SIMULATION MODELING. INPUT DATA COLLECTION AND ANALYSIS

Delia UNGUREANU
Francisc SISAK
Dominic Mircea KRISTALY
Sorin MORARU

Automatics Department, "Transilvania" University of Brasov, M.Viteazu street, no.5, 500174, Brasov, Romania, phone: +40 0268 418836
delia@deltanet.ro, sisak@unitbv.ro, kdominic@vision-systems.ro, smoraru@vision-systems.ro

The steps of the process for conducting a simulation modeling and analysis project include: problem formulation, project planning, system definition, input data collection, model translation, verification, validation, experimental design, analysis. Simulation project involves the collection of input data, analysis of the input data, and use of the analysis of the input data in the simulation model. The input data may be either obtained from historical records or collected in real time as a task in the simulation project. The analysis involves the identification of the theoretical distribution that represents the input data. The use of the input data in the model involves specifying the theoretical distributions in the simulation program code. If we successfully fit the observed data to a theoretical distribution, then any data value from the theoretical distribution may drive the simulation model.

Keywords: simulation modeling, data distribution, analysis project.

1. INTRODUCTION

Simulation modeling and analysis is a technique for improving or investigating process performance. It is a cost-effective method for evaluating the performance of resource allocation and alternative operating policies. It may also be used to evaluate the performance of capital equipment before investment. These benefits have resulted in simulation modeling and analysis projects in virtually every service and manufacturing sector.

In a simulation project, the ultimate use of input data is to drive the simulation. This process involves the collection of input data, analysis of the input data, and use of the analysis of the input data in the simulation model. The input data may be either obtained from historical records or collected in real time as a task in the simulation project. The analysis involves the identification of the theoretical distribution that represents the input data. The use of the input data in the model involves specifying the theoretical distributions in the simulation program code. This process is represented in Fig.1.

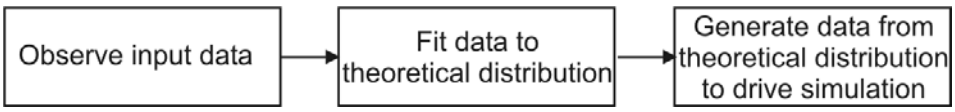

Figure 1- Role of theoretical probability distributions in simulation.

The collection of input data is often considered the most difficult process involved in conducting a simulation modeling and analysis project.

In the data collection and analysis process is necessary to solve some problems: sources for input data, collecting input data, deterministic versus probabilistic input data, discrete versus continuous input data, common input data distributions, analyzing input data.

2. COLLECTING INPUT DATA

There are many sources that we can use to acquire input data, like the following: historical records, manufacturer specifications, vendor claims, operator estimates, management estimates, automatic data capture, direct observation, etc. The data collection phase is the most difficult part of the simulation process.

If the operator is knowledgeable about the system, it may be possible to obtain some performance estimates that can be used as input data. The most physically and mentally demanding form of data collection is direct observation. The input data may be collected either manually or with the assistance of electronic devices.

An important issue for simulation input data concerning time intervals is the *time unit* that should be used. It is usually less labor intensive to collect the data correctly in the first place using a relative, *interarrival* time approach. A second time collection issue is what *types of units* to use. The simulation practitioner should know that we want unbiased data and we do not want to disrupt the process. If the data are biased in either manner, it can lead to a model that may yield inaccurate results.

While collecting the input data, we should realize that there are different classifications of data. One method of classifying data is whether it is *deterministic or probabilistic*. Each individual project will call for a unique set or type of input data. Some of the types of input data may be deterministic, and other types are probabilistic. Deterministic data means that the event involving the data occurs in the same manner or in a predictable manner each time. This means that this type of data needs to be collected only once because it never varies in value. A probabilistic process does not occur with the same type of regularity. In this case, the process will follow some probabilistic distribution. Thus, it is not known with the same type of certainty that the process will follow an exactly known behavior.

Another classification of input data is whether the data are *discrete or continuous*. Discrete-type data can take only certain values. Usually this means an integer. Continuous distributions can take any value in the observed range. This means that fractional numbers are a definite possibility.

3. INPUT DATA DISTRIBUTIONS

We present a few of the most common input data distributions. There are many more different types of probabilistic distributions that we may actually encounter. Sometimes we may encounter these distributions only as a result of a computerized data fitting program. These types of programs are geared toward returning the best mathematical fit among many possible theoretical distributions. In these types of cases, a particular result does not necessarily mean that there is a rational reason why the data best fit a specific distribution. Sometimes a theoretical distribution that does make sense will be almost as good a fit. In these cases, we will have to decide

whether it makes more sense to use the best mathematical fit or a very close fit that makes sense.

- *Bernoulli Distribution* - is used to model a random occurrence with one of two possible outcomes. These are frequently referred to as a success or failure. The Bernoulli distribution is illustrated in Fig.2. The mean and variance of the Bernoulli distribution are:

$$mean = p$$
$$var = p\,(1 - p)$$

where:
p = the fraction of successes;
$(1 - p)$ = the fraction of failures.

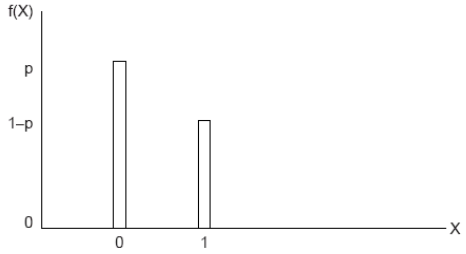

Figure 2 – Bernoulli distribution

- *Uniform Distribution* (Fig.3) – that means that over the range of possible values, each individual value is equally likely to be observed. Uniform distributions can be used as a first cut for modeling the input data of a process if there is little knowledge of the process. The uniform distribution may be either discrete or continuous.

The mean and variance of a uniform distribution are:

$$mean = \frac{(a+b)}{2}$$
$$\text{var} = \frac{(b-a)^2}{12}$$

where:
a is the minimum value
b is the maximum value.

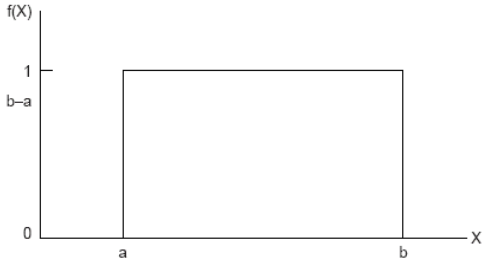

Figure 3 - Uniform distribution.

- *Exponential Distribution* (Fig.4)- is commonly utilized in conjunction with interarrival processes in simulation models because the arrival of entities in many systems has been either proven or assumed to be a random or Poisson process. This means that a random number of entities will arrive within a specific unit of time. The number of arrivals that can be expected to arrive during the unit of time is randomly distributed around the average value. The statistical equations for the mean and variance of the exponential distribution are:

$mean = B$
$var = B2$

The probability is represented by:

$$f(x) = \frac{1}{B}e^{-x}, \text{ or } x = B * \ln[1 - F(x)]$$

where:
B is the average of the data sample;
x is the data value.

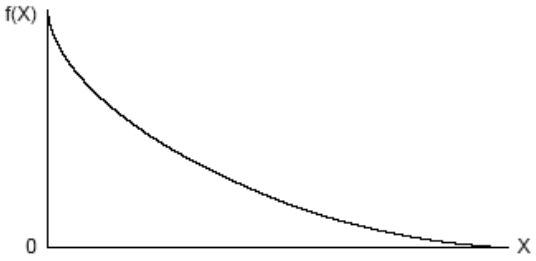

Figure 4 – Exponential distribution

▪ *Triangular Distribution* - it may be used in situations where the practitioner does not have complete knowledge of the system but suspects that the data are not uniformly distributed. In particular, if the practitioner suspects that the data are normally distributed, the triangular distribution may be a good first approximation. The triangular distribution has only three parameters: the minimum possible value, the most common value, and the maximum possible value (Fig.5). Because the most common value does not have to be equally between the minimum and the maximum value, the triangular distribution does not necessarily have to be symmetric. The mean and variance of the triangular distribution are:

$$mean = \frac{a+m+b}{3}$$

$$\text{var} = \frac{(a^2+m^2+b^2-ma-ab--mb)}{18}$$

where : a = minimum value
m = most common value
b = maximum value

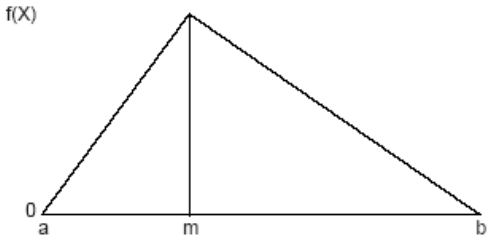

Figure 5 - Triangular distribution.

▪ *Normal Distribution* (Fig.6)- The time duration for many service processes follows the normal distribution. The reason for this is that many processes actually consist of a number of subprocesses. Regardless of the probability distribution of each individual subprocess, when the subprocess times are added together, the resulting time durations frequently become normally distributed. The normal distribution has two parameters: the mean and the standard deviation. The normal distribution is also symmetric. This means that there are an equal number of observations less than and greater than the data mean. The pattern or distribution of the observations on each side is also similar. The somewhat formidable mathematical formula for the normal distribution probability is:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where:
μ is the mean:
σ is the standard deviation.

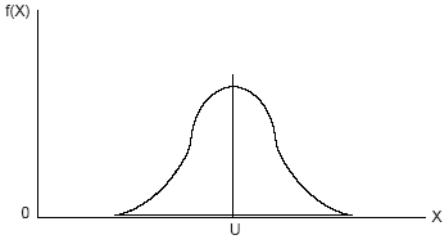

Figure 6 - Normal distribution.

The normal distribution is frequently discussed in terms of the standard normal or Z distribution. This is a special form of the normal distribution where the mean is 0 and the standard deviation is 1. For normally distributed data, the standard normal distribution can be converted to an actual data value with the following equation:

$x = \mu \pm \sigma Z$, where μ is the true population mean and σ is the true population standard deviation.

We can also manipulate this equation to find out the values for specific

cumulative percentages. This is performed with the use of the standard normal or Z table. The following table is a common standard normal or Z table. The small table below illustrates how to use the table. This table is known as a right-hand tail table. The top row and left column contain the Z values. The interior of the table contains the cumulative percentage of observations for the standard normal distribution. The table is symmetric. The cumulative percentages are subtracted from 1.

Z value	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839

- *Poisson Distribution* (Fig.7) - is used to model a random number of events that will occur in an interval of time. The Poisson distribution has only one parameter, λ.

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where:

λ is both mean and variance;

x is the value of the random variable.

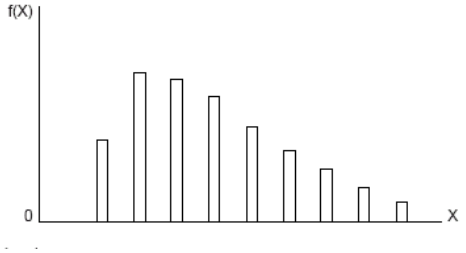

Figure 7 - Poisson distribution.

- *Weibull Distribution* (Fig.8) - is often used to represent distributions that cannot have values less than zero. The Weibull distribution possesses two parameters. These are an α shape parameter and a β scale parameter. The lengthy probability function for the Weibull is:

$$f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}\text{ , for } x>0,\ 0 \text{ otherwise}$$

where: α is a shape parameter and β is a scale parameter.

The mean and variance are represented mathematically by:

$$mean = \frac{B}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$$

$$\text{var} = \frac{\beta^2}{\alpha}\left\{2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^2\right\}$$

where: α is a shape parameter,

β is a scale parameter,

Γ is given by $\Gamma = \int_0^{\infty} x^{\alpha-1}e^{-x}dx$.

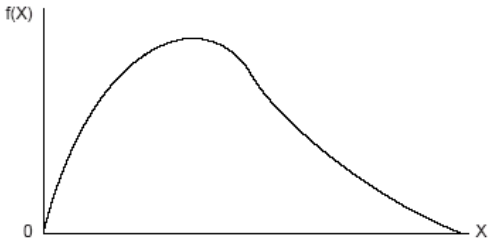

Figure 8 - Weibull distribution.

▪ *Gamma Distribution* (Fig.9) - is another distribution that may be less common to the practitioner. The probability density equation for the gamma distribution is:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},$$

for x>0, 0 otherwise

where α, β, and Γ are defined as in the Weibull distribution.

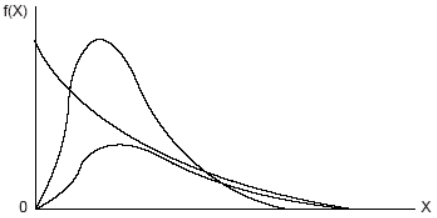

Figure 9 – Gamma distribution

The gamma distribution can degenerate to the same mathematical representation as the exponential distribution. The gamma distribution cannot go below 0.

▪ *Beta Distribution* (Fig.10) - is a bit different from most of the previously presented distributions. The beta distribution (Fig. 10) holds the distinction of being able to cover only the range between 0 and 1. The mathematical formula for the beta distribution probability density is:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \text{ for } 0<x<1, \text{ 0 elsewhere}$$

where α and β are shape parameters 1 and 2, respectively, and Γ is defined as for the Weibull and gamma distributions. The mean and variance of the beta distribution are:

$$mean = \frac{\alpha}{\alpha+\beta}$$

$$\text{var} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

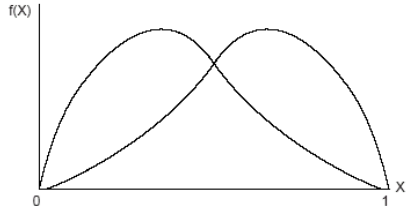

Figure 10 – Beta distribution

▪ *Geometric Distribution* (Fig.11) - In contrast to the previous less common distributions, the geometric distribution is discrete. This means that the values taken on by the geometric distribution must be whole numbers. The geometric distribution (Fig.11) has one parameter p, which is considered the probability of success on any given attempt; (1 - p) is the probability of failure on any given attempt. The probability of x – 1 failures before a success on the xth attempt is represented by:

$$p(x) = p(1-p)^{x-1}, \text{ x=1,2,3…}$$

The mean and variance of the geometric distribution are represented by:

$$mean = \frac{1-p}{p}$$

$$\text{var} = \frac{1-p}{p^2}$$

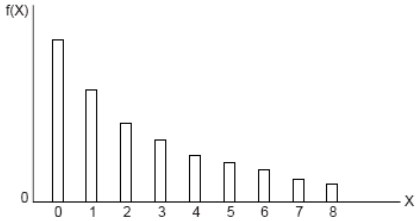

Figure 11 – Geometric distribution

Some types of input data are actually a combination of a deterministic component and a probabilistic component. These types of processes generally have a minimum time, which constitutes the deterministic component. The remaining component of the time follows some sort of distribution.

4. ANALYZING INPUT DATA

The process of determining type of the distribution for a set of data usually involves what is known as the essence of fit test. These tests are based on some sort of comparison between the observed data distribution and a corresponding theoretical distribution. If the difference between the observed data distribution and the corresponding theoretical distribution is small, then it may be stated with some level of certainty that the input data could have come from a set of data with the same parameters as the theoretical distribution. There are four different methods for conducting this comparison:

- *Graphic approach* - The most fundamental approach to attempting to fit input data is the graphic approach. This approach consists of a visual qualitative comparison between the actual data distribution and a theoretical distribution from which the observed data may have come. The steps for using the graphic approach include:
  - Create a histogram of observed data
  - Create a histogram for the theoretical distribution
  - Visually compare the two histograms for similarity
  - Make a qualitative decision as to the similarity of the two data sets

There are two common approaches for determining how to handle the cell issue: *equal-interval approach* - we set the width of each data cell range to be the same value, or *equal-probability approach* – is a more statistically robust method for determining the number of cells is the equal-probability approach

- *Chi-square test* - is commonly accepted as the preferred goodness of fit technique. Like the graphic comparison test, the chi-square test is based on the comparison of the actual number of observations versus the expected number of observations.
- *Kolmogorov–Smirnov test* - should be utilized only when the number of data points is extremely limited and the chi-square test cannot be properly applied. The reason for this is that it is generally accepted that the KS test has less ability to properly fit data than other techniques such as the chi-square test.
- *Square error* - uses a summed total of the square of the error between the observed and the theoretical distributions. The error is defined as the difference between the two distributions for each individual data cell.

A very common question among data acquisition are:

- how much data needs to be collected – is necessary to observe: the right data, the different values that are likely to occur, the need to have enough data to perform a goodness of fit test.
- is possible to fit the observed data to a theoretical distribution - possible causes for this difficulty include: not enough data were collected, data are a combination of a

number of different distributions.

5. SOFTWARE IMPLEMENTATIONS FOR DATA FITTING

Fitting a significant number of observed data sets to theoretical distributions can become a very time consuming task. In some cases, it may be mathematically very difficult to fit the observed data to some of the more exotic probability distributions. For this reason, most practitioners utilize some sort of data fitting software. Two commonly available programs among others to perform this function are:

- *Arena input analyzer*
- *Expert fit*

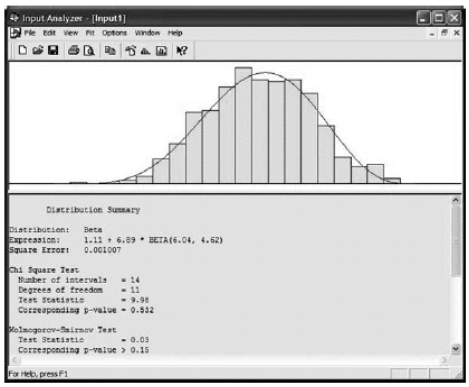

Figure 12 - ARENA Input Analyzer.

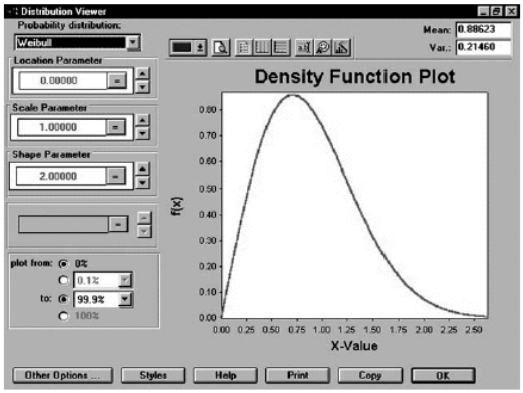

Figure 13 – Expert fit

6. CONCLUSIONS

In this paper we covered the collection and analysis of input data. This phase of the simulation project is often considered the most difficult. Data may be collected from historical records or can be collected in real time. In order to use observed data in a simulation model, it is preferred to firstly be fit to a theoretical probability distribution. The theoretical probability distribution is then used to generate values to drive the simulation model.

7. RESULTS

In order to determine the best theoretical distribution fit for the observed data, we can use one of many comparison methods. These include the graphic, chi-square, KS, and square error methods. In some cases, it may not be possible to obtain a satisfactory theoretical fit. Additional data should be collected, and a new attempt to fit the data should be made. If this is not possible, it may be necessary to use observed data to drive the simulation model.

8. REFERENCES

[1] Hildebrand, D.K. and Ott, L., *Statistical Thinking for Managers,* PWS-Kent, Boston, 1991.

[2] Johnson, R.A., Freund, J.E., Miller, I., *Miller and Freund's Probability and Statistics for Engineers,* Pearson Education, New York, 1999.

[3] Law, A.M., Kelton, D.W., *Simulation Modeling and Analysis,* 3rd ed., McGraw-Hill, New York, 2000.