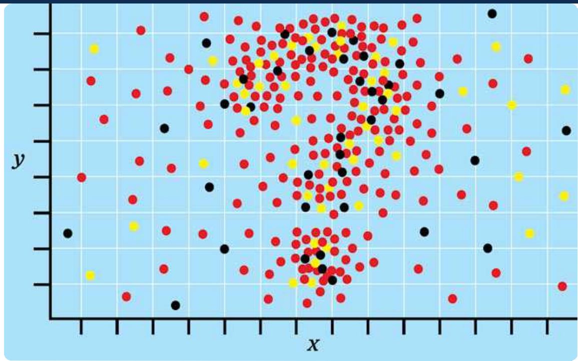
SAND2018-2188C



Model Credibility in Statistical Reliability Analysis with Limited Data

Caleb King and Lauren Hund

Statistical Sciences Group

Sandia National Laboratories

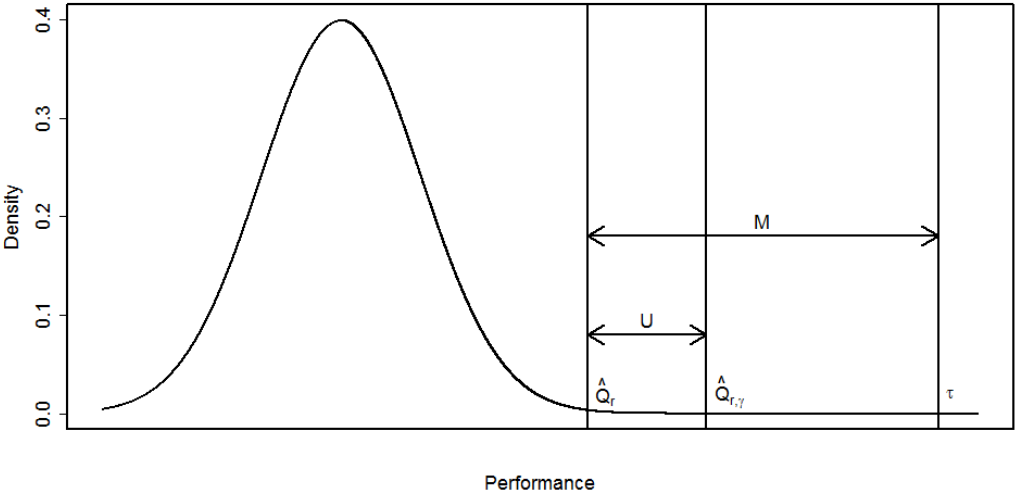
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

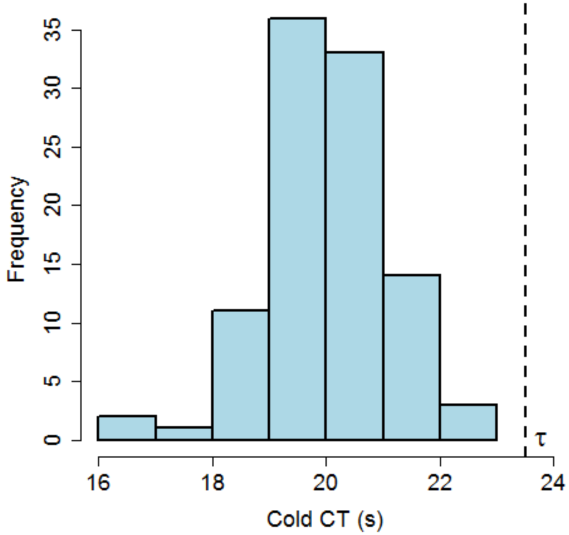
Outline

- Motivating Examples
  - Performance Reliability
  - Environmental Reliability
- Complexity of Reliability Demonstration
- Model Credibility
  - What is it and how is it done?
  - Data-driven approach
  - “Evidence-driven” approach
- Summary and Conclusions

Performance Reliability Demonstration

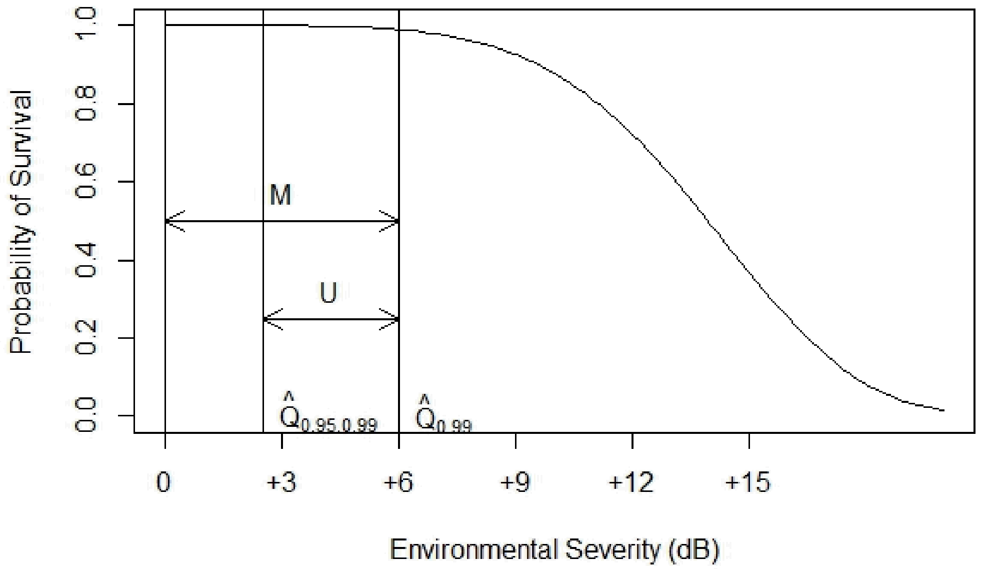
- **Goal:** To show that the product meets its requirement(s) with the specified reliability.
- **Ideal Statement:** "We can demonstrate with XX% confidence that YY% of our product will pass its requirement(s)."
  - Based on the definition of a statistical tolerance bound.
  - May be assessed using metrics, such as a "tolerance ratio" (TR=M/U)
- Example: A hypothetical launch safety device on a missile has a requirement to close within 23.5s of launch with 99.5% reliability.
  - N=100 units tested. None of the units yielded closing times greater than the requirement.
  - Can we make the statement that we are 95% confident that 99.5% of units will pass the 23.5s requirement?

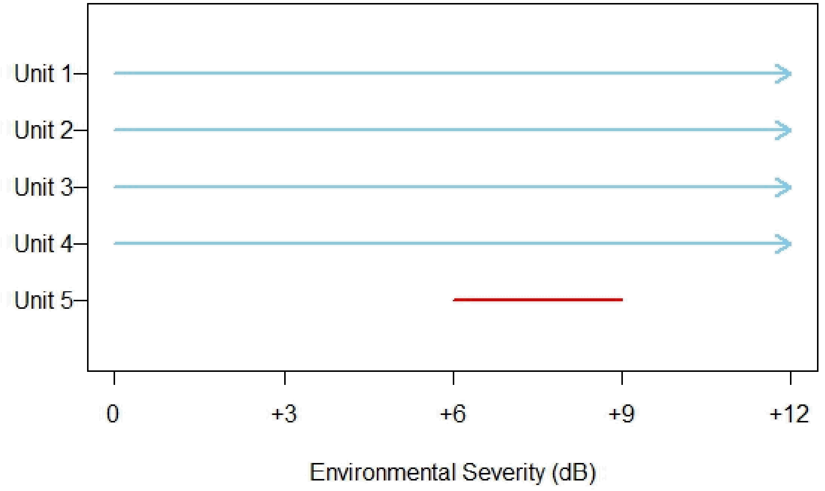




Environmental "Reliability" Demonstration

- **Goal:** To show that the product can continue to meet its requirements with the specified reliability at more severe environments than originally planned.
- **Ideal Statement:** "We can demonstrate that, up to ZZ dB of environmental severity, at least YY% of our product will continue to pass the requirement with XX% confidence.
  - Can be viewed as "performance reliability as a function of environment."
- Example: A hypothetical component in a bomb will undergo shock testing to assess it's sensitivity to more extreme shock environments than it is expected to see in it's usual lifetime.
  - The component can be operated multiple times without draining its useful life.
  - The component must pass a functional requirement with 99% reliability.
  - Only N=5 units are available for testing.
  - Tests were performed at shock levels +3, +6, +9, and +12 dB higher than nominal. Only one failure was observed and that at +9 dB.
  - Can we determine the maximum severity level at which 99% of units will pass their requirements with 95% confidence?





Complexity of Reliability Demonstration

- **Key Question:** Is there enough information/data to make the ideal statement?
  - Performance -> Have enough samples been collected?
  - Environmental -> Have enough samples been collected? Is there enough resolution in the selected test levels?
- In practice, there are several barriers to the “full demonstration” indicated by the ideal statements:

Test data have measurement uncertainty, are limited to a restrictive set of inputs and conditions, and are relatively few in number.

Statistical models may be inadequate or insufficiently justified for modeling the data.

Not all uncertainties can be straightforwardly quantified.

Quantities of interest may be poorly defined.

Experts state of knowledge is imperfect.

Model Credibility

- **Definition:** The degree to which one believes a selected statistical model adequately describes the behavior seen in the observed data.
  - Analysts know that statistical model uncertainty decreases confidence in the results, but often lack a method to communicate this uncertainty.

The [normal] tolerance interval... is not distributionally robust to even small deviations from normality" (Fernholz and Gillespie 2001).

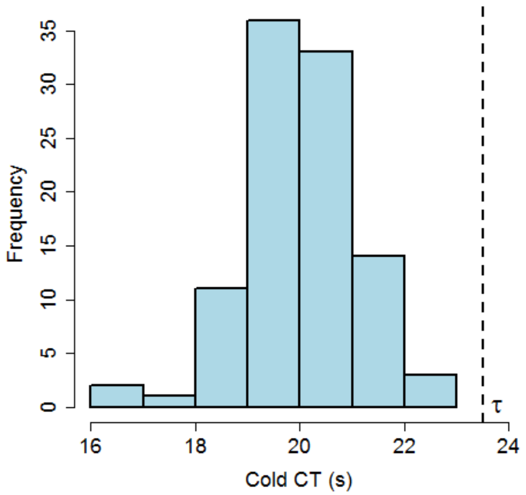
“Estimating tail parameters is analogous to estimating parameters `exterior to the data'... Many times estimates are made by assuming the data is sampled independently from a parametric family. This can lead to disastrous results” (Scholz 2005).

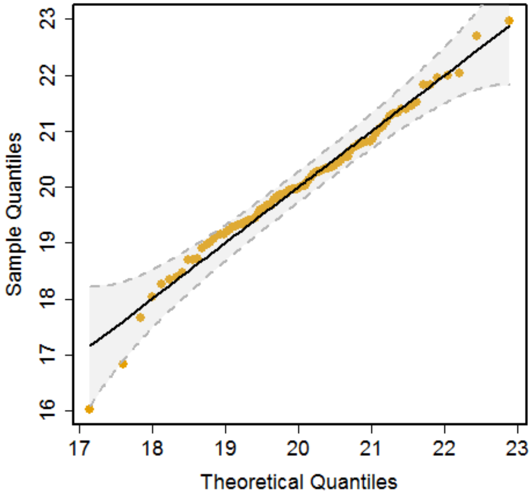
“… extrapolation is often required in reliability engineering/statistical analysis. Extrapolation is always risky…” (Meeker and Escobar 2004)

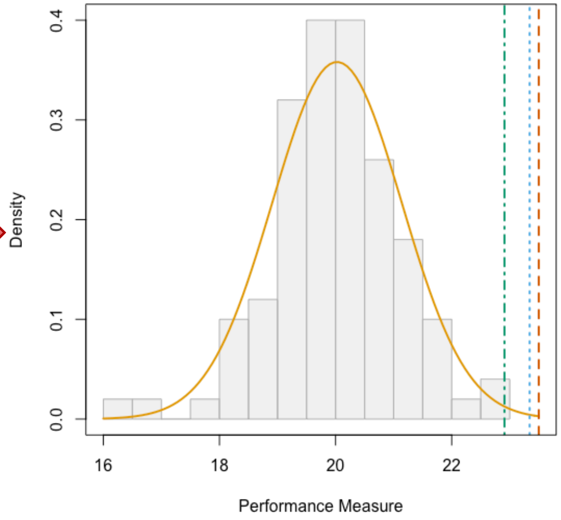
“[Obtaining] a numerical estimate of reliability based on knowledge of full probability distributions in conjunction with QMU would place great demands on our ability to characterize uncertainties. In view of this, it is inevitable that there would be pressure to adopt ‘short cuts’ by simply assuming the forms of PDFs or using PDFs that are not based on some but inadequate supporting data. The response to such pressure would **make or break nuclear certification**. No analysis that is based on speculation or that neglects significant possibilities can lead to genuine confidence, but instead will frequently lead to over-confidence or under-confidence, both of which carry severe costs” (Sharp et. al 2003).

Model Credibility in Practice

- **Ideal:** Relate statistical model selection to **model validation** in engineering applications.
  - **Validity:** Is the model an accurate representation of the real world for the intended uses of the model? (Oberkampf and Barone 2006).
- **Common practice:** Extrapolative prediction from un-validated models.
  - Use statistical tools to select a (normal) model for the data.
    - QQ plots: fit of model over all data, rather than "for the intended uses of the model."
    - Goodness of fit tests: can show evidence of lack of model fit, but cannot validate a model.
  - Extrapolate to the tails using the model.
    - Conclusion: demonstrate 99.5% reliability with 95% confidence, *assuming the model is correct.*

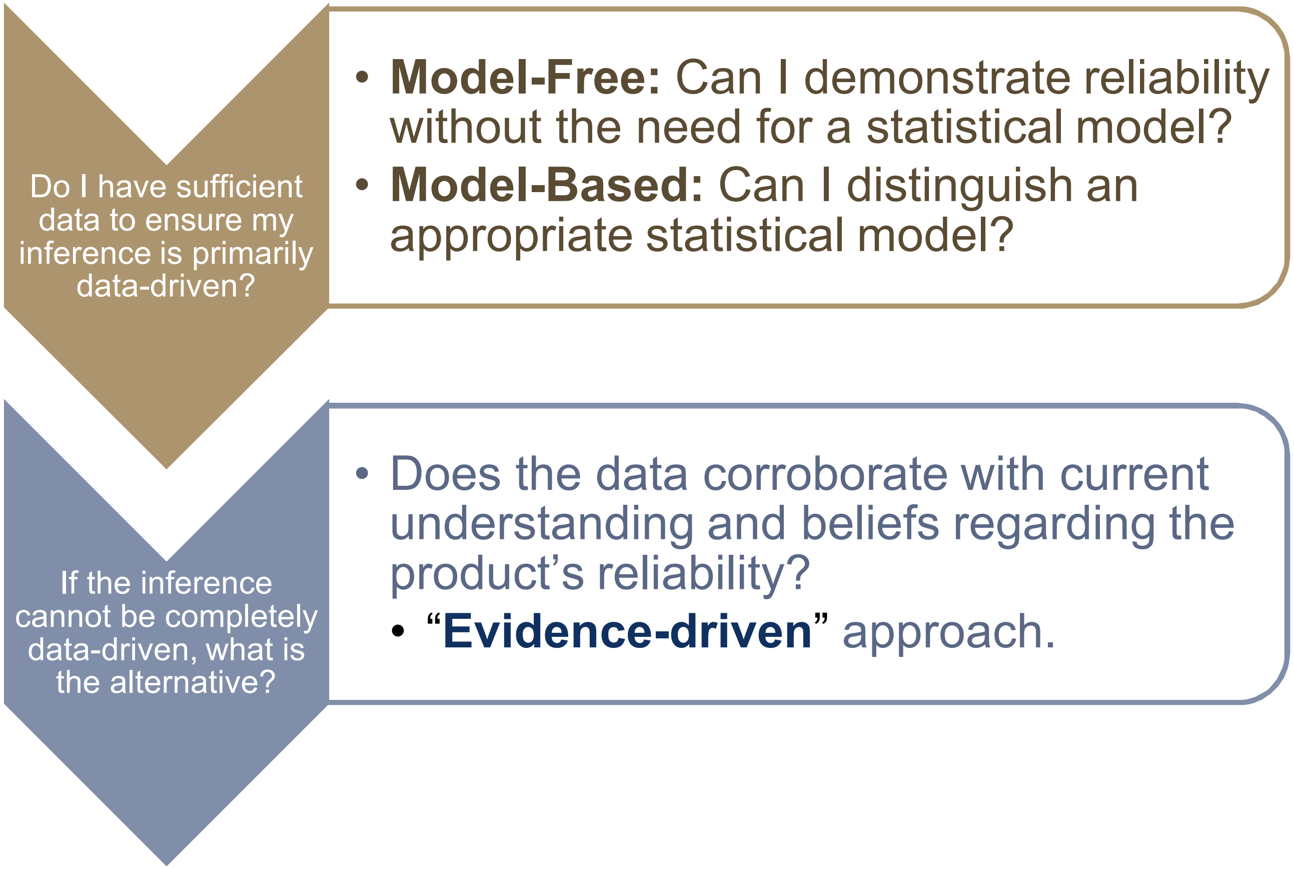






Metrics for Model Credibility

- When assessing the ability to demonstrate reliability, a good approach is to ask the following questions:



Do I have sufficient data to ensure my inference is primarily data-driven?

- **Model-Free:** Can I demonstrate reliability without the need for a statistical model?
- **Model-Based:** Can I distinguish an appropriate statistical model?

If the inference cannot be completely data-driven, what is the alternative?

- Does the data corroborate with current understanding and beliefs regarding the product’s reliability?
  - “**Evidence-driven**” approach.

Data-Driven Metrics

1. **Degree of extrapolation:** Is extrapolation outside the range of the observed data occurring?

2. **Model fit in the tails:** How consistent are the observed tails of the data with the fitted model?

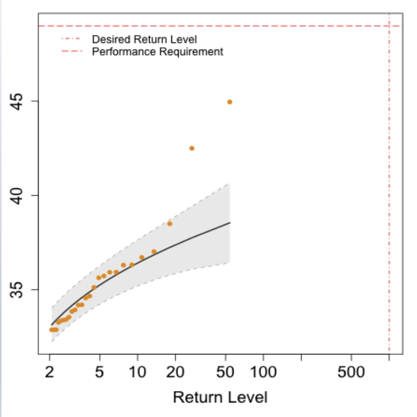
3. **Model adequacy:** Does the model adequately describe the data, even though it may not match perfectly?

4. **Sensitivity to model choice:** How much do the tail estimates change when the modeling assumptions are relaxed?

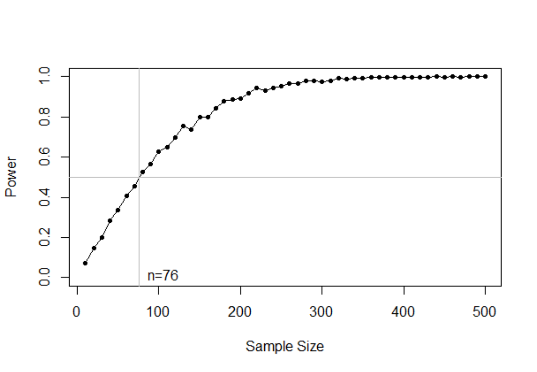
Binomial confidence (Wilks 1941)

$$n^{*} = \log(1-\gamma)\,/\log(r)$$

Return-level plots



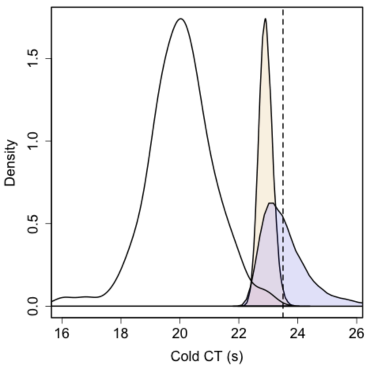
N-index (Lindsay and Liu 2009)



Model-validation

Reliability metric:

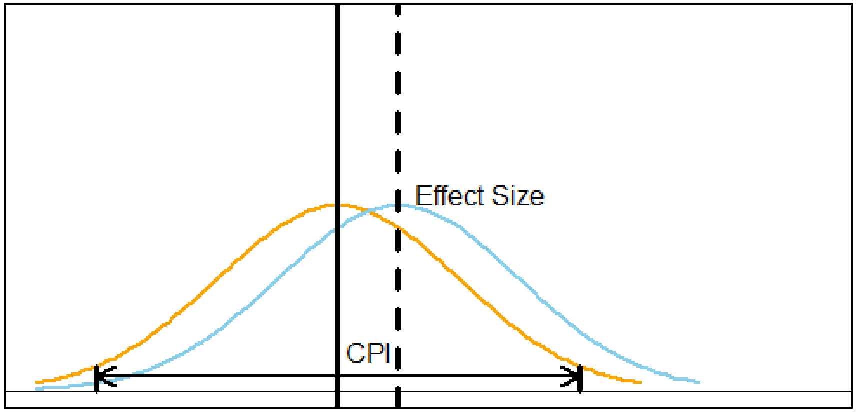
$$R = P(Q_1 - Q_2 < \epsilon)$$



Evidence-Driven Approach

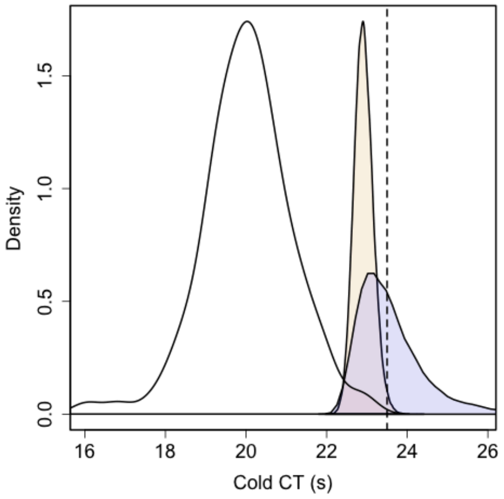
- **Key Idea:** Are the findings from the data consistent with the current understanding of the product's behavior?
  - The goal is then to ***assert*** product reliability when it is not (yet) possible to fully ***demonstrate*** reliability.
  - Data are part of an **evidence package** for asserting reliability rather than being the sole evidence given.
  - This approach can also help with test planning; allowing test planners to allocate their resources more efficiently when they are constrained.

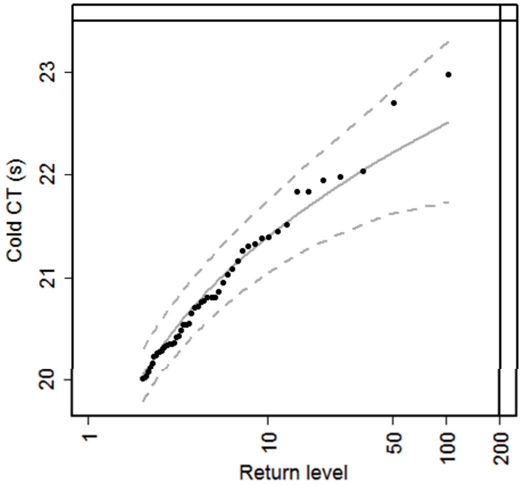
Critical Prior Interval (CPI): The range of prior values capable of rendering the claimed findings no longer credible (Matthews 2018).



Example of Using Data-Driven Metrics

- **Closure time:** estimate 99.5th percentile with 95% confidence.
  - **Degree of extrapolation**: When n = 100, extrapolation is occurring beyond the 97th percentile. We would need a sample 6 times larger to avoid extrapolation.
  - **Validation metrics**: Median of the 99.5th percentile estimate is 23.4 s under model making weaker assumptions versus 22.9 s under normal model.





We conclude:

- The model cannot be evaluated where prediction will occur, and
- The demonstration decision is sensitive to selection of the normal model.

Performance Reliability – Evidence-Driven Approach

- **Practical solutions:** use summary statistics to evaluate performance margin with heuristic measure of uncertainty.

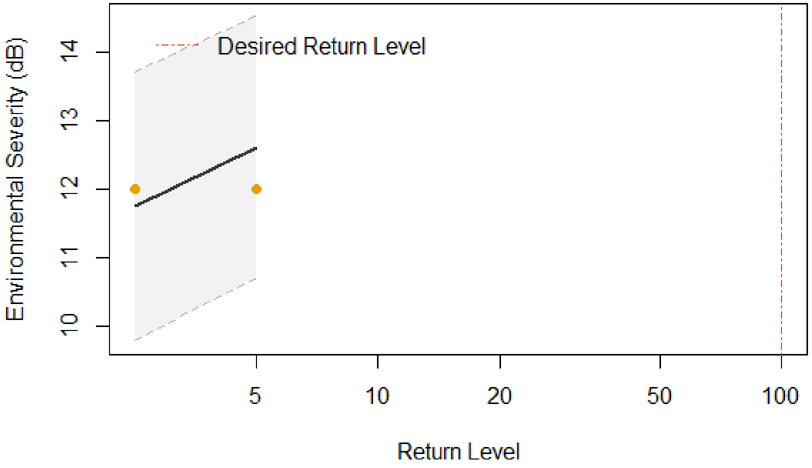
Component	Closure Time
n	100
Sample Mean	20.02 s
Sample Standard Deviation	1.12 s
Median	19.99 s
Inter-Quartile Range	19.38 s, 20.71 s
Range	16.03 s, 22.98 s
Outliers	16.03 s, 16.83 s, 22.98 s
Requirement	23.5 s
K-Factor	3.10
95% Confidence Bound on K-Factor	2.71 s

We can conclude:

- There does appear to be positive margin in the closure time. The confidence bound is quite close to the raw k-factor estimate, indicating reduced uncertainty. Decision-makers would have to decide whether the asserted margin of 3.1 sample standard deviations is sufficient for this component.

Environmental Reliability Example

- **Severity level:** estimate level at which 99th percentile corresponds to requirement with 95% confidence.
  - **Degree of extrapolation**: When n = 4, extrapolation is occurring beyond the 48th percentile. We would need a sample 75 times larger to avoid extrapolation.
  - **Validation metrics**: The estimated level varies greatly based on the choice of model.
    - Weibull – 2.47 dB
    - Logistic – -0.74 dB



We conclude:

- The model cannot be evaluated anywhere near where the prediction will occur, and
- The demonstration decision is extremely sensitive to selection of the model.

Environmental Reliability – Evidence-Driven Approach

- **Practical solutions:** Use summary statistics to evaluate performance margin with heuristic measures of uncertainty.

Component	Failure Mode
n	4
Test levels	+ 3 dB, + 6 dB, + 9 dB, +12 dB
# of failures observed	1
When failures occurred	Failure at +9 dB
Highest level at which no failures occurred (E_{pass})	+6 dB
95% lower bound at E_{pass}	47%

We can conclude:

- There is some evidence for positive margin, though there is still quite a bit of uncertainty. Other information is required for the evidence package to make a final assertion.

Evidence-Driven Approach to Test Planning

- An evidence-driven approach can prove very useful for test planning purposes.
- In performance reliability assessment…
  - Test planners can allocate more resources toward testing under conditions where there is less evidence.
- In environmental reliability assessment…
  - Test planners can start at higher levels where they expect their product to survive or are uncertain of their product's performance.
- In summary, an evidence-driven approach allows test planners to focus limited resources on areas with little evidence.
  - Reduces the need to test in areas with sufficient evidence.

Summary and Conclusions

- The ideal approach to reliability demonstration is often not feasible given on the constraints on data and available information.
- Model credibility when analyzing data is an important, yet often overlooked aspect of reliability analysis.
  - Is your inference data-driven or model-driven?
- Utilizing the concept of an "evidence-driven" package to assert reliability is a practical alternative for when it is not (yet) possible to fully demonstrate reliability.
  - Also avoids the risk of making statistically rigorous statements based on unjustifiable assumptions.

References

- Meeker, W.Q. and Escobar, L.A. (2003), “Reliability: The Other Dimension of Quality,” *Quality Technology & Quantitative Management*, 1 (1): 1-25.
- Lindsay, B. and Liu, J. (2009), “Model Assessment Tools for a Model False World,” *Statistical Science*, 24 (3): 303-318.
- Matthews, R.A.J. (2018), “Beyond ‘significance’: principles and practice of the Analysis of Credibility,” *Royal Society Open Science*, 5: 171047.
- Fernholz, L. T. and Gillespie, J. A. (2001) “Content-corrected tolerance limits based on the bootstrap.” *Technometrics*, 43(2):147–155.
- Scholz, F. (2005) “Nonparametric tail extrapolation.” Boeing Information & Support Services ISSTECH-95-014, Seattle, WA.
- Sharp, D., Wallstrom, T., and Wood-Schulz, M. (2003) “Physics package confidence: “one” vs. “1.0”.” Proceedings of the NEDPC 2003.
- Pilch, M., Trucano, T. G., and Helton, J. C. (2006) “Ideas underlying quantification of margins and uncertainties (QMU): a white paper.” Unlimited Release SAND2006-5001, Sandia National Laboratory, Albuquerque, New Mexico, 87185:2.
- Pilch, M., Trucano, T. G., and Helton, J. C. (2011) “Ideas underlying the quantification of margins and uncertainties.” Reliability Engineering & System Safety, 96(9):965–975.