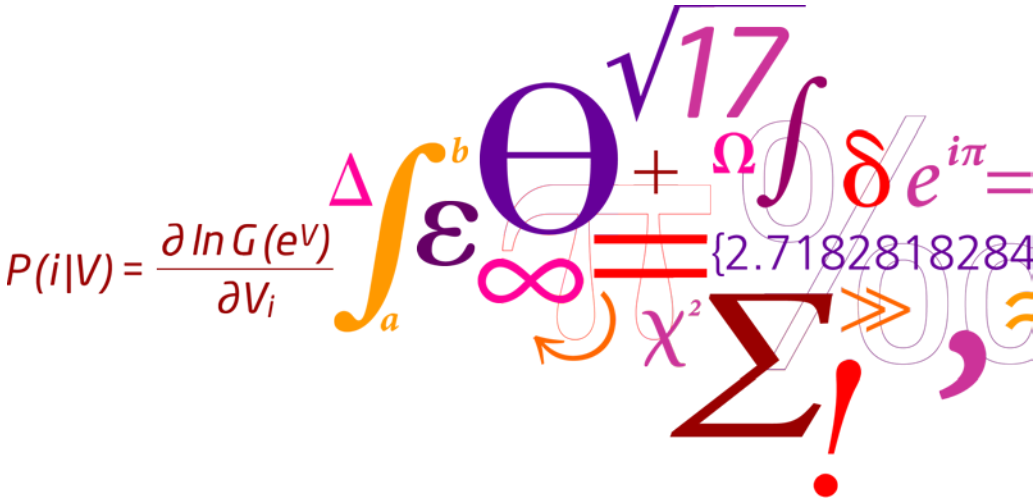
# Distribution Fitting for Very Large Railway Delay Data Sets with Discrete Values

Trafikdage 2018, Aalborg

*Steven Harrod[a], Georgios Pournaras[c], Bo Friis Nielsen[b]*

[a] *Department of Management Engineering*
*Technical University of Denmark*
[b] *Institut for Matematik og Computer Science, Technical University of Denmark*
[c] *Ansaldo STS - a Hitachi company group*




DTU Management Engineering

# Fitting Distributions to Data Who Cares?

- Primary delays – for simulation and modeling
- Aggregate delays
  - –for performance estimation and forecasting
  - –for validation of simulation models
- Both of management interest

To the absolute last session of Trafikdage!

# Velkommen!

- Review some previous literature
- Very large data sets fail tests
- Examine lognormal mixed distributions
- Demonstrate an alternative test
- Remarks and conclusions

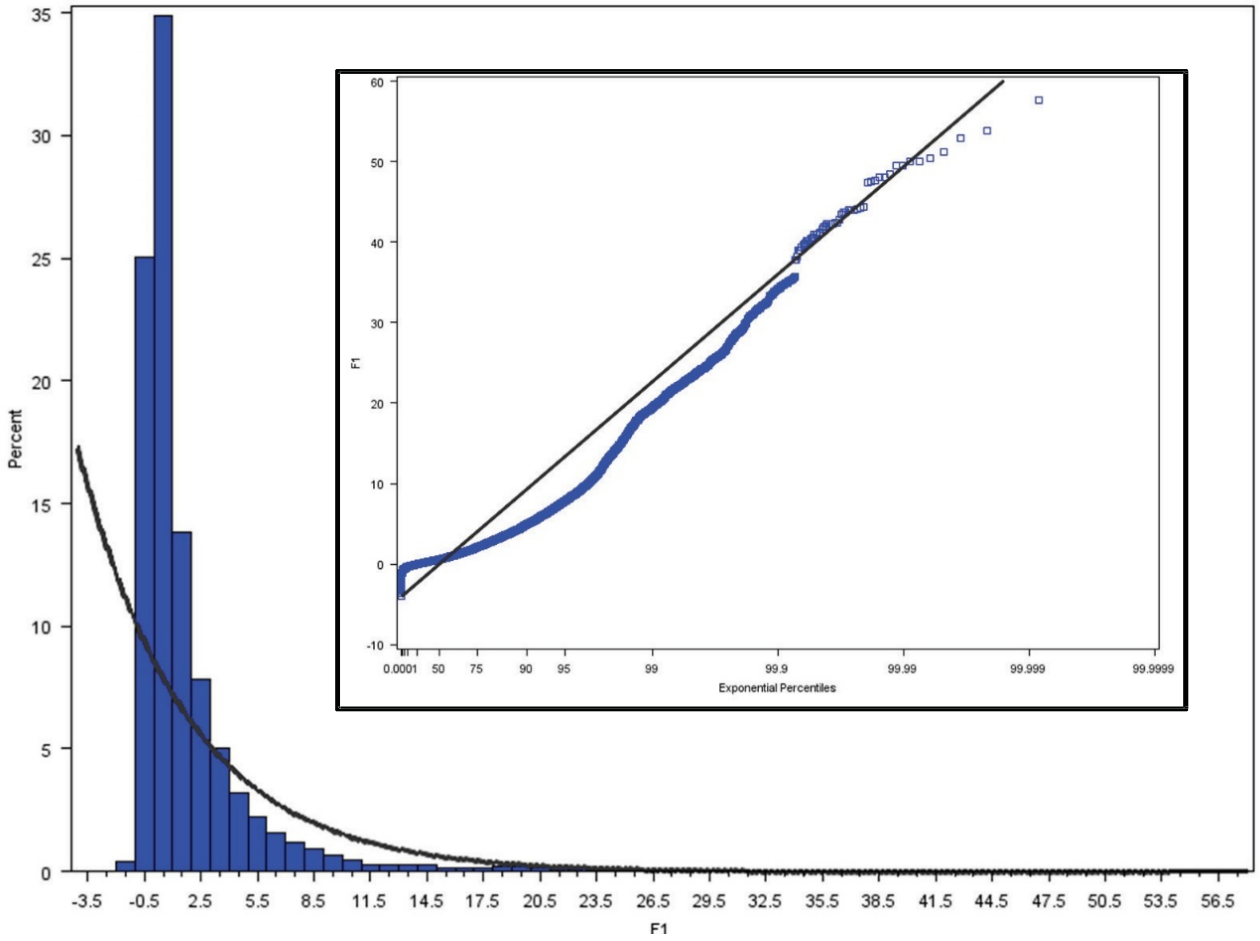
# Some Previous Literature

| Literature | Sample Size | Fitted Dist. | Fit Test |
|---|---|---|---|
| Goverde 2005 | n=103 | Exponential | K-S pass |
| Yuan and Hansen 2002 | n=416 | Exponential | K-S pass |
| Nie and Hansen 2005 | n=4320 | none | K-S fail |
| Yang et al. 2017 | n=11452 | Lognormal | K-S p=.06 |
| Wen et al. 2017 | n=1249 | Lognormal | |

# Big Data
# Challenges and Solutions

- Example Kystbane departure delays
  - –September-November 2014
  - –n=75.244!
  - –ØK, ØD, and ØP trains northbound
- No distribution satisfies fit tests
- Fits become acceptable when random sample approaches n=275

# Fit Attempt to n=75.244

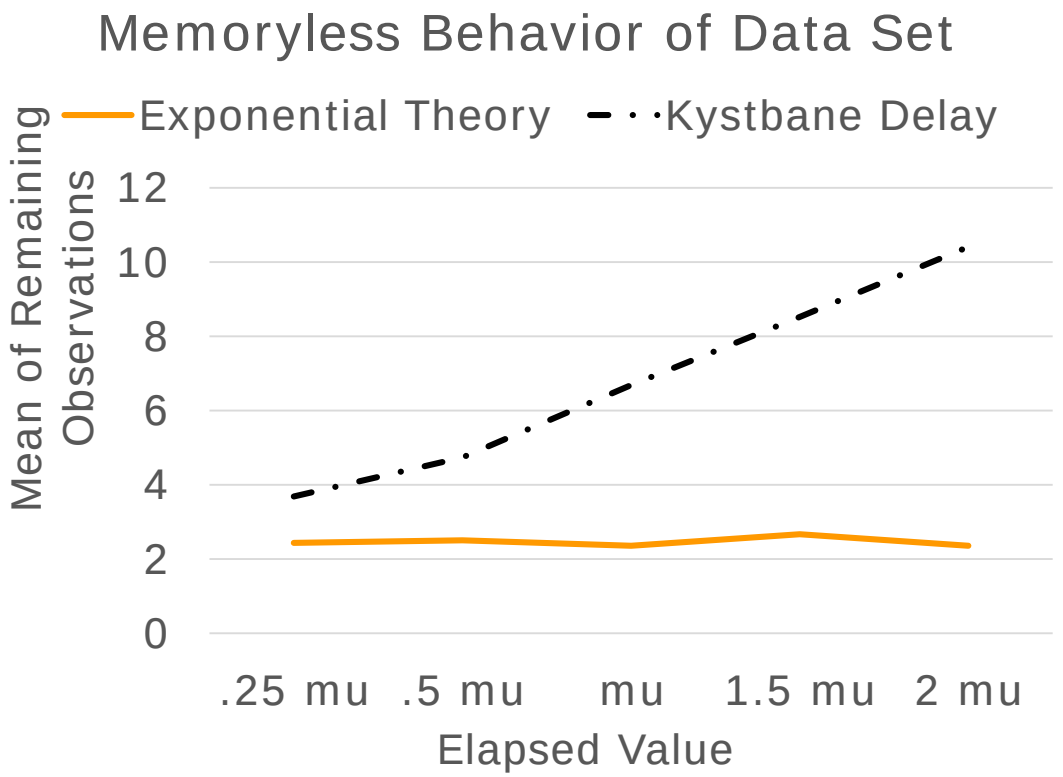


# Brown and Cudeck 1992

- “Statistical goodness-of-fit tests are often more a reflection on the size of the sample than on the adequacy of the model”

# Second Problem
# Process is Not Memoryless

- An exponential process should be memoryless
- Delay data is not



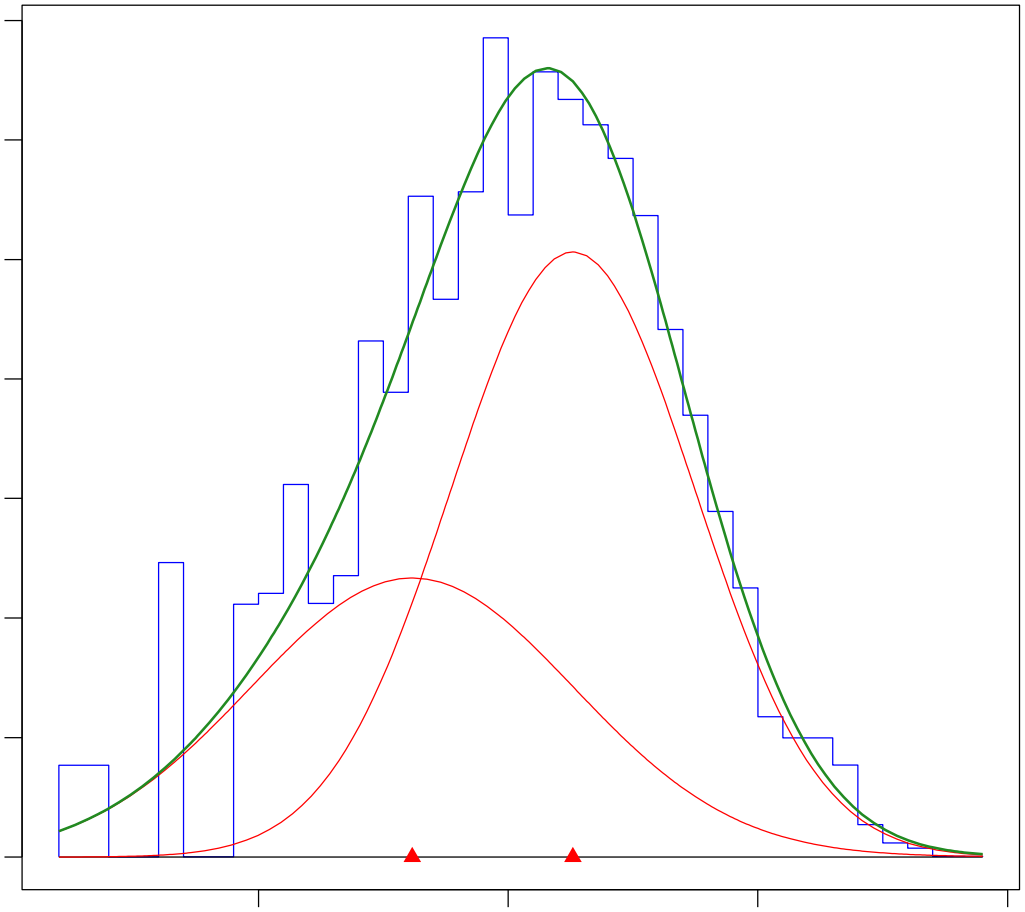
# Try a Mixed Distribution LogNormals

- Suggested from phone exchange research
- Transform data by log base 10
- Remove earliness data
- Fit a mixed distribution using mixdist package in R statistical software
  - *$p(x) = w1f1(x) + w2f2(x) + \ldots$

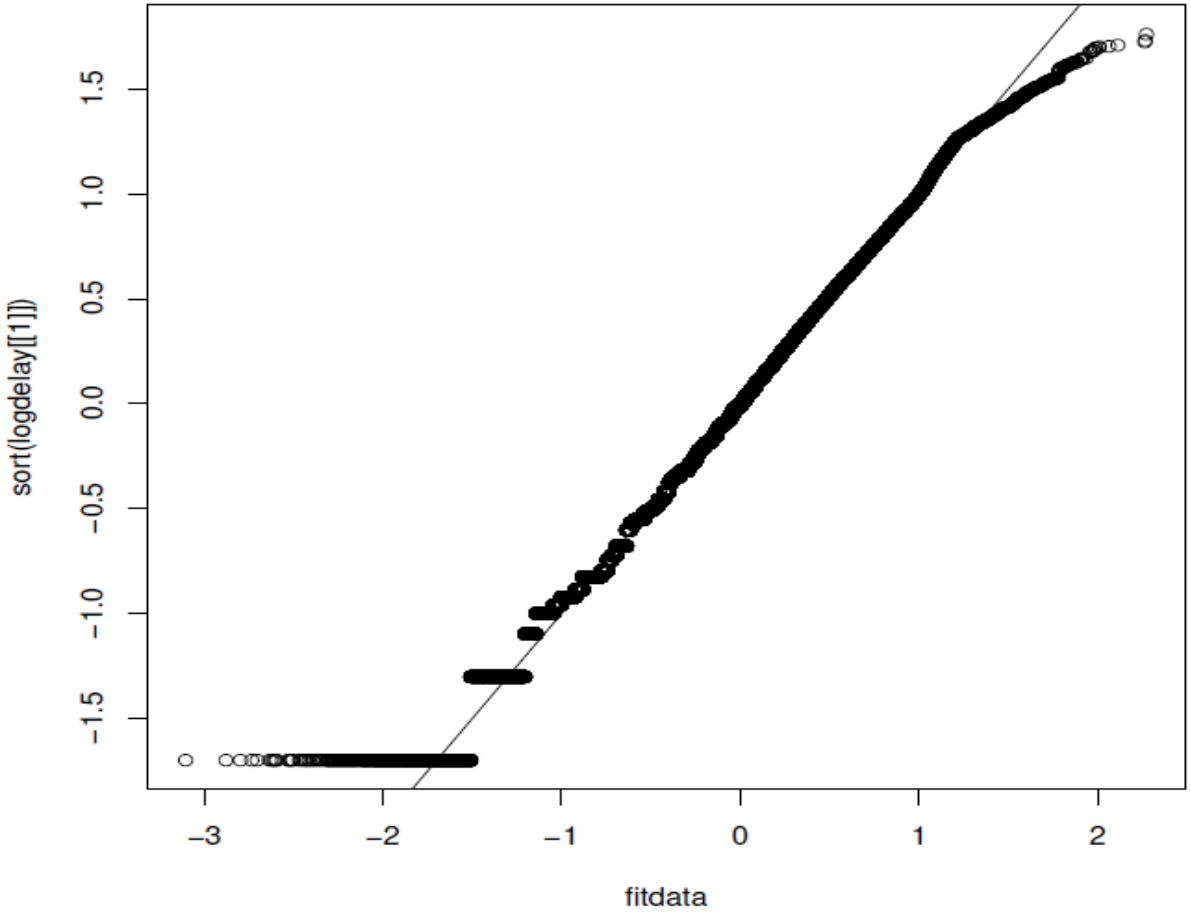
# First Result

- Distribution fit is robust to any statistical software parameters
- Two distribution components found
- $p(x) = 0.3792\ N(\mu = -0.3845, \sigma = 0.6478) + 0.6208\ N(0.2593, 0.4891)$.
- Probability plot shows a large deviation at lower tail

# Mixed Distribution Fit



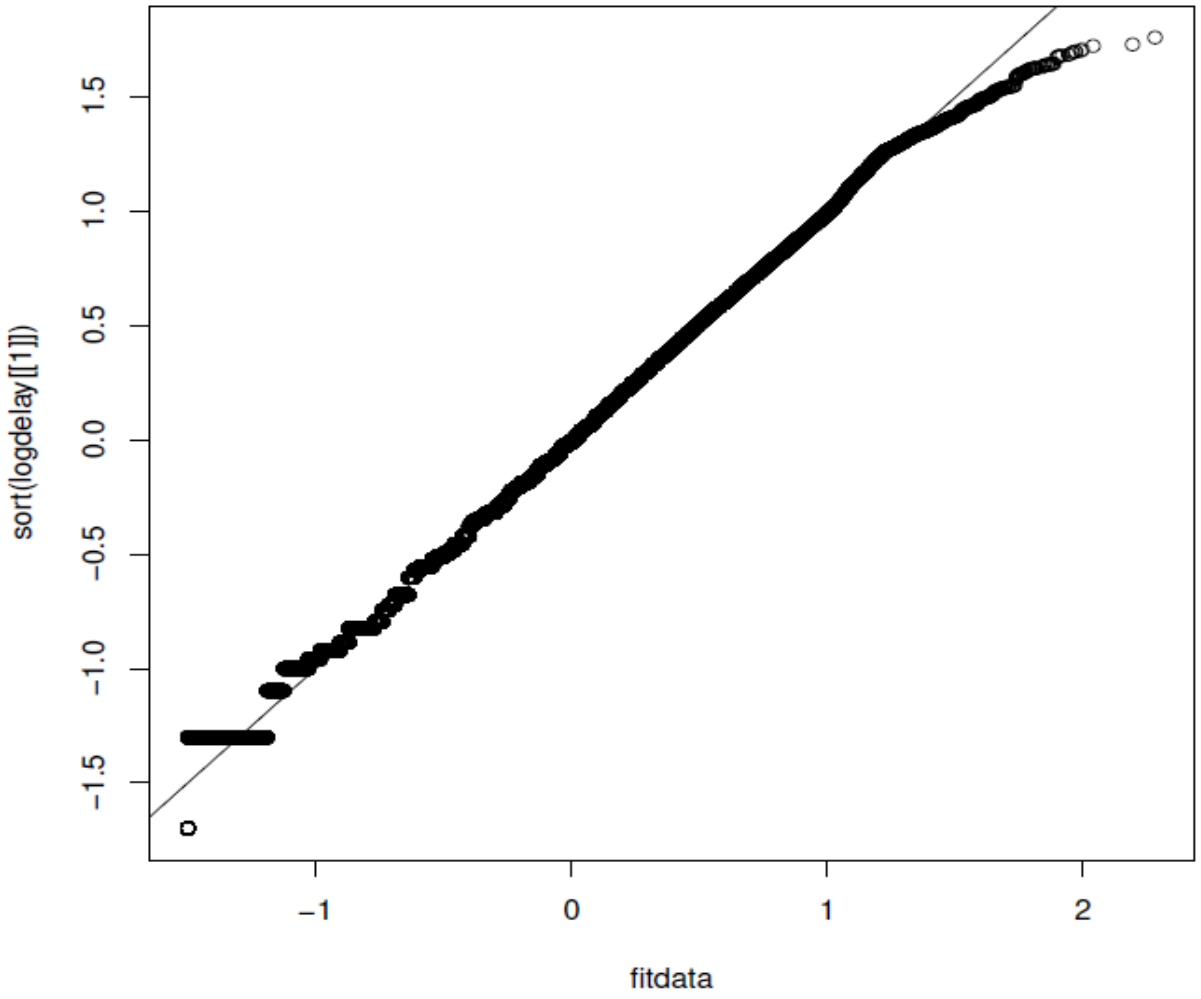
# Probability Plot of First Fit



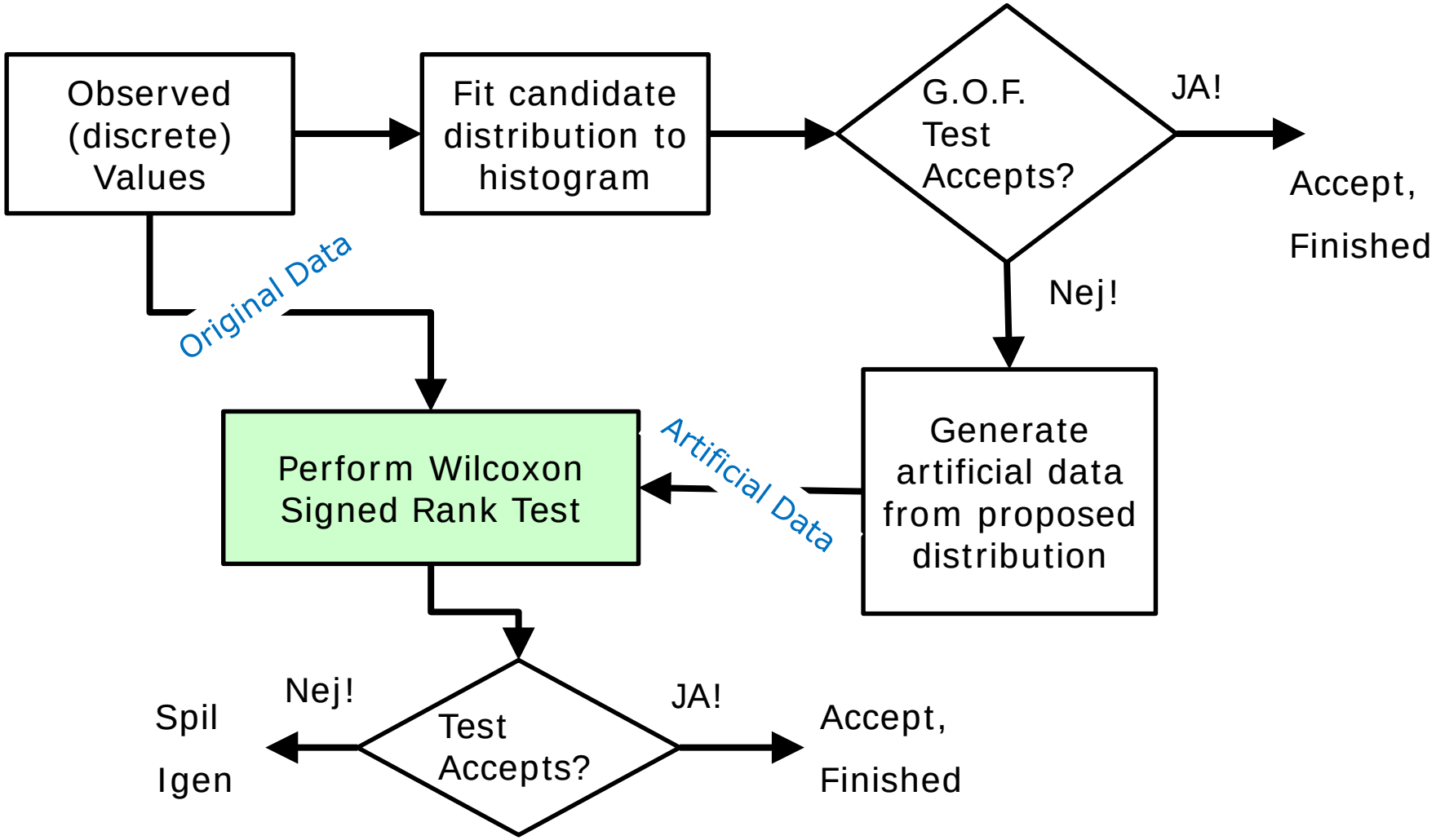
# Manually Adjust the Distribution

- Lower tail is distorted by signal systems discrete time measurement
- Large number of false “late” measures
- Bound distribution at a minimum delay
- $p(x) = 0.3792\ max(-1.5, N(\mu = -0.3845, \sigma = 0.6478)) + 0.6208\ max(-1.5, N(0.2593, 0.4891))$
- Still have a problem with standard goodness of fit tests failing

# Fit of Bounded Distribution



# Wilcoxon Fit Test



# Conclusion

- A single exponential distribution is not appropriate for this data
- A lognormal mixed distribution shows promise of a close fit for large data sets
- Some manual adjustment is still required due to the rounding and discrete output of the signal system.
- The Kystbane departure delays, transformed log base 10, fit mixed distribution
  $p(x) = 0.3792\ max(-1.5, N(\mu = -0.3845, \sigma = 0.6478)) + 0.6208\ max(-1.5, N(0.2593, 0.4891))$
- They satisfy a Wilcoxon fit test with p=0.678

# Tak!