Module 4

CHAPTER 8

Input Modeling

8.1 Introduction

- Input models represent the uncertainty in a stochastic simulation. Input modeling is the process/practice of selecting probability distributions (input models) to represent random input processes.
- Collecting data on the appropriate elements of the system of interest is one of the initial steps in successful input modelling
- For queuing system simulation, input models are the distributions of time of arrivals and service times.
- For an inventory system simulation, input models include distributions of demand (demand per unit time) and of lead – time.
- For reliability system simulation, input models include the distribution of time to failure of a component.
- The simulation analyst must understand that there is no "true" input model for any stochastic input; the main goal of analyst is to obtain an approximation that captures the key characteristics of the underlying input process.

8.2 Development of Useful Model of Input Data

MU - May 16

Q. Mention steps in input modeling. MU - May 05, May 06, May 08, Dec. 08

Q. What one used to obtain information about a process in the absence of input data ? MU - May 11, Dec. 11, May 13

Q. Explain steps involved in development of useful model of input data. MU - May 12, May 13, Dec. 13

There are four steps in the development of a useful model of input data and steps are

- Collect data from the real system

- Identify a probability distribution to represent the input process.
- Choose parameters for the distribution.
- Evaluate the chosen distribution and parameters for goodness of fit.

Data Collection : Data is collected from real system. Collecting data from real system require considerable amount of time and resources. When data is not available one can take expert opinion or can make educated guess based from knowledge of the process.

Identification of the Distributions : Identify a probability distribution to represent the input process. To identify this we first develop histogram/frequency distribution. Some frequency distributions provide good estimation.

Choose Parameters : Choose parameters for determining the specific instance of the distribution family. These parameters are estimated from the data.

Evaluation of Distributions and Parameters : The selected/indentified distribution and the associated parameters are evaluated for goodness of fit. They can be evaluated by graphical methods or statistical tests. If the test is not satisfied for the chosen distribution then choose different family of distribution and repeats the procedure.

8.3 Data Collection

MU - May 16

Q. How would you collect data to be used as input to simulation model ? MU - Dec. 05, May 08, Dec. 08

- It is one of the biggest tasks in solving a real problem.
- Though very important but it is hard to achieve.
- The data may be too scare or too abundant.
- According to GIGO (Garbage In Garbage Out) principle even when model structure is valid simulation results can be misleading, if the input data are inaccurately collected or inappropriately analyzed.
- The following suggestions that may enhance and facilitate data collection are :

Plan Ahead : Begin by a practice or pre – observing session, and watch for unusual circumstances. Devise forms for the same purpose. If possible video tape the system and extract the data later by viewing the data. If data has already been collected by someone else then allocate lot of time for converting data into usable format.

Analyze Data : Analyze data as it is being collected and check for adequacy. The data should be adequate to provide the distributions needed as input to the simulation. Recognize and discard all the useless and superfluous data.

Combine Homogeneous Data Sets : Check data for homogeneity over successive time periods or also during the same time period on successive days.

Data Censoring : The quantity of interest when not observed in its entirety is called as data censoring. It exposes the system to the danger of leaving out long process times.

Check for Relationships : Observe the variable to find out any kind of relationship between them. For this generally a scatter diagram can be built. Sometimes an eyeball scan of the same diagram can give us an idea regarding the relationship between variables of interest.

Check for Autocorrelation : Autocorelation is a relationship between values separated from each other by a given time lag. So sequence of observations may appear they are independents. They should be checked for auto correlation, to find any relationship exists. In simulation we can find out there is any relationship between successive customers or successive time periods. The Durbin–Watson statistic test is used to detect the presence of autocorrelation

- **Collect Input Data, Not Performance Data :** Be aware of collecting input data, we have to collect input data not a output data. This input data is mainly have qualities that are beyond the control of the system. In queuing system customer arrival time is input data.

8.4 Identifying Distribution

8.4.1 Histogram

- A histogram is a graph that shows the frequency, or the number of times, something happens within a specific interval.
- Histograms allow a visual interpretation of numerical data by indicating the number of data points that lie within a range of values, called a class.
- The shape of a distribution is determined by using frequency distribution (histogram).
- Followings are the steps to develop histogram :
  - i. Draw a horizontal line. This will be where we denote our classes.
  - ii. Place evenly spaced marks along this line that correspond to the classes.
  - iii. Label the marks so that the scale is clear and give a name to the horizontal axis.

  - iv. Draw a vertical line just to the left of the lowest class.
  - iv. Choose a scale for the vertical axis that will accommodate the class with the highest frequency.
  - v. Label the marks so that the scale is clear and give a name to the vertical axis.
  - vi. Construct bars for each class. The height of each bar should correspond to the frequency of the class at the base of the bar.

8.4.2 Selecting the Family of Distribution

- The main reason for making a histogram is to infer the known PDF or PMF.
- A family of distribution is selected based on: the context of the input variable and shape of the histogram.
- There are numerous probability distributions. We use the physical basis of the distribution as a guide, some of which are mentioned below :

i. Binomial	ii. Negative Binomial
iii. Poisson	iv. Normal
v. Lognormal	vi. Exponential
vii. Weibull	viii. Discrete or Continuous Uniform
ix. Triangular	x. Empirical
xi. Gamma	xii. Beta
xiii. Erlang.	

- It should also be kept in mind that there is no "true" distribution for any stochastic input process. The only goal is to obtain a good approximation.
- There are numerous probability distributions. We use the physical basis of the distribution as a guide, some of which are mentioned below :

  (i) Binomial : Models the number of successes in n trials, when trials are independent with common success probability, p; **E.g. :** Number of defective parts found in a lot of n parts.

  (ii) Negative Binomial : Models the number of trials, until the k^{th} success. **E.g. :** number of computer chips that must be inspected in order to find defective chips.

  (iii) Poisson : Models the number of independent events that occur in a fixed amount of time or space.

  E.g. : Number of customers coming to a bank during one hour.

(iv) Normal : Models the distribution of a process that is the sum of a number of components processes.

E.g. : Time taken to assemble a product that is a sum of the times required for each operation on it.

(v) Lognormal : Models the distribution of a process that is the product of a number of component processes.

E.g. : Rate of return on a compounded investment, which is the product of returns for the number of periods.

(vi) Exponential : Models the time between independent events, or a process time that is memory less. It is a highly variable distribution. It is sometime overused because it often leads to mathematically tractable models. In case the time between events is exponentially distributed then the number of events in fixed interval of time is poisson.

E.g. : The time between arrivals from a large population of potential customers who act independently of each other.

(vii) Weibull : Models the time to failure for components.

E.g. : The time to failure of a hard disk.

(viii) Discrete or Continuous Uniform : Models complete uncertainty. In this case all outcomes are equally likely. This distribution often is used inappropriately, when there are no data.

(ix) Triangular : Models a process when only the minimum, most-likely, and maximum values of the distribution are known.

(x) Empirical : An empirical distribution is based on actual observed data. It is used when there is no any appropriate theoretical distribution.

(xi) Gamma : This is an extremely flexible distribution used to model non-negative random variables that is bounded on one side.

(xii) Beta : This distribution is used to model bounded random variables.

(xiii) Erlang : Models processes that can be viewed as the sum of several exponentially distributed processes. This distribution is a special case of the gamma distribution.

E.g. : A computer network fails when a computer and two backup computers fail, and each has a time to failure that is exponentially distributed.

8.4.3 Quantile - Quantile Plot

- The construction of histograms allows us to select the family of distributions.
- But it does nothing to check the fit of the distribution.
- In case of less number of data points the histogram will face even grater problems.
- So, in order to avoid all these problems concerning the histogram, we make use of the quantile-quantile plots.
- Quantile-Quantile plot, q-q plot as it is called is a useful tool for evaluating distribution fit.
- If X is a random variable with CDF F, then the q-quantile of X is the γ such that

$$F(\gamma) = P(X \leq \gamma) = q, \quad \text{for } 0 < q < 1$$

- When F has an inverse,

$$\gamma = F^{-1}(q)$$

- Let $\{x_i, i = 1, 2, ..., n\}$ be a sample of data from X and $\{y_j, j = 1, 2, ..., n\}$ be the observations in ascending order. Then :

$$y_j \text{ is approximately } F^{-1}\left(\frac{j - 0.5}{n}\right)$$

where j is the ranking or the order number.

- The plot of y_j versus $F^{-1}[(j - 0.5/n)]$ is approximately a straight line if F is a member of an appropriate family of distributions.
- The line has slope 1 if F is a member of an appropriate family of distributions with appropriate parameter values.
- In case the assumed distribution is inappropriate, the points will deviate from a straight line, usually in a systematic manner.
- A **Q–Q plot** is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.
- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- q-q plot can also be used to check homogeneity.
- It can be used to check whether a single distribution can represent two sample sets :
- Given two random variables X and $x_1, x_2, ..., x_n$ and Z and $z_1, z_2, ..., z_n$.

8.5 Parameter Estimation

Parameter Estimation is the next step after selecting a family of distributions.

If observations in a sample of size n are $X_1, X_2, \ldots, X_n$ (discrete or continuous), the sample mean and variance are :

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$

If the data are discrete and have been grouped in a frequency distribution :

$$\bar{X} = \frac{\sum_{j=1}^{n} f_j X_j}{n} \qquad S^2 = \frac{\sum_{j=1}^{n} f_j X_j^2 - n\bar{X}^2}{n-1}$$

Where f_j is the observed frequency of value X_j

When raw data are unavailable (data are grouped into class intervals), the approximate sample mean and variance are :

$$\bar{X} = \frac{\sum_{j=1}^{n} f_j m_j}{n} \qquad S^2 = \frac{\sum_{j=1}^{n} f_j m_j^2 - n\bar{X}^2}{n-1}$$

Where f_j is the observed frequency in the j^{th} class interval, m_j is the midpoint of the j^{th} interval, c is the number of class intervals

A parameter is an unknown constant, but an estimator is a statistic.

Distribution	Parameter	Estimator
Poisson	α	$\hat{\alpha} = \bar{X}$
Exponential	λ	$\hat{\lambda} = \frac{1}{\bar{X}}$
Gamma	β, θ	$\hat{\beta}, \hat{\theta} = \frac{1}{\bar{X}}$
Normal	μ, σ^2	$\hat{\mu} = \bar{X}, \hat{\sigma}^2 = S^2$
Lognormal	μ, σ^2	$\hat{\mu} = \bar{X}, \hat{\sigma}^2 = S^2$

8.6 Goodness of Fit Test

Q. Which tests are used to test "Goodness of fit". Describe anyone of them.
MU - Dec 05, May 08, May 09

Q. What do you understand by "Goodness of Fit Test"? Write the procedure for the same.
MU - May 15

- For a set of given data it is important to check if the data matches the chosen distribution.
- For this various graphical as well as analytical methods are used.
- Most of the techniques require that the type of distribution function of X is known and either its parameters are being estimated or a hypothesis concerning its parameters is being tested.
- Goodness-of-fit tests are one of these analytical methods that are used to check for the fit of the distribution.
- Goodness-of-fit tests do not give a definite answer with respect to the real distribution.
- These tests however provide helpful guidance for evaluating the suitability of potential input model.
- No single correct distribution in a real application exists
- For a correct distribution it is very important to consider the effect of the sample size on the results :
  - If very little data are available (small number of data points), it is unlikely to reject any candidate distributions.
  - If a lot of data are available (large number of data points), it is likely to reject all candidate distributions.
- We need to be aware of mistakes that can occur during decision-making :
  - Type I Error : α
  - Type II Error : β

Statistical Decision	State of the null Hypothesis	
	H_o True	H_o False
Reject H_o	Type I error	Correct
Accept H_o	Correct	Type II error

- The two most important tests used for testing the fitness of the distribution chosen are as :
  - Kolmogorov-Smirnov test
  - Chi-square test.

8.6.1 Chi-square Test

- It is one of the oldest and the most frequently used tests.
- It compares the histogram of the data to the shape of the candidate density or mass function.
- It is valid for large sample sizes when parameters are estimated by maximum likelihood.
- This test works well for both discrete as well as continuous distributional assumptions.
- In case of discrete data we have natural intervals in which we have to group the observed values.
- It follows the following algorithm :
  - (i) Define the hypothesis for the chi-square test :
    - H_0 : The random variable X, conforms to the distributional assumption with the parameter (s) given by the estimate (s).
    - H_1 : The random variable X does not conform
  - (ii) Arrange the n observations into sets of k class intervals or cells.
  - (iii) Compute the expected frequency for each class interval using the formula $E_i = nP_i$, where P_i is the theoretical, hypothesized probability.
  - (iv) The test statistic is computed by :

$$\chi_0^2 = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i}$$

where O_i is the observed number in the i^{th} class interval

E_i is the expected number in the i^{th} class interval

k is the number of classes

  - (v) Determine the critical value for the test with the specified level of significance and the degrees of freedom k – s – 1 where 'k' is number of intervals and 's' is the number of estimated parameters.
  - (vi) If the observed chi-square χ_0^2, based on the data is greater than the critical theoretical chi-square table value, χ_0^2, k – s – 1 then the null hypothesis may be rejected.
- If the distribution tested is discrete and combining adjacent cell is not required (so that E_i > minimum requirement)

- Each value of the random variable should be a class interval, unless combining is necessary.

$$P_i = p(x_i) = P(X = x_i)$$

Ex. 8.6.1 : Records related to the monthly number of job-related injuries at a company were being studied. The values for the past 100 months were as follows :

Injuries Per Month	Frequencies of Occurrence
0	35
1	40
2	13
3	6
4	4
5	1
6	1

Apply Chi-Square test to test these data to test hypothesis that the underlying distribution is poisson, where $\alpha = 0.05$, $\chi^2_{0.5,2} = 5.99$. MU - May 16

Soln :

The hypothesis is :

H_0 : The underlying distribution is poisson

H_1 : The underlying distribution is not poisson

But the estimator of poisson distribution $\hat{\lambda} = \bar{X}$

Where,

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n} = \frac{\sum_{i=1}^{6} f_i X_i}{100} = \frac{111}{100} = 1.11$$

$$\therefore \hat{\lambda} = \bar{X} = 1.11$$

Now, the PMF of poisson distribution

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$$

$$= 0, \text{otherwise}$$

The PMF and test statistics is computed in the following table :

x_i	O_i	$P_i = \frac{e^{-1.11} 1.11^x}{x_i!}$	$E_i = np_i$	$\chi_0^2 = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i}$
0	35	0.3296	32.96	0.126
1	40	0.3658	36.58	0.320
2	13	0.2030	20.30	2.627
3	6	0.0751	07.51	
4	4	0.0209	02.09	
5	1	0.0046	0.46	
≥ 6	1	0.0010	0.087	0.333
Totals	**100**			

Notice that we have grouped cells i = 3, 4, 5 ≥ 6 together into a single cell with $O_i = 12$ and $E_i = 10.13$

The test statistics

$$\chi_0^2 = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i} = 3.416$$

The critical value for the specified significance level $\alpha = 0.05$ with degrees of freedom $k - s - 1 = 4 - 1 - 1 = 2$

Where k is the number of intervals

$$\chi^2_{0.05,2} = 5.99$$

$$\chi_0^2 = 3.416 < \chi^2_{0.05,2} = 5.99$$

Therefore, we do not reject H_0.

8.6.1.1 Chi-square Test with Equal Probabilities

- If the distribution tested is continuous, then class intervals should be of equal probability rather than equal width.
- One advantage of having equiprobable cells is that each cell receives that same weighing division factor np_i.
- The important issue here is to determine the number of intervals k.
- Determining a number of cells and cell sizes to collect the data is an art.

As each k interval has equal probability so $P_i = \frac{1}{k}$.

Since $E_i = np_i \geq 5$

Substituting $Pi = \frac{1}{k}$ in the above equation leads to

$$\frac{n}{k} \geq 5$$

Or $$k \leq \frac{n}{5}$$

- This equation leads to determine the recommended maximum number of class intervals.
- If the distribution tested is continuous we have :

$$P_i = \int_{a_i - 1}^{a_i} f(x)\,dx = F(a_i) - F(a_i - 1)$$

Where $a_i - 1$ and a_i are the endpoints of the i^{th} class interval

f(x) is the assumed pdf, F(x) is the assumed cdf

- Recommended number of class intervals (k) :

Sample Size, n	Number of Class Intervals, k
20	Do not use the Chi-Square test
50	5 to 10
100	10 to 20
> 100	$n^{1/2}$ to n/5

Drawbacks of Chi-Square Test

- It requires the data to be placed in class intervals, changing the number of class and interval width affects the result of the test.
- Grouping of the data may affect accept-reject decision about the hypothesis.
- The estimation of the parameters from the data results in decrease in the degree of freedom.
- Test statistic is known only approximately and the power of test statistic is at time rather low.
- Valid for large sample size only.

Ex. 8.6.2 : The time required to 50 different employees to compute and recorded number of hours worked during the week was measured with the following results in minutes which is given as follows :

Employee	Time (minutes)	Employee	Time (minutes)
1	1.88	26	0.04
2	0.54	27	1.49
3	1.90	28	0.66
4	0.15	29	2.03
5	0.02	30	1.00
6	2.81	31	0.39
7	1.50	32	0.34
8	0.53	33	0.01
9	2.62	34	0.10
10	2.67	35	1.10
11	3.53	36	0.24
12	0.53	37	0.26
13	1.80	38	0.45
14	0.79	39	0.17
15	0.21	40	4.29
16	0.80	41	0.80
17	0.26	42	5.50
18	0.63	43	4.91
19	0.36	44	0.35
20	2.03	45	0.36
21	1.42	46	0.90
22	1.28	47	1.03
23	0.82	48	1.73
24	2.16	49	0.38
25	0.05	50	0.48

Using Chi-Square test, test the hypothesis that these service times are exponentially distributed. Let the number of class interval be k = 6. Use level of significance $\alpha = 0.05$.

Soln :

The hypothesis is :

H_0 : The underlying distribution is exponential

H_1 : The underlying distribution is not exponential

The estimator of exponential distribution

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

Where,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{60.3}{50} = 1.206$$

Hence,

$$\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{1.206} = 0.829$$

As per the assumption of continuous distribution we have to perform the Chi-Square test with intervals of equal probability.

Finding the end points of the class intervals;

Now, k = 6 i.e. probability of each interval

Hence, p = 0.166 (1/6)

Now the end point of each interval are computed from cdf of the exponential distribution.

Let a_i be the end point of each interval then,

$$F(a_i) = 1 - e^{-\lambda a_i}$$

Since $F(a_i)$ is cumulative area from 0 to a_i

$$F(a_i) = ip$$

$$ip = 1 - e^{-\lambda a_i}$$

$$e^{-\lambda a_i} = 1 - ip$$

Taking log on both sides

$$\ln(e^{-\lambda a_i}) = \ln(1 - ip)$$

$$-\lambda a_i = \ln(1 - ip)$$

$$a_i = \frac{-1}{\lambda} \ln(1 - ip) \text{ where } i = 0, 1, \ldots k$$

Now using the above equation for a_i, we can find the end points for the interval for the value i = 0, 1, ...6,

$$a_0 = \frac{-1}{\lambda} ln\,(1-0) = 0$$

$$a_1 = \frac{-1}{0.829} ln\,[1-(1)(0.1666)] = 0.22$$

$$a_2 = \frac{-1}{0.829} ln\,[1-(2)(0.1666)] = 0.49$$

$$a_3 = \frac{-1}{0.829} ln\,[1-(3)(0.1666)] = 0.839$$

$$a_4 = \frac{-1}{0.829} ln\,[1-(4)(0.1666)] = 1.33$$

$$a_5 = \frac{-1}{0.829} ln\,[1-(5)(0.1666)] = 2.174$$

$$a_6 = \frac{-1}{0.829} ln\,[1-(6)(0.1666)] = \infty$$

So the interval are [0, 0.22], [0.22, 0.49], [0.49, 0.839], [0.839, 1.33], [1.33, 2.174] and [2.174, ∞].

Applying Chi-Square test, we get

i	Class Interval	Observed Frequency	Expected Frequency $E_i = np_i$	$\chi_0^2 = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i}$
1	[0, 0.22]	8	8.33	0.0133
2	[0.22, 0.49]	11	8.33	0.8533
3	[0.49, 0.839]	9	8.33	0.0533
4	[0.839, 1.33]	5	8.33	1.3333
5	[1.33, 2.174]	10	8.33	0.3333
6	[2.174, ∞]	7	8.33	0.2133

The test statistics is given by :

$$\chi_0^2 = \frac{\sum_{i=1}^{k} (O_i - E_i)^2}{E_i} = 2.7998$$

The critical value for the specified significance level $\alpha = 0.05$ with degrees of freedom $k - s - 1 = 6 - 1 - 1 = 4$

$$\chi^2_{0.05,4} = 9.49$$

$$\chi_0^2 = 2.7998 < \chi^2_{0.05,4} = 9.49$$

H_0 is not rejected.

8.6.2 Kolmogorov – Smirnov Test

- Kolmogorov - Smirnov test came up because of the various drawback of Chi-Square test.
- So, it is a much stronger test the Chi-Square test.
- This test basically formalizes the idea behind examining the q-q plot.
- This test does not require any grouping of data unlink Chi-Square test.
- The test compares the continuous cdf, F(x), of the hypothesized distribution with the empirical cdf, $S_N(x)$, of the N sample observations.
- It is based on the maximum difference statistics.

$$D = \max |F(x) - S_N(x)|$$

- This test is particularly useful when :
  - Sample sizes are small.
  - No parameters have been estimated from the data.
- When parameter estimates have been made the critical values are biased, too large. This leads to smaller type I error than those specified.
- Anderson-Darling test is similar to Kolmogorov-Smirnov test. The Anderson-Darling test to compare the fit of an observed cumulative distribution function to an expected cumulative distribution function. This test gives more weight to the tails than the Kolmogorov-Smirnov test. Anderson-Darling test is much more sensitive to the tails of a distribution.

Ex. 8.6.3 : The highway between Atlanta, Georgia and Athens, Georgia, has a high incidence of accidents along its 100 km. Public safety officers say that the occurrence accident along the highway is randomly (uniformly) distributed, but the news media says otherwise.

The Georgia department of public safety published records for the month of September.

These records indicated the point at which 30 accidents occurred, as follows :

88.3	40.7	36.3	27.3	36.3
91.7	67.3	7.0	45.2	23.3
98.8	90.1	17.2	23.7	97.4
32.4	87.8	69.8	62.6	99.7
20.6	73.1	21.6	6.0	45.3
76.6	73.2	27.3	87.6	87.2

Use the Kolmogorov – smirnov test to determine whether the distribution is uniformly distributed given, $D_{0.05, 30} = 0.24$.

Soln :

The data points as which 30 accidents involving an injury of death occurred.

The data points represent the distance from city limits of Georgia.

The hypothesis is :

H_0 : The data point is uniformly distributed.

H_1 : the data points are not uniformly distributed.

(i) Normalize the data points to (0, 1) for k-s test. The results are in the column $R_{(i)}$ of the table given below.

(ii) Apply the k-s test.

(a) Arrange the data point in increasing order and compute D^+ and D^-.

i	$R_{(i)}$	$\frac{i}{N}$	$\frac{i}{N} - R_{(i)}$	$\frac{(i-1)}{N}$	$R_{(i)} - \frac{(i-1)}{N}$
1	0.06	0.033	–	0	0.06
2	0.07	0.066	–	0.033	0.037
3.	0.172	0.1	–	0.066	0.106
4	0.206	0.133	–	0.1	0.106
5	0.216	0.166	–	0.133	0.083
6	0.233	0.2	–	0.166	0.067
7	0.237	0.233	–	0.2	0.037
8	0.273	0.237	–	0.233	0.04
9	0.273	0.273	0.027	0.266	0.007
10	0.324	0.273	0.009	0.3	0.024

i	$R_{(i)}$	$\frac{i}{N}$	$\frac{i}{N} - R_{(i)}$	$\frac{(i-1)}{N}$	$R_{(i)} - \frac{(i-1)}{N}$
11	0.363	0.324	0.003	0.333	0.03
12	0.368	0.363	0.032	0.366	0.002
13	0.407	0.368	0.026	0.4	0.007
14	0.452	0.407	0.014	0.433	0.019
15	0.453	0.452	0.047	0.466	–
16	0.626	0.453	–	0.5	0.126
17	0.673	0.626	–	0.533	0.14
18	0.698	0.673	–	0.566	0.132
19	0.731	0.698	–	0.6	0.131
20	0.732	0.731	–	0.633	0.099
21	0.766	0.732	–	0.66	0.1
22	0.872	0.733	–	0.700	0.172
23	0.876	0.766	–	0.733	0.143
24	0.878	0.878	–	0.766	0.112
25	0.883	0.883	–	0.8	0.083
26	0.901	0.901	–	0.833	0.068
27	0.917	0.917	–	0.866	0.051
28	0.974	0.974	–	0.9	0.074
29	0.988	0.988	–	0.933	0.055
30	0.997	0.997	0.003	0.96	0.031

$$D^+ = \max_{1 \le i \le 30} \left\{ \frac{i}{N} - R_{(i)} \right\} = 0.047$$

$$D^- = \max_{1 \le i \le 30} \left\{ R_{(i)} - \frac{(i-1)}{N} \right\} = 0.172$$

(a) $D = \max (D^+, D^-) = \max (0.047, 0.172) = 0.172$

(b) Obtain the critical value D_α for the specified significance level $\alpha = 0.05$ and $N = 30$ from $D_{0.05, 30} = 0.24$

(c) Since $D = 0.172 < D_{0.05, 30} = 0.24$. Hence H_0 is not rejected.

8.6.3 P-values and Best-fits

- A goodness of fit test uses a significance level always to test the null hypothesis.
- Significance level is the probability of falsely rejecting H_0, the random variable conforms to the distributional assumption.
- The most commonly used levels are 0.1, 0.05, 0.01.
- Initially a small set of standard values was used to produce tables of critical values.
- Now there are software packages that compute a p-value.
- p-value is the significance level at which one would just reject H_0 for the given test statistic value.
- The p-value corresponds to the fit of the distribution.
- A large p-value tends to indicate a good fit (we would have to accept a large chance of error in order to reject). While a small p-value suggests a poof fit (to accept we would to consider the no risk case).
- But now many input modeling software implements the best-fit option so it is not suggested not to use the above approach of p-value.
- By evaluating all feasible models, the best fit option recommends as input model to user. For evaluation purpose it considers data is continuous or discreet, data is bounded or unbounded etc.
- Things to be cautious about while using p-value and best-fit approach :
  - Software may not know about the physical basis of the data, distribution families it suggests may be inappropriate.
  - Automated best-fit procedures tend to choose the more flexible distribution since more flexibility allows close conformance to the data which does not necessarily lead to the most appropriate input model.
  - p-value is just a summary measure and does not say much about where the lack of fit occurs instead a graphical tool can work better in this case.

8.7 Selecting Input Models without Data

There are a number of ways to select the input model even if there is no data available. Some possible sources to obtain information about the process are :

- ***Engineering Data :*** Often product or process has performance rating provided by the manufacturer or company rules specify time or production standards. E. g : A bulb can work for 8 hours at a stretch.

- **Expert Option :** We should talk to and seek opinions of 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 as well.
- **Physical or Conventional Limitations :** Physical limits on performance, limits or bounds generally narrow the range of the input process.
- **The Nature of the process :** The description of the distributions can be used to justify the input model selected.
- When the data is not available uniform, beta and triangular distributions are used as input distribution (models) because of their characteristics.
- The uniform distribution can be a bad choice because it's the upper and the lower bounds are rarely fits as the central values in the real processes.
- The triangular distributions should be used when in addition to the lower and upper bounds even the most likely value is given.
- If beta distribution is being used, then we should be sure to plot the density function of the selected distributions.

8.8 Multivariate and Time Series Models

Q. Discuss multivariate and time series models. **MU - May 06**

Q. What is time-series input model ? Explain AR (1) and EAR (1) model. **MU May 11, Dec. 13**

- The random variables are considered to be independent of any other variables within the context of the problem.
- However, variables may be related.
- If they appear as input, the relationship should be investigated and taken into consideration.
- **Covariance and Correlation**
  - Measures the linear relationship between the variables.
- Various models are :
- **Multivariate Input Models**
  - Fixed, finite number of random variables.
  - For example, lead time and annual demand for an inventory model
  - An increase in demand result in lead time increase, hence variables are dependent.

Multivariate : dependent finite variables

- **Time – series Input Models**
  - Infinite sequence of random variables
  - For example, time between arrivals of orders to buy and sell stocks
  - Buy and sell orders tend to arrive in burst, hence times between arrivals are dependent.

8.8.1 Covariance and Correlation

Q. Explain covariance and correlation. **MU - Dec. 11**

- Covariance between X_1 and X_2 is defined as :

$$\text{cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1, X_2) - \mu_1 \mu_2$$

- Depending on the value of cov (X1, X2) we have

$$\text{Where cov}(X_1, X_2) \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases} \Rightarrow \beta \begin{cases} = 0 \\ < 0 \\ > 0 \end{cases} \qquad -\infty < \text{cov}(X_1, X_2) < \infty$$

- The correlation standardizes the covariance to be between – 1 to 1

$$\rho = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

- As in case of covariance, different correlation values give different β values :

$$\text{corr}(X_1, X_2) = 0 \text{ implies } \beta = 0$$
$$< 0 \text{ implies } \beta < 0$$
$$> 0 \text{ implies } \beta > 0$$

- Covariance and correlation are measure of the linear dependence between X_1 and X_2.
- The model that describes the relationship between X_1 and X_2 is given by :

$$(X_1 - \mu_1) = \beta (X_2 - \mu_2) + \varepsilon$$

Where ε is the random variable with mean 0 and is independent of X_2

There are following conditions depending on the value of β. They are as follows :

- $\beta = 0$, X_1 and X_2 are statistically independent
- $\beta > 0$, X_1 and X_2 tend to be above or below their means together.
- $\beta < 0$, X_1 and X_2 tends to be on opposite sides of their means.

- The closer the value of ρ to – 1 or 1, the stronger the linear relationship is between X_1 and X_2.

- Now assume a sequence of random variables $X_1, X_2, X_3 \ldots$ that are identically distributed but could be dependent.
- Such a sequence is called as a time series and to cov (X_t, X_{t+h}) and corr (X_t, X_{t+h}) as the lag-h autocovariance and lag-h autocorrelation.
- If the value of autocovariance depends only on h and not t, then we say that the time series is covariance stationary.
- These are represented as :

$$\rho_h = \text{corr}(X_t, X_{t+h})$$

Ex. 8.8.1 : The following data was available for the past 10 years on demand and lead time

Lead time	6.5	4.3	6.9	6.0	6.9	6.9	5.8	7.3	4.5	6.3
Demand	103	83	116	97	112	104	106	109	92	96

Estimate correlation and covariance.

Soln :

We know that

$$\bar{X} = \frac{\sum_{j=1}^{n} X_{ij}}{n}$$

Hence, using this formula we get, $\bar{X}_1 = 6.14$ and $\bar{X}_2 = 101.80$

Also, $\hat{\sigma}_1 = \sqrt{\frac{\sum_{j=1}^{n} X_{1j}^2 - n\bar{X}_1^2}{n-1}}$ and $\hat{\sigma}_2 = \sqrt{\frac{\sum_{j=1}^{n} X_{2j}^2 - n\bar{X}_2^2}{n-1}}$

Using these formulae, we get $\hat{\sigma}_1 = 1.02$ and $\hat{\sigma}_2 = 9.93$

To estimate correlation :

$$\sum_{j=1}^{10} X_{1j} X_{2j} = 6328.5$$

Hence, covariance is given by :

$$\hat{\text{cov}} = \frac{[6328.5 - (10)(6.14)(101.80)]}{10 - 1} = 8.66 \text{ and}$$

$$\hat{\rho} = \frac{8.66}{(1.02)(9.93)} = 0.86$$

Here in this model lead time and demand are strongly dependent. Also, since the demand is a discrete valued quantity, so the continuous normal distribution is at best an approximation. The following algorithm is used to generate bivariate normal random variables:

- Generate Z_1 and Z_2, independent standard normal random variables.
- Set $X_1 = \mu_1 + \sigma_1 Z_1$
- Set $X_2 = \mu_2 + \sigma_2 [\rho Z_1 + \sqrt{(1-\rho^2)}\, Z_2]$

8.8.2 Multivariate Inputs Model

MU - May 16

Q. When will you use AR(1) and EAR(i) model ? MU - Dec. 04, May 08, May 09

Q. Explain Time series input model. MU - May 10, May 11, May 14, Dec. 14, May 15

- If X_1 and X_2 are normally distributed, dependence between them can be modeled by the bivariate normal distribution with μ_1, μ_2, σ_1^2, σ_2^2 and correlation ρ.
- To estimate μ_1, μ_2, σ_1^2, σ_2^2 we use "Parameter Estimation".
- To estimate ρ, suppose we have n independent and identically distributed pairs (X_{11}, X_{21}), (X_{12}, X_{22}), ... (X_{1n}, X_{2n}).
- Then the sample covariance is :

$$\hat{cov}(X_1, X_2) = \frac{1}{n-1} \sum_{j=1}^{n} (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2)$$

- The sample correlation is :

$$\hat{\rho} = \frac{\hat{cov}(X_1, X_2)}{\hat{\sigma}_1 \hat{\sigma}_2}$$

Where $\hat{\sigma}_1$ are $\hat{\sigma}_2$ sample variances.

Time series input models

- If $X_1, X_2, X_3, \ldots$ is a sequence of identically distributed, but dependent and covariance stationary random variables, then we can represent the process as follows :
  - Autoregressive order -1 model, AR (1)
  - Exponential autoregressive order-1 model, EAR (1)
- Both have the characteristics that :
  - $\rho_h = corr(X_t, X_{t+h}) = \rho^h$, for h = 1, 2, ...
  - Lag-h autocorrelation decreases geometrically as the lag increases ; hence, observations far apart in time are nearly independent.

AR(1) model

- Consider the time series model :

$$X_t = \mu + \phi (X_{t-1} - \mu) + \varepsilon_t, \text{ for } t = 2, 3, \ldots$$

Where $\varepsilon_2, \varepsilon_3, \ldots$ are independent and identically (normally) distributed with $\mu_\varepsilon = 0$ and variance $= \sigma_\varepsilon^2$

- If initial value X_1 is chosen appropriately, then :
  - $X_1, X_2, \ldots$ are normally distributed with mean $= \mu$, and variance $= \sigma^2 / (1 - \phi^2)$
  - Autocorrelation $\rho_h = \phi^h$
- To estimate ϕ, μ, σ_ε^2 :

$$\hat{\mu} = \bar{X} \qquad \hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2 (1 - \hat{\phi}^2) \qquad \phi = \frac{\hat{cov}(X_t, X_{t+1})}{\hat{\sigma}^2}$$

Where $\hat{cov}(X_t, X_{t+1})$ is the lag-1 autocovariance

- The following algorithm generates a stationary AR (1) time series, given all the parameter values :

(i) Generate X_1 from the normal distribution with mean μ and variance $\frac{\hat{\sigma}_\varepsilon^2}{(1 - \hat{\phi}^2)}$ set t = 2

(ii) Generate ε_t from the normal distribution with mean 0 and variance σ_ε^2.

(iii) Set $X_t = \mu + \phi (X_{t-1} - \mu) + \varepsilon$

(iv) Set t = t + 1 and go to step (ii).

EAR (1) model

- Consider the time series model :

$$X_t = \begin{cases} \phi X_{t-1}, & \text{with probablity } \phi \\ \phi X_{t-1} + \varepsilon_t, & \text{with probablity } 1 - \phi \end{cases} \quad \text{for } t = 2, 3, \ldots$$

Where $\varepsilon_2, \varepsilon_3, \ldots$ are independent and identically (normally) distributed with $\mu_\varepsilon = 1 / \lambda$ and variance $0 \le \phi < 1$

- If X_1 is chosen appropriately, then :
  - $X_1, X_2, \ldots$ are exponentially distributed with mean $= 1 / \lambda$
  - Autocorrelation $\rho_h = \phi^h$
- Only positive correlation can be represented using this model.

- To estimate ϕ, λ we have :

$$\hat{\lambda} = 1/\bar{X}, \hat{\phi} = \hat{\rho} = \frac{\hat{cov}(X_t, X_{t+1})}{\hat{\sigma}^2}$$

Where $\hat{cov}(X_t, X_{t+1})$ is the lag-1 autocovariance.

- The following algorithm generates a stationary EAR (1) time series, given the parameter values :

(i) Generate X_t from the exponential distribution with mean $= 1/\lambda$. Set $t = 2$

(ii) Generate U from the uniform distribution on [0, 1]. If $U \leq \phi$, then set $X_t = \phi X_{t-1}$ Otherwise, generate ε_t from the exponential distribution with mean $1/\lambda$ and set

$$X_t = \phi X_{t-1} + \varepsilon_t$$

(iii) set $t = t + 1$ and go to step (ii).

Review Questions

Q. 1 Discuss the steps involved in development of a model of input data.

Q. 2 Mention steps in input modeling.

Q. 3 How would you collect input data ?

Q. 4 How would you collect data to be used as input to simulation model ?

Q. 5 What do you understand by "Goodness of fit test"? Write the procedure for the same.

Q. 6 Which tests are used to test "Goodness of fit" Describe anyone of them.

Q. 7 Discuss Multivariate and Time Series Models.

Q. 8 Explain time series input models with suitable example.

Q. 9 When will you use AR(1) and EAR(i) model ?

Q. 10 Explain quantile - quantile plot and state its use.

Q. 11 Mention the steps in input modeling. How would you collect input data?

Q. 12 When will you use AR (1) and EAR (I) model? What do you understand by covariance and correlation ?

Q. 13 Which tests are used to test goodness of fit? Describe anyone of them.

8.9 University Questions and Answers

May 2010

Q. 1 *Explain and give* the algorithms to generate the AR(1) and EAR(1) time series models. **(Section 8.8.2)** **(10 Marks)**

May 2011

Q. 2 Explain Input modeling in detail. **(Section 8.2)** **(10 Marks)**

Q. 3 Explain Multivariate Input Models. **(Section 8.8)** **(10 Marks)**

Dec. 2011

Q. 4 Explain the steps in the development of a model of input data. **(Section 8.2)** **(4 Marks)**

Q. 5 Explain the AR(1) time series model along with the algorithm. **(Section 8.8.2)** **(5 Marks)**

Q. 6 Explain covariance and correlation. **(Section 8.8.1)** **(5 Marks)**

May 2012

Q. 7 Explain steps involved in development of useful model of input data. **(Section 8.3)** **(10 Marks)**

May 2013

Q. 8 What one used to obtain information about a process in the absence of input data ? Explain data collection for input modeling. **(Sections 8.2 and 8.3)** **(10 Marks)**

Dec. 2013

Q. 9 Explain data collection and analysis for input modeling. **(Section 8.3)** **(10 Marks)**

Q. 10 Explain data collection and analysis for input modeling. **(Section 8.3)** **(10 Marks)**

Q. 11 What is time-series input model ? Explain AR (1) and EAR (1) model. **(Section 8.8)** **(10 Marks)**

May 2014

Q. 12 Explain Time series input model. **(Section 8.8.2)** **(5 Marks)**

Dec. 2014

Q. 13 Explain time-series model. **(Section 8.8.2)** **(5 Marks)**

May 2015

Q. 14 What do you understand by "Goodness of Fit Test"? Write the procedure for the same. **(Section 8.6)** **(10 Marks)**

Q. 15 Explain Time series Model. **(Section 8.8.2)** **(5 Marks)**

May 2016

Q. 1(c) Explain data collection and analysis in input modeling. **(5 Marks)**

(Ans. : Refer sections 8.2 and 8.3)