sample UQ Methods Without Resampling

- 5 methods: SD, TI-EN with 90%, 95%, 99% confidence settings, EON-90%
- tail probability magnitudes $10^{-1}$, $10^{-2}$, $10^{-3}$, $10^{-4}$, $10^{-5}$
- # samples $N = 2, 3, 4, \ldots, 20$
- 16 diverse distribution shapes below
- 10K trials for each combination of the above factors

8 analytical PDFs

- Narrow Exponential
- Wide Exponential
- Log-Normal
- Normal
- Wide Weibull
- Narrow Weibull
- Bi-Modal

8 empirical PDFs
Performance Metrics for Estimation of Tail Probability

Quantify performance of the UQ methods in two ways:

- **Reliability** of conservatively bounding the true probability
  - proportion of random trials where tail probability estimate > true EP

- **Combined Performance Metric**: Reliability + Accuracy of estimate relative to true EP
  
  Error metric = \( \Delta \log = \log(EP_{estimated}) - \log(EP_{true}) \)
  
  (# of orders of magnitude that the predicted probability is off by)

  - **unpenalized**
    
    EP peformance metric = \[ \frac{\sum N^+ \Delta \log + \sum N^- |\Delta \log|}{N^+} \]

  - **10Xpenalty** for negative errors (under-estimation of exceedance probability)
    
    EP-10X peformance metric = \[ \frac{\sum N^+ \Delta \log + 10 \sum N^- |\Delta \log|}{N^+} \]

- all error types usually not equally bad; preference weighting via penalty factor in numerator
- larger average magnitude of given error type drives numerator and metric up
- larger proportion of + (conservative) errors in denominator drives metric down
- lower metric value = better performance
Example results for $EP = 10^{-4}$ on Exponential Distribution

- fairly difficult distribution for tail probability estimation

**Reliability (higher = better)**

**Accuracy + Reliability (lower = better)**

- Reliability decreases with added samples, for all methods (for non-Normal PDFs)
- Tradeoffs exist between reliability and accuracy
  — higher reliability correlates strongly with more conservative/less-accurate

- Each method has an optimum # of samples $N_{opt}$ for best combined accuracy + reliability (right figure); further samples yield worse performance

- Superdistribution (SD) method had best performance with $N_{opt} = 4$, other methods with any number of samples up to 20 performed worse than SD with $N_{opt} = 4$. 
Highlights of Sparse-Sample UQ Methods’ Results

• Prior slide’s performance trends vs. #samples for the methods are representative for many of the 16 distributions and 4 of the 5 EP magnitudes

• EP magnitude, distribution shape, and # of samples strongly affect absolute performance all methods – results are highly variable over the factor space

• Most distinctive and consequential trend difference between methods occurs vs. EP magnitude
  
  • SD reliability gets better as EP magnitude decreases; declines as EP magnitude increases
  
  • Reverse trend occurs for TI-EN and EON methods
  
  • Cross-over point exists at $10^{-1}$ EP magnitude where TI-EN methods have better reliability and often better average and optimal performance per the reliability + accuracy EPmetrics

• ~100 million numerical tests showed Superdistribution (SD) method performed best overall (per reliability + accuracy EPmetrics)
### SD Performance Results

#### Optimal # of samples $N_{opt}$ for SD

<table>
<thead>
<tr>
<th>EP</th>
<th>$N \leq 10^{-1}$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$N \leq 10^{-1}$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$N \leq 10^{-1}$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$N \leq 10^{-1}$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>2</td>
<td>6</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>TI$^1$</td>
<td>2$^A$</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Log-Normal</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>TI$^2$</td>
<td>2$^B$</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5 d.o.f. t</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>TI$^1$</td>
<td>3$^B$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Weibull Wide</td>
<td>-</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>TI$^2$</td>
<td>2$^B$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Exponential Narrow</td>
<td>-</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>TI$^2$</td>
<td>2$^C$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Exponential Wide</td>
<td>-</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>TI$^2$</td>
<td>2$^C$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>TI$^1$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Empirical 4</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>TI$^{2a}$</td>
<td>2$^C$</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>T1$^1$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>T1$^1$</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Empirical 6</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>T1$^2$</td>
<td>2$^B$</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>T1$^2$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Empirical 3</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>T1$^2$</td>
<td>2$^B$</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>T1$^2$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Empirical 2</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>T1$^1$</td>
<td>2$^A$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>20</td>
<td>20</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>T1$^1$</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Empirical 5</td>
<td>-</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>T1$^2$</td>
<td>2$^A$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>T1$^2$</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Empirical 8</td>
<td>-</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>T1$^2$</td>
<td>2$^B$</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>T1$^2$</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Empirical 1</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>19</td>
<td>T1$^1$</td>
<td>2$^A$</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>20</td>
<td>20</td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>T1$^1$</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Empirical 7</td>
<td>-</td>
<td>3</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>T1$^2$</td>
<td>2$^B$</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>T1$^1$</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Bi-modal</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>20$^3$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20$^3$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20$^3$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Weibull Narrow</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>TI-EN99.99 better$^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TI-EN99.99 better$^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TI-EN99.99 better$^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Some SD reliabilities with “sweet spot” # of samples if know approx. EP magnitude

- $N = 5, P = 10^{-5}$: 14 of 16 dists... reliability > 80%
- $N = 4, P = 10^{-5}$: 15 of 16 dists... reliability > 80%
- $N = 4, P = 10^{-4}$: 14 of 16 dists... reliability > 80%
- $N = 3, P = 10^{-3}$: 14 of 16 dists... reliability > 80%
- $N = 2, P = 10^{-3}$: 15 of 16 dists... reliability > 95%
- $N = 2, P = 10^{-2}$: 15 of 16 dists... reliability > 85%
- $N = 2, P = 10^{-1}$: 5 of 16 dists... reliability > 80% -- SD breaks down for $10^{-1}$
Building Block 2 — Statistical Resampling Methods

• Statistical Resampling methods have traditionally been used to reduce bias and variability of statistics estimated from sample data.

• They can be used to improve the reliability of the Sparse-Sample UQ methods.

**Bootstrapping:**
• Generate artificial sample sets of size \( N \) by drawing from the original set with replacement.
• Apply Sparse-Sample UQ methods to each generated sample set and average estimates.

**Jackknifing:**
• Create sub-samples of the original sample (many subsample combinations can exist).
• Apply Sparse-Sample UQ methods to each generated sample set and average estimates.
Illustration of SD with Statistical Jackknifing

- N=4 samples ➔ four unique $r=3$-sample subsets for SD estimation
  - “4 choose 3” (4C3) NCr-Jackknife
  - Average the four 3-sample SD estimates of the tail probability
  - Example shows much-improved reliability of a conservative estimate vs. Optimal SD
    - optimal SD reliability @ N=4 is 0.8, whereas 4C3 SD-Jackknifing reliability = ~1
    - 25% improvement to ~perfect reliability for same # of samples with SD-Jackknifing
- One more sample enables 5C4 NCr-Jackknifing where $r = 4 = N_{SD_{opt}}$
  - EPmetric improvement indicates accuracy improvement because reliability can only improve trivially from ~1

- Example
  - Exponential distribution, EP = $10^{-4}$
  - Jackknife subsampling and averaging improves accuracy and conservatism performance of all methods tried for this distribution and EP magnitude, but best results are with SD

SD-Jackknife reliability+accuracy performance does not degrade with added samples, whereas SD-alone does after optimal 4 samples

Reliability improvement for various SD-Jackknife subsamplings and same total # samples (N=8 case)

Reliability of obtaining a conservative estimate increases with total # samples vs. decreasing reliability with SD alone
Highlights for Sparse-Data UQ Methods with Resampling

- Bootstrapping and Jackknifing were applied with sparse-sample UQ methods SD, EON90, TI-EN 90 & 95
  - EP= $10^{-4}$, 6 distributions spanning easiest to most difficult for tail probability estim.
- Bootstrapping did not help significantly, but Jackknifing usually did.
- We focus on SD-Jackknifing because it performed better than Jackknifing with the other sparse-sample methods
- Jackknifing always increased reliability vs. SD-alone, for a given number of total samples – good
- But combined accuracy + reliability performance of SD-Jackknifing was sometimes better and sometimes worse than using SD alone
  - SD-J only ensured to have better accuracy + reliability performance than SD-alone if the NCr Jackknifing sub-sample size is $r = N_{SDopt}$
  - requires knowing $N_{SDopt}$ which requires knowing PDF shape and EP magnitude
- Seek a resampling strategy for improving SD that is not dependent on detailed prior knowledge
Favored Strategy: SD with “Complete” Jackknifing

- “Complete” Jackknifing (CJ) uses all possible NCr cases for a given # samples N -- not dependent on sub-sample size r because all possible r’s used
  - e.g., N=4 samples enables 4C2 or 4C3 or averaging 4C2 and 4C3 results (= CJ)

- SD-CJ gave reliabilities of ~1 for any # samples tried (3 to 11), for 5 of the 6 PDFs and EP = 10^{-4} magnitude studied

- SD-CJ reliability + accuracy performance transitions from the worst-performing NCr curve to consecutively better ones as the total number of samples increases (figure).

- SD-CJ is cautiously indicated (for many of the EP’s and distributions not tested) to have
  - the highest reliability of any method tried so far for a given number of samples
  - best robustness to unknown PDF shape and tail-probability magnitude
  - continually improving accuracy as samples are added

![Graph showing EP metric vs. number of samples](image)
More Results (post V&V Symposium paper)

- Spot-check with $10^{-1}$, $10^{-2}$ EP magnitudes for Exponential distribution
  - For $10^{-1}$ use TIEN-95 with Complete Jackknifing
  - For $10^{-2}$ use TIEN-95 or SD with Complete Jackknifing
  - Get reliabilities of 0.9 to 1 for $N = 3$ to 7 samples
  - The more samples in this range, the better the accuracy while maintaining reliability $\geq 0.9$
Still a work in progress

Tentative Strategy and Reliability projections from all investigations to date:

- If only N=2 samples affordable, use SD and get reasonable reliability for $EP \leq 10^{-2}$
  - $> 85\%$ for 15 of 16 PDF when $EP \leq 10^{-2}$
  - $(> 80\%$ for 5 of 16 PDF when $EP = 10^{-1}$; TIEN99.99 $> 90\%$ for 15 of 16 PDF when $EP = 10^{-1}$)

- For N=3 to 7 samples, get $\geq 90\%$ reliability for 15 of the 16 distributions & EP magnitudes as follows:
  - If strongly suspect $EP \geq 10^{-2}$ use TIEN-95 with Complete Jackknifing
  - If strongly suspect $EP \leq 10^{-2}$ use SD with Complete Jackknifing
  - Very high reliabilities of $\sim 1$ occur at the smaller # of samples 3, 4, etc. which unavoidably implies an accuracy tradeoff of conservative over-estimation, e.g., estimate $10^{-3}$ for a true EP of $10^{-6}$
  - The more samples in this range ($\leq 7$), the better the accuracy while maintaining high reliability $\geq 90\%$

Need to verify/refine these projections with $\sim 100$ Million more tests over the full matrix using Complete Jackknifing with SD and TI-EN methods

Quantify accuracy-reliability performance for selecting method and # samples