Chapter 2

System Element Models

2.1 Probability Distributions as Models

In building a suitable model for a queueing system, we start with its elements. Of the elements mentioned in Chapter 1, number of servers, system capacity, and discipline are normally deterministic (unless, the number of available servers becomes a random variable, which is also possible in some cases). But there are uncertainties related to arrivals and service which result in the underlying processes being stochastic.

The similarity of the arrival and service processes can be brought out by identifying similar components, such as inter-arrival times and service times; arrival epochs and departure epochs.

Of these pairs departure epochs are *almost always* from a nonempty system, whereas arrival epochs are *mostly* independent of the state of the system (exceptions are possible). Therefore, first we discuss the possibilities of using certain probability distributions to represent the process of inter-arrival times and service times. In the case of Poisson process discussed below, it is also convenient to consider the distribution of the number of events occurring in a given length of time.

To start with, we should note that depending on the properties of the basic process and convenience, we may use either continuous or discrete distributions. In many situations continuous distributions may be easier to handle analytically (algebra of discrete distributions could be cumbersome.); nevertheless, it is worthwhile to note that continuous and discrete models are mutual analogs and most of the properties carry through in both cases.

Keeping a common notation we use $Z_1, Z_2, \ldots$ as nonnegative random variables representing either inter-arrival times or service times of consecutive customers. Further, let

$$F(x) = P(Z_n \leq x), \quad n = 1, 2, \ldots$$

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_2

We also assume that $\{Z_n\}_{n=1}^{\infty}$ are independent and identically distributed random variables. Let

$$E[Z_n] = b, \quad n = 1, 2, \ldots$$

and define the Laplace–Stieltjes transform of $F(x)$ as

$$\psi(\theta) = \int_0^\infty e^{-\theta x} dF(x) \quad Re(\theta) \geq 0.$$

Clearly, we get

$$-\psi'(0) = b.$$

It should be noted that when b is the mean inter-occurrence time, $1/b$ is the rate of occurrence of the event.

In considering the suitability of a probability model for a random phenomenon, moment properties of the model distribution become useful. Many times the first two moments appear as the parameters of the model. Furthermore, the first few moments describe the shape of the density curve, thus, making them suitable measures in selecting the model (e.g., coefficient of variation (CV) = s.d./mean; coefficient of skewness = (third moment)/(s.d.)^3; coefficient of kurtosis = (fourth moment)/(s.d.)^4).

The commonly used distribution models for arrivals and service are: deterministic (when arrivals are at specified time epochs, or inter-arrival times or service times are of constant length); exponential (as distribution models for inter-arrival times or service times); Poisson (as the distribution of the number of arrivals during a specified length of time); Erlang (as distribution models for inter-arrival times or service times); and variants of these distributions. We introduce deterministic, exponential, Poisson, and Erlang distributions in the following discussion and the remainder in Appendix A.

Deterministic Distribution (D)

Let

$$\begin{aligned} F(x) &= 0 \quad x < b \\ &= 1 \quad x \geq b \end{aligned} \tag{2.1.1}$$

We get $E(Z_n) = b$ and $\psi(\theta) = e^{-\theta b}$. Also, $V(Z_n) = 0$.

This seemingly simple distribution is suitable when arrivals take place at equal intervals of time (interval length b) or service takes exactly b units of time. In practice, however, it may be hard to achieve this exactness. Early or late arrivals, early or late service completions will be closer to reality. In such cases, the assumption of a deterministic distribution should be considered a reasonable approximation of the real system.

If we are interested in an exact model for the early or late occurrence of events, we may consider the displacement from the deterministic epoch as a random variable with some distribution like the uniform or the normal. Under these conditions, it is possible to have the kth scheduled event occurring later than the occurrence of the $(k+1)$th scheduled event.

Exponential Distribution, Poisson Process (M)

Let

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0. \tag{2.1.2}$$

Then we get,

$$\begin{aligned} f(x) &= \frac{d}{dx}F(x) = \lambda e^{-\lambda x} \\ E[Z_n] &= \frac{1}{\lambda} \end{aligned}$$

and

$$\psi(\theta) = \frac{\lambda}{\theta + \lambda}.$$

Also, $V(Z_n) = \frac{1}{\lambda^2}$ and $CV(Z_n) = 1$.

Let $X(t)$ be the number of events occurring in time t such that the inter-occurrence times have the distribution given by $F(x)$. Symbolically, for the stochastic process $X(t)$ we can write

$$X(t) = \max\{n | Z_1 + Z_2 + \ldots + Z_n \leq t\}.$$

Let

$$\begin{aligned} P_n(t) &= P(X(t) = n | X(0) = 0) \\ &= P(Z_1 + Z_2 + \ldots + Z_n \leq t) \\ &\quad -P(Z_1 + Z_2 + \ldots + Z_{n+1} \leq t), \end{aligned}$$

where $F_n(t) = P(Z_1 + Z_2 + \ldots + Z_n \leq t)$ is obtained as the n-fold convolution of $F(t)$ with itself. Using the Laplace transform of $F(t)$ we find

$$\int_0^\infty e^{-\theta t} dF_n(t) = \left(\frac{\lambda}{\theta + \lambda} \right)^n.$$

On inversion this gives

$$\begin{aligned} F_n(t) &= \int_0^t e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} dy \\ &= 1 - \sum_{r=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^r}{r!}. \end{aligned} \tag{2.1.3}$$

Thus, we get

$$\begin{aligned} P_n(t) &= F_n(t) - F_{n+1}(t) \\ &= [1 - \sum_{r=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^r}{r!} \\ &\quad -[1 - \sum_{r=0}^{n} e^{-\lambda t} \frac{(\lambda t)^r}{r!}] \\ &= e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \end{aligned} \tag{2.1.4}$$

which is a Poisson distribution with mean λt. Hence, $X(t)$ is known as a *Poisson process.*

Define the probability generating function of $X(t)$ as

$$\Pi(z,t) = \sum_{n=0}^{\infty} z^n P_n(t) \quad |z| \leq 1.$$

For the Poisson process we get

$$\Pi(z,t) = e^{-\lambda(1-z)t}.$$

Also, $E[X(t)] = \lambda t$ and $V[X(t)] = \lambda t$.

The Poisson process is a special case of the Markov process which is introduced in the next chapter. It is widely used in stochastic modeling because of its properties with reference to the occurrence of events and the properties of the exponential distribution representing the corresponding inter-occurrence times of events. Two of them are given below: (a) the first describes the *memoryless property* of the exponential distribution and (b) the second generates the Erlang distribution.

(a) When $P(Z_n \leq x) = 1 - e^{-\lambda x}$ $(\lambda > 0)$

$$\begin{aligned} P(Z_n \leq t + x | Z_n > t) &= \frac{P(t < Z_n < t + x)}{P(Z_n > t)} \\ &= \frac{[1 - e^{-\lambda(t+x)}] - [1 - e^{-\lambda t}]}{e^{-\lambda t}} \\ &= 1 - e^{-\lambda x}. \end{aligned} \quad (2.1.5)$$

The implication of this property is that if an interval, such as service time, can be represented by an exponential distribution and the interval is ongoing at time t, the remaining time in the interval has the same distribution as the original one, regardless of the start of the interval. This property is commonly known as the *memoryless* property of the exponential distribution.

(b) The discussion leading to equation (2.1.3) implies that the time required for the occurrence of a given number of Poisson events has a distribution given by that expression, i.e., if Y_n is the waiting time until the nth occurrence and $\{Z_1, Z_2, \ldots\}$ are the inter-occurrence times

$$\begin{aligned} Y_n &= Z_1 + Z_2 + \ldots + Z_n \\ F_n(t) &= P(Y_n \leq t) \\ &= \int_0^t e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} dy \end{aligned}$$

and

$$f_n(y) = e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} dy \quad (y > 0). \quad (2.1.6)$$

The distribution given by (2.1.6) is a gamma distribution with parameters n and λ. In queueing theory it is commonly called the *Erlang distribution* with scale parameter n. It is symbolically denoted by E_n. Equation (2.1.3) also establishes a useful identity

$$\int_y^\infty e^{-\lambda x}\frac{(\lambda x)^{n-1}}{(n-1)!}\lambda dx = \sum_{r=0}^{n-1} e^{-\lambda y}\frac{(\lambda y)^r}{r!}. \tag{2.1.7}$$

For modeling purposes, Poisson process is considered an appropriate model for events occurring "at random." The reasons for such a characterization rests on its properties described in Appendix A; specifically, independence of events occurring in nonoverlapping intervals of time, the constant rate of occurrence independent of time, the independent and identically distributed nature of the inter-occurrence times, and its relationship with the uniform distribution as expressed in (A.1.4) of Appendix A. The significance of the Erlang distribution stems from the phase interpretation that can be provided for generating a suitable arrival or service process.

Consider a Poisson arrival process and suppose a queueing system admits every kth customer into the system instead of all arrivals. Now the inter-arrival time between effective arrivals to the queueing system is the sum of k exponential random variables with mean $1/\lambda$, hence, it has the distribution given by (2.1.6). Similarly, consider a service process in which a customer goes through k phases of service, each phase being exponentially distributed with mean $1/\lambda$. The total service time has the distribution (E_k), given by (2.1.6) with n replaced by k.

To facilitate comparison with the Poisson and deterministic processes consider the Erlang distribution $F(x)$ with mean $1/\lambda$. This can be accomplished by starting with an exponential distribution with parameter $k\lambda$. Then we get

$$\begin{aligned} F(x) &= \int_0^x e^{-k\lambda y}\frac{(k\lambda)^k y^{k-1}}{(k-1)!}dy \\ f(x) &= e^{-k\lambda x}\frac{(k\lambda)^k x^{k-1}}{(k-1)!}. \end{aligned} \tag{2.1.8}$$

For $k = 1$, we have the exponential distribution, which generates a Poisson process. To determine the form of $f(x)$ as $k \to \infty$, we use its transform $\psi(\theta)$. We have

$$\psi(\theta) = \left(\frac{k\lambda}{k\lambda+\theta}\right)^k = \frac{1}{(1+\theta/k\lambda)^k} \to e^{-\theta/\lambda} \text{ as } k \to \infty.$$

The resulting transform is the transform of a constant $1/\lambda$, and hence generates the deterministic distribution given in (2.1.1). Depending on the values of k, even a moderately large value of k (e.g., $k = 10$ or 15) may be sufficient for the Erlang to exhibit the property of a deterministic distribution.

2.2 Identification of Models

In the formulation of a queueing model, one starts with the identification of its elements and their properties. The system structure is easily determined. What remains is the determination of the form and properties of the input and service processes. Four major steps are essential in this analysis (i) collection of data, (ii) tests for stationarity in time, (iii) tests for independence, and (iv) distribution selection and/or estimation.

2.2.1 Collection of Data

To estimate parameters of system elements, one has to establish a sampling plan identifying the data elements to be collected with reference to specific parameters. For instance, the number of arrivals in a time period gives the arrival rate or the mean inter-arrival time, which are reciprocals of each other. Sometimes there is a tendency to use empirical performance measures to estimate parameters intrinsic to the model. For instance, in an $M/M/1$ queue, noting that the traffic intensity (which is the ratio of arrival to service rate) provides the utilization factor for the system, we may use the empirical utilization factor as its estimate. Some of the pitfalls of this approach are indicated by Cox (1965) who notes that if ρ is the traffic intensity, the efficiency of this approach is given by $1 - \rho$. Also, see the discussion by Burke following Cox's article on the bias resulting from estimating the load factor in an $M/M/$s loss system as (average number of customers in systems)/(1-probability of loss).

The length and the mode of observation are problems of interest in a sampling plan. If the arrival process is Poisson, Birnbaum (1954) has shown that observing the system until a specific number of events have occurred gives a better sample than observing for a specific amount of time. But when nothing is known regarding the processes, no such statements can be made and the efficiency of different schemes should be considered in individual cases. Another aspect of the sampling plan is the mode of observations; for discussions of what are known as the snap reading method and systematic sampling, the reader is referred to Cox (1965), and page 86 of Cox (1962), respectively.

2.2.2 Tests for Stationarity

Cox and Lewis (1966) give a comprehensive treatment of tests for stationarity in stochastic processes. In addition to the treatment of data on the occurrence of events as a time series and the determination of second-order properties of the counting processes, they consider statistical problems related to renewal processes and provide tests of significance in some general, as well as some specific cases. Lewis (1972) updates this study and considers topics such as trend analysis of nonhomogeneous Poisson processes.

In many queueing systems (such as airport and telephone traffic), the nonstationarity of the arrival process leads to a periodic behavior. Furthermore, even though the process is nonstationary when the entire period is considered, it might be possible to consider it as a piecewise stationary process in which stationary periods can be identified (e.g., a rush hour). Under such circumstances, a procedure that can be used to test the stationarity of the process, as well as to identify stationary periods, is the Mann–Whitney–Wilcoxon test (see, for example, Conover (1971), or Randles and Wolfe (1979), or a test appropriately modified to handle ties in ranks as in Putter (1959)). The data for the test can be obtained by considering two adjacent time intervals $(0, t_1]$ and $(t_1, t_2]$ and observing the number of arrivals during such intervals for several time periods. Let $X_1, X_2, \ldots, X_n$ be the number of arrivals during the first interval for n periods, and let $Y_1, Y_2, \ldots, Y_m$ be the number of arrivals during the second interval for m periods (usually $m = n$). If F and G represent the distributions of the $X's$ and $Y's$, respectively, then the hypothesis to be tested is $F = G$ against the alternative $F \neq G$, for which the Mann–Whitney–Wilcoxon statistic can be used. Using this test, successive stationary periods can be delineated and the system can be studied in detail within such periods (see Moore (1975), who gives an algorithm for the procedure).

To analyze cyclic trends of the type discussed above, we may also use the periodogram method described by Lewis (1972) for the specific case of a nonhomogeneous Poisson process. Another test in the framework of the nonhomogeneous Poisson process is proposed by Joseph et al. (1990). They consider the output of an M/G/∞-queue where G is assumed to be known.

2.2.3 Tests for Independence

While formulating queueing models, for simplification and convenience, several assumptions of independence are made about its elements. Thus, most of the models assume that inter-arrival times and service times are independent sequences of independent and identically distributed random variables. If there are reasons to make such assumptions, statistical tests can be used for verification. Some of the tests that can be used to verify independence of a sequence of observations are tests for serial independence in point processes, described in Lewis (1972), and various tests for trend analysis and renewal processes, given by Cox and Lewis (1966). To verify the assumption of independence between inter-arrival and service times, nonparametric tests seem appropriate. Spearman's rho and Kendall's tau (Randles and Wolfe (1979), Hollander et al. (2013)) are used to test the correlation between two sequences of random variables, whereas Cramer-von Mises type statistics (see Koziol and Nemec (1979) and references cited therein) are used to test for bivariate independence directly from the definition of independence applied to random variables.

2.3 Distribution Selection

The next step in the model identification process is the determination of the best model for arrival and service processes. The distribution selection problem is based on the nature of data and availability of model distributions. For this problem, readers are referred to books on applied statistics and data analysis (e.g., Venables and Ripley (2002)). It is advisable to start with simple distributions such as the Poisson and exponential, since analysis under such assumptions is considerably simpler. After all, a mathematical model is essentially an approximation of a real process. The simpler the model is, the easier it is to analyze and to extract information from it. Thus, the selection of a distribution should be made with due consideration to the tradeoff between the advantages of the sophistication of the model and our ability to derive useful information from it.

Distributions such as Erlang and hyperexponential, are closely related to the exponential and with an appropriate selection of parameter values, they represent a wide variety of distributions. As noted in Appendix A, Erlang with coefficient of variation ≤ 1 and hyper-exponential with coefficient of variation ≥ 1, form a family of distributions with a broad range of distribution characteristics while retaining the convenience of analysis based on Markovian properties.

Once the distribution model is chosen, the next step is the determination of parameter values that bind the model to the real system. Normally either the maximum likelihood method or the method of moments is used for parameter estimation; the former is preferred because of its desirable statistical properties whereas the latter is used for its ease of implementation. A discussion of parameter estimation and hypothesis testing in queueing theory is given in Chapter 10.

2.4 Review Exercises

1. Determine the mean, variance, and coefficient of variation (CV) for the following distributions introduced in this chapter and Appendix A.

   (i) Deterministic, (2.1.1)

   (ii) Exponential, (2.1.2)

   (iii) Hyperexponential, (A.3.1)

   (iv) Erlang, (2.1.6), (A.4.1)

   (v) Mixed Erlang, (A.5.1), (A.5.2)

   (vi) Geometric, (A.8.1)

   (vii) Binomial, (A.8.3)

   (viii) Negative binomial, (A.8.4)

2. Determine the Laplace transform or the probability-generating function, as the case may be, for the distributions listed under Ex. 1.

3. Determine the probability-generating function for

   (i) The Poisson process, (A.2.1)

   (ii) The Compound Poisson process, (A.2.2)

4. Redo Exercise 1 using the Laplace transform or probability-generating function, as the case may be.

5. Determine for a specific value of t, the mean, variance, and coefficient of variation for

   (i) Poisson process

   (ii) Compound Poisson process

6. Establish the identity (2.1.7)

7. Establish the result (A.1.3)

8. Establish the result (A.2.4)

9. Determine the maximum likelihood estimates of the mean value parameters in distributions listed under Ex. #1.

http://www.springer.com/978-0-8176-8420-4

An Introduction to Queueing Theory
Modeling and Analysis in Applications
Bhat, U.N.
2015, XIV, 339 p. 18 illus., 6 illus. in color., Hardcover
ISBN: 978-0-8176-8420-4
A product of Birkhäuser Basel