UNIT 6. INPUT MODELING

Input models are the driving force for a simulation. Generally, input models are the distributions of inter-arrival time, service time, demand time, lead (delay in supply) time, time to failure of a component etc. Choosing an appropriate distribution for input data is a challenge in any simulation model development. There are four steps in the development of a useful model of input data:

1. **Collect data from the real system of interest:** This requires huge amount of time and resources. In some situations, it is not possible to collect data, especially in a short duration. When data are not available, expert opinion and knowledge of the process must be made to make some valid guesses.
2. **Identify a probability distribution to represent the input process:** When data are available, the probability distribution can be identified by developing histogram of the data. With the help of histogram and structural knowledge of the process, a family of well-known distributions is chosen.
3. **Choose parameters that determine a specific instance of the distribution family:** When data are available, the parameters can be estimated from the data.
4. **Evaluate the chosen distribution and the associated parameters for goodness of fit:** To check whether the distribution identified based on the input really resembles the original model, the goodness of fit is used. It uses graphical or statistical tests. Most widely used tests are Chi-Square test and Kolmogorov-Smirnov tests. If chosen distribution is not a good approximation of the data, the analyst has to go to 2nd step and to choose different distribution. The procedure is repeated till a good fit between assumed distribution and the collected data is found. Sometimes, the empirical form of the distribution also can be used.

Nowadays, the softwares are available for accomplishing the steps 2, 3 and 4. Still, it is advised to understand the softwares for better model preparation. But, no software is available of input modeling (step 1) when there is a relationship between two or more variables of interest or when data are not available.

6.1 DATA COLLECTION

Data collection is one of the biggest tasks in solving a real problem and also in simulation. The focus of input modeling is on the statistical aspects of fitting probability distributions to data. These distributions will later provide the inputs to the simulation. But, for fitting the distribution, there need to be accurate and relevant data available and it should be well understood. Such data are often handed over to input modeling software, which will fit some probability model to the data. But, there is no guarantee that the probability distribution suggested by the software is really fitting the actual data. Some of the drawbacks of this approach are discussed below –

1. **Stale Data :** A simulation study always tries to reduce the time in every step. That does not mean that, an old data (stale data) can be used for the study. For example, to study the current health status of a patient, it is not advisable to refer his/her 10 year old medical report. Instead, it would be ideal to use the reports of recent one year or so.

Thus, out-of-date study should not be used. The possible feeding of such old data to the input modeling software definitely will not result into appropriate distribution.

2. **Unexpected Data:** Sometimes, a software may generate unexpected data. For example, the time required to pass through a security gate may be showed up as negative. But, it is obvious that, time cannot be negative, rather, it may be very small in terms of minutes. Hence, a keen observation is required on the data generated by the software.
3. **Time-varying Data:** Based on the data available, some distribution may be decided by the software. For example, number of calls received at a call center per day, can be treated as Poisson distribution. But, it is obvious that, the number of calls will not be equal if it is divided into per hour-basis. Because, normally, early morning hours, calls will be less. As the time passes by in a day, the calls may increase. Again, later in the night, number of calls will be less. Thus, the analyst has to verify whether the distribution suggested by the software really depicts the actual scenario or not.
4. **Dependent Data:** The data available may not always be independent. For example, a sale of cool-drinks is seasonal. Hence, if the distribution is fit based on an average sale throughout the year, it may not give the clear idea about the actual sales.

All these situations illustrates that just having data is not enough for effective input modeling. The data should be sufficiently new and clean (not containing errors). There are methodologies for cleaning the data as well. Following are few suggestions for data collection:

1. A useful expenditure of time is in planning: This could be done through pre-observation of the session. It may need to undergo several modifications. Though, watching the circumstances will definitely lead to proper data collection.
2. Try to analyze the data as they are being collected. Check whether the data being collected are adequate to identify the distribution needed for the simulation. If any set of data are found to be useless, then itself, it can be discarded instead of storing it further.
3. Try to combine homogeneous data sets. Check data for homogeneity in successive time periods and during the same time period on successive days.
4. Be aware of the possibility of data censoring, in which a quantity of interest is not observed in its entirety. This problem occurs when the analyst is interested in the time required to complete the process, but the process begins prior to or finishes after the completion of observation period.
5. To discover whether there is a relationship between two variables, build a scatter diagram.
6. Consider the possibility that a sequence of observations that appear to be independent actually has autocorrelation.
7. Keep in mind the difference between input data and output data, and be sure to collect input data.

6.2 IDENTIFYING THE DISTRIBUTION WITH DATA

Here, we discuss methods for selecting families of input distributions when data are available. We assume that the data are independent and identically distributed. The specific distribution within a family is then specified by estimating its parameters.

6.2.1 Histograms

A frequency distribution and histogram are useful in identifying the shape of a distribution. A histogram is constructed as follows –

1. Divide the range of the data into equal intervals.
2. Label the horizontal axis (x – axis) to match the type of data (of the intervals selected).
3. Find the frequency of occurrences within each interval.
4. Label the vertical axis (y – axis) so that the total occurrences can be plotted for each interval.
5. Plot the frequencies on the vertical axis.

The number of class intervals depends on the number of observations and on the amount of dispersion in the data. It is found that, keeping the number of class intervals as a square root of the sample size would be ideal. One should take care that the histogram should not be too ragged (very small class intervals) or coarse (too large class intervals).

6.2.2 Selecting the family of distributions

When the probability mass functions and probability density functions of the well-known distributions are plotted, it will show-up some shape. The purpose of preparing a histogram is to check whether the data matches with any known pmf or pdf. In simulation study, exponential, normal and Poisson distributions are common and are not difficult to identify through the shapes of histograms.

6.3 PARAMETER ESTIMATION

Once the family of distributions has been decided, the parameters of the distribution have to be estimated. In most of the cases, the sample mean and the sample variance are used to estimate the parameters of a hypothesized distribution.

If the observations in a sample of size n are $X_1, X_2, \ldots. X_n$, then the sample mean $\overline{X}$ is defined as –

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

And the sample variance is given as –

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

Such numerical estimates of the distribution parameters will help in deciding the exact distribution from the family of distributions.

6.4 GOODNESS OF FIT TESTS

Goodness-of-fit tests helps in evaluating the suitability of a potential input model. But, in real time application, there is no single correct distribution. Hence, the result of the goodness-of-fit test has to be just taken as guidance, but not for decision making. Also, goodness-of-fit tests are hugely affected by the sample size. If the sample size is too small, the test may not reject any candidate distribution. If the sample size is too large, then test may not accept any distribution. Hence, proper choice of sample size should be made.

We use Chi-square test and Kolmogorov – Smirnov test for goodness-of-fit. Here, the hypothesis about the distributional forms of input data is tested.

6.4.1 Chi – Square Test

Chi-square test for goodness-of-fit tests the hypothesis that a random sample of size *n* of the random variable *X*, follows a specific distribution or not. The test is valid for large sample sizes and for both discrete and continuous distributions. Here, the given *n* values are arranged into a set of *k* class intervals or *cells*. The test statistic is given by –

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

Here, O_i is the observed frequency in the i^{th} class interval
E_i is the expected frequency in that class interval

The expected frequency for each class interval is computed as $E_i = np_i$ where p_i is the theoretical, hypothesized probability associated with the i^{th} class interval.

It can be shown that χ_0^2 approximately follows chi-square distribution with *k-s-1* degrees of freedom, where *s* is the number of parameters of the hypothesized distribution estimated by the sample statistics. The hypotheses are as follows –

H_0 : The random variable *X* follows the specific distribution with the parameters given by the estimates.
H_1 : The random variable *X* do not follows the specific distribution.

The H_0 is rejected if the calculated value χ_0^2 is greater than table value of $\chi^2_{\alpha,k-s-1}$ for a given level of significance α.

NOTE that, if the expected frequencies (E_i) are less than 5, then it must be combined with the expected frequencies in adjacent class intervals. The corresponding O_i values should also be combined and the value of *k* has to be reduced by 1, for each cell that is combined.

Example 1: The number of vehicles arriving at a particular signal in a 5-minute period between 7:00 am and 7:05 am was monitored for 50 work-days over a 20-week period. Following table shows the data:

Arrivals per period	0	1	2	3	4	5	6	7	8	9	10	11
Frequency	12	10	19	17	10	8	7	5	5	3	3	1

Using χ^2 test, check whether the given data follows Poisson distribution, for $\alpha = 0.05$.

Solution: To check whether the given data follows Poisson distribution, the hypotheses are formed as –

H_0: The random variable is Poisson distributed.
H_1: The random variable is not a Poisson variate.

Note that, here, the random variable X is number of arrivals in a given duration.

The probability mass function for Poisson variate is –

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

We need to first compute the parameter λ for the given values. λ is nothing but a mean.

So, $$\lambda = \bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$, here f_i is the frequency.

In this case, $\bar{x} = \frac{\sum x_i O_i}{\quad}$