ENE 3031

Computer Simulation

Week 9: Input Modeling

(attributed by Seong-Hee Kim and Barry L. Nelson)

Chuljin Park

Assistant Professor

Industrial Engineering

Hanyang University

Input Modeling

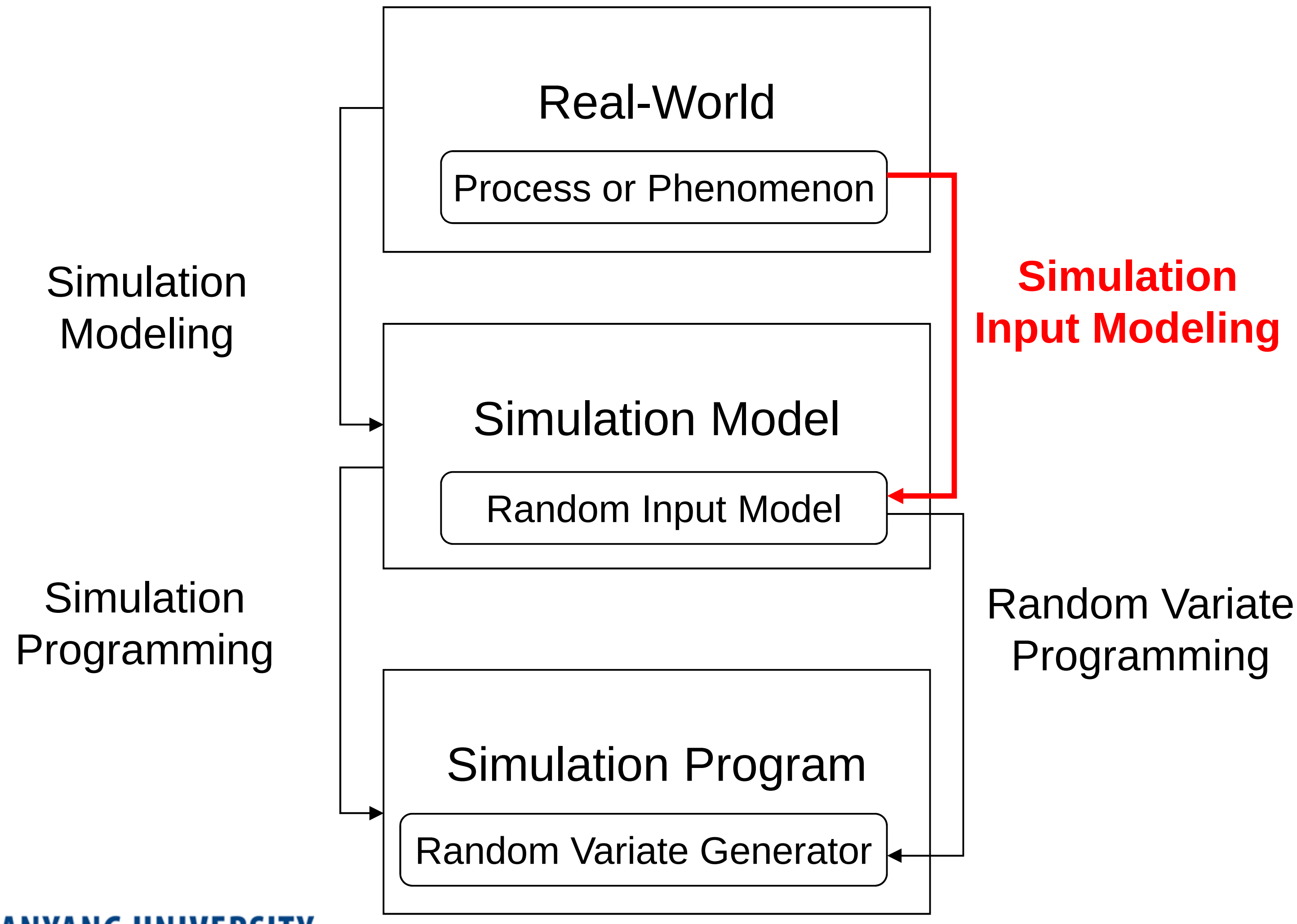
- Input models represent the **uncertainty** in a stochastic simulation.
- The fundamental requirements for an input model are:
  - It must be capable of representing the physical realities of the process.
  - It must be easily tuned to the situation at hand.
  - It must be amenable to random variate generation.

- There is no “true” model for any stochastic input. The best that we can hope is to obtain an approximation that yields useful results.
- A key distinction in input modeling problems is the presence or absence of data:
  - When we have data, then we *fit* a model to the data. Good software is available for this.
  - When no data are available then we have to creatively use what we can get to construct an input model.

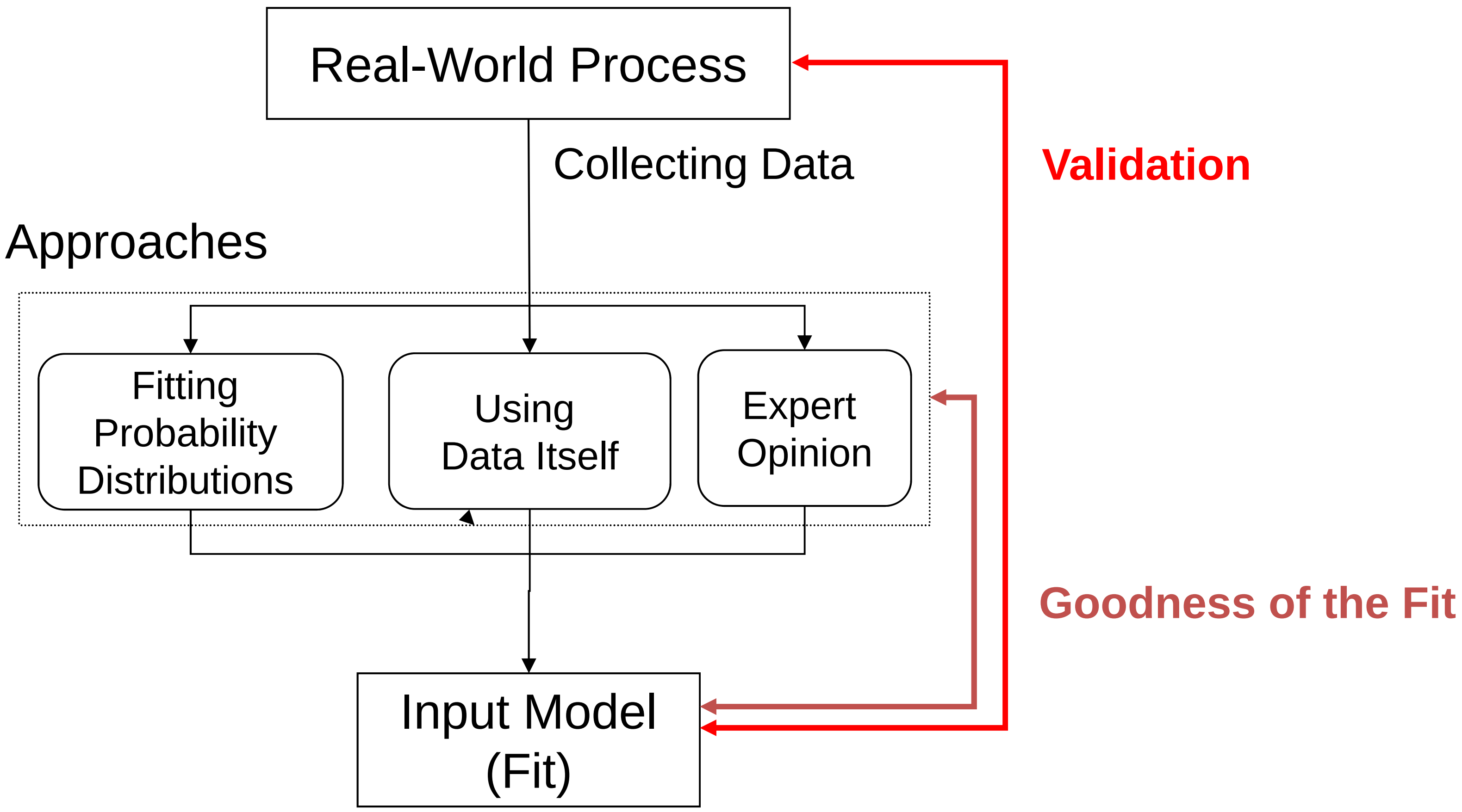
Outline

- Input modeling with data
  - physical basis for distributions
  - fitting and checking
  - ExpertFit
- Input modeling without data
  - sources of information
  - incorporating expert opinion

Simulation Model Development



Input Model Development



Why do we use input models?
Reliability Example

- Suppose you are a supplier of a component that is supposed to last for one year, a component that you know has a mean time to failure of 2 years.
- A client is willing to pay $1000 for your component, but wants you to pay a penalty of $5000 if failure occurs in less than one year
- Ignore the uncertainty: $1000 profit per component
- Life time ~ Expo with mean 2 years

Does the particular input model matter?

Reliability Example (Continued)

- Life Time ~ Expo with mean 2 years:
- Life Time ~ Unif(0,4years) :
- Uniform has the right mean (2 years) but it results in much smaller loss and causes you to underprice the component.

Input Modeling with Data

1. Select one or more candidate distributions, based on physical characteristics of the process and graphical examination of the data.
2. Fit the distribution to the data (determine values for its unknown parameters).
3. Check the fit to the data via tests and graphical analysis.
4. If the distribution does not fit, select another candidate and go to 2, or use an empirical distribution.

Physical Basis for Distributions

- Most probability distributions were invented to represent a particular physical situation.
- If we know the physical basis for a distribution, then we can match it to the situation we have to model
- A number of examples follow…

– : Models the number of successes in *n* trials, when the trials are independent with common success probability, *p*. *Example: the number of defective components found in a lot of n components*.

– : Models the number of trials required to achieve *k* "successes." *Example: the number of components that we must inspect to find 4 defective components.*

- : Models the number of independent events that occur in a fixed amount of time or space. *Ex: number of customers that arrive to a store during 1 hour, or number of defects found in 30 cubic meters of sheet metal.*
- : Models the distribution of a process that can be thought of as the *sum* of a number of component processes. *Ex: the time to assemble a product which is the sum of the times required for each assembly operation.*

– : Models the distribution of a process that can be thought of as the *product* of a number of component processes. *Example: the rate of return on an investment, when interest is compounded, is the product of the returns for a number of periods. Also widely used to model stock prices.*

– : Models the time between independent events, or a process time which is *memoryless*. *Example: the time to failure for a system that has constant failure rate over time. Note: if the time between events is exponential, then the number of events is Poisson.*

HANYANG UNIVERSITY

– : The sum of *k* identical exponential random variables. A special case of the gamma...

– : An extremely flexible distribution used to model nonnegative random variables.

– : An extremely flexible distribution used to model bounded (fixed upper and lower limits) random variables.

– : Models the time to failure for components; can model increasing or decreasing failure rate hazard. *Ex: the time to failure for a disk drive.*

– : Models complete uncertainty, since all outcomes are equally likely.
– : Models a process when only the minimum, most likely and maximum values of the distribution are known. *Ex: the minimum, most likely and maximum inflation rate we will have this year.*
– : Reuses the data themselves by making each observed value equally likely. Can be interpolated to obtain a continuous distribution.

HANYANG UNIVERSITY

Fitting

- Common methods for fitting distributions are *maximum likelihood*, *method of moments*, and *least squares*.
  - While the method matters, the variability in the data often overwhelms the differences in the estimators.
  - Remember: There is no "true distribution" just waiting to be found!

- Ways to check fit include the χ^2 ,K-S and Anderson-Darling tests, and density-histogram and *q-q* plots.
  - Beware of goodness-of-fit tests because they are unlikely to reject *any* distribution when you have little data, and are likely to reject *every* distribution when you have lots of data.
  - Tests represent lack of fit by a *summary statistic*, while plots show where the lack of fit occurs and whether it is important.
  - χ^2 tests and density-histogram plots are sensitive to how we group the data.

Tests and *p*-values

- In the typical test...
  H_0: the chosen distribution fits
  H_1: the chosen distribution does not fit
- Test statistics calculate difference between data and a chosen distribution.
- Thus, a small test statistics supports a good fit.
- The *p*-value of a test is the Type I error level (significance) at which we would *just reject H_0 for the given data.*
- Thus, a large (> 0.10) *p*-value supports H_0 that the distribution fits.

ExpertFit

- Stand-alone software package for fitting distributions to data.
- Data > Data Summary: good for getting sample mean and sample variance
- Models > Automated Fitting: sort candidates from the best
- Models > View/Delete Models > Show Model Parameters: give fitted distributions

- Reports *test-statistic* for
  - Anderson-Darling test
  - Kolmogorov-Smirnov test
  - χ^2 test
- Also provide density-histogram, p-p and q-q plots.

Data Summary/Fitted Model

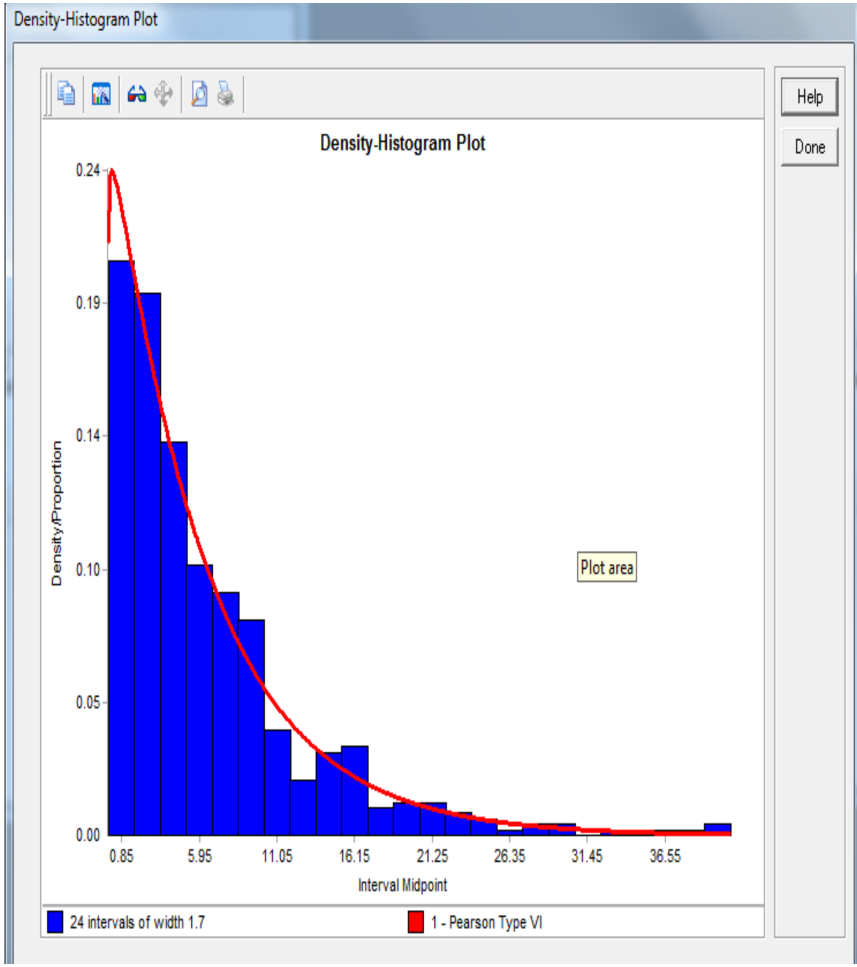
Data-Summary Table

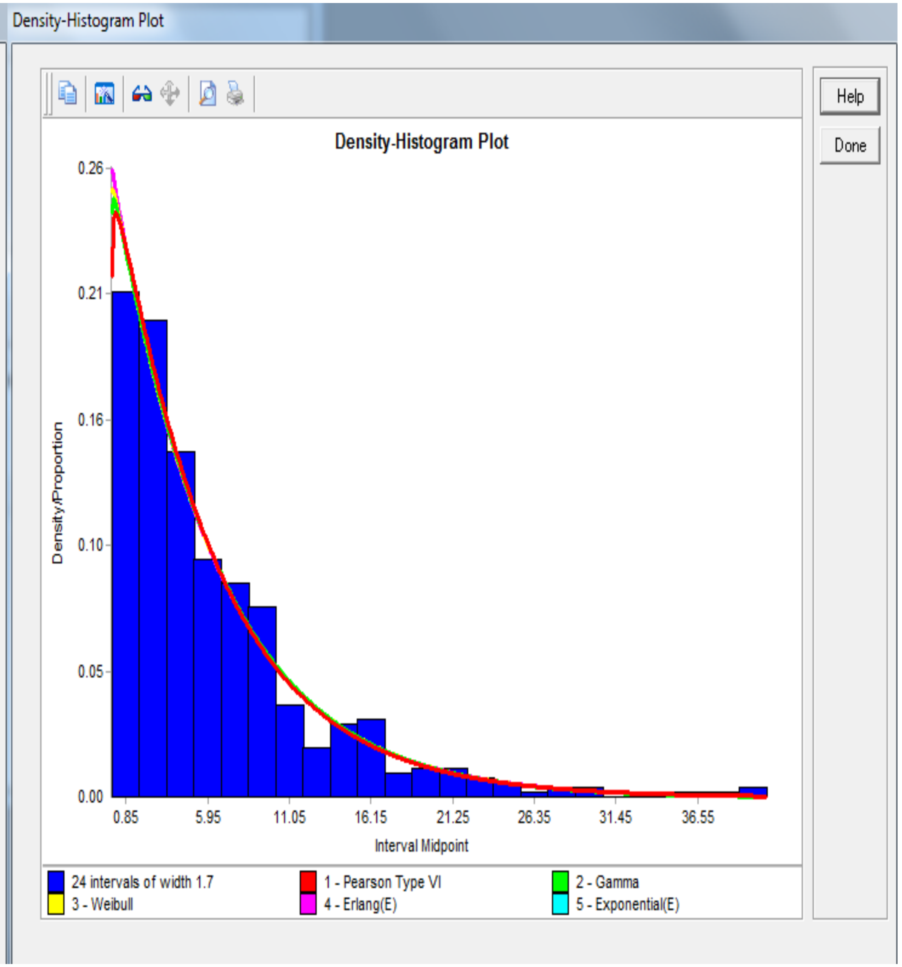
Data Characteristic	Value
Source file	3044a
Observation type	Real valued
Number of observations	500
Minimum observation	0.00417
Maximum observation	40.21170
Mean	6.56469
Median	4.46179
Variance	44.04246
Coefficient of variation	1.01093
Skewness	2.03511

Fitted Models

Model	Parameters		
6 - Erlang	Location	Default	0.00000
	Scale	ML estimate	6.56469
	Shape	ML estimate	1
7 - Exponential	Location	Default	0.00000
	Scale	ML estimate	6.56469
8 - Beta	Lower endpoint	OPT estimate	1.43349 e -4
	Upper endpoint	OPT estimate	80.42180
	Shape #1	ML estimate	0.93629
	Shape #2	ML estimate	10.45035
9 - Johnson SB	Lower endpoint	OPT estimate	0.00000
	Upper endpoint	OPT estimate	43.87742
	Shape #1	ML estimate	1.54858
	Shape #2	ML estimate	0.67884
10 - Lognormal	Location	Default	0.00000
	Scale	ML estimate	3.71528
	Shape	ML estimate	1.29332
11 - Random Walk	Location	Default	0.00000
	Scale	ML estimate	1.79107

Density-Histogram Plot





Usage Notes

- The "Automated Fitting" option tries all relevant distributions and ranks distributions from the best.
- Be sure to try different numbers of histogram cells; it affects the *test-statistic* of the χ^2 test, and your perception of the fit.

- Since exponential is a special case of Erlang which is a special case of gamma, “Automated Fitting” rarely selects exponential or Erlang. Similarly, exponential is a special case of Weibull.
- Raw data can be read in from text files (looks for .dat), one value per line.

Distributions in Simio

Random.Bernoulli	Random.Lognormal
Random.Beta	Random.NegativeBinomial
Random.Binomial	Random.Normal
Random.Erlang	Random.PearsonVI
Random.Exponential	Random.Pert
Random.Gamma	Random.Poisson
Random.Geometric	Random.Triangular
Random.JohnsonSB	Random.Uniform
Random.JohnsonUB	Random.Weibull
Random.LogLogistic	

Usage Notes

- Parameter definitions often do not match with Simio definitions.
- Check Help > User's Guide before setting parameters.
- Be sure to check if your entered input distribution generates random variates from the right distribution.

q-q Plot

- One way to generate data from cdf *F* is via

$$Y = F^{-1}(R)$$

- The *q-q* plot displays the *sorted* data

$$Y_1 \le Y_2 \le \cdots \le Y_n$$

vs.

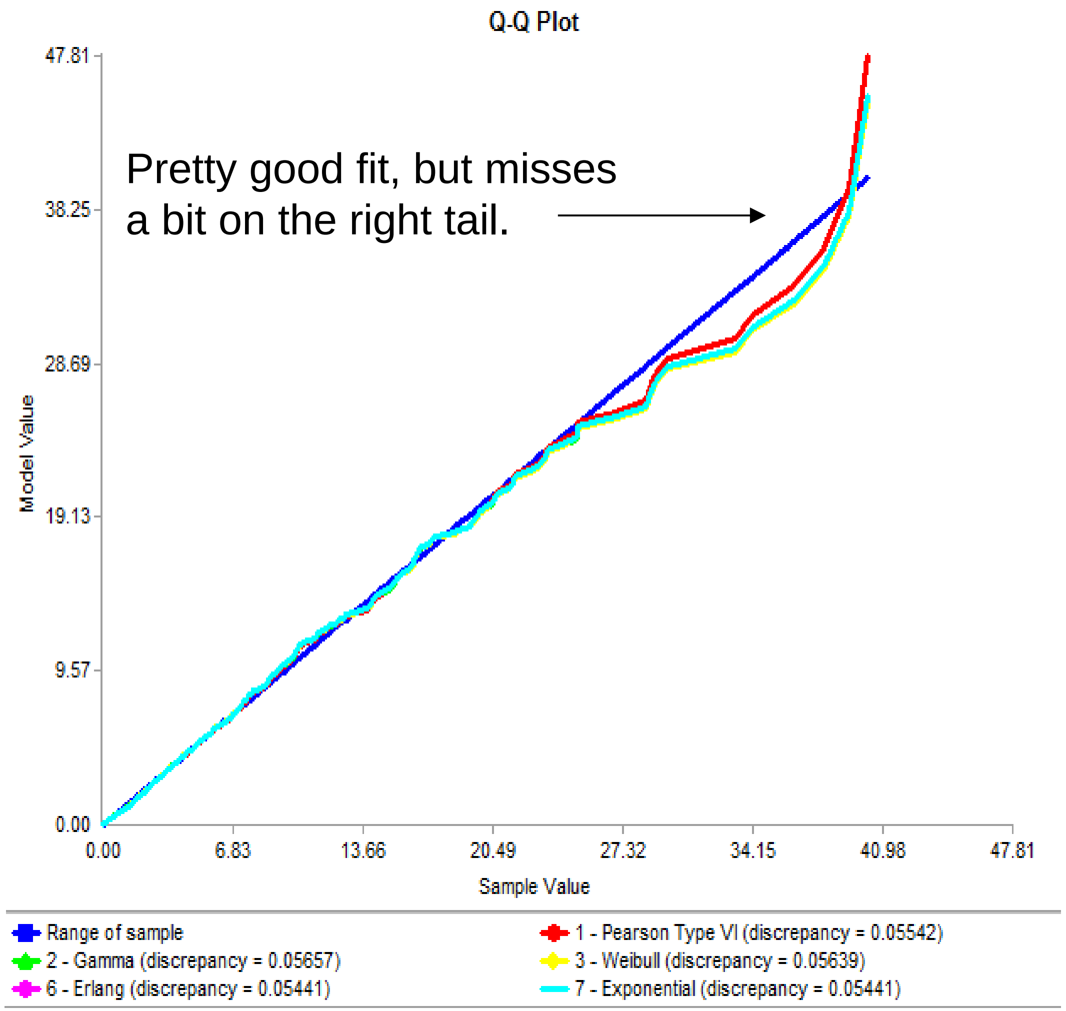
$$F^{-1}\left(\frac{j-1/2}{n}\right), j = 1,2,\ldots n$$

Features of the *q-q* Plot

- It does not depend on how the data are grouped.
- It is much better than a density-histogram when the number of data points is small.
- Deviations from a straight line show where the distribution does not match.

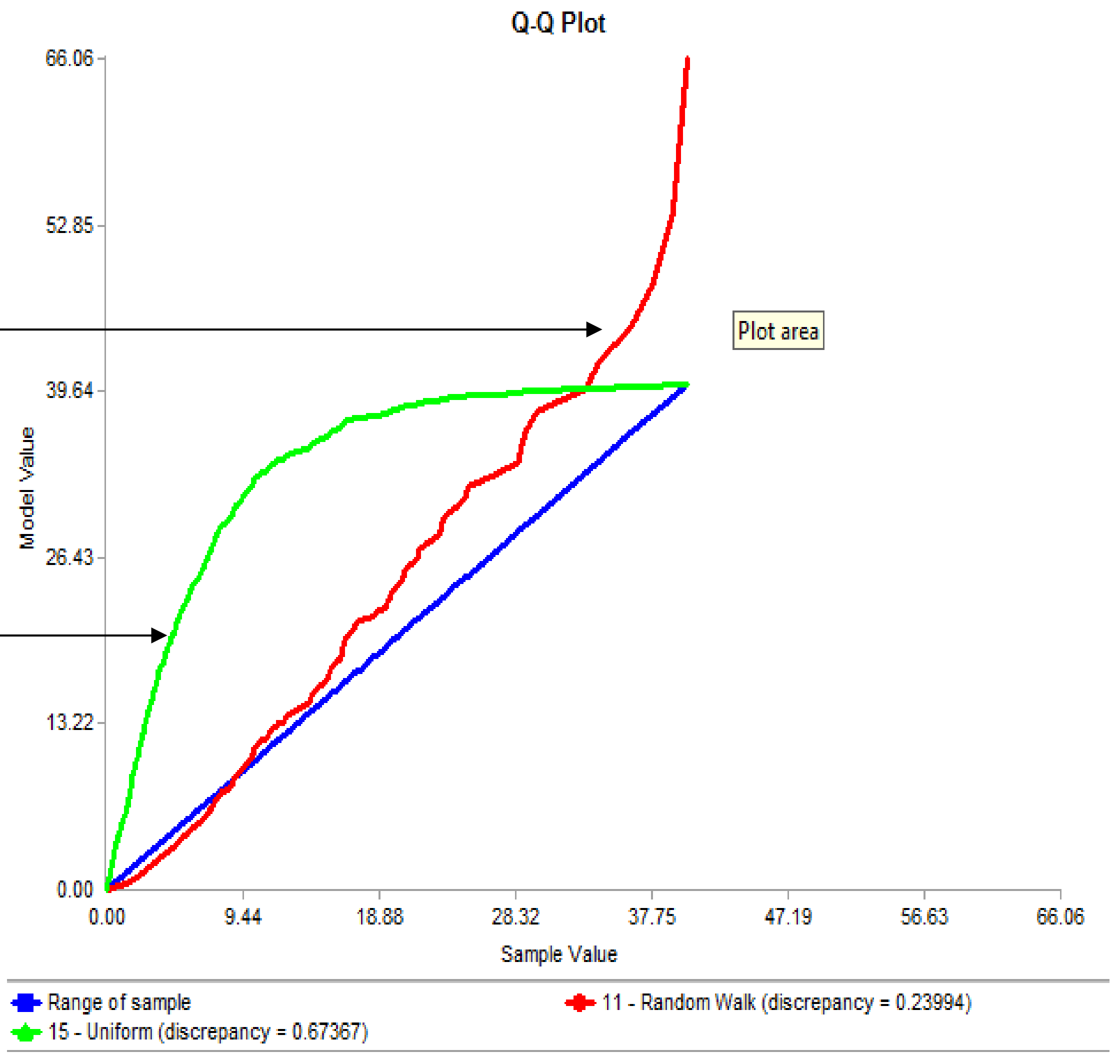
Mode (from the top menu) > Advance > Comparisons > Probability Plots > Q-Q Plot

A straight line implies the *family* of distributions is correct; a 45^{o} line implies correct *parameters.*



misses badly in the right tail

a curved line implies a wrong dist'n family



Fitting with GoF tests & Q-Q plots

Chi-square test

Features:

- A formal comparison of a histogram or line graph with the fitted density or mass function
- Sensitive to how we group the data.

K-S and A-D tests

Features:

- Comparison of an empirical distribution function with the distribution function of the hypothesized distribution.
- Does not depend on the grouping of data.
- A-D detects discrepancies in the tails and has higher power than K-S test

- Beware of goodness-of-fit tests because they are unlikely to reject *any* distribution when you have little data, and are likely to reject *every* distribution when you have lots of data.
- Avoid histogram-based summary measures, if possible, when asking the software for its recommendation!