Chapter 2
Preliminary Concepts

Abstract This chapter describes the mathematical and probability concepts that are the foundation for the remaining chapters of the book. This includes the Poisson, Exponential and Erlang distributions, the Postulates that define a queuing system, and also the difference, differential, equilbrium and reduced equations. The equilibrium and/or the reduced equations are used to generate the probability distribution on the number of units in the system, and the various performance measures of the system.

2.1 Introduction

This chapter gives an overview on some of the key mathematical and probability concepts that are used in queuing theory. The chapter introduces the concepts that are used in the subsequent text so that they do not need to be repeated throughout the book. This includes a definition of the Poisson, Exponential and Erlang distributions and how they are related to each other. The chapter also lists the Postulates that are needed to define a queuing system. The postulates are used to identify a particular queuing system by way of difference equations. The difference equations yield the differential and equilibrium equations and finally the reduced equations. The equilibrium and/or the reduced equations are needed to generate the probability distribution of n units in the system, and then the various performance measures.

2.2 Some Useful Relations

Some of the identities that are used in developing the queuing models are listed here. Equations 2.1–2.3 are identities of infinite sums that apply when $0 < \theta < 1$. Equations 2.4–2.8 are identities that concern finite sums, and Eq. 2.9 is an identity that pertains to the exponent term.

N. T. Thomopoulos, *Fundamentals of Queuing Systems*,
DOI: 10.1007/978-1-4614-3713-0_2, © Springer Science+Business Media New York 2012

$$\sum\nolimits_{k \geq 0} \theta^k = 1/(1-\theta) \tag{2.1}$$

$$\sum\nolimits_{k \geq 0} k\theta^k = \theta/(1-\theta)^2 \tag{2.2}$$

$$\sum\nolimits_{k \geq 0} k^2\theta^k = \theta(1+\theta)/(1-\theta)^3 \tag{2.3}$$

$$\sum\nolimits_{k=1}^{N} 1 = \mathrm{N} \tag{2.4}$$

$$\sum\nolimits_{k=1}^{N} k = \mathrm{N}(\mathrm{N}+1)/2 \tag{2.5}$$

$$\sum\nolimits_{k=1}^{N} k^2 = \mathrm{N}(\mathrm{N}+1)(2\mathrm{N}+1)/6 \tag{2.6}$$

$$\sum\nolimits_{k=0}^{N} x^k = \left(1-\mathrm{x}^{\mathrm{N}+1}\right)/(1-\mathrm{x}) \quad \mathrm{x} \neq 1 \tag{2.7}$$

$$\sum\nolimits_{k=0}^{N} kx^k = \mathrm{x}\left[1-(\mathrm{N}+1)\mathrm{x}^{\mathrm{N}} + \mathrm{N}\mathrm{x}^{\mathrm{N}+1}\right]/(1-\mathrm{x})^2 \quad \mathrm{x} \neq 1 \tag{2.8}$$

$$\mathrm{e}^{\mathrm{ax}} = \sum\nolimits_{k \geq 0} (ax)^k/\mathrm{k}! \tag{2.9}$$

2.3 Exponential Distribution

Consider a random variable t that is continuous with t ≥ 0 and follows the exponential distribution. The probability density of t is $\mathrm{f(t)} = \theta\mathrm{e}^{-\theta\mathrm{t},}$ and the corresponding cumulative distribution is $\mathrm{F(t)} = 1 - \mathrm{e}^{-\theta\mathrm{t}}$. For the exponential variable t, the expected value and variance are $\mathrm{E(t)} = 1/\theta$, and $\mathrm{V(t)} = 1/\theta^2$, respectively.

2.4 Poisson Distribution

The Poisson probability distribution has a discrete variable n where n = 0, 1, 2,.... The probability distribution of n is $\mathrm{P_n} = \theta^{\mathrm{n}}\mathrm{e}^{-\theta}/\mathrm{n}!$. The expected value and variance of n are $\mathrm{E(n)} = \theta$, and $\mathrm{V(n)} = \theta$, respectively.

The Poisson distribution can also be defined in units of time t. In this situation, the discrete variable n represents the number of occurrences in time t. The probability of n units in time t becomes,

$\mathrm{P(n,t)} = (\theta\mathrm{t})^n\mathrm{e}^{-(\theta\mathrm{t})}/\mathrm{n}!$.

2.5 Relation Between the Exponential and Poisson Distributions

The Poisson distribution and the exponential distribution are related as shown here. Recall the exponential probability density is

$$f(t) = \theta e^{-\theta t}$$

Suppose τ is exponential with expected value $1/\theta$, and n is Poisson with mean θ. From the exponential,

$$\begin{aligned} P(\tau > t) &= 1 - F(t) \\ &= e^{-\theta t} \\ &= P(n = 0 \text{ in } t) \\ &= P(0, t) \end{aligned}$$

where the latter is Poisson. Also, note below, where the probability of n units in time t, P(n,t), becomes Poisson.

$$P(0, t) = e^{-\theta t}$$

$$P(1, t) = \int_{\tau=0}^{t} P(0, \tau)\ f(t - \tau) d\tau = \theta t e^{-\theta t}$$

$$P(2, t) = \int_{\tau=0}^{t} P(1, \tau)\ f(t - \tau) d\tau\ =\ (\theta t)^2 e^{-\theta t}/2!$$

$$\ldots$$

$$P(n, t) = \int_{\tau=0}^{t} P(n - 1, \tau)\ f(t - \tau) d\tau\ =\ (\theta t)^n e^{-\theta t}/n!$$

In the following discussion, θ is replaced with λ for arrival times. So when the arrivals to a system are exponential with an average time of $1/\lambda$ the number of units that arrive to the system in a unit of time is Poisson distributed with an average of λ. In the same way, if the number of units that arrive to a system is Poisson with parameter λ, the time between arrivals is exponential with an average of $1/\lambda$.

In the following discussion, θ is replaced with μ for service times. Hence, when the time to process the units is exponential and the average service time is $1/\mu$, the number of units that are serviced, during a continuously busy span of time, is Poisson distributed with an average of μ in a unit of time. If the units coming out of a continuously busy service facility is Poisson with a parameter of μ, the time to service the units are exponential with an average of $1/\mu$.

2.6 Convolution of Two Poisson Variables

Consider two Poisson random variables, x_1 and x_2, with parameters θ_1 and θ_2, respectively. Now assume another variable $x = x_1 + x_2$ is formed. Note the convolution below.

$$\begin{aligned}
P(x) &= \sum_{x1=0}^{x} P(x_1)P(x - x_1) \\
&= \sum_{x1=0}^{x} \left[e^{-\theta 1}\theta_1^{x1}/x_1!\right]\left[e^{-\theta 2}\theta_2^{x-x1}/(x - x_1)!\right] \\
&= e^{-(\theta 1+\theta 2)}\theta_2^{x} \sum_{x1=0}^{x} (\theta_1/\theta_2)^{x1}/[x_1!(x - x_1)!] \\
&= e^{-(\theta 1+\theta 2)}(\theta_1 + \theta_2)^{x}/x!
\end{aligned}$$

Thus, x is also Poisson with parameter $(\theta_1 + \theta_2)$.

2.7 Erlang Distribution

In some queuing systems, the time associated with arrivals and service times is assumed as an Erlang continuous random variable. The Erlang variable has two parameters, θ and k. The parameter k represents the number of exponential variables that are summed together to form the Erlang variable. Note, if y is an exponential variable with $E(y) = 1/\theta$, and x is the sum of k y's, then

$x = (y_1 + \ldots + y_k)$,

and the expected value of x becomes,

$E(x) = kE(y) = k/\theta$.

Further, the variance of x, denoted as V(x), is derived from adding k variances of y, V(y), as below:

$V(x) = kV(y) = k/\theta^2$

Note, when $k = 1$, the Erlang variable is the same as an exponential variable where the mode is zero and the density is skewed to the right. As k increases, the mode moves further away from zero and becomes less skewed to the right. As k increases, the shape of the Erlang density starts to resemble a normal density, via the central limit theorem.

2.8 Memory-Less Property of the Exponential Distribution

Recall, when a random variable t is exponential, the probability density is

$f(t) = \theta e^{-\theta t}$

and the cumulative distribution is

$$F(t) = 1 - e^{-\theta t}$$

For a time increment h, the probability that t is larger than h becomes

$$P(t > h) = e^{-\theta h}$$

At $t = (t' + h)$, the probability t is larger than $(t' + h)$ is

$$P(t > t' + h) = e^{-\theta(t' + h)}$$

The conditional probability of $t > (t' + h)$ given $t > t'$ is

$$P(t > t' + h|t > t') = e^{-\theta(t' + h)}/e^{-\theta t'} = e^{-\theta h}$$

Note the probabilities $P(t > t' + h|t > t')$ and $P(t > h)$ are the same, i.e.,

$$P(t > t' + h|t > t') = P(t > h) = e^{-\theta h}$$

Because the two probabilities are the same, the exponential distribution is called a memory-less probability distribution.

2.9 Cumulative Distribution for a Small Increment h

Consider time t that follows the exponential distribution, and observe, for a particular time increment h, the cumulative distribution of h becomes $F(h) = 1 - e^{-\theta h}$. Note the expression for F(h) can be converted using Eq. (2.9) above in the following way.

$$\begin{aligned} F(h) &= P(t<h) = 1 - e^{-\theta h} \\ &= 1 - [(-\theta h)^0/0! + (-\theta h)^1/1! + (-\theta h)^2/2! + \ldots] \\ &= 1 - [1 + (-\theta h)^1/1! + (-\theta h)^2/2! + \ldots] \\ &= \theta h - [(-\theta h)^2/2! + (-\theta h)^3/3! + \ldots] \\ &= \theta h + o(h) \end{aligned}$$

where

$$o(h) = -[(-\theta h)^2/2! + (-\theta h)^3/3! + \ldots]$$

Note o(h) is a function that approaches zero faster than h. That is

$$\underset{h\to 0}{Lim}\{o(h)/h\} = 0$$

Thereby, as h approaches zero, o(h) also approaches zero. This expression concerning the probability distribution of h is applied subsequently to define the postulates in the queuing analysis.

2.10 Probability Postulates

Assume a queuing system where the arrivals follow an exponential distribution and the average time between arrivals is $1/\lambda$. As shown above, the probability that the time between two arrivals is h or less becomes $[\lambda h + o(h)]$. Also, we assume the service time follows an exponential distribution with an average service time of $1/\lambda$. Hence, the probability is $[\mu h + o(h)]$ that the service time is less than h. Also consider the two events: A = event of an arrival in time interval h, and D = event of a departure in time interval h.

Now note the probabilities listed below that concern the events of A and D during the time interval from t to t + h, and denoted here as (t, t + h). Recall, h approaches zero.

$P[A \text{ in } (t,t + h)] = [\lambda h + o(h)]$
$P[D \text{ in } (t,t + h)] = [\mu h + o(h)]$
$P[\text{neither } A \text{ or } D \text{ in } (t,t + h)] = [1 - \lambda h - o(h)][1 - \mu h - o(h)] = [1 - \lambda h - \mu h + o(h)]$
$P[2 \text{ or more } A \text{ and/or } D \text{ in } (t,t + h)] = o(h)$

These four probabilities are the postulates that define most of the queuing systems that follow.

2.11 Difference Equations

Consider a queuing system with one service facility, infinite queue length, with exponential arrival times with an average of $1/\lambda$ and exponential service times with an average of $1/\mu$. The difference equations specify how the system operates. This is the first step to define a queuing system. The difference equations specify how the probability of n units in the system may change as time goes from t to (t + h), denoted as (t,t + h), and where h is a very small increment of time. The number of units n in the system at any time period are integers of $n \geq 0$. As described earlier, o(h) is a function that approaches zero faster than h. The difference equations are below:

$n = 0 \qquad P_0(t + h) = (1 - \lambda h)P_0(t) + \mu h P_1(t) + o(h)$
$n \geq 1 \qquad P_n(t + h) = (1 - \lambda h - \mu h)P_n(t) + \lambda h P_{n-1}(t) + \mu h P_{n+1}(t) + o(h)$

2.12 Differential Equations

Differential equations are obtained from the difference equations when the time increment h approaches zero. They are needed in an interim manner to subsequently yield the equilibrium equations. To convert, the three identities listed below are applied. The first shows how the derivative is formed. The second expresses the probability without the increment of h, and the third concerns the function o(h).

$$\underset{h\to 0}{Lim}\{[\mathrm{P_n(t+h)} - \mathrm{P_n(t)}]/\mathrm{h}\} = \mathrm{P_n(t)}'$$
$$\underset{h\to 0}{Lim}\{[\lambda\mathrm{h} + \mu\mathrm{h})\mathrm{P_n(t)}]/\mathrm{h}\} = [(\lambda + \mu)\mathrm{P_n(t)}]$$
$$\underset{h\to 0}{Lim}\{\mathrm{o(h)/h}\} = 0$$

Thus, as h approaches zero in the difference equations, the following set of the differential equations evolve:

$$\begin{array}{ll} n = 0 & P_0(t)' = (-\lambda)P_0(t) + \mu P_1(t) \\ n \geq 1 & P_n(t)' = (-\lambda-\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) \end{array}$$

2.13 Equilibrium Equations

The equilibrium equations are obtained by studying what happens to the differential equations when time t approaches infinity under equilibrium conditions. The two identities below are used in this process.

$$\underset{t\to\infty}{Lim}\{\mathrm{P_n(t)}'\} = 0$$
$$\underset{t\to\infty}{Lim}\{\mathrm{P_n(t)}\} = \mathrm{P_n}$$

Applying the above identities to the differential equations yields the following equilibrium equations:

$$\begin{array}{ll} n = 0 & 0 = -\lambda P_0 + \mu P_1 \\ n \geq 1 & 0 = -(\lambda + \mu)P_n + \lambda P_{n-1} + \mu P_{n+1} \end{array}$$

2.14 Reduced Equations

Algebra is needed at this point to transform the equilibrium equations to reduced equations as is shown here. Note below where the equilibrium equations for n = 0, 1, 2, say, are listed on the left-hand-side, and the corresponding reduced equations are on the right-hand-side. When n = 0, the equilibrium equation and the associated reduced equation are the same. For $n \geq 1$, the reduced equation for n is derived from the corresponding (n) equilibrium equation and the (n − 1) reduced equation.

$$
\begin{array}{lll}
n = 0 & 0 = -\lambda P_0 + \mu P_1 & \Rightarrow \quad 0 = -\lambda P_0 + \mu P_1 \\
n = 1 & 0 = -(\lambda + \mu)P_1 + \lambda P_0 + \mu P_2 & \Rightarrow \quad 0 = -\lambda P_1 + \mu P_2 \\
n = 2 & 0 = -(\lambda + \mu)P_2 + \lambda P_1 + \mu P_3 & \Rightarrow \quad 0 = -\lambda P_2 + \mu P_3
\end{array}
$$

The general form for the reduced equations becomes the following:

$$0 = -\lambda P_{n-1} + \mu P_n \qquad n \geq 1$$

2.15 Probability of n Units in the System (P_n)

The common notation in queuing is to use n as the number of units in the system at an arbitrary moment in time. In this way, n is discrete where n is zero or larger. One measure of interest in studying queuing systems is the probability of n units in the system, and this is denoted as P_n for $n \geq 0$.

2.16 Performance Measures

Some of the other measures of interest in queuing systems are listed below:

P_o = probability the system is empty
Ls = expected number of units in the service facility
Lq = expected number of units in the queue
L = expected number of units in the system
Ws = expected time in the service facility
Wq = expected time in the queue
W = expected time in the system
Wq′ = expected time in the queue given the arrival is delayed
SL = service level = probability the arrival is not delayed in the queue
Ploss = probability an arrival is lost (does not enter the system)

2.17 Wait Time in Queue Given a Delay (Wq′)

Using conditional expectation notation, $Wq' = W_{q|D}$ where D is the event that the arrival is delayed waiting in the queue before being serviced. Using the same notation, D′ = event the arrival is not delayed. In general,

$W_{q|D}$ = wait time in queue given delay
$W_{q|D'}$ = wait time in queue given no delay
and
P(D) = probability of a delay
P(D′) = probability of no delay
The relation between the waiting time (Wq) and the conditional waiting times ($W_{q|D'}$,$W_{q|D}$) is below:
$Wq = W_{q|D'}P(D') + W_{q|D}P(D)$
Since, $W_{q|D'} = 0$,
$Wq' = W_{q|D} = Wq/P(D)$

2.18 Little's Law

In 1961, John Little published a paper showing that the expected number of units in the system, L, is related to the expected time in the system, W, by $L = \lambda W$, as long as the arrival rate λ is constant. In the same way, the following three relations are established:

$L = \lambda W$	= expected number of units in the system
$Ls = \lambda Ws$	= expected number of units in the service facility
$Lq = \lambda Wq$	= expected number of units in the queue

Using Little's Law,

$W = L/\lambda$	= expected time in the system
$Ws = Ls/\lambda$	= expected time in the service facility
$Wq = Lq/\lambda$	= expected time in the queue

2.19 Kendall's Notation

In queuing theory, Kendall's notation is the standard way to describe and classify the queuing systems. This method of classifying the systems was first suggested by D. G. Kendall in 1953 as a three-factor *A/B/C* notation system for identifying queues. It has since been extended to include up to six different factors.

The 3 factor notation (A/B/C) signifies the following:

A = arrival process
B = service time process
C = number servers

The six (6) factors (A/B/C/K/N/D) go even further where the latter three factors denote the following:

K = number places in system	(assume K = infinity unless specify other)
N = calling population	(assume N = infinity unless specify other)
D = service discipline	(assume non-priority unless specify other)

The arrival and service time factors (A,B)are denoted as below:

M = Markovian (Poisson or Exponential)
D = deterministic
Ek = Erlang with k stages
G = general

The service discipline (D)may take on the notation given below:

FIFO = first-in first-out
LIFO = last-in first-out
Random
Priority = preemptive or non-preemptive

Bibliography

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics*, *24*(3), 338–354.

Little, J. D. C. (1961). A proof of the queuing formula $L = \lambda W$. *Operations Research*, *9*(3), 383–387.