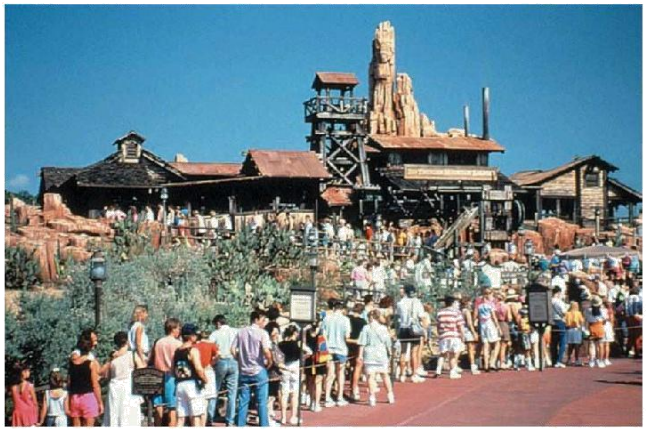
LEARNING OBJECTIVES

- **LO1** Describe the characteristics of arrivals, waiting lines, and service systems 737
- **LO2** Apply the single-server queuing model equations 742
- **LO3** Conduct a cost analysis for a waiting line 745
- **LO4** Apply the multiple-server queuing model formulas 745
- **LO5** Apply the constant-service-time model equations 750
- **LO6** Perform a limited-population model analysis 751



Jeff Greenberg/PhotoEdit Inc.

Paris's EuroDisney, Tokyo's Disney Japan, and the U.S.'s Disney World and Disneyland all have one feature in common—long lines and seemingly endless waits. However, Disney is one of the world's leading companies in the scientific analysis of queuing theory. It analyzes queuing behaviors and can predict which rides will draw what length crowds. To keep visitors happy, Disney makes lines appear to be constantly moving forward, entertains people while they wait, and posts signs telling visitors how many minutes until they reach each ride.

Queuing Theory

Queuing theory

A body of knowledge about waiting lines.

Waiting line (queue)

Items or people in a line awaiting service.

The body of knowledge about waiting lines, often called **queuing theory**, is an important part of operations and a valuable tool for the operations manager. **Waiting lines** are a common situation—they may, for example, take the form of cars waiting for repair at a Midas Muffler Shop, copying jobs waiting to be completed at a Kinko's print shop, or vacationers waiting to enter the Space Mountain ride at Disney. Table D.1 lists just a few OM uses of waiting-line models.

TABLE D.1 Common Queuing Situations

SITUATION	ARRIVALS IN QUEUE	SERVICE PROCESS
Supermarket	Grocery shoppers	Checkout clerks at cash register
Highway toll booth	Automobiles	Collection of tolls at booth
Doctor's office	Patients	Treatment by doctors and nurses
Computer system	Programs to be run	Computer processes jobs
Telephone company	Callers	Switching equipment forwards calls

Bank	Customers	Transactions handled by teller
Machine maintenance	Broken machines	Repair people fix machines
Harbor	Ships and barges	Dock workers load and unload

Waiting-line models are useful in both manufacturing and service areas. Analysis of queues in terms of waiting-line length, average waiting time, and other factors helps us to understand service systems (such as bank teller stations), maintenance activities (that might repair broken machinery), and shop-floor control activities. Indeed, patients waiting in a doctor's office and broken drill presses waiting in a repair facility have a lot in common from an OM perspective. Both use human and equipment resources to restore valuable production assets (people and machines) to good condition.

Characteristics of a Waiting-Line System

In this section, we take a look at the three parts of a waiting-line, or queuing, system (as shown in Figure D.1):

- **1.** *Arrivals or inputs to the system:* These have characteristics such as population size, behavior, and a statistical distribution.
- **2.** *Queue discipline, or the waiting line itself:* Characteristics of the queue include whether it is limited or unlimited in length and the discipline of people or items in it.
- **3.** *The service facility:* Its characteristics include its design and the statistical distribution of service times.

We now examine each of these three parts.

LO1 *Describe* the characteristics of arrivals, waiting lines, and service systems.

Arrival Characteristics

The input source that generates arrivals or customers for a service system has three major characteristics:

- **1.** *Size* of the arrival population
- **2.** *Behavior* of arrivals
- **3.** *Pattern* of arrivals (statistical distribution)

Size of the Arrival (Source) Population

Population sizes are considered either unlimited (essentially infinite) or limited (finite). When the number of customers or arrivals on hand at any given moment is just a small portion of all potential arrivals, the arrival population is considered **unlimited, or infinite**. Examples of unlimited populations include cars arriving at a big-city car wash, shoppers arriving at a supermarket, and students arriving to register for classes at a large university. Most queuing models assume such an infinite arrival population. An example of a **limited, or finite**, population is found in a copying shop that has, say, eight copying machines. Each of the copiers is a potential "customer" that may break down and require service.

Unlimited, or infinite, population

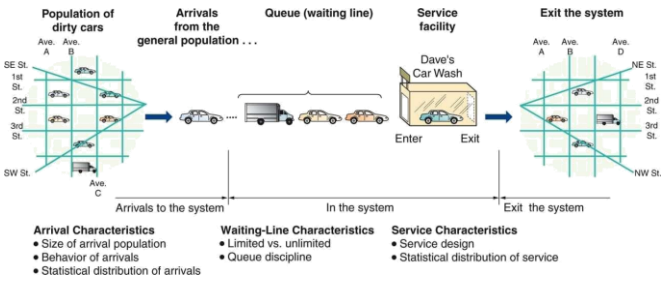
A queue in which a virtually unlimited number of people or items could request the services, or in which the number of customers or arrivals on hand at any given moment is a very small portion of potential arrivals.

Limited, or finite, population

A queue in which there are only a limited number of potential users of the service.

Pattern of Arrivals at the System

Figure D.1 Three Parts of a Waiting Line, or Queuing System, at Dave's Car Wash



Customers arrive at a service facility either according to some known schedule (for example, one patient every 15 minutes or one student every half hour) or else they arrive *randomly*. Arrivals are considered random when they are independent of one another and their occurrence cannot be predicted exactly. Frequently in queuing problems, the number of arrivals per unit of time can be estimated by a probability distribution known as the **Poisson distribution**.[1] For any given arrival time (such as 2 customers per hour or 4 trucks per minute), a discrete Poisson distribution can be established by using the formula:

Poisson distribution

A discrete probability distribution that often describes the arrival rate in queuing theory.

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \text{ for } x = 0, 1, 2, 3, 4, \ldots \quad (D\text{-}1)$$

where

$P(x)$ = probability of x arrivals

x = number of arrivals per unit of time

λ = average arrival rate

e = 2.7183 (which is the base of the natural logarithms)

With the help of the table in Appendix II, which gives the value of $e^{-\lambda}$ for use in the Poisson distribution, these values are easy to compute. Figure D.2 illustrates the Poisson distribution for $\lambda = 2$ and $\lambda = 4$. This means that if the average arrival rate is $\lambda = 2$ customers per hour, the probability of 0 customers arriving in any random hour is about 13%, probability of 1 customer is about 27%, 2 customers about 27%, 3 customers about 18%, 4 customers about 9%, and so on. The chances that 9 or more will arrive are virtually nil. Arrivals, of course, are not always Poisson distributed (they may follow some other distribution). Patterns, therefore, should be examined to make certain that they are well approximated by Poisson before that distribution is applied.

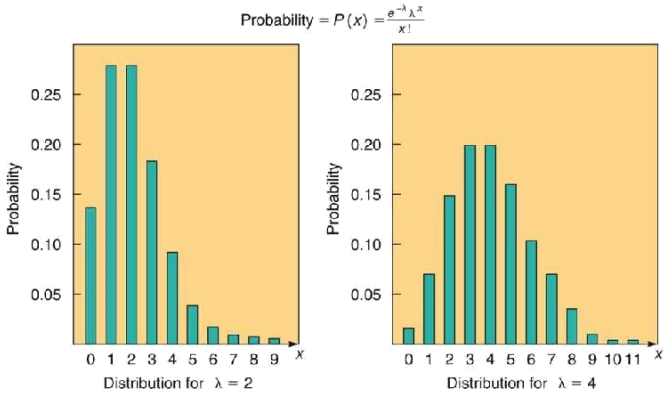
Behavior of Arrivals

Most queuing models assume that an arriving customer is a patient customer. Patient customers are people or machines that wait in the queue until they are served and do not switch between lines. Unfortunately, life is complicated by the fact that people have been known to balk or to renege. Customers who *balk* refuse to join the waiting line because it is too long to suit their needs or interests. *Reneging* customers are those who enter the queue but then become impatient and leave without completing their transaction. Actually, both of these situations just serve to highlight the need for queuing theory and waiting-line analysis.

Waiting-Line Characteristics

The waiting line itself is the second component of a queuing system. The length of a line can be either limited or unlimited. A queue is *limited* when it cannot, either by law or because of physical restrictions, increase to an infinite length. A small barbershop, for example, will have only a limited number of waiting chairs. Queuing models are treated in this module under an assumption of *unlimited* queue length. A queue is *unlimited* when its size is unrestricted, as in the case of the toll booth serving arriving automobiles.

Figure D.2 Two Examples of the Poisson Distribution for Arrival Times



$$\text{Probability} = P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

STUDENT TIP

Notice that even though the mean arrival rate might be $\lambda = 2$ per hour, there is still a small chance that as many as 9 customers arrive in an hour.

[1]When the arrival rates follow a Poisson process with mean arrival rate, λ, the time between arrivals follows a negative exponential distribution with mean time between arrivals of $1/\lambda$. The negative exponential distribution, then, is also representative of a Poisson process but describes the time between arrivals and specifies that these time intervals are completely random.

A second waiting-line characteristic deals with *queue discipline*. This refers to the rule by which customers in the line are to receive service. Most systems use a queue discipline known as the **first-in, first-out (FIFO) rule**. In a hospital emergency room or an express checkout line at a supermarket, however, various assigned priorities may preempt FIFO. Patients who are critically injured will move ahead in treatment priority over patients with broken fingers or noses. Shoppers with fewer than 10 items may be allowed to enter the express checkout queue (but are *then* treated as first-come, first-served). Computer-programming runs also operate under priority scheduling. In most large companies, when computer-produced paychecks are due on a specific date, the payroll program gets highest priority.[2]

First-in, first-out (FIFO) rule

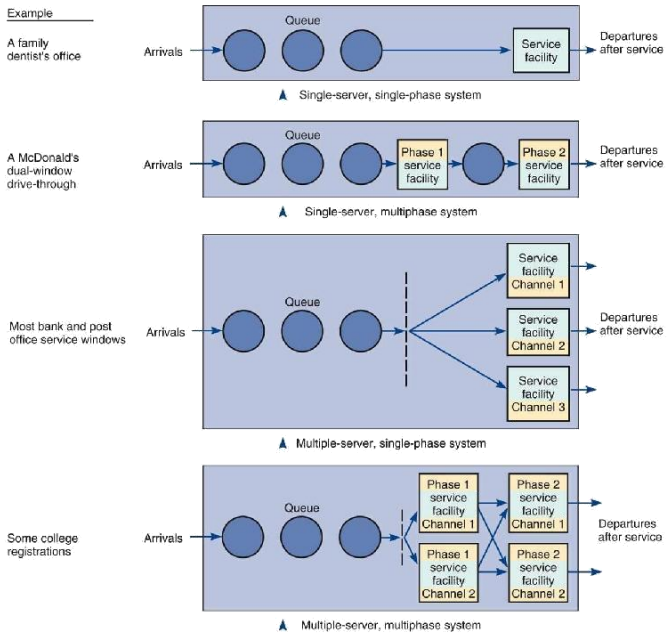
A queue discipline in which the first customers in line receive the first service.

Service Characteristics

The third part of any queuing system is the service characteristics. Two basic properties are important: (1) design of the service system and (2) the distribution of service times.

Basic Queuing System Designs

Figure D.3 Basic Queuing System Designs



[2]The term *FIFS* (first-in, first-served) is often used in place of FIFO. Another discipline, LIFS (last-in, first-served), also called last-in, first-out (LIFO), is common when material is stacked or piled so that the items on top are used first.

Service systems are usually classified in terms of their number of *servers* (number of channels) and number of *phases* (number of service stops that must be made). See Figure D.3. A **single-server** (or single-channel) **queuing system**, with one server, is typified by the drive-in bank with only one open teller. If, on the other hand, the bank has several tellers on duty, with each customer waiting in one common line for the first available teller, then we would have a **multiple-server** (or multiple channel) **queuing system**. Most banks today are multiple-server systems, as are most large barbershops, airline ticket counters, and post offices.

Multiple-server queuing system

A service system with one waiting line but with several servers.

In a **single-phase system**, the customer receives service from only one station and then exits the system. A fast-food restaurant in which the person who takes your order also brings your food and takes your money is a single-phase system. So is a driver's license agency in which the person taking your application also grades your test and collects your license fee. However, say the restaurant requires you to place your order at one station, pay at a second, and pick up your food at a third. In this case, it is a **multiphase system**. Likewise, if the driver's license agency is large or busy, you will probably have to wait in one line to complete your application (the first service stop), queue again to have your test graded, and finally go to a third counter to pay your fee. To help you relate the concepts of servers and phases, Figure D.3 presents these four possible configurations.

Single-server queuing system

A service system with one line and one server.

Single-phase system

A system in which the customer receives service from only one station and then exits the system.

Multiphase system

A system in which the customer receives services from several stations before exiting the system.

Service Time Distribution

Service patterns are like arrival patterns in that they may be either constant or random. If service time is constant, it takes the same amount of time to take care of each customer. This is the case in a machine-performed service operation such as an automatic car wash. More often, service times are randomly distributed. In many cases, we can assume that random service times are described by the **negative exponential probability distribution.**

Negative exponential probability distribution

A continuous probability distribution often used to describe the service time in a queuing system.

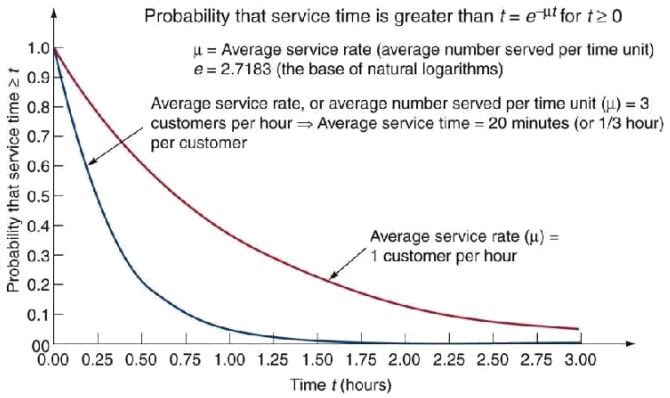
Figure D.4 shows that if *service times* follow a negative exponential distribution, the probability of any very long service time is low. For example, when an average service time is 20 minutes (or three customers per hour), seldom if ever will a customer require more than 1.5 hours in the service facility. If the mean service time is 1 hour, the probability of spending more than 3 hours in service is quite low.

Measuring a Queue's Performance

Queuing models help managers make decisions that balance service costs with waiting-line costs. Queuing analysis can obtain many measures of a waiting-line system's performance, including the following:

- **1.** Average time that each customer or object spends in the queue
- **2.** Average queue length
- **3.** Average time that each customer spends in the system (waiting time plus service time)
- **4.** Average number of customers in the system
- **5.** Probability that the service facility will be idle
- **6.** Utilization factor for the system
- **7.** Probability of a specific number of customers in the system

Figure D.4 Two Examples of the Negative Exponential Distribution for Service Times



STUDENT TIP

Although Poisson and exponential distributions are commonly used to describe arrival rates and service times, other probability distributions are valid in some cases.

OM in Action Zero Wait Time Guarantee at This Michigan Hospital's ER

Associated Press

Other hospitals smirked a few years ago when Michigan's Oakwood Healthcare chain rolled out an emergency room (ER) guarantee that promised a written apology and movie tickets to patients not seen by a doctor within 30 minutes. Even employees cringed at what sounded like a cheap marketing ploy.

But if you have visited an ER lately and watched some patients wait for hours on end—the *official* average wait is 47 minutes—you can understand why Oakwood's patient satisfaction levels have soared. The 30-minute guarantee was such a huge success that fewer than 1% of the 191,000 ER patients asked for free tickets. The following year, Oakwood upped the stakes again, offering a 15-minute guarantee. Then Oakwood started its Zero Wait Program in the ERs. Patients who enter any Oakwood emergency department are *immediately* cared for by a healthcare professional.

Oakwood's CEO even extended the ER guarantee to on-time surgery, 45-minute meal service orders, and other custom room services. "Medicine is a service business," says Larry Alexander, the head of an ER in Sanford, Florida. "And people are in the mindset of the fast-food industry."

How did Oakwood make good on its promise to eliminate the ER queue? It first studied queuing theory, then reengineered its billing, records, and lab operations to drive down service time. Then, to improve service capability, Oakwood upgraded its technical staff. Finally, it replaced its ER physicians with a crew willing to work longer hours.

Sources: Wall Street Journal (October 19, 2010); *Time* (January 26, 2011); and *Crain's Detroit Business* (March 4, 2002).

Queuing Costs

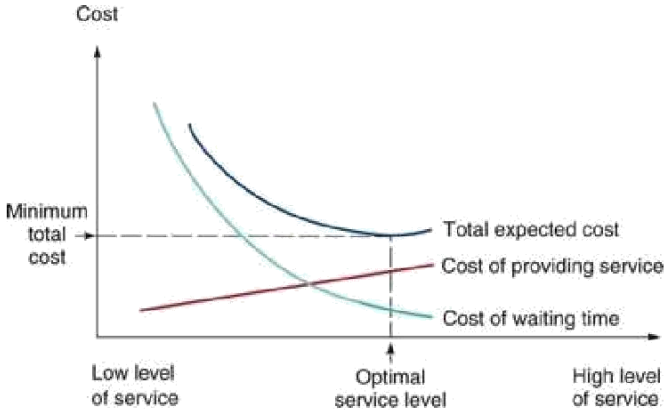
As described in the *OM in Action* box "Zero Wait Time Guarantee at This Michigan Hospital's ER," operations managers must recognize the trade-off that takes place between two costs: the cost of providing good service and the cost of customer or machine waiting time. Managers want queues that are short enough so that customers do not become unhappy and either leave without buying, or buy, but never return. However, managers may be willing to allow some waiting if it is balanced by a significant savings in service costs.

STUDENT TIP

The two costs we consider here are cost of servers and cost of lost time waiting.

One means of evaluating a service facility is to look at total expected cost. Total cost is the sum of expected service costs plus expected waiting costs.

Figure D.5 The Trade-off Between Waiting Costs and Service Costs



As you can see in Figure D.5, service costs increase as a firm attempts to raise its level of service. Managers in *some* service centers can vary capacity by having standby personnel and machines that they can assign to specific service stations to prevent or shorten excessively long lines. In grocery stores, for example, managers and stock clerks can open extra checkout counters. In banks and airport check-in points, part-time workers may be called in to help. As the level of service improves (that is, speeds up), however, the cost of time spent waiting in lines decreases. (Refer to Figure D.5.) Waiting cost may reflect lost productivity of workers while tools or machines await repairs or may simply be an estimate of the cost of customers lost because of poor service and long queues. In some service systems (for example, an emergency ambulance service), the cost of long waiting lines may be intolerably high.

The Variety of Queuing Models

A wide variety of queuing models may be applied in operations management. We will introduce you to four of the most widely used models. These are outlined in Table D.2, and examples of each follow in the next few sections. More complex models are described in queuing theory textbooks[3] or can be developed through the use of simulation (the topic of Module F). Note that all four queuing models listed in Table D.2 have three characteristics in common. They all assume:

- **1.** Poisson distribution arrivals
- **2.** FIFO discipline
- **3.** A single-service phase

STUDENT TIP

This is the main section of Module D. We illustrate four important queuing models.

In addition, they all describe service systems that operate under steady, ongoing conditions. This means that arrival and service rates remain stable during the analysis.

Model A (M/M/1): Single-Server Queuing Model with Poisson Arrivals and Exponential Service Times

The most common case of queuing problems involves the *single-server*, or single-channel, waiting line. In this situation, arrivals form a single line to be serviced by a single station (see Figure D.3 on p. 739). We assume that the following conditions exist in this type of system:

- **1.** Arrivals are served on a first-in, first-out (FIFO) basis, and every arrival waits to be served, regardless of the length of the line or queue.
- **2.** Arrivals are independent of preceding arrivals, but the average number of arrivals (*arrival rate*) does not change over time.
- **3.** Arrivals are described by a Poisson probability distribution and come from an infinite (or very, very large) population.
- **4.** Service times vary from one customer to the next and are independent of one another, but their average rate is known.

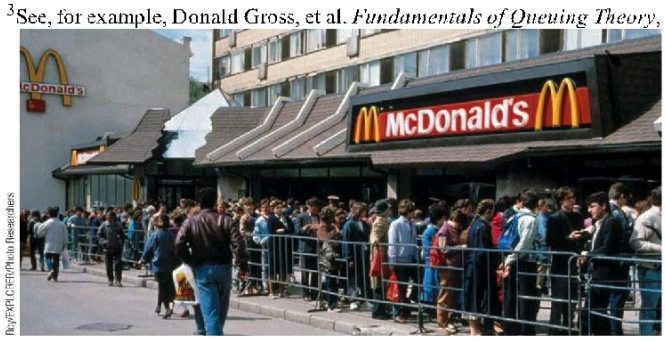
LO2 *Apply* the single-server queuing model equations

TABLE D.2 Queuing Models Described in This Chapter

MODEL	NAME (TECHNICAL NAME IN PARENTHESES)	EXAMPLE	NUMBER OF SERVERS (CHANNELS)	NUMBER OF PHASES	ARRIVAL RATE PATTERN	SERVICE TIME PATTERN	POPULATION SIZE	QUEUE DISCIPLINE
	Single-server	Information counter at						

A	system (M/M/1)	department store	Single	Single	Poisson	Exponential	Unlimited	FIFO
B	Multiple-server (M/M/S)	Airline ticket counter	Multi-server	Single	Poisson	Exponential	Unlimited	FIFO
C	Constant service (M/D/1)	Automated car wash	Single	Single	Poisson	Constant	Unlimited	FIFO
D	Limited population (finite population)	Shop with only a dozen machines that might break	Single	Single	Poisson	Exponential	Limited	FIFO

[3]See, for example, Donald Gross, et al. *Fundamentals of Queuing Theory*, 4th ed. New York: Wiley, 2008.



The giant Moscow McDonald's boasts 900 seats, 800 workers, and $80 million in annual sales (vs. less than $2 million in a U.S. outlet). Americans would balk at the average waiting time of 45 minutes, but Russians are used to such long lines. McDonald's represents good service in Moscow.

- **5.** Service times occur according to the negative exponential probability distribution.
- **6.** The service rate is faster than the arrival rate.

When these conditions are met, the series of equations shown in Table D.3 can be developed. Examples D1 and D2 illustrate how Model A (which in technical journals is known as the M/M/1 model) may be used.[4]

TABLE D.3 Queuing Formulas for Model a: Single-Server System, also Called M/M/1

λ = mean number of arrivals per time period

μ = mean number of people or items served per time period (average service rate)

L_s = average number of units (customers) in the system (waiting and being served)

$$= \frac{\lambda}{\mu - \lambda}$$

W_s = average time a unit spends in the system (waiting time plus service time)

$$= \frac{1}{\mu - \lambda}$$

L_q = average number of units waiting in the queue

$$= \frac{\lambda^2}{\mu(\mu - \lambda)}$$

W_q = average time a unit spends waiting in the queue

$$= \frac{\lambda}{\mu(\mu - \lambda)} = \frac{L_q}{\lambda}$$

ρ = utilization factor for the system

$$= \frac{\lambda}{\mu}$$

P_0 = probability of 0 units in the system (that is, the service unit is idle)

$$= 1 - \frac{\lambda}{\mu}$$

$P_n > k$ = probability of more than k units in the system, where n is the number of units in the system

$$= \left(\frac{\lambda}{\mu}\right)^{k+1}$$

[4]In queuing notation, the first letter refers to the arrivals (where M stands for Poisson distribution); the second letter refers to service (where M is again a Poisson distribution, which is the same as an exponential rate for service—and a D is a constant service rate); the third symbol refers to the number of servers. So an M/D/1 system (our Model C) has Poisson arrivals, constant service, and one server.

Example D1 A SINGLE-SERVER QUEUE

Tom Jones, the mechanic at Golden Muffler Shop, is able to install new mufflers at an average rate of 3 per hour (or about 1 every 20 minutes), according to a negative exponential distribution. Customers seeking this service arrive at the shop on the average of 2 per hour, following a Poisson distribution. They are served on a first-in, first-out basis and come from a very large (almost infinite) population of possible buyers.

We would like to obtain the operating characteristics of Golden Muffler's queuing system.

APPROACH This is a single-server (M/M/1) system, and we apply the formulas in Table D.3.

SOLUTION

λ = 2 cars arriving per hour

μ = 3 cars serviced per hour

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{2}{3 - 2} = \frac{2}{1}$$

= 2 cars in the system, on average

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{3 - 2} = 1$$

= 1-hour average time in the system

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{2^2}{3(3 - 2)} = \frac{4}{3(1)} = \frac{4}{3}$$

= 1.33 cars waiting in line, on average

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{2}{3(3 - 2)} = \frac{2}{3}\text{hour}$$

= 40-minute average waiting time per car

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$

= 66.6% of time mechanic is busy

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{2}{3}$$

= .33 probability there are 0 cars in the system

Probability of More Than *k* Cars in the System

k	$P_{n>k} = (2/3)^{k+1}$

0 .667 ← Note that this is equal to $1 - P_0 = 1 - .33 = .667$.

1 .444

2 .296

3 .198 ← Implies that there is a 19.8% chance that more than 3 cars are in the system.

4 .132

5 .088

6 .058

7 .039

INSIGHT Recognize that arrival and service times are converted to the same rate. For example, a service time of 20 minutes is stated as an average *rate* of 3 mufflers *per hour*. It's also important to differentiate between time in the *queue* and time in the *system*.

LEARNING EXERCISE If $\mu = 4$ cars/hour instead of the current 3 arrivals, what are the new values of L_s, W_s, L_q, W_q, and P_0? [Answer: 1 car, 30 min., .5 cars, 15 min., 50%, .50.]

RELATED PROBLEMS D.1, D.2, D.3, D.4, D.6, D.7, D.8, D.9a–e, D.10, D.11a–c, D.12a–d

EXCEL OM Data File **ModDExD1.xls** can be found at www.pearsonhighered.com/heizer.

ACTIVE **MODEL** D.1 This example is further illustrated in Active Model D.1 at www.pearsonhighered.com/heizer.

Once we have computed the operating characteristics of a queuing system, it is often important to do an economic analysis of their impact. Although the waiting-line model described above is valuable in predicting potential waiting times, queue lengths, idle times, and so on, it does not identify optimal decisions or consider cost factors. As we saw earlier, the solution to a queuing problem may require management to make a trade-off between the increased cost of providing better service and the decreased waiting costs derived from providing that service.

Example D2 examines the costs involved in Example D1.

Example D2 ECONOMIC ANALySIS OF ExAMPLE D1

Golden Muffler Shop's owner is interested in cost factors as well as the queuing parameters computed in Example D1. He estimates that the cost of customer waiting time, in terms of customer dissatisfaction and lost goodwill, is $15 per hour spent *waiting* in line. Jones, the mechanic, is paid $11 per hour.

APPROACH First compute the average daily customer waiting time, then the daily salary for Jones, and finally the total expected cost.

SOLUTION Because the average car has a $\frac{2}{3}$-hour wait (W_q) and because there are approximately 16 cars serviced per day (2 arrivals per hour times 8 working hours per day), the total number of hours that customers spend waiting each day for mufflers to be installed is:

$$\frac{2}{3}(16) = \frac{32}{3} = 10\frac{2}{3}\text{hour}$$

Hence, in this case:

$$\text{Customer waiting-time cost} = \$15\left(10\frac{2}{3}\right) = \$160\text{ per day}$$

LO3

Conduct a cost analysis for a waiting line

The only other major cost that Golden's owner can identify in the queuing situation is the salary of Jones, the mechanic, who earns $11 per hour, or $88 per day. Thus:

$$\text{Total expected costs} = \$160 + \$88$$

$$= \$248\text{ per day}$$

This approach will be useful in Solved Problem D.2 on page 756.

INSIGHT L_q and W_q are the two most important queuing parameters when it comes to cost analysis. Calculating customer wait times, we note, is based on average time waiting in the queue (W_q) times the number of arrivals per hour (λ) times the number of hours per day. This is because this example is set on a daily basis. This is the same as using λ_q, since $L_q = W_q\lambda$.

LEARNING EXERCISE If the customer waiting time is actually \$20 per hour and Jones gets a salary increase to \$15 per hour, what are the total daily expected costs? [Answer: \$333.33.]

RELATED PROBLEMS D.12e–f, D.13, D.22, D.23, D.24

Model B (M/M/S): Multiple-Server Queuing Model

Now let's turn to a multiple-server (multiple-channel) queuing system in which two or more servers are available to handle arriving customers. We still assume that customers awaiting service form one single line and then proceed to the first available server. Multiple-server, single-phase waiting lines are found in many banks today: a common line is formed, and the customer at the head of the line proceeds to the first free teller. (Refer to Figure D.3 on p. 739 for a typical multiple-server configuration.)

LO4 *Apply* the multiple-server queuing model formulas

The multiple-server system presented in Example D3 again assumes that arrivals follow a Poisson probability distribution and that service times are exponentially distributed. Service is first-come, first-served, and all servers are assumed to perform at the same rate. Other assumptions listed earlier for the single-server model also apply.

The queuing equations for Model B (which also has the technical name M/M/S) are shown in Table D.4. These equations are obviously more complex than those used in the single-server model, yet they are used in exactly the same fashion and provide the same type of information as the simpler model. (*Note:* The POM for Windows and Excel OM software described later in this chapter can prove very useful in solving multiple-server and other queuing problems.)

Photo courtesy of Costco Wholesale, 2012

To shorten lines (or wait times), each Costco register is staffed with two employees. This approach has improved efficiency by 20–30%.

TABLE D.4 Queuing Formulas for Model b: Multiple-Server System, also Called M/M/S

M = number of servers (channels) open

λ = average arrival rate

μ = average service rate at each server (channel)

The probability that there are zero people or units in the system is:

$$P_0 = \frac{1}{\left[\sum_{n=0}^{M-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{M!}\left(\frac{\lambda}{\mu}\right)^M \frac{M\mu}{M\mu - \lambda}} \quad \text{for } M\mu > \lambda$$

The average number of people or units in the system is:

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

The average time a unit spends in the waiting line and being serviced (namely, in the system) is:

$$W_s = \frac{\mu(\lambda/\mu)^M}{(M-1)!(M\mu-\lambda)^2}P_0 + \frac{1}{\mu} = \frac{L_s}{\lambda}$$

The average number of people or units in line waiting for service is:

$$L_q = L_s - \frac{\lambda}{\mu}$$

The average time a person or unit spends in the queue waiting for service is:

$$W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda}$$

Example D3 A MULTIPLE-SERVER QUEUE

The Golden Muffler Shop has decided to open a second garage bay and hire a second mechanic to handle installations. Customers, who arrive at the rate of about $\lambda = 2$ per hour, will wait in a single line until 1 of the 2 mechanics is free. Each mechanic installs mufflers at the rate of about $\mu = 3$ per hour.

The company wants to find out how this system compares with the old single-server waiting-line system.

APPROACH Compute several operating characteristics for the $M = 2$ server system, using the equations in Table D.4, and compare the results with those found in Example D1.

SOLUTION

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

$$= \frac{1}{1+\frac{2}{3}+\frac{1}{2}\left(\frac{4}{9}\right)\left(\frac{6}{6-2}\right)} + \frac{1}{1+\frac{2}{3}+\frac{1}{3}} = \frac{1}{2}$$

= .5 probability of zero cars in the system

Then:

$$L_s = \frac{(2)(3)^{(2/3)^2}}{1![2(3)-2]^2}\left(\frac{1}{2}\right) + \frac{2}{3} = \frac{8/3}{16}\left(\frac{1}{2}\right) + \frac{2}{3} = \frac{3}{4}$$

$$= .75 \text{ average number of cars in the system}$$

$$W_s = \frac{L_s}{\lambda} = \frac{3/4}{2} = \frac{3}{8} \text{ hour}$$

$$= 22.5 \text{ minutes average time a car spends in the system}$$

$$L_q = L_s - \frac{\lambda}{\mu} = \frac{3}{4} - \frac{2}{3} = \frac{9}{12} - \frac{8}{12} = \frac{1}{12}$$

$$= .083 \text{ average number of cars in the queue (waiting)}$$

$$W_q = \frac{L_q}{\lambda} = \frac{.083}{2} = .0415 \text{ hour}$$

$$= 2.5 \text{ minutes average time a car spends in the queue (waiting)}$$

INSIGHT It is very interesting to see the big differences in service performance when an additional server is added.

LEARNING EXERCISE If $\mu = 4$ per hour, instead of $\mu = 3$, what are the new values for P_0, L_s, W_s, L_q, and W_q? [Answers: 0.6, .53 cars, 16 min, .033 cars, 1 min.]

RELATED PROBLEMS D.7h, D.9f, D.11d, D.15, D.20

EXCEL OM Data File **ModDExD3.xls** can be found at **www.pearsonhighered.com/heizer**.

ACTIVE **MODEL** D.2 This example is further illustrated in Active Model D.2 at **www.pearsonhighered.com/heizer**.

We can summarize the characteristics of the two-server model in Example D3 and compare them to those of the single-server model in Example D1 as follows:

	SINGLE SERVER	TWO SERVERS (CHANNELS)
P_0	.33	.5
L_s	2 cars	.75 car
W_s	60 minutes	22.5 minutes
L_q	1.33 cars	.083 car
W_q	40 minutes	2.5 minutes

The increased service has a dramatic effect on almost all characteristics. For instance, note that the time spent waiting in line drops from 40 minutes to only 2.5 minutes.

Use of Waiting-Line Tables

Imagine the work a manager would face in dealing with $M = 3$, 4, or 5 server waiting-line models if a computer was not readily available. The arithmetic becomes increasingly troublesome. Fortunately, much of the burden of manually examining multiple-server queues can be avoided by using Table D.5. This table, the result of hundreds of computations, represents the relationship between three things: (1) a ratio, λ/μ, (2) number of servers open, and (3) the average number of customers in the queue, L_q (which is what we'd like to find). For any combination of the ratio λ/μ and $M = 1$, 2, 3, 4, or 5 servers, you can quickly look in the body of the table to read off the appropriate value for L_q.

TABLE D.5 Values of L_q for M = 1 − 5 Servers (channels) and Selected Values of λ/μ

POISSON ARRIVALS, EXPONENTIAL SERVICE TIMES					
	NUMBER OF SERVERS (CHANNELS), M				
λ/μ	1	2	3	4	5
.10	.0111				
.15	.0264	.0008			
.20	.0500	.0020			
.25	.0833	.0039			
.30	.1285	.0069			
.35	.1884	.0110			
.40	.2666	.0166			
.45	.3681	.0239	.0019		
.50	.5000	.0333	.0030		
.55	.6722	.0449	.0043		
.60	.9000	.0593	.0061		
.65	1.2071	.0767	.0084		
.70	1.6333	.0976	.0112		
.75	2.2500	.1227	.0147		
.80	3.2000	.1523	.0189		
.85	4.8166	.1873	.0239	.0031	
.90	8.1000	.2285	.0300	.0041	
.95	18.0500	.2767	.0371	.0053	
1.0		.3333	.0454	.0067	
1.2		.6748	.0904	.0158	

1.4	1.3449	.1778	.0324	.0059
1.6	2.8444	.3128	.0604	.0121
1.8	7.6734	.5320	.1051	.0227
2.0		.8888	.1739	.0398
2.2		1.4907	.2770	.0659
2.4		2.1261	.4305	.1047
2.6		4.9322	.6581	.1609
2.8		12.2724	1.0000	.2411
3.0			1.5282	.3541
3.2			2.3856	.5128
3.4			3.9060	.7365
3.6			7.0893	1.0550
3.8			16.9366	1.5184
4.0				2.2164
4.2				3.3269
4.4				5.2675
4.6				9.2885
4.8				21.6384

Example D4 illustrates the use of Table D.5.

Example D4 USE OF WAITING-LINE TABLES

Alaska National Bank is trying to decide how many drive-in teller windows to open on a busy Saturday. CEO Ted Eschenbach estimates that customers arrive at a rate of about $\lambda = 18$ per hour, and that each teller can service about $\mu = 20$ customers per hour.

APPROACH Ted decides to use Table D.5 to compute L_q and W_q.

SOLUTION The ratio is $\lambda / \mu = \frac{18}{20} = .90.$ Turning to the table, under $\lambda / \mu = .90,$, Ted sees that if only $M = 1$ service window is open, the average number of customers in line will be 8.1. If two windows are open, L_q drops to .2285 customers, to .03 for $M = 3$ tellers, and to .0041 for $M = 4$ tellers. Adding more open windows at this point will result in an average queue length of 0.

It is also a simple matter to compute the average waiting time in the queue, W_q, since $W_q = L_q/\lambda$. When one service window is open, W_q = 8.1 customers/(18 customers per hour) = .45 hours = 27 minutes waiting time; when two tellers are open, W_q = .2285 customers/(18 customers per hour) = .0127 hours $\cong \frac{3}{4}$ minute; and so on.

INSIGHT If a computer is not readily available, Table D.5 makes it easy to find L_q and to then compute W_q. Table D.5 is especially handy to compare L_q for different numbers of servers (M).

LEARNING EXERCISE The number of customers arriving on a Thursday afternoon at Alaska National is 15/hour. The service rate is still 20 customers/hour. How many people are in the queue if there are 1, 2, or 3 servers? [Answer: 2.25, .1227, .0147.]

RELATED PROBLEM D.5

You might also wish to check the calculations in Example D3 against tabled values just to practice the use of Table D.5. You may need to interpolate if your exact value is not found in the first column. Other common operating characteristics besides L_q are published in tabular form in queuing theory textbooks.

Long check-in lines (left photo) such as at Los Angeles International (LAX) are a common airport sight. This is an M/M/S model—passengers wait in a single queue for one of several agents. But at Anchorage International Airport (right photo), Alaska Air has jettisoned the traditional wall of ticket counters. Instead, 1.2 million passengers per year use self-service check-in machines and staffed "bag drop" stations. Looking nothing like a typical airport, the new system doubled the airline's check-in capacity and cut staff needs in half, all while speeding travelers through in less than 15 minutes, even during peak hours.

Model C (M/D/1): Constant-Service-Time Model

LO5 *Apply* the constant-service-time model equations

Some service systems have constant, instead of exponentially distributed, service times. When customers or equipment are processed according to a fixed cycle, as in the case of an automatic car wash or an amusement park ride, constant service times are appropriate. Because constant rates are certain, the values for L_q, W_q, L_s, and W_s are always less than they would be in Model A, which has variable service rates. As a matter of fact, both the average queue length and the average waiting time in the queue are halved with Model C. Constant-service-model formulas are given in Table D.6. Model C also has the technical name M/D/1 in the literature of queuing theory.

TABLE D.6 Queuing Formulas for Model C: Constant Service, also Called M/D/1

Average length of queue: $L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)}$

Average waiting time in queue: $W_q = \frac{\lambda}{2\mu(\mu - \lambda)}$

Average number of customers in system: $L_s = L_q + \frac{\lambda}{\mu}$

Average time in system: $W_s = W_q + \frac{1}{\mu}$

Example D5 gives a constant-service-time analysis.

Example D5 A CONSTANT-SERVICE-TIME MODEL

Inman Recycling, Inc., collects and compacts aluminum cans and glass bottles in Reston, Louisiana. Its truck drivers currently wait an average of 15 minutes before emptying their loads for recycling. The cost of driver and truck time while they are in queues is valued at $60 per hour. A new automated compactor can be purchased to process truckloads at a *constant* rate of 12 trucks per hour (that is, 5 minutes per truck). Trucks arrive according to a Poisson distribution at an average rate of 8 per hour. If the new compactor is put in use, the cost will be amortized at a rate of $3 per truck unloaded.

APPROACH CEO Tony Inman hires a summer college intern to conduct an analysis to evaluate the costs versus benefits of the purchase. The intern uses the equation for W_q in Table D.6.

SOLUTION

Current waiting cost/trip = (1/4 hr waiting now)($60/hr cost) = $15/trip

New system: $\lambda = 8$ trucks/hr arriving $\mu = 12$ trucks/hr served

Average waiting time in queue = $$W_q = \frac{\lambda}{2\mu(\mu - \lambda)} = \frac{8}{2(12)(12 - 8)} = \frac{1}{12}\text{hr}$$

Waiting cost/trip with new compactor = (1/12 hr wait)($60/hr cost) = $5/trip

Savings with new equipment = $15(current system) − $5(new system) = $10/trip

Cost of new equipment amortized: = $3/trip

Net savings = $7/trip

INSIGHT Constant service times, usually attained through automation, help control the variability inherent in service systems. This can lower average queue length and average waiting time. Note the 2 in the denominator of the equations for L_q and W_q in Table D.6.

LEARNING EXERCISE With the new constant-service-time system, what are the average waiting time in the queue, average number of trucks in the system, and average waiting time in the system? [Answer: 0.0833 hours, 1.33, 0.1667 hours.]

RELATED PROBLEMS D.14, D.16, D.21

EXCEL OM Data File **ModDExD5.xls** can be found at **www.pearsonhighered.com/heizer**.

ACTIVE **MODEL** D.3 This example is further illustrated in Active Model D.3 at **www.pearsonhighered.com/heizer**.

Little's Law

A practical and useful relationship in queuing for any system in a *steady state* is called Little's Law. A steady state exists when a queuing system is in its normal operating condition (e.g., after customers waiting at the door when a business opens in the morning are taken care of). Little's Law can be written as either:

$$L = \lambda W \text{ (which is the same as } W = L/\lambda) \quad \text{(D-2)}$$

or:

$$L_q = \lambda W_q \text{ (which is the same as } W_q = L_q/\lambda) \quad \text{(D-3)}$$

The advantage of these formulas is that once two of the parameters are known, the other one can easily be found. This is important because in certain waiting-line situations, one of these might be easier to determine than the other.

Little's Law is also important because it makes no assumptions about the probability distributions for arrivals and service times, the number of servers, or service priority rules. The law applies to all the queuing systems discussed in this module, except the limited-

population model, which we discuss next.

Model D: Limited-Population Model

When there is a limited population of potential customers for a service facility, we must consider a different queuing model. This model would be used, for example, if we were considering equipment repairs in a factory that has 5 machines, if we were in charge of maintenance for a fleet of 10 commuter airplanes, or if we ran a hospital ward that has 20 beds. The limited-population model allows any number of repair people (servers) to be considered.

LO6 *Perform* a limited-population model analysis

This model differs from the three earlier queuing models because there is now a *dependent* relationship between the length of the queue and the arrival rate. Let's illustrate the extreme situation: If your factory had five machines and all were broken and awaiting repair, the arrival rate would drop to zero. In general, then, as the *waiting line* becomes longer in the limited population model, the *arrival rate* of customers or machines drops.

TABLE D.7 Queuing Formulas and Notation for Model D: Limited-Population Formulas

Service factor: $X = \frac{T}{T+U}$	Average number of units running: $J = NF(1 - X)$
Average number waiting: $L_q = N(1 - F)$	Average number being serviced: $H = FNX$
Average waiting time: $W_q = \frac{L_q(T+U)}{N - L_q} = \frac{T(1-F)}{XF}$	Number in population: $N = J + L_q + H = FNX$
Notation	
D = probability that a unit will have to wait in queue	N = number of potential customers
F = efficiency factor	T = average service time
H = average number of units being served	U = average time between unit service requirements
J = average number of units in working order	W_q = average time a unit waits in line
L_q = average number of units waiting for service	X = service factor
M = number of servers (channels)	

Table D.7 displays the queuing formulas for the limited-population model. Note that they employ a different notation than Models A, B, and C. To simplify what can become time-consuming calculations, finite queuing tables have been developed that determine D and F. D represents the probability that a machine needing repair will have to wait in line. F is a waiting-time efficiency factor. D and F are needed to compute most of the other finite model formulas.

A small part of the published finite queuing tables is illustrated in this section. Table D.8 provides data for a population of $N = 5$.[5]

To use Table D.8, we follow four steps:

- **1.** Compute X (the service factor), where $X = T/(T + U)$.
- **2.** Find the value of X in the table and then find the line for M (where M is the number of servers).
- **3.** Note the corresponding values for D and F.
- **4.** Compute L_q, W_q, J, H, or whichever are needed to measure the service system's performance.

Example D6 illustrates these steps.

Example D6 A LIMITED-POPULATION MODEL

Past records indicate that each of the 5 massive laser computer printers at the U.S. Department of Energy (DOE), in Washington, DC, needs repair after about 20 hours of use. Breakdowns have been determined to be Poisson distributed. The one technician on duty can service a printer in an average of 2 hours, following an exponential distribution. Printer downtime costs $120 per hour. Technicians are paid $25 per hour. Should the DOE hire a second technician?

APPROACH Assuming the second technician can also repair a printer in an average of 2 hours, we can use Table D.8 (because there are $N = 5$ machines in this limited population) to compare the costs of 1 vs. 2 technicians.

SOLUTION

- **1.** First, we note that $T = 2$ hours and $U = 20$ hours.
- **2.** Then, $X = \frac{T}{T+U} = \frac{2}{2+20} = \frac{2}{22} = .091$ (close to .090 [to use for determining D and F]).
- **3.** For $M = 1$ server, $D = .350$ and $F = .960$.
- **4.** For $M = 2$ servers, $D = .044$ and $F = .998$.
- **5.** The average number of printers working is $J = NF(1 - X)$. For $M = 1$, this is $J = (5)(.960)(1 - .091) = 4.36$. For $M = 2$, it is $J = (5)(.998)(1 - .091) = 4.54$.
- **6.** The cost analysis follows:

NUMBER OF TECHNICIANS	AVERAGE NUMBER PRINTERS DOWN (N − J)	AVERAGE COST/HR FOR DOWNTIME (N − J)($120/HR)	COST/HR. FOR TECHNICIANS (AT $25/HR)	TOTAL COST/HR
1	.64	$76.80	$25.00	$101.80
2	.46	$55.20	$50.00	$105.20

INSIGHT This analysis suggests that having only one technician on duty will save a few dollars per hour ($105.20 − $101.80 = $3.40). This may seem like a small amount, but it adds up to over $7,000 per year.

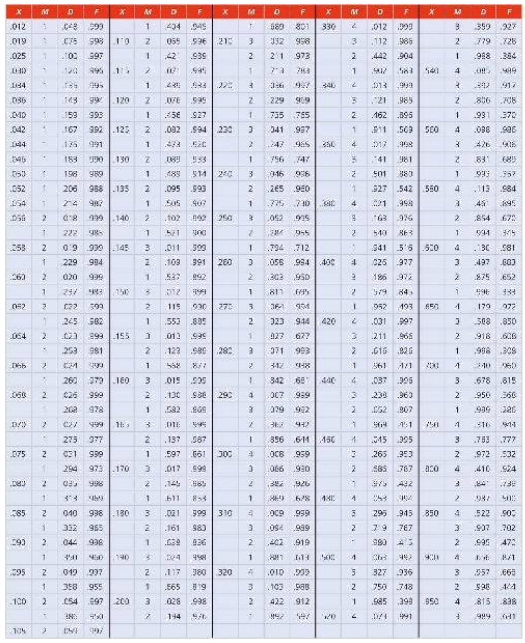
LEARNING EXERCISE DOE has just replaced its printers with a new model that seems to break down after about 18 hours of use. Recompute the costs. [Answer: For $M = 1$, $F = .95$, $J = 4.275$, and total cost/hr = $112.00. For $M = 2$, $F = .997$, $J = 4.487$, and total cost/hr = $111.56.]

RELATED PROBLEMS D.17, D.18, D.19

EXCEL OM Data File **ModDExD6.xls** can be found at **www.pearsonhighered.com/heizer**.

[5]Limited, or finite, queuing tables are available to handle arrival populations of up to 250. Although there is no definite number that we can use as a dividing point between limited and unlimited populations, the general rule of thumb is this: If the number in the queue is a significant proportion of the arrival population, use a limited population queuing model. For a complete set of N-values, see L. G. Peck and R. N. Hazelwood, *Finite Queuing Tables*. New York: Wiley, 1958.

TABLE D.8 Finite Queuing Tables for a Population of $N = 5$*



*See notation in Table D.7.

*See notation in Table D.7.

Other Queuing Approaches

STUDENT TIP

When the assumptions of the 4 models we just introduced do not hold true, there are other approaches still available to us.

Many practical waiting-line problems that occur in service systems have characteristics like those of the four mathematical models already described. Often, however, *variations* of these specific cases are present in an analysis. Service times in an automobile repair shop, for example, tend to follow the normal probability distribution instead of the exponential. A college registration system in which seniors have first choice of courses and hours over other students is an example of a first-come, first-served model with a preemptive priority queue discipline. A physical examination for military recruits is an example of a multiphase system, one that differs from the single-phase models discussed earlier in this module. A recruit first lines up to have blood drawn at one station, then waits for an eye exam at the next station, talks to a psychiatrist at the third, and is examined by a doctor for medical problems at the fourth. At each phase, the recruit must enter another queue and wait his or her turn. Many models, some very complex, have been developed to deal with situations such as these.

Summary

Queues are an important part of the world of operations management. In this module, we describe several common queuing systems and present mathematical models for analyzing them.

The most widely used queuing models include Model A, the basic single-server, single-phase system with Poisson arrivals and exponential service times; Model B, the multiple-server equivalent of Model A; Model C, a constant-service-rate model; and Model D, a limited-population system. All four models allow for Poisson arrivals; first-in, first-out service; and a single-service phase. Typical operating characteristics we examine include average time spent waiting in the queue and system, average number of customers in the queue and system, idle time, and utilization rate.

A variety of queuing models exists for which all the assumptions of the traditional models need not be met. In these cases, we use more complex mathematical models or turn to a technique called *simulation*. The application of simulation to problems of queuing systems is addressed in Module F.

Key Terms

- Queuing theory (p. 736) Waiting line (queue) (p. 736) Unlimited, or infinite, population (p. 737) Limited, or finite, population (p. 737) Poisson distribution (p. 738) First-in, first-out (FIFO) rule (p. 739) Single-server queuing system (p. 739) Multiple-server queuing system (p. 740) Single-phase system (p. 740) Multiphase system (p. 740) Negative exponential probability distribution (p. 740)

Discussion Questions

1.

Name the three parts of a typical queuing system.

2.

When designing a waiting line system, what "qualitative" concerns need to be considered?

3.

Name the three factors that govern the structure of "arrivals" in a queuing system.

4.

State the seven common measures of queuing system performance.

5.

State the assumptions of the "basic" single-server queuing model (Model A, or M/M/1).

6.

Is it good or bad to operate a supermarket bakery system on a strict first-come, first-served basis? Why?

7.

Describe what is meant by the waiting-line terms *balk* and *renege*. Provide an example of each.

8.

Which is larger, W_s or W_q? Explain.

9.

Briefly describe three situations in which the first-in, first-out (FIFO) discipline rule is not applicable in queuing analysis.

10.

Describe the behavior of a waiting line where $\lambda > \mu$. Use both analysis and intuition.

11.

Discuss the likely outcome of a waiting line system where $\mu > \lambda$ but only by a tiny amount (e.g., $\mu = 4.1$, $\lambda = 4$).

12.

Provide examples of four situations in which there is a limited, or finite, waiting line.

13.

What are the components of the following queuing systems? Draw and explain the configuration of each.

- a) Barbershop
- b) Car wash
- c) Laundromat
- d) Small grocery store