OPERATIONS MANAGEMENT -- FOCUSING ON QUALITY & COMPETITIVENESS

CHAPTER 16
WAITING LINE MODELS FOR SERVICE IMPROVEMENT

LEARNING OBJECTIVES

This chapter shows various mathematical formulations for evaluating the operating characteristics waiting lines. By the end of the chapter, you should be able to answer the following questions:

1. What elements of a waiting line must be known to use queuing theory models?

2. What are the basic structures of waiting lines?

3. What operating characteristics are typically calculated when evaluating the performance of a service?

4. Give some examples of how each of the following waiting models are used: basic single-server model, constant service time model, finite queue length model, finite calling population model, and multiple-channel, single-phase model.

You should also know how to use the formulas representing each model and be able to evaluate the results.

Click here when you are ready to take the readiness assessment test.

LECTURE NOTES -- WAITING LINE MODELS

Introduction

Think back to the supermarket example we discussed in Chapter 14. One of the principles designed into supermarket checkout systems is that an express line can be used in conjunction with first-come, first-served prioritization to process customers more rapidly. In this chapter, we’ll use queuing theory to determine how many checkout stands are necessary to provide acceptable customer service in general. Speed of service has become increasingly important as consumers evaluate service quality
with such criteria.

Elements of Waiting Line Analysis

Waiting lines form because people or things arrive at the servicing function faster than they can be served every single time. This does not mean that the service function is understaffed, rather it means that waiting lines form because customers do not arrive at a constant rate, nor are they all served in an equal amount of time. A waiting line is continually increasing and decreasing in length, and in the long run approaches an average rate of customer arrivals and an average time to serve the customers.

Decisions about waiting lines and the management of waiting lines are based on the averages for customer arrivals and service times. They are used in queuing formulas to compute operating characteristics, such as the average number of customers waiting in line and the average time a customer must wait in line.

Different sets of formulas are used to calculate operating characteristics, depending on the type of waiting line system being investigated. The last part of this lecture will show examples using these formulas.

Elements of a Waiting Line

The basic elements of a waiting line, or queue, are

- arrivals
- servers
- the waiting line structure

Following is a brief description of each waiting line component.

The Calling Population

The calling population is the source of customers to the queuing system, and it can be either infinite
or finite.

- Infinite - a large enough population that one more customer can always arrive to be served.

- Finite - a countable number of potential customers.

**The Arrival Rate**

The arrival rate is the rate at which customers arrive at the service facility during a specified period of time.

- It expresses the frequency of customer arrivals at a waiting line system.

- It typically follows a Poisson distribution.

- Average arrival rate = $\lambda$

**Service Times**

Service is expressed in terms of time, but it can be converted to a rate to be compatible with the arrival rate. Service can be represented by a number of probability distributions, but it typically follows the negative exponential distribution.

- Service times often follow a negative exponential distribution.

- Average service rate = $\mu$

**Arrival Rate Less than Service Rate**
Arrival rate ($\lambda$) must be less than service rate $\mu$ or the system never clears out.

**Queue Discipline and Length**

Queue discipline is the order in which customers are served. First come, first served is the most common, but random and last-in, first out is possible in some manufacturing or service systems. Queues lengths can be infinite or finite.

- Infinite is most common.

- Finite is limited by some physical structure, like a driveway that can accommodate only a limited number of cars.

**Basic Waiting Line Structures**

Waiting lines are generally categorized into four basic structures, according to the nature of the service facilities: single-channel, single-phase; single-channel, multiple-phase; multiple-channel, single-phase; and multiple-channel, multiple-phase.

- **Channels** are the number of parallel servers.

- **Phases** denote number of sequential servers the customer must go through.

Figure 16.2 illustrates the basic waiting line structures.
Operating Characteristics

Operating characteristics are the criteria that can be used to evaluate the performance of a queuing system. They are descriptive, not optimal decision results.

- Mathematics of queuing theory does not provide the optimal or best solutions.

- Computed operating characteristics describe system performance.

- Steady state is the constant, average value for performance characteristics that the system will reach after a long time.

- The results of operating characteristic calculations should be used to evaluate if the performance of the system satisfactorily satisfies customers and company policy.

- Typical operating characteristics computed are:
**Notation Description**

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L</strong></td>
</tr>
<tr>
<td><strong>L_q</strong></td>
</tr>
<tr>
<td><strong>W</strong></td>
</tr>
<tr>
<td><strong>W_q</strong></td>
</tr>
<tr>
<td><strong>P_0</strong></td>
</tr>
<tr>
<td><strong>P_n</strong></td>
</tr>
<tr>
<td><strong>ρ</strong></td>
</tr>
</tbody>
</table>

The concepts we’ve reviewed here have been used to derive mathematical equations that characterize waiting lines. Several applications of the equations will be reviewed in the next sections.

**Single-Channel, Single-Phase Models**

The simplest, most basic of the waiting line structures illustrated in Figure 16.2 is the single-channel, single-phase model. There are several variations of the model that will be reviewed:

- All assume a Poisson arrival rate

- Variations to be reviewed use:
  - exponential service times
  - constant service times
  - exponential service times with finite queue length
  - exponential service times with finite calling population
The Basic Single-Server Model

The assumptions of the basic single-server model are

- Poisson arrival rate
- exponential service times
- first-come, first-served queue discipline
- infinite queue length
- infinite calling population

- $\lambda = \text{mean arrival rate}$
- $\mu = \text{mean service rate}$

Formulas used for calculating the operating characteristics are shown in the following table:

**Formulas For the Single-Server Model**

Probability that no customers are in the system:

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right)$$

Probability of exactly $n$ customers in the system:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

Average number of customers in the system:

$$L = \mu$$

Average number of customers in waiting line:

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

Average time a customer spends in the system:

$$W = \frac{1}{\mu - \lambda}$$
Average time a customer spends waiting in line to be served:

\[ W_q = \frac{\lambda}{\mu (\mu - \lambda)} \]

Single-Server Example:

A drive-in market has one checkout counter where one employee operates the cash register. The combination of the cash register and the operator is the server in this queuing system; the customers who line up at the counter to pay for their selections form the waiting line. Customers arrive at a rate of 24 per hour according to a Poisson distribution, and service times are exponentially distributed with a mean rate of 30 customers per hour. What are the operating characteristics for this waiting line system?

Solution:

Given \( \lambda = 24 \) per hour, \( \mu = 30 \) customers per hour, each operating characteristic is calculated in the table below:

<table>
<thead>
<tr>
<th>Operating Characteristic</th>
<th>Calculation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability that no customers are in the system:</td>
<td>( P_0 = \left(1 - \frac{\lambda}{\mu}\right) )</td>
<td>( P_0 = \left(1 - \frac{24}{30}\right) = 0.20 )</td>
</tr>
<tr>
<td>Average number of customers in the system:</td>
<td>( L = \frac{C \cdot Q}{2} )</td>
<td>( L = \frac{24}{30 - 24} = 4 )</td>
</tr>
<tr>
<td>Average number of customers in waiting line:</td>
<td>( L_q = \frac{\lambda^2}{\mu (\mu - \lambda)} )</td>
<td>( L_q = \frac{24^2}{30(30 - 24)} = 3.2 )</td>
</tr>
<tr>
<td>Average time a customer spends in the system:</td>
<td>( W = \frac{1}{\mu - \lambda} )</td>
<td>( W = \frac{1}{30 - 24} = 0.167 ) hr. = 10 min</td>
</tr>
</tbody>
</table>
Average time a customer spends waiting in line to be served:

\[ W_q = \frac{\lambda}{\mu(\mu - \lambda)} \]

\[ W_q = \frac{24}{30(30 - 24)} = 0.133 \text{ hr} \]

\[ = 8 \text{ min} \]

We see that, on average, the number of customers in the system, 4, is fairly high. Also, an average wait of 8 minutes and average time in the system of 10 minutes is relatively high (try waiting in a line that long some time!) Let’s evaluate if it is cost effective to add additional employees or an extra service station.

Waiting Line Cost Analysis

To improve customer service management wants to test two alternatives to reduce customer waiting time:

1. Hire another employee to pack up purchases.
2. Install another checkout counter and hire another employee to service that counter.

Alternative 1

Here are the facts used to evaluate alternative 1:

- The extra employee costs $150/week
- Each one-minute reduction in customer waiting time avoids $75 in lost sales
- The extra employee will increase service rate to 40 customers per hour

When several alternatives must be tested using similar calculations, the best method is to use a spreadsheet and enter formulas only once. Input data can be changed to accommodate the evaluation of alternatives. The solution for the problem can be see by clicking here.

The operating characteristics for alternative 1 are summarized below:
• \( W_q = 0.038 \) hours = 2.25 minutes, originally was 8 minutes
• Time savings = 8 minutes - 2.25 minutes = 5.75 minutes
• 5.75 minutes \times $75/\text{minute/week} = $431.25 \text{ per week}
• The new employee saves $431.25 - 150.00 = $281.25/\text{wk}

Alternative 2

Here are the facts used to evaluate alternative 2:

• The new counter costs $6000 plus $200 per week for a checker.
• Customers divide themselves equally between the two checkout lines
• Arrival rate is reduced from \( \lambda = 24 \) to \( \lambda = 12 \).
• Service rate for each checker is \( \mu = 30 \).

To review the spreadsheet calculations for this problem, click here.

The operating characteristics for alternative 2 are summarized below:

• \( W_q = 0.022 \) hours = 1.33 minutes, originally was 8 minutes
• 8.00 minutes -1.33 minutes = 6.67 minutes
• 6.67 minutes \times $75/\text{minute/week} = $500.00/\text{wk} - 200.00 = $300/\text{wk}
• Counter is paid off in 6000/300 = 20 weeks
• Counter saves $300/\text{wk after week 20}; choose alternative 2.

If the cost of lost sales are as high as $75 per minute of waiting time, it makes good sense to install an extra counter. Even if costs cannot be computed, the performance of the system with a single counter is very poor; it is probably in the business’s best interest to reduce customer waiting time by implementing alternative 2.

IN-LINE EXERCISE

a. Evaluate the waiting line performance of the drive-in market if a third counter is installed. The third counter changes the arrival rate to \( \lambda = 8 \) customers per hour.
   The service rate is the same at \( \mu = 30 \) customers per hour. What is the average time spent in the waiting line (not in the system, which includes the service time in addition to the waiting time)?
Answer

Constant Service Times

The single-server model with Poisson arrivals and constant service times is used most frequently for systems with automated equipment and machinery. In the case of constant service times, there is no variability in service times.

- Constant service times occur with machinery and automated equipment
- Constant service times are a special case of the single-server model with general or undefined service times
- Operating characteristics for constant service times can be calculated with the formulas below:

**Formulas For Constant Service Time Single-Server Model**

Probability that no customers are in the system:

\[ P_0 = \left(1 - \frac{\lambda}{\mu}\right) \]

Average number of customers in the system:

\[ L = \frac{\lambda}{\mu} + L_q \]

Average number of customers in waiting line:

\[ L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} \]

Average time a customer spends in the system:

\[ W = W_q + \frac{1}{\mu} \]
Average time a customer spends waiting in line to be served:

\[ W_q = \frac{L_q}{\lambda} \]

**Constant Service Time Example:**

An automatic car wash accommodates one car at a time, and it requires a constant time of 4.5 minutes for a wash. Cars arrive at the car wash at an average rate of 10 per hour (Poisson-distributed). What is the average length of the waiting line and the average time at the car wash.

**Given:**

- **Automated car wash with service time = 4.5 min**
- **Cars arrive at rate \( \lambda = 10 \) per hour (Poisson)**
- \( \mu = 60 \) minutes per hour/4.5 minutes = 13.3 per hour

**Solution:**

The spreadsheet can be viewed by clicking here, or you can review the calculations in the table below.

| Average number of customers in waiting line: | \( L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} \) | \( L_q = \frac{10^2}{2(13.33)(13.33 - 10)} = 1.13 \) cars |
| Average time a customer spends waiting in line to be served: | \( W_q = \frac{L_q}{\lambda} \) | \( W_q = \frac{1.13}{10} = 0.113 \) hr. = 6.84 min |

The amount of time in line, 6.84 minutes is high for a car wash. The company should consider installing a second wash bay.

**Finite Queue Length**
For some waiting line systems, the length of the queue may be limited by the physical area in which the queue forms; space may permit only a limited number of customers to enter the queue. Such a waiting line is referred to as a finite queue.

- A physical limit exists on the length of queue.
- $M = \text{maximum number in queue}$.
- Service rate does not have to exceed arrival rate ($\mu > \lambda$) to obtain steady-state conditions.
- The operating characteristics of this variation of the single-server model can be calculated with the following formulas:

**Formulas For Finite Queue Length Single-Server Model**

Probability that no customers are in the system:

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

Probability of exactly $n$ customers in the system:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 \text{ for } n \leq M$$

Average number of customers in the system:

$$L = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} \left(\frac{\lambda}{\mu}\right)^{M+1} \frac{M + 1}{1 - (\frac{\lambda}{\mu})^{M+1}}$$

Average number of customers in waiting line:

$$L_q = L - \frac{\lambda(1 - P_N)}{\mu}$$

Average time a customer spends in the system:

$$W = \frac{L}{\lambda(1 - F_M)}$$
Average time a customer spends waiting in line to be served:

\[ W_q = W - \frac{1}{\mu} \]

**Finite Queue Example:**

Slick’s Quick Lube has space for only one vehicle in service and three vehicles lined up to wait for service. The mean time between arrivals for customers seeking lube service is 3 minutes. The mean time required to perform the lube operation is 2 minutes. Both the interarrival times and the service times are exponentially distributed. The maximum number of vehicles in the system is 4. Determine the average waiting time, the average queue length, and the probability that a customer will have to drive on.

**Solution:**

The spreadsheet solution can be found by clicking here, or you can review the calculations in the table below:

\[ \lambda = \frac{60}{3} = 20 \text{ cars per hour} \]
\[ \mu = \frac{60}{2} = 30 \text{ cars per hour} \]

Probability that no customers are in the system:

\[ P_0 = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{M+1}} = \frac{1 - \frac{20}{30}}{1 - \left(\frac{20}{30}\right)^{5}} = 0.38 \]

Probability of 4 customers in the system and the next must drive on:

\[ P_n = \left(\frac{\lambda}{\mu}\right)^n \cdot P_0 \text{ for } n \leq M \]
\[ P_4 = \left(\frac{20}{30}\right)^4 (0.38) = 0.076 \]

Average number of customers in the system:

\[ L = \frac{\lambda/\mu}{1 - \frac{\lambda}{\mu}} \cdot \frac{(M + 1) \left(\frac{\lambda}{\mu}\right)^{M+1}}{1 - \left(\frac{\lambda}{\mu}\right)^{M+1}} = \frac{20/30}{1 - 20/30} - \left(\frac{20}{30}\right)^{5} = 1.24 \]
Average number of customers in waiting line:
\[ L_q = L - \frac{\lambda(1 - P_h)}{\mu} \]
\[ L_q = 1.24 - \frac{20(1 - 0.076)}{30} = 0.62 \]

Average time a customer spends in the system:
\[ W = \frac{L}{\lambda(1 - P_h)} \]
\[ W = \frac{1.24}{30(1 - 0.076)} = 0.067 \text{ hr.} \]
\[ = 4.03 \text{ min} \]

Average time a customer spends waiting in line to be served:
\[ W_q = W - \frac{1}{\mu} \]
\[ W_q = 4.03 - \frac{1}{30} = 0.033 \text{ hr} \]
\[ = 2.03 \text{ min} \]

The probability that a customer must drive on because the system is filled is 0.076, or about 7.6%. The average wait time in line is 2.03 minutes. In general, the system shows satisfactory performance.

**IN-LINE EXERCISE**

b. For the example just illustrated, what is the probability that three cars will be in the system?

**Answer**

**Finite Calling Populations**

In the single-server model with finite calling populations, the population of customers from which arrivals originate is limited, such as the number of police cars at a station answer calls.

- Arrivals originate from a finite (countable) population
- \( N = \text{population size} \)
- The operating characteristics for the single-server model with finite calling population can be computed with the formulas below.
Formulas For Finite Calling Population Single-Server Model

Probability that no customers are in the system:
\[ P_0 = \frac{1}{\sum_{n=0}^{N} \frac{N!}{(N-n)!} \left( \frac{\lambda}{\mu} \right)^n} \]

Probability of exactly \( n \) customers in the system:
\[ P_n = \frac{N!}{(N-n)!} \left( \frac{\lambda}{\mu} \right)^n P_0 \]

Average number of customers in the system:
\[ L = L_q + (1 - P_0) \]

Average number of customers in waiting line:
\[ L_q = N - \left( \frac{\lambda + \mu}{\lambda} \right) (1 - P_0) \]

Average time a customer spends in the system:
\[ W_q = \frac{L_q}{(N - L)\lambda} \]

Average time a customer spends waiting in line to be served:
\[ W = W_q - \frac{1}{\mu} \]

Finite Calling Population Example:

The finite calling population model can be used to evaluate maintenance system waiting times. A manufacturing company operates a job shop that has 20 machines. Each machine operates an average of 200 hours before breaking down, and the mean repair time is 3.6 hours. The breakdown rate is Poisson distributed and the service times are exponentially distributed. Is the staffing for the present repair system adequate?

Solution:

- \( N = 20 \) machines
• 20 machines which operate an average of 200 hrs before breaking down, \( \lambda = 1/200 \text{ hr} = 0.005/\text{hr} \n\)

• Mean repair time = 3.6 hrs, \( \mu = 1/3.6 \text{ hr} = 0.2778/\text{hr} \n\)

The equations for this model are too difficult to work manually. Click here to see the spreadsheet solution.

Summary of solution:

• \( P_0 = 0.652 \), the probability of idle
• \( L_q = 0.169 \) machines waiting
• \( W_q = 1.74 \) hours waiting for repair

The system seems to be performing reasonably well. Maintain staffing as is.

Multiple-Channel, Single-Phase Models

An example of a multiple-channel, single-phase model is an airline ticket and check-in counter. A single line of customers queues to multiple servers in parallel. Its assumptions are:

• Two or more independent servers serve a single waiting line

• Poisson arrivals, exponential service, infinite calling population

• The number of channels \( x \) the service rate must exceed the arrival rate, \( s\mu > \lambda \n\)

• The operating characteristics for the multiple-server model can be computed using the following equations:
Formulas for the Basic Multiple-Server Model

Probability that no customers are in the system:

\[ P_0 = \frac{1}{\left[ \sum_{n=0}^{s-1} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n \right] + \frac{1}{s!} \left( \frac{\lambda}{\mu} \right)^s \left( \frac{s\mu}{s\mu - \lambda} \right)} \]

Probability of exactly \( n \) customers in the system:

\[ P_n = \begin{cases} \frac{1}{s!} \left( \frac{\lambda}{\mu} \right)^n P_0, & \text{for } n > s \\ \frac{1}{\Gamma(n)} \left( \frac{\lambda}{\mu} \right)^n, & \text{for } n \leq s \end{cases} \]

Average number of customers in the system:

\[ L = \frac{\lambda \mu (\lambda / \mu)^s}{(s - 1)(s\mu - \lambda)^s} P_0 + \frac{\lambda}{\mu} \]

Average number of customers in waiting line:

\[ L_q = L - \frac{\lambda}{\mu} \]

Average time a customer spends in the system:

\[ W = \frac{L}{\lambda} \]

Average time a customer spends waiting in line to be served:

\[ W_q = W - \frac{\lambda}{\mu} \]

Utilization rate:

\[ \rho = \frac{\lambda}{s\mu} \]

As you can see, the formulas are quite complex. The formula for \( P_0 \) may be computed as shown in the spreadsheet example below, or it may be gotten directly by looking it up in Table 16.2. To determine the correct value of \( P_0 \) in Table 16.2, calculate \( \rho = \lambda / s\mu \), where \( s \) = the number of channels. Look down the first column of the table to find \( \rho \). Look across to find the correct number of channels. \( P_0 \) is the number at the intersection of the two.
### IN-LINE EXERCISE

c. Using Table 16.2, determine $P_0$ for a 4-channel system with an arrival rate, $\lambda = 9$ customers per hour, and service rate, $\mu = 3$ customers per hour.

### Answer
Multiple-Server Example

A customer service area in a department store has a single waiting line served by three customer representatives. Customers are served on a first-come, first-served basis as a representative becomes free. Customers arrive at an average of 10 customers per hour (Poisson-distributed), and an average of 4 customers can be served per hour by a customer service representative (Poisson distributed). What are the operating characteristics of the system?

Solution:

\[ \lambda = 10 \text{ customers/area} \]
\[ \mu = 4 \text{ customers/hour per service representative} \]
\[ s\mu = (3)(4) = 12 \]

Click here to see the spreadsheet solution.

Summary of operating characteristic results:

- \( P_0 = 0.045 \), the probability that no customers are in the system
- \( L = 6 \) customers in the service departments
- \( L_q = 3.5 \) customers waiting to be served
- \( W_q = 0.35 \) hour, or 21 minutes waiting in line
- \( P_w = 0.70 \), the probability that a customer must wait for service

The performance of the service system is very poor. The waiting time is too long, and too many customers must wait. Consider the following possibility to improve the system:

- Add a 4th server to improve service
- Recompute operating characteristics:

\[ P_0 = 0.073 \text{ probability of no customers} \]
\[ L = 3.0 \text{ customers} \]
\[ W = 0.30 \text{ hour, 18 minutes in service} \]
\[ L_q = 0.5 \text{ customers waiting} \]

\[ W_q = 0.05 \text{ hours, 3 minutes waiting, versus 21 earlier} \]

\[ P_w = 0.31 \text{ probability that a customer must wait} \]

Waiting time is significantly reduced to 3 minutes. The service department improves dramatically with a fourth service representative, but the store may consider adding another to do even better.

The examples in this chapter illustrate various formulas and methods of evaluating whether to expand service. Judgment must be used when making the calculations.

**INTERNET EXPLORATION**


2. Find more queue-related sites on the web by searching operations research sources.

**HOMEWORK**

1. Problem 16-2, pg. 774.
2. Problem 16-6, pg. 775.
3. Problem 16-12, pg. 776.
4. Problem 16-14, pg. 776.
5. Problem 16-24, pg. 778.

**HOMEWORK SOLUTIONS**

1. Problem 16-2, pg. 774.
2. Problem 16-6, pg. 775.
3. Problem 16-12, pg. 776.
4. Problem 16-14, pg. 776.
5. Problem 16-24, pg. 778.

**PRACTICE EXAM**