How To Cross A River

creating input distributions

Poll: have you used Stat::Fit®

“All models are wrong, but some are useful.” George Box

- “Is the model illuminating and useful?”
- Building a model tells you what you don’t know
- Models are an iterative process
- Models require design, verification, validation and stochastic completeness
  - Input parameters properly represented (support, moments)
  - Replications provide sufficient confidence limits
  - Output stable to small input changes

Beware of Using Averages

- Some processes include ruin (adsorption)
  - The average depth of a river may be 3 feet, but the bottom?
  - Beware of a non-ergodic process where the ensemble average does not represent your true process over time.

- Some processes interact with randomness
  - Queueing processes

- Some processes have no empirical average (fat tails)
  - Pareto process (income)
  - Cauchy process (change of bitcoin value)

- If all you have is an average, STOP, do not pass go

What kind of data do you have?

- Continuous vs. Discrete
- Is it Bounded
  - Unbounded
  - Bounded with a lower minimum
  - Doubly bounded
- Independent (iid)
  - Identically distributed
- Most time series cannot be represented with an iid distribution

What information do you have?

- What is the physical model?
- Is the process unimodal (or sum of unimodal) distributions
- Do you have limits on the parameters?
  - Do you have a value at the limit?
- Do you have expert advice on most likely value?
- Are the data from the process independent?

Stat::Fit® Uses and Data Irregularities

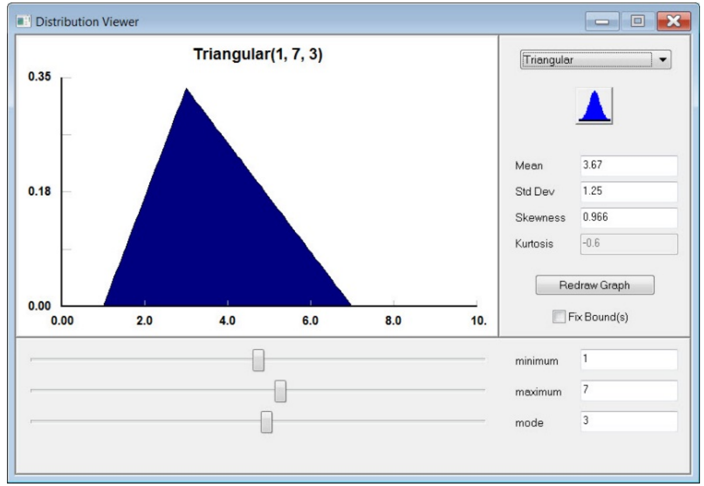
- Is the physical situation being represented?
- Are outliers distorting your process?
- Is more than one process buried in your data?
- Has your data been rounded or pinned?
- Do you have any data?
  - No Data Representations

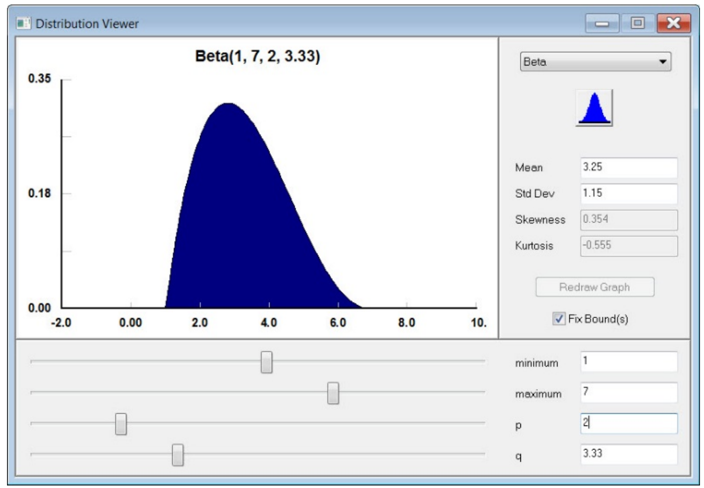
Doubly Bounded

- Skewed
  - Triangular
  - Beta
  - Johnson SB
  - Uniform

If minimum, maximum, and most likely value (mode) can be guessed, a Triangular can be used. For more control over the variance, can use Beta and still get range

For Beta, use mean or
mode = min + (p-1)/(p+q-1)
p,q > 1
increasing p,q to decrease standard deviation



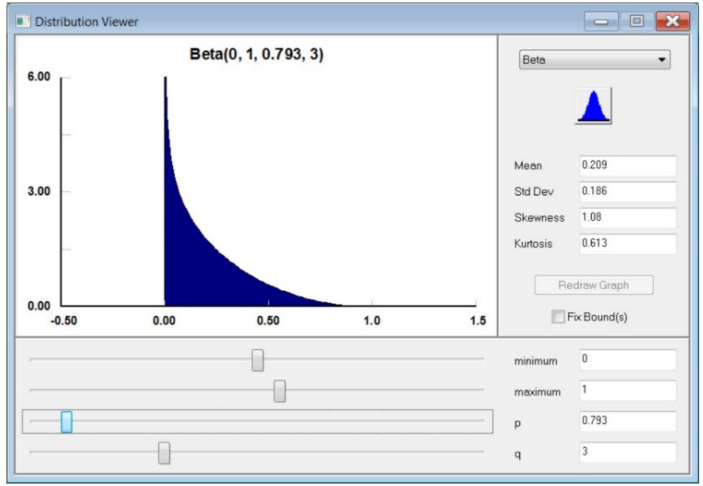


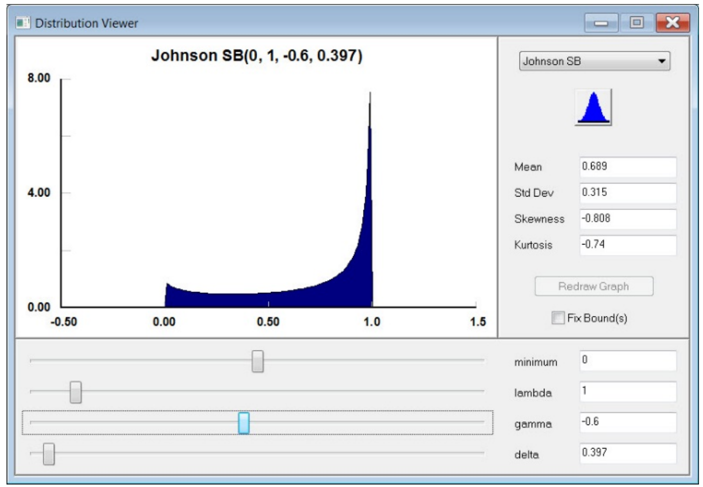
Doubly Bounded

- Skewed
  - Triangular
  - Beta
  - Johnson SB
  - Uniform

If distribution is positive near one or both bounds, then you can try either Beta or Johnson SB. In that way, you can quickly map a reliability problem. It requires a minimum, maximum, and a good guess.

The Johnson SB distribution goes to zero at both bounds. The Beta distribution goes to zero if p,q is greater than 1, a finite value if p,q are equal to 1, and infinity if p,q less than 1. p controls the left side, q the right side.





Lower Bound

- Time to An Event
  - Exponential
  - Weibull
  - Gamma

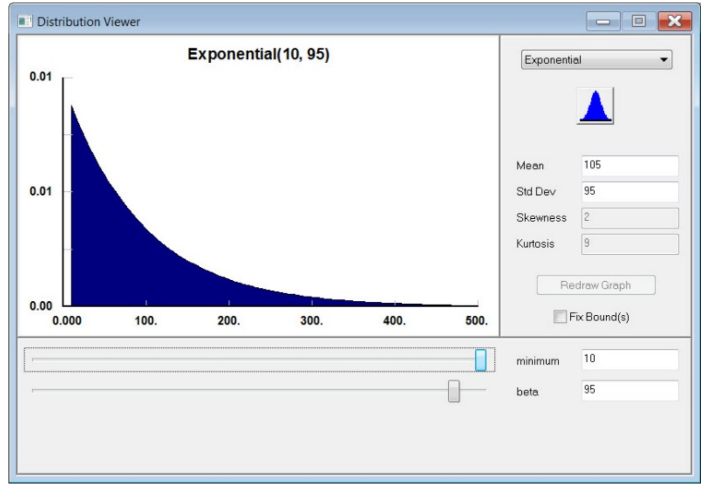
To model the time to a random event, use the Exponential distribution. It requires a minimum and a mean.

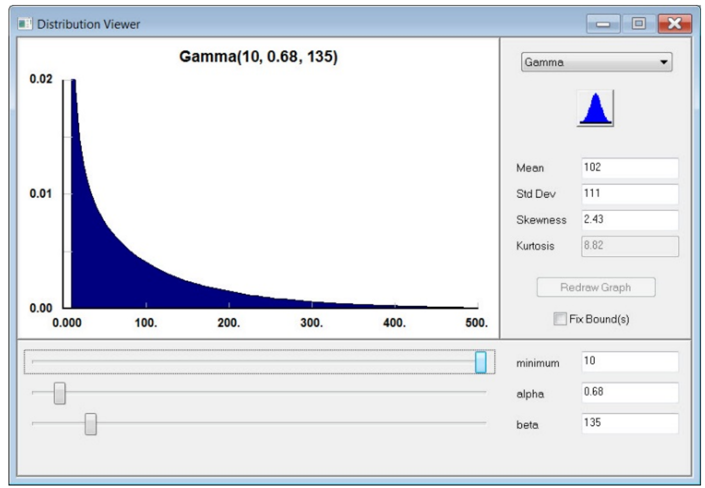
To model the time to a complex event, like the time to equipment failure or arrival from different unknown sources, you can use either a Weibull or Gamma distribution. It requires a minimum, mean, and standard deviation.

$$\beta = \frac{(\text{STANDARD DEVIATION})^2}{(\text{MEAN} - \text{MIN})}$$

$$\alpha = \frac{(\text{MEAN} - \text{MIN})}{\beta}$$

Gamma(MIN, ALPHA, BETA)





Lower Bound

- Time to Task Completion
  - Weibull
  - Gamma

To model the time to a task completion, use the Gamma distribution because the parameters are easy to calculate. It requires the minimum, mean, and standard deviation. Or the mode may be substituted for the mean. An alpha of 1.5 is a good starting point.

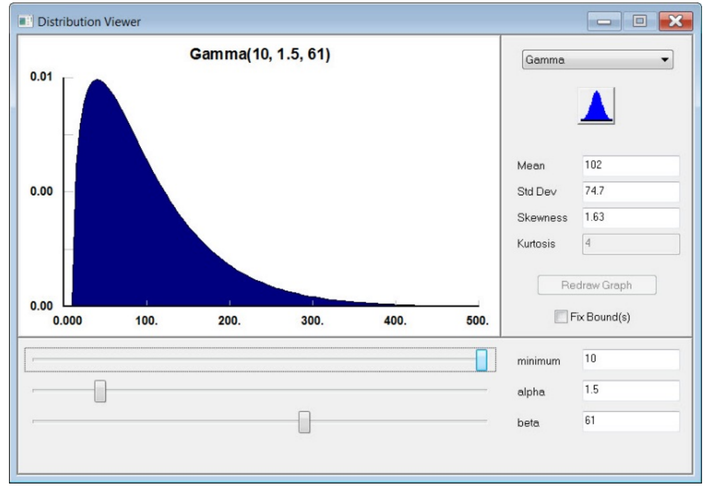
$$\beta = \frac{(\text{STANDARD DEVIATION})^2}{(\text{MEAN} - \text{MIN})}$$

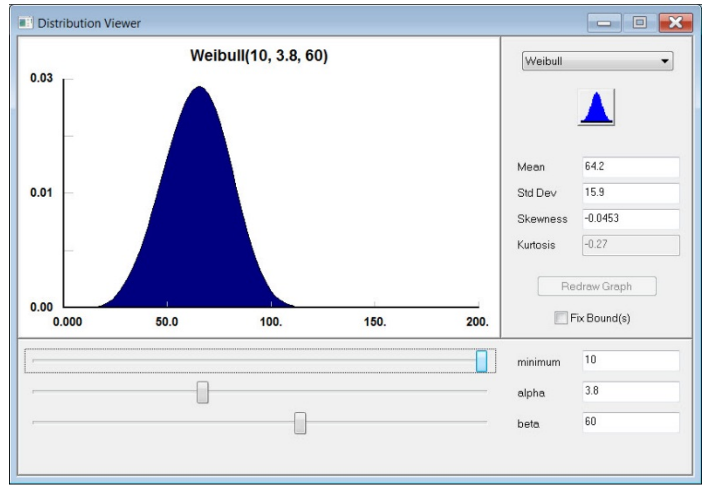
$$\alpha = \frac{(\text{MEAN} - \text{MIN})}{\beta}$$

$$\alpha = \frac{\text{MODE} - \text{MIN}}{\beta} + 1$$

Gamma(MIN, ALPHA, BETA)

To get negative skewness, use a Weibull distribution. Start with an alpha of 3.8 (which replicates Normal) and increase as needed.

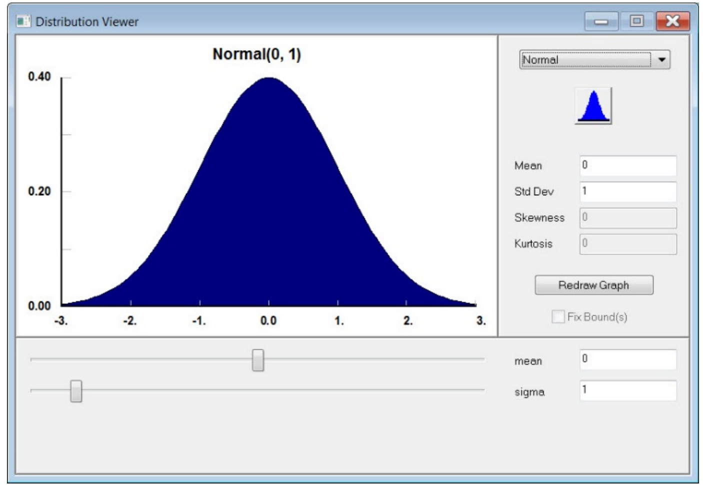


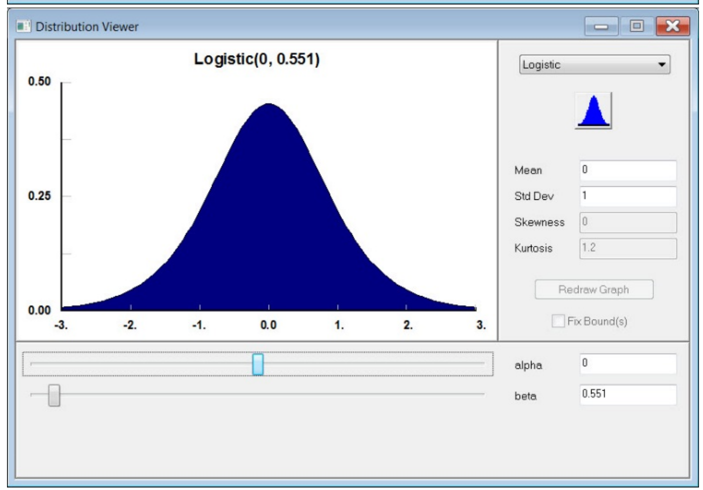


Unbounded

- Symmetrical
  - Normal
  - Logistic
  - Cauchy
- Skewed
  - Johnson SU
  - Extreme Value IA
  - Extreme Value IB

With a mean and standard deviation, a Normal distribution can be used but that is usually a poor expression of the process. It rarely fits any data, because it usually underestimates the tails. Use Logistic instead. (Limit tails to prevent anomalous variates)





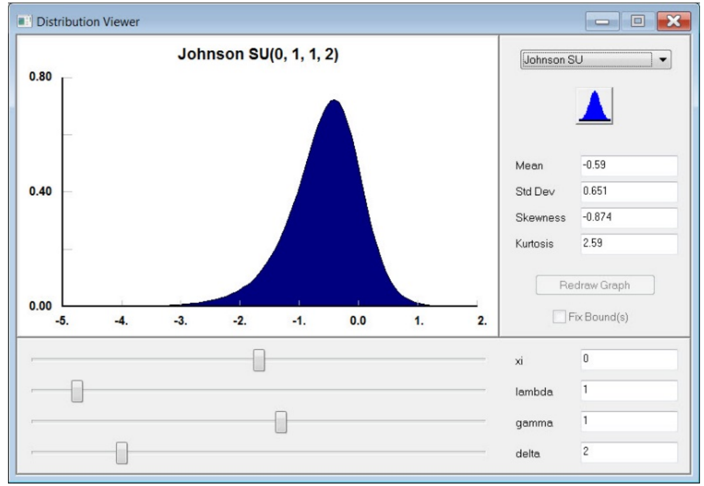
Unbounded

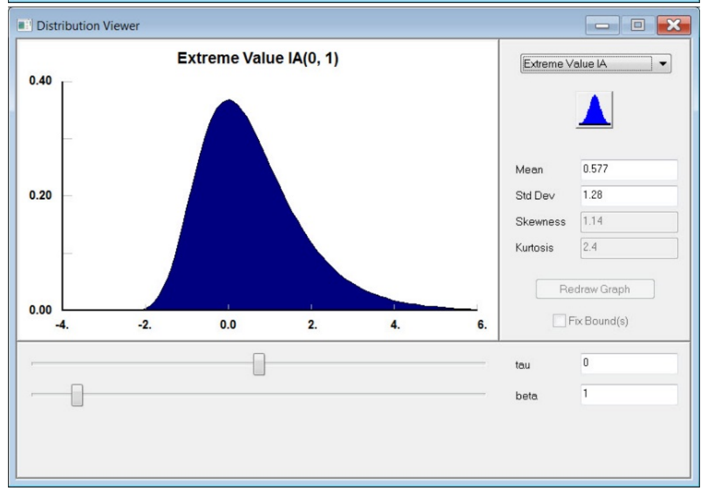
- Symmetrical
  - Normal
  - Logistic
  - Cauchy

- Skewed
  - Johnson SU
  - Extreme Value IA
  - Extreme Value IB

To model unbounded and skewed distributions, start with a Johnson SU. It requires at least mean and a standard deviation as well as a visual sense of what is needed. It also allows negative skewness.

To model the largest value of a parameter in each period, e.g. highest river height, use an Extreme Value IA. It requires a mean and some visualization. Extreme Value IB is the smallest value.





Remember

Some rivers can only be crossed with risk; leave room in the model for the unexpected.