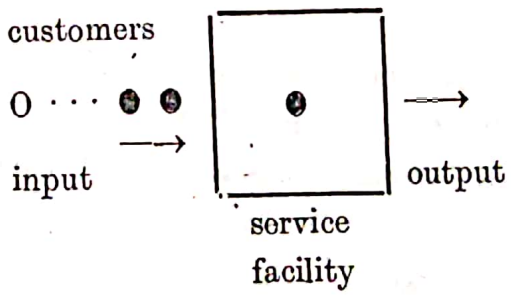
CHAPTER 1

POISSON QUEUES I

1.1 Introduction

In queueing theory we study situations where *units* of some kind *arrive* at a *service facility* for receiving *service* of some description, some of the units having to *wait* for service, and *depart* after service. A *queue* or *waiting line* develops whenever the service facility cannot cope with the number of units requiring service. The units calling for service are often called *customers* in a generic sense. A *system* is generally defined as something having an input, an output and a transformation process in between, which changes the input into the output. Thus, the arrangement described above qualifies to be regarded as a *queueing system*, where the customers requiring service form the input, the serviced customers the output and the service rendered is the transformation process.



The subject of queueing theory had its origin in the pioneering work done by Agner Krarup Erlang, an engineer at the Copenhagen Telephone Exchange around the beginning of this century on the application of probability theory to telephone traffic problems. It soon drew attention of many other probability theorists and has remained a popular field of research almost throughout the eighty years since then. There are many situations in real life where queues of the type described above develop. Thus, we can have a queue of broken-down machines waiting for repair at a repair shop, a queue

of customers at a departmental store cash counter, or such dangerous queues as the one formed by planes circling above an airport waiting to land. Often we have cases where a physical queue is absent, such as the waiting list of passengers for a railway or airline ticket or of persons who register their names, say, for the purchase of a car, which is not readily available and is to be supplied from future production.

1.2 Characterization

A queueing system is specified by stating the following details about it.

(a) *The input or the arrival pattern.* Let the successive arrivals to the system occur at times $t_1, t_2, \ldots$; then $u_r = t_{r+1} - t_r$ $(r = 1, 2, \ldots)$ are the inter-arrival times. We assume that the u_r are independent and identically distributed (i.i.d.) random variables (r.v.'s) and the input process is specified by stating their probability distribution function (P.D.F.), $A(u)$. $A(u)$ is called the inter-arrival time distribution or, simply, the arrival distribution or the input distribution. Such information as whether arrivals occur singly or in groups may be added here.

(b) *The queue discipline.* It is the statement of the rule of selection of customers for service out of those waiting, and may include rules regarding formation of queues.

(c) *The service mechanism.* It is stated by specifying the number of servers $c \geqslant 1$ and the probability distribution function $B(v)$ of the service times of the successive customers $v_1, v_2, \ldots$, which are assumed to be i.i.d. r.v.'s and also independent of the inter-arrival times u_r $(r = 1, 2, \ldots)$. We may also state whether the service is rendered to customers singly or in groups of fixed or varying size.*

1.3 Topics of Study

In any analysis of a queueing system, one or more of the following aspects of it are investigated.

(a) *The queue length.* This means the number of customers waiting before service. Sometimes the term is used to mean the total number of customers present in the system, including the one or more customers who are being served at the moment. Generally, the precise meaning is clear from the context. However, some authors prefer to avoid confusion by using the term queue length in the former sense, *i.e.* to mean the queue proper (before

*Some recent studies on queues with correlated arrivals or service times have been made.

service), and by calling it the number in the system or the system size when those being served are intended to be included. We shall be following this arrangement.

(b) *The waiting time.* This means the time spent by a customer in waiting before service. This term, too, is used in a wider sense by some authors to mean the time spent in waiting or being served, and they will call the time spent in waiting before service the queueing time. We shall prefer the former meaning, which seems to be more popular, and we shall call the waiting time in the wider sense the total time spent in the system, which is clearly the waiting time (before service) *plus* service time.

(c) *The busy period.* By a busy period is meant a period of time over which the server is continuously busy. Clearly, a busy period starts with an arrival at an empty service facility and ends when the number of customers in the system drops to zero for the first time. At this instant an *idle period* starts, which ends at the commencement of the next busy period.

The queue length, the waiting time and the busy period are studied through their probability distributions, from which moments like mean, variance, etc. can be obtained.

1.4 Notation

We shall use Kendall's (1953) notation, explained below, to denote the various queueing systems to be discussed. Occasionally, when necessary, we shall enhance the notation following the extension of Kendall's notation suggested by Lee (1966).

A queueing system will be denoted by a triad $\cdot\,|\,\cdot\,|\,\cdot$, in which the first two members will be letters which stand for the forms of the input and the service time distributions, respectively, and the third member is a number signifying the number of servers. The forms of the arrival and service distributions commonly used are described below.

(a) M (*Markovian or the negative exponential distribution*). This is the most commonly used distribution in queueing literature for reasons to be discussed later. When the letter M is used in the first place in the triad mentioned above, it will be taken to mean that the input distribution is given by

$$A(u) = 1 - e^{-\lambda u}, \quad u \geqslant 0 \tag{1}$$
$$= 0, \qquad \text{otherwise}$$

or, that the corresponding inter-arrival time density function is given by

$$a(u) = \lambda e^{-\lambda u},\ u \geqslant 0 \tag{2}$$
$$= 0, \qquad \text{otherwise}$$

The mean of the above distribution is $1/\lambda$, which is, therefore, the mean inter-arrival time. It follows that λ is the mean arrival rate, *i.e.* the average number of arrivals per unit time. From the well-known relation between the (negative) exponential and the Poisson distributions, it can be seen that the number of arrivals in a given time t will have the following Poisson distribution :

$$P\{n \text{ arrivals in time } t\} = (\lambda t)^n e^{-\lambda t}/n!,\ n = 0, 1, 2, \ldots \tag{3}$$

Because of the above, the case of exponential inter-arrival times is equivalently stated as Poisson arrivals or Poisson input.

In the same way, when M stands at the second place in the triad, it will mean that the service time distribution is

$$B(v) = 1-e^{-\mu v},\ v \geqslant 0 \tag{4}$$
$$= 0, \qquad \text{otherwise}$$

the corresponding density function being given as before, with λ, u in (2) replaced by μ, v. As before, we see that $1/\mu$ is the mean service time and μ the mean service rate, and

$$P\{n \text{ services in time } t\} = (\mu t)^n e^{-\mu t}/n!,\ n = 0, 1, 2, \ldots \tag{5}$$

Note : *Special features of the negative exponential distribution.* The exponential distribution possesses the following two properties, which make it very convenient for use in the analysis of queueing systems.

(i) *The Markovian or the "forgetfulness" property.* The conditional probability of an arrival during the time interval $(u, u+du)$, given that an arrival has not taken place in $[0, u_0]$, is equal to

$$P\{u < u_r < u+du \mid u_r > u_0\} = a(u)du/[1-A(u_0)] \tag{6}$$
$$= \lambda e^{-\lambda(u-u_0)}\, du$$

which shows that the entire inter-arrival time and the *residual* inter-arrival time after time u_0 have the same exponential distribution. Taking $u_0 = u$, we see that the *probability of an arrival during* $(u, u+du)$ *does not depend upon the time since the last arrival.* That is why Poisson arrivals are also called *random arrivals.*

Similarly, in the case of exponential service times the residual service time and the entire service time have the same distribution and the probability of a service completion is independent of the elapsed service time. Due to this fact, the exponential distribution is not suitable for service times. On the other hand, the exponential distribution is often a good fit for inter-arrival times, especially where arrivals are due to chance phenomena like machine failure, or come from an unstructured system like bus arrivals at a bus stop far from the starting point. In many instances 'customers' arriving at a business shop exhibit such random arrival pattern. There is a saying in Hindi which means that a customer and death have no fixed time.

(ii) In the case of negative-exponentially distributed inter-arrival times, since the arrivals in a given interval follow the Poisson distribution, we have

$$P\{\text{one arrival during } (t, t+\Delta t)\} = (\lambda\Delta t)e^{-\lambda\Delta t}/1! \tag{7}$$
$$= \lambda\Delta t\{1-\lambda\Delta t+\ldots\}$$
$$= \lambda\Delta t+o(\Delta t)$$

where $o(\Delta t)$ stands for terms which are negligible in comparison to Δt, in the sense that

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0 \tag{8}$$

Similarly,

$$P\{\text{no arrival during } (t, t+\Delta t)\} = (\lambda\Delta t)^0 e^{-\lambda\Delta t}/0! \tag{9}$$
$$= 1-\lambda\Delta t+o(\Delta t),$$

$$P\{\text{two or more arrivals during } (t, t+\Delta t)\} = o(\Delta t) \tag{10}$$

In the same way, if service times are also exponentially distributed, we have

$$P\{\text{one service completion during } (t, t+\Delta t)\} = \mu\Delta t+o(\Delta t) \tag{11}$$

$$P\{\text{no service completion during } (t, t+\Delta t)\} = 1-\mu\Delta t+o(\Delta t) \tag{12}$$

$$P\{\text{two or more service completions during } (t, t+\Delta t)\} = o(\Delta t) \tag{13}$$

$$P\{\text{no arrival and no service during } (t, t+\Delta t)\} = 1-(\lambda+\mu)\Delta t+o(\Delta t) \tag{14}$$

$$P\{\text{one arrival and one service during } (t, t+\Delta t)\} = o(\Delta t) \tag{15}$$

(b) *D (deterministic or constant inter-arrival/service times).* The letter D in the first (second) place in the triad in Kendall's notation indicates that the inter-arrival time (service time) is constant. In the case of constant

inter-arrival times, also called regular arrivals or regular input, when each inter-arrival time is a, say, we have the input distribution

$$A(u) = 0, \quad u < a$$
$$= 1, \quad u \geqslant a \quad (0 < a < \infty) \tag{16}$$

and the density function

$$a(u) = \delta(u-a), \tag{17}$$

where $\delta(\cdot)$ is the Dirac delta function. When service times are constant, we have similar expressions for $B(v)$ and $b(v)$.

(c) E_k (*k-Erlang distribution*). We say that the inter-arrival times have a k-Erlang distribution, or an Erlang distribution of order k, if each inter-arrival time is made up of k phases, the time spent in each phase having a negative exponential distribution with a common mean $1/(k\lambda)$, say, independently of other phases, so that the mean inter-arrival time is $1/\lambda$. Thus, the inter-arrival time density $a(u)$ is given by

$$\begin{aligned} a(u)du &= \mathrm{P}\{u < u_r < u+du\} \\ &= \mathrm{P}\{(k-1) \text{ phases are completed in time } u \text{ and} \\ &\qquad \text{one phase in } (u, u+du)\} \\ &= \frac{(k\lambda u)^{k-1}e^{-k\lambda u}}{(k-1)!} \cdot k\lambda du \\ &= \frac{(k\lambda)^k u^{k-1}e^{-k\lambda u}}{(k-1)!}\, du. \end{aligned} \tag{18}$$

This is clearly the k-fold convolution of the exponential distribution with itself and, therefore, gives the exponential distribution for $k = 1$. Also, it can be seen from (18), or by using the convolution theorem, that the Laplace transform of $a(u)$ is

$$\int_0^\infty e^{-su}a(u)du = \left(\frac{k\lambda}{k\lambda+s}\right)^k = \left(1+\frac{s}{k\lambda}\right)^{-k} \to e^{-s/\lambda} \tag{19}$$

as $k \to \infty$. Now, since the Laplace transform of the density function of a r.v. X is the expectation $E(e^{-sX})$, we see that $e^{-s/\lambda}$ corresponds to the case of constant inter-arrival times $= 1/\lambda$, the mean value. So, for $k \to \infty$, the Erlangian input becomes the regular input, given by (16) or (17).

The case of k-Erlang service time distribution can be dealt with similarly, the density function $b(v)$ being given by (18) with λ, u changed to μ, v, respectively.

(d) G (*general distribution*). The letter G is used for a general or arbitrary distribution. However, following the convention, we shall write GI (standing for "general independent") for general inter-arrival times and G for general service time distribution.

From the notation explained above, we see that

(i) M/M/1 stands for Poisson arrivals, exponential service times and one server,

(ii) M/D/c stands for Poisson arrivals, constant service times and c servers,

(iii) M/E_k/2 stands for Poisson arrivals, k-Erlang service times and two servers,

(iv) GI/G/c stands for the most general case of general input, general service times and c servers.

Lee (1966) has extended the above notation to $\cdot\,|\,\cdot\,|\,\cdot : (\cdot\,|\,\cdot)$, where the triad is as explained earlier, and within the parentheses the first member is a number, showing the maximum number of units that can be accommodated in the system (queue *plus* service) and the second member is an abbreviation for the queue discipline, like FIFO, LIFO, SIRO, etc., explained below. We shall now describe some of the queue disciplines encountered.

Queue discipline. The most commonly practised queue discipline is "first come, first served" or, as it is sometimes called, "first in, first out", abbreviated, respectively, as FCFS or FIFO. In this case, the waiting customers are selected for service in the order of their arrivals. Various other possibilities are there. Sometimes we have just the opposite rule of "last come, first served" or "last in, first out", in short LCFS or LIFO. In some cases we can have a random selection for service, or "service in random order", briefly, SIRO. And there are also a number of priority disciplines, according to which the customers are divided into two or more priority classes and a higher priority unit is chosen for service before a lower priority unit. Further, under a "preemptive priority" rule, a lower priority unit is preempted or taken out of service whenever a higher priority unit arrives during its service, the preempted lower priority unit being attended to only after serving all the units of higher priority classes. Under "non-preemptive priority" or the "head-of-the-line priority" rule, a service once started is allowed to continue up to completion, irrespective of arrivals of higher priority units, which are attended to only after completion of the service in

progress. There are also "dynamic priority" disciplines, where the priority increases due to waiting. For a detailed account of priority disciplines in queues, one may refer to Jaiswal (1968).

It may be noted that out of the three queue characteristics of the queue length, the busy period and the waiting time, only the waiting time is affected by the queue discipline.

Impatient customers. In addition to the queue discipline, some traits of behaviour of waiting customers have often to be taken into account. Thus, a customer is said to "balk" if, seeing the size of the queue, he decides not to join the queue and is thus lost to the system. If after waiting for some time, he gets impatient and leaves the system before his service, he is said to "renege".

Systems of queues. In a multi-server queueing system like M/M/c, the *servers* are operating *in parallel* and are fed by a single queue. Alternatively, we may have a separate queue for each server, *i.e.* a number of *queues in parallel*. In such a case, the customers exhibit the phenomenon of shifting between queues time and again, for example, to be always in the shortest queue. This is called *jockeying*. We have an arrangement of *queues in series* or *queues in tandem* when the output of one queue becomes the input to the next queue and so on until the last queue in the series. We have a case of *cyclic queues* if the output of the last queue again becomes input to the first queue. Sometimes we have *queueing networks* consisting of queues arranged both in series and in parallel.

1.5 The Queueing System M/M/1

(a) *The queue length*

Let us consider the single-server queue with Poisson arrivals and exponential service times. Let us write

(1) $$P_n(t) = P\{\text{there are } n \text{ units in the system at time } t\}$$

So, using eq. 1.4 (7)–(15), we have, for $n \geqslant 1$,

(2) $$\begin{aligned} P_n(t+\Delta t) &= P_n(t)P\{\text{no arrival and no service in } (t, t+\Delta t)\} \\ &\quad + P_{n-1}(t)P\{\text{one arrival in } (t, t+\Delta t)\} \\ &\quad + P_{n+1}(t)P\{\text{one service in } (t, t+\Delta t)\} + o(\Delta t) \\ &= P_n(t)\{1-(\lambda+\mu)\Delta t\} + P_{n-1}(t)\cdot\lambda\Delta t + P_{n+1}(t)\cdot\mu\Delta t + o(\Delta t) \end{aligned}$$

whence

(3) $$\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t} = -(\lambda+\mu)P_n(t)+\lambda P_{n-1}(t)+\mu P_{n+1}(t)+\frac{o(\Delta t)}{\Delta t}$$

Proceeding to limits as $\Delta t \to 0$, we have

(4) $$\frac{d}{dt}P_n(t) = -(\lambda+\mu)P_n(t)+\lambda P_{n-1}(t)+\mu P_{n+1}(t), \quad n = 1, 2, \ldots$$

For $n = 0$, we shall have

(5) $$\begin{aligned} P_0(t+\Delta t) &= P_0(t)P\{\text{no arrival in } (t, t+\Delta t)\} \\ &\quad + P_1(t)P\{\text{one service in } (t, t+\Delta t)\} + o(\Delta t) \\ &= P_0(t)\{1-\lambda\Delta t\} + P_1(t)\cdot\mu\Delta t + o(\Delta t) \end{aligned}$$

whence, as before,

(6) $$\frac{d}{dt}P_0(t) = -\lambda P_0(t)+\mu P_1(t).$$

Eq. (4) and (6) are differential-difference equations, which can be solved to evaluate $P_n(t)$, thus giving the time-dependent probability distribution of the number in the system. This solution is quite involved and will be discussed in the next chapter. However, for most of practical applications of the queueing model, a "steady state" solution suffices. We say, in general, that a system has reached a steady state or a state of "statistical equilibrium" when fluctuations with respect to time can be disregarded. Hence, to obtain a steady state solution we shall regard the $P_n(t)$, $n = 0, 1, 2, \ldots$ as independent of time. Consequently we have $dP_n(t)/dt = 0$, and we shall drop the argument t in $P_n(t)$ and write simply P_n. Thus, (4) and (6) yield

(7) $$0 = -(\lambda+\mu)P_n+\lambda P_{n-1}+\mu P_{n+1}, \quad n = 1, 2, \ldots$$

(8) $$0 = -\lambda P_0+\mu P_1.$$

Eq. (7) and (8) are easy to solve recursively, though there are other methods to solve them. (For details see Section 1.7.) It is customary to write

(9) $$(\lambda/\mu) = \rho$$

and call this ratio the "traffic intensity" of the system. From (8), we have

(10) $$P_1 = \rho P_0$$

Now, writing eq. (7) for $n = 1$, substituting for P_1, and solving the resulting equation for P_2, we have

(11) $$P_2 = \rho^2 P_0$$

Similarly, taking eq. (7) for $n = 2$ and substituting for P_2 and P_1, we have

$$P_3 = \rho^3 P_0 \tag{12}$$

and so on for P_4, etc.

It is easy to show by mathematical induction that

$$P_n = \rho^n P_0, \quad n = 1, 2, \ldots \tag{13}$$

Hence, the steady-state probability distribution P_n is known in terms of P_0. To evaluate P_0, we use the fact that the total probability over the entire distribution is one. Thus, if we have no upper limit on the number of customers that can join the queue, we have

$$\sum_{n=0}^{\infty} P_n = 1 \tag{14}$$

which is called the normalizing equation.

From (13) and (14), we have

$$P_0(1+\rho+\rho^2+\ldots) = 1 \tag{15}$$

Now, the infinite geometric series $1+\rho+\rho^2+\ldots$ is convergent only when $\rho < 1$. We conclude that the steady state exists only when $\rho < 1$. Assuming that $\rho < 1$ we have, from (15),

$$P_0 = 1-\rho \tag{16}$$

so that, from (13),

$$P_n = (1-\rho)\rho^n, \quad (\rho < 1, n = 0, 1, 2, \ldots) \tag{17}$$

which is a geometric probability distribution.

The mean number in the system, which we shall denote by L, is given by

$$\begin{aligned} L &= \sum_{n=0}^{\infty} nP_n \\ &= \sum_{n=0}^{\infty} \{n.(1-\rho)\rho^n\} \\ &= (1-\rho)(\rho+2\rho^2+3\rho^3+\ldots) \\ &= \rho(1-\rho)(1-\rho)^{-2} \end{aligned} \tag{18}$$

Thus,

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \tag{19}$$

The mean number in the queue (waiting for service), denoted by L_q, is similarly given by

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty} (n-1)P_n \\ &= \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \end{aligned} \tag{20}$$

(b) *The waiting time*

To study the waiting time, let us assume that the queue discipline is "first come, first served". Then, the mean waiting time in the queue and in the system, denoted by W_q and W, respectively can be found from simple arguments.

Let us first take W_q. The customers found in the queue at any time are those that arrived during the waiting time of the customer at the head of the line. Therefore, the mean number of customers in queue is the mean number of arrivals during a mean waiting time, that is,

$$L_q = \lambda W_q, \tag{21}$$

whence

$$W_q = L_q/\lambda = \lambda/\{\mu(\mu-\lambda)\}. \tag{22}$$

By similar reasoning, we have

$$L = \lambda W, \tag{23}$$

giving

$$W = L/\lambda = 1/(\mu-\lambda). \tag{24}$$

Now, to find the distribution of the time spent in system (w_s), let a typical customer arrive to find n customers in the system. Then, for w_s to lie in $(w, w+dw)$ all the $(n+1)$ customers including the new arrival himself should complete their service within time $w+dw$. Thus, the waiting time density $f(w)$ is given by

$$\begin{aligned} f(w)dw &= P\{w < w_s < w+dw\} \\ &= \sum_{n=0}^{\infty} P_n \cdot P\{n \text{ services in time } w\} \cdot P\{1 \text{ service in } dw\} \\ &= \sum_{n=0}^{\infty} (1-\rho)\rho^n \cdot \frac{(\mu w)^n e^{-\mu w}}{n!} \cdot \mu dw \end{aligned} \tag{25}$$

$$= (1-\rho)\mu e^{-\mu w} dw \sum_{n=0}^{\infty} (\rho\mu w)^n/n!$$

$$= (1-\rho)\mu e^{-\mu w}\, dw . e^{\rho\mu w}$$

$$= (1-\rho)\mu e^{-(1-\rho)\mu w}\, dw$$

Hence, the time spent in system has an exponential distribution with mean $= 1/\{(1-\rho)\mu\} = 1/(\mu-\lambda)$, which agrees with (24).

To find the distribution of the waiting time in the queue, w_q, we see that

$$P(w_q = 0) = P_0 = 1-\rho \tag{26}$$

and, arguing as before, the density function for the non-zero values of w_q, is given by

$$f_q(w)dw = P\{w < w_q < w+dw\} \tag{27}$$

$$= \sum_{n=1}^{\infty} P_n . P\{(n-1) \text{ services in } w\} . P\{1 \text{ service in } dw\}$$

$$= \sum_{n=1}^{\infty} (1-\rho)\rho^n . \frac{(\mu w)^{n-1} e^{-\mu w}}{(n-1)!} \cdot \mu dw$$

$$= (1-\rho)\rho\mu e^{-\mu w}\, dw . e^{\mu\rho w}$$

$$= \rho\mu(1-\rho) e^{-(1-\rho)\mu w}\, dw$$

From (26), (27) we see that the distribution of w_q consists of a discrete atom at zero and a continuous portion which is a truncated exponential distribution over $(0+, \infty)$. The mean waiting time in the queue is

$$W_q = 0.(1-\rho) + \int_{0+}^{\infty} w . \rho\mu(1-\rho) e^{-(1-\rho)\mu w}\, dw \tag{28}$$

$$= \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

(c) *The busy period*

The distribution of busy period, which involves time-dependent probabilities, will be discussed in the next chapter. However, the mean duration of a busy period can be obtained from simple arguments as follows.

The server's entire time is made up of alternating busy and idle periods; so the number of busy and idle periods in a long period of time T can be taken to be equal. Now, an idle period is the time from the instant the server

becomes idle until the next arrival which, according to the Markovian property of the exponential inter-arrival time distribution, is distributed as the entire inter-arrival time. Hence the mean duration of an idle period is $1/\lambda$. Now, the probability that the server is idle is $P_0 = 1-\rho$. Therefore, over the time T the server is expected to be idle for a duration $T(1-\rho)$, which is equivalent to $T\lambda(1-\rho)$ idle periods. Thus the server's busy time $T\rho$ is made up of $T\lambda(1-\rho)$ busy periods, and the mean duration of a busy period is

$$T\rho/[T\lambda(1-\rho)] = 1/(\mu-\lambda) \tag{29}$$

It follows that the mean number of customers served in a busy period is

$$\mu/(\mu-\lambda) = (1-\rho)^{-1} \tag{30}$$

(d) *Higher Moments*

In the foregoing discussion of the queue length, waiting time, etc., we have kept the analysis up to the derivation of the mean, which is the first moment, about the origin, of the distribution in question. However, it is often easy to evaluate higher moments like the second, third, fourth moments about the origin or other values, and so on, by using their definitions or by successive differentiation of the Laplace transforms or generating functions as the case may be. We give below derivations of the variances.

(i) *System size.* Let N be the r.v. denoting the number in system. Then, the variance of N is

$$\text{Var}[N] = E[N^2] - (E[N])^2 \tag{31}$$

where $E[X]$ stands for expected value of X, and $E[N] = L$, given by (18), (19) Also, by definition,

$$E[N^2] = \sum_{n=0}^{\infty} n^2 P_n \tag{32}$$

$$= (1-\rho) \sum_{n=0}^{\infty} n^2 \rho^n \quad \text{(on substituting for } P_n\text{)}$$

$$= (1-\rho) \left[\frac{\rho}{(1-\rho)^2} + \frac{2\rho^2}{(1-\rho)^3} \right]$$

$$= \frac{\rho}{1-\rho} + \frac{2\rho^2}{(1-\rho)^2}$$

Thus, (31) simplifies to

$$\text{Var}[N] = \frac{\rho}{1-\rho} + \frac{\rho^2}{(1-\rho)^2} \tag{33}$$

It is interesting to note that the factorial moment, of order 2, of the system size distribution is

(34) $$E[N(N-1)] = E[N^2]-E[N] = \frac{2\rho^2}{(1-\rho)^2}$$

(ii) *Queue length.* If N_q, N_{serv} denote the number in queue and number in service, respectively, we have

(35) $$N = N_q+N_{serv}$$

and from the description of the system it is clear that N_q and N_{serv} are independent. Hence,

(36) $$\mathrm{Var}[N] = \mathrm{Var}[N_q]+\mathrm{Var}[N_{serv}].$$

Now, the r.v. N_{serv} has the probability distribution

(37) $$\begin{cases} P(N_{serv} = 0) = P_0 = 1-\rho \\ P(N_{serv} = 1) = \rho \end{cases}$$

whence it is easy to show that

(38) $$\mathrm{Var}[N_{serv}] = \rho(1-\rho)$$

From (35) and (38), we have

(39) $$\mathrm{Var}[N_q] = \frac{\rho}{1-\rho}+\frac{\rho^2}{(1-\rho)^2}-\rho(1-\rho)$$

(iii) *Time in system.* Since w_s, the waiting time in system, has a negative exponential distribution with mean $1/(\mu-\lambda)$, (see (25)), its variance is given by

(40) $$\mathrm{Var}[w_s] = 1/(\mu-\lambda)^2$$

Incidentally, it is easy to see that the r-th moment of w_s *about the origin* is

(41) $$\mu_r' = \int_0^\infty w^r f(w)dw = r!/(\mu-\lambda)^r$$

$f(w)$ in (41) being given by (25).

(iv) *Waiting time in queue.* Proceeding as in (35), (36), we have for variance of w_q, the waiting time in queue,

(42) $$\mathrm{Var}[w_q] = \mathrm{Var}[w_s]-\mathrm{Var}[v], \text{ } v \text{ being the service time} = \frac{1}{(\mu-\lambda)^2}-\frac{1}{\mu^2}$$

Note that (42) can also be easily derived from (27).

1.6 State-Dependent Parameters

We now consider the same queueing model as in Section 1.5 except that the mean arrival and service rates λ and μ will now be taken to depend upon the state n of the system and, therefore, will be denoted by λ_n and μ_n respectively. By choosing different forms of the functions λ_n and μ_n, we shall obtain various interesting models of queues with Poisson arrivals and exponential service times.

Proceeding as before, eq. 1.5 (4), (6), now become

(1) $$\frac{d}{dt}P_n(t) = -(\lambda_n+\mu_n)P_n(t)+\lambda_{n-1}P_{n-1}(t)+\mu_{n+1}P_{n+1}(t), \quad (n = 1, 2, \ldots)$$

(2) $$\frac{d}{dt}P_0(t) = -\lambda_0P_0(t)+\mu_1P_1(t)$$

Hence, the steady-state equations are

(3) $$0 = -(\lambda_n+\mu_n)P_n+\lambda_{n-1}P_{n-1}+\mu_{n+1}P_{n+1}, \quad (n = 1, 2, \ldots)$$

(4) $$0 = -\lambda_0P_0+\mu_1P_1$$

From (4), we have

(5) $$P_1 = \frac{\lambda_0}{\mu_1}P_0$$

Writing eq. (3) for $n = 1$, substituting for P_1 from (5) and solving for P_2, we have

(6) $$P_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}P_0$$

and so on.

By mathematical induction, we can prove that

(7) $$P_n = \frac{\lambda_0\lambda_1\ldots\lambda_{n-1}}{\mu_1\mu_2\ldots\mu_n}P_0$$

In the general case when there is no upper limit on n, we have the normalizing equation

(8) $$P_0\left[1+\frac{\lambda_0}{\mu_1}+\frac{\lambda_0\lambda_1}{\mu_1\mu_2}+\ldots+\frac{\lambda_0\lambda_1\ldots\lambda_{n-1}}{\mu_1\mu_2\ldots\mu_n}+\ldots\right] = 1$$

For a steady-state solution to be possible, the infinite series inside the brackets in (8) should be convergent. Let it have a sum S; then

(9) $$P_0 = 1/S$$

From (7) and (9) we have the steady-state solution

$$P_n = \frac{\lambda_0\lambda_1\ldots\lambda_{n-1}}{\mu_1\mu_2\ldots\mu_n} S^{-1}, \quad n = 0, 1, 2, \ldots \tag{10}$$

where an empty product is taken as unity.

We proceed to consider a few particular cases of the model.

(a) *The queueing system* $M/M/1/N$

To obtain the solution for the queueing system with mean arrival and service rates λ and μ, respectively, where no more than N units are allowed to enter the system at any time, we set

$$\begin{aligned} \lambda_n &= \lambda \quad (0 \leqslant n < N) \\ &= 0 \quad (n \geqslant N) \\ \mu_n &= \mu \quad (n = 1, 2, \ldots) \end{aligned} \tag{11}$$

Then, from (8), we see that S is the finite sum

$$S = 1+\rho+\rho^2+\ldots+\rho^N = \frac{1-\rho^{N+1}}{1-\rho} \tag{12}$$

where $\rho = \lambda/\mu$, the traffic intensity. Since S is finite, we conclude that the steady state exists in all cases, *i.e.*, for all values of ρ. Substituting for S in (10), we have

$$P_n = (1-\rho)\rho^n/(1-\rho^{N+1}), \quad \rho \neq 1, \quad n = 0, 1, 2, \ldots, N \tag{13}$$

When $\rho = 1$, (12) gives $S = N+1$ and, therefore,

$$P_n = 1/(N+1), \quad \rho = 1, \quad n = 0, 1, \ldots, N \tag{14}$$

which can also be obtained from (13) by letting $\rho \to 1$.

The mean number in the system is

$$\begin{aligned} L &= \sum_{n=0}^{N} nP_n \\ &= \rho[1-(N+1)\rho^N+N\rho^{N+1}]/[(1-\rho)(1-\rho^{N+1})], \quad (\rho \neq 1) \\ &= N/2, \quad (\rho = 1) \end{aligned} \tag{15}$$

The mean number in the queue is

$$\begin{aligned} L_q &= \sum_{n=1}^{N} (n-1)P_n \\ &= \rho^2[1-N\rho^{N-1}+(N-1)\rho^N]/[(1-\rho)(1-\rho^{N+1})], \quad (\rho \neq 1) \\ &= N(N-1)/\{2(N+1)\}, \quad (\rho = 1) \end{aligned} \tag{16}$$

(b) *The queueing system* $M/M/c/\infty$

Let there be c severs (channels), the service times at each of which are exponentially distributed with mean service rate μ and let arrivals be Poisson with mean rate λ. To obtain the solution for this case, we set

$$\begin{aligned} \lambda_n &= \lambda \quad \text{for all } n \\ \mu_n &= n\mu \quad \text{for } n < c \\ &= c\mu \quad \text{for } n \geqslant c \end{aligned} \tag{17}$$

The justification for the above substitutions is that when n servers are busy, the probability of a departure during $(t, t+\Delta t)$ is the probability that any one of the n servers completes service which, by the addition theorem of probability, is $n\mu\Delta t$.

In this case, we have, from (8),

$$S = 1+\frac{\lambda}{\mu}+\frac{\lambda^2}{2!\,\mu^2}+\ldots+\frac{\lambda^{c-1}}{(c-1)!\,\mu^{c-1}}+\frac{\lambda^c}{c!\,\mu^c}\left\{1+\frac{\lambda}{c\mu}+\frac{\lambda^2}{(c\mu)^2}+\ldots\right\} \tag{18}$$

For a c-channel system it is customary to write the traffic intensity

$$\rho = \lambda/(c\mu) \tag{19}$$

Thus (18) gives

$$\begin{aligned} S &= 1+c\rho+\frac{(c\rho)^2}{2!}+\ldots+\frac{(c\rho)^{c-1}}{(c-1)!}+\frac{(c\rho)^c}{c!}\{1+\rho+\rho^2+\ldots\} \\ &= \sum_{j=0}^{c-1}(c\rho)^j/j!+\{(c\rho)^c/c!\}(1-\rho)^{-1}, \text{ provided } \rho < 1 \end{aligned} \tag{20}$$

The series in (20) is convergent only when $\rho < 1$, *i.e.*, when $\lambda/c\mu < 1$. We conclude that the steady state exists only when $\rho < 1$.

Using (17) in (7), we have

$$\begin{aligned} P_n &= P_0 c^n\rho^n/n! \text{ for } n < c \\ &= P_0 c^c\rho^n/c! \text{ for } n \geqslant c \end{aligned} \tag{21}$$

where

$$P_0 = S^{-1} = \left[\sum_{j=0}^{c-1}\{(c\rho)^j/j!\}+\{(c\rho)^c/(c!\,(1-\rho))\}\right]^{-1}. \tag{22}$$

It is easy to see that the probability of having to wait is

$$\begin{aligned} P(n \geqslant c) &= P_0(c^c/c!)\sum_{n=c}^{\infty}\rho^n \\ &= P_0(c\rho)^c/\{c!\,(1-\rho)\} = P_c/(1-\rho) \end{aligned} \tag{23}$$

The expected number waiting in the queue is

$$L_q = \sum_{n=c}^{\infty} (n-c)P_n$$

$$= P_0\{(c\rho)^c/c!\} \sum_{n=c}^{\infty} (n-c)\rho^{n-c} \quad (24)$$

$$= P_0\rho(c\rho)^c/\{c!(1-\rho)^2\}$$

$$= \frac{\lambda\mu(\lambda/\mu)^c P_0}{(c-1)!(c\mu-\lambda)^2}$$

The mean number of customers being served or the mean number of busy channels is

$$L_s = \sum_{n=0}^{c-1} nP_n + \sum_{n=c}^{\infty} cP_n \quad (25)$$

$$= \sum_{n=1}^{c} nP_n + \sum_{n=c+1}^{\infty} cP_n$$

$$= \sum_{n=1}^{c} n.P_0 \frac{(c\rho)^n}{n!} + \sum_{n=c+1}^{\infty} c.P_0 \frac{c^c\rho^n}{c!}$$

$$= c\rho\, P_0 \left[\sum_{n=1}^{c} \frac{(c\rho)^{n-1}}{(n-1)!} + \frac{c^c}{c!} \cdot \frac{\rho^c}{1-\rho} \right]$$

$$= c\rho\, P_0 \left[\sum_{j=0}^{c-1} \frac{(c\rho)^j}{j!} + \frac{(c\rho)^c}{c!(1-\rho)} \right]$$

$$= c\rho\, P_0 . P_0^{-1}$$

$$= c\rho = \lambda/\mu$$

Hence, the mean number in the system is

$$L = L_q + L_s = L_q + (\lambda/\mu) \quad (26)$$

The mean waiting time in queue is

$$W_q = L_q/\lambda = \mu(\lambda/\mu)^c P_0/\{(c-1)!(c\mu-\lambda)^2\} \quad (27)$$

The mean time spent in the system is

$$W = L/\lambda = W_q + (1/\mu) \quad (28)$$

(c) *The system M/M/∞*

This is the system where every arrival is able to find a server and is sometimes called the ample-server system. Possible applications are self-service systems where the customer carries the server with himself. In such systems

no queueing is involved and some authors like Conolly (1975) have pointed out that they should not be called queueing systems. Here we have

$$\lambda_n = \lambda \quad \text{for all } n \quad (29)$$

$$\mu_n = n\mu, \quad n = 0, 1, 2, \ldots$$

Therefore,

$$S = 1 + (\lambda/\mu) + \frac{(\lambda/\mu)^2}{2!} + \ldots = e^{\lambda/\mu} \quad (30)$$

Note that the series is convergent and, therefore, the steady state exists for all values of λ/μ. It is easily seen that we have

$$P_n = \frac{(\lambda/\mu)^n e^{-\lambda/\mu}}{n!} \quad (31)$$

so that the number in the system has a Poisson distribution with mean (λ/μ).

Thus

$$L = \lambda/\mu \quad (32)$$

$$W = L/\lambda = 1/\mu \quad (33)$$

which is otherwise obvious : since there is no queueing, the mean time in system equals the mean service time.

(d) *Queue with balking*

Let us consider the M/M/1 queueing system with mean arrival and service rates λ and μ, such that when there are n units in the system, an incoming customer joins the queue with probability $1/(1+n)$. So we take

$$\lambda_n = \lambda/(n+1), \quad n = 0, 1, 2, \ldots \quad (34)$$

$$\mu_n = \mu \quad \text{for all } n$$

Then,

$$S = 1 + (\lambda/\mu) + \frac{(\lambda/\mu)^2}{2!} + \ldots = e^{\lambda/\mu} \quad (35)$$

as before. Hence, in this case also, the number in the system has the Poisson distribution, given by (31). It follows that the mean number in the system is λ/μ, as before. But in the present case queueing is allowed, and we have the mean number in the queue

$$L_q = \sum_{n=1}^{\infty} (n-1)P_n = (\lambda/\mu) + e^{-\lambda/\mu} - 1 \quad (36)$$

(e) *The system $M/M/c/c$ (loss system)*

This is the system with c channels, where no queueing is allowed, so that the maximum number allowed in the system is c. An application of this model is a telephone exchange with c trunk lines with no facility for holding subscribers who fail to get a free line. Taking λ as the mean arrival rate and μ as the mean service rate for each channel, we have

$$\lambda_n = \lambda \quad \text{for } n < c \tag{37}$$
$$= 0 \quad \text{for } n \geqslant c$$
$$\mu_n = n\mu \quad \text{for } n \leqslant c$$

Therefore, S is the finite sum

$$S = 1+(\lambda/\mu)+\frac{(\lambda/\mu)^2}{2!}+\ldots+\frac{(\lambda/\mu)^c}{c!} \tag{38}$$

Thus, the steady state exists for all values of λ/μ, and we have

$$P_n = \frac{(\lambda/\mu)^n}{n!}S^{-1}, \quad (0 \leqslant n \leqslant c) \tag{39}$$

where S is given by (38). The probability that all the channels are busy and, therefore, the incoming calls are lost is given by

$$P_c = \frac{\frac{(\lambda/\mu)^c}{c!}}{1+(\lambda/\mu)+\frac{(\lambda/\mu)^2}{2!}+\ldots+\frac{(\lambda/\mu)^c}{c!}} \tag{40}$$

(40) is called *Erlang's loss formula* and is still in use in telephone traffic work.

(f) *The machine interference model*

Let the customers come from a finite source of k members, such that the probability that any member will call for service, independent of other members, in the interval $(t, t+\Delta t)$ is $\lambda\Delta t+o(\Delta t)$, and let there be a single server with mean service rate μ, the service times being exponentially distributed. Then, we have

$$\lambda_n = (k-n)\lambda \quad \text{for } n \leqslant k \tag{41}$$
$$= 0 \quad \text{for } n > k$$
$$\mu_n = \mu \quad \text{for } n = 1, 2, \ldots$$

Therefore,

$$S = 1+k(\lambda/\mu)+k(k-1)(\lambda/\mu)^2+\ldots+k!\,(\lambda/\mu)^k \tag{42}$$

and

$$P_0 = 1/S \tag{43}$$

$$P_n = k(k-1)\ldots(k-n+1)(\lambda/\mu)^n P_0, \ (n = 1, 2, \ldots, k) \tag{44}$$

Possible applications of this model are situations where there are k automatic machines under the charge of a single operator and the machines fail and require the operator's attention from time to time, or situations where k industrial workers require service at a facility like a grinding machine, tool crib, etc.

(g) *The brand-share model*

Consider a market consisting of N consumers, out of which at any time n consumers are using a particular brand, say brand A, of a product. Let the consumers switch into and out of brand A in a Poisson fashion with rates λ and μ, respectively, so that when the brand-share of A is n, we can write

$$\lambda_n = (N-n)\lambda \quad \text{for } n \leqslant N \tag{45}$$
$$= 0 \quad \text{for } n > N$$
$$\mu_n = n\mu \quad \text{for } n \leqslant N$$

Thus, we have

$$S = 1+\binom{N}{1}(\lambda/\mu)+\binom{N}{2}(\lambda/\mu)^2+\ldots+\binom{N}{N}(\lambda/\mu)^N \tag{46}$$
$$= \{1+(\lambda/\mu)\}^N$$

whence

$$P_0 = 1/S = \{\mu/(\lambda+\mu)\}^N \tag{47}$$

$$P_n = \binom{N}{n}(\lambda/\mu)^n P_0 \tag{48}$$

Clearly P_n is the probability that A has a brand share n, while

$$P_N = \{\lambda/(\lambda+\mu)\}^N \tag{49}$$

is the probability of market saturation by brand A. It may be noted that P_n follows the binomial probability distribution :

$$P_n = \binom{N}{n}q^{N-n}p^n \tag{50}$$

where

$$p = \lambda/(\lambda+\mu), \ q = 1-p \tag{51}$$

Hence, the expected brand share of A is

$$E(n) = Np = N\lambda/(\lambda+\mu) \tag{52}$$

Ex. 1. A television repairman finds that the time spent on his jobs has an exponential distribution with mean 30 minutes. If he repairs sets in the order in which they come in, and if the arrivals of sets are approximately Poisson with an average rate of 10 per 8-hour day, what is the repairman's expected idle time each day ? How many jobs are ahead of the average set just brought in ? [By permission from Sasieni, Yaspan and Friedman (1959)]

We have an M/M/1 queueing system with the mean arrival rate

$$\lambda = \frac{10}{8} = \frac{5}{4} \text{ sets per hour}$$

and the mean service rate

$$\mu = 2 \text{ sets per hour}$$

Therefore,

$$\rho = \frac{\lambda}{\mu} = \frac{5}{8}$$

The probability that the server is idle

$$= P_0 = 1-\rho = \frac{3}{8}$$

Hence the repairman's (server's) expected idle time per day

$$= \frac{3}{8} \times \text{duration of a day (}i.e.\text{, working hours)}$$

$$= \frac{3}{8} \times 8 \text{ hours} = 3 \text{ hours}$$

Also, expected number of units in the system

$$= L = \frac{\rho}{1-\rho} = \frac{5/8}{1-5/8} = \frac{5}{3} = 1\frac{2}{3} \text{ jobs (sets)}$$

Ex. 2. A bank has two tellers working on savings accounts. The first teller handles withdrawals only and the second teller handles deposits only. It has been observed that the service time distributions for both deposits and withdrawals are exponential with mean service time 3 minutes per customer. Depositors are found to arrive in a Poisson fashion throughout the day with mean arrival rate 16 per hour. Withdrawers also arrive in a Poisson fashion with mean arrival rate 14 per hour. What would be the

effect on the average waiting time for depositors and withdrawers if each teller could handle both withdrawals and deposits ? What would be the effect if this could only be accomplished by increasing the mean service time to 3·5 minutes ? [By permission from Sasieni, Yaspan and Friedman (1959)]

Originally we have for depositors, $\lambda = 16$ arrivals per hour, $\mu = 20$ services per hour, so that $\rho = \lambda/\mu = 4/5 < 1$ (steady state exists).

Therefore, $$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{16}{20(20-16)} = \frac{1}{5} \text{ hour} = 12 \text{ minutes}$$

And, for withdrawers, $\lambda = 14$ per hour, $\mu = 20$ per hour, $\rho = 7/10 < 1$

Therefore, $$W_q = \frac{14}{20\times 6} \text{ hour} = 7 \text{ minutes}$$

In the second case, we have an M/M/2 system with $\lambda = 16+14 = 30$ per hour, $\mu = 20$ per hour for each server, so that

$$\rho = \lambda/2\mu = \frac{3}{4} < 1 \text{ (so steady state exists)}$$

Therefore, from 1.6 (22), taking $c = 2$, we have

$$P_0 = \left[\sum_{j=0}^{1} \frac{(3/2)^j}{j!} + \left(\frac{3}{2}\right)^2 \Big/ \left\{2!\left(1-\frac{3}{4}\right)\right\}\right]^{-1}$$

$$= \left[1+\frac{3}{2}+\frac{9}{2}\right]^{-1} = \frac{1}{7}$$

Hence, from 1.6(27),

$$W_q = \frac{\mu(\lambda/\mu)^2}{(2\mu-\lambda)^2} P_0 = \frac{20\times\frac{9}{4}}{100} \times \frac{1}{7} \text{ hour} = 3\frac{6}{7} \text{ min.}$$

In the third case, when the mean service time increases to 3·5 min., we have $\mu = \frac{120}{7}$ services per hour, and $\lambda = 30$ per hour, as before. Therefore, $\rho = \lambda/2\mu = 7/8 < 1$. Thus, we have

$$P_0 = \left[1+\frac{7}{4}+\frac{(7/4)^2}{2\times\frac{1}{8}}\right]^{-1} = \left[1+\frac{7}{4}+\frac{49}{4}\right]^{-1} = 1/15$$

whence

$$W_q = \frac{\frac{120}{7}\,(7/4)^2}{\left(\frac{240}{7}-30\right)^2} \times \frac{1}{15} \text{ hour} = \frac{343}{30} \text{ min.} = 11\,\frac{13}{30} \text{ min.}$$

1.7 Alternative Solution Methods

We give below two alternative methods for the solution of the set of equations 1.5(7)–(8) for the steady-state queue length probabilities P_n in M/M/1.

(a) *Difference-equation method*

The equations can be written

$$\mu P_{n+2} - (\lambda+\mu)P_{n+1} + \lambda P_n = 0, \quad n = 0, 1, 2, \ldots \tag{1}$$

$$\mu P_1 - \lambda P_0 = 0 \tag{2}$$

Now, (1) is a difference equation with constant coefficients. Using the difference operator E defined by

$$E^r P_n = P_{n+r}, \quad r = 1, 2, \ldots$$

we can write (1) in the form

$$[\mu E^2 - (\lambda+\mu)E + \lambda]P_n = 0$$

so that the characteristic equation is

$$\mu E^2 - (\lambda+\mu)E + \lambda = 0$$

or,

$$(\mu E - \lambda)(E-1) = 0$$

which gives

$$E = 1, \rho, \text{ where } \rho = \lambda/\mu$$

Hence, the solution of (1) is

$$P_n = A + B\rho^n \tag{3}$$

where A, B are arbitrary constants.

Now, substituting from (3) in (2), we have

$$A + B\rho = \rho(A+B)$$

which gives $A = 0$ and hence $B = P_0$. Thus we have, from (3),

$$P_n = P_0\rho^n \tag{4}$$

which is the same as 1.5(13), and we can proceed further as before to show that $P_0 = 1-\rho\ (\rho < 1)$, which completes the solution.

(b) *Generating-function method*

Let us define the generating function

$$P(z) = \sum_{n=0}^{\infty} P_n z^n, \quad |z| \leqslant 1 \tag{5}$$

Multiplying eq. 1.5(7) by z^n, summing over $n = 1, 2, \ldots$, and adding 1.5(8), we have

$$0 = -(\lambda+\mu)\sum_{n=0}^{\infty} P_n z^n + \mu P_0 + \lambda \sum_{n=1}^{\infty} P_{n-1} z^n + \mu \sum_{n=0}^{\infty} P_{n+1} z^n \tag{6}$$

or,

$$0 = -(\lambda+\mu)P(z) + \mu P_0 + \lambda z P(z) + (\mu/z)\{P(z) - P_0\} \tag{7}$$

or,

$$\{\lambda z^2 - (\lambda+\mu)z + \mu\}P(z) + \mu(z-1)P_0 = 0 \tag{8}$$

This yields

$$P(z) = \frac{\mu(1-z)P_0}{\lambda z^2 - (\lambda+\mu)z + \mu} \tag{9}$$

The denominator of the right-hand member of (9) can be factorized as

$$\lambda z^2 - (\lambda+\mu)z + \mu = (\lambda z - \mu)(z-1)$$

so that (9) can be written

$$P(z) = \frac{\mu P_0}{\mu - \lambda z} = P_0(1-\rho z)^{-1} = P_0 \sum_{n=0}^{\infty} (\rho z)^n, \text{ since } \rho z < 1 \tag{10}$$

Recalling the definition (5) and equating the coefficients of z^n on the two sides, we have

$$P_n = P_0\rho^n \tag{11}$$

P_0 can be evaluated as in 1.5. Alternatively, we may note that $P(1) = 1$, and then (10) gives $P_0 = 1-\rho$, as before.

Ex. 3. Find the steady-state probability P_n for the finite-waiting-space multichannel queueing system M/M/c/N, where the number of channels is c and the maximum number allowed in the system is N ($N > c$). Also deduce expressions for L, L_q, W, W_q and the expected number of busy channels.

In the notation of Section 1.6, we have

$$\lambda_n = \lambda \quad \text{when } 0 \leqslant n < N$$
$$= 0 \quad \text{when } n \geqslant N$$
$$\mu_n = n\mu \quad \text{when } n < c$$
$$= c\mu \quad \text{when } n \geqslant c$$

Therefore, writing $\lambda/c\mu = \rho$, we have as in 1.6(21)

$$P_n = P_0 c^n \rho^n / n!, \quad (0 \leqslant n < c)$$
$$= P_0 c^c \rho^n / c!, \quad (c \leqslant n \leqslant N)$$

where (as in 1.6(18))

$$S = P_0^{-1} = \sum_{j=0}^{c-1} \frac{(c\rho)^j}{j!} + \frac{(c\rho)^c}{c!} \sum_{i=0}^{N-c} \rho^i$$

$$= \sum_{j=0}^{c-1} \frac{(c\rho)^j}{j!} + \frac{(c\rho)^c}{c!} \cdot \frac{1-\rho^{N-c+1}}{1-\rho} \quad \text{when } \rho \neq 1$$

$$= \sum_{j=0}^{c-1} \frac{c^j}{j!} + \frac{c^c}{c!}(N-c+1) \quad \text{when } \rho = 1$$

Note that the steady state exists for all values of ρ.

Now, the mean number in the queue (waiting) is

$$L_q = \sum_{n=c+1}^{N} (n-c)P_n = \{P_0(c\rho)^c/c!\} \sum_{n=c+1}^{N} (n-c)\rho^{n-c}$$

$$= \{P_0(c\rho)^c/c!\}\{\rho + 2\rho^2 + 3\rho^3 + \ldots + (N-c)\rho^{N-c}\}$$

$$= \frac{P_0(c\rho)^c\rho}{c!(1-\rho)^2}[1-\rho^{N-c}-(1-\rho)(N-c)\rho^{N-c}]$$

The average number being served or the mean number of busy

$$L_s = \sum_{n=0}^{c-1} nP_n + \sum_{n=c}^{N} cP_n$$

$$= \sum_{n=1}^{c} nP_0 \cdot \frac{(c\rho)^n}{n!} + \sum_{n=c+1}^{N} cP_0 \frac{c^c\rho^n}{c!}$$

$$= c\rho P_0 \left[\sum_{n=1}^{c} \{(c\rho)^{n-1}/(n-1)!\} + \frac{c^c}{c!} \sum_{n=c+1}^{N} \rho^{n-1} \right]$$

$$= c\rho P_0 \left[\sum_{j=0}^{c-1} (c\rho)^j/j! + \frac{(c\rho)^c}{c!} \sum_{n=0}^{N-c-1} \rho^n \right]$$

$$= c\rho P_0 \left[P_0^{-1} - \frac{(c\rho)^c}{c!} \cdot \rho^{N-c} \right] = c\rho(1-P_N) = (\lambda/\mu)(1-P_N)$$

Therefore, the mean number in the system is

$$L = L_q + L_s = L_q + (\lambda/\mu)(1-P_N)$$

The effective arrival rate at the system is $\lambda' = \lambda(1-P_N)$. Therefore, from Little's formula,

$$W = L/\lambda' = W_q + (1/\mu)$$
$$W_q = L_q/\lambda'$$

Ex. 4. At a port there are six unloading berths and four unloading crews. When all the berths are full, arriving ships are diverted to an overflow facility 20 miles down the river. Tankers arrive according to a Poisson process with a mean of one every 2 hours. It takes an unloading crew, on the average, ten hours to unload a tanker, the unloading time following an exponential distribution.

(a) On the average, how many tankers are at the port ?

(b) On the average, how long does a tanker spend at the port ?

(c) What is the average arrival rate at the overflow facility ?

Here we have an M/M/c/N system with $c = 4$ and $N = 6$. Also, $\lambda = \frac{1}{2}$ arrival per hour, and $\mu = \frac{1}{10}$ service per hour. Thus, $c\rho = \lambda/\mu = 5$, and $\rho = 5/4$. Therefore, as in Ex. 3,

$$1/P_0 = \sum_{j=0}^{3} \frac{5^j}{j!} + \frac{5^4}{4!} \cdot \frac{1-(5/4)^3}{1-(5/4)} = 1+5+\frac{25}{2}+\frac{125}{6}+\frac{625}{24}\times\frac{61}{16}$$

$$\approx 138{\cdot}62$$

Therefore,

$$P_0 = 1/138{\cdot}62 \approx 0{\cdot}0072$$

$$P_N = P_0 c^c \rho^N / c! = 0{\cdot}0072 \times 256 \times \left(\frac{5}{4}\right)^6 / 24 = 0{\cdot}293$$

Hence,

(a) The mean number of tankers at the port is

$$L = L_q + (\lambda/\mu)(1-P_N)$$

Now,

$$L_q = \frac{P_0(c\rho)^c\rho}{c!(1-\rho)^2}[1-\rho^{N-c}-(1-\rho)(N-c)\rho^{N-c}]$$

$$= \frac{0.0072\times 5^4\times(5/4)}{24\times(1/16)}\left[1-\frac{25}{16}+\frac{1}{4}\times 2\times\frac{25}{16}\right] = 0{\cdot}820$$

Thus, $L = 0{\cdot}820 + 5(1-0{\cdot}293) = 4{\cdot}36$ tankers approx.

(b) The average time spent at the port is

$$W = L/\{\lambda(1-P_N)\} = 4{\cdot}36/(\tfrac{1}{2}\times 0{\cdot}707) = 12{\cdot}33 \text{ hours}$$

Thus the average time spent in waiting before unloading is

$$W_q = W - 10 \text{ hours} = 2{\cdot}33 \text{ hours}$$

(c) The average arrival rate at the overflow facility is

$$\lambda P_N = \tfrac{1}{2}\times 0{\cdot}293 \approx 0{\cdot}15 \text{ tanker per hour}$$

PROBLEM SET 1

1. Arrivals at a telephone booth are considered to be Poisson with an average time of 10 minutes between one arrival and the next. The length of a phone call is assumed to be distributed exponentially with mean 3 minutes.

(a) What is the probability that a person arriving at the booth will have to wait ?

(b) The telephone department will install a second booth when convinced that an arrival would expect to wait for at least 3 minutes for the phone. By how much should the flow of arrivals increase in order to justify a second booth ? [By permission from Sasieni, Yaspan and Friedman (1959)]

[(a) 0·3; (b) *new arrival rate one every six minutes*]

2. A supermarket has two girls ringing up sales at the counters. If the service time for each customer is exponential with mean 4 minutes, and if people arrive in a Poisson fashion at the rate of 10 an hour,

(a) What is the probability of having to wait for service ?

(b) What is the expected percentage of idle time for each girl ?

(c) Find the average queue length and the average number of units in the system. [By permission from Sasieni, Yaspan and Friedman (1959)]

[(a) 1/6; (b) $66\frac{2}{3}\%$; (c) 1/12; 2/3]

3. In a railway marshalling yard, goods trains arrive at the rate of 30 trains per day. Assuming that the inter-arrival time follows an exponential distribution and the service-time distribution is also exponential with an average of 36 minutes, calculate the following :

(i) the mean queue size (line length).

(ii) the probability that the queue size exceeds 10.

[$\rho/(1-\rho) = 3$ *trains*; $\rho^{10} = 0{\cdot}056$]

4. At what average rate must a clerk at a supermarket work in order to ensure a probability of 0·90 that the customer will not have to wait longer than 12 minutes (a) before service, (b) including service ? Assume that there is only one counter, at which customers arrive in a Poisson fashion at an average rate of 15 per hour. The length of service by the clerk has an exponential distribution. [By permission from Sasieni, Yaspan and Friedman (1959)]

[(a) 2·48 *minutes per service*; (b) 2·26 *minutes per service*]

5. Problems arrive at a computer centre in a Poisson fashion at an average rate of five per day. The rules of the computer centre are that any man waiting to get his problem solved must aid the man whose problem is being solved. If the time to solve a problem with one man has an exponential distribution with mean time of 1/3 day, and if the average solving time is inversely proportional to the number of people working on the problem, approximate the expected time in the centre for a person entering the line. [By permission from Sasieni, Yaspan and Friedman (1959)]

[8 *hours*]

6. A telephone company is planning to install telephone booths in a new airport. It has established the policy that a person should not have to wait more than 10 per cent of the times he tries to use a phone. The demand for use is estimated to be Poisson with an average of 30 per hour. The average phone call has an exponential distribution with a mean time of 5 minutes. How many phone booths should be installed ? [6]

7. Show that for the k-Erlang distribution given by 1·4(18), the mean, variance and mode are, respectively, $1/\lambda$, $1/k\lambda^2$, and $(1-1/k)(1/\lambda)$.

8. For an M/M/2 queueing system in steady state, show that

$$\lambda P_0 = \mu P_1$$
$$(\lambda+\mu)P_1 = \lambda P_0 + 2\mu P_2$$
$$(\lambda+2\mu)P_n = \lambda P_{n-1} + 2\mu P_{n+1} \quad (n = 2, 3, \ldots)$$

Hence prove that $P_n = 2\rho^n P_0$ $(n = 1, 2, \ldots)$, where $\rho = \lambda/2\mu$, and $P_0 = (1-\rho)/(1+\rho)$, and deduce that $L = 2\rho/(1-\rho^2)$, $L_q = 2\rho^3/(1-\rho^2)$, and that the probability of having to wait is $2\rho^2/(1+\rho)$.

9. Find the steady-state probability P_n, given that

$$\lambda_n = \begin{cases} (N-n)\lambda, & 0 \leqslant n < N \\ 0, & n \geqslant N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 1 \leqslant n < c \\ c\mu, & n \geqslant c \end{cases}$$

Describe a practical situation represented by the above model.

$$\left[P_n = \binom{N}{n} (\lambda/\mu)^n P_0, \ 0 \leqslant n < c \right.$$

$$\left. = \binom{N}{n} (n!/c!\, c^{n-c})(\lambda/\mu)^n P_0, \ c \leqslant n \leqslant N \right]$$

10. In Problem 3, assume that the yard can admit 9 trains at a time (there being 10 lines, one of which is reserved for shunting purposes). Calculate the probability that the yard is (i) empty, (ii) completely occupied, and find the mean number of trains in the yard.

[0·26; 0·02; 2·4 *trains*]

11. For an M/M/1 system, show by simple arguments that $W_q = L/\mu$, and $W = (L+1)/\mu$. Hence use the expression for L to deduce those for W_q and W.

12. At a railway station, only one train is handled at a time. The railway yard is sufficient only for two trains to wait while another is given signal to leave the station. Trains arrive at the station at an average of 6 per hour, and the railway station can handle them on an average of 12 per hour. Assuming Poisson arrivals and exponential service distribution, find the steady-state probabilities for the various number of trains in the system. Also find the average waiting time of a new train coming into the yard.

[$P_0 = 0{\cdot}533$, $W_q = 3\frac{2}{3}$ *minutes*]

CHAPTER 2

POISSON QUEUES II

2.1 Introduction

In Chapter 1, we obtain the steady-state solutions of various queueing systems with Poisson arrivals and exponential service times like M/M/1, M/M/c, etc. In this chapter we shall discuss the time-dependent or transient solutions of similar queueing systems.

2.2 The Queueing System M/M/1—Transient Solution

As we saw in Section 1.5, the probabilities $P_n(t)$ that there are n units in the system at time t are given by (Cf. 1.5(4), (6))

$$\frac{d}{dt} P_n(t) = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t), \quad n = 1, 2, \ldots \tag{1}$$

$$\frac{d}{dt} P_0(t) = -\lambda P_0(t) + \mu P_1(t) \tag{2}$$

Define the p.g.f. (probability generating function)

$$G(z, t) = \sum_{n=0}^{\infty} P_n(t) z^n \tag{3}$$

which clearly converges inside and on the unit circle $|z| = 1$. Multiplying (1) by z^n, summing over n and adding (2), we have, on simplification,

$$\frac{d}{dt} G(z, t) = -(\lambda+\mu-\lambda z-\mu/z)G(z, t) + \mu(1-1/z)P_0(t) \tag{4}$$

Let the system start at time $t = 0$ with i units in the system, so that

$$P_n(0) = \delta_{in} \tag{5}$$

$$= \begin{cases} 1, & \text{when } n = i \\ 0, & \text{when } n \neq i \end{cases}$$

δ_{in} is called the Kronecker delta. Thus, we have

$$G(z, 0) = z^i \tag{6}$$

Let $g(z, s)$ be the Laplace transform (L.T.) of $G(z, t)$, defined by

$$g(z, s) = \int_0^{\infty} e^{-st} G(z, t)\, dt, \ \mathrm{Re}(s) > 0 \tag{7}$$

and let the L.T.'s of other functions be similarly denoted by the corresponding lower case letters. From (4), on taking L.T.'s, we have

$$(s+\lambda+\mu-\lambda z-\mu/z)g(z,s) = z^{i}+\mu(1-1/z)p_0(s) \tag{8}$$

or,

$$g(z,s) = \frac{z^{i+1}-\mu(1-z)p_0(s)}{(s+\lambda+\mu)z-\lambda z^2-\mu} \tag{9}$$

Since $g(z, s)$ converges inside and on the unit circle for $\mathrm{Re}(s) > 0$, the zeros of the denominator inside and on $|z| = 1$ must coincide with the corresponding zeros of the numerator. The zeros of the denominator are

$$\alpha_1(s) = (2\lambda)^{-1}[s+\lambda+\mu-\sqrt{(s+\lambda+\mu)^2-4\lambda\mu}] \tag{10}$$

$$\alpha_2(s) = (2\lambda)^{-1}[s+\lambda+\mu+\sqrt{(s+\lambda+\mu)^2-4\lambda\mu}] \tag{11}$$

Now, let

$$f(z) = (s+\lambda+\mu)z \qquad g(z) = -(\lambda z^2+\mu) \tag{12}$$

Then, on the unit circle $|z| = 1$,

$$|f(z)| = |s+\lambda+\mu| > |\lambda+\mu| = g(z) \tag{13}$$

Hence, by Rouché's theorem, $f(z)$ and $f(z)+g(z)$ have the same number of zeros inside the unit circle. But $f(z)$ has only one zero, and so $f(z)+g(z)$ also has only one zero inside the unit circle. This zero must be α_1. Thus we have

$$p_0(s) = \frac{\alpha_1^{i+1}}{\mu(1-\alpha_1)} \tag{14}$$

Therefore,

$$g(z,s) = \frac{z^{i+1}-(1-z)\cdot\dfrac{\alpha_1^{i+1}}{1-\alpha_1}}{-\lambda(z-\alpha_1)(z-\alpha_2)} \tag{15}$$

or,

$$g(z,s) = \frac{z^{i+1}(1-\alpha_1)-(1-z)\alpha_1^{i+1}}{\lambda\alpha_2(z-\alpha_1)(1-z/\alpha_2)(1-\alpha_1)}$$

$$= \frac{(z^{i+1}-\alpha_1^{i+1})-z\alpha_1(z^{i}-\alpha_1^{i})}{\lambda\alpha_2(z-\alpha_1)(1-z/\alpha_2)(1-\alpha_1)}$$

$$= \frac{(z^{i}+\alpha_1 z^{i-1}+\ldots+\alpha_1^{i})-z\alpha_1(z^{i-1}+\alpha_1 z^{i-2}+\ldots+\alpha_1^{i-1})}{\lambda\alpha_2(1-z/\alpha_2)(1-\alpha_1)}$$

$$= \frac{(z^{i}+\alpha_1 z^{i-1}+\ldots+\alpha_1^{i})(1-\alpha_1)+\alpha_1^{i+1}}{\lambda\alpha_2(1-z/\alpha_2)(1-\alpha_1)}$$

or, using binomial expansion,

$$g(z,s) = \frac{1}{\lambda\alpha_2}(z^{i}+\alpha_1 z^{i-1}+\ldots+\alpha_1^{i})\sum_{k=0}^{\infty}(z/\alpha_2)^k + \frac{\alpha_1^{i+1}}{\lambda\alpha_2(1-\alpha_1)}\sum_{k=0}^{\infty}(z/\alpha_2)^k \tag{16}$$

Note that $|z/\alpha_2| < 1$.

No, $p_n(s)$, the L.T. of $P_n(t)$ is the coefficient of z^n in $g(z, s)$. The contribution to the coefficient of z^n from the second term on the right-hand side of (16) is

$$\frac{\alpha_1^{i+1}}{\lambda\alpha_2^{n+1}(1-\alpha_1)} = \frac{\alpha_1^{i+1}}{\lambda\alpha_2^{n+1}}(1+\alpha_1+\alpha_1^2+\ldots) \tag{17}$$

$$= \frac{1}{\lambda}\cdot\frac{\alpha_1^{n+i+2}}{(\alpha_1\alpha_2)^{n+1}}\sum_{k=0}^{\infty}\alpha_1^k$$

$$= \frac{1}{\lambda}\left(\frac{\lambda}{\mu}\right)^{n+1}\sum_{k=n+i+2}^{\infty}\alpha_1^k$$

$$= \frac{1}{\lambda}\left(\frac{\lambda}{\mu}\right)^{n+1}\sum_{k=n+i+2}^{\infty}\left(\frac{\mu}{\lambda}\right)^k\frac{1}{\alpha_2^k}, \quad \text{since } \alpha_1\alpha_2 = \mu/\lambda$$

And, the contribution from the first term is

$$\sum_{m=(i-n)^+}^{i}\frac{1}{\lambda\alpha_2}\cdot\frac{\alpha_1^m}{\alpha_2^{n-i+m}} = \sum_{m=(i-n)^+}^{i}\frac{(\mu/\lambda)^m}{\lambda\alpha_2^{n-i+2m+1}} \tag{18}$$

where

$$(x)^+ = \max(x, 0) \tag{19}$$

Therefore,

$$p_n(s) = \frac{1}{\lambda}\left[\sum_{m=(i-n)^+}^{i}\frac{(\mu/\lambda)^m}{\alpha_2^{n-i+2m+1}} + \left(\frac{\lambda}{\mu}\right)^{n+1}\sum_{k=n+i+2}^{\infty}\frac{(\mu/\lambda)^k}{\alpha_2^k}\right] \tag{20}$$

Now, from Appendix E, eq. (18), we have, for the inverse Laplace transform,

$$\mathcal{L}^{-1}[(s+\sqrt{s^2-a^2})^{-\nu}] = \nu a^{-\nu}t^{-1}I_\nu(at) \tag{21}$$

where $I_\nu(z)$ is the modified Bessel function of the first kind and order ν, given by

$$I_\nu(z) = \sum_{k=0}^{\infty}\frac{(z/2)^{\nu+2k}}{k!\,\Gamma(\nu+k+1)} \tag{22}$$

Therefore,

$$\mathcal{L}^{-1}\frac{1}{\alpha_2^k} = e^{-(\lambda+\mu)t}\cdot(2\lambda)^k \cdot k(2\sqrt{\lambda\mu})^{-k}t^{-1}I_k(2\sqrt{\lambda\mu}\,t) \tag{23}$$

$$= e^{-(\lambda+\mu)t}\cdot(\sqrt{\lambda/\mu})^k t^{-1}\cdot kI_k(2\sqrt{\lambda\mu}\,t)$$

Hence, from (20), taking the inverse Laplace transforms, we have

$$P_n(t) = \frac{e^{-(\lambda-\mu)t}}{\lambda}\left[\left(\sqrt{\frac{\lambda}{\mu}}\right)^{n-i+1}\sum_{m=(i-n)^+}^{i}\frac{n-i+2m+1}{t}\cdot I_{n-i+2m+1}(2\sqrt{\lambda\mu}\,t)\right. \tag{24}$$

$$\left.+\left(\frac{\lambda}{\mu}\right)^{n+1}\sum_{k=n+i+2}^{\infty}\left(\sqrt{\frac{\mu}{\lambda}}\right)^k\cdot\frac{k}{t}I_k(2\sqrt{\lambda\mu}\,t)\right]$$

On using the well-known recurrence relation

$$(2\nu/z)I_\nu(z) = I_{\nu-1}(z) - I_{\nu+1}(z) \tag{25}$$

and after some obvious simplification, we have

$$P_n(t) = e^{-(\lambda+\mu)t}\left[(\sqrt{\mu/\lambda})^{i-n}I_{|n-i|}+(\sqrt{\mu/\lambda})^{i-n+1}I_{n+i+1}\right. \tag{26}$$

$$\left.+(1-\lambda/\mu)(\lambda/\mu)^n\sum_{k=n+i+2}^{\infty}(\sqrt{\mu/\lambda})^k I_k\right]$$

where $I_\nu \equiv I_\nu(2\sqrt{\lambda\mu}\,t)$. Again, since for integral ν,

$$I_{-\nu}(z) = I_\nu(z) \tag{27}$$

as can be seen from (22), the $I_{|n-i|}$ in (26) can be replaced by I_{n-i}.

2.3 The Busy Period Distribution

For finding the distribution of busy period, we assume that at time $t = 0$ the system starts with the arrival of a customer which makes the number in the system $n = 1$ and that there is an absorbing barrier at the state $n = 0$. Then, clearly,

$$P_0(t) = P(\text{busy period} \leqslant t) = F(t), \text{ say} \tag{1}$$

is the probability distribution function (P.D.F.) of the busy period duration. Under these assumptions, the $P_n(t)$ are given by

$$\frac{d}{dt}P_n(t) = -(\lambda+\mu)P_n(t)+\lambda P_{n-1}(t)+\mu P_{n+1}(t),\quad n = 2, 3, \ldots \tag{2}$$

$$\frac{d}{dt}P_1(t) = -(\lambda+\mu)P_1(t)+\mu P_2(t) \tag{3}$$

$$\frac{d}{dt}P_0(t) = \mu P_1(t) \tag{4}$$

Using the generating function

$$G(z, t) = \sum_{n=1}^{\infty}P_n(t)z^n \tag{5}$$

we have, from (2) and (3),

$$\frac{d}{dt}G(z, t)+(\lambda+\mu-\lambda z-\mu/z)G(z, t) = -\mu P_1(t) \tag{6}$$

Now, taking L.T.'s on both sides of (6), and denoting the transforms by corresponding small letters, we have

$$(s+\lambda+\mu-\lambda z-\mu/z)g(z, s) = z-\mu p_1(s) \tag{7}$$

or,

$$g(z, s) = \frac{z-\mu p_1(s)}{s+\lambda+\mu-\lambda z-\mu/z} \tag{8}$$

The two zeros of the denominator are given by 2.2(10), (11). As in Section 2.2, by Rouché's theorem the numerator of the right-hand member of (8) vanishes at $z = \alpha_1$. Thus,

$$\mu p_1(s) = \alpha_1 \tag{9}$$

From (1), (4) and (9), the busy period density function is

$$f(t) = \frac{d}{dt}F(t) = \mu P_1(t) = \mathcal{L}^{-1}\alpha_1 \tag{10}$$

$$= \mathcal{L}^{-1}\left[\frac{(s+\lambda+\mu)-\sqrt{(s+\lambda+\mu)^2-4\lambda\mu}}{2\lambda}\right]$$

$$= \mathcal{L}^{-1}\left[\frac{2\mu}{(s+\lambda+\mu)+\sqrt{(s+\lambda+\mu)^2-4\lambda\mu}}\right]$$

$$= 2\mu\cdot\frac{1}{2\sqrt{\lambda\mu}}\cdot\frac{1}{t}\cdot I_1(2\sqrt{\lambda\mu}\,t)\cdot e^{-(\lambda+\mu)t}$$

$$= \sqrt{\mu/\lambda}\,e^{-(\lambda+\mu)t}\cdot\frac{1}{t}I_1(2\sqrt{\lambda\mu}\,t)$$

where $I_1(\cdot)$ is the modified Bessel function of the first kind and order 1, and we have used 2.2 (21) for evaluating the inverse Laplace transform.

The Initial Busy Period

The initial busy period may start with any initial number $n = i$ in the system instead of $n = 1$ as we assumed above. So, proceeding as above and taking $G(z, 0) = z^i$, we see that the busy period density is

(11) $$f_i(t) = \mathcal{L}^{-1}(\alpha_1^i)$$

$$= e^{-(\lambda+\mu)t} \cdot i(\sqrt{\mu/\lambda})^i \cdot \frac{1}{t} I_i(2\sqrt{\lambda\mu}\, t).$$

Alternatively, (11) may be proved as follows.

We know that the busy period duration does not depend on the queue discipline. The initial busy period under consideration starts with i customers in the system; let these customers be numebered 1, 2, ..., i. Let us first serve the customer '1' followed by service of all customers that arrive during his service time, then followed by all customers who arrive during their service times, and so on until all such arrivals have been serviced. Then, we treat similarly customers 2, 3, ..., i. Thus it is clear that the distribution of a busy period starting with i customers is the i-fold convolution of that of an 'ordinary' busy period starting with one customer. Let the corresponding probability density functions (p.d.f.) of these distributions be, respectively, $f_i(t)$ and $f(t)$. Now, we have shown in (10) that

(12) $$\mathcal{L}f(t) = \alpha_1$$

Therefore, by the convolution theorem of Laplace transforms,

(13) $$\mathcal{L}f_i(t) = \alpha_1^i$$

from which (11) follows.

2.4 The System M/M/∞

The steady-state solution of the service system M/M/∞ was discussed in Section 1.6. The probabilities, $P_n(t)$, for n units in the system at time t for this system are given by

(1) $$\frac{d}{dt} P_n(t) = -(\lambda+n\mu)P_n(t)+\lambda P_{n-1}(t)+(n+1)\mu P_{n+1}(t), \quad (n = 1, 2, \ldots)$$

(2) $$\frac{d}{dt} P_0(t) = -\lambda P_0(t)+\mu P_1(t)$$

Define the generating function

(3) $$G(z, t) = \sum_{n=0}^{\infty} P_n(t) z^n, \quad |z| \leqslant 1$$

Multiplying (1) by z^n, summing over n and adding (2) as usual, we have

$$\frac{\partial G}{\partial t} = -\sum_{n=0}^{\infty} (\lambda+n\mu)P_n(t)z^n+\lambda \sum_{n=1}^{\infty} P_{n-1}(t)z^n+\mu \sum_{n=0}^{\infty} (n+1)\, P_{n+1}(t)z^n$$

where $G \equiv G(z, t)$

or,

$$\frac{\partial G}{\partial t} = -\lambda G(z, t)-\mu z \frac{\partial G}{\partial z} +\lambda z G(z, t)+\mu \frac{\partial G}{\partial z}$$

or,

(4) $$\frac{\partial G}{\partial t} -\mu(1-z) \frac{\partial G}{\partial z} = -\lambda(1-z)G(z, t)$$

To solve this partial differential equation of Lagrange type, we write the corresponding simultaneous equations

(5) $$\frac{dt}{1} = \frac{dz}{-\mu(1-z)} = \frac{dG}{-\lambda(1-z)G}$$

The solution of the equation obtained from the first two members of the above equation is given by

(6) $$(1-z)e^{-\mu t} = c_1$$

and the solution of the equation obtained from the last two members is

(7) $$Ge^{-(\lambda/\mu)z} = c_2$$

where c_1, c_2 are arbitrary constants. Hence, the general solution of (4) is

(8) $$G(z, t) = e^{(\lambda/\mu)z}\phi\{(1-z)e^{-\mu t}\}$$

where ϕ is an arbitrary function.

Now, if the initial number in the system is i, we have the boundary condition

(9) $$G(z, 0) = z^i$$

From (8) and (9), we have

$$z^i = e^{(\lambda/\mu)}\, \phi(1-z)$$

whence

(10) $$\phi(z) = (1-z)^i \exp[-(\lambda/\mu)(1-z)]$$

Using (10) in (8), we have

(11) $$G(z, t) = \exp[(\lambda/\mu)z]\{1-(1-z)e^{-\mu t}\}^i \cdot \exp[-(\lambda/\mu)\{1-(1-z)e^{-\mu t}\}]$$
$$= (q+pz)^i \exp[-(\lambda q/\mu)(1-z)]$$

where

$$p = e^{-\mu t},\ q = 1-p \tag{12}$$

Taking the coefficient of z^n in (11), we have

$$P_n(t) = \sum_{k=0}^{n} \binom{i}{n-k} \frac{(\lambda/\mu)^k}{k!} \cdot e^{-\lambda q/\mu} p^{n-k} q^{i-n+2k} \tag{13}$$

The mean number in the system at time t is given by

$$L = \left.\frac{\partial G}{\partial z}\right|_{z=1} = ip + \frac{\lambda q}{\mu} = \frac{\lambda}{\mu} + \left(i - \frac{\lambda}{\mu}\right) e^{-\mu t} \tag{14}$$

And, the variance V of the number in the system at time t is given by

$$V = \left.\frac{\partial^2 G}{\partial z^2}\right|_{z=1} + L - L^2 \tag{15}$$

which simplifies to

$$\begin{aligned} V &= ipq + \frac{\lambda q}{\mu} \\ &= (\lambda/\mu) + (i - \lambda/\mu)e^{-\mu t} - ie^{-2\mu t} \end{aligned} \tag{16}$$

It may be noted from (11) that $G(z, t)$ is the product of the generating functions of the binomial distribution

$$P(X = k) = \binom{i}{k} q^{i-k} p^k,\ k = 0, 1, \ldots, i \tag{17}$$

and the Poisson distribution

$$P(X = k) = e^{-\lambda q/\mu} (\lambda q/\mu)^k / k!,\ k = 0, 1, 2, \ldots \tag{18}$$

Thus, the distribution of the number in the system at time t is the convolution of these two distributions. This gives us an alternative way of writing (13) by taking the convolution of the two distributions given by (17) and (18). Also, the mean and variance of the number in the system at time t can then be obtained by adding the means and variances of the above binomial and Poisson distributions, which directly gives the results given in (14) and (16).

For the steady-state distribution of the number in system, we let $t \to \infty$ in (11) and obtain

$$G(z) = \lim_{t \to \infty} G(z, t) = \exp\left[-(\lambda/\mu)(1-z)\right] \tag{19}$$

which is the generating function of the Poisson distribution with mean λ/μ, giving the steady-state probabilities

$$P_n = e^{-\lambda/\mu} (\lambda/\mu)^n / n!\ (n = 0, 1, 2, \ldots) \tag{20}$$

which are as derived in 1.6(31).

PROBLEM SET 2

1. From 2.2(26) deduce the steady-state probability P_n by making $t \to \infty$.

[From the asymptotic behaviour of $I_k(z)$, we have

$$I_k(z) \sim \exp(z)/\sqrt{2\pi z} \tag{1}$$

which is independent of k. Hence, for large t,

$$\begin{aligned} e^{-(\lambda+\mu)t} I_k(2\sqrt{\lambda\mu}\, t) &\sim \exp[-(\sqrt{\lambda} - \sqrt{\mu})^2 t]/\sqrt{t} \\ &\to 0 \text{ as } t \to \infty \end{aligned} \tag{2}$$

Thus the first two terms on the right-hand side of 2.2(26) vanish. Also, it is well-known that

$$\exp[(x/2)(y + 1/y)] = \sum_{k=-\infty}^{\infty} y^k I_k(x) = \sum_{k=0}^{m-1} y^k I_k(x) + \sum_{k=m}^{\infty} y^k I_k(x) + \sum_{k=1}^{\infty} y^{-k} I_k(x) \tag{3}$$

since $I_{-k}(x) = I_k(x)$ for integral k.

Taking $x = 2\sqrt{\lambda\mu}\, t$, $y = \sqrt{\mu/\lambda}$, and using (1) and the fact that $\lambda/\mu < 1$, we see that the first and the third sums on the right-hand side of (3) vanish as $t \to \infty$. Thus (3) reduces to

$$\sum_{k=m}^{\infty} (\sqrt{\mu/\lambda})^k I_k(2\sqrt{\lambda\mu}\, t) = \exp[(\sqrt{\lambda\mu}\, t)(\sqrt{\mu/\lambda} + \sqrt{\lambda/\mu})] = \exp[(\lambda+\mu)t]$$

Taking $m = n + i + 2$, we see from 2.2(26) that

$$P_n(t) \to (1 - \lambda/\mu)(\lambda/\mu)^n \text{ as } t \to \infty.]$$

2. Obtain the distribution of an ordinary busy period in M/M/1 system and use it to show that the mean duration of a busy period is $1/(\mu - \lambda)$.

[Mean busy period $= \int_0^\infty t f(t)\, dt$, where $f(t)$ is the busy period density function given by 2.3(10). Use Erdélyi *et al.* (1954a, p. 195, eq. (1)), *viz.*

$$\mathcal{L}[I_\nu(at)] = a^\nu (s^2 - a^2)^{-1/2} (s + \sqrt{s^2 - a^2})^{-\nu},$$

Alternatively, mean busy period $= -\bar{f}'(0)$, where $\bar{f}(s) = \mathcal{L}f(t) = \alpha_1$ from 2.3(10).]

3. Show that the variance of the (ordinary) busy period is $(\lambda+\mu)/(\mu-\lambda)^3$.

[If B stands for the busy period, Var $[B] = E[B^2]-(E[B])^2$, where $E[B] = 1/(\mu-\lambda)$, and $E[B^2] = \int_0^\infty t^2 f(t)dt$, with $f(t)$ given by 2.3(10). Now use Erdélyi *et al.* (1954a), p. 195, Eq. (2), *viz.*

$$\mathcal{L}[t\ I_\nu(at)] = a^\nu(s+\nu\sqrt{s^2-a^2})(s^2-a^2)^{-3/2}\,(s+\sqrt{s^2-a^2})^{-\nu}.$$

Alternatively, $E[B^2] = \tilde{f}''(0)$, with $\tilde{f}(s)$ given by 2.3(10), which simplifies to $E[B^2] = 2\mu/(\mu-\lambda)^3$.]

4. In an M/M/c queueing system, show that the duration of a busy period defined as the time during which all the c channels are continuously busy has the density function

$$e^{-(\lambda+c\mu)t}t^{-1}\sqrt{c\mu/\lambda}\,I_1(2\sqrt{c\lambda\mu}\,t).$$

Hence show that the mean duration of such a period is $1/(c\mu-\lambda)$.

5. Show that Eq. 2.4(17), (18) give, respectively, the distributions of the customers who are still in the system at time t out of the initial number i in the system at time zero and out of those who arrive in $(0, t]$.

CHAPTER 3

NON-POISSON QUEUES

3.1 Introduction

In this chapter we shall consider queueing systems in which either the distribution of inter-arrival times or that of service times is a general or arbitrary distribution. As was pointed out in Section 1.4, the negative-exponential distribution possesses the Markovian or the forgetfulness property, so that in the case of an M/M/1 queue we do not have to take into account the time since the last arrival or the elapsed service time of the unit in service. This is not the case with queues where one or both of the arrival and service time distributions are different from exponential. It is explained in the next section.

3.2 Imbedded Markov Chain

To be specific, let us first take the case of an M/G/1 queueing system. Here the inter-arrival time distribution is exponential, which has the Markovian property. However, the service time distribution being general, the probability of a service completion in time $(t, t+\Delta t)$ *depends* upon the time for which the service has already been given. To overcome this complication, we study the system not in continuous time but at the discrete time points at which customers depart after service. Between any two consecutive departures, we have to consider only arrivals which are Markovian. Suppose we are considering the queue length process or the number in the system. The queue length at the departure epochs is said to constitute a Markov chain imbedded in the non-Markovian queue length process in continuous time. The departure instants are called *regeneration points* and are such that a knowledge of the state of the process at any regeneration point renders previous history of the process irrelevant.

In the case of the GI/M/1 queue with general arrivals and exponential service times, the arrival epochs are the regeneration points. Clearly for an M/M/1 system, *all* time points are regeneration points and the whole process in continuous time is Markovian.