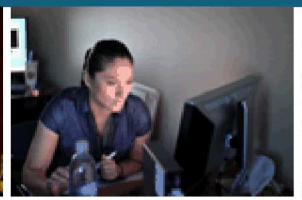
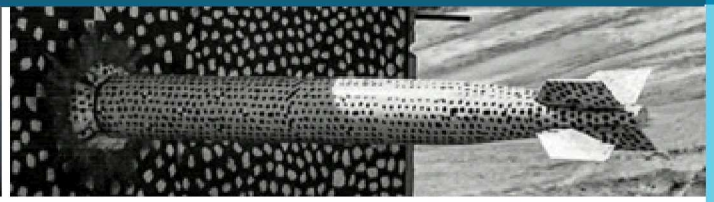
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2020-8985C

Heuristic Perspectives on Parametric Survival Analysis





PRESENTED BY

Thor D. Osborn, PhD, MBA, CAP

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Five Point Synopsis

- The choice of distribution used for survival (or time-to-event) analysis is often motivated by precedent, ease of use, or empirically demonstrated best fit to the data
- However, each of the commonly used parametric survival distributions represents a different fundamental underlying process mechanism
- Choosing the model based on accepted practical considerations fails to leverage process knowledge that could offer insight into the characteristic mechanism
- Conversely, the model that best fits the data may offer insight into the dominant mechanism governing the process at hand, accelerating comprehension
- Simulation of the common distributions via atomistic representations of their respective core mechanisms exposes informative heuristics for choosing distribution models and interpreting model fit

Addressing a Common Gap

- How often have you seen statements similar to these when reading scholarly journals or technical works?
  - The xxx distribution / hazard function can accommodate an appropriate shape for matching…
    - e.g., (Adelian et al., 2015), (Billinton & Allen, 1987)
  - The yyy distribution has often been used to describe…
    - e.g., (George, Seals, & Aban, 2014)
  - The zzz distribution fits these data well…
    - e.g., (Surendran & Tota-Maharaj, 2015), (Zare et al., 2014)
- Such statements imply that the author has made a conventional, non-controversial choice of distribution to describe the phenomena of interest – however:
  - The relative suitability of the chosen distribution *vs.* alternatives may not be addressed
  - Insight from the fundamental mechanism underlying the distribution may be lost

Common Distributions in Parametric Survival Analysis

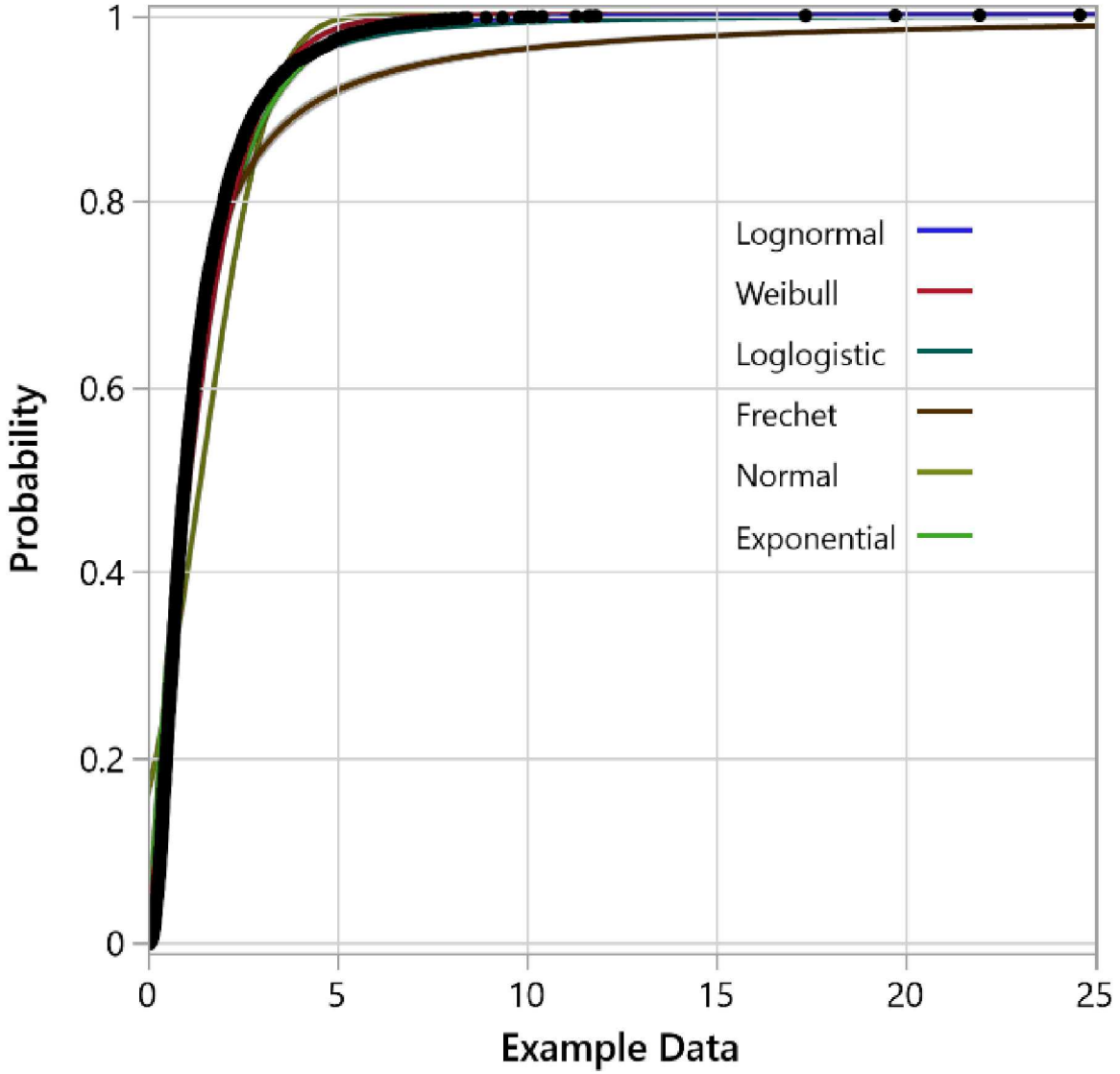
Five of the most common distributions used in parametric survival analysis:

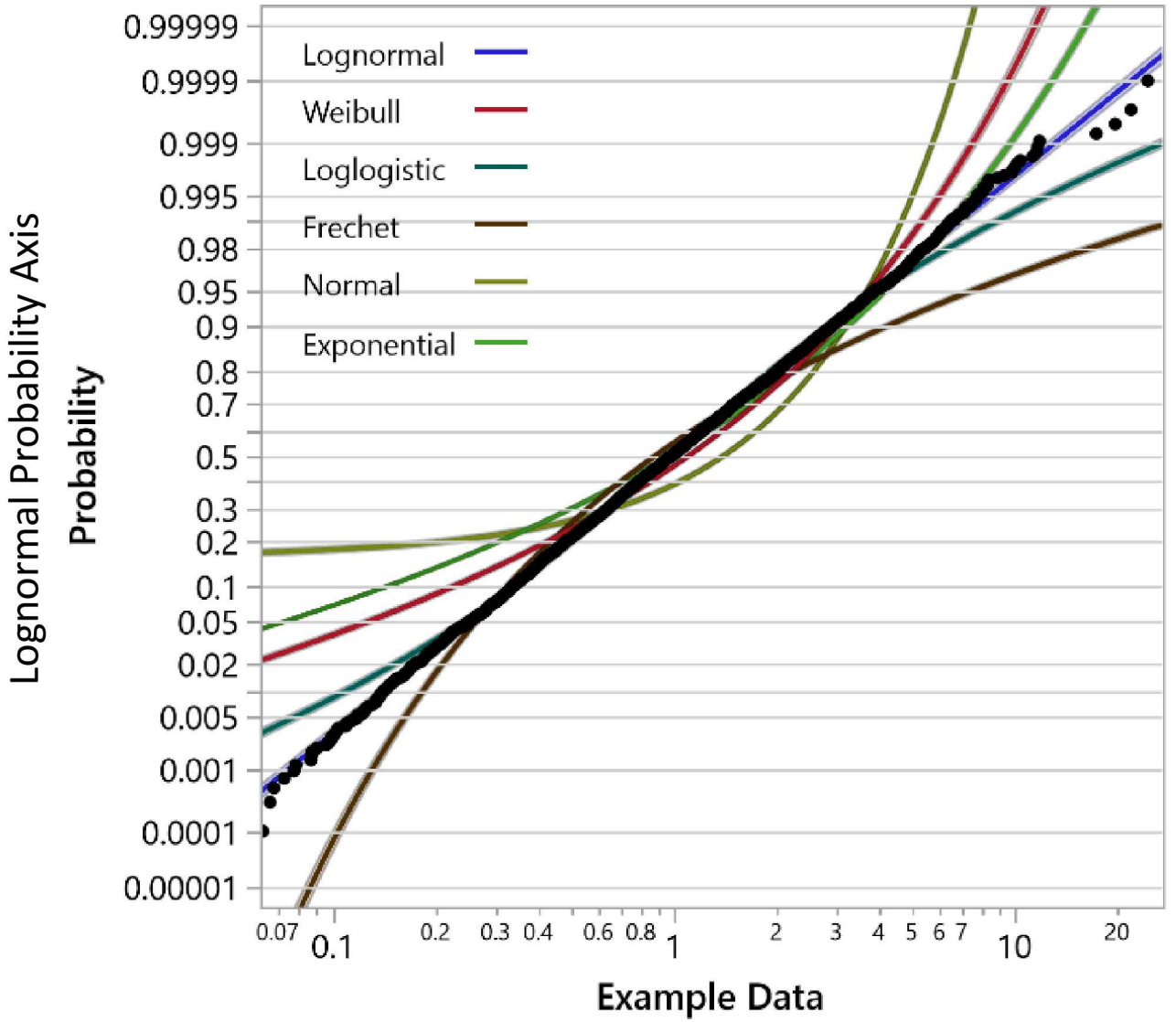
- Lognormal – the logarithm of the distribution is normally distributed
- Exponential – constant hazard rate (event probability)
- Weibull – shortest time to failure for elements of a system depending on all elements to function
- Fréchet – longest time to failure for elements of a system depending on any of multiple elements to function
- Loglogistic – the logarithm of the distribution is logistically distributed – time to event for a system comprised of cooperatively interacting elements

The shapes of the distributions differ because they model fundamentally different system archetypes

Demonstration – Fit All Common Distributions to a Sample Data Set

Demo Result – Fit All Common Distributions to a Sample Data Set





Underlying Mechanisms

Mechanistic Perspective on the Normal Distribution

The Normal distribution is not commonly used for survival analysis; however, it provides a familiar platform for introducing the mechanistic perspective.

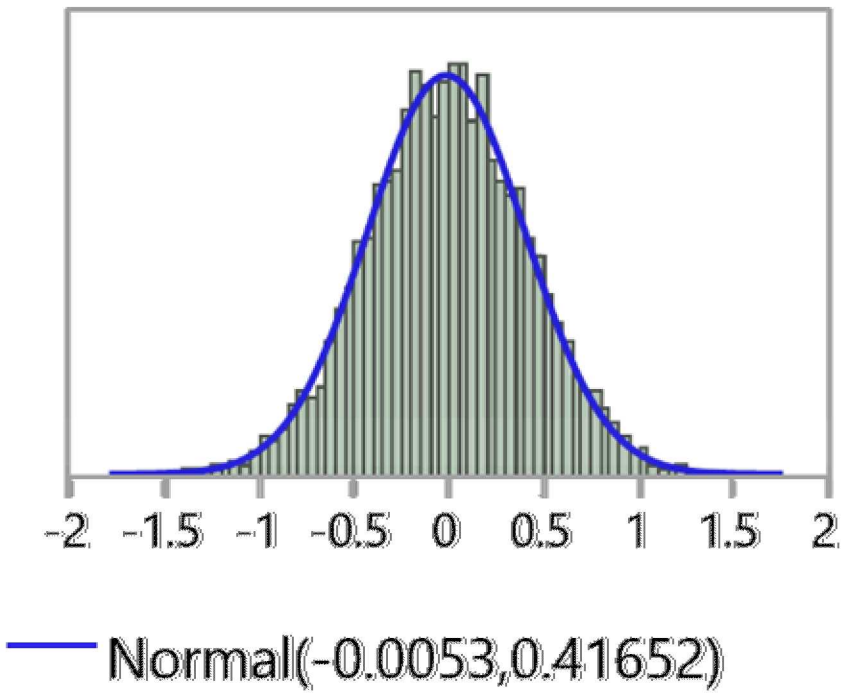
The Central Limit Theorem has been framed in various ways – one construction is that the distribution of sample means produced by randomly drawing samples from any fixed distribution will yield the Normal distribution.

An implication of this perspective is that the normal distribution may be contemplated as a summation of many small uncorrelated effects (ϵ_i).

$$x_t = x_0 + \sum_{i=1}^{t} \epsilon_i$$

Demonstration – Generate and Fit Synthetic Normal Data

Demo Result – Generate and Fit Synthetic Normal Data



Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	-0.005325	-0.016873	0.0062233
Dispersion	σ	0.41652	0.4085138	0.4248486

Measure	
-2*LogLikelihood	5430.1772
AICc	5434.1796
BIC	5447.2115

Mechanistic Perspective on the Lognormal Distribution

A simple mathematical form results if we adopt Kalecki's approach to Gibrat's law of proportionate effect, as recast by Sutton (Kalecki, 1945; Sutton, 1997):

Each small random fluctuation ϵ_i increases or decreases x in proportion to the current basis.

$$x_t - x_{t-1} = \epsilon_t x_{t-1}$$

The value of x at time t results from the multiplicative effect of many small fluctuations on the original value of x at time 0.

$$x_t = x_0 \prod_{i=1}^{t} (1 + \epsilon_i)$$

Logarithmic transformation yields the corresponding summation.

$$\ln x_t = \ln x_0 + \sum_{i=1}^{t} \ln(1 + \epsilon_i)$$

For infinitesimal fluctuations $\epsilon_i \ll 1, \ln(1 + \epsilon_i)$ may be approximated as ϵ_i based on the Taylor series expansion.

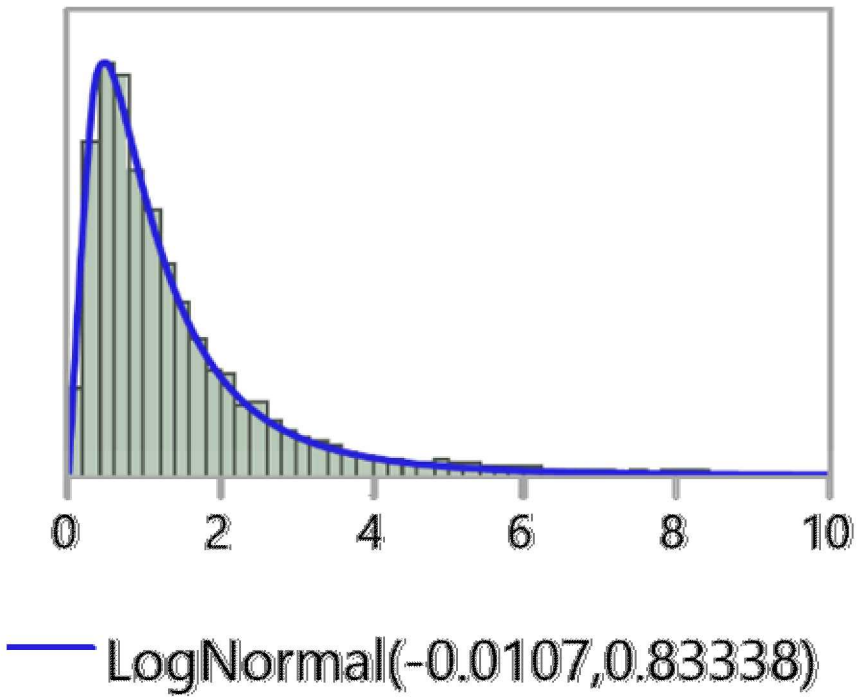
$$\ln x_t = \ln x_0 + \sum_{i=1}^{t} \epsilon_i$$

Rearranging, the growth of x over the time interval is clearly lognormal, as taking the logarithm reveals a summation of small fluctuations.

$$\frac{x_t}{x_0} = e^{\sum_{i=1}^{t} \epsilon_i}$$

Demonstration – Generate and Fit Synthetic Lognormal Data

Demo Result – Generate and Fit Synthetic Lognormal Data



Fitted LogNormal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Scale	μ	-0.010656	-0.03376	0.012448
Shape	σ	0.8333788	0.8173079	0.8499834

Measure	
-2*LogLikelihood	12260.154
AICc	12264.157
BIC	12277.189

Mechanistic Perspective on the Weibull and Fréchet Distributions

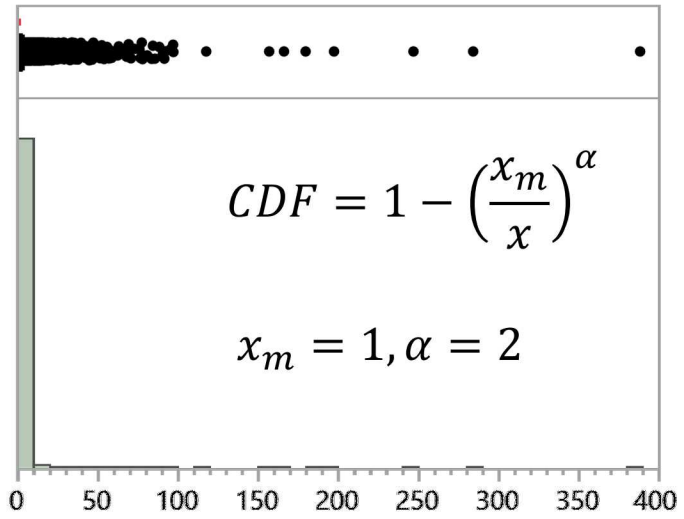
The Fréchet distribution represents maximal extreme values. For example, the Fréchet may be used for a set of samples of observations drawn from a random process where each sample is represented by its maximum value.

The Weibull distribution represents minimal extreme values. For example, the Weibull may be used for a set of samples drawn from a random process where each sample is represented by its minimum value.

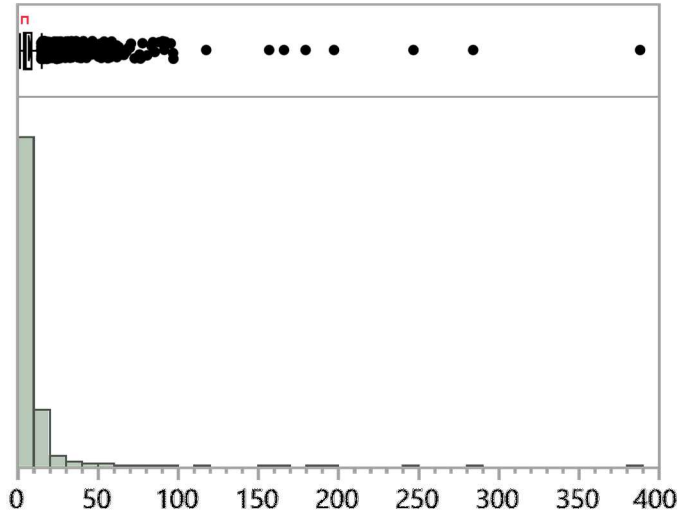
Demonstration – Generate and Fit Synthetic Fréchet Data

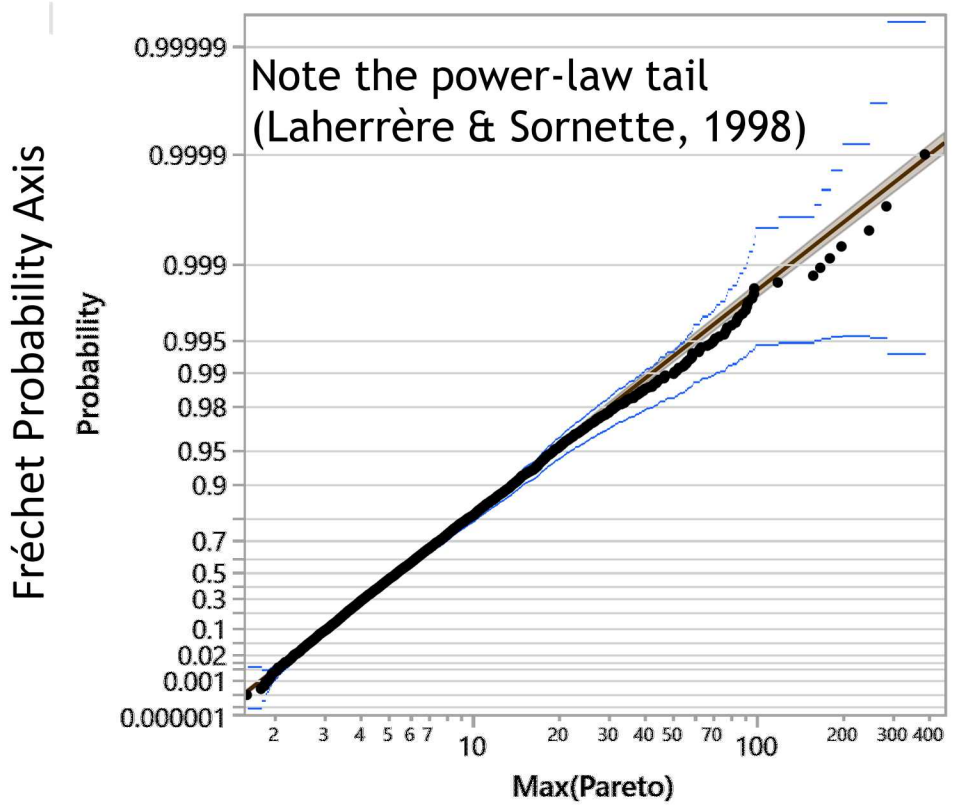
Demo Result - Generate and Fit Synthetic Fréchet Data

Pareto



Max(Pareto)



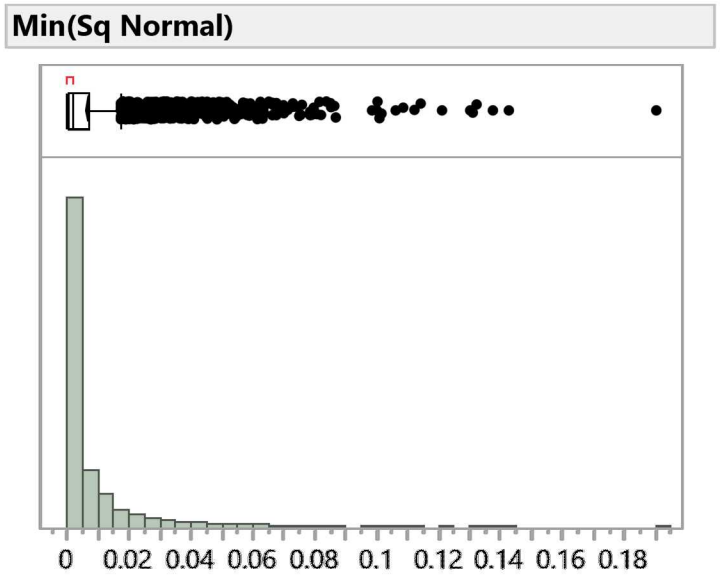


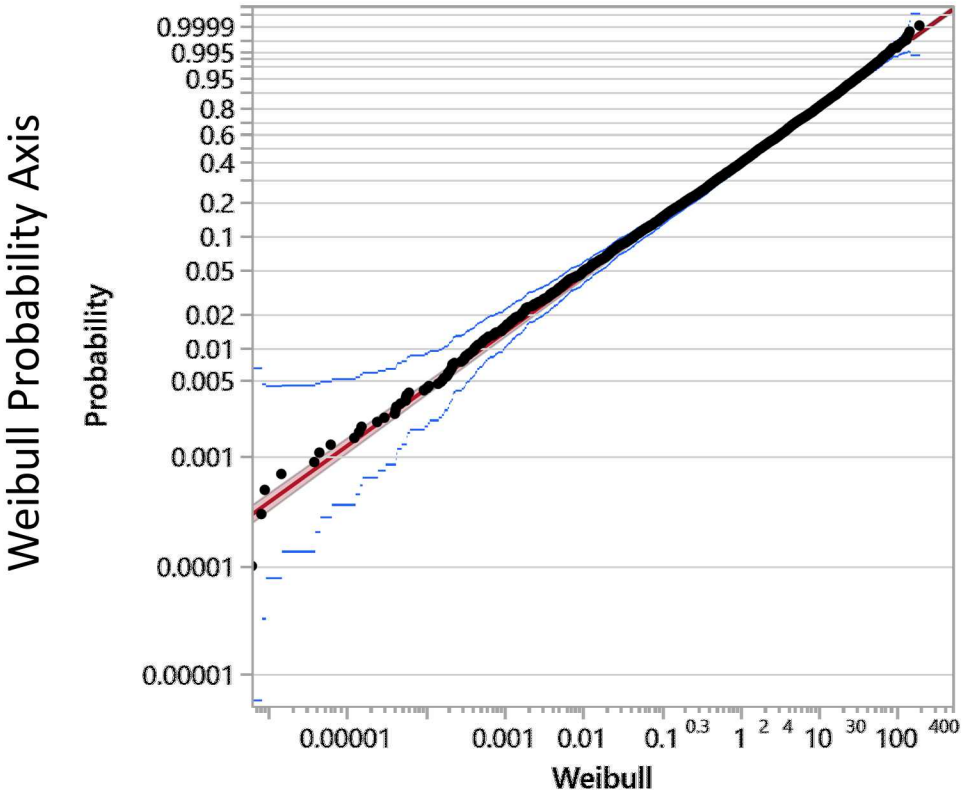
Model Comparisons

Distribution	AICc	-2Loglikelihood	BIC
Frechet	26698.497	26694.495	26711.529
Loglogistic	27427.480	27423.478	27440.512
Lognormal	27753.804	27749.801	27766.836
Weibull	30666.313	30662.311	30679.345
Exponential	30899.259	30897.258	30905.776

Demonstration – Generate and Fit Synthetic Weibull Data

Demo Result – Generate and Fit Synthetic Weibull Data





Model Comparisons

Distribution	AICc	-2Loglikelihood	BIC
Weibull	24272.346	24268.344	24285.378
Loglogistic	25004.558	25000.556	25017.590
Lognormal	25285.879	25281.877	25298.911
Frechet	27997.787	27993.784	28010.819
Exponential	29120.290	29118.289	29126.806

Mechanistic Perspective on the Loglogistic Distribution

The demonstrations up to this point have all used independent samples. The Loglogistic distribution is similar to the Lognormal, but occurs when the data in each sample are correlated.

This can be attained for small samples over short sequences by using autocorrelated data.

If Y1 and Y2 are independent normally-distributed random variables then correlated normally-distributed random variables X1 and X2 may be generated as follows (Cordes, 2019):

$$X_1 = \cos\phi \cdot Y_1 + \sin\phi \cdot Y_2$$

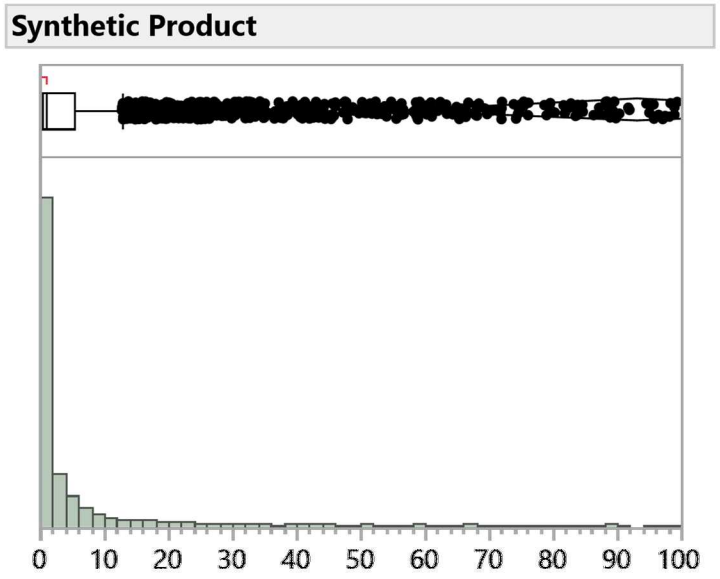
$$X_2 = \sin\phi \cdot Y_1 + \cos\phi \cdot Y_2$$

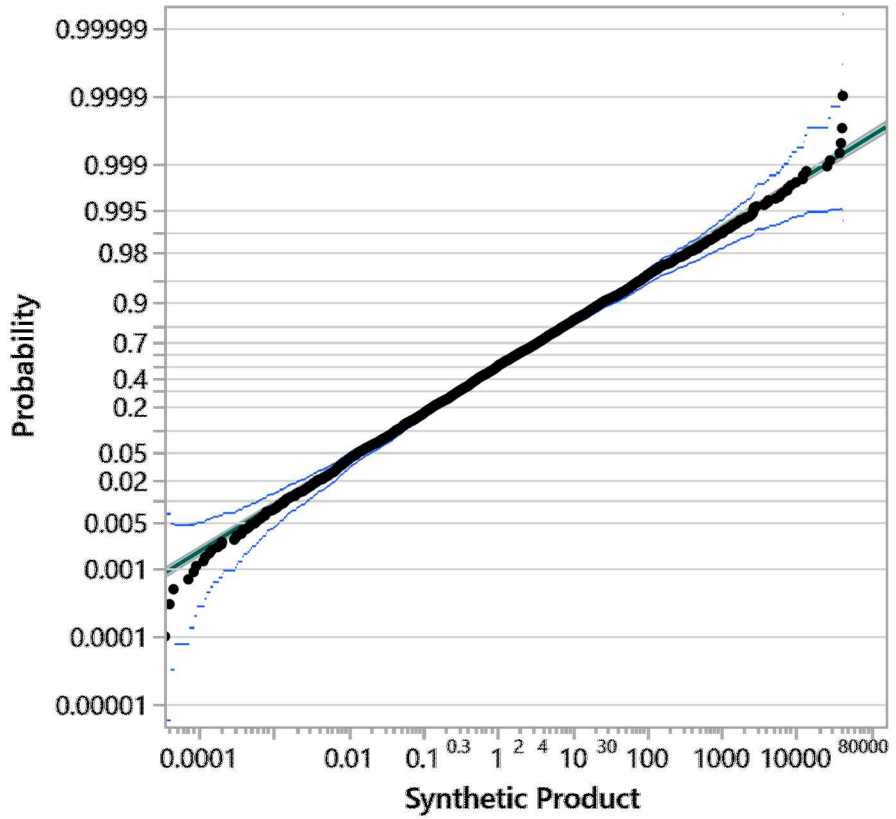
Where the value of ϕ necessary to produce the correlation coefficient is determined by:

$$\phi = \frac{1}{2}\sin^{-1}\rho \cdot X1 \cdot X2$$

Demonstration – Generate and Fit Synthetic Loglogistic Data

Demo Result – Generate and Fit Synthetic Loglogistic Data





Model Comparisons

Distribution	AICc	-2Loglikelihood	BIC
Loglogistic	24380.438	24376.435	24393.470
Lognormal	24491.281	24487.279	24504.313
Frechet	25478.359	25474.356	25491.391
Weibull	25729.508	25725.506	25742.540
Exponential	55339.088	55337.087	55345.604

Example – San Francisco Zoning Variance Analysis

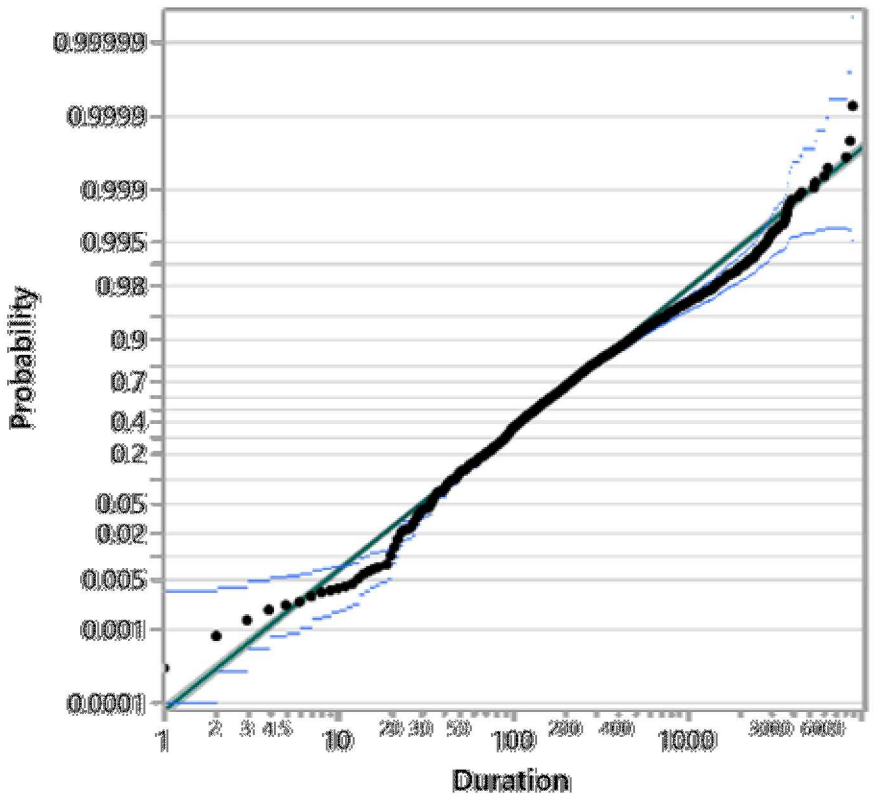
The San Francisco Zoning Variance Process

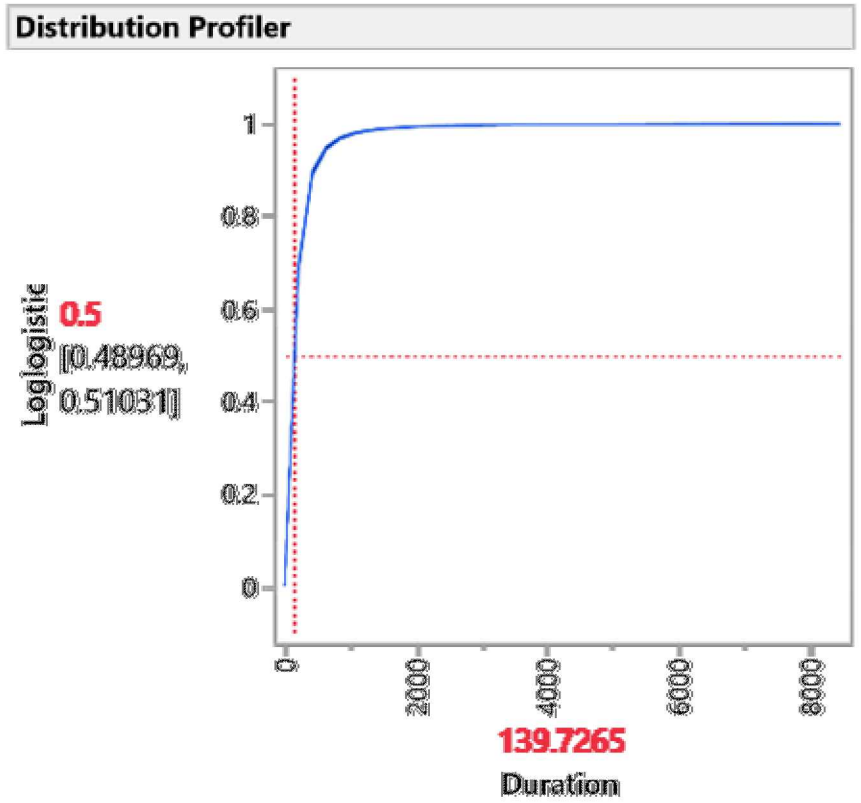
Process Step	San Francisco
Assemble Preliminary Variance Application and Exhibits	Applicant fills out application form and gathers necessary drawings, evidence, and justification per variance requirements
Preliminary Review and Revision	Applicant has Intake Appointment with a Planner to ensure application meets requirements
Submit Plan	Applicant submits revised application and materials to Planning Department
Verify Need for Variance	Assigned Planner checks plan against Planning Code, San Francisco General Plan and Planning Department policies
Community Notification	Planning Department notifies property owners within 300 feet of subject property
Community Input	Assigned Planner gathers comments and concerns from the neighborhood during the notification period
Public Hearing	Conducted by Zoning Administrator
Final Determination	Zoning Administrator issues Decision

Process is characterized by substantial interaction among the Applicant, Assigned Planner, and Local Property Owners, converging to a single formal decision by the Zoning Administrator

Survival Analysis of San Francisco Zoning Variance Cases

Distribution	Scale
☑ Nonparametric	○
☐ Lognormal	○
☐ Weibull	○
☑ Loglogistic	◉
☐ Frechet	○
☐ Normal	○
☐ SEV	○
☐ Logistic	○
☐ LEV	○
☐ Exponential	○
☐ LogGenGamma	
☐ GenGamma	
☐ TH Lognormal	
☐ TH Weibull	
☐ TH Loglogistic	
☐ TH Frechet	
☐ DS Lognormal	
☐ DS Weibull	
☐ DS Loglogistic	
☐ DS Frechet	





Statistics

Model Comparisons

Distribution	AICc	-2Loglikelihood	BIC
Loglogistic	85336.196	85332.194	85349.829
Lognormal	85601.223	85597.221	85614.855

Loglogistic outcome makes sense given the cooperative action among multiple parties necessary to resolve each case

Summary of Survival Distribution Heuristics

Heuristics for Distribution Selection and Interpretation of Empirical Results

Distribution	Characteristic Utility	Example
Lognormal *Product of many small fluctuations*	Where the final outcome of a process is the result of a long sequence of small, independent incremental steps, each building on the result of all prior impacts	Plant growth / growth of terminal organs (Koyama, Yamamoto, & Ushio, 2016); Age of disease onset (Limpert, Stahel, & Abbt, 2001)
Fréchet *Maximum extreme values*	Where the final outcome of a process represents the greatest duration among an ensemble of independent subprocesses	Annual maximum daily rainfall (Papalexiou & Koutsoyiannis, 2013)
Weibull *Minimum extreme values*	Where the final outcome of a process represents the least duration among an ensemble of independent subprocesses	Failure of a complex, non-redundant system
Loglogistic *Cooperation among groups*	Where the final outcome of a process results from the collective action of two or more entities having mutual influence over each other	San Francisco zoning variance approval process; Job offer acceptance process

Your Feedback Is Welcome!

- Was this presentation interesting?
- Do you think that considering the fundamental behavior addressed by each common parametric survival distribution will help you make better distribution model choices in your work?
- Do you think that considering these fundamental behaviors will help you to better understand the processes you are analyzing?

Please address your comments to:

Thor D. Osborn

Sandia National Laboratories

tdosbor@sandia.gov

References

Adelian, R., Jamali, J., Zare, N., Ayatollahi, S. M. T., Pooladfar, G. R., & Roustaei, N. 2015. Comparison of Cox's Regression Model and Parametric Models in Evaluating the Prognostic Factors for Survival after Liver Transplantation in Shiraz during 2000-2012. ***International journal of organ transplantation medicine***, 6(3): 119-125.

Billinton, R., & Allen, R. N. 1987. ***Reliability Evaluation of Engineering Systems: Concepts and Techniques***. New York: Plenum Press.

Cordes, J. M. 2019. Generating Correlated Random Variables: 1-4: Cornell.

Du, X., Li, M., Zhu, P., Wang, J., Hou, L., Li, J., Meng, H., Zhou, M., & Zhu, C. 2018. Comparison of the flexible parametric survival model and Cox model in estimating Markov transition probabilities using real-world data. ***PLOS ONE***, 13: e0200807.

Focke, W. W., Westhuizen, I. v. d., Musee, N., & Loots, M. T. 2017. Kinetic interpretation of log-logistic dose-time response curves. ***Scientific Reports***: 1-11.

George, B., Seals, S., & Aban, I. 2014. Survival analysis and regression models. ***Journal of nuclear cardiology : official publication of the American Society of Nuclear Cardiology***, 21(4): 686-694.

Kalecki, M. 1945. On the Gibrat Distribution. ***Econometrica***, 13(2): 161-170.

Koyama, K., Yamamoto, K., & Ushio, M. 2016. A lognormal distribution of the lengths of terminal twigs on self-similar branches of elm trees. ***Proceedings of the Royal Society B***, 284.

Laherrère, J., & Sornette, D. 1998. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. ***European Physical Journal B***, 2: 525-539.

Limpert, E., Stahel, W. A., & Abbt, M. 2001. Log-normal distributions across the sciences: keys and clues. ***BioScience***, 51(5): 341-352.

Papalexiou, S. M., & Koutsoyiannis, D. 2013. Battle of extreme value distributions: A global survey on extreme daily rainfall. ***Water Resources Research***, 49: 187-201.

Surendran, S., & Tota-Maharaj, K. 2015. Log logistic distribution to model water demand data. ***Procedia Engineering***, 119: 798-802.

Sutton, J. 1997. Gibrat's legacy. ***Journal of Economic Literature***, 35(March): 40-59.

Wikipedia. 2020. Natural logarithm, Vol. 2020: Wikipedia.

Zare, A., Mahmoodi, M., Mohammad, K., Zeraati, H., Hosseini, M., & Naieni, K. H. 2014. Comparison between parametric and semi-parametric cox models in modeling transition rates of a multi-state model: application in patients with gastric cancer undergoing surgery at the Iran cancer institute. ***Asian Pac J Cancer Prev***, 14(11): 6751-6755.