Input Data Analysis: Specifying Model Parameters & Distributions

Christos Alexopoulos
David Goldsman
School of Industrial & Systems Engineering
Georgia Tech

Overview

- Deterministic vs. random inputs
- Data collection
- Distribution fitting
  - Model "guessing"
  - Fitting parametric distributions
    - Assessment of independence
    - Parameter estimation
    - Goodness-of-fit tests
- No data?
- Non-stationary arrival processes
- Multivariate / correlated input data
- Case study

Deterministic vs. Random Inputs

- *Deterministic*: Nonrandom, fixed values
  - Number of units of a resource
  - Entity transfer time (?)
  - Interarrival, processing times (?)
- *Random*: Model as a distribution, “draw” or “generate” values from to drive simulation
  - Interarrival, processing times
  - What distribution? What distributional parameters?
  - Causes simulation output to be random, too
- Don’t just assume randomness away!

Collecting Data

- Generally hard, expensive, frustrating, boring
  - System might not exist
  - Data available on the wrong things — might have to change model according to what's available
  - Incomplete, "dirty" data
  - Too much data (!)
- Sensitivity of outputs to uncertainty in inputs
- Match model detail to quality of data
- Cost — should be budgeted in project
- Capture variability in data — model validity
- Garbage In, Garbage Out (GIGO)

Using Data: Alternatives and Issues

- Use data "directly" in simulation
  - Read actual observed values to drive the model inputs (interarrivals, service times, part types, …)
  - All values will be "legal" and realistic
  - But can never go outside your observed data
  - May not have enough data for long or many runs
  - Computationally slow (reading disk files)
- Or, fit probability distribution to data
  - "Draw" or "generate" synthetic observations from this distribution to drive the model inputs
  - Can go beyond observed data (good and bad)
  - May not get a good "fit" to data — validity?

Fitting Distributions: Some Important Issues

- Not an exact science — no “right” answer
- Consider theoretical vs. empirical
- Consider range of distribution
  - Infinite both ways (e.g., normal)
  - Positive (e.g., exponential, gamma)
  - Bounded (e.g., beta, uniform)
- Consider ease of parameter manipulation to affect means, variances
- Simulation model sensitivity analysis
- Outliers, multimodal data
  - Maybe split data set

Main Steps (continued)

- Guess model using:
  - Summary statistics, such as
    - Sample mean $\overline{X}_n$
    - Sample variance S_n^2
    - Sample median
    - Sample coefficient of variation $S_n / \overline{X}_n$
    - Sample skewness

Estimates

$$\mathrm{CV}(X) = \sigma/\mu = \sqrt{\mathrm{Var}(X)}\,/\,\mathrm{E}(X)$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^3}{S_n^3}$$

Estimates

$$\mathrm{E}(X-\mu)^3 / \sigma^3$$

      - Skewness close to zero indicates a symmetric distribution
      - A skewed distribution with unit coefficient of variation is likely the exponential
  - Histograms (play with interval width to get a reasonably smooth histogram). They resemble the unknown density
  - Box plots

Main Steps (continued)

- If a parametric models seems plausible:
  - Estimate parameters
  - Test goodness-of-fit

Fitting Parametric Distributions

- Assume that the sample data are independent identically distributed data from some distribution with density (probability) function

$$X_1, X_2, \ldots, X_n \sim f(x;\theta)$$
$$\theta = (\theta_1, \ldots, \theta_m)$$

- All data are complete (no censoring)
- How can we test independence?
  - Using the scatter-plot of $(X_i, X_{i+1}),\ i = 1, \ldots, n-1$
  - By means of von-Neumann's test

Von Neumann's Test

The test statistic is

$$U_n = \sqrt{\frac{n^2-1}{n-2}} \times \left[\hat{\rho}_1 + \frac{(X_1-\bar{X}_n)^2 + (X_n-\bar{X}_n)^2}{2\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}\right]$$

where

$$\hat{\rho}_1 = \frac{\sum_{i=1}^{n-1}(X_i-\bar{X}_n)(X_{i+1}-\bar{X}_n)}{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}$$

estimates the correlation between adjacent observations.

If the data are independent and $n \geq 20$, $U_n \approx N(0,1)$

We reject the hypothesis of independence when

$$|U_n| > z_{\beta/2}$$

where β is the type-I error

Types of Parameters

- *Location* parameters — they shift the density function
- *Shape* parameters — they change the shape of the density function
- *Scale* parameters
- Example: For the $N(\mu,\sigma^2)$ distribution
  - μ is the location parameter because
    $X \sim N(\mu,\sigma^2) \Leftrightarrow X\text{-}\mu \sim N(0,\sigma^2)$
  - σ is the scale parameter because
    $X \sim N(\mu,\sigma^2) \Leftrightarrow X/\sigma \sim N(\mu,1)$
- Example: In the Weibull(α,λ) distribution
  - α is a shape parameter
  - λ is the scale parameter

Parameter Estimation Methods

- Method of moments
- Maximum likelihood estimation

Method of Moments

- Equate the first *m* sample (non-central) moments to the theoretical moments and solve the resulting system for the unknown parameters:

$$E(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ k = 1,\ldots,m$$

Method of Moments (continued)

- Example: The normal distribution

$$E(X) = \mu = \bar{X}_n$$

$$E(X^2) = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

give

$$\hat{\mu} = \bar{X}_n \quad \text{and} \quad \hat{\sigma} = S_n$$

Maximum Likelihood Estimation

- The likelihood function is the joint density (probability function) of the data:

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

- The Maximum Likelihood Estimator of θ maximizes $L(\theta)$ or, equivalently, the log-likelihood $\ln L(\theta)$:

$$\ln L(\hat{\theta}) \geq \ln L(\theta) \text{ for all } \theta$$

Maximum Likelihood Estimation (continued)

- Example: The exponential distribution

$$\ell(\lambda) \equiv \ln L(\lambda) = \ln\left(\prod_{i=1}^{n} \lambda e^{-\lambda X_i}\right) = n \ln \lambda - \lambda \sum_{i=1}^{n} X_i$$

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \Rightarrow \hat{\lambda} = 1/\bar{X}_n$$

Check that $d^2\ell / d\lambda^2 = -1/\lambda^2 < 0$;

this guarantees that $\hat{\lambda}$ is a maximizer

Maximum Likelihood Estimation (continued)

- Example: The normal distribution

$$\hat{\mu} = \bar{X}_n$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}S_n^2$$

Maximum Likelihood Estimation (continued)

- Example: The Uniform(0, b) distribution

  We wish to find the MLE of b

  The likelihood function is

$$L(b) = \begin{cases} 1/b^n & \text{for } 0 \le X_i \le b \Leftrightarrow b \ge \max X_i \\ 0 & \text{otherwise} \end{cases}$$

  Notice that $L(b)$ is discontinuous; so don't take derivatives…

  Check that $L(b)$ is maximized at

$$\hat{b} = \max X_i$$

Maximum Likelihood Estimation (continued)

Example: The Weibull distribution

The density is given by

$$f(x) = \alpha\lambda(\lambda x)^{\alpha-1}\exp[-(\lambda x)^{\alpha}],$$

where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter

The m.l.e.s satisfy the following equations:

$$\frac{\sum_{i=1}^{n} X_i^{\widehat{\alpha}} \ln X_i}{\sum_{i=1}^{n} X_i^{\widehat{\alpha}}} - \frac{1}{\widehat{\alpha}} = \frac{\sum_{i=1}^{n} \ln X_i}{n} \quad \text{and} \quad \widehat{\lambda} = \left(\frac{\sum_{i=1}^{n} X_i^{\widehat{\alpha}}}{n}\right)^{-1/\widehat{\alpha}}$$

We can solve the first equation by Newton's method

Maximum Likelihood Estimation (continued)

- MLEs are "nice" because they are
  - Asymptotically ($n \to \infty$) unbiased
  - Asymptotically normal
  - Invariant, i.e., if g is continuous,

$$\lambda = g(\theta) \Rightarrow \hat{\lambda} = g(\hat{\theta})$$

Example: The MLE of the variance ($\sigma^2 = 1/\lambda^2$) for the exponential distribution is $\bar{X}_n^{\,2}$

Testing Goodness-of-Fit

We want to test the null hypothesis

$H_0: \; X_1, \ldots, X_n$ are from $\hat{f}(x) = f(x; \hat{\theta})$

$\alpha =$ Type I Error $=$ Pr(reject $H_0 \mid H_0$ is true)

$\beta =$ Type II Error $=$ Pr(accept $H_0 \mid H_0$ is false)

Power $= 1 - \beta =$ Pr(reject $H_0 \mid H_0$ is false)

p-value $=$ smallest value of type I error that leads to rejection of H_0

Testing Goodness-of-Fit (continued)

- Graphical approaches
  - The Q-Q plot graphs the quantiles of the fitted distribution vs. the sample quantiles. It emphasizes poor fitting at the tails
  - The P-P plot graphs the fitted CDF vs. the empirical CDF

$$\bar{F}(x) = \frac{\text{number of } X_i \le x}{n}, -\infty < x < \infty$$

Computation: Sort $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Then

$$\bar{F}(X_{(i)}) = \frac{i}{n}$$

It emphasizes poor fitting at the middle of the fitted CDF

Testing Goodness-of-Fit (continued)

- Statistical Tests
  - The chi-square test
  - The Kolmogorov-Smirnov test
  - The Anderson-Darling test

The Chi-square Test

- Split the range of X into k adjacent intervals
- Let

$$I_i = [a_{i-1}, a_i) = \text{ ith interval}$$

$$O_i = \text{number of observations in interval } i$$

$$E_i = \text{ expected number of observations in interval } i$$

$$= n[\hat{F}(a_i) - \hat{F}(a_{i-1})]$$

CDF of fitted distribution

The Chi-square Test (continued)

- The null hypothesis is rejected (at level α) if

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-s-1,\alpha}$$

where s is the number of parameters replaced by their MLEs

  - One should use $E_i \geq 5$
  - The test has maximum power if the E_i are equal (the intervals are equiprobable)

The Kolmogorov-Smirnov Test

- It generally assumes that all parameters are known
- Sort the data and define the empirical CDF

$$\bar{F}(x) = \frac{\text{number of } X_i \leq x}{n}$$

$$= \begin{cases} 0 & \text{if } x < X_{(1)} \\ \dfrac{i}{n} & \text{if } X_{(i)} \leq x < X_{(i+1)},\ 1 \leq i \leq n-1 \\ 1 & \text{if } x > X_{(n)} \end{cases}$$

The Kolmogorov-Smirnov Test (continued)

- The null hypothesis is rejected (at level α) if

$$D_n = \sup\left|\hat{F}(x) - \bar{F}(x)\right|$$

$$= \max\left\{\max\left[\frac{i}{n} - \hat{F}(X_{(i)})\right], \max\left[\hat{F}(X_{(i)}) - \frac{i-1}{n}\right]\right\} > \underbrace{d_{n,\alpha}}_{\text{tabulated}}$$

The Kolmogorov-Smirnov Test (continued)

- We usually simplify the above inequality by computing a modified test statistic and a modified critical value c_α:

$$\text{Adjusted Test Statistic} > \underbrace{c_\alpha}_{\text{tabulated}}$$

- When parameters are replaced by MLEs modified K-S test statistics exist for the following distributions:
  - Normal
  - Exponential
  - Weibull
  - Log-logistic

The Kolmogorov-Smirnov Test (continued)

Modified Critical Values c_α for Adjusted K-S Statistics

Case	Adjusted Test Statistic	α 0.15	0.10	0.05	0.025	0.01
All parameters known	$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n$	1.138	1.224	1.358	1.480	1.628
Nor$(\bar{X}_n, S_n^2)$	$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right) D_n$	0.775	0.819	0.895	0.995	1.035
Expo$(1/\bar{X}_n)$	$\left(D_n - \frac{0.2}{\sqrt{n}}\right)\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right)$	0.926	0.990	1.094	1.190	1.308

The Kolmogorov-Smirnov Test (continued)

Modified Critical Values for the K-S Test for the Weibull Distribution

	α			
n	0.10	0.05	0.025	0.01
10	0.760	0.819	0.880	0.944
20	0.779	0.843	0.907	0.973
50	0.790	0.856	0.922	0.988
∞	0.803	0.874	0.939	1.007

The Kolmogorov-Smirnov Test (continued)

Modified Critical Values for the K-S Test for the Log-logistic Distribution

	α			
n	0.10	0.05	0.025	0.01
10	0.679	0.730	0.774	0.823
20	0.698	0.755	0.800	0.854
50	0.708	0.770	0.817	0.873
∞	0.715	0.780	0.827	0.886

The Anderson-Darling Test

- The null hypothesis is rejected (at level α) if

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[\hat{F}(x) - \overline{F}(x)]^2}{\hat{F}(x)[1 - \hat{F}(x)]} \hat{f}(x)\, dx$$

$$= -\frac{1}{n} \sum_{i=1}^{n} (2i-1) \left\{ \ln \hat{F}(X_{(i)}) + \ln\left[1 - \hat{F}(X_{(n-i+1)})\right] \right\} - n > \underbrace{a_{n,1-\alpha}}_{\text{tabulated}}$$

- It generally assumes that all parameters are known

The Anderson-Darling Test (continued)

- We usually simplify the above inequality by computing a modified test statistic and a modified critical value a_α:

$$\text{Adjusted Test Statistic} > \underbrace{a_\alpha}_{\text{tabulated}}$$

- When parameters are replaced by MLEs, modified A-D test statistics exist for:
  - The normal distribution
  - The exponential distribution
  - The Weibull distribution
  - The log-logistic distribution

The Anderson-Darling Test (continued)

Modified Critical Values a_α for Adjusted A-D Statistics

		α			
Case	Adjusted Test Statistic	0.10	0.05	0.025	0.01
All parameters known	A_n^2 for $n \geq 5$	1.933	2.492	3.070	3.857
Nor$(\bar{X}_n, S_n^2)$	$\left(1 + \frac{4}{n} - \frac{25}{n^2}\right) A_n^2$	0.632	0.751	0.870	1.029
Expo$(1/\bar{X}_n)$	$\left(1 + \frac{0.6}{n}\right) A_n^2$	1.070	1.326	1.587	1.943
Weibull$(\widehat{\alpha}, \widehat{\beta})$	$\left(1 + \frac{0.2}{\sqrt{n}}\right) A_n^2$	0.637	0.757	0.877	1.038
Log-logistic$(\widehat{\alpha}, \widehat{\beta})$	$\left(1 + \frac{0.25}{\sqrt{n}}\right) A_n^2$	0.563	0.660	0.769	0.906

No Data?

- Happens more often than you would like
- No good solution; some (bad) options:
  - Interview “experts”
    - Min, Max: Uniform
    - Average, % error or absolute error: Uniform
    - Min, Mode, Max: Triangular
      - Mode can be different from Mean — allows asymmetry (skewness)
  - Interarrivals — independent, stationary
    - Exponential — still need some value for mean
  - Number of “random” events in an interval: Poisson
  - Sum of independent “pieces”: normal
  - Product of independent “pieces”: lognormal

Non-stationary Arrival Processes

- External events (often arrivals) whose rate varies over time
  - Lunchtime at fast-food restaurants
  - Rush-hour traffic in cities
  - Telephone call centers
  - Seasonal demands for a manufactured product
- It can be critical to model this non-stationarity for model validity
  - Ignoring peaks, valleys can mask important behavior
  - Can miss rush hours, etc.
- Good model: *Non-stationary Poisson process*

Non-stationary Arrival Processes (continued)

- Two issues:
  - How to specify/estimate the *rate function*
  - How to generate from it properly during the simulation (will be discussed during the Output Analysis session)
- Several ways to estimate rate function — we'll just do the *piecewise-constant* method
  - Divide time frame of simulation into subintervals of time over which you think rate is fairly flat
  - Compute observed rate within each subinterval
  - Be *very* careful about time units!
    - Model time units = minutes
    - Subintervals = half hour (= 30 minutes)
    - 45 arrivals in the half hour; rate = 45/30 = 1.5 *per minute*

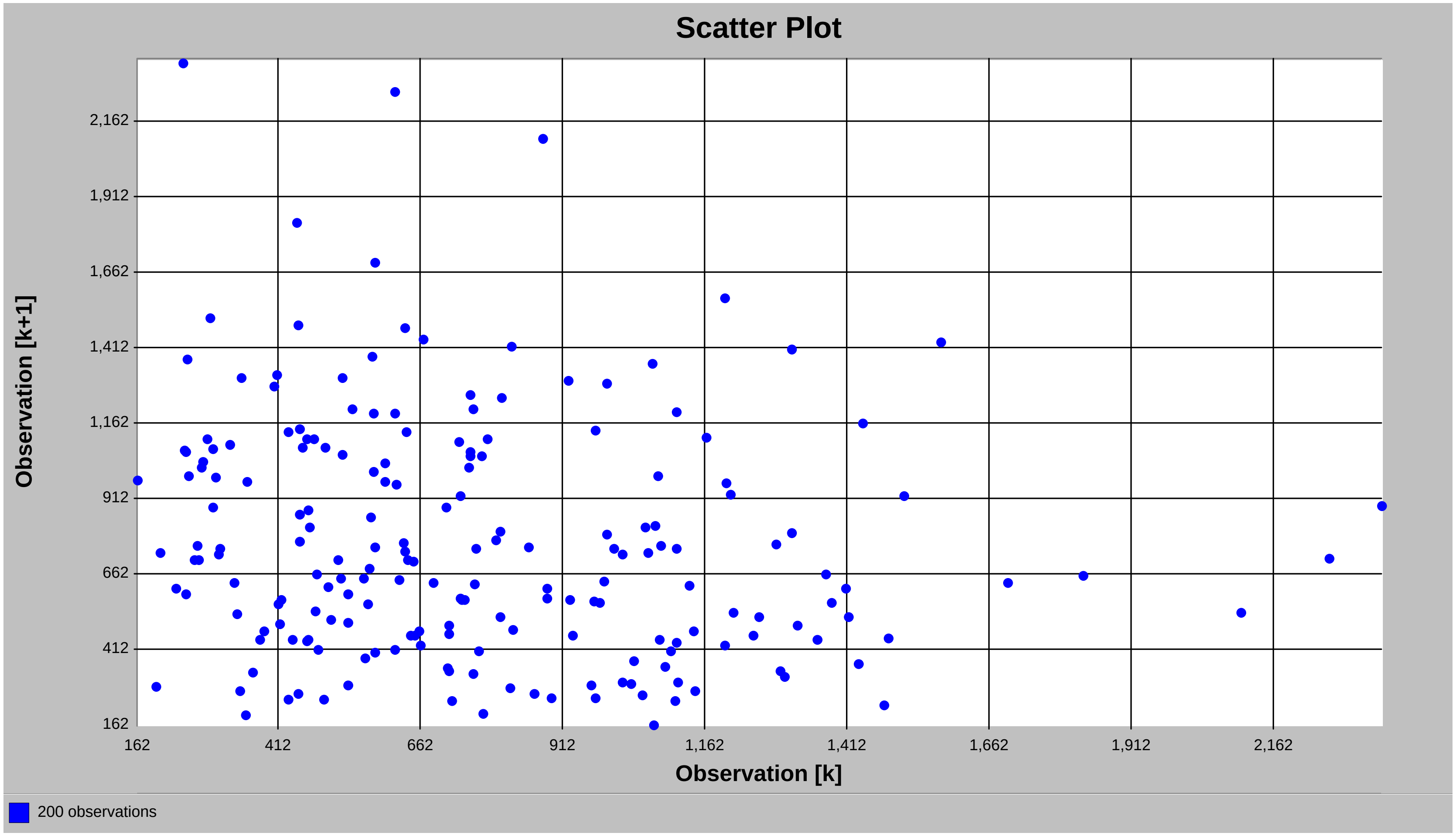
Multivariate and Correlated Input Data

- Usually we assume that all generated random observations across a simulation are independent (though from possibly different distributions)
- Sometimes this isn't true:
  - A "difficult" part may require longer service times by a set of machines
  - This indicates positive correlation
- Ignoring such relations can invalidate model

Case Study: Times-to-Failure

- A data set contains 200 times-to-failure for a piece of equipment
- We use ExpertFit®
- To assess independence, we create a scatter plot

Case Study — Scatter Plot



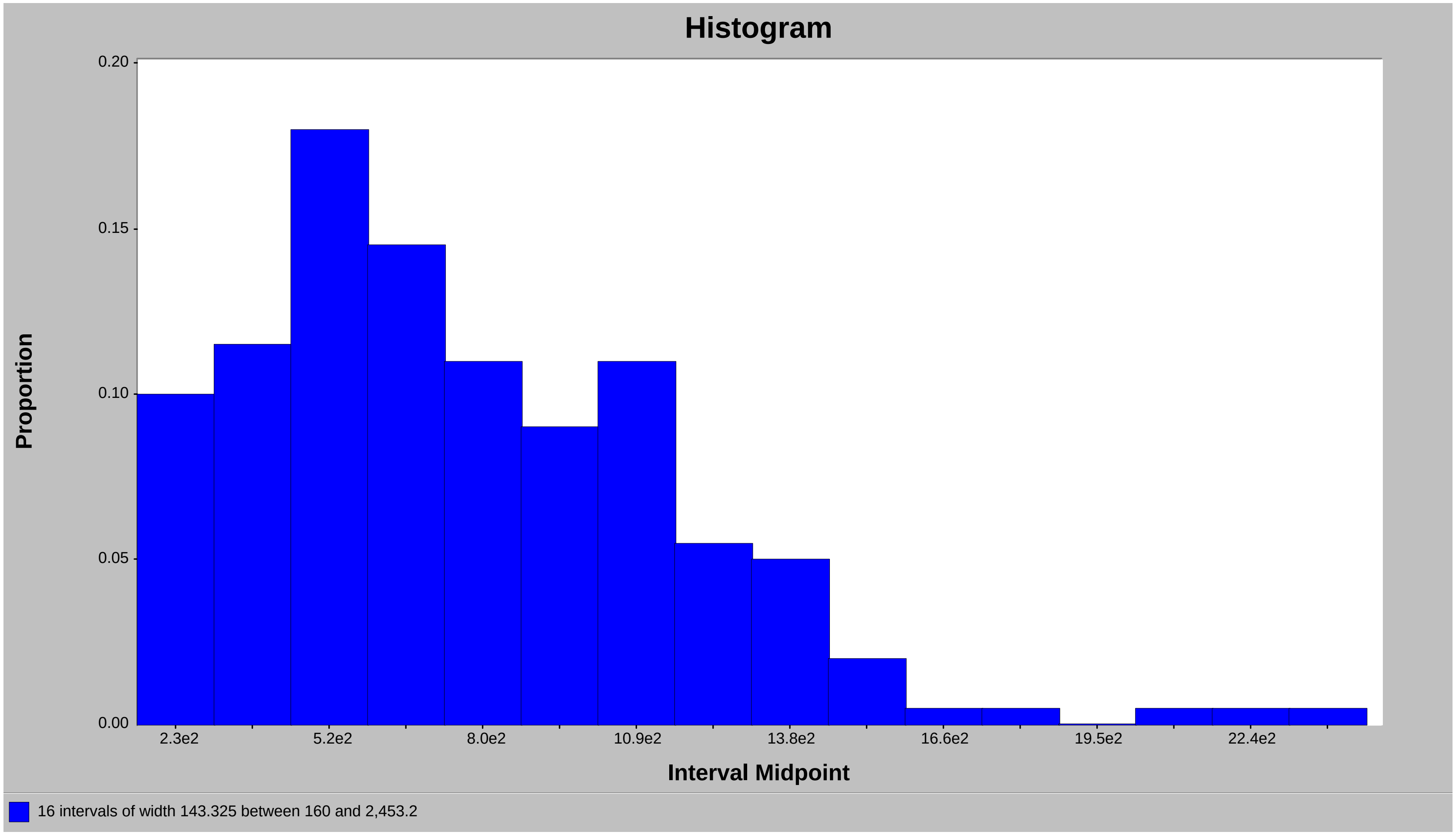
The data appear to be independent

Case Study — Data Summary

Data Characteristic	Value
Source file	TTF.DAT
Observation type	Real valued
Number of observations	200
Minimum observation	162.26205
Maximum observation	2,351.98858
Mean	768.91946
Median	709.90162
Variance	157,424.22579
Coefficient of variation	0.51601
Skewness	1.02670

- Can the data be from
  - The normal distribution?
  - The exponential distribution?

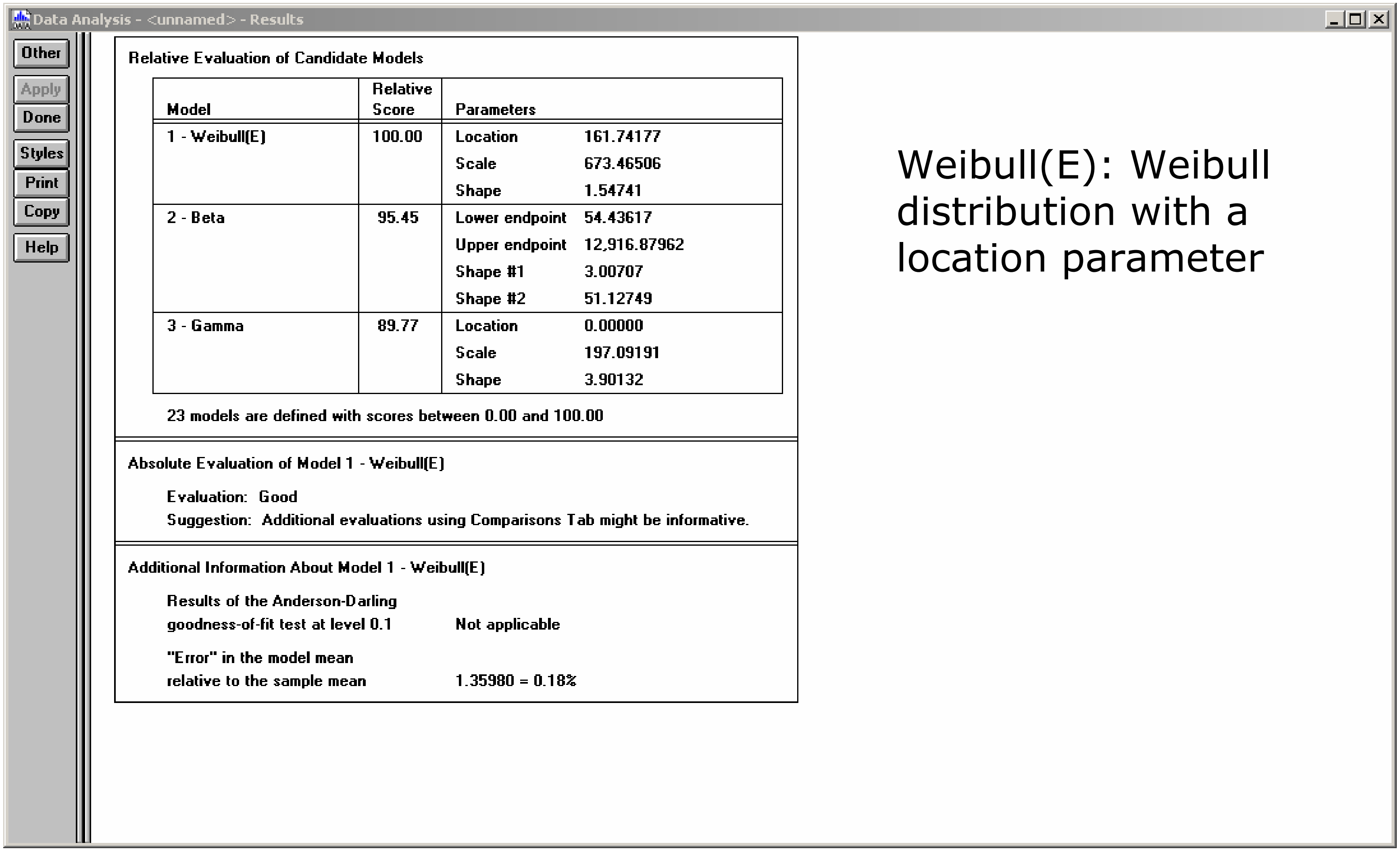
Case Study — Histogram with 16 Intervals



Case Study — Model Guessing

- We will allow ExpertFit to choose a continuous distribution automatically
- We will tell it that
  - the left limit for the underlying random variable is zero and
  - the tight limit is infinity

Case Study — ExpertFit's Choice…



Relative Evaluation of Candidate Models

Model	Relative Score	Parameters	
1 - Weibull(E)	100.00	Location	161.74177
		Scale	673.46506
		Shape	1.54741
2 - Beta	95.45	Lower endpoint	54.43617
		Upper endpoint	12,916.87962
		Shape #1	3.00707
		Shape #2	51.12749
3 - Gamma	89.77	Location	0.00000
		Scale	197.09191
		Shape	3.90132

23 models are defined with scores between 0.00 and 100.00

Absolute Evaluation of Model 1 - Weibull(E)

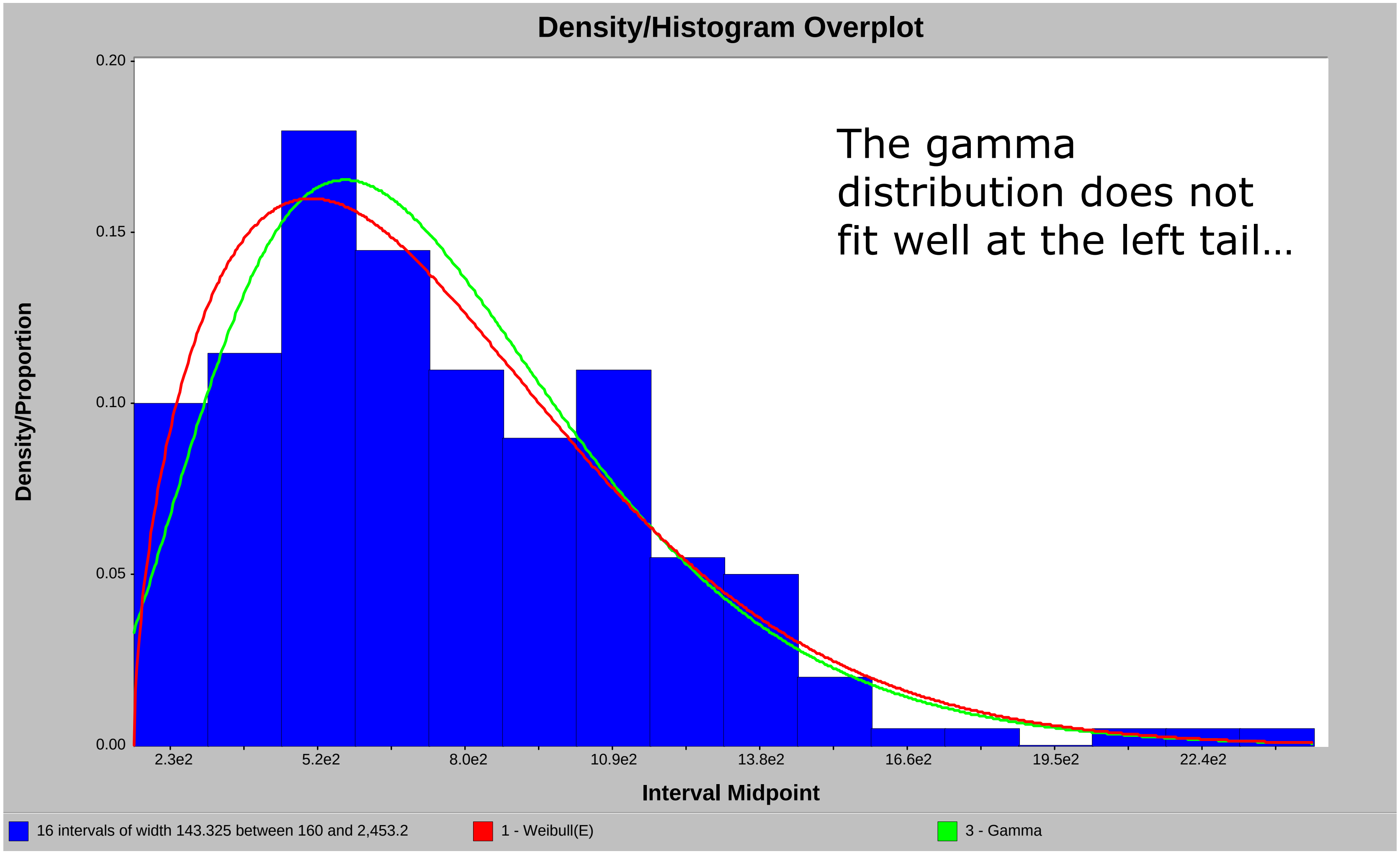
Evaluation: Good
Suggestion: Additional evaluations using Comparisons Tab might be informative.

Additional Information About Model 1 - Weibull(E)

Results of the Anderson-Darling goodness-of-fit test at level 0.1	Not applicable
"Error" in the model mean relative to the sample mean	1.35980 = 0.18%

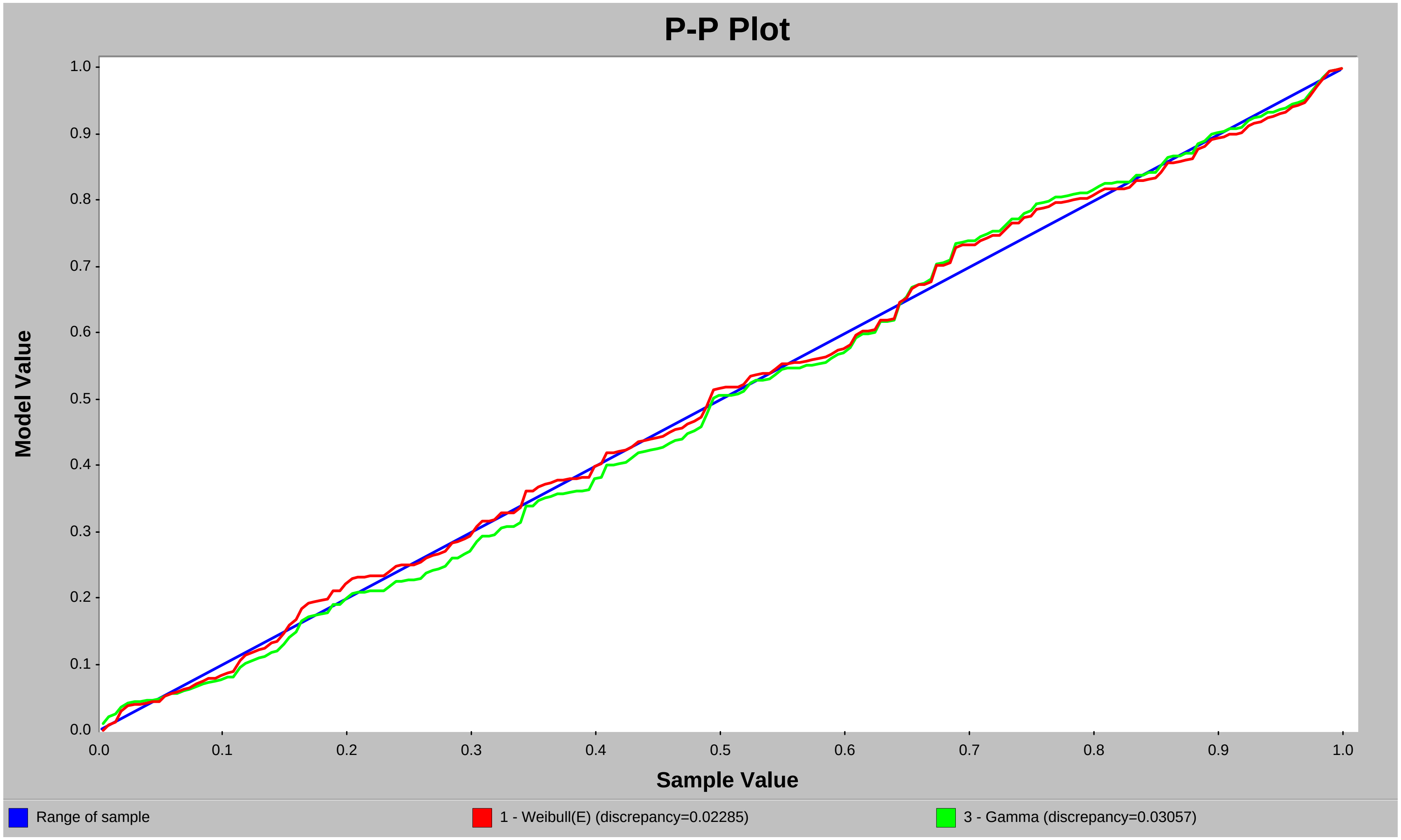
Weibull(E): Weibull distribution with a location parameter

Case Study — Histogram Comparisons

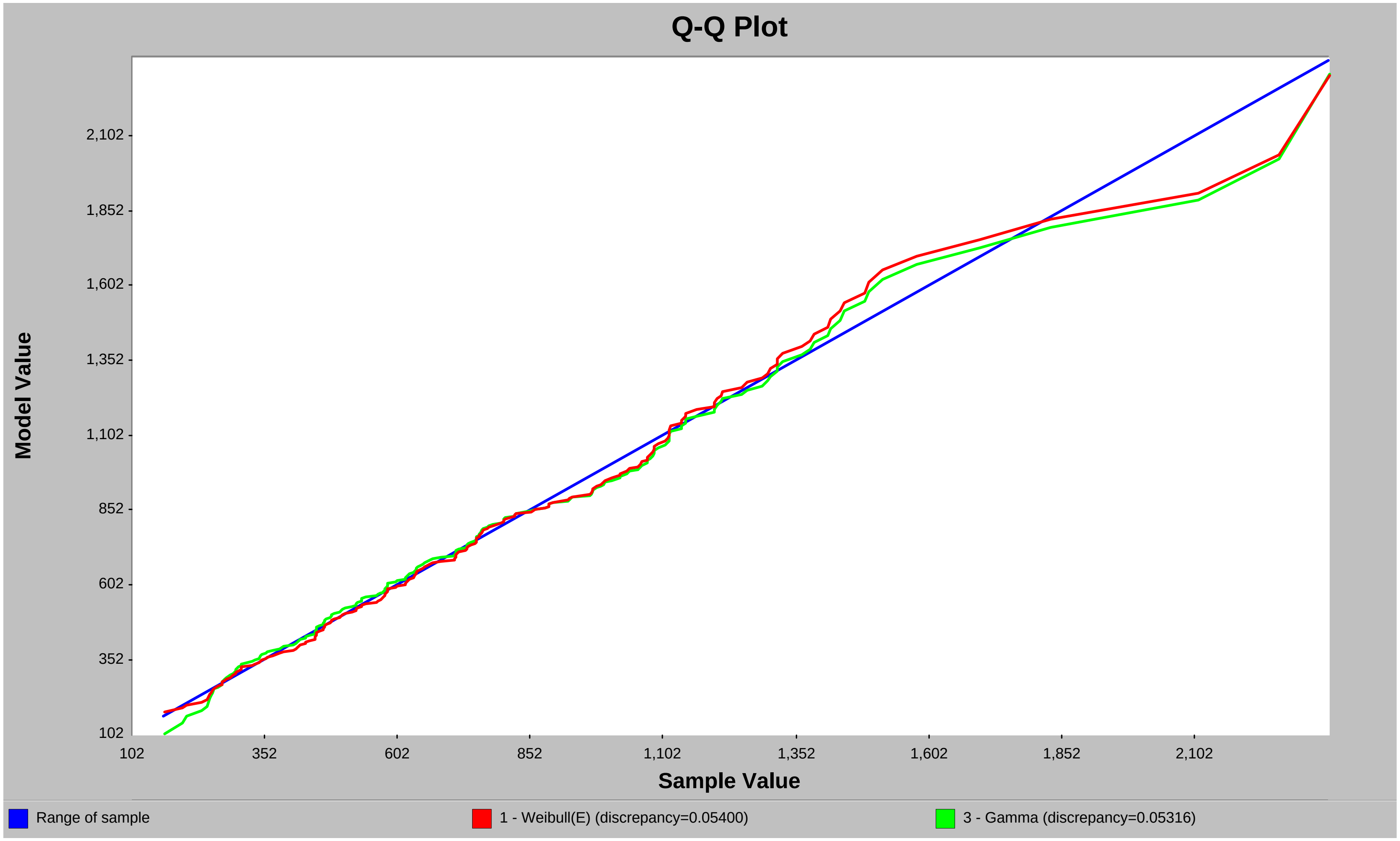


The gamma distribution does not fit well at the left tail…

Case Study — Graphical Goodness-of-Fit Tests



Case Study — Graphical Goodness-of-Fit Tests
(continued)



Case Study — A-D & K-S Goodness-of-Fit Tests

Anderson-Darling Test With Model 1 - Weibull(E)

Sample size 200
Test statistic 0.33184

Note: No critical values exist for this special case.
The following critical values are for the case where
all parameters are known, and are conservative.

Sample Size	Critical Values for Level of Significance (alpha) 0.250	0.100	0.050	0.025	0.010	0.005
200	1.248	1.933	2.492	3.070	3.857	4.500
Reject?	No					

Kolmogorov-Smirnov Test With Model 1 - Weibull(E)

Sample size 200
Normal test statistic 0.04426
Modified test statistic 0.62593

Note: No critical values exist for this special case.
The following critical values are for the case where
all parameters are known, and are conservative.

Sample Size	Critical Values for Level of Significance (alpha) 0.150	0.100	0.050	0.025	0.010
200	1.128	1.213	1.346	1.467	1.613
Reject?	No				

Anderson-Darling Test With Model 3 - Gamma

Sample size 200
Test statistic 0.48640

Note: The following critical values are approximate.

Sample Size	Critical Values for Level of Significance (alpha) 0.250	0.100	0.050	0.025	0.010	0.005
200	0.474	0.638	0.761	0.884	1.047	1.176
Reject?	Yes	No				

Kolmogorov-Smirnov Test With Model 3 - Gamma

Sample size 200
Normal test statistic 0.04957
Modified test statistic 0.70106

Note: No critical values exist for this special case.
The following critical values are for the case where
all parameters are known, and are conservative.

Sample Size	Critical Values for Level of Significance (alpha) 0.150	0.100	0.050	0.025	0.010
200	1.128	1.213	1.346	1.467	1.613
Reject?	No				

Case Study — Chi-square Goodness-of-Fit Tests

Equal-Probable Chi-Square Test With Model 1 - Weibull(E)

Number of intervals	20
Expected (model) count	10
Test statistic	14.6

Warning: The test may not be statistically valid because a method other than maximum likelihood was used to estimate parameters.

Degrees of Freedom	Observed Level of Significance	Critical Values for Level of Significance (alpha)				
		0.25	0.15	0.10	0.05	0.01
16	0.554	19.369	21.793	23.542	26.296	32.000
19	0.748	22.718	25.329	27.204	30.144	36.191
	Reject?	No				

Beware:

Outcomes depend on the number of intervals!

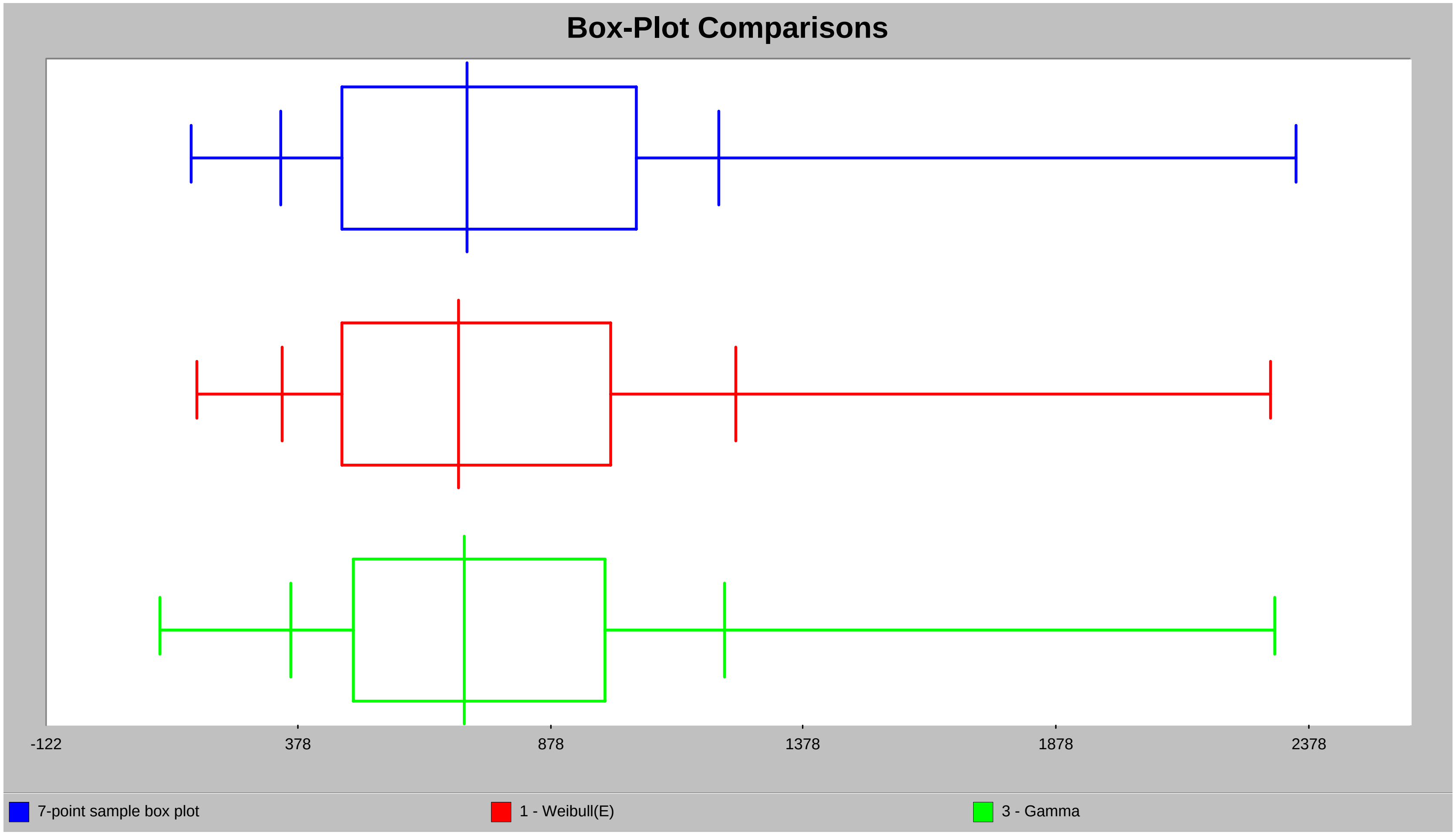
What distribution gives a better fit?

Equal-Probable Chi-Square Test With Model 3 - Gamma

Number of intervals	20
Expected (model) count	10
Test statistic	28

Degrees of Freedom	Observed Level of Significance	Critical Values for Level of Significance (alpha)				
		0.25	0.15	0.10	0.05	0.01
17	0.045	20.489	22.977	24.769	27.587	33.409
19	0.083	22.718	25.329	27.204	30.144	36.191
	Reject?	Yes			No	

Case Study — Additional Graphical Comparisons



Case Study — Arena Code for the Winner…

Arena Representation of Model 1 - Weibull(E)

Use:

161.741769 + WEIB(673.465060, 1.547408, <stream>)

We haven't used this yet…

Estimate for location parameter. Check the translation…

Estimate for shape parameter

Estimate for scale^{-1} parameter