Operations Research II: Stochastic Models
Some Solved Examples: Queueing

Example 1: Telephone calls arrive at a ticket reservation office at a Poisson rate of 2 per minute. The time required for the single operator to take an order is uniformly distributed between 15 and 25 seconds. Incoming calls that arrive while the operator is busy are lost. In the long run, what fraction of the time does the operator spend taking orders?

We think of this as a regenerative process in which a “cycle” consists of an order being taken (during which arrivals are ignored), followed by a wait until the next arrival (call).

Thus, the expected length of a cycle is $E(Y + Z)$, where $Y$ is the Uniform(15,25) time (in seconds) it takes to take an order, and $Z$ is the exponential(1/30) time (in seconds) it takes to wait for the next call to arrive. (The intercall time is exponential with rate 2 per minute, which is 1/30 per second.) Thus, the expected cycle length is

$$E(Y + Z) = \frac{15 + 25}{2} + \frac{1}{1/30} = 50$$

The expected time per cycle taking orders is simply $E(Y) = 20$ seconds. Thus, the fraction of time taking orders is

$$\frac{E(\text{time per cycle taking orders})}{E(\text{length of cycle})} = \frac{E(Y)}{E(Y + Z)} = \frac{20}{50}$$

Example 2: Cars come in three sizes: Tercels (8 feet long), Mustangs (10 feet long), and stretched limos (20 feet long). Assume that 1/4 of all cars are Tercels, and 1/4 are limos. In a very bad traffic jam in New York city, a long stretch of cars sit bumper-to-bumper, extending across a draw-bridge. The draw-bridge opens, dumping into the Hudson the poor unfortunate car that sits on top of the crack joining the two sides of the bridge. What is the expected length of the car that gets trashed?

This is an example of the Renewal Paradox. The expected length, $X$, of a car is

$$\mu = E(X) = \left(\frac{1}{4}\right)8 + \left(\frac{1}{2}\right)10 + \left(\frac{1}{4}\right)20 = 12 \text{ feet},$$

the variance is $\sigma^2 = \text{var}(X) = \left(\frac{1}{4}\right)64 + \left(\frac{1}{2}\right)100 + \left(\frac{1}{4}\right)400 - 12^2 = 22 \text{ feet}^2$. We want the expected length of a renewal epoch covering $t$, for a large value of $t$, which is approximately:

$$\mu + \frac{\sigma^2}{\mu} = 12 + \frac{22}{12} = 83/6$$

feet.

Example 3: Consider an M/M/1 queue in which customers arrive at a rate of 20 per hour, and the average service time is 150 seconds. What is the average number of customers in the system, and how long, on average, does each spend in the system and in line? Answer the question also if the arrival rate increases by 10%.

We know that $\lambda = 20$ per hour and $\mu = 1/150$ per second, or $\mu = 2/5$ per minute, or $\mu = 24$ per hour. Let’s work in common units of hours$^{-1}$.

For the M/M/1 queue, we know the expected system size (average number of customers in the system) is

$$L = \frac{\lambda}{\mu - \lambda} = \frac{20}{24 - 20} = 5$$

From Little’s Law, then we get $5 = L = \lambda W = 20(\text{hr}^{-1}) \cdot W$, so the average time spent in the system is $W = 5/20 = 1/4$ hours (or 15 minutes).

The average time spent in line is $W_Q = W - \frac{1}{\mu} = (1/4) - (1/24) = 5/24$ hours (12.5 minutes).

If the arrival rate increases by 10% to 22 per hour, then $L = \frac{22}{24 - 22} = 11$ (the expected system size more than doubles), $W = L/\lambda = 11/22 = 1/2$ hour, and $W_Q = W - \frac{1}{\mu} = (1/2) - (1/24) = 11/24$ hours.

[For a similar example, see Example 8.2 of the text.]

Example 4: A tourist information desk has a single server, who answers questions for tourists, who stand in line to be served. Assume the time it takes to answer a tourist’s question(s) is exponentially distributed with mean of 3 minutes. Tourists arrive to the desk at a rate of 18 per hour. But each tourist is carrying a tourbook, and, while standing in line, looks in the book for the answer to their question: suppose it takes, on average, 4 minutes for a
tourist to find the answer in the book, and that the search time to do so is exponentially distributed. If a tourist finds the answer to his/her question while standing in line, he/she leaves the line.

Let \( X(t) \) be the number of tourists in the “system” (in line, or at the desk of the single server). Then, we claim that \( X(t) \) defines a CTMC on state space \( S = \{0, 1, 2, 3, \ldots \} \).

Think about what happens when we are in state \( i > 1 \): The next event is caused either (i) by an arrival (this happens at rate \( \lambda = 18 \) per hour), or (ii) by a service completion (this happens at rate \( \mu = 20 \) per hour, since the average service time is 3 minutes), or (iii) by a tourist leaving the line, upon answering their own question by reading the tourbook (this happens at total rate \( (i - 1) \cdot 15 \) per hour, since there are \( i - 1 \) tourists in line, each leaving at rate \( 1/(4 \text{ minutes}) = 15 \) per hour). Thus, we get the rate transition diagram below.

![Rate Transition Diagram](image)

The balance equations allow us to compute the \( \pi_i \)'s; setting flow out equals flow in, we get

\[
18\pi_0 = 20\pi_1 \\
(18 + 20)\pi_1 = 18\pi_0 + (20 + 15)\pi_2 \\
(18 + (20 + 15))\pi_2 = 18\pi_1 + (20 + 15 \cdot 2)\pi_3 \\
(18 + (20 + 15 \cdot 2))\pi_3 = 18\pi_2 + (20 + 15 \cdot 3)\pi_4 \\
\vdots \\
1 = \pi_0 + \pi_1 + \pi_2 + \cdots
\]

In principle, we can solve for the \( \pi_i \)'s, but we leave the answers below in terms of the \( \pi_i \)'s.

(a). What is the long-run fraction of time that the server is busy?

The server is busy whenever \( X(t) \geq 1 \) (i.e., \( X(t) \neq 0 \)). Thus, the fraction of time the server is busy is \( 1 - \pi_0 \).

(b). What fraction of tourists end up answering their own question, without ever reaching the server’s desk?

Tourists arrive at rate 18 per hour; they depart by leaving the line (after answering their own question) at rate \( 0 \cdot \pi_0 + 0 \cdot \pi_1 + 15 \cdot \pi_2 + 2 \cdot 15 \cdot \pi_3 + 3 \cdot 15 \cdot \pi_4 + \cdots \). Thus, the fraction that leave the line because they answer their own question is \( 15 \cdot (\pi_2 + 2\pi_3 + 3\pi_4 + \cdots) \)

(c). What fraction of tourists arrive to the desk and immediately get served (without waiting in a line)?

An arrival that comes when \( X(t) = 0 \) goes immediately to service. Thus, the fraction of arrivals that go immediately to service is \( \pi_0 \), the probability that the server is idle (the system is empty).

**Example 5:** Consider the queueing network shown below. Arrivals to each node are according to Poisson processes with rates \( \lambda_1 = 2 \), \( \lambda_2 = 3 \), and \( \lambda_3 = 4 \), all in units of hours\(^{-1}\). When leaving station 1, a customer next goes to station 3 with probability \( p = 1/2 \), and returns to station 1 with probability \( q = 1/3 \). At station 1 there is a single server with exponential service time of mean 5 minutes. At station 2 there are an infinite number of servers, each working with exponential rate 10 per hour. At station 3 there is a single server with exponential rate 10 per hour.

![Queueing Network Diagram](image)

(a). Assume that the current situation is that there are no customers at station 1, two customers at station 2, and three customers at station 3. What is the probability that the next change in state will occur at a time between 5 minutes from now and 10 minutes from now?
(b). What is the long-run probability that there are 2 customers at each station?
(c). What is the variance of the number of customers at station 1, in the long run?

**Example 6:** Consider the queueing network shown below, where each box ("node") represents a queueing station with infinite capacity (unlimited "chairs"). Assume that node 1 has 3 servers, each capable of completing a service in an exponential length of time with mean 1/4; node 2 has four servers, each working at exponential rate $\mu_2 = 2$; and node 3 has a single server working at exponential rate $\mu_3 = 4$. Assume that there are a total of $m = 3$ customers in the network.

(a). Define a state space for which this system can be modelled as a CTMC.
Let $X(t) = (i, j, k)$, the state at time $t$, indicate that, at time $t$, there are $i$ customers at node 1, $j$ customers at node 2, and $k$ customers at node 3.

The state space is then $S = \{(3, 0, 0), (2, 1, 0), (2, 0, 1), (1, 1, 1), (1, 0, 2), (0, 3, 0), (0, 2, 1), (0, 1, 2), (0, 0, 3)\}.

(b). Assume that the network is currently in the situation of having one customer at node 3 and two customers at node 2. What is the probability that there is no change in the situation in the next 5 time units?
The holding time, $T_{0,2,1}$, in state $(0,2,1)$ is exponential($\nu_{0,2,1}$), where $\nu_{0,2,1} = 2\mu_2 + \mu_3 = 8$ is the rate at which the chain leaves state $(0,2,1)$.
(c). What is the routing matrix $R$ for the queueing network?

$$R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

(d). Compute $\frac{\pi_{1,0,2}}{\pi_{0,3,0}}$.
First, we compute the $\gamma$ vector of effective net arrival rates, using: $\gamma = \gamma R$. These 3 equations in 3 unknowns readily give that $\gamma_1 = \gamma_2 = \gamma_3$. We can pick them to be any constant (since normalization is done later, in making the $\pi$’s sum to 1); thus, we let $\gamma = (1, 1, 1)$.

This gives us

$$\rho_1 = \frac{1}{4}, \quad \rho_2 = \frac{1}{2}, \quad \rho_3 = \frac{1}{4},$$

and allows us to compute the ratio

$$\frac{\pi_{1,0,2}}{\pi_{0,3,0}} = \frac{K \cdot \rho_1^1 \cdot \rho_2^0 \cdot \rho_3^3}{K \cdot 1 \cdot \rho_1^1 \cdot 1} = \frac{3}{4}$$

(e). In terms of the $\pi$’s, what is the average rate at which customers arrive at node 2?
The average rate at which customers arrive at node 2 is the rate of departure from node 1, which is

$$\mu_1(\pi_{1,2,0} + \pi_{1,1,1} + \pi_{1,0,2}) + 2\mu_1(\pi_{2,1,0} + \pi_{2,0,1}) + 3\mu_1\pi_{3,0,0},$$

where $\mu_1 = 4$ (since the mean service time at node 1 is 1/4).
(f). What is the variance of the number of customers at node 2 (in the long run)? (You may leave your answer in terms of the $\pi$’s.)
Let $X_2(\infty)$ denote the number of customers at node 2 in the long run. Let $p_i = P(X_2(\infty) = i)$, for $i = 0, 1, 2, 3$. Then, we want to compute

$$\text{var}(X_2(\infty)) = 0^2 p_0 + 1^2 p_1 + 2^2 p_2 + 3^2 p_3 - (0p_0 + 1p_1 + 2p_2 + 3p_3)^2,$$

where $p_1 = \pi_{2,1,0} + \pi_{1,1,1} + \pi_{0,1,2}$, $p_2 = \pi_{1,2,0} + \pi_{0,2,1}$, and $p_3 = \pi_{0,3,0}$. (Simply enumerate all cases in which there can be $i$ customers at node 2.)

**Example 7:** [Problem 1, Chapter 8, page 534.] For an M/M/1 queue, compute (a). the expected number of arrivals during a service period, and (b). the probability that no customers arrive during a service period.
Let $S$ be a service period; we know that $S$ is exponential($\mu$). Let $Y$ be the number of arrivals that occur during a service period of length $S$.

(a). We want to compute $E(Y)$. If we knew what the length of $S$ is, it would be easy: Given that $S = x$, the expected number of arrivals is simply $\lambda x$, since the arrivals form a Poisson process of rate $\lambda$.

Thus, we condition on the value of $S$:

$$E(Y) = E[E(Y|S)] = \int_0^\infty E(Y|S = x)\mu e^{-\mu x} dx = \int_0^\infty \lambda x \mu e^{-\mu x} dx = \frac{\lambda}{\mu}$$

(b). We want to compute $P(Y = 0)$. Again, we condition on $S$:

$$P(Y = 0) = \int_0^\infty P(Y = 0|S = x)\mu e^{-\mu x} dx = \int_0^\infty e^{-\lambda x} \frac{\lambda^0}{0!} e^{-\mu x} dx = \frac{\mu}{\lambda + \mu} \int_0^\infty (\lambda + \mu) e^{-(\lambda + \mu)x} dx = \frac{\mu}{\lambda + \mu}$$

**Example 8:** You are considering renting some data processing equipment. Model 1 has service rate $\mu_1 = 100$ customers/hour and rents for $c_1 = $1000/month. Model 2 has service rate $\mu_2 = 200$ customers/hour and rents for $c_2 = $1800/month. The arrival rate of customers (orders to be processed) is $\lambda = 80$ customers per hour. The waiting time cost is judged to be $1.00/customer-hour. Assume Poisson arrivals during normal working hours (about 200 hours/month), exponential service, and that the equipment operates only during these hours.

(a). Assume you rent Model 1. What is the probability that an order requires between 5 and 6 minutes to be processed?

For model 1, $\mu_1 = 100$ hr$^{-1}$, $\lambda = 80$ hr$^{-1}$. This is an M/M/1 queue, so the total waiting time in the system of a customer is a random variable $W$ with an exponential distribution with parameter $\mu_2 - \lambda = 20$ hr$^{-1}$.

We want to compute

$$P(1/12 \leq W \leq 1/10) = \int_{1/12}^{1/10} 20e^{-20x} dx = e^{-5/3} - e^{-2}$$

(b). Which model should you rent? (Ignore carryover effects, e.g., over a weekend.) You need to compute the expected cost per hour for each model, and then compare.

The rental cost per hour for Model $i$ is $c_i/200$. The waiting cost per hour is $1$ per customer-hour, and a typical customer waits in the system $L_i = \lambda/\mu_i$ time. Thus, the total expected cost per hour is $L_i + (c_i/200) = 4 + 5 = 9$ dollars per hour for Model 1, and $L_2 + (c_2/200) = 2/3 + 9 = 9.67$ dollars per hour for Model 2. Thus, we should rent Model 1.