CHAPTER **3**

Elements of Simulation

Introduction

It will be useful, at this point, to gather together and define a number of terms that occur frequently in the simulation of discrete systems.

Entities

Entities are objects of interest that are acted upon or created by the simulated processes. Customers, data packets, paperwork forms, components, and manufactured products are all examples of entities. Entities have attributes, such as size, color, type, and value that can determine routing through the simulated system and subsequent processing. The arrival times of entities, and their attributes, are gathered during model building phase of simulation. These arrivals may be periodic, scheduled, event triggered, or random. Entities may be aggregated or transformed into other entities by a process. Entities can also be created, and can be split into component entities, by processes.

Attributes

Attributes are properties of entities that are held to be significant for the purposes of simulation. Arrival rates, sizes, colors, and logical values are some examples of attributes. Some of these attributes will have an effect upon the direction of flow that entities take in a system or on their processing time in a given operation. An airline provides different check-in services to passengers, according to the ticket class held, for example. In this case, the class of ticket is an attribute of the passenger.

Arrivals

An arrival is the entry of an entity into a process domain. Arrivals have statistical properties in discrete systems. The time between successive arrivals is termed the interarrival time. If the arrivals are random, the interarrival time is modeled by a probability distribution function. An Exponential distribution would be a possible case. Simple cases are also possible, such as a single scheduled arrival, or a periodic arrival. Arrivals may also be triggered by a process event, such as a level or a logical condition. In such cases, the arrival is not random, since it is dependent upon another operation. The withdrawal for testing of every one-hundredth item from a production series, for instance, would create a non-random arrival at a testing station. Some arrivals may appear periodic, but are really random. The arrival of the morning newspaper appears to occur periodically, for example, but the exact time between arrivals will be statistically distributed, albeit with a narrow deviation from the mean value of 1440 minutes.

Activities

Activities are functions performed by process elements that use time or create events. They define what entities are doing during a time interval. Operations perform activities on their input entities. These operations may depend, to some extent, upon the attributes of the entities. The checkout of a basket of grocery items, for example, is an activity that uses an amount of time that varies with the number of items in the basket. Performance of activities requires resources.

Resources

Resources are auxiliary personnel or matériel required to process entities. Resources may be simply required for temporary use, such as a vessel to cook in, or may be consumed, such as broth to cook. Resources can also have failure and scheduled downtime properties. Resource failure is typically statistical, and is associated with a statistical model. The statistical failure of a conveyor would be an example. Resource downtime is sometimes scheduled, and is often periodic. Working shifts and weekends are examples. Resources can also have downtime for recycling after an amount of use, including a single use. A hospital bed would be an example. Resources are available to activities (operations) in certain specified quantities. When there are no resources available, for whatever reason, the activity dependent upon them will be delayed or paused. Bins of limited size for work in process would be an example, one that occurs by design in order to minimize in-process inventory. When a bin is full, the process becomes blocked. Resources can also have parameters, such as sizes or rates of delivery. The sizes of bins for work-in-process would be an example.

Paths

Paths are the routes taken by entities or resources in a system. They may be fixed, random, or dependent. A fixed path is determined by the design of the process or system. A random path, on the other hand, is taken by an entity or resource according to a statistical distribution. The paths taken by good and defective product after testing would be examples of this. A dependent path is one taken as the result of a logical operation upon an attribute of an entity. Paths have attributes, as well, such as speed, capacity, etc. In paths for which speed is an attribute, transit time delay will be inverse to the speed. A conveyer would be an example. Paths can also have conflicts (interferences), such as the crossing paths of fork lift trucks in a warehouse. Knowing the dimensions of the region of conflict and the intensity of use, a statistical model of likely interference delays can be developed.

State

The state of a system is the collection of variables needed to describe the future behavior in response to input. The state of a queuing system includes the queued entities and the entities being serviced. An arrival alters the state of the queue, and thus of the system. The completion of a service also affects the queue, since a new queued entity is withdrawn for service. The serviced entity also progresses to the next queue in the system, altering its state in turn. If we were to halt a simulation and then restart it, it is the state of the system that would have to be saved and then restored before the simulation could resume.

Events

An event is an instantaneous occurrence that can alter the state of the system. There are two types of events. *Exogenous events* are those having their origin external to the system under study. They arrival of a data packet at a router would exemplify such an event. *Endogenous events* are those having their origin internal to the system under study. The completion of service of an entity would be an example of such an event. Simulation events are saved in an event list until their time to be executed occurs. In an event driven simulation program, the simulation clock is advanced to the next event time at each computational cycle. This is in contrast to older, and slower, time driven simulation systems, in which the simulation clock advances in fixed intervals.

Components

Components are parts of a system. They can be process elements or entire subsystems.

Parameters

Parameters are quantities that can take on arbitrary values, to be assigned by the operator of the model or simulation. They typically remain constant for one simulation study. The arrival rates of entities and service times, usually specified by statistical distributions, would be examples. However, the maximum queue length of a process operation should also be considered a parameter, because it could be varied between simulation runs in order to determine if there is any performance sensitivity to this parameter under the simulated conditions.

Variables

Variables are quantities that can assume only those values permitted by the form of the model. An example might be a discrete variable such as CreditCardType, which could take on specified values of “Regular,” “gold,” and “Platinum.” This style of formal type definitions serves to reduce run-time errors, especially in projects being assembled by teams. The length of a queue is also a discrete variable, because it increases or decreases by exactly one unit in response to an arrival or a completion of service, respectively. The number of customers being served is also a variable, because some queues may be empty. An empty queue idles its attached server, though servers can also be idled because of downtime for breaks or other causes.

Functional Properties

The functional properties of a process are the mathematical relationships that describe the behavior of the dependent variables of a process for the given parameters, initial state, and inputs. These are the representations of a process. A server exhibiting exponential service time would be an example.

Constraints

Constraints are limitations placed upon the permitted values of variables or upon the allocation of resources. A limit on queue length, a work in process bin size, or a maximum number of servers would qualify as constraints.

Criterion Function

A criterion function is a mathematical expression that is used to measure the performance of a system with respect to prescribed goals. A least sum of squared error measure, for instance, can be used to measure the degree to which a probability distribution correctly models process data gathered in the field. A minimization of operating cost criterion could also be used for the optimization of system design by simulated evaluation of competing process models or operating strategies

Basics of Queuing Systems

Introduction

The theory of queuing systems provides a basis for understanding the simulation of discrete systems. The basic queuing model, shown in Figure 3.1, consists of an arrival from a statistically distributed calling population, an input queue, and a server. The calling population is thought to contain arrivals with interarrival times distributed according to a specified probability density function. This population is like a set of balls in a lottery, which are to be withdrawn. Each ball contains a number, which specifies the interarrival time. These numbers are distributed according to a mathematical function, some numbers occurring more often than others. Adding a withdrawn interarrival time to the simulation clock time produces an arrival event that is scheduled to take place. Arrivals are queued, and entities to be serviced are withdrawn from the input queue. The service time is also a statistical function, and it may depend upon certain properties of the arriving entities. It may be helpful to think

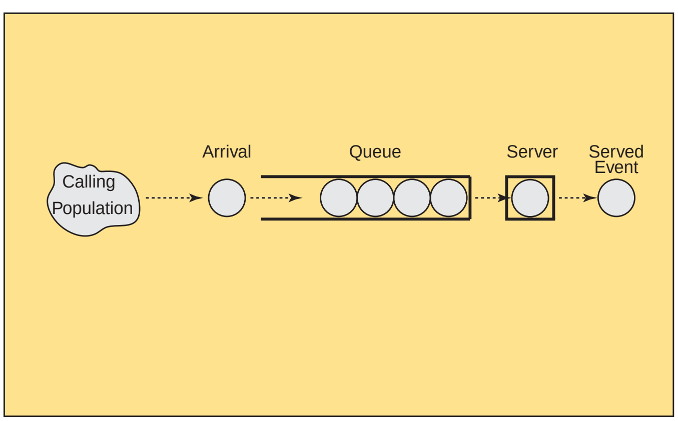

Figure 3.1 Basic queuing model

of drawing two balls for each arriving entity in a kind of double lottery. The first indicates the interarrival time, to be treated as indicated above. The second drawing, from a differently distributed population, indicates a parameter upon which the service time will depend, or it may indicate the service time itself for this entity. The queue server will hold the entity for this elapsed time after it is withdrawn from the queue.

It is important to understand that the *average* rate of process inputs cannot exceed the *average* service rate in a queuing system. That is, the average number of entities arriving per unit time cannot exceed the average number leaving after being serviced. If this condition were to be violated, the queue length would grow without limit until the queue would overflow at its maximum length. The *instantaneous* rates can vary, however, and it is this instantaneous fluctuation that the queue is designed to absorb. If arrivals and service completions were synchronized, there would be no need for a queue at all. Imagine a doctor's office where patients always arrive at the appointed time and the doctor is always ready at exactly this time, having just completed the previous case. There would be no need for a waiting room, and the doctor would serve the maximum possible number of patients each day. It is the variability of arrivals and service times that necessitates a queue, and this variability is expressed by the corresponding statistical distributions. In general, the larger the deviation from the mean (average) value, the larger the average length of the queue will be. This queue length can be significant, for instance, if entities unprocessed at the end of a shift need to be specially processed at high cost. Such costs can be a factor in determining when to add shifts, overtime labor, or outside contractors.

Static Queuing Analysis

A basic queuing system model was shown in Figure 3.2. For the present discussion, inputs will be assumed to arrive at a Poisson distributed average rate of **l** units per unit time. This discrete distribution is often used to model the number of entities arriving during a unit of time and will be discussed later. It has a single parameter, the mean value. Then **l** will be the mean arrival rate. It will also be assumed that the server will service available units at an exponentially distributed rate of **m** units per unit of time. (This distribution will also be discussed later) If an event occurs according to a Poisson distributed rate, the time between events will be exponentially distributed. This distribution has a single parameter, the mean service time. Thus, if customers arrive at an average rate of 20 per hour and are served at an average rate of 25 per hour, we will have, $l = 20$ and $m = 25$. Now, if customers arrive at precisely spaced intervals of $t_a = 1/l$ (0.05 hours), and it always takes exactly $t_s = 1/m$ (0.04 hours) to service them, there would be no queue. Each customer would be serviced immediately upon arrival, and the service would be completed 0.01 hours before the next customer arrival. This would be a very rare occurrence. Normally a queue would form. This permits certain analysis to be performed.

Total time delay

The average total time delay, W, before completion of service on a given arrival is the sum of the time spent in the queue, W_q, plus the time spent in service, W_s. Mathematically,

$$W = W_s + W_q \tag{3.1}$$

We know, however, that the time spent in service is just $1/\mu$, because the service rate is μ . If there are **n** customers in the queue upon arrival, the time in the queue will be **n** times the unit service time. That is,

$$W = nW_s + W_s = (n+1)W_s \tag{3.2}$$

Using the given service time $1/\mu$,

$$W = (n+1)(1/\mu) \tag{3.3}$$

Thus the expected delay will depend upon the state of the queue and the mean service rate. This will be a familiar concept to those who have been stuck in highway traffic queues.

Number of entities in system

The expected (mean) number of entities in the system, L, will be equal to the arrival rate, l, multiplied by the expected total waiting time, W. This is,

$$L = \lambda W = \lambda/(\mu-\lambda) \tag{3.4}$$

Note that as the average arrival rate, λ, approaches the average service rate, μ, the length of the queue becomes infinite.

Number of entities in queue

The expected (mean) number of entities in the queue is given by,

$$L_q = \lambda W_q \tag{3.5}$$

In terms of the rates, the number of queued entities thus becomes,

$$L_q = \lambda^2/\mu(\mu-\lambda) \tag{3.6}$$

The total waiting time is, on average,

$$W = W_q + 1/\mu \tag{3.7}$$

Substituting for W_q,

$$W = 1/(\mu-\lambda) \tag{3.8}$$

These conclusions lead to Little's result:

$$L = \lambda W = \frac{\lambda}{\mu - \lambda} \tag{3.9a}$$

$$L_q = \lambda W_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{3.9b}$$

$$W_q = \frac{1}{\lambda} L_q = \frac{\lambda}{\mu(\mu - \lambda)} \tag{3.9c}$$

$$W = W_q + \frac{1}{\mu} = \frac{1}{\mu - \lambda} \tag{3.9d}$$

There are four variables and three independent variables. Thus, if one is known, the rest can be calculated.

Statistical Models

The previous discussion dealt only with *average* values and could not be used for the calculation of instantaneous queue lengths required in simulation studies. Real world behavior may require more parameters, other than average values, and may not be limited to only Poisson input and Exponential service models. It will be the task of the model developer either to fit a well known mathematical model to process data or to formulate an entirely new model to fit the data. Discrete system simulations will then use these probability models to generate instantaneous simulated process states.

The fundamental statistical model is the probability density function of a random variable **f(X)**. This describes the probability that the random variable takes on the value **X**. Table 3.1 shows the number of cars entering a car wash with the specified interarrival times, designated by **X**, on a given day. Four-minute spaces occur six times, for instance. For these

Table 3.1 Mean value of car wash probability distribution

X	*Frequency*	*Probability*	*X*Prob*
1.000	0.000	0.000	0.000
2.000	1.000	0.032	0.065
3.000	1.000	0.032	0.097
4.000	6.000	0.194	0.774
5.000	13.000	0.419	2.097
6.000	7.000	0.226	1.355
7.000	3.000	0.097	0.677
8.000	0.000	0.000	0.000
9.000	0.000	0.000	0.000
10.000	0.000	0.000	0.000
		Mean =	**5.065**

Table 3.2 Calculation of variance of collected data

X	*Frequency*	*Probability*	*X*Prob*	X - E(X)	[X-E(X)]^2*P(X)
1.000	0.000	0.000	0.000	-4.065	0.000
2.000	1.000	0.032	0.065	-3.065	0.303
3.000	1.000	0.032	0.097	-2.065	0.137
4.000	6.000	0.194	0.774	-1.065	0.219
5.000	13.000	0.419	2.097	-0.065	0.002
6.000	7.000	0.226	1.355	0.935	0.198
7.000	3.000	0.097	0.677	1.935	0.363
8.000	0.000	0.000	0.000	2.935	0.000
9.000	0.000	0.000	0.000	3.935	0.000
10.000	0.000	0.000	0.000	4.935	0.000
	E(X)=	**5.065**		**Var =**	**1.222**

data, the probability that the input takes on the value **X** = 5 is 13/(1+1+6+13+7+3) or 13/31. This is the probability that the time between cars entering the carwash is 5 minutes. If we were to plot the number of times **X** takes on each value indicated in Table 3.1, the result would be the *frequency distribution* for **X**. The corresponding discrete *probability density function* could be calculated from this by dividing the frequencies each by 31, the total of the samples. This has been graphed in Figure 3.2. The Probability column in the table is generated by dividing the numbers in the frequency column by 31. Each number in he rightmost column is the product of the corresponding numbers in the Frequency and Probability columns. Adding this column then gives the expected (mean) value shown. The expected (mean) value **E(X_i)** of the probability density function of the discrete random variable **X_i (I=1,2,…,n)**, sometimes called the average, is defined as,

$$E(X_i) = \sum_j x_j p_{x_i}(X_i = x_j) \tag{3.10}$$

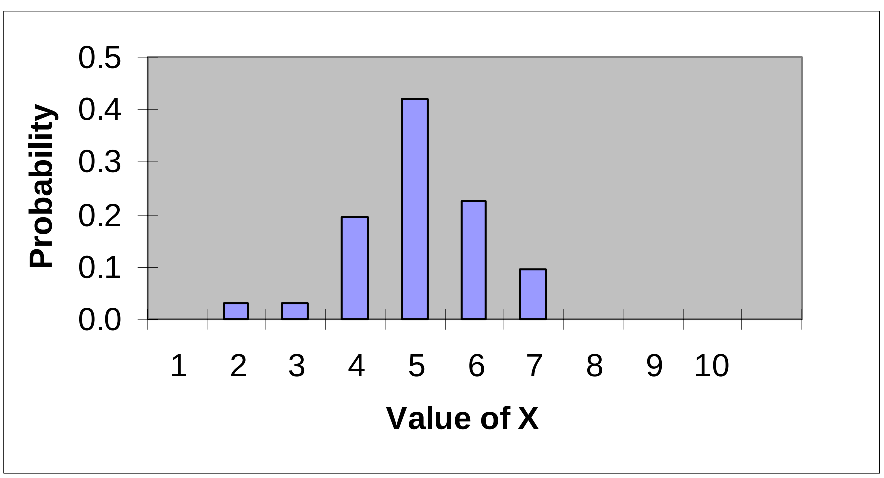

Figure 3.2 Probability density function for collected data

Table 3.3 Calculation of normally distributed model for collected data

X	*Freq.*	*Prob.*	*X*Prob*	X - E(X)	[X-E(X)]²*P(X)	Norm(X,5.065,1.1054)
1.000	0.000	0.000	0.000	-4.065	0.000	0.0004
2.000	1.000	0.032	0.065	-3.065	0.303	0.0077
3.000	1.000	0.032	0.097	-2.065	0.137	0.0630
4.000	6.000	0.194	0.774	-1.065	0.219	0.2269
5.000	13.000	0.419	2.097	-0.065	0.002	0.3603
6.000	7.000	0.226	1.355	0.935	0.198	0.2524
7.000	3.000	0.097	0.677	1.935	0.363	0.0780
8.000	0.000	0.000	0.000	2.935	0.000	0.0106
9.000	0.000	0.000	0.000	3.935	0.000	0.0006
10.000	0.000	0.000	0.000	4.935	0.000	0.0000

These calculations are, so far, independent of the shape of the probability density function. In words, this says that the mean value is equal to the sum of the probabilities, each multiplied by the corresponding value of **X**. For the given data, this would be found by adding up the probabilities of Table 3.1, each multiplied by its corresponding value of **X**. The products, shown in the table, add up to the mean value **E(X)** = 5.065. This is the mean interarrival time between the cars. The significance of the mean is that it is the most probable value of the random variable.

The variance σ_X^2 of a distribution function having mean **E(X)** is given by,

$$\sigma_X{}^2 = E[X - E(X)]^2 = E(X^2) - E^2(X) \tag{3.11}$$

That is, the variance is the sum of the squared deviations from the mean value, each multiplied by the probability of that value. This is 1.222, for the tabulated data, as calculated in

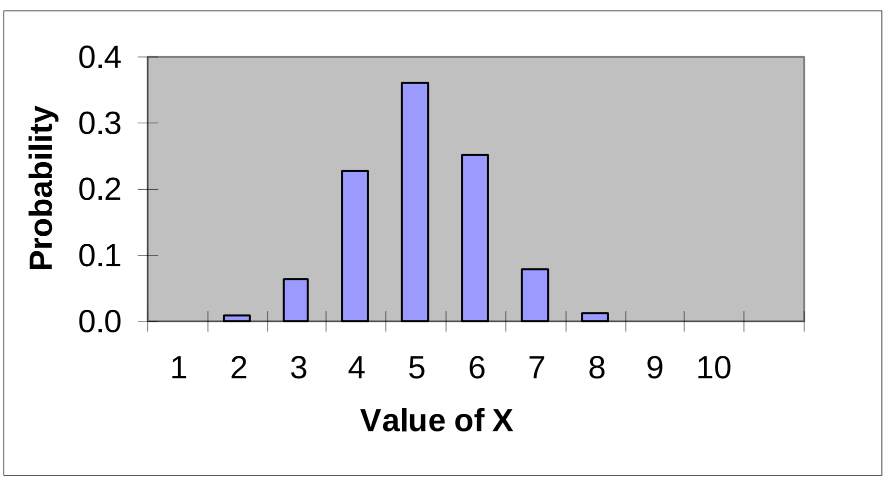

Figure 3.3 Normally distributed data modeling collected data

Table 3.2 and shown in Figure 3.2. The mean value 5.065 was used for **E(X)** in the calculations shown.

The significance of the variance is that it is the most probable range of deviation from the mean. Note that a Normal distribution with a mean of 5.065 and standard deviation of 1.1054 (the square root of the variance) has a probability distribution very close to that of the tabulated data, as calculated in Table 3.3 and shown in Figure 3.3 (Compare Prob. and Norm. columns). In this case we can say that we have a model for the tabulated data. This model can be used in a simulation in order to generate data similar to that gathered in the field.

Note that the Normal distribution is more symmetrical than the distribution measured in the field, and the probabilities differ slightly. When an ideal distribution fits data gathered in the field closely, the model probably needs no further development for use in simulation. In the case shown, a Normal distribution with a mean of 5.065 and a standard deviation of 1.1054 would probably be considered good enough to represent the data in the simulation model for a car wash.

A method for model development has thus emerged. The method begins with collection of field data for interarrival times or service times. These are then grouped, so that there are a number of occurrences of each time in each group. The number of occurrences are then normalized to probabilities by dividing each by the total number of samples. This gives a probability of occurrence for each specified time. These are then plotted and compared with available standard probability density functions. Some adjustment of the parameters of the standard functions may be needed to get a close fit. The fitted standard function can then be used to generate the events in a simulation.

Standard Distributions

Mathematicians have evolved a number of probability distributions that can be used to model business processes. These distributions are of two types: continuous and discrete. Continuous distributions represent the probability of a random variable **X** taking on any real number value. In a discrete distribution, the random variable **X** can take on only integer (whole number) values. Thus continuous distributions are frequently used to model interarrival times and service times, while discrete distributions could be used to model the number of items or entities arriving at a given time or in a specified time interval. The Normal, Exponential, and Weibul distributions are examples of continuous distributions. The Binomial, Poisson, and Geometric distributions are representative of discrete distributions.

Poisson Distribution

The Poisson distribution is often used to characterize the number of events taking place over a specific time or a specific sample size, such as the number of cars arriving at a toll gate in a 1 minute interval or the number of items in a shopping cart. Note that, the time between events occuring according to a Poisson distributed rate will be Exponentially distributed. The Poisson distribution has a single-parameter (λ), and requires a whole number input. The Poisson distributed probability of **X** events taking place in a specified time interval is given by,

$$p(X) = \frac{e^{-\lambda}\lambda^{X}}{X!} \qquad 3.12$$

The Mean and Variance of the distribution are both equal to λ. Some examples for different values of λ are shown in Figure 3.4a and 3.4b. Notice that the mean of each distribution lies at the chosen value of λ. Alternatively, the Beta distribution sets non-zero discrete limits.

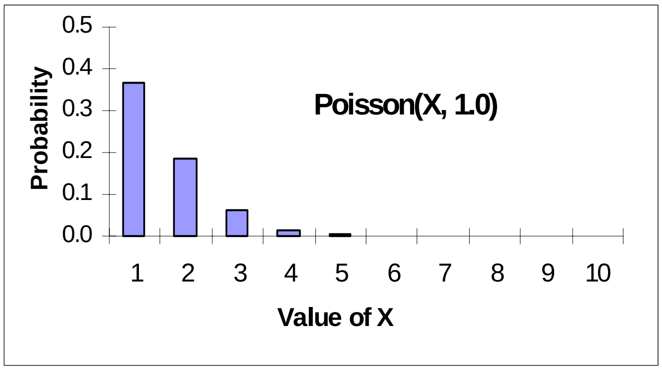

Figure 3.4a Poisson probability distribution for λ = 1

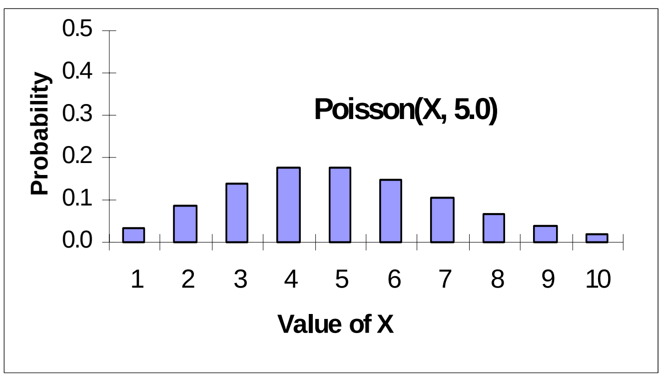

Figure 3.4b Poisson probability distribution for λ = 5

<u>Exponential Distribution</u>

The Exponential distribution is often used to characterize the interarrival time of entities entering the service queue or the service time for an entity after removal from the queue. The time between customer arrivals can frequently be modeled by the exponential distribution. The Exponential distribution has also been used to model the time between failures of a piece of equipment. As has already been mentioned, the time between Poisson distributed events is Exponentially distributed. The Exponential distribution has a single parameter (μ) and requires a real number input **(X)**. The Exponentially distributed probability of an input taking on value **X** taking is given by,

$$p(X) = \frac{1}{\mu} e^{-X/\mu} \qquad 3.13$$

The Mean of the distribution is equal to μ, and the Variance is equal to μ^2. An example for mean μ = 0.5 is shown in Figure 3.5.

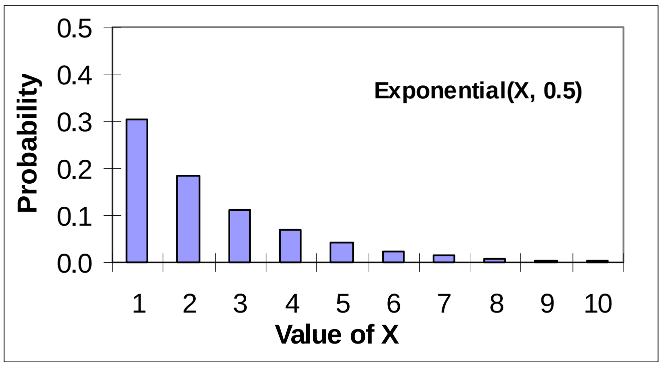

Figure 3.5 **Exponential probability distribution for parameter μ = 0.5**

The cumulative distribution function **F(X)** is given by

$$F(X) = 1 - e^{-X/\mu} \qquad 3.14$$

and represents the probability that the event occurs between **0** and the given **X** value. Subtracting two values of **F(X)** gives the probability that **X** lies between the lower and upper values selected. Thus, for discrete (integer) values of **X**,

$$p(X) = F(X + 0.5) - F(X - 0.5) \qquad 3.15$$

This fact can be used to generate the equivalent of a discrete distribution.

The Normal Distribution

The Normal distribution is commonly used to characterize events that have a time of occurrence, or other measurable variable, that takes on values about a central, Mean, value. The probability distribution is a continuous distribution having two parameters, the Mean and the Standard Deviation. **X**. The Normally distributed distributed probability of an input taking on value **X** taking is given by,

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad 3.16$$

where the Mean of the distribution is equal to μ, and the Variance is equal to σ^2. An example for $\mu = 5$ and $\sigma = 0.1$ is shown in Figure 3.6. If negative values cause problems, the Lognormal distribution can be used.

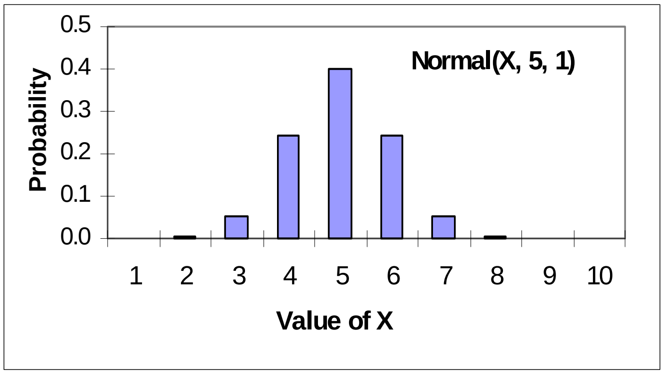

Figure 3.6 **Normal probability distribution for parameters μ= 5 and σ=1.**

As before, subtracting two values of the cumulative distribution **F(X)** will give the probability that **X** lies between the lower and upper values selected. That is, for discrete (integer) values of **X**,

$$p(X) = F(X + 0.5) - F(X - 0.5) \qquad 3.17$$

This fact can be used to generate the equivalent of a discrete distribution.

Summary

The basic queuing model for an element of a discrete process, is characterized by entity arrival rates and service times, each represented by a statistical distribution. As the average arrival rate for a given process approaches the average service rate, the number of entities in the queue will increase. Probability distributions can be obtained by recording the interarrival times of entities and the service completion times for servers. The number of occurrences of these times are then normalized to probabilities by dividing by their frequencies of occurrence. The corresponding probability density function is then plotted and compared with available standard functions. To complete the model, the parameter(s) of the standard function is (are) adjusted for best fit to the experimentally observed data.

Questions

(1) Identify the entities, attributes, activities, resources, paths, and states, for a carwash.

(2) Identify the entities, attributes, activities, resources, paths, and states, for a "Seattle style" coffee shop.

(3) Identify the entities, attributes, activities, resources, paths, and states, for a warehouse.

(4) Identify the entities, attributes, activities, resources, paths, and states, for a purchasing system.

(5) Identify the entities, attributes, activities, resources, paths, and states, for a production system for manufacturing and packaging cookies.

(6) Identify the entities, attributes, activities, resources, paths, and states, for a customer service system.

(7) Develop a spread sheet to verify the calculations and graphs shown for the data of Table 3.1.

Design Homework

(1) The number of customer groups of size 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 entering a restaurant for the evening was found to be 16, 14, 9, 5, 2, 1, 0, 0, 0, 0, respectively. Please make a statistical model to fit these data.

(2) The number of shopping baskets containing 1 - 20 items arriving at the express check-out of a supermarket was found to be 1, 4, 9, 13, 18, 16, 14, 10, 7, 4, 2, 1, 1, 0, 0, 1, 0, 1, 0, 0, respectively. Please make a statistical model to fit these data.

(3) The time people take to complete their transactions at an automatic teller machine, from 3 to 12 minutes was observed. The number of people taking each of these times was recorded to be 37, 14, 5, 2, 1, 0, 0, 0, 0, 0, respectively. Please make a statistical model to fit these data.

(4) Find and fit a statistical model for the data tabulated below.

Bin	*Frequency*
1	0
2	1
3	1
4	6
5	13
6	7
7	3
8	0
9	0
10	0
More	0

Laboratory Homework

(1) Record the arrival time (time of day) of customers at a car wash. Do this around the middle of a busy day. Then repeat this procedure at another time of day, or another day, in which the car wash is not as busy. Be sure to record this for at least 50 customers in each case.

(2) Tabulate the set of recorded interarrival times and processing times in a spread sheet, by subtracting the recorded times of day. Use the histogram function of the spread sheet program to plot discrete distributions for the processing times in each of the operations.

(3) Make a statistical model that fits the data presented in Problem 2 as closely as possible.

(4) Record the arrival time (time of day) of customers at a "Seattle style" coffee shop. Do this early on a busy morning. Then repeat this procedure at another time of day, or another day, in which the coffee shop is not as busy. Be sure to record this for at least 50 customers in each case. Save the data in a spreadsheet for later use.

(5) Tabulate the set of recorded interarrival times and processing times in a spread sheet, by subtracting the recorded times of day. Use the histogram function of the spread sheet program to plot discrete distributions for the processing times in each of the operations.

(6) Make a statistical model that fits the data presented in Problem 5 as closely as possible.

(7) Make a statistical model for the number of items in a shopping basket. Use the methods described in the problems above.

(8) Make a statistical model for the time people take to complete their transactions at an automatic teller machine. Use the methods described in the problems above.