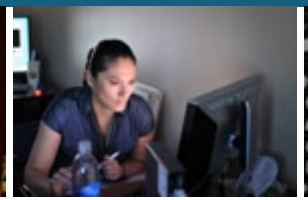
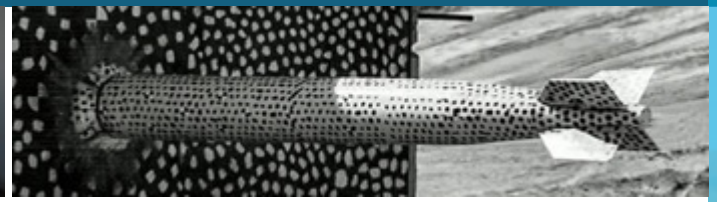
SAND2022-0680C

Sandia National Laboratories

# Heuristic Perspectives on Parametric Survival Analysis





*PRESENTED BY*

Thor D. Osborn, PhD, MBA, CAP

tdosbor@sandia.gov

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

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.

SAND2022-XXXX

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

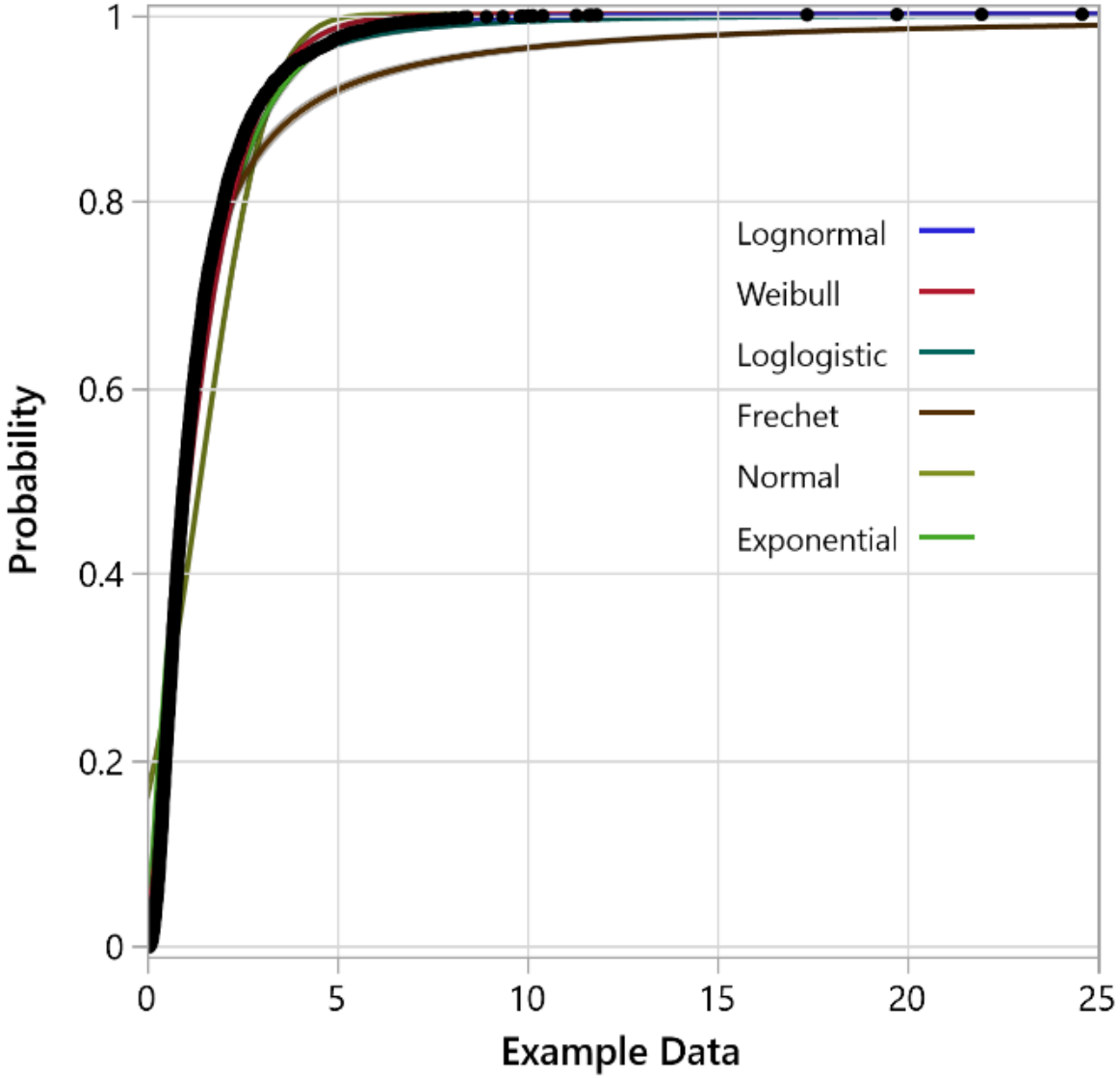
Five of the most common distributions used in parametric survival analysis:

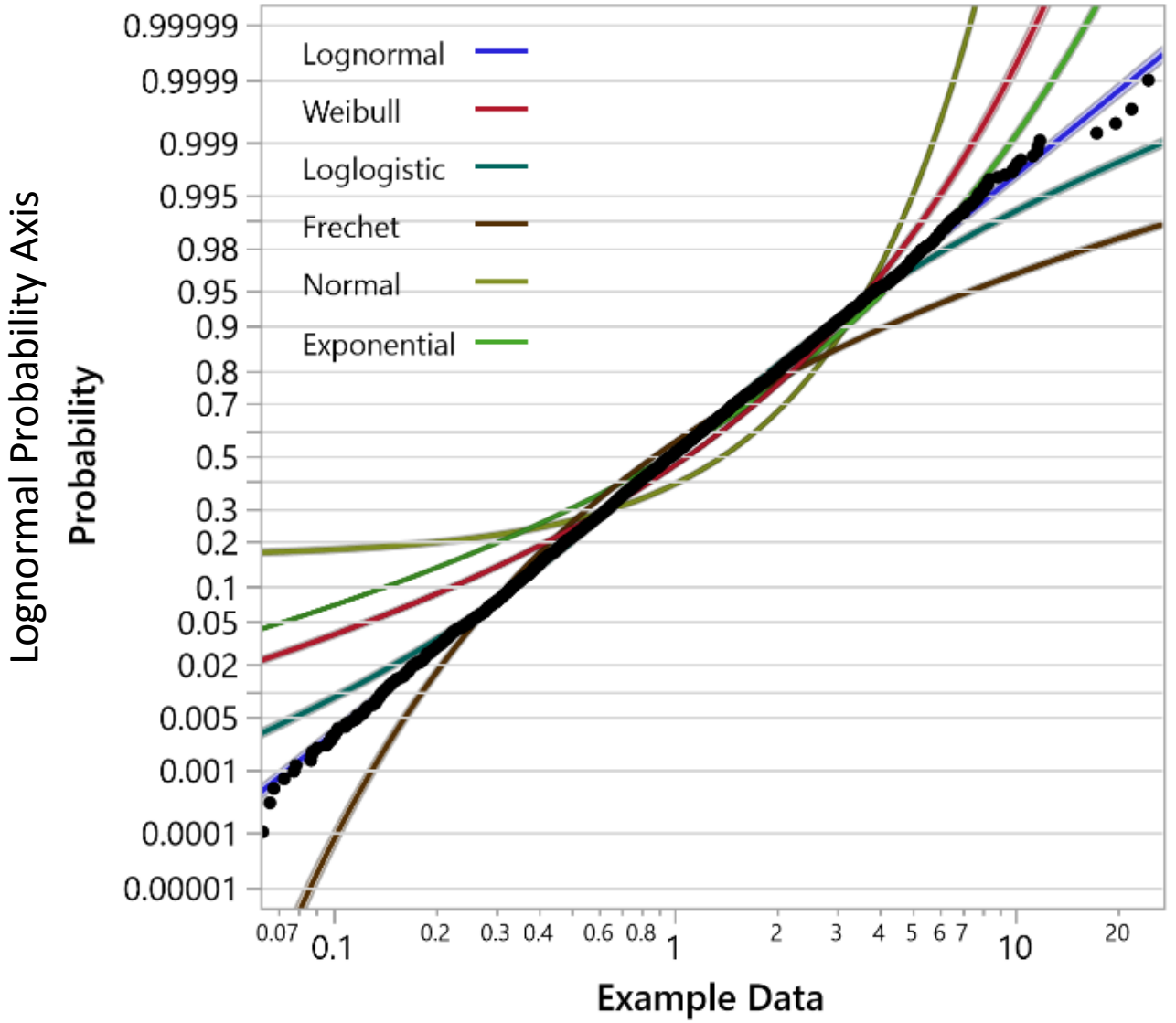
- Lognormal – the logarithm of the distribution is normally distributed
- Exponential – constant hazard rate (event probability); special case of Weibull
- Weibull – shortest time to failure for elements of a system depending on all elements to function
- Fréchet – longest time to failure for elements 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

# Example – Fit All Common Distributions to a Sample Data Set

Sample is N = 5000 points random Lognormal



# 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.

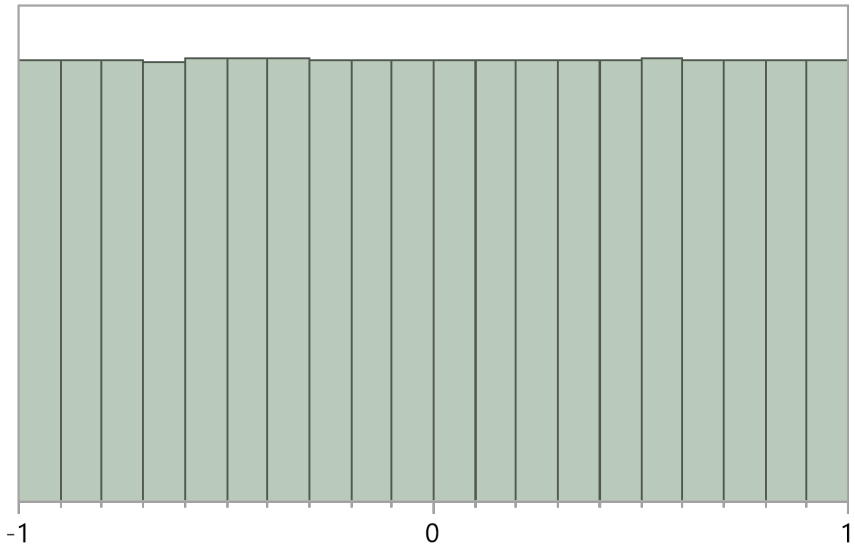
The Central Limit Theorem has been framed in various ways – one construction is that the distribution of sample means produced by randomly drawing samples from any fixed distribution will yield the Normal distribution.

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

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

# Demonstration – Generate Synthetic Normal Data

Start with 5,000,000 random uniform observations between ±1




| Summary Statistics | |
|---|---|
| Mean | -0.000134 |
| Std Dev | 0.5772901 |
| Std Err Mean | 0.0002582 |
| Upper 95% Mean | 0.0003718 |
| Lower 95% Mean | -0.00064 |
| N | 5000000 |

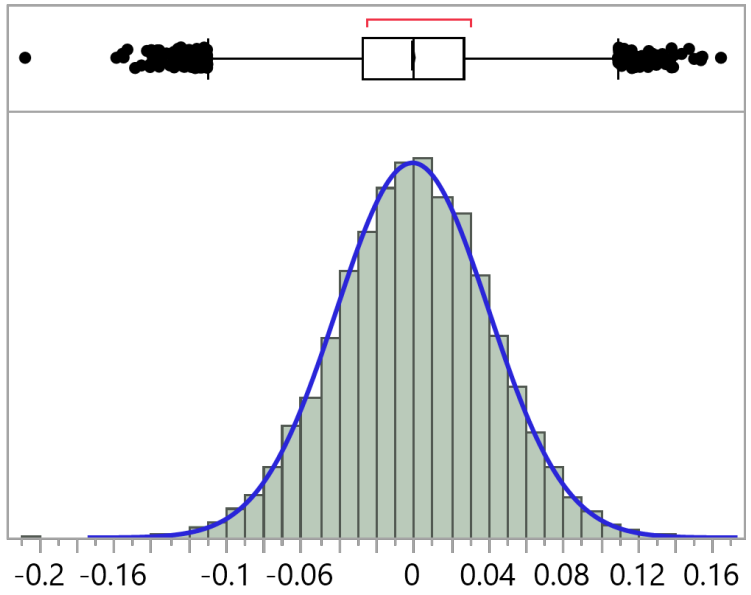
$$\approx \frac{2}{\sqrt{12}} \approx 0.57735$$

Empirical SD matches theoretical expectation for a uniformly distributed variable over a span of two (2) units

Carve into 25,000 samples of 200 observations each and take sample means: Each result is equivalent to summing 200 uniform random uncorrelated fluctuations between ±0.005

| | Sample | Mean(Data) |
|---|---|---|
| 1 | 1 | -0.015466011 |
| 2 | 2 | 0.0431644838 |
| 3 | 3 | -0.02016792 |
| 4 | 4 | 0.0032842523 |
| 5 | 5 | 0.0300040112 |
| 6 | 6 | 0.0094529965 |
| 7 | 7 | 0.0147012687 |
| 8 | 8 | 0.0582764187 |
| 9 | 9 | 0.0210799995 |
| 10 | 10 | -0.02462875 |
| 11 | 11 | 0.031939952 |
| 12 | 12 | -0.013106945 |
| 13 | 13 | -0.05453438 |
| 14 | 14 | -0.070809957 |
| 15 | 15 | -0.030572554 |

# Demonstration – Fit Synthetic Normal Data




Normal(-0.0001,0.04072)

**Summary Statistics**

| | |
|---|---|
| Mean | -0.000134 |
| Std Dev | 0.0407185 |
| Std Err Mean | 0.0002575 |
| Upper 95% Mean | 0.0003705 |
| Lower 95% Mean | -0.000639 |
| N | 25000 |

**Fitted Normal**

**Parameter Estimates**

| Type | Parameter | Estimate | Lower 95% | Upper 95% |
|---|---|---|---|---|
| Location | μ | -0.000134 | -0.000639 | 0.0003705 |
| Dispersion | σ | 0.0407185 | 0.0403647 | 0.0410786 |

| Measure | |
|---|---|
| -2*LogLikelihood | -89107.75 |
| AICc | -89103.75 |
| BIC | -89087.49 |

**Goodness-of-Fit Test**

KSL Test

| D | | Prob>D |
|---|---|---|
| 0.003675 | > | 0.1500 |

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

# 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 $\epsilon_i$ 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 $\epsilon_i$ based on the Taylor series expansion.

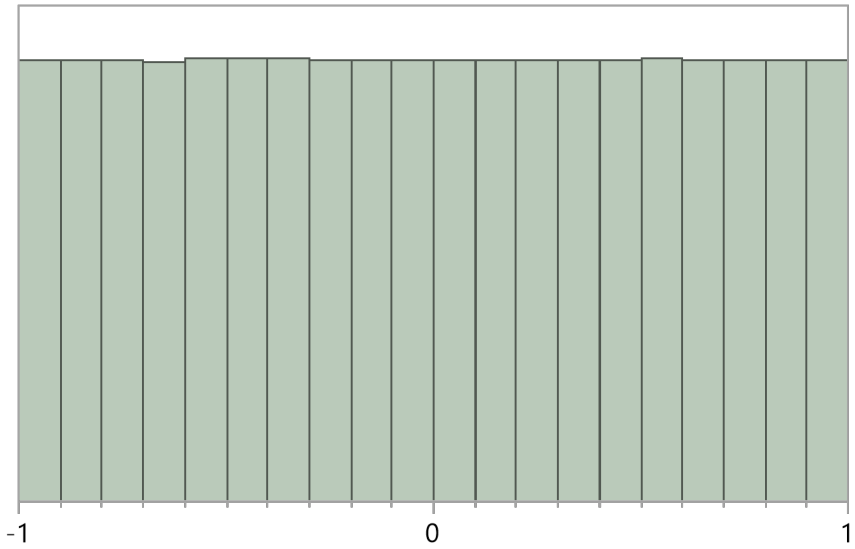
$$\ln x_t = \ln x_0 + \sum_{i=1}^{t} \epsilon_i$$

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

Start with 5,000,000 random uniform observations between ±1




| Summary Statistics | |
|---|---|
| Mean | -0.000134 |
| Std Dev | 0.5772901 |
| Std Err Mean | 0.0002582 |
| Upper 95% Mean | 0.0003718 |
| Lower 95% Mean | -0.00064 |
| N | 5000000 |

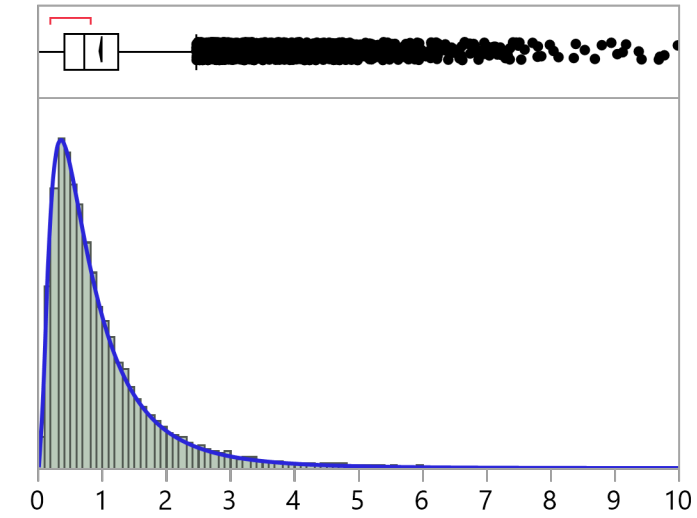
| | Sample | Fluctuation Product |
|---|---|---|
| 1 | 1 | 0.54451099 |
| 2 | 2 | 1.70781912 |
| 3 | 3 | 0.48427478 |
| 4 | 4 | 0.76704634 |
| 5 | 5 | 1.25331796 |
| 6 | 6 | 0.87479432 |
| 7 | 7 | 0.95993534 |
| 8 | 8 | 2.4141732 |
| 9 | 9 | 1.10827065 |
| 10 | 10 | 0.4256548 |
| 11 | 11 | 1.3612483 |
| 12 | 12 | 0.54600655 |
| 13 | 13 | 0.24923194 |
| 14 | 14 | 0.17580188 |
| 15 | 15 | 0.38964198 |

Carve into 25,000 samples of 200 observations each and take within-sample products as:

$$Product_j = \prod_{i=1}^{200}(1 + \frac{Observation_i}{10})$$

# Demonstration – Fit Synthetic Lognormal Data

**Fluctuation Product**




**Quantiles**

| | | |
|---|---|---|
| 100.0% | maximum | 19.3797595 |
| 99.5% | | 5.63618996 |
| 97.5% | | 3.46749643 |
| 90.0% | | 2.02321679 |
| 75.0% | quartile | 1.23962293 |
| 50.0% | median | 0.7183036 |
| 25.0% | quartile | 0.412731 |
| 10.0% | | 0.25050556 |
| 2.5% | | 0.14144813 |
| 0.5% | | 0.08358822 |
| 0.0% | minimum | 0.01071355 |

**Summary Statistics**

| | |
|---|---|
| Mean | 0.9932891 |
| Std Dev | 0.9555733 |
| Std Err Mean | 0.0060436 |
| Upper 95% Mean | 1.0051349 |
| Lower 95% Mean | 0.9814433 |
| N | 25000 |

**Fitted LogNormal**

**Parameter Estimates**

| Type | Parameter | Estimate | Lower 95% | Upper 95% |
|---|---|---|---|---|
| Scale | μ | -0.336956 | -0.347074 | -0.326838 |
| Shape | σ | 0.8162028 | 0.8091006 | 0.8234096 |

| Measure | |
|---|---|
| -2*LogLikelihood | 43944.519 |
| AICc | 43948.52 |
| BIC | 43964.773 |

**Goodness-of-Fit Test**

Kolmogorov's D

| D | | Prob>D |
|---|---|---|
| 0.004710 | > | 0.1500 |

Note: Ho = The data is from the LogNormal distribution. Small p-values reject Ho.

# Mechanistic Perspective on the Weibull and Fréchet Distributions

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

# Example – Generate Synthetic Fréchet Data

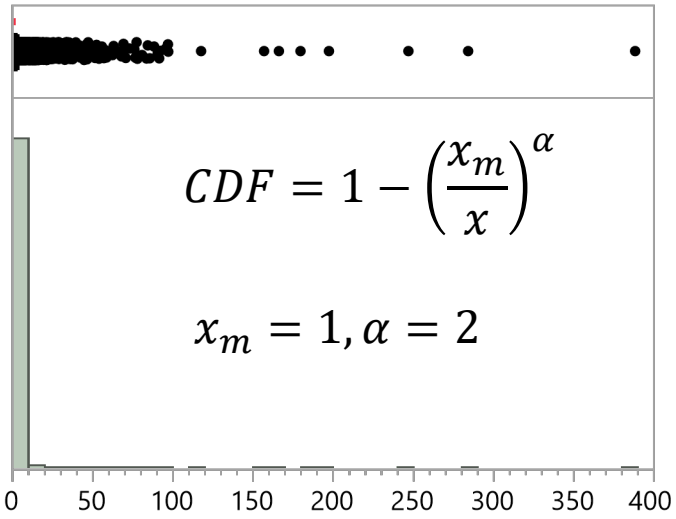
100,000 random values grouped in samples of 20 (5,000 samples)

| | Random Uniform | Pareto | Sample |
|---|---|---|---|
| 1 | 0.4369458593 | 1.3326770589 | 1 |
| 2 | 0.0374987419 | 1.0192937167 | 1 |
| 3 | 0.4430181412 | 1.3399218676 | 1 |
| 4 | 0.3010679858 | 1.1961414306 | 1 |
| 5 | 0.3553666447 | 1.2454996598 | 1 |
| 6 | 0.2121213775 | 1.1266015425 | 1 |
| 7 | 0.1621927414 | 1.0925163388 | 1 |
| 8 | 0.6296886236 | 1.6432985531 | 1 |
| 9 | 0.9866800073 | 8.6645897917 | 1 |
| 10 | 0.1778434103 | 1.1028659512 | 1 |
| 11 | 0.2534119368 | 1.1573360442 | 1 |
| 12 | 0.1026394516 | 1.0556416428 | 1 |
| 13 | 0.1302897299 | 1.0722910984 | 1 |
| 14 | 0.473981397 | 1.3787942012 | 1 |
| 15 | 0.0568378926 | 1.0296907892 | 1 |
| 16 | 0.2189468192 | 1.1315133892 | 1 |
| 17 | 0.6204989851 | 1.6232803412 | 1 |
| 18 | 0.232053366 | 1.141128309 | 1 |
| 19 | 0.9349860428 | 3.9219016576 | 1 |
| 20 | 0.3415118409 | 1.2323271452 | 1 |
| 21 | 0.0720837915 | 1.0381153663 | 2 |
| 22 | 0.3525312331 | 1.2427695058 | 2 |
| 23 | 0.9204713975 | 3.545996678 | 2 |

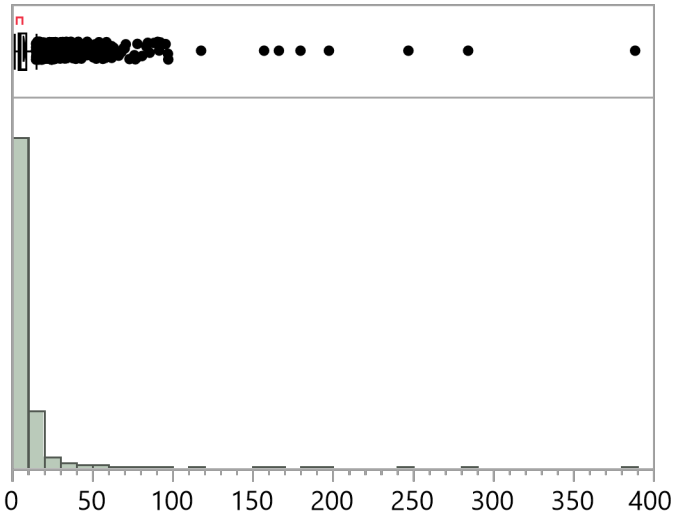
Sample 1

Max (Sample j)
for
j = 1 to 5000

**Pareto**

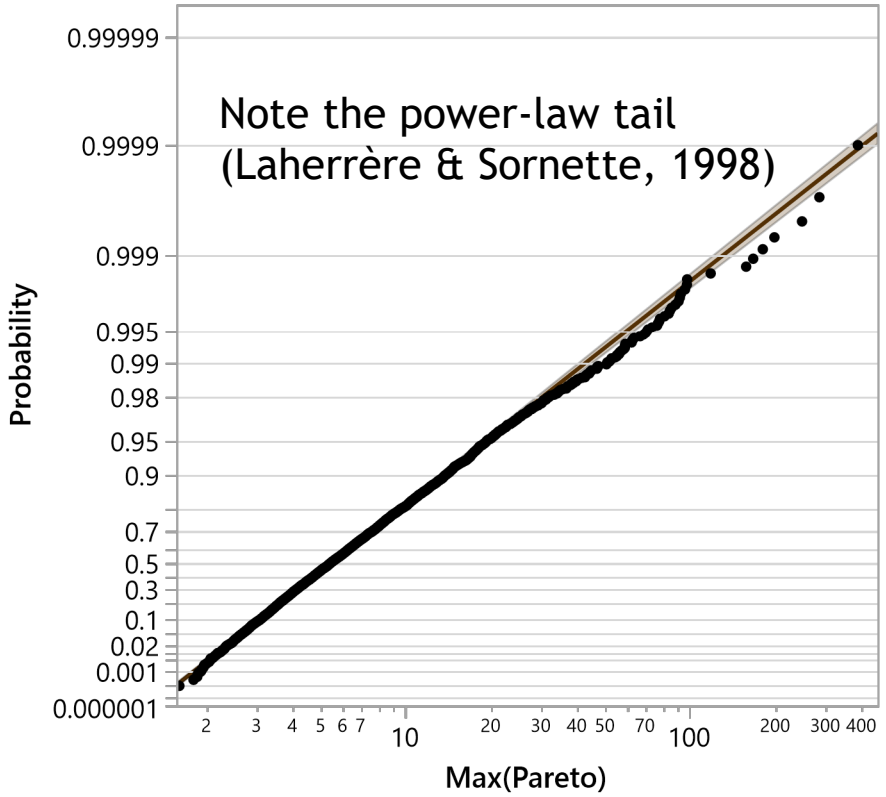


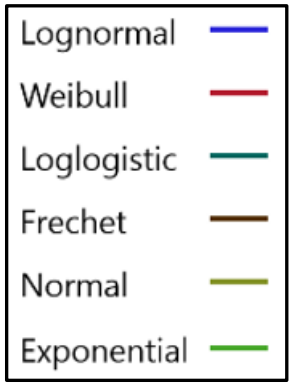

**Max(Pareto)**

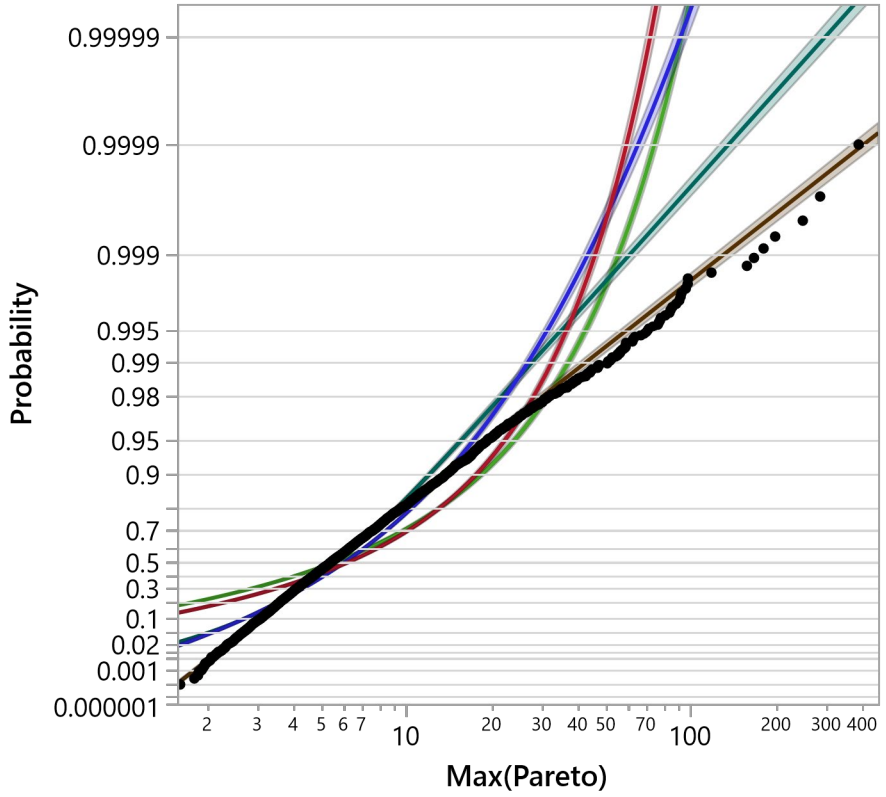


# Example - Fit Synthetic Fréchet Data

Fréchet Probability Axis




**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 |

Fréchet clearly fits these data better than the other common survival distributions shown

# Example – Generate Synthetic Weibull Data

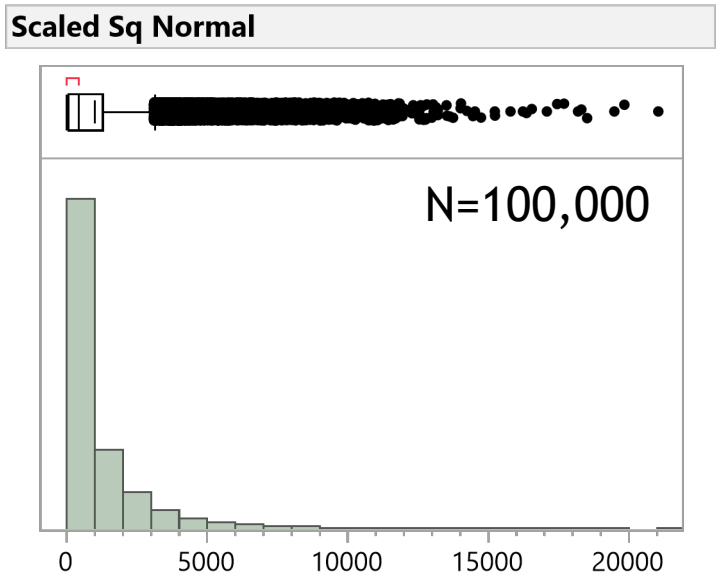
100,000 random values grouped in samples of 20 (5,000 samples)

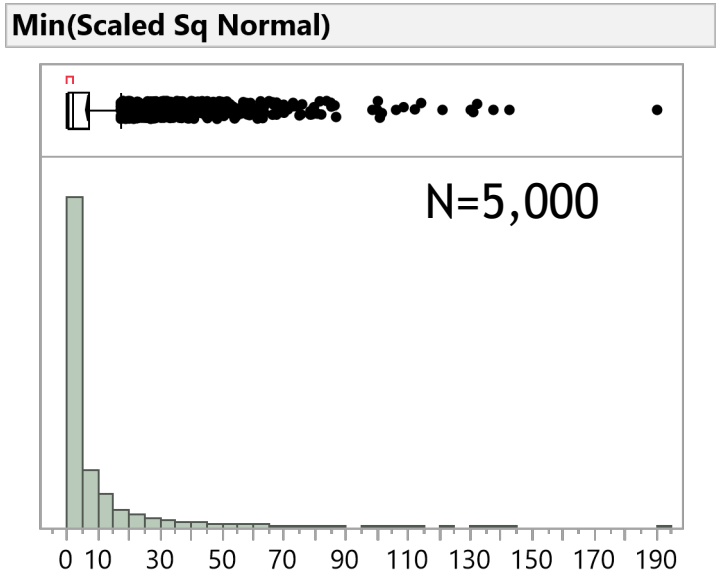
| | Sq Normal | Scaled Sq Normal | Sample |
|---|---|---|---|
| 1 | 2.5843429656 | 2584.3429656 | 1 |
| 2 | 0.3582783086 | 358.27830863 | 1 |
| 3 | 1.3867246571 | 1386.7246571 | 1 |
| 4 | 0.0000898804 | 0.0898803892 | 1 |
| 5 | 0.0005692624 | 0.5692623855 | 1 |
| 6 | 0.6056701697 | 605.67016966 | 1 |
| 7 | 0.9390482565 | 939.04825651 | 1 |
| 8 | 0.5607257843 | 560.72578425 | 1 |
| 9 | 0.3485738455 | 348.57384549 | 1 |
| 10 | 0.114736994 | 114.73699404 | 1 |
| 11 | 0.0080481447 | 8.0481447307 | 1 |
| 12 | 0.1781835378 | 178.18353778 | 1 |
| 13 | 0.6991829743 | 699.18297434 | 1 |
| 14 | 6.1874438e-8 | 0.0000618744 | 1 |
| 15 | 1.0528093712 | 1052.8093712 | 1 |
| 16 | 4.5210171687 | 4521.0171687 | 1 |
| 17 | 6.7637043405 | 6763.7043405 | 1 |
| 18 | 0.7180125971 | 718.01259714 | 1 |
| 19 | 2.5695985864 | 2569.5985864 | 1 |
| 20 | 1.2419774543 | 1241.9774543 | 1 |
| 21 | 0.9385030838 | 938.50308378 | 2 |
| 22 | 0.6699276249 | 669.92762489 | 2 |
| 23 | 0.204695956 | 204.695956 | 2 |

Square of random Normal scaled by 1,000 for readability of sample minimum

Sample 1

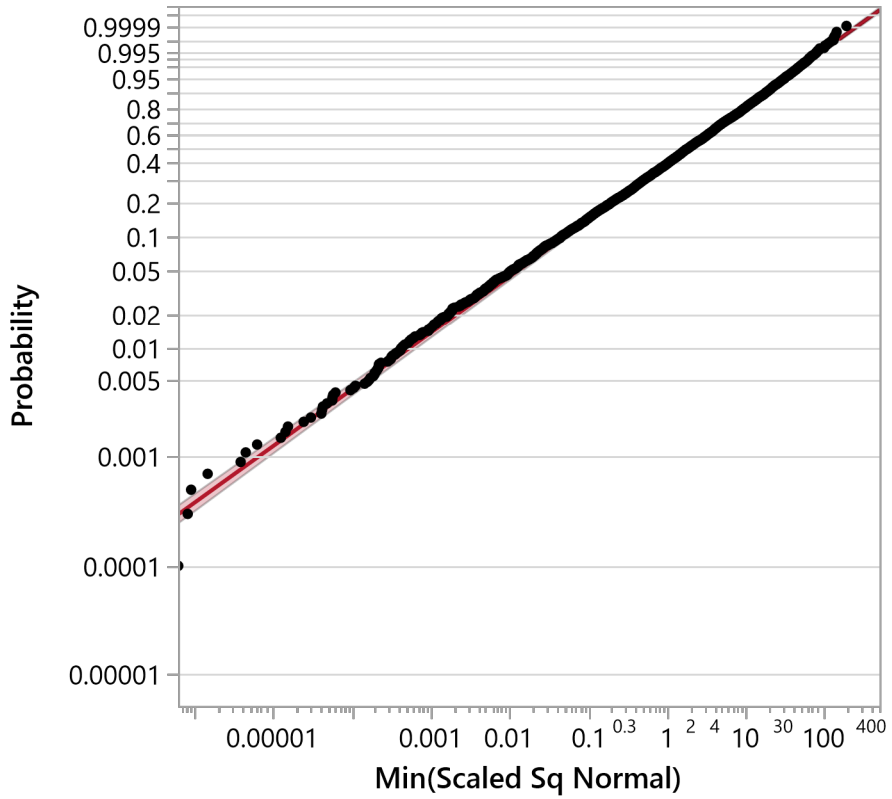
Min (Sample j) for j = 1 to 5000

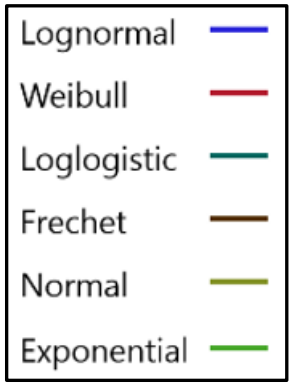


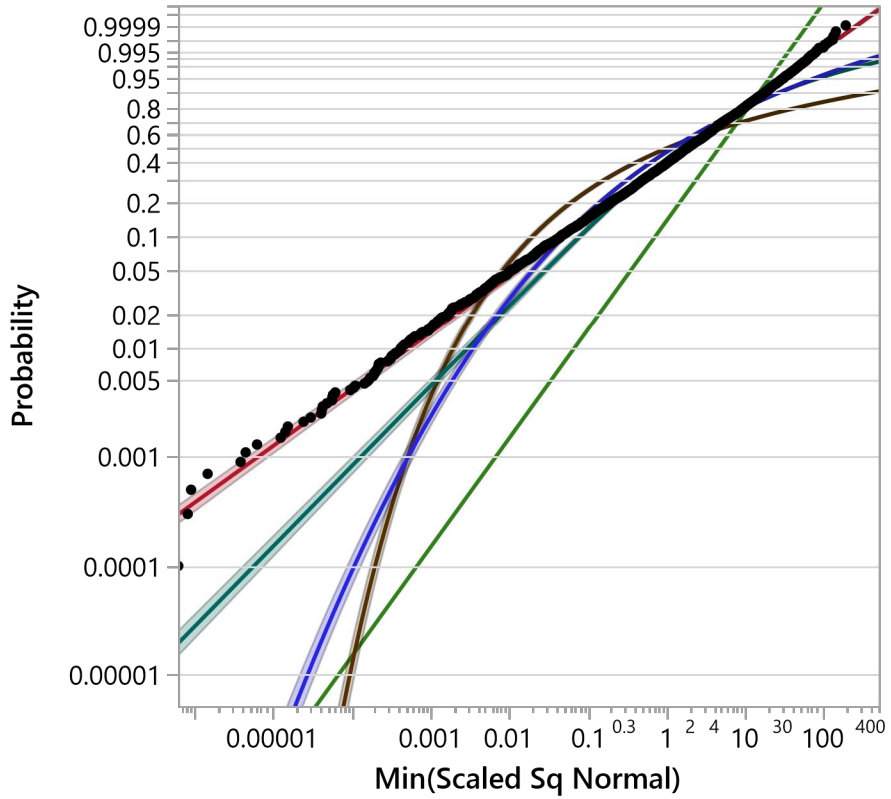



# Example – Fit Synthetic Weibull Data

Weibull Probability Axis




**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 |

Weibull clearly fits these data better than the other common survival distributions shown

# 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):

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

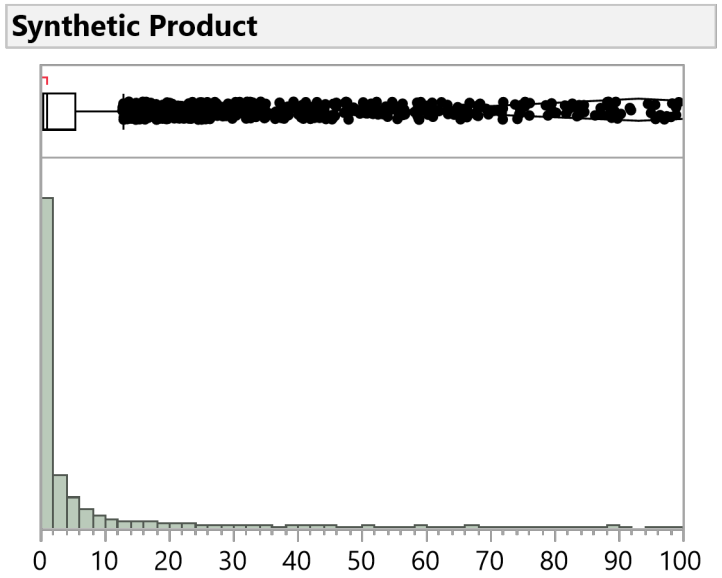
$$X_2 = \sin\phi \cdot Y_1 + \cos\phi \cdot Y_2$$

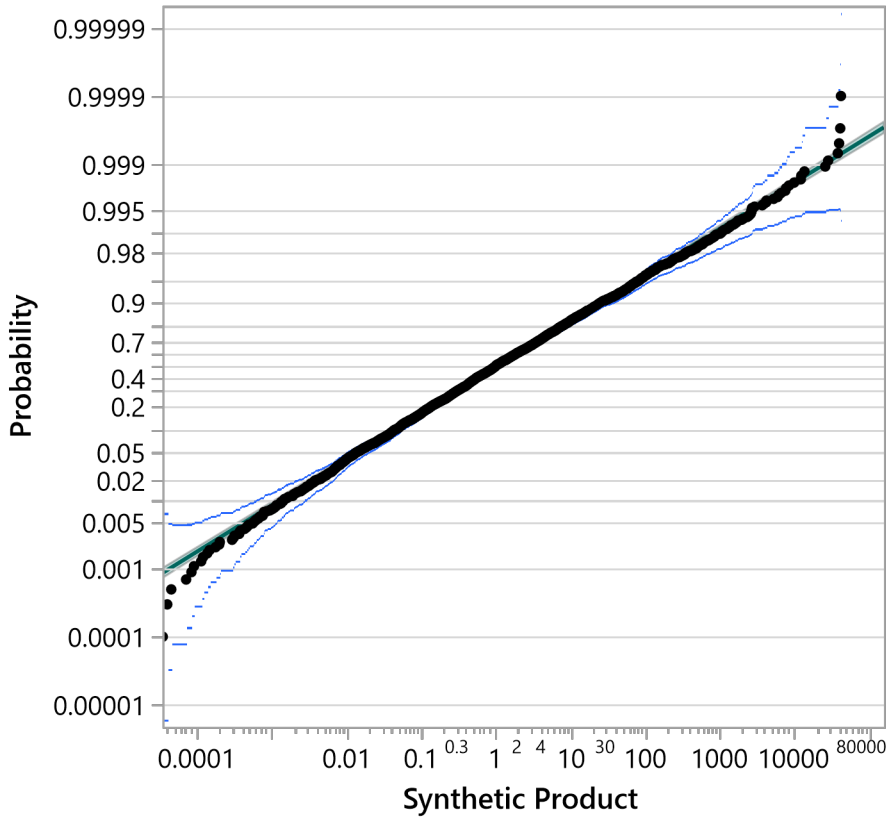
Where the value of $\phi$ necessary to produce the correlation coefficient is determined by:

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

# Demonstration – Generate and Fit Synthetic Loglogistic Data

Similar to generation of synthetic Lognormal data except that fluctuations are pre-processed using the correlation scheme described by Cordes so that resulting products are based on *correlated* fluctuations.




**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

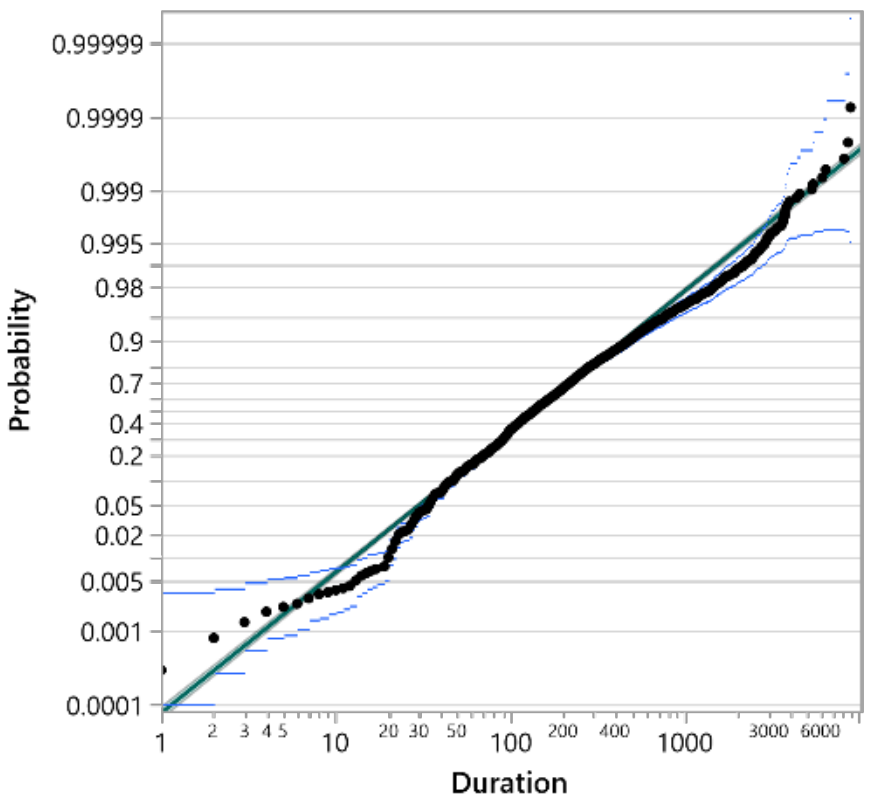
# The San Francisco Zoning Variance Process

| 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 |

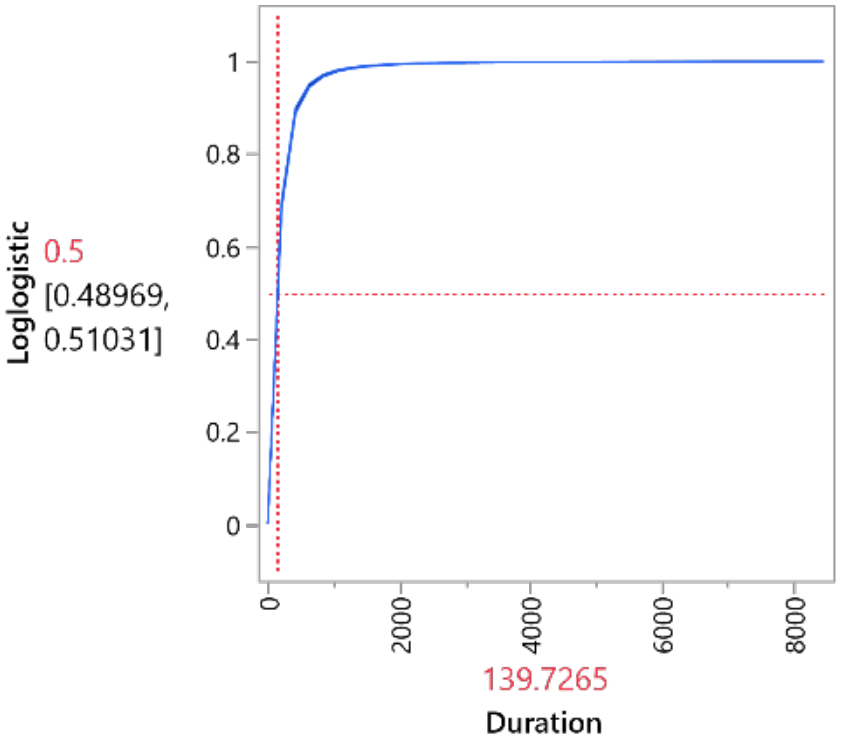
Process is characterized by substantial interaction among the Applicant, Assigned Planner, and Local Property Owners, converging to a single formal decision by the Zoning Administrator

# 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 | |




**Distribution Profiler**




**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

# Closing Remarks

# Heuristics for Distribution Selection and Interpretation of Empirical Results

| Distribution | Characteristic Behavior | 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 |

# Bibliography


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.

Babińska, M., Chudek, J., Chełmecka, E., Janik, M., Klimek, K., & Owczarek, A. (2015). Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome. 43(1), 33. doi:https://doi.org/10.1515/slgr-2015-0040

Billinton, R., & Allan, R. N. (1992). Reliability Evaluation of Engineering Systems: Springer.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building (MRC Technical Summary Report #1954). Retrieved from https://apps.dtic.mil/docs/citations/ADA070213

Brown, G., & Sanders, J. W. (1981). Lognormal Genesis. Journal of Applied Probability, 18(2), 542-547.

Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society B, 34(2), 187-220.

Department of City Planning. (2018). Zoning Handbook. Retrieved from www.nyc.gov/planning.

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. doi:10.1038/s41598-017-02474-w

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. doi:10.1007/s12350-014-9908-2

Hill, A. V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. The Journal of Physiology, 40(suppl), i-vii. doi:10.1113/jphysiol.1910.sp001386

K. Jayaraman. (1999). A Statistical Manual for Forestry Research. Retrieved from http://www.fao.org/3/X6831E/X6831E00.htm#TOC

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

Kleiber, C., & Kotz, S. (2003). Statistical Size Distributions in Economics and Actuarial Sciences: John Wiley & Sons, Inc.

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.

San Francisco Planning Department. (2012). Application Packet for Variance from the Planning Code. Retrieved from http://www.sfplanning.org.

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.

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. doi:10.7314/apjcp.2013.14.11.6751