

Chapter 2

Elements of Statistics for Simulation

Idalia Flores De La Mota and Antoni Guasch

2.1 Introduction

This chapter presents statistical concepts and definitions that are used when designing a simulation model. We start, in the first instance, by considering a conceptual model, then the need to verify the initial data for the model, followed, if necessary, by the data that can be adjusted to some probability distribution, where validating this adjustment also involves statistical concepts. Later, once the results are in, we consider the experiments that need to be done, as well as the replications, ending up with the analysis of the data obtained.

As described in the previous chapter, the construction of a simulation model requires data collection. This fundamental phase of constructing simulation models can, according to Trybula (1994), take between 10 and 40% of the time required for the study. Fortunately there have been a lot of studies looking for ways to shorten this time, and Skoogh and Johansson (2008) have proposed a methodology that we will discuss further on. So it is necessary to have enough of the required data in the shortest time possible, for which we need to have an idea of what type of data about the system under study are required. According to Robinson and Bhatia (1995), the data can be classified as shown in the following Table 2.1.

Category A data are very convenient as the only work involved is their analysis and validation. Category B data require an additional effort as they have to be collected during the simulation study. Lastly the category C data, which corresponds to an estimation of the data, require strategy as well as careful and scrupulous design in order to maintain the quality of the model.

For good data collection, once we have identified into which of the three categories they fall, we need to answer some questions, such as:

What information do we have? A common fault in simulation studies that are not well delimited during the planning stage is because of more data than necessary or than can be validated by the available data being extracted from the simulation.

Table 2.1 Data classification for a simulation. *Source* Robinson and Bhatia (1995)

Type of data		
Category	Availability	Cases
Category A	Available	Automated recording systems, previous measured data
Category B	Not available but collectable	The system has not been studied previously and there are no records
Category C	Not available or collectable	New processes or equipment

Another problem that has to be solved is the reliability of our data, therefore some questions that can support this process are:

- What data do we need?
- How will we get these data?
- How long does each stage of data collection take, approximately?
- With what information and how will the simulation results be validated?
- What configurations of the model should be run?
- How many runs should there be and how large should they be?

To collect information about the system's structure and operating procedures, the following considerations are required:

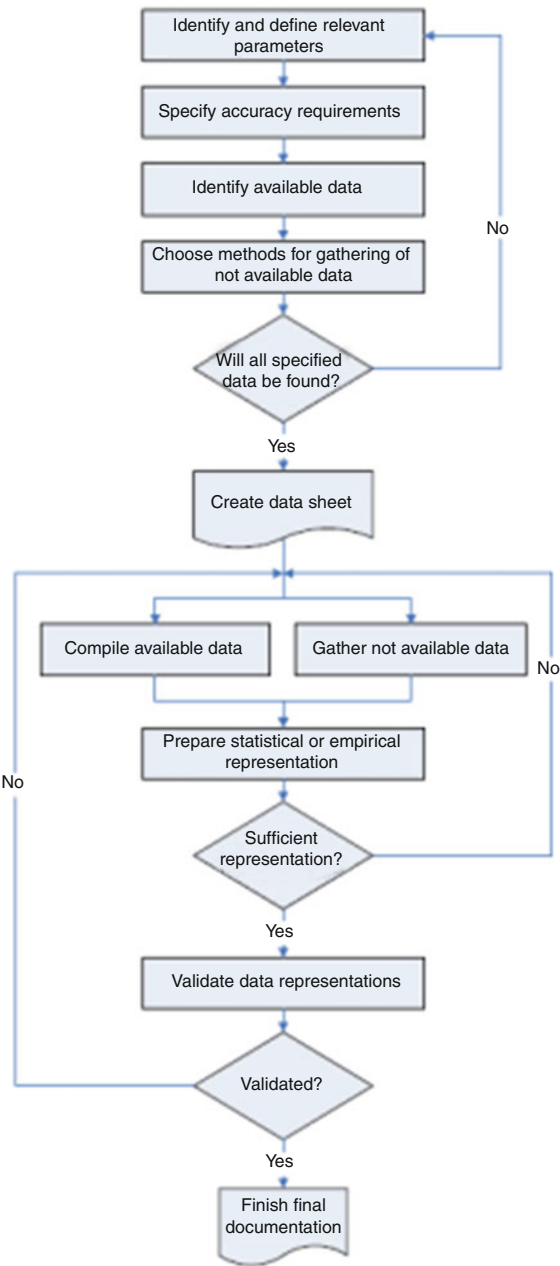
- A single document or an interview with one person is not enough. It is essential for the simulation analyst to talk to as many experts in the system as necessary to get a full understanding of the system to be modeled.
- Part of the information provided will invariably be wrong. If a certain part of the system is particularly important, then at least two experts in the system will be needed.
- The system's operating procedures may not be formalized.

Considering all of the above and in accordance with the methodology proposed by Skoogh and Johansson (2008) which is presented in the following figure, we propose the following steps (Fig. 2.1):

The steps listed in the above paragraphs are:

1. Identifying and defining relevant parameters.
2. Specifying the requirements for accuracy.
3. Identifying the availability of the data.
4. Choosing methods for collecting data that are not available.
5. Were all the specified data found?
6. Creating a datasheet
7. Compiling the available data
8. Collecting the data that are not available
9. Preparing a statistical or empirical representation, as the case may be.
10. Is the representation sufficient?
11. Validating the representations of the data.

Fig. 2.1 Methodology proposed for increasing the accuracy and speed in the administration of the data for a simulation. *Source* Skoogh and Johansson (2008)



12. Was the validation sufficient?
13. Finishing the documentation of the data.

Because of the above, probability and statistical tools are indispensable, so here is an outline of some important concepts and definitions.

- *Random variable*: X is a random variable if it can take any value from a finite (discrete random variable) or an infinite range (continuous random variable). Despite the exact sequence of values not being known, the range of variation and probability of obtaining a certain value is known.
- *Probability distribution*: makes it possible to relate a set of values or measures with their relative frequency of appearance.
- *Probability density function* $f(x_i)$: describes the probability of a random variable X having a certain value x_i

$$f(x_i) = P(X = x_i)$$

- *Cumulative distribution function* $F(x_i)$: describes the probability of a random variable X having a value that is smaller than or equal to a certain value x_i

$$F(x_i) = P(X \leq x_i)$$

- *Probabilistic or stochastic model*: this type of model uses one or more random variables to formalize the system's dynamics of interest. In consequence, during the experimentation phase, the model will not generate a single output, but rather the results generated are useful for getting estimations of the variables that characterize the real behavior of the system.
- *Sampling*: this is the act of taking samples from a population. A sample must be representative of the population from which it is obtained for it to be useful when inferring statistics about said population.
- *Random sample*: If the random variables X_1, X_2, \dots, X_n have the same probability function (density) as that of the distribution of the population and their joint probability (distribution) function is equal to the product of the marginals, then X_1, X_2, \dots, X_n form a set of n independent identically distributed (IID) random variables that constitute a random sample of the population.
- *Time between arrivals*. In the simulation of discrete events the time between arrivals is defined as the time that elapses between one arrival of an event to the system and the arrival of another event.

The collection of data (if possible) serves to specify the parameters of the model and the *probability distributions* (for example, for the failure time and repair time of the machine). The simulation of a system or process where there are components that are inherently random requires the generation of *random variables*. In the following sections we discuss how these values can be conveniently and efficiently generated from a desired probability distribution for their use in the simulation

models. We also include the distribution functions that are more frequently employed in simulation models and the cases they are used in are specified. Care must be taken not to commit two common errors at this level: replacing a probability distribution by its mean value or using an unsuitable distribution.

All the discrete simulation packages that are available contain a random number generator however it is important to briefly discuss what this phase of the modeling consists of.

2.2 Generation of Random Numbers

Random simulation methods were initially applied by mathematicians and physicists to solve certain deterministic problems that could be expressed as mathematical equations whose solutions could not be easily obtained by the usual numerical or analytical methods. In many significant mathematical problems, we can find a stochastic process with a probability distribution or parameters that satisfy the requirements of the equations. The deterministic problems that stochastic simulation has been used for include the evaluation of multiple integrals, the solution of very high order differential equations, complex queuing problems and schedule programming. Although there are analytical methods for these cases, simulation methods have been found to be more effective.

Another type of problem that leads to the simulation of random variables arises in those situations where there is stochastic behavior and that require some type of sampling, which in practice is either impossible or inconvenient, as in the case of future data. Although we cannot get the data, we know something about the population from which it comes. For stochastic simulation, it will then be necessary to build a probabilistic model that is tailored to the study. This means that shall be indispensable identify one (or several) probability distribution(s) tailored to each case, which makes it possible to generate values that behave similarly to the phenomenon in question. Nowadays, most statistical and simulation packages include a random number generator, however we still cover the issue in this chapter as it a good idea to have a better understanding of what it means to generate these numbers.

The methodology for generating random numbers has a long and interesting history. The first methods were developed practically by hand, such as tossing coins, choosing cards, throwing dice or taking numbered balls out of an urn (Fig. 2.2). A lot of lotteries currently operate this way. At the start of the 20th Century, statisticians followed gamblers into their interest in random numbers and mechanical devices were built for a speedier generation of random numbers. In 1938, Kendall and Babington-Smith used a fast-spinning disk to prepare a table of 100,000 random digits. Sometime later they developed electrical circuits based on randomly pulsating vacuum tubes to throw out random numbers at a rate of 50 numbers a second. The Royal Mail used this type of machine: Electronic Random Number Indicator Equipment (ERNIE), to choose the winners of the Premium

Fig. 2.2 Selection of random numbers



Bonds lottery. The Rand Corporation used another similar device to generate a table of one million random digits.

Many other approaches have been used to randomly select numbers, such as the selection of numbers at random from the telephone directory or from census reports, or using digits taken from the decimal expansion of π .

As computers as well as simulation were being used more, there has been more interest in methods for generating random numbers that are compatible with the way computers work. Thus, research in the 1940s and 1950s focused on numerical or arithmetic ways of generating random numbers. Said methods are sequential, with every new number determined by one or more of its predecessors and, new ones are generated according to a mathematical formula, as we will see in some examples given in the following section.

2.3 Properties of a Good Random Number Generator

Because of the characteristics of simulation, it is necessary to generate random numbers that represent the behavior of the problem to be simulated. Although there are many ways of generating random numbers, for the majority of real applications the generator shall have a series of properties that make it truly useful and similar to the real processes:

- It must produce random numbers.
- It must be fast.
- It must have a long period before repeating its cycle.
- It must generate random numbers that can be reproduced.
- It must not require a lot of computer storage space.
- It must not degenerate.

Each of these properties are explained as follows:

In the production of random numbers, the fact that they are random means that they have to be independent of each other. They should initially come from a

uniform distribution, which means that they are not strictly random, as their generation is based on a function but, for all practical purposes, this is what best fits the concept. This is the reason why they are called pseudo-random because, in reality, there are no cycles.

Large-scale simulation models generally require a lot of random numbers, so the generating method must be fast while the time and memory used in the computer must not be excessive.

In practical terms, all the generating methods produce numbers that sooner or later repeat their cycle at some point. This means that the sequence of numbers is repeated. Then, it is important for the generating method that is chosen to produce all the random necessary numbers before a cycle is completed. This suggests that the selection of the method will depend on the specific application. If 100 numbers are needed, a 200-long cycle will not cause any problems.

Also, it is important for the random number generating method not to degenerate, in other words, for the method not to repeat the same number indefinitely. For example, some methods degenerate at the zero value.

Consequently, we have to look for algorithmic procedures for the generation of number that are at least apparently random. Von Neumann's idea was to produce numbers that look random employing the computer's arithmetic operations. Starting from a seed or initial value $(u_0, u_{-1}, \dots, u_{-p+1})$, a sequence is generated by means of $u_i = d(u_{i-1}, \dots, u_{i-p})$ for a certain function d . With the seed having been chosen, the sequence is set.

The first question to be resolved is what do we mean by random numbers, which is why it is important have a good definition. Starting out from the modified version of Kolmogorov and Uspenskii's classic definition (1987) associated with the idea of algorithmic complexity, we have the following definition:

A sequence of numbers is random if it cannot be efficiently produced by a program that is shorter than the string itself.

For example, the sequence 0010010010... is interpreted as being non-random, given the fact that we can give a shorter algorithm than the string itself. The discussion of these ideas leads to interesting proposals. For example, a criterion for the definition of random numbers can be introduced that is similar to Turing's for recognizing an artificial intelligence, which brings us to the following definition:

A succession of numbers is random if nobody using reasonable computer resources in a reasonable time is able to distinguish between the series and a truly random sequence better than throwing a coin to decide which one it is.

The precise expression of this definition leads to the ones known as PT-perfect generators (Lécuyer 1990), of great interest in cryptography, but not in simulation, because of its slowness.

2.4 Generation of Random Numbers with a Uniform Distribution Between Zero and One

Discrete case

The importance of generating random numbers consists of them representing the value of a random variable; thus, if the variable is discrete and can only take n given values which are x_i , $1 \leq i \leq n$, whose probability is p_i , we know that:

$$\sum_{i=1}^n p_i = 1$$

These probabilities can be obtained in advance or else be determined through a series of observations, on which basis the different probabilities are established, as shown in the following example:

Example 2.1

At the junction of streets A and B, the following observation was made of the vehicles traveling along street A. Table 2.2 gives these observations:

The stochastic variable x_i can take one of the following three values: turning to the right, turning to the left or not turning. Observe that the possible values do not necessarily have to be numerical, they could, in this case, be actions. For it to be discrete it is necessary for the number of possible results to be finite, with the probabilities that are given in Table 2.3 (Fig. 2.3).

To generate random numbers with these probabilities, we consider the cumulative probability graph given in Fig. 2.4.

We generate a sequence r_i of random numbers with uniform distribution between zero and one and, depending on the range where the random number is found, this will be the value we associate with it, as shown in Table 2.4.

Congruential or residual methods

Congruential random number generating methods first occurred to Lehmer in 1951. These methods are based on what mathematicians call congruence relations. Although there are a lot of variants, the most popular ones are the multiplicative congruential generators or power residue methods.

This method, like the one above, requires a first number, after which a string of random numbers is generated by the recursive application of the following formula:

$$X_{i+1} \equiv \alpha X_i \pmod{m}$$

This relation is read as “ X_{i+1} is congruent with αX_i in module m ”. By definition, two integers A and B are said to be congruent in module m (with integer m), if $(A-B)$ is divisible between m and if A and B produce identical residues when divided by m , this means that A is congruent with B in module m , if and only if there is a value k in the integers, so that

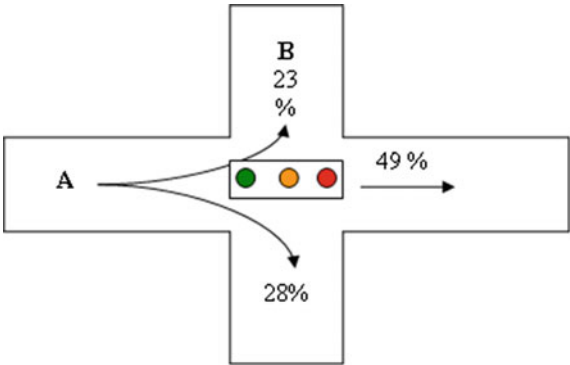
Table 2.2 Observations at the junction

x_i	Remarks	Probability (p_i)	Probability accumulated
To the right	28	0.28	0.28
To the left	23	0.23	0.51
Do not turn	49	0.49	1.00

Table 2.3 Probabilities associated with the junction

Value	Right	Left	Does not turn
Probability	0.28	0.23	0.49

Fig. 2.3 Probabilities associated with the junction in the form of a graph



$(A-B) = km$.

Any sequence of numbers can be obtained by multiplying the preceding number by a constant and then reducing product by module m . The module m operation means dividing αX_i between m and keeping the residue as the value of X_{i+1} .

For example, let $\alpha = 5$, $X_0 = 3$ and $m = 32$. The value of X_1 will be 15 as:

$$X_1 \equiv (5)(3)(\text{mod } 32)$$
$$15/32 = 0 \text{ with } 15 \text{ as residue}$$

$X_2 = 11$ can be obtained in the same way. The distribution of the X_i is uniform and a source of random numbers.

Another way that is even simpler consists of using the “random” that different programming languages have. In the case of congruential methods, their length depends on the chosen module.

These are only some of the methods for generating of random numbers. There are many more and if the problem in question requires special treatment as regards randomness, it is important to consider the possibility of the random numbers being generated by the modeler or else through a programming language when the simulation is being executed and not using software that already includes it.

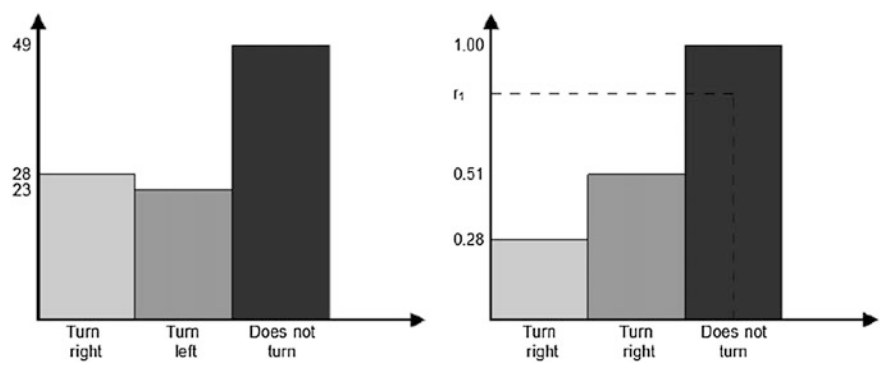


Fig. 2.4 Cumulative probability

Table 2.4 Ranges and cumulative probability

r_i range value of x_i	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
	0.274 (0,0.28) right	0.911 (0.51,1) Does not turn	0.046 (0, 0.28) right	0.466 (0.28, 0.51) left	0.4976 (.0.28, 0.51) left

2.5 Selection of a Distribution Function

In some cases, the initial problem in the simulation of random variables is precisely the choice of a suitable distribution function. There are four considerations you may take into account for this selection:

1. *The special characteristics of each specific distribution.* This means the particular behavior that the phenomenon in question may have. For example, if the data has only two distinct values, the proper distribution will doubtless be a Bernoulli. If it is a question of discrete data, we can immediately leave to one side all the continuous distributions and vice versa. If it is a sampling, we should observe whether this is with or without replacement, in which case binomial or hypergeometric distribution, respectively, is used. Another important datum is the symmetry or asymmetry of the data of the phenomenon in question; for example, normal distribution is symmetric, whereas triangular distribution may or may not be. The times between events tend to be distributed as exponential, for continuous time.
2. *The accuracy with which a distribution can represent a set of experimental data.* This is only verified through graphs, such as the data histogram, which was obtained from the frequencies which were observed and goodness-of-fit tests.
3. *The facility with which the distribution fits the data, in other words, the estimation process for the corresponding parameters.* In some distributions, the process of obtaining estimators is extremely complicated and time-consuming,

particularly for models composed by equations with non-linear parameters. In these cases, we can resort to a simpler approximate distribution or make use of iterative algorithms, in order to get estimators that are sufficiently suitable for the particular needs of each problem.

4. *Computational efficiency in generating random variables.* As we have already mentioned, in some cases you have to do a lot of calculations to generate a set of variables. The simpler these calculations, the more efficient will be the calculation to obtain a large number of variables, which is important as large samples are desirable if we want to extract reliable conclusions.

The use of probability and statistics is an integral part of a simulation study and are used to understand how to model a probabilistic system that meets the following characteristics:

- To validate the simulation model.
- To choose the probability distributions to start with.
- To obtain random samples from the distributions.
- To make a statistical analysis of the simulation results.
- To design the simulation experiments.

There can be different probability distributions depending on the system and the problem to be solved, as shown in the following (Table 2.5):

2.6 Continuous Distribution Functions

A summary table is included for each one of the functions presented. This summary table includes the probability density function, the cumulative distribution function, the mean and the variance. The mean (or expected value) and the variance of a continuous random variable X that follows a density function, is calculated by means of:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Table 2.5 Initial probability distributions

Type of system	Sources of chance
Manufacturing	Process times, time between breakdowns, arrivals of orders
Communications	Time between arrivals of messages, length, type, end destination
Transport	Size of the load, transport time, loading and unloading times
Hospital processes	Time between arrivals of patients, type of illness, length of consultation

$$\sigma^2 = V(X) = E(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

2.6.1 Exponential Distribution Function

This distribution is used to model the time between entities. It is also employed for modeling service times that are highly variable; for example, the length of time of a phone call. This distribution is related to Poisson distribution, given that if an arrivals rate (arrivals per unit of time = λ) follows a Poisson distribution, the time between arrivals follows an exponential distribution of parameter $\beta = 1/\lambda$.

Normal, log-normal and gamma distribution functions tend to be more frequently used for modeling those activities where, under normal operating conditions, the time consumed usually shows (physically justifiable) variations in respect of an average value (Fig. 2.5 and Table 2.6).

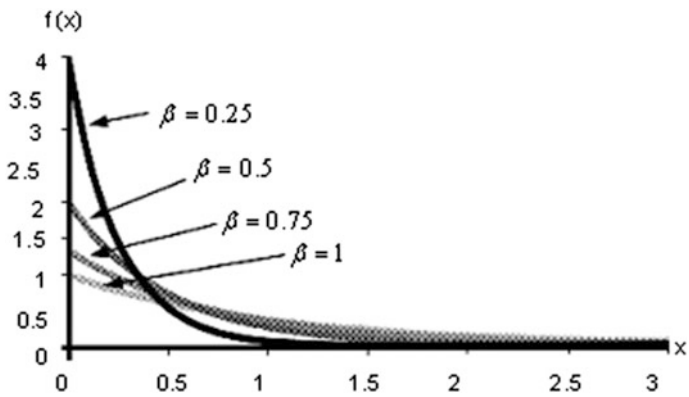


Fig. 2.5 Exponential distribution

Table 2.6 Exponential distribution function

Exponential	<i>Expo</i> (β)
Possible interest	Time between arrivals of customers when the mean frequency of arrivals is constant
Probability density	$f(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 1 - e^{-\frac{x}{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Mean	β
Variance	β^2

2.6.2 Gamma Distribution Function

In general, the time that a production unit requires to carry out a repetitive raw materials processing operation or the time consumed in a repetitive activity of transporting material between two work stations usually follows a constant value with small variations caused by certain physical aspects. These could be conclusively modeled but, in order to simplify the task, are usually described as the result of a random activity through statistical models.

In accordance with the parameters of the gamma probability distribution function (pdf), it shows a very similar graph to that of the normal pdf, but with a certain asymmetry that answers to the presence of data with values that are higher than the average value. This asymmetry makes it possible to model sequences of activities (for example, processing units or transport units) that are done in parallel, so that each one of them answers to a normal pdf, but the time consumed in the sequence of activities shows an asymmetry slanted towards the values that are higher than the average (Table 2.7).

Figure 2.6 shows different shapes of the gamma distribution in accordance with the variation of their parameters α and β , which are, respectively, shape and scale parameters.

The gamma distribution function represents a very good statistical modeling tool for modeling real systems submitted to the occurrence of certain events; for example, probability of machine failure, which increases the appearance of values higher than the average value.

Table 2.7 Gamma distribution function

Gamma	$Gamma(\alpha, \beta)$
Probability density	$f(x) = \begin{cases} \frac{\beta^{-\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$ <p>$\Gamma(\alpha)$ is the gamma function</p> $\Gamma(\alpha) = \int t^{\alpha-1} e^{-t} dt$ <p>If α is a positive integer $\Gamma(\alpha) = (\alpha - 1)!$</p>
Cumulative distribution	$F(x) = \begin{cases} 1 - e^{-\frac{x}{\beta}} \sum_{j=0}^{\alpha-1} \frac{(x/\beta)^j}{j!} & x \geq 0 \\ 0 & x < 0 \end{cases}$ <p>If α is a positive integer; otherwise there is no closed formula</p>
Mean	$\alpha\beta$
Variance	$\alpha\beta^2$

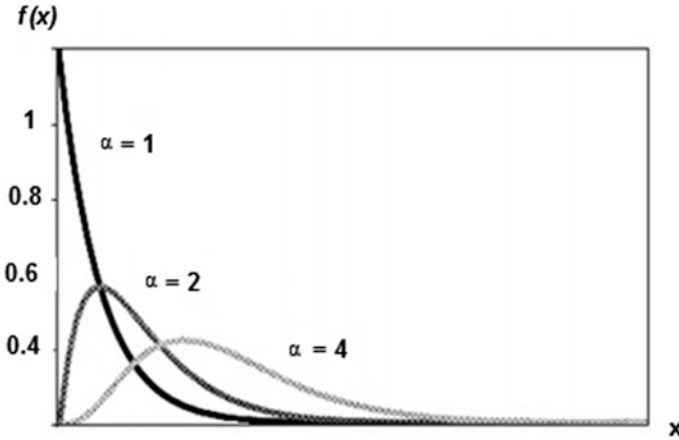


Fig. 2.6 Gamma distribution function ($\beta = 1$)

2.6.3 Log-Normal Distribution Function

In general, the log-normal distribution function is used for modeling a multiplicative sequence of operations; for example, the repercussion of the failure of one machine is therest of the machines being shutdown. The gamma distribution function is used for modeling an additive sequence of operations, while the log-normal distribution function can be used for modeling the time required to do a manual task (Fig. 2.7 and Table 2.8).

2.6.4 Normal Distribution Function

This is used for modeling systems where 70% of the sampled data is found at a distance of less than σ (standard deviation) from the average value μ , and the frequency of appearance of the data is found symmetrically distributed in respect of the average value.

One example for using a normal distribution function is the modeling of the production time of the machines, when the possibility of different types of faults or errors is not considered.

Figure 2.8 represents the histogram of a normal distribution function, in which the difference of the gamma and log-normal pdfs, the data practically does not present huge variations in respect of an average value (Table 2.9).

The cumulative distribution function cannot be accurately calculated. As a consequence, numerical methods were employed to obtain tables for the function. Given that it is not practical to obtain a table for all the possible values of μ and σ^2 , a table is constructed for the standard normal distribution (of parameters $\mu = 0$,

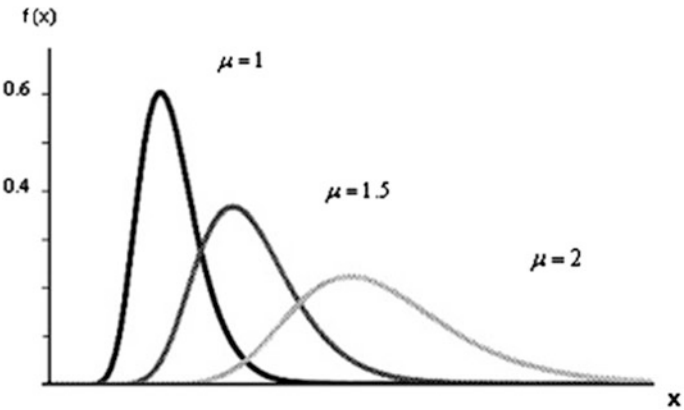


Fig. 2.7 Log-normal function $\sigma = 1$

Table 2.8 Log-normal distribution function

Log-normal	$LN(\mu, \sigma^2)$
Probability density	$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$
Cumulative distribution	There is no closed formula
Mean	$e^{\mu + \sigma^2/2}$
Variance	$e^{2\mu + \mu^2}(e^{\sigma^2} - 1)$

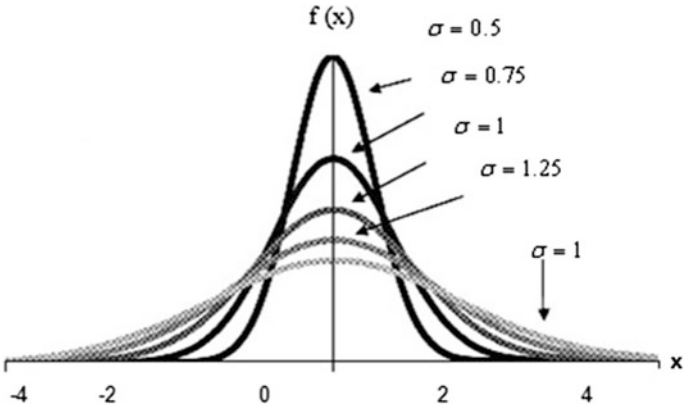


Fig. 2.8 Normal distribution function ($\mu = 0$)

$\sigma^2 = 1$). If X is a random variable of normal distribution of values of, μ and σ^2 the random variable $Z = (X-\mu)/\sigma$ follows a normal distribution of mean 0 and variance 1.

Table 2.9 Normal distribution function

Normal	$N(\mu, \sigma^2)$
Probability density	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
Cumulative distribution	There is no closed formula
Mean	μ
Variance	σ^2

2.6.5 *Triangular Distribution Function*

Triangular distribution provides a first approximation when there is not very much available information. This distribution is defined with the minimum value, the maximum and the mode. It is also used to specify activities that have a minimum, maximum and more probable time (Fig. 2.9 and Table 2.10).

2.6.6 *Uniform Distribution Function*

The uniform distribution is a continuous distribution that is used to specify a random variable, which has the same probability of having its value at any point on a range of values. It is defined by specifying a lower bound and an upper bound b for the range. Uniform distribution is not, in general, a valid representation of a

Fig. 2.9 Triangular distribution function

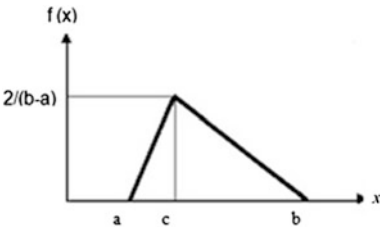


Table 2.10 Triangular distribution function

Triangular	$Trian(a, b, c)$
Probability density	$f(x) \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} & c < x \leq b \\ 0 & o.c \end{cases}$
Cumulative distribution	$F(x) \begin{cases} 0 & x < a \\ \frac{(x-a)^2}{(b-a)(c-a)} & a \leq x \leq c \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)} & c < x \leq b \\ 1 & b < x \end{cases}$
Mean	$\frac{a+b+c}{3}$
Variance	$\frac{a^2+b^2+c^2-ab-ac-bc}{18}$

Fig. 2.10 Uniform distribution function

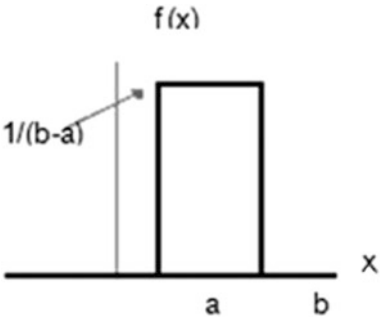


Table 2.11 Uniform distribution function

Uniform	$U(a, b)$
Probability density	$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & o.c. \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & b < x \end{cases}$
Mean	$\frac{(a+b)}{2}$
Variance	$\frac{(b-a)^2}{12}$

random phenomenon. It is used when the distribution is unknown and there is only information about the extreme values (Fig. 2.10 and Table 2.11).

2.6.7 Weibull Distribution Function

The Weibull distribution function is a family of distributions that depend on two parameters: the shape parameter α and the scale parameter β . When the shape parameter $\alpha = 1$, both the Weibull distribution and the Gamma distribution are reduced to the negative exponential distribution. This is used for modeling process times and also for modeling the reliability of an item of equipment by defining the time that elapses until the equipment breaks down. An additional parameter can be introduced by replacing the Weibull random variable X by $X-a$, where a is a location parameter that represents a threshold or guarantee time (Table 2.12).

As you can see in Fig. 2.11, this distribution has different shapes depending on the value of the scale parameter β , although the shape parameter α can also be made to vary and likewise obtain different shapes.

In numerous situations, the empirical probability distribution has such a shape that there is not a standard distribution that properly represents the behavior of the process. The options that can be posed for its formalization are various:

Table 2.12 Weibull distribution function

Weibull	Weibull (α, β)
Probability density	$f(x) = \begin{cases} \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 1 - e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Mean	$\frac{\beta}{\alpha} \Gamma(\frac{1}{\alpha})$
Variance	$\frac{\beta^2}{\alpha} \left\{ 2\Gamma(\frac{2}{\alpha}) - \frac{1}{\alpha} [\Gamma(\frac{1}{\alpha})]^2 \right\}$

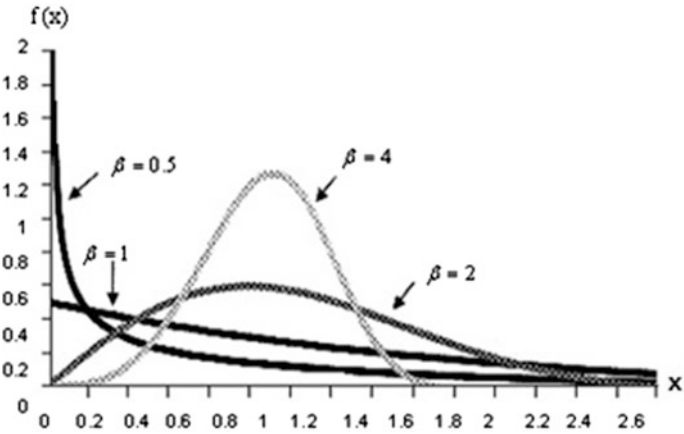


Fig. 2.11 Weibull distribution function ($\alpha = 1/2$)

- Directly employing the empirical probability distribution with the advantages and drawbacks already described for this option.
- Rejecting the values of the sample that are clearly atypical: This is possible, always provided the loss of information can be assumed in the context of the process that is being modeled.

If the histogram has several dominant areas, one can try to separate and adjust it in several cases. In other words, a different distribution shall be adjusted in each one of the dominant areas (Law 2006; Altioik and Melamed 2007) obtaining a multi-modal distribution. If p_j is the proportion of samples in each dominant area and $f_j(x)$ the probability density function in each one of the areas, the overall probability density function shall be

$$f(x) = \sum_{j=1}^n p_j f_j(x)$$

Fig. 2.12 Histogram of the weight in kg/mm of steel coils provided by a supplier (reproduced with the kind permission of Siderúrgica del Mediterráneo S.A.)

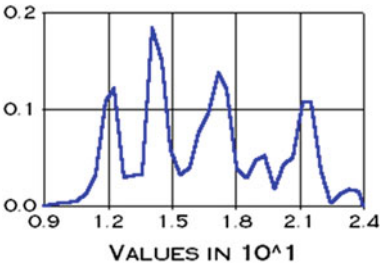


Table 2.13 Set of adjusted pdfs

Range (kg/mm)	% of samples	Adjusted pdf
[9–13.5]	18.56	Normal, $\sigma = 0.64302$ $\mu = 12.193$
[13.5–15.5]	24.03	Log-normal, $\sigma = 0.02787$ $\mu = 2.6743$
[15.5–18.5]	28.78	Triangular, $m = 17.51$ $a = 15.358$ $b = 18.626$
[18.5–24]	28.61	Normal, $\sigma = 1.2976$ $\mu = 21.095$

Example 2.2 Modeling of the weight of steel coils

A lot of steel works, especially ones that produce steel for the car industry, work with coils, in other words, their raw materials can be unprocessed steel coils. In the factory there are semi-processed coils and the end product that is delivered to the customer, i.e. processed coils (end product). One of the most important aspects that must be borne in mind when planning production operations is the definition of the physical characteristics of the coil: width, length, thickness, weight. Figure 2.12 is the histogram for a sample of the weight in kilograms per millimeter of width of 1200 steel coils. From this value, the weight of the coil can be obtained by multiplying the diameter of the coil and its length by its width, if its thickness is known.

It is not possible to obtain a unimodal probability density function that fits the histogram. Accordingly, the option was taken to obtain a different adjustment for each one of the four dominant areas. The final result was (Table 2.13).

Figure 2.13 shows the adjustment obtained for one of the dominant areas.

2.7 Discrete Distribution Functions

2.7.1 Bernoulli Distribution Function

Bernoulli distribution is applied in cases where there are two possible states. The probability of one state is p and that of another state $q = 1 - p$. The phenomena that define them are, among others:

- Whether or not the piece that exits the process is defective.
- Whether or not an employee comes to work.

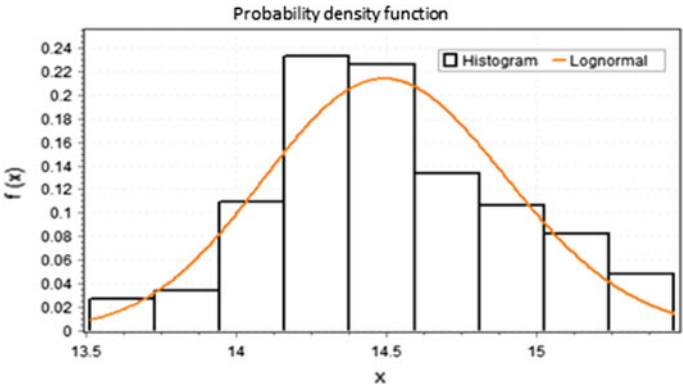


Fig. 2.13 Histogram and pdf adjusted for the second dominant area (reproduced with the kind permission of Siderúrgica del Mediterráneo S.A.)

- Whether or not an operation requires a secondary process, a reoperation (Fig. 2.14 and Table 2.14).

2.7.2 Discrete Uniform Distribution Function

This is used when all the values in the $[i, j]$ range have an equal probability. It is employed as a first model, when we only have information about the limits of the range (Table 2.15).

2.7.3 Binomial Distribution Function

Binomial distribution is a discrete distribution that expresses the result of n separate experiments. It is essentially the sum of n Bernoulli experiments. Let us suppose

Fig. 2.14 Bernoulli probability function ($p = 0.6$)

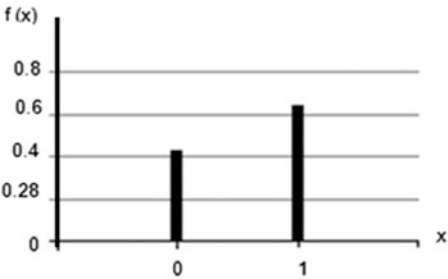


Table 2.14 Bernoulli distribution function

Bernoulli	<i>Bernoulli</i> (p)
Probability function	$f(x) = \begin{cases} 1-p & \text{if } x=0 \\ p & \text{if } x=1 \\ 0 & \text{o.c} \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i} & \text{if } 0 \leq x \leq n \\ 1 & \text{if } x > n \end{cases}$
Mean	p
Variance	$p(1-p)$

Table 2.15 Discrete uniform distribution function

Discrete uniform	<i>UD</i> (i, j)
Probability function	$f(x) = \begin{cases} \frac{1}{j-i+1} & x \in \{i, i+1, \dots, j\} \\ 0 & \text{o.c} \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 0 & \text{if } x < i \\ \frac{\lfloor x \rfloor - i + 1}{j - i + 1} & \text{if } i \leq x \leq j \\ 1 & \text{if } x > j \end{cases}$
Mean	$\frac{(i+j)}{2}$
Variance	$\frac{(j-i+1)^2-1}{12}$

that an experiment that has two possible results is done n times ($n > 0$). Also, let us suppose that the probability of obtaining a particular result, (let us call it result a) for any experiment is p , and the probability of the other result is $q = 1 - p$ (let us call it result b).

Therefore, result a may appear a number of times between 0 and n , as can result b . Binomial distribution specifies the probability of result a occurring k times. Some phenomena that can be defined using this distribution are:

- The number of defective pieces in a batch.
- The number of customers of a particular type that enter the system (Table 2.16).

2.7.4 Poisson Distribution Function

The frequency of the appearance of events in an arrivals process can be formalized by specifying the time between two successive arrivals or the number of arrival events per range.

Table 2.16 Binomial distribution function

Binomial	$Bin(n, p)$
Probability function	$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0, 1, \dots, n\} \\ 0 & o.c \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{i=0}^{[x]} \binom{n}{i} p^i (1-p)^{n-i} & \text{if } i \leq x \leq j \\ 1 & j < x \end{cases}$ Where $[x]$ indicates the largest integer $\leq x$
Mean	np
Variance	$np(1-p)$

- *Time between 2 successive arrival events:* in general, the time between two consecutive independent arrival events usually responds to an exponential distribution.
- *Number of arrival events per range:* instead of describing the time between arrival events, the number of events in a range of constant time is described. Note, for example, that it is not possible to describe by means of an exponential distribution the arrival of material at a production unit when it is transported on *pallets* with a number of variable pieces, as the time between the arrival of one piece and the next one is 0. Poisson distribution is one of the most used to describe this type of behavior. This distribution was originally developed for modeling the phone calls of a telephone exchange. Other phenomena that can be modeled are:

1. The number of temporary entities that arrive per unit of time.
2. The total number of defects in a piece.
3. The number of times that a resource is interrupted per unit of time (Table 2.17).

2.7.5 Geometric Distribution Function

Geometric distribution describes the number of experiments with p probability of success, which must be carried out until a particular result is obtained. Some examples of phenomena that can be modeled with this distribution are:

- The number of machine cycles until it breaks down
- The number of pieces inspected until one is found with defects
- The number of customers served until one of a particular type is found (Table 2.18).

Table 2.17 Poisson distribution function

Poisson	$Poisson(\lambda)$
Probability function	$f(x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x \in \{0, 1, \dots\} \\ 0 & o.c. \end{cases}$
Cumulative distribution	$F(x) = \begin{cases} 0 & x < 0 \\ e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} & 0 \leq x \end{cases}$
Mean	λ
Variance	λ

Table 2.18 Geometric distribution function

Geometric	$Geom(p)$
Probability function	$f(x) = \begin{cases} p(1-p)^x & x \in \{0, 1, \dots\} \\ 0 & o.c. \end{cases}$
Cumulative distribution	
Mean	$\frac{(1-p)}{p}$
Variance	$\frac{(1-p)}{p^2}$

2.8 Development of a Statistical Model

The development of a statistical model is very important for the design of a simulation model. In the case of discrete event simulation there are items such as the arrivals of customers or elements of interest into a system (that is going to be simulated), the times between the arrivals of the customers or elements into the system, time in the system, service time etc. These concepts shall be covered in more detail in later chapters.

The steps that a statistical model contains are:

1. Collection and analysis of the data
2. Adjustment of a distribution function
3. Validation of the adjustment

Example 2.3

1. Data collection

In modeling the random part, only the data referring to the process have to be recorded, without considering either the causes of the random activity or its effect. The time between arrivals at a tollbooth is presented below (Table 2.19).

2. Data analysis

IID: in simulation we assume that the values of the data samples are IID: Independent Identically Distributed Values, which means:

Independent: the set of values is not correlated

Identically distributed: follow the same probability distribution (Fig. 2.15).

Table 2.19 Time between arrivals MM1 in sec*100

0.50	3.35	20.85	7.81	0.44	0.03	3.82	7.09	3.02	2.80
2.08	6.53	52.53	10.23	0.76	0.00	28.21	15.51	4.86	10.41
5.25	11.67	46.23	28.06	6.05	4.82	46.36	2.90	5.47	17.42
7.20	41.15	9.54	4.88	19.10	9.17	0.83	7.43	9.98	4.11
10.28	23.44	6.19	2.39	7.57	12.97	12.62	7.65	18.49	6.95
1.08	9.89	5.49	2.16	14.18	11.89	12.73	0.51	14.61	27.01
1.91	18.77	4.98	6.41	1.45	1.71	5.21	2.89	8.38	3.50
2.86	17.60	4.89	11.74	15.31	36.64	3.62	21.78	2.15	6.70
17.13	0.11	17.58	1.30	2.44	9.59	1.74	5.02	6.46	18.76
1.49	7.92	4.03	3.13	1.67	23.31	3.13	9.35	0.10	0.51

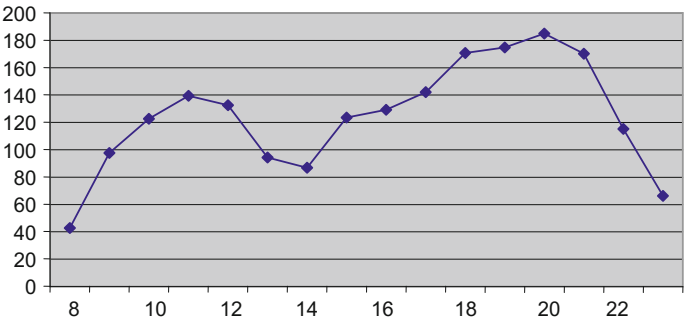


Fig. 2.15 Number of arrivals per time segment

3. Validation of the adjustment

To be able to validate the adjustment of the data, goodness of fit tests are used such as the chi square, Kolmogorov Smirnov or the Anderson Darling test. At the present time, one or more than these tests are built into popular simulation packages, such as Promodel, which has Stat Fit that, as mentioned in said software, has the following use:

- Curve fitting. It helps you to find the best distribution to represent the data. **Stat::fit** uses the most commonly known goodness of fit tests such as:
 - a. Anderson-Darling
 - b. Chi-Square
 - c. Kolmogorov-Smirnov
- Determining the number of replications to run a simulation model.
- Determining the size of the sample for taking process and transportation times.

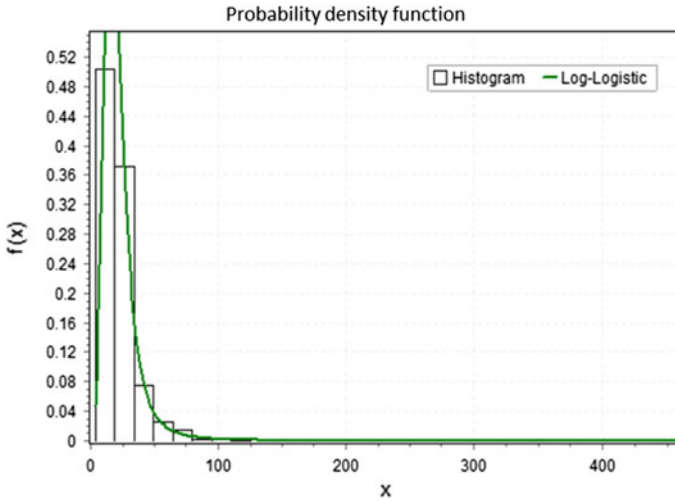


Fig. 2.16 Histogram and pdf LogLogistic

- Graphing the input data, graphing all the probability distributions that can be used, drawing up descriptive statistics for the data.
- An excellent option to disseminate statistical thought.¹

In the case of SIMIO, this is left to the user's criteria. It does not have built-in data adjustment software.

Example 2.4 Modeling of the time between arrivals at a port terminal

Figure 2.16 shows the histogram and the adjusted probability density function (LogLogistic, $\alpha = 3.4$ $\beta = 19.23$) for a sample of 1687 values for times between arrivals in hours of boats in a port terminal of the Port of Barcelona (Spain).

A total of 8 values from the sample have a time between arrivals of more than 110 h. If the objective of the study is to analyze the behavior of the terminal in periods of normal or high workload, the elimination of these 8 values in the sample (0.5% of the total) will not have a significant impact on the results of the study, given that they correspond to periods of little activity. Figure 2.17 shows the histogram for the real sample without the above 8 values and the adjusted probability density function (LogLogistic, $\alpha = 3.6$ $\beta = 19.01$).

¹<http://www.promodel.com.mx/statfit.php>.

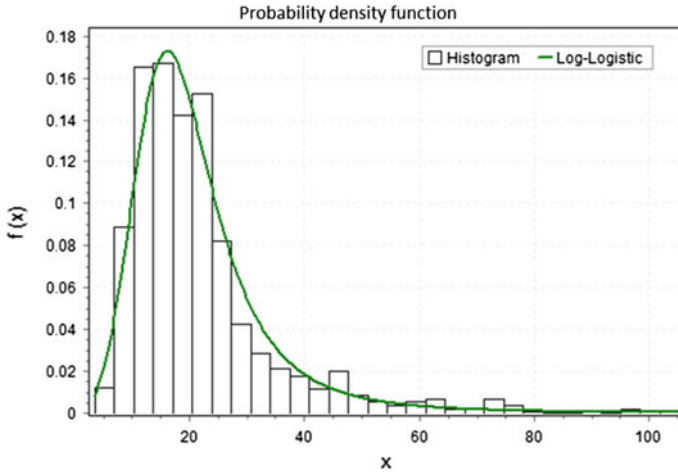


Fig. 2.17 Histogram of the real sample

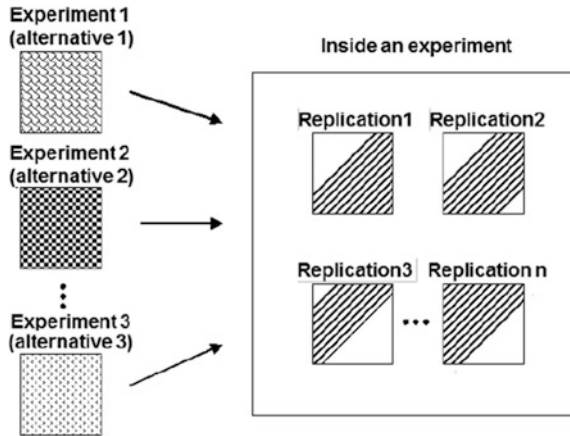
2.9 Statistical Analysis of the Simulation Results

Once the simulation has been done it is important for a statistical analysis of the results of the simulation to be carried out. In many simulation studies too much time and money is spent on the development of the model and the programming and very little effort is made to properly analyze the results. In fact, a very common mode of operation is to make a single simulation run of a somewhat arbitrary length and then use the results of said run as true, however, owing to the randomness of the variables there can be large variations in the results that entail wrong inferences about the real problem.

That is why it is necessary to develop one or more experiments and one or more replications in each one of them. We can see this pattern in schematic form Fig. 2.18.

Figure 2.18 illustrates the difference between an experiment and a replication. As can be appreciated, the replications are in the same experiment. When carrying out a simulation study, a series of parameters are considered, for example, the time between arrivals. Along these lines, a change is made in the time between arrivals or a variation can also be considered in the number of servers. The implementation of each one of these changes corresponds to the development of different experiments to evaluate different scenarios. The replications are run from the same model without making any changes to the parameters: These replications will give different results owing to the use of different series of random numbers. Every replication produces statistical results that differ from those produced by other replications, and said results can be analyzed throughout the entire set of replications.

Fig. 2.18 Experiments and replications in a simulation model. *Source* Law 2006



In this part of the experiment we are interested in being able to answer the following questions:

1. How many experiments need to be carried out?
2. How many replications are needed for the simulation?

The random nature of the simulation entails a stochastic process in the results that we can call Y_1, Y_2, \dots, Y_m for a simple run. For example, Y_i could be a delay in the arrival of the i -th job in the simple queuing system. C_i is also defined as the cost of operating an inventory in the i -th month. Y_{ij} are random variables q_j that are not, in general, independent or identically distributed (IID). Thus, many of the formulas of classic statistics shall not be directly applicable to the analysis of the simulation's output data.

Example 2.5 (Law 2010)

A queuing system is considered where Y_1 is the delay of customer 1, Y_2 is the delay of customer 2, etc. In this system the delays in the queue shall not be independent, as a long delay for a customer waiting in a queue will tend to even further delay the next waiting customer. Assuming that the simulation starts at zero time without any customers in the system, as usually done. Then, delays in the queue at the start of simulation will tend to be shorter than delays at the end, hence the delays are not identically distributed.

Let $y_{11}, y_{12}, \dots, y_{1m}$ be results from running the simulation for random variables Y_1, Y_2, \dots, Y_m with the specific random numbers u_{11}, u_{12}, \dots . If the simulation is run with a different set of random numbers u_{21}, u_{22}, \dots , then a set of different results $y_{21}, y_{22}, \dots, y_{2m}$ is obtained from the same random variables. Y_1, Y_2, \dots, Y_m .

The two sets of results are not equal as different random numbers were used in two runs and, accordingly, two different samples were produced from the same probability distributions.

In general, assuming that n independent replications or simulation runs each one of size m , that would, in the example, mean simulating the delays of m customers, obtaining the following observations:

2.10 Conclusions

This chapter shows some of the basic and essential concepts of statistics in order to develop a simulation model with everything it requires, initial data, data adjustment, runs and experiments, model verification and validation, as well as the interpretation of the results.

As simulation is a stochastic process there is always something more we could say about it. If the reader would like to further information, we recommend they consult the following references.

References

- Altioik, T., & Melamed, B. (2007). *Simulation modeling and analysis with ARENA*. New York: Academic Press.
- Banks, J. (ed.). (1998). *Handbook of simulation*. New York: Wiley.
- Barton, R. (2004). Designing simulation experiments. In *Proceedings of the 2004 winter simulation conference* (pp. 73–79).
- Carson, J., & Banks, J. (1993). *Discrete- event system simulation*. Englewood Cliffs: Prentice Hall.
- Coss, B. R. (2003). *Simulación*. Limusa: Un enfoque práctico.
- Currie, C., & Cheng, R. (2013) A practical introduction to analysis of simulation output data. In R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill & M. E. Kuhl (Eds.), *Proceedings of the 2013 winter simulation conference* (pp. 328–341).
- Flores, I., & Elizondo, M. (2007). *Apuntes de simulación*, Facultad de Ingeniería UNAM.
- González, M. C. (1996). *Modelos y simulación*. UNAM: ENEP Acatlán.
- Gordon, G. (1991). *Simulación de sistemas*, 6ta. reimpresión de la 1ª. Edición, Diana.
- Guasch, A., Piera, M. A., Casanovas, J., & Figueras, Y. J. (2003). *Modelado y simulación: aplicación a procesos logísticos de fabricación y servicios*. 2a. ed., Barcelona, Ediciones UPC.
- Kelton, D., & Barton, R. (2003). Experimental design for simulation. In *Proceedings of the 2003 winter simulation conference* (pp. 59–65).
- Kolmogorov, A. N. Y., & Uspenskii, V.A. (1987). *Algorithms and randomness*, Ed. Teor. Veroyatnost. i Primenen.
- Law, A. (2003). How to conduct a successful simulation study. In *Proceedings of the 2003 winter simulation conference* (pp. 66–70).
- Law, A. (2004). Statistical analysis of simulation output data: The practical state of the art. In *Proceedings of the 2004 winter simulation conference* (pp. 67–72).
- Law, A. (2006). *Simulation modeling and analysis with expertfit software*. New York: Mc. Graw-Hill.
- Law, A. (2010). Statistical analysis of simulation output data: The practical state of the art. In *Proceedings of the 2010 winter simulation conference* (pp. 65–73).
- Law, A., & Kelton, D. (2000). *Simulation modelling and analysis*. New York: Mc. Graw-Hill.
- Lécuyer, P. (1990). Random numbers for simulation. *Communications of the ACM*, 33, Núm. 10, 85–97.
- Lehmer, D. H. (1951). Mathematical methods in large-scale computing units. In *Proceedings of a second symposium on large-scale digital calculating machinery* (pp. 141–146). Cambridge: Harvard University Press.
- Robinson, S., & Bhatia, V (1995). Secrets of successful simulation projects. In *Conference: Simulation conference proceedings, Winter WSC '95* Proceedings of the 27th conference on Winter simulation, pp. 61–67

- Skoogh, A., & Johansson, B. (2008). A methodology for input data management in discrete event simulation projects. In S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson & J. W. Fowler (Eds.), *Proceedings of the 2008 winter simulation conference* (pp. 1727–1735).
- Taha, H. (2004). *Investigación de Operaciones*, 7^a. ed., Prentice-Hall.
- Trybula, W. J. (1994). Building simulation models without data. *Proceedings of the IEEE International Conference of Systems, Man and Cybernetics*, 209–214.

Robust Modelling and Simulation

Integration of SIMIO with Coloured Petri Nets

De La Mota, I.F.; Guasch, A.; Mujica Mota, M.; Piera, M.A.

2017, XVII, 162 p. 112 illus., 70 illus. in color.,

Hardcover

ISBN: 978-3-319-53320-9