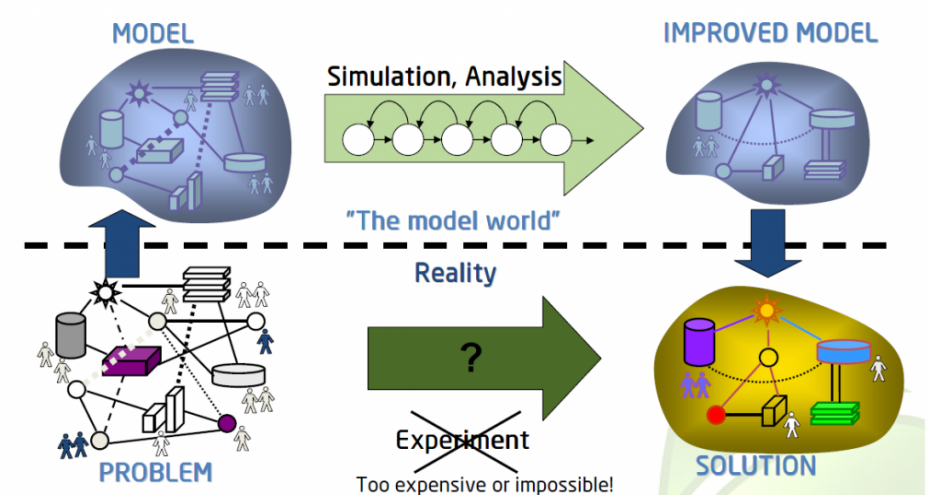
Modeling & Simulation: Input Data Modelling

Course: Modeling & Simulation - EEN14253

Dr. Jagat Jyoti Rath

Department of Electrical Engineering, MNNIT, Allahabad



What is Input Modeling?

What?

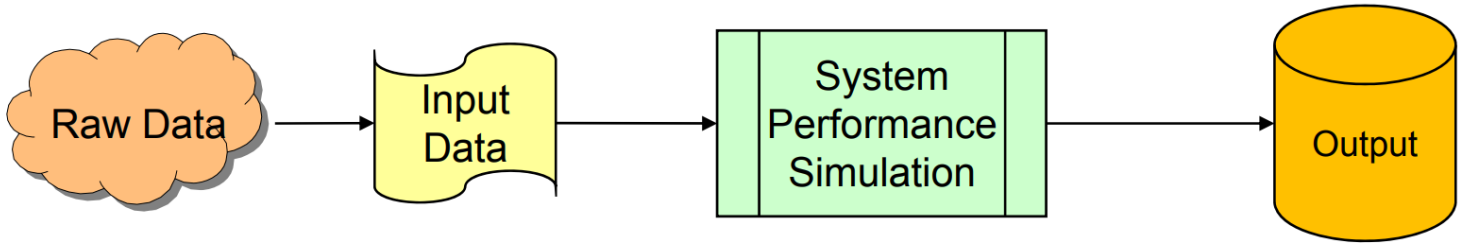
Input modeling for a Discrete Event Simulation (DES) involves analyzing and characterizing real-world data to create accurate probabilistic models that define system inputs.

The different steps for Input Modeling are:

- Data Collection
- Exploratory Data Analysis
- Identification of Suitable Probability Distributions
- Parameter Estimation
- Goodness of Fit
- Verification and Validation

Data Collection

- One of the biggest tasks in solving a real problem



- Even when model structure is valid simulation results can be misleading, if the input data is
  1. inaccurately collected
  2. inappropriately analyzed
  3. not representative of the environment

1. **Inaccurately collected:** For example, arrival times at a service desk are recorded manually but with inconsistent timestamps due to human error or forgotten entries, leading to incorrect estimates of the arrival distribution.
2. **Inappropriately analyzed:** Fitting a Normal distribution to highly skewed service time data without testing goodness-of-fit can result in a poor representation of variability, which affects system performance estimates.
3. **Not representative of the environment:** Collecting machine breakdown data during a low-production season and using it to model peak-season operations.

Exploratory Data Analysis

Visualizing and understanding the main characteristics of a dataset before formal modeling

A histogram can be visualized as:

Exploratory Data Analysis

Visualizing and understanding the main characteristics of a dataset before formal modeling

A histogram can be visualized as:

Histograms

A histogram is used to display the distribution of data values along the real number line.

- The number of class intervals depends on:
  1. The number of observations
  2. The dispersion of the data
- Histogram for continuous data corresponds to the probability density function of a theoretical distribution
- Histogram for discrete data corresponds to the probability mass function of a theoretical distribution.
- Same data can be represented with different bin sizes.

Tools for Identifying the Distribution

- **Scatter Plot**: A scatter diagram is a quality tool that can show the relationship between paired data
- Consider two random variables X and Y. The correlation between these two random variables can be represented as:

- Similarly positive or negative correlation can be represented as:

Identifying the Distribution

A family of distributions is selected based on:

- The context of the input variable
- Shape of the histogram

The easy to analyze distributions are:

- Exponential
- Normal
- Poisson
- Triangular
- Weibull

Remember the physical characteristics of the process

- Is the process naturally discrete or continuous valued?
- Is it bounded?
- Value range?
- No single and exact distribution for any stochastic input process
- Goal: obtain a good approximation

Selecting Family of Distributions

Using physical basis of the distribution as a guide:

- Number of independent events that occur in a fix amount of time or space
- Distribution of a process that is the sum of a number of component processes
- Distribution of a process that is the product of a number of component processes
- Time between independent events or a process time Time between independent events, or a process time that is memoryless
- Time to failure for components
- Modeling uncertainty
- Resamples from actual data collected

Quantile

A quantile is a cutoff point that divides a probability distribution or dataset into intervals with equal probabilities. Given data is sorted in ascending order:

- The 0.25 quantile (also called the 25th percentile or first quartile, Q_1) is the value below which 25% of the data falls.
- The 0.5 quantile is the median — 50% of the data lies below it.
- The 0.75 quantile (75th percentile or third quartile, Q3) is the value below which 75% of the data lies.

If X is a random variable with a CDF $F(x)$, then the q-quantile of X is a value γ such that

In other words:

Let $x_i, i = 1, 2, 3\ldots\ldots, n$ be a sample drawn from the random variable X and $y_j, j = 1, 2, 3, \ldots., n$ be the ordered version of this sample in ascending order.

Q-Q plot to identify a distribution

- Compare these sorted sample values y_j to the theoretical quantiles of a known distribution
- For each j-th value, the theoretical quantile is computed as:

  where j is the plotting position.
- Using the theoretical quantile, make the Q-Q plot:

- If the data fits the distribution F, the points will lie close to the 45 degree line i.e. slope is 1.

Q-Q plot to identify a distribution

- Compare these sorted sample values y_j to the theoretical quantiles of a known distribution
- For each j-th value, the theoretical quantile is computed as:

  where j is the plotting position.
- Using the theoretical quantile, make the Q-Q plot:
- If the data fits the distribution F, the points will lie close to the 45 degree line i.e. slope is 1.

While evaluating linearity for a Q-Q plot consider:

- The observed values never fall exactly on a straight line
- The ordered values are ranked and hence not independent, unlikely for the points to be scattered about the line
- Variance of the extremes is higher than the middle
- Linearity of the plots in the middle part is more significant

Q-Q plot to identify a distribution

Ex Consider, the door installation times by a robot collected as

$$T_i = (14.1, 13.7, 14.5, 15.2, 13.9, 14, 15, 14.3, 13.8, 14.6)$$

Using Q-Q plot justify if the above installation times follow an exponential distribution.

Parameter Estimation for a Distribution

Basic Parameters of Importance

Given n observations, $X_1, X_2, X_3, \ldots, X_n$ (discrete or continuous), the sample mean and variance is given as:

Given a model, the parameters are the actual values which yield a distribution. Typically, from given data, the parameters of a model (i.e. for a distribution) needs to be estimated. Given the standard distributions, the corresponding parameters are:

- Poisson
- Exponential
- Uniform
- Gaussian
- Weibull
- Triangular
- Erlang

Estimation of Parameters

From here onward, θ indicates the vector of parameters for any distribution. The two main approches to estimate parameters of probability distributions are:

- Maximum Likelihood Estimation
- Maximum A Posteriori Estimation

Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$. Then:

Likelihood Function

The joint probability density function of $x_1, x_2, \ldots, x_n$ which is a function of $\theta_1, \theta_2, \ldots, \theta_n$ is given as:

The above function, $L(\theta_1, \theta_2, \ldots, \theta_m)$ is called the likelihood function.

Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$. Then:

Likelihood Function

The joint probability density function of $x_1, x_2, \ldots, x_n$ which is a function of $\theta_1, \theta_2, \ldots, \theta_n$ is given as:

The above function, $L(\theta_1, \theta_2, \ldots, \theta_m)$ is called the likelihood function.

Maximum Likelihood estimator

If there exists an estimator $u_1(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)$ that maximizes the likelihood function, then

is the maximum likelihood estimator of x_i.

Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$. Then:

Likelihood Function

The joint probability density function of $x_1, x_2, \ldots, x_n$ which is a function of $\theta_1, \theta_2, \ldots, \theta_n$ is given as:

The above function, $L(\theta_1, \theta_2, \ldots, \theta_m)$ is called the likelihood function.

Maximum Likelihood estimator

If there exists an estimator $u_1(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)$ that maximizes the likelihood function, then

is the maximum likelihood estimator of x_i.

Maximum Likelihood Estimates

The corresponding observed values of the statistics i.e. $u_1(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)$ are called the maximum likelihood estimates of θ_i.

Maximization

The goal in MLE is to use maximization and hence get

- $\hat{\theta}$ i.e. estimates of the parameter

Using log function properties of monotonicity, we transform the above maximization task as:

Thus,

- To develop a MLE, we first express the LF in terms of LLF.
- Choose parameters $\hat{\theta}$ that maximize the value of LLF.
- To compute maximum first order derivative needs to be computed.

MLE for Exponential Distribution

Given the PDF of the exponential distribution

$$f(x) = \lambda e^{-\lambda x} \quad : \quad x \geq 0, \lambda > 0$$

Find the MLE of the exponential distribution parameter λ

Maximum Likelihood Estimates

MLE for Poisson Distribution

Find the MLE of the Poisson distribution parameter λ.

Maximum Likelihood Estimates

MLE for Gaussian Distribution

Find the MLE of the Gaussian distribution parameters μ, σ^2.

Goodness of Fit Tests

Conduct hypothesis testing on input data distribution using

- Kolmogorov-Smirnov test
- Chi-square test

Note there is no single distribution which is correct for input data.
Typical mistakes in identifying the correct distribution:

- Type I error
- Type II error

Statistical Decision	State of the null hypothesis	
	H_0 True	H_0 False
Accept H_0	Correct	Type II Error Incorrectly accept H_0 False negative
Reject H_0	Type I Error Incorrectly reject H_0 False positive	Correct

Chi-Square Test

Idea

Comparing the histogram of the input data to the shape of the candidate density or mass function.

Valid for large sample sizes when parameters are estimated by maximum-likelihood method. The steps for the Chi-square test are as follows:

- **STEP 1**: Arrange **n** observations into **k** classes. To choose the proper sample size k a guide is given as:
  1. If $n \leq 100$, k is chosen between 10 to 20
  2. If $n \leq 50$, k is chosen between 5 to 10
  3. If $n \geq 100$, k is chosen between $\sqrt{n}$ to n/5
- **STEP 2**: Estimate the parameters of the distribution using MLE discussed earlier.
- **STEP 3**: Compute the expected frequency E_i for each class as:
- **STEP 4**: Combine categories so that $E_i \geq 5$ for each class. This is mandatory to ensure validity of the Chi-square test.
- **STEP 5**: Compute the test:

Chi-Square Test

- **STEP 6**: Get the degrees of freedom as

- **STEP 7**: Compare with Chi-Square Table or compute p-value to verify hypothesis

The hypothesis for the test is as follows:

- H_0: The data follows the specified distribution (Null Hypothesis)
- H_1: The data does not follow the specified distribution.

Given a significance value α,

$$\text{if the p-value} < \alpha \text{ reject the null hypothesis.}$$

Ex: Given $\alpha = 0.05$, check whether the number of calls received per minute at a call center follows a Poisson distribution.
Number of calls: 0,1,2,3,4,5 or more
Frequency: 10,26, 35, 25,15,9.
The Chi-square critical values are given as:
df: 1,2,3,4,5,6,7
$\chi^2(\alpha = 0.05)$: 3.841, 5.991, 7.815, 9.488, 11.070, 12.592, 14.067

Selection of Input Models in Absence of Data

If data is not available, some possible sources to obtain information about the process are:

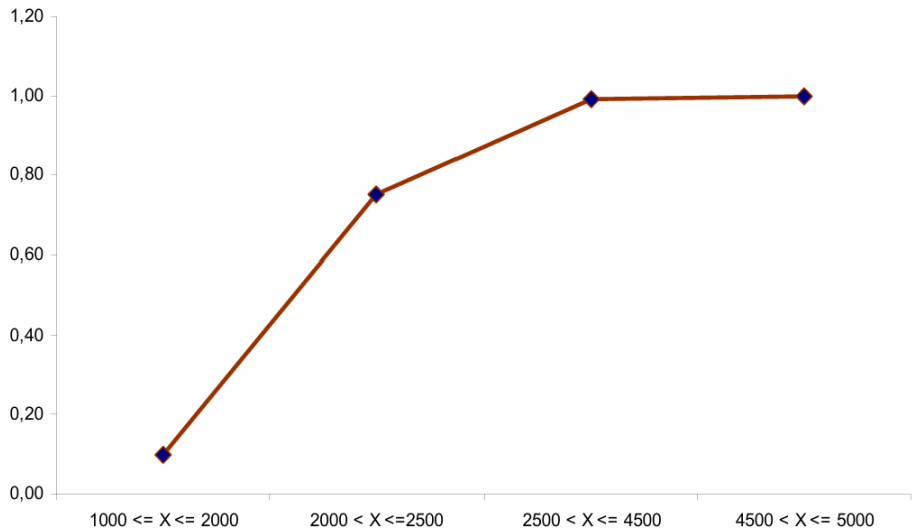
- **Engineering data**: Often product or process has performance ratings provided by the manufacturer or company rules specify time or production standards
- **Expert option**: People who are experienced with the process or similar processes, often, they can provide optimistic, pessimistic and most-likely times, and they may know the variability as well.
- **Physical or conventional limitations**: physical limits on performance, limits or bounds that narrow the range of the input process.
- **The nature of the process**

Ex: Example: Production planning simulation.

- Input of sales volume of various products is required, salesperson of product XYZ says that:
  - No fewer than 1000 units and no No fewer than 1000 units and no more than 5000 units will be sold.
  - Given her experience, she believes there is a 90% chance of selling more than 2000 units a 25% more than 2000 units, a 25% chance of selling more than 2500 units, and only a 1% chance of selling more than 4500 units.
- Translating these information into a cumulative probability of being less than or equal to those goals less than or equal to those goals for simulation input

Selection of Input Models in Absence of Data

i	Interval (Sales)	PDF	Cumulative Frequency, ci
1	$1000 \leq X \leq 2000$	0.1	0.10
2	$2000 < X \leq 2500$	0.65	0.75
3	$2500 < X \leq 4500$	0.24	0.99
4	$4500 < X \leq 5000$	0.01	1.00



Multivariate and Time Series Input Models

- As of now, any random variable considered was independent of any other variable in context of the problem.
- However, multiple variables which appear as input may be related

Multivariate and Time Series Input Models

- As of now, any random variable considered was independent of any other variable in context of the problem.
- However, multiple variables which appear as input may be related

Multivariate Series

A multivariate input model describes the joint behavior of two or more random variables. It captures not just their individual distributions, but also their dependence structure.

- If we are modeling the arrival time and service time in a call center, the random variables can be:
  1. Individually distributed: Exponential, Normal, etc.
  2. Jointly distributed: Correlated (e.g., busy times = longer service times)

Multivariate and Time Series Input Models

- As of now, any random variable considered was independent of any other variable in context of the problem.
- However, multiple variables which appear as input may be related

Multivariate Series

A multivariate input model describes the joint behavior of two or more random variables. It captures not just their individual distributions, but also their dependence structure.

- If we are modeling the arrival time and service time in a call center, the random variables can be:
  1. Individually distributed: Exponential, Normal, etc.
  2. Jointly distributed: Correlated (e.g., busy times = longer service times)

Time-Series

A time-series process refers to a sequence of data points or observations that are ordered in time. These observations typically occur at successive, evenly spaced time intervals.

- Stationary Time Series: A stationary process has constant statistical properties over time, such as a constant mean, variance, and autocorrelation.
- Non-Stationary Time Series: A non-stationary time series has properties (e.g., mean or variance) that change over time.

Covariance and Correlation

Consider a bi-variate linear model that describes the relationship between two random variables X_1 and X_2

The covariance between the variables can be given as:

Given the values of covariance obtained, the coefficient β can be related as

- If the variables are independent
- If the variables are positively correlated
- If the variables are negatively correlated

The correlation between X_1 and X_2 is given as

Input Models for Multivariate

For a given two random variables expressed as time-series (X_t, X_{t+h}),

- The expression $cov(X_t, X_{t+h})$ is the lag-h autocovariance.
- The expression $corr(X_t, X_{t+h})$ is the lag-h autocorrelation.

If the autocovariance value depends only on h and not on t, the time series is *covariance stationary*.

Ex: Let X_1 the average lead time to deliver and X_2 the annual demand for a product, be two random variables given as:
X_1 = 6.5, 4.3, 6.9, 6.0, 6.9, 6.9, 5.8, 7.3, 4.5, 6.3
X_2 = 103, 83, 116, 97, 112, 104, 106, 109, 92, 96
Find if the variables are correlated.

Input Models for Time Series

Consider, $X_1, x_2222, X_3, \ldots$ is a sequence of identically distributed, but dependent and covariance-stationary random variables, then the process can be represented as:

- Autoregressive model of order 1 : **AR(1)**
- Exponential Autoregressive model of order 1 : **EAR(1)**

Both have the characteristics that

Lag-h autocorrelation decreases geometrically as the lag increases, hence, observations far apart in time are nearly independent.

Autoregressive Model

It models a variable using its own past values. It assumes that the current value of a time series is linearly dependent on its previous values and a stochastic error term. The generic AR model of order p is represented as :

Consider a time series model represented via a normal distribution as:

Input Models for Time Series

If the initial value X_1 is chosen appropriately:

- $X_1, X_2, \ldots$ are normally distributed with
- The autocorrelation is given as:

The estimates of the parameters are given as: