Input Data Analysis (Part 2)

Dr.Çağatay ÜNDEĞER

Öğretim Görevlisi
Bilkent Üniversitesi Bilgisayar Mühendisliği Bölümü
&
...

e-mail : cagatay@undeger.com
cagatay@cs.bilkent.edu.tr

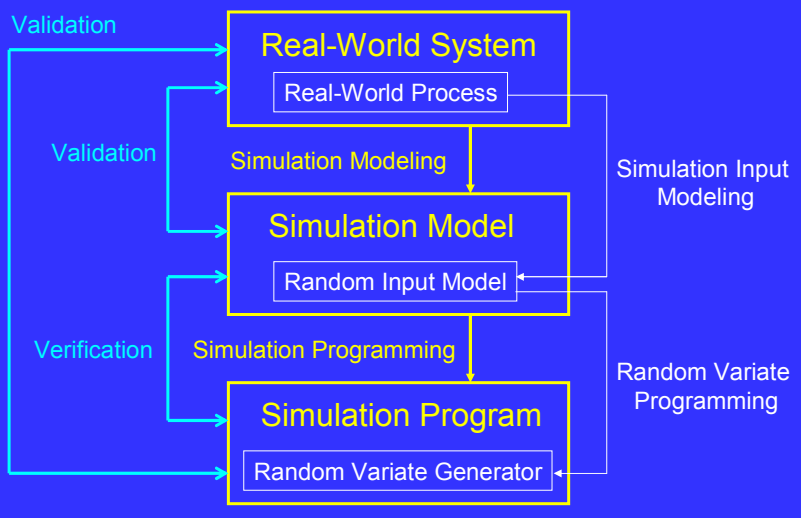
Input Data Analysis (Outline)

- Simulation Input Modeling
- Input Data Collection
  - Data Collection Problems
  - Practical Suggestions
  - Effect of Period of Time
- Input Modeling Strategy
  - Histograms
  - Probability Distributions
  - Selecting a Probability Distribution
  - Evaluating Goodness of Fit

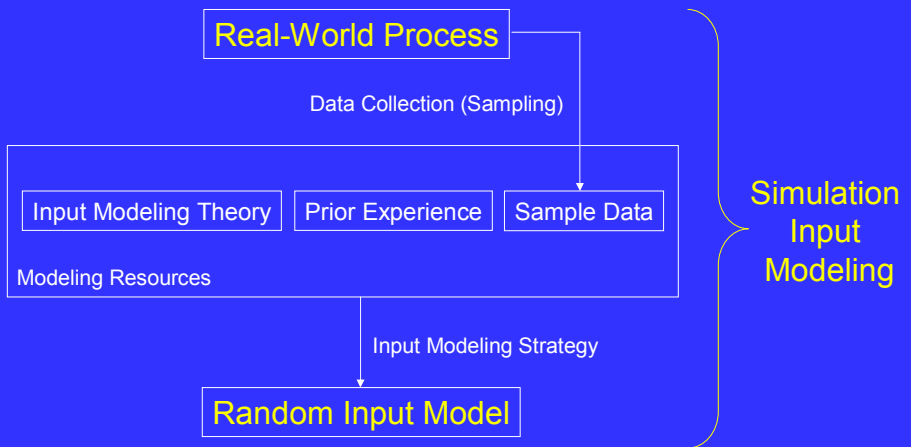
Input Data Analysis

- A valid simulation model involves:
  - Real-world system under consideration
    Real-World System
  - A theoretical model of the system
    Simulation Model
  - Computer-based representation of the model
    Simulation Program

Input Data Analysis



Simulation Input Modeling



Data Collection

- Most difficult aspect of simulation input modeling is;
  - Gathering data of sufficient;
    - Quality,
    - Quantity, and
    - Variety
  - In order to perform a reasonable analysis.

Data Collection Problems

- Sampled data may include:
  - Randomness, and
  - Annoyances.

Sources of Annoyance

- Data recorded in an order rather than in which it was observed.
- Data recorded with insufficient precision or rounded to the closest integer.
- Data recorded with obviously erroneous values.
- Data recorded with insufficient information (meta-data) about the data.
- Data grouped into intervals (e.g. histograms).

Practical Suggestions

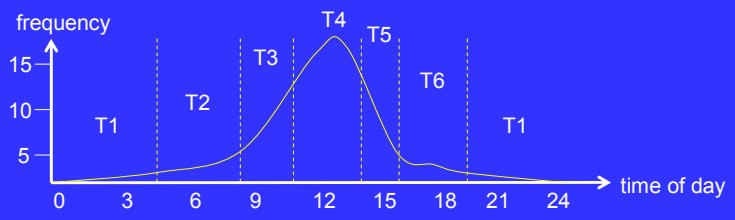
- Collect between 100-200 observations.
  - Less will have noticeable effects.
  - More will not gain much.
- For real values, record them with high precision.
- When interested in interval times, record event times and later calculate interval times.
- If there is any suspicion that real-world process depends on time of day or day of week, collect a number of samples from different time periods.

Effect of Period of Time

- Many process related with human activities are not stable even within small time periods.
- For example, arrivals rates in airports, restaurants, banks will be significantly effected by time of day.
- Period of time may not be important if we are interested in a small portion of time period (e.g. worst case scenario for times having peak demans).

Effect of Period of Time

- If period of time is significant;
  - Collect data from a whole range of different time periods,
  - Examine data collected, and
  - Divide data into intervals for different time periods if required.

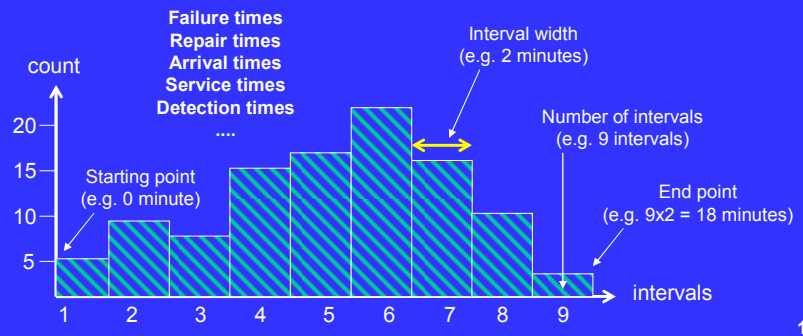


Input Modeling Strategy

- Examine the available data in details
  - Detect any problems in collected data
  - Correct the problems if required
  - Familier yourself with the shape of the data
    - Divide data into time intervals if required
    - Prepare histograms
    - Prepare cummulative probability distributions
- Determine alternative probability distributions that are possible to fit the data
- Fit alternative probability distribution to data
- Evaluate goodness of fit
- Select a probability distribution to represent data

Histograms

- A graphical display of tabulated frequencies (a set of data intervals & sample counts for them).
- Data samples are commonly represented as times for occurance of some events or completion of a process.

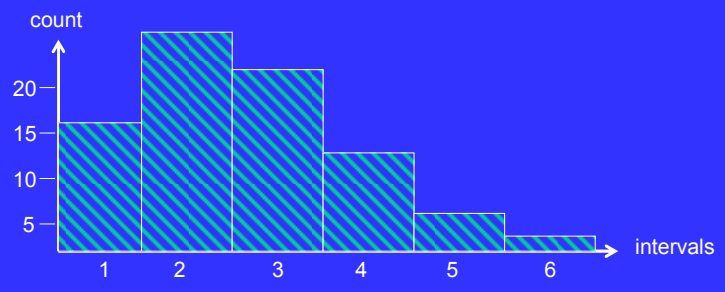


Histograms

- No definite rule to select correct histogram parameters.
- Iterate through;
  - Adjusting starting point and interval width,
  - And setting the number of intervals to cover all the data.
- Select an appropriate histogram for representing the data samples.

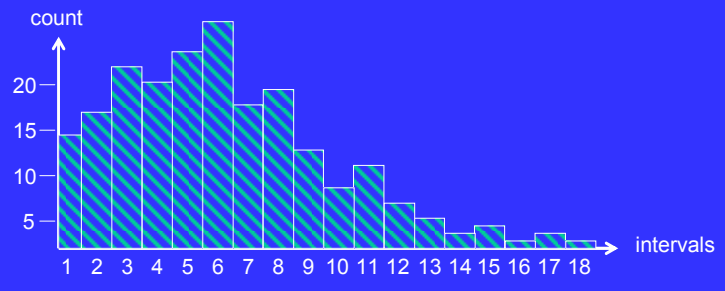
Histograms

- If interval widths are so large,
  - Chart will be too coarse, and
  - Details of the shape of the data will be lost.



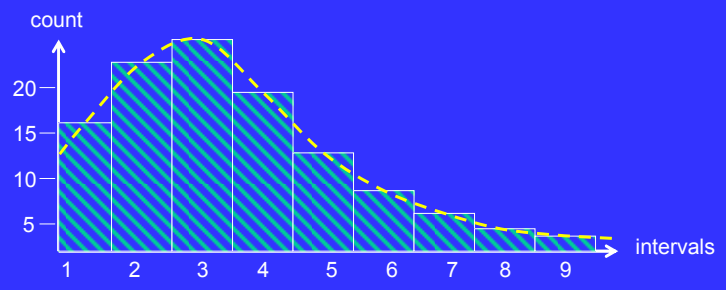
Histograms

- If interval widths are so small,
  - Chart will be too noisy, and
  - Overview of the shape of the data will be lost.



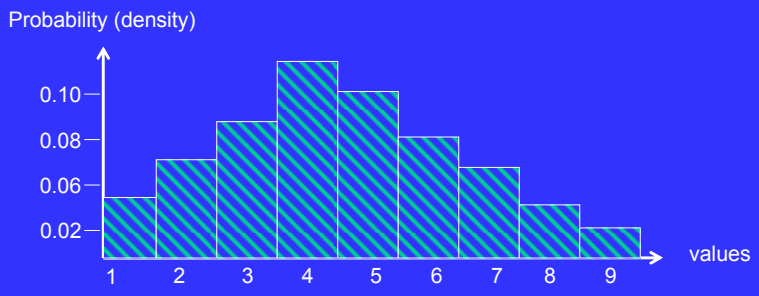
Histograms

- There is no best histogram.
- As a suggestion try to cover at least 3 to 5 samples in each interval.



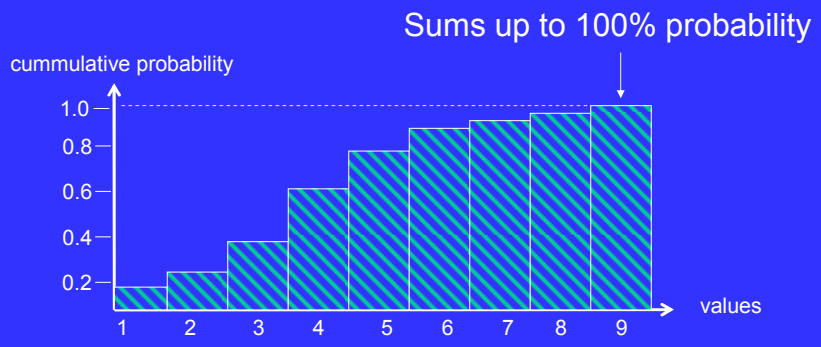
Probability Distributions

- Describes the values and probabilities associated with a random event (probability distribution function, probability density function).



Cummulative Probability Distributions

- Describes the values and cummulative probabilities associated with a random event.



Probability Distribution Types (WRT Values)

- Types according to number of values:
  - Discrete distributions:
    - Finite or countable number of different values
  - Continuous distributions:
    - Uncountable number of different values
- Types according to range of values:
  - Nonnegative distributions
  - Bounded distributions
  - Unbounded distributions

Probability Distribution Types (WRT Representation Style)

- Following probability distributions are commonly used in simulation input modeling:
  - Standard distributions
  - Emperical distributions

Standard Distributions

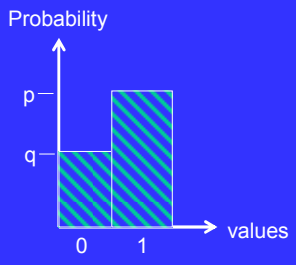
- Matematical parametric distributions that typically have location and scale parameters, and zero, one or two shape parameters.

Common Standard Distributions

Nonnegative Cont.	Nonnegative Dis.	Unbounded cont.	Bounded Cont.	Bounded Dis.
Chi-square	Geometric	Cauchy	Beta	**Bernoulli**
Erlang	Logarithmic	Error	Johnson S_B	Binomial
Exponential	Nagative binomial	Exponential power	Power function	**Uniform**
F	**Poisson**	Extreme value	**Triangular**	
Gamma		Johnson S_U	**Uniform**	
Inverse gaussian		Laplace		
Inverted weibull		Logistic		
Log-laplace		**Normal**		
Log-normal		Pareto		
Pearson type 5		Student's t		
Pearson type 6				
Random walk				
Rayleigh				
Wald				
Weibull				

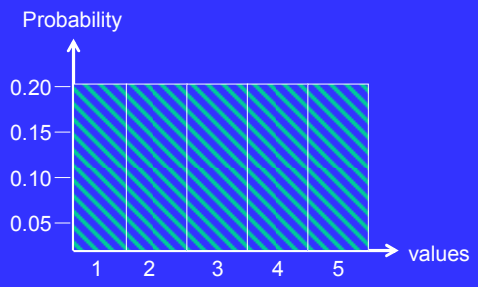
Bernoulli Distributions

- A discrete bounded probability distribution, which takes;
  - Value 1 with success probability *p*, and
  - Value 0 with failure probability *q = 1 − p*.



Uniform Distributions

- A bounded probability distribution, which takes values between a and b, where a>b, and probabilities of all the values are equal.
- Can be continuous or discrete.

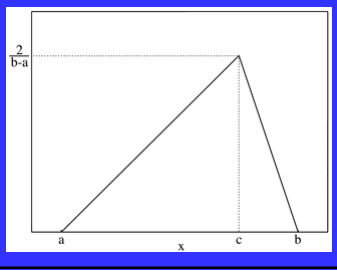


Triangular Distributions

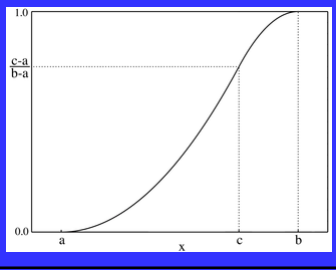
- A bounded contiuous probability distribution with lower limit *a*, mode *c,* and upper limit *b*.

$$f(x|a,b,c) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c \le x \le b \\ 0 & \text{for any other case} \end{cases}$$

Probability density function

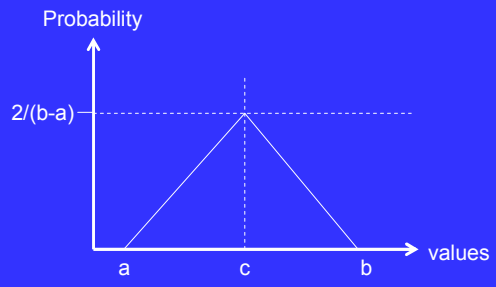


Cumulative distribution function



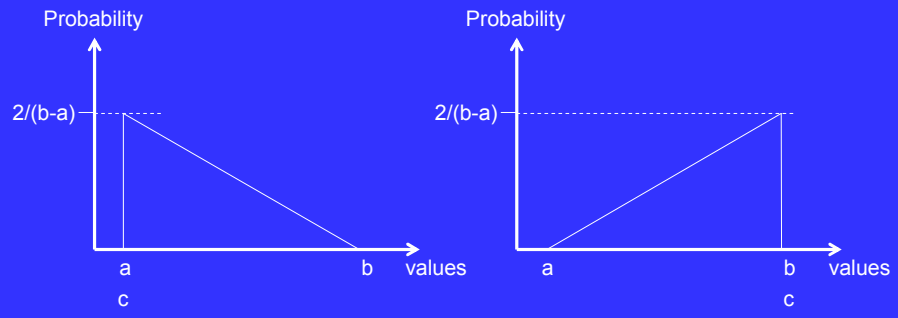
Symmetric Triangular Distributions

- A triangular distribution with c located at the center of a and b.



Two Points Triangular Distributions

- A triangular distribution with known a and b, and c is equal to either a or b.

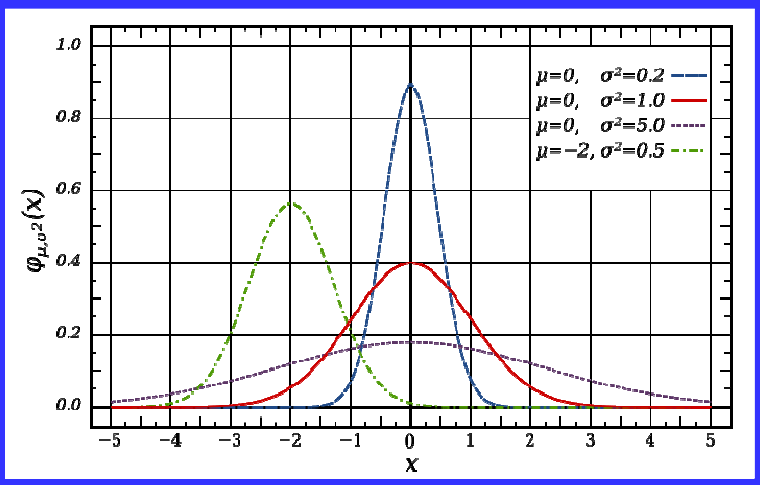


Normal (Gaussian) Distributions

- An important family of unbounded continuous probability distributions, applicable in many fields.
- Defined by two parameters:
  - Location: μ, mean (average)
  - Scale : σ^2, variance (standard deviation squared)
- Standard normal distribution:
  - Normal distribution with
    - A mean of 0, and a variance of 1.

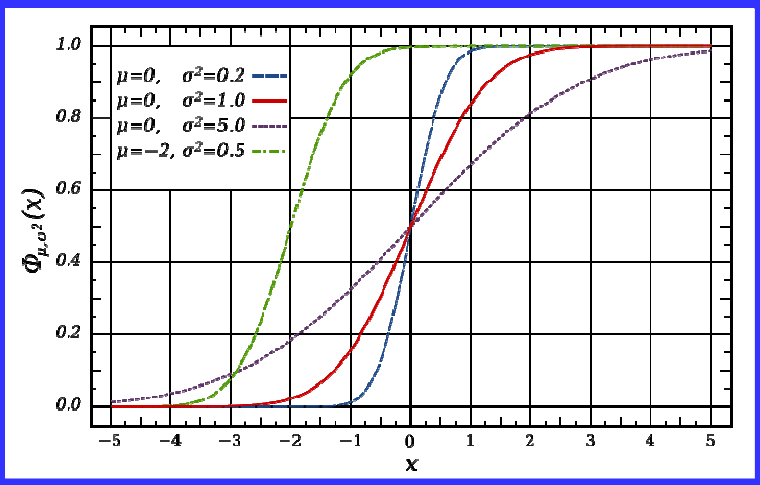
Normal (Gaussian) Distributions

A bell-like shaped probability density function



Normal (Gaussian) Distributions

Cummulative probability distribution



Normal (Gaussian) Distributions

- Importance:
  - A model of quantitative phenomena in the natural and behavioral sciences due in part to the central limit theorem.
  - Many measurements, ranging from psychological to physical phenomena can be approximated, to varying degrees, by the normal distribution.
  - Most widely used family of distributions in statistics.
  - Many statistical tests are based on the assumption of normality.

Normal (Gaussian) Distributions

Probability density function

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R},$$

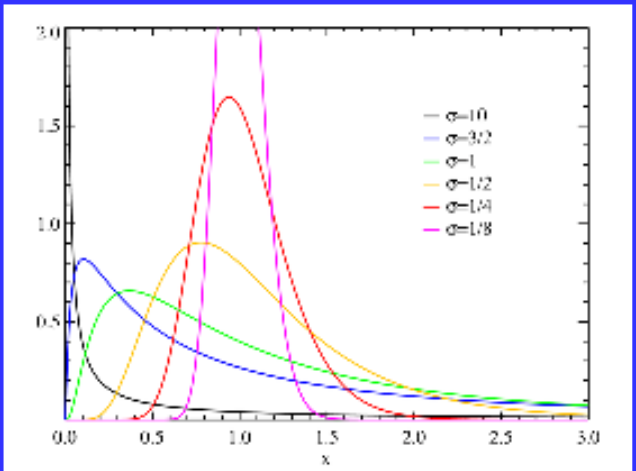
$\sigma > 0$ is standard deviation
μ is mean (expected value)

Log-Normal Distributions

- A nonnegative continuous probability distribution having single-tailed distribution of any random variable whose logarithm is normally distributed.
- Defined by two parameters:
  - Mean : μ
  - Standard deviation : σ

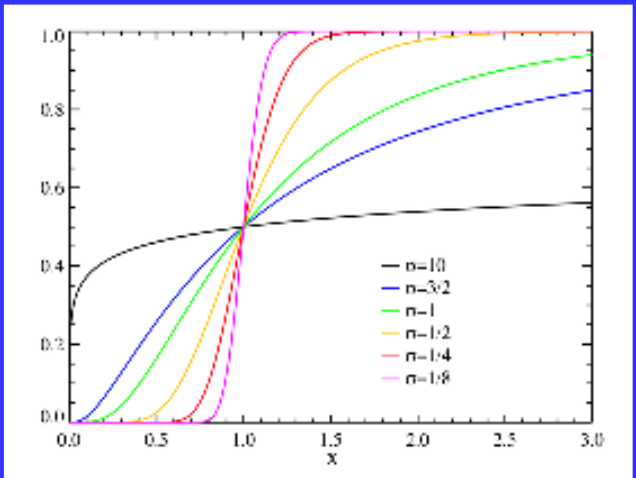
Log-Normal Distributions

Probability density function ($\mu = 0$)



Log-Normal Distributions

Cummulative probability distribution ($\mu = 0$)



Log-Normal Distributions

Probability density function for x > 0

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

$\sigma > 0$ is standard deviation
μ is mean

Exponential Distributions

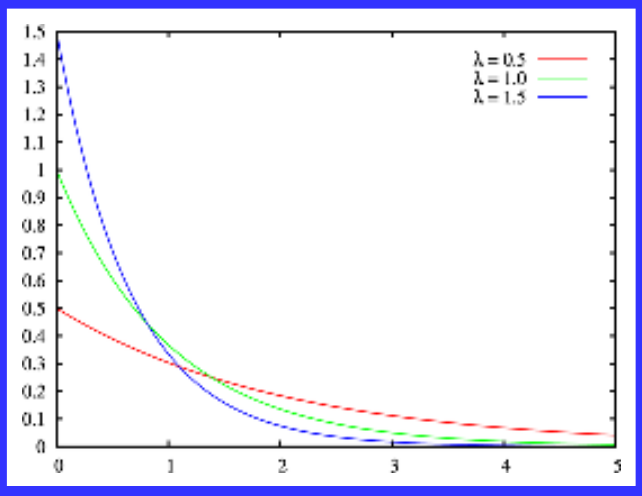
- A nonnegative continuous distribution with parameter λ, which describes the times between events in a Poisson process.
- Occurs naturally when describing the lengths of the inter-arrival times of events in a homogeneous Poisson process.

Exponential Distributions

- Poisson process:
  - A process in which events occur continuously and independently of one another at a constant average rate.
- Defined by one parameter:
  $\lambda > 0$: often called the *rate parameter*.

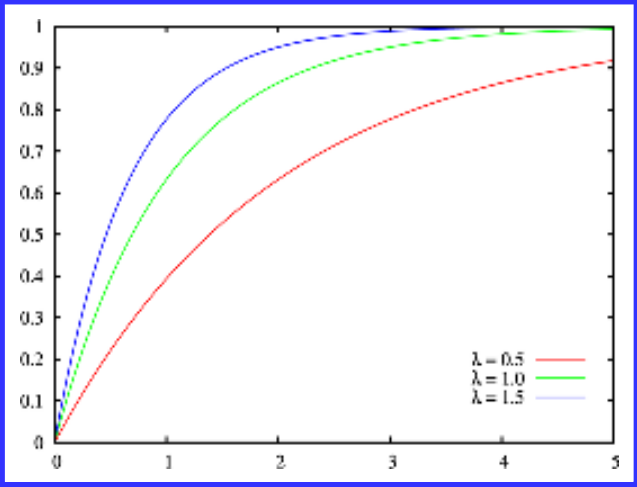
Exponential Distributions

Probability density function



Exponential Distributions

Cummulative probability distribution



Exponential Distributions

Probability density function

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

λ > 0 is *rate parameter*

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

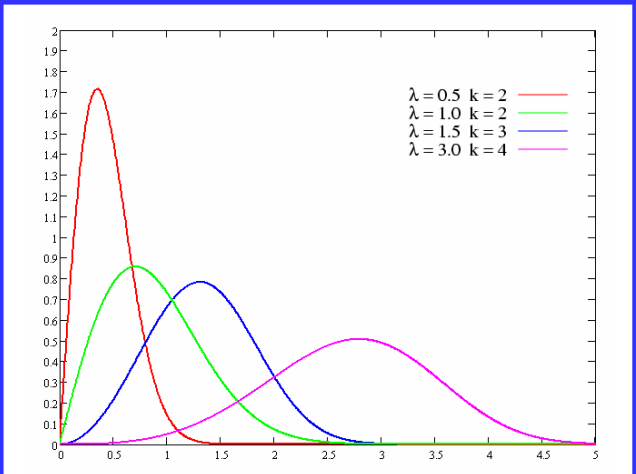
β > 0 is a scale parameter of the distribution, and
is the reciprocal (multiplicative inverse) of the *rate parameter*

Weibull Distributions

- A nonnegative continuous probability distribution used to describe the size distribution of particles.
- Defined by two parameters:
  - Shape : k
  - Scale : λ

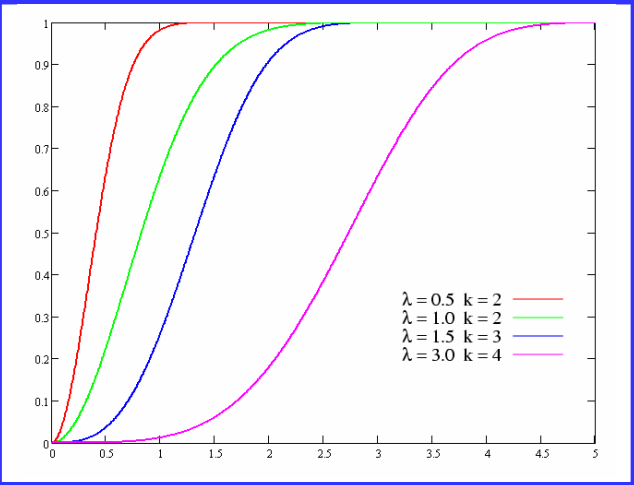
Weibull Distributions

Probability density function



Weibull Distributions

Cummulative probability distribution



Weibull Distributions

Probability density function

$$f(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

k is shape
λ is scale

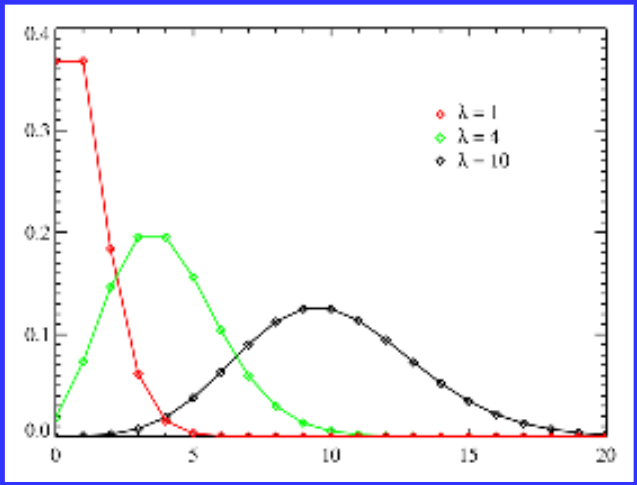
When *k* = 1, the Weibull distribution reduces to the exponential distribution.
When *k* = 3.4, the Weibull distribution appears similar to the normal distribution.

Poisson Distributions

- A nonnegative discrete distribution.
- Expresses the probability of a number of events occurring in a fixed period of time.
- Focuses on a number of discrete event occurrences (sometimes called "arrivals") that take place during a time-interval of given length.
- Defined by two parameters:
  - k : number of occurrences of an event
  - $\lambda > 0$: expected number of occurrences in the fixed interval.

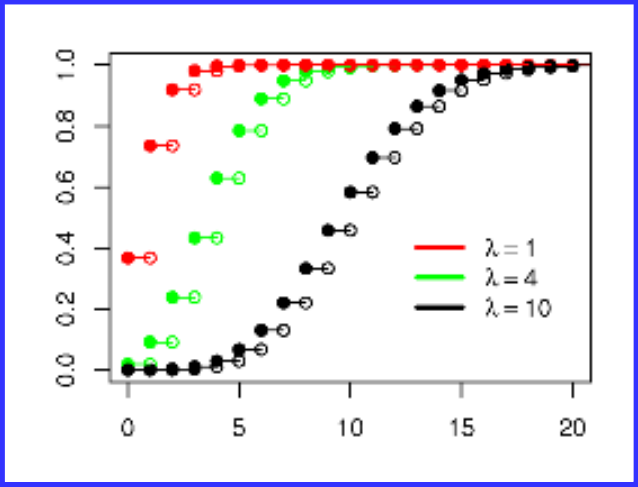
Poisson Distributions

Probability density function



Poisson Distributions

Cummulative probability distribution



Poisson Distributions

Probability density function

The probability that there are exactly *k* occurrences of event is equal to

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

k : number of occurrences of an event
$\lambda > 0$: expected number of occurrences in the fixed interval.

Emperical Distributions

- An alternative to standard distributions is using emperical distributions.
- A cumulative probability distribution function, which assigns a probability of 1/*n* to each element of a sample set that contains *n* number of samples.

Emperical Distributions

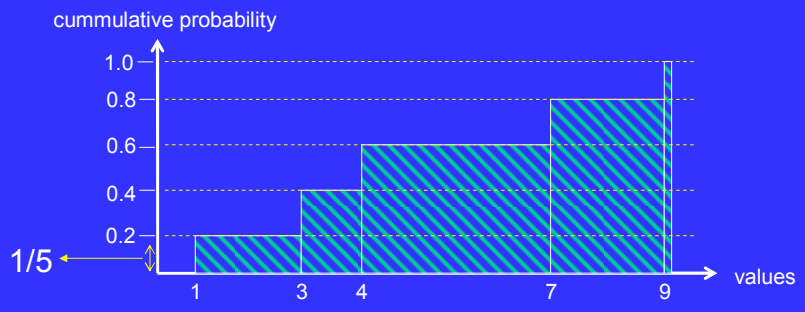
- In general form:
  - Probability of a value less than or equal to x (in other words cummulative probability of x)

$$F_n(x) = \frac{\text{number of elements in the sample} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

if $X_i \leq x$ then *1* else *0*

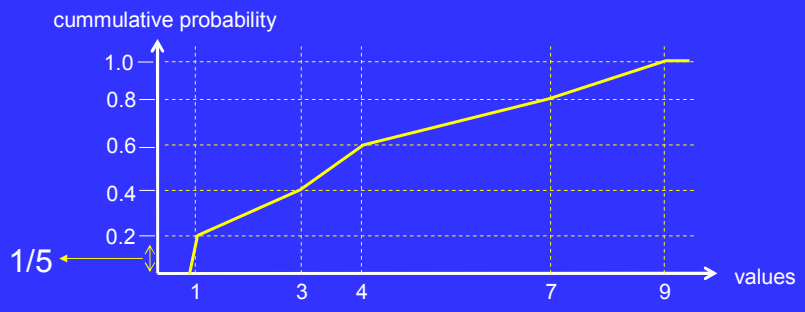
Emperical Distributions

- General form is a step function that rises at each unique observed sample value proportional to the total number of such values.



Emperical Distributions

- An alternative formalization used for continuous distributions replaces the step with a linear interpolation between subsequent points.

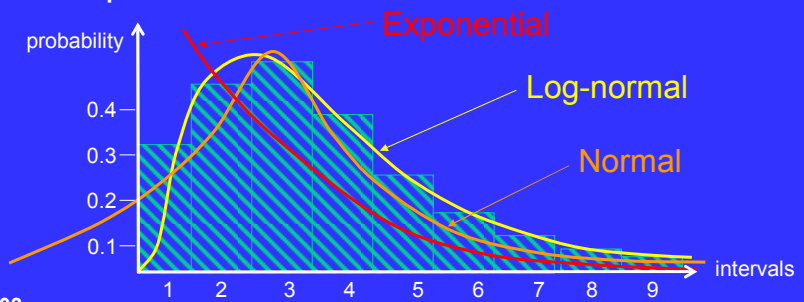


Emperical Distributions

- Drawbacks:
  - Can only represent bounded distributions within the observed sample range.
    - We may add an estimated tail for beginning and end points.
  - Quality of representation is completely dependent on the quality of sample available.
  - The probability that history repeats itself exactly is zero.

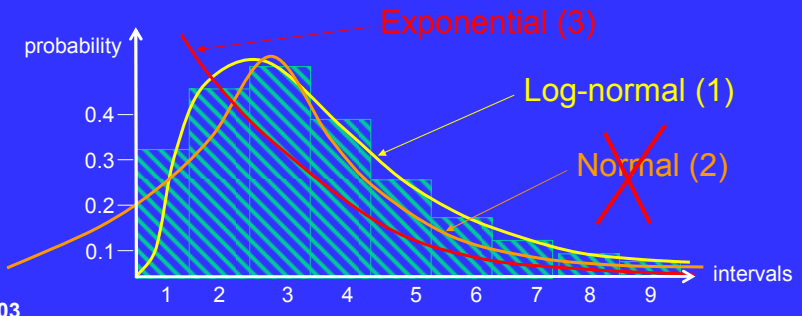
Selecting a Probability Distribution

- Use knowledge of randomness to determine any definite limits on the values it can produce.
- Fit as many standard distributions to the data as possible.



Selecting a Probability Distribution

- Use a set of criteria to rank goodness of fit of the fitted distributions to the data.
- If any of the top-ranked models are terribly inconsistent with the range of limits of value, rule them out.



Selecting a Probability Distribution

- Use a reasonable set of criteria to determine if the best of the fitted distributions is a reasonable representation of data.
- If best one provides a reasonable representation of data,
  - Use it in simulation,
- Otherwise,
  - Use an emperical distribution to represent data directly.

Evaluating Goodness of Fit

- Consider a number of measures of goodness of fit rather than a single one
  - Since each will be unrealible in some cases.
- Do not depend on goodness of fit measures
  - That rely on overly clean data samples (e.g. ignored problematic samples) or
  - On user supplied parameters (e.g. histogram configurations).
  - Since they can provide inconsistent results.

Evaluating Goodness of Fit

- In the context of simulation input modeling,
  - Classical goodness of fit methods in statistics are not completely appropriate for final assessment of quality of fit.
  - Statistical methods have definite assumptions that are sometimes not true for simulation modeling.
  - So, graphical heuristic methods should also be used to assess which is best and which is good enough.

Evaluating Goodness of Fit

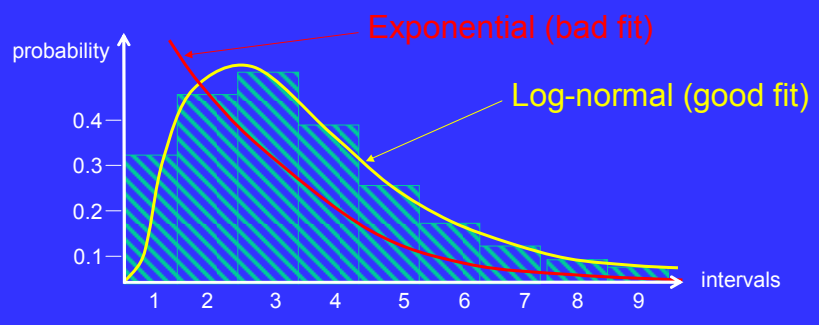
- For evaluation;
  - Histograms, and
  - Emperical cummulative probability distribution function of sample data can be used.

Evaluating Goodness of Fit (Histograms)

- A histogram is an estimate of probability density function of the event in real-world system.
- So it is reasonable to plot and compare the density functions of fitted models over the histogram.

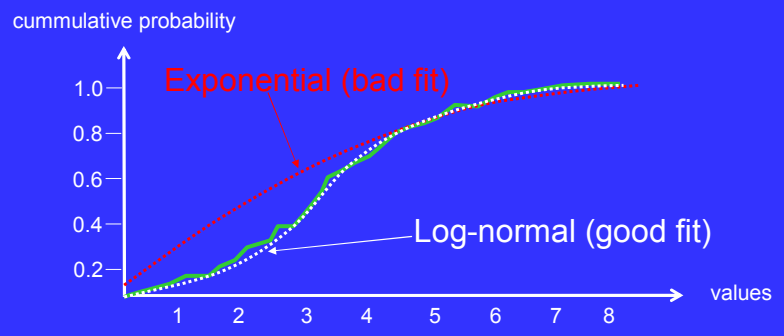
Evaluating Goodness of Fit (Histograms)

- Histograms represent sample counts, but they can easily be converted to density units by dividing to the total number of samples.



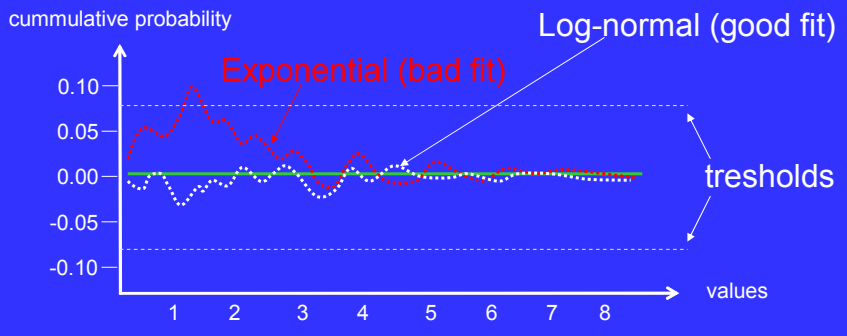
Evaluating Goodness of Fit (Cummulative Dist. Func.)

- Emprical cummulative distribution functions of sample data can be used to compare with the fitted models.



Evaluating Goodness of Fit (Cummulative Dist. Func.)

- Differences (errors) between emperical and fitted model cummulative probabilities can be plotted on to a graph.



Evaluating Goodness of Fit (Test Statistics)

- Test statistic:
  - An operational procedure of goodness of fit
  - To calculate a function of the data observed and the fitted model, and
  - To compare the errors with a critical value in order to accept or reject the hypothesis.
- Reject means there is sufficient evidence to say that the two distributions are not the same.
- Unless extreme errors, better not to reject, but rank to find a best or good enough solution.

Evaluating Goodness of Fit (Test Statistics: Chi-Squared)

- We divide the range of sample data into M intervals (similar to histograms).
- The first and last interval may be extended to [+/-] infinite to cover the entire range of random variable.
- Compute differences between sample data and fitted model (errors) in each interval.
- Sum up errors of all the intervals to get total error.
- This test is universally applicable.

Evaluating Goodness of Fit (Test Statistics: Chi-Squared)

M = Number of intervals

Ts = Number of samples

O_k = Number of samples in k^{th} interval

E_k = Probability of k^{th} interval in fitted model x Ts

$$\text{Total error} = \sum_{k=1}^{M} \frac{(O_k - E_k)^2}{E_k}$$

$(O_k - E_k)^2$ → squared error

E_k → weight that is inversely proportional to the number of expected points (more weight is placed on rare events)

Evaluating Goodness of Fit (Test Statistics: Chi-Squared)

- By changing the interval configuration, conflicting results can be produced.
- Therefore, results should not be trusted standalone.