Administración de las Operaciones

Cualquiera que haya tenido que esperar frente a un semáforo, en la cola de un banco o de un restaurante de comidas rápidas, ha vivido la dinámica de las filas de espera.

MODELOS DE LÍNEAS DE ESPERA

El análisis de líneas de espera es de interés para los gerentes porque afecta el diseño, la planificación de la capacidad, la planificación de la distribución de espacios, la administración de inventarios y la programación. Aquí discutiremos por qué se forman colas y filas, las aplicaciones de los métodos de la administración de las operaciones y la estructura de los modelos de filas de espera. Veremos también cómo abordan los gerentes sus decisiones con esos modelos.

Roberto CARRO PAZ
Daniel GONZÁLEZ GÓMEZ

16

Facultad de Ciencias Económicas y Sociales

Universidad Nacional de Mar del Plata

El Sistema de
Producción y Operaciones

CRÉDITOS FOTOGRÁFICOS:
La totalidad de las fotografías incluidas en este trabajo han sido tomadas por los autores.

MODELOS DE LÍNEAS DE ESPERA

Cualquiera que haya tenido que esperar frente a un semáforo, en la cola de un banco o de un restaurante de comidas rápidas, ha vivido la dinámica de las filas de espera. El análisis de líneas de espera es de interés para los gerentes porque afecta el diseño, la planificación de la capacidad, la planificación de la distribución de espacios, la administración de inventarios y la programación.

Junto a los árboles de decisiones, con frecuencia los modelos de líneas de espera son útiles para la planificación de la capacidad. Frente a ciertos centros de trabajo, como el mostrador de pasajes de un aeropuerto, un centro de máquinas o un centro de cómputos central, tienden a formarse líneas de espera. Es así porque los tiempos de llegada entre dos trabajos o clientes sucesivos varían y el tiempo de procesamiento también varía de un consumidor al siguiente. Los modelos de líneas de espera usan distribuciones de probabilidad para ofrecer estimaciones del tiempo de retraso promedio de los clientes, la longitud promedio de las filas de espera y la utilización del centro de trabajo. Los gerentes suelen usar esta información para elegir la capacidad más efectiva en términos de costos, hallando un equilibrio entre el servicio al cliente y el costo de la capacidad agregada.

Se conoce como **línea de espera** a una hilera formada por uno o varios **clientes** que aguardan para recibir un servicio. Los clientes pueden ser personas, objetos, máquinas que requieren mantenimiento, contenedores con mercancías en espera de ser embarcados o elementos de inventario a punto de ser utilizados. Las líneas de espera se forman a causa de un desequilibrio temporal ente la demanda de un servicio y la capacidad del sistema para suministrarlo.

Una posibilidad de 0,0625 significa que la posibilidad de tener más de tres clientes en una fila de registro en un aeropuerto, en un cierto momento del día, es una probabilidad de uno en 16. Si este check-in de American Airlines en el aeropuerto de Bruselas puede vivir con cuatro o más pasajeros en línea aproximadamente el 6% del tiempo, un agente de servicio será suficiente. Si no, se deben sumar más puestos de registro y personal.

En la mayoría de los problemas de líneas de espera que se presentan en la vida real, la tasa de demanda varía; es decir, los clientes llegan a intervalos imprevisibles. Lo más común es que también haya variaciones en el ritmo de producción del servicio, dependiendo de las necesidades del cliente.

Los pacientes que aguardan al médico en su consultorio y los taladros descompuestos que esperan en una instalación de reparación tienen mucho en común desde una perspectiva de Producción y Operaciones. Ambas utilizan recursos humanos y recursos de equipos para mantener los valiosos activos de producción (gente y máquinas) en buenas condiciones.

Los administradores de operaciones reconocen el trueque que se lleva a cabo entre el costo de ofrecer un buen servicio y el costo del tiempo de espera del cliente o la máquina. Los administradores desean que las filas de espera sean lo suficientemente cortas, de tal forma que los clientes no se sientan descontentos y se vayan sin comprar, o que compren pero nunca regresen. Sin embargo, los administradores están dispuestos a permitir alguna espera, si ésta es proporcional a un ahorro significativo en los costos del servicio. Cuando la empresa intenta elevar su nivel de servicio, se observa un incremento en los costos.

USOS DE LA TEORÍA DE LÍNEAS DE ESPERA

La teoría de las líneas de espera es aplicable a empresas de servicios o manufactureras, porque relaciona la llegada de los clientes y las características de procesamiento del sistema de servicios con las características de salida de dicho sistema. El sistema de servicio puede consistir en la operación de cortar el cabello en una peluquería, o bien, en el departamento de partes, con una máquina determinada para atender un pedido de producción. Otros ejemplos de clientes y servicios son las filas de los espectadores que esperan frente a un estadio de fútbol para comprar entradas, los camiones que aguardan para ser descargados en una planta de acopio de cereales, las máquinas en espera de ser reparadas por una cuadrilla de mantenimiento y los pacientes que hacen antesala para ser atendidos por un médico. Cualquiera que sea la situación, los problemas referentes a líneas de espera tienen algunos elementos en común.

ESTRUCTURA DE LOS PROBLEMAS DE LÍNEAS DE ESPERA

El análisis de los problemas de líneas de espera comienza con una descripción de los elementos básicos de la situación. Cada situación específica tendrá características diferentes, pero cuatro elementos son comunes a todas ellas:

1. Un insumo, o población de clientes, que genera clientes potenciales.
2. Una línea o fila de espera formada por los clientes.
3. La instalación de servicio, constituida por una persona (o una cuadrilla), una máquina (o grupo de máquinas) o ambas cosas si así se requiere para proveer el servicio que el cliente solicita.
4. Una regla de prioridad para seleccionar al siguiente cliente que será atendido por la instalación de servicio.

La figura 16.1 ilustra estos elementos básicos. El sistema de servicio describe el número de filas y la disposición de las instalaciones. Una vez que el servicio ha sido suministrado, los clientes atendidos salen del sistema.

Población de clientes

La fuente de insumos para el sistema de servicio es una población de clientes. Si el número potencial de nuevos clientes para el sistema de servicio resulta afectado notablemente por el número de clientes que ya se encuentran en el sistema, se dice que esa fuente de insumos es finita. Por ejemplo, supongamos que una cuadrilla de mantenimiento se le asigna la responsabilidad de reparar 10 máquinas y dejarlas en buen estado de funcionamiento. Esa población generará los clientes para la cuadrilla de mantenimiento, de acuerdo con una función matemática de las tasas de falla de las máquinas. A medida que un mayor número de máquinas falle y entre en el sistema de servicio, ya sea para esperar su turno o para ser reparada de inmediato, la población de clientes disminuirá y, por consiguiente, se registrará un descenso de la tasa a la cual dicha población es capaz de generar otro cliente. En consecuencia, se dice que la población de clientes es finita.

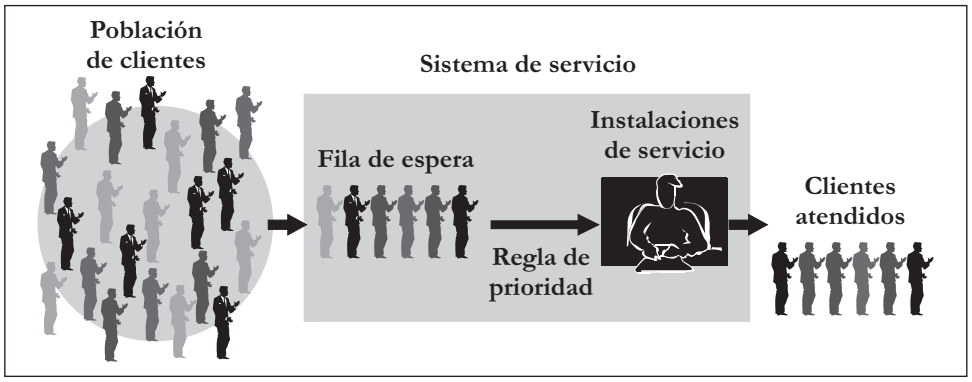

Figura 16.1
Elementos básicos de los modelos de líneas de espera.

En forma alternativa, la **población de clientes infinita** es aquella en la que el número de clientes que entran al sistema no afecta la tasa a la cual dicha población genera nuevos clientes. Por ejemplo, considere una operación de pedidos por Internet para la cual la población de clientes está constituida por los compradores que han recibido un catálogo de los productos que vende la compañía vía lista de distribución por correo electrónico. En vista de que la población de clientes es muy grande y sólo una pequeña fracción de los compradores hace pedidos en un momento determinado, el número de nuevos pedidos que genera no resulta afectado en forma notable por el número de pedidos que están en espera de servicio o que son atendidos por el sistema que imparte dicho servicio. En este caso se dice que la población de clientes es finita.

Los clientes de las líneas de espera pueden ser **pacientes o impacientes**, lo cual nada tiene que ver con el lenguaje que un cliente que espera largo tiempo en una fila, durante un día caluroso, podría utilizar. En el contexto de los problemas de líneas de espera, un cliente paciente es el que entra al sistema y permanece allí hasta ser atendido; un cliente impaciente es el que o bien decide no entrar al sistema (arrepentido) o sale de éste antes de haber sido atendido (desertor). Para simplificar los métodos, en este suplemento utilizaremos como supuesto que todos los clientes son pacientes.

Sistema de servicio

El sistema de servicio suele describirse en términos del número de filas y la disposición de las instalaciones.

Número de filas. Las filas de espera se diseñan en forma de **una sola fila** o **filas múltiples**. La figura 16.2 muestra un ejemplo de cada una de esas disposiciones. En general, se utiliza una sola fila en mostradores de aerolíneas, cajas de los bancos y algunos restaurantes de comida rápida, mientras que las filas múltiples son comunes en los supermercados y espectáculos públicos como teatros o canchas de fútbol. Cuando se dispone de servidores múltiples y cada uno de ellos puede manejar transacciones de tipo general, la disposición de una sola fila mantiene a todos ellos uniformemente ocupados y proyecta en los clientes una sensación de que la situación es equitativa. Estos piensan que serán atendidos de acuerdo con su orden de llegada, no por el grado en que hayan podido adivinar los diferentes tiempos de espera al formarse en una fila en particular. El diseño de filas múltiples es preferible cuando algunos de los servidores proveen un conjunto de servicios limitado. En esta disposición, los clientes eligen los servicios que necesitan y esperan en la fila donde se suministra dicho servicio, como sucede en los supermercados es las que hay filas especiales para los clientes que pagan en efectivo o para los que compran menos de 10 artículos.

Algunas veces, los elementos que esperan su turno no están organizados nítidamente en filas. Las máquinas que necesitan ser reparadas en el taller de producción de una fábrica pueden permanecer en sus respectivos sitios y el equipo de mantenimiento es el que tiene que acudir a cada lugar. A pesar de todo, podemos considerar que esas máquinas forman una sola fila o filas múltiples, según el número de cuadrillas de reparación y sus respectivas especialidades. Asimismo, los usuarios que llaman por teléfono para pedir un taxi también forman una fila, aunque cada uno se encuentre en un lugar diferente.

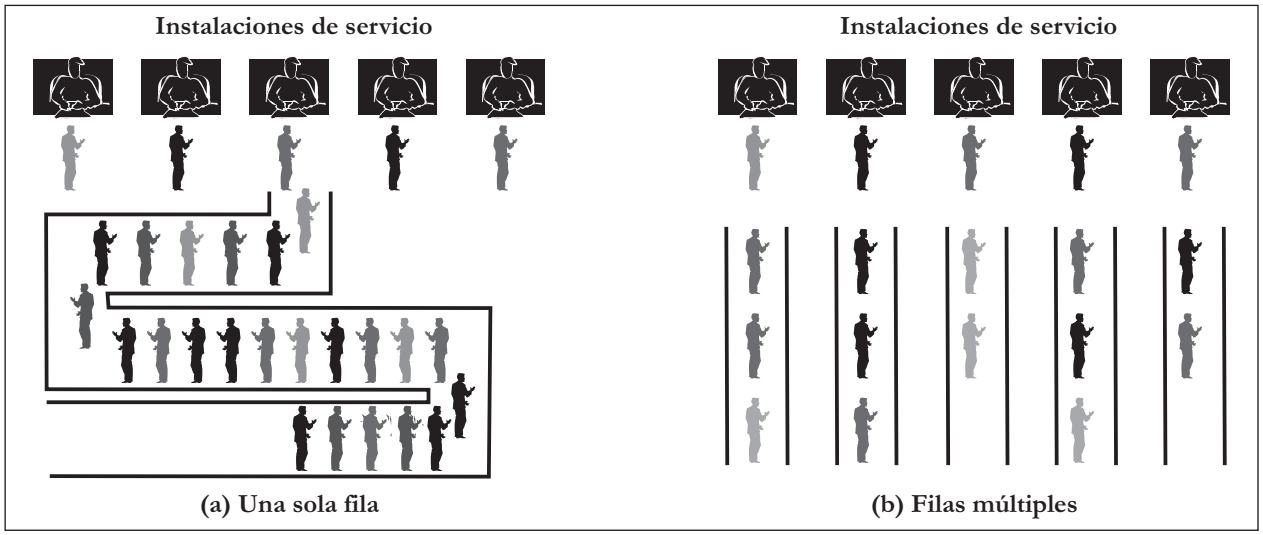

Figura 16.2
Diversas disposiciones de filas de espera

Disposición de instalaciones de servicio. Las instalaciones de servicio consisten en el personal y/o el equipo necesario para proporcionar dicho servicio al cliente. La figura 16.3 muestra algunos ejemplos de los cinco tipos básicos de disposiciones para las instalaciones de servicio. Los gerentes deben elegir una disposición adecuada según el volumen de sus clientes y el carácter de los servicios ofrecidos. Algunos servicios requieren un solo paso, también conocido como fase, en tanto que otros requieren una secuencia de pasos.

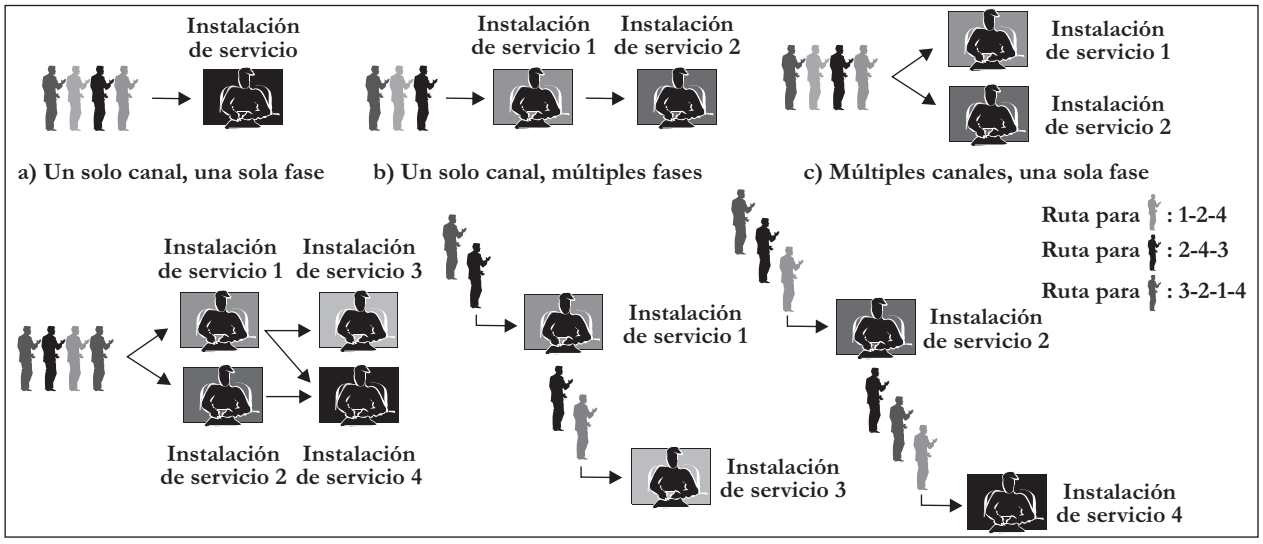

Figura 16.3
Ejemplos de disposiciones para instalaciones de servicio

En el sistema de **un solo canal y una sola fase**, todos los servicios solicitados por un cliente suelen impartirse por una instalación con un solo servidor. En ese caso, los clientes forman una sola fila y circulan uno por uno a través de la instalación de servicio. Ejemplos de esto son los servicios de lavado automático donde los automovilistas no necesitan salir de sus vehículos, o bien, cualquier máquina en la cual sea necesario procesar varios lotes de partes.

La disposición con **un solo canal y múltiples fases** se usa cuando es más conveniente que los servicios se impartan en secuencia por varias instalaciones, pero el volumen de la clientela y otras restricciones limitan el diseño a un solo canal. Los clientes forman una sola fila y avanzan en forma secuencial, pasando de una instalación de servicio a la siguiente. Un ejemplo de esta disposición son las instalaciones donde se realizan los trámites para la obtención de documentos personales como los pasaportes, donde la primera instalación de toma el formulario de inscripción, la segunda toma la huellas digitales y la tercera saca la foto del solicitante.

La disposición de **múltiples canales y una sola fase** se usa cuando la demanda es suficientemente grande para justificar que se suministre el mismo servicio en más de una instalación, o bien, cuando los servicios ofrecidos por las instalaciones son diferentes. Los clientes forman una o varias filas, dependiendo del diseño. En el diseño de una sola fila, los clientes son atendidos por el primer servidor disponible, como sucede en las oficinas del ANSES o en los bancos. Si cada canal tiene su propia fila de espera, los clientes aguardan hasta que el servidor de su respectiva fila pueda atenderlos.

La disposición de **múltiples canales y múltiples fases** se presenta cuando los clientes pueden ser atendidos por una de las instalaciones de la primera fase, pero después requieren los servicios de una instalación de la segunda fase, y así sucesivamente. En algunos casos, los clientes no pueden cambiar de canales después de iniciado el servicio; en otros sí pueden hacerlo. Un ejemplo de esta disposición son los negocios de servicios de lavado como Laverap. Las máquinas lavadoras son las instalaciones de la primera fase y las secadoras son las instalaciones de la segunda fase. Algunas lavadoras y secadoras están diseñadas para recibir cargas de mayor volumen, con lo cual es posible brindar al cliente la posibilidad de elegir entre varios canales.

En el problema más complejo de filas de espera intervienen clientes cuyos servicios requeridos tienen secuencias únicas; por consiguiente, el servicio no puede dividirse claramente en distintas fases. En esos casos se utiliza una **disposición mixta.** En esta disposición, las filas de espera suelen formarse frente a cada instalación, como en un taller de producción intermitente donde cada trabajo personalizado tal vez requiera el uso de diversas máquinas y diferentes rutas.

Regla de prioridad

La regla de prioridad determina a qué cliente se deberá atender a continuación. En la mayoría de los sistemas de servicio que conocemos, se aplica la regla de "a quien llega primero, se atiende primero" (FCFS; del inglés, *first-come, first-served*). El cliente que está en primer lugar en la fila de espera tiene la más alta prioridad, y el que llega al final tiene la prioridad más baja. En otras disciplinas para determinar órdenes de prioridad, se concede la prioridad al cliente que tenga la fecha prometida de vencimiento más próxima (EDD; del inglés, *earliest due date*) o al que corresponda el tiempo de procesamiento más corto (SPT; del ingles, *shortest processing time*).

Una disciplina prioritaria consiste en una regla que permite a un cliente de más alta prioridad interrumpir el servicio de otro cliente. Por ejemplo, en la sala de emergencias de un hospital, los pacientes que llegan con heridas que representan amenazas más graves para la vida son atendidos primero, sin importar en qué orden hayan llegado. La construcción de modelos de sistemas con disciplinas de prioridad complejas se realiza comúnmente por medio de una simulación por computadora.

Años de escuchar las quejas de los clientes han enseñado a las aerolíneas algunas lecciones de recolección de equipaje. Cuando Aeropuertos Argentina 2000 diseñó su área de recolección de equipaje en el Aeroparque Jorge Newbery de la ciudad de Buenos Aires, la puso cerca de las puertas, de tal forma que los pasajeros que desembarcaran no tuvieron que hacer una caminata demasiado larga. Pero aunque los pasajeros llegan al área rápidamente, deben esperar por su equipaje. En el aeropuerto internacional de Ezeiza, en cambio, los pasajeros tienen que caminar cierta distancia hasta el área de recolección, pero cuando llegan, sus maletas generalmente están ahí. Aún cuando los viajeros internacionales duran más tiempo total recogiendo su equipaje, la empresa ha encontrado que no se quejan tanto por la demora del equipaje de la forma en que lo hacen los pasajeros de cabotaje.

DISTRIBUCIONES DE PROBABILIDAD

Las fuentes de variación en los problemas de filas de espera provienen del carácter aleatorio de la llegada de los clientes y de las variaciones que se registran en los distintos tiempos de servicio. Cada una de esas fuentes suele describirse mediante una distribución de probabilidades.

Distribución de llegadas

La llegada de clientes a las instalaciones de servicio es aleatoria. La variabilidad en los intervalos de llegada de los clientes a menudo se describe por medio de una curva de distribución de Poisson, la cual especifica la probabilidad de que n clientes lleguen en T periodos de tiempo:

$$P_{(n)} = \frac{(\lambda T)^n}{n!} e^{-\lambda T} \qquad \text{para } n = 0, 1, 2, \ldots$$

donde: $P_{(n)}$ = probabilidad de n llegadas en T periodos de tiempo
λ = número promedio de llegadas de clientes por periodo
e = 2,7183

La medida de distribución de Poisson es λT, y la varianza también es λT. La distribución de Poisson es una distribución discreta; es decir, las probabilidades corresponden a un número específico de llegadas por unidad de tiempo.

Ejemplo del cálculo de la probabilidad de llegadas de clientes. Los clientes se presentan en la sección de atención de quejas por productos defectuosos de una casa de electrodomésticos a razón de dos clientes por hora. ¿Cuál es la probabilidad de que se presenten cuatro clientes durante la próxima hora?

Solución. En este caso $\lambda = 2$ clientes por hora, $T = 1$ hora y $n = 4$ clientes. La probabilidad de que lleguen cuatro clientes en la próxima hora es:

$$P_{(4)} = \frac{[2\,(1)]^4}{4!}\, e^{-2(1)} = \frac{16}{24}\, e^{-2} = 0{,}090$$

Otra forma de especificar la distribución de llegadas consiste en hacerlo en términos de tiempos entre **llegadas de clientes**; es decir, el intervalo de tiempo entre la llegada de dos clientes sucesivos. Si la población de clientes genera a éstos de acuerdo con una distribución de Poisson, entonces la **distribución exponencial** describe la probabilidad de que el próximo cliente llegue durante los siguientes T periodos de tiempo.

Distribución de tiempo de servicio

La distribución exponencial describe la probabilidad de que el tiempo de servicio del cliente en una instalación determinada no sea mayor que T periodos de tiempo. La probabilidad: puede calcularse con la siguiente fórmula:

$$P_{(t \leqslant T)} = 1 - e^{-\mu T}$$

donde: μ = número medio de clientes que completan el servicio en cada periodo
t = tiempo de servicio del cliente
T = tiempo de servicio propuesto como objetivo

La media de la distribución del tiempo de servicio es $1/\mu$, y la varianza es $(1/\mu)^2$. A medida que T incrementa, la probabilidad de que el tiempo de servicio del cliente sea menor que T se va aproximando a 1,0.

Ejemplo del cálculo de la probabilidad del tiempo de servicio. El empleado de la sección de atención de quejas por productos defectuosos puede atender, en promedio, a tres clientes por hora. ¿Cuál es la probabilidad de que un cliente requiera menos de 10 minutos de ese servicio?

Solución. Es necesario expresar todos los datos en las mismas unidades de tiempo. Puesto que $\mu = 3$ clientes por hora, convertimos los minutos en horas, o sea, $T = 10$ minutos $= 10/60$ hora $= 0{,}167$ hora. Entonces:

$$P_{(t \leqslant T)} = 1 - e^{-\mu T}$$

$$P_{(t \leqslant 0{,}167\,h)} = 1 - e^{-3(0{,}167)} = 1 - 0{,}61 = 0{,}39$$

Algunas características de la distribución exponencial no siempre se adaptan a una situación real. El modelo de distribución exponencial se basa en la suposición de que cada tiempo de servicio es independiente de los tiempos que lo precedieron. Sin embargo, en la vida real, la productividad puede mejorar a medida que los servidores humanos aprenden a hacer mejor su trabajo.

La suposición fundamental en este modelo es que los tiempos de servicio muy pequeños, igual que los muy grandes, son posibles. No obstante, las situaciones de la vida real requieren a menudo un tiempo de duración fija para su puesta en marcha, algún límite para la duración total del servicio o un tiempo de servicio casi constante.

USO DE MODELOS DE FILAS DE ESPERA PARA ANALIZAR OPERACIONES

Los gerentes de operaciones suelen utilizar modelos de filas de espera para establecer el equilibrio entre las ventajas que podrían obtener incrementando la eficiencia del sistema de servicio y los costos que esto implica. Además, los gerentes deberían considerar los costos por no hacer mejoras al sistema: las largas filas de espera o los prolongados tiempos de espera resultantes de esto provocan que los clientes se arrepientan o deserten. Por lo tanto, es preciso que los gerentes estén interesados en las siguientes características de operación del sistema:

1. **Longitud de la fila.** El número de clientes que forman una fila de espera refleja alguna de estas dos condiciones: las hileras cortas significan que el servicio al cliente es bueno o que la capacidad es excesiva, y las hileras largas indican una baja eficiencia del servidor o la necesidad de aumentar la capacidad.
2. **Número de clientes en el sistema.** El número de clientes que conforman la fila y reciben servicio también se relaciona con la eficiencia y la capacidad de dicho servicio. Un gran número de clientes en el sistema provoca congestionamientos y puede dar lugar a la insatisfacción del cliente, a menos que el servicio incremente su capacidad.
3. **Tiempo de espera en la fila**. Las filas largas no siempre significan tiempos de espera prolongados. Si la tasa de servicio es rápida, una fila larga puede ser atendida eficientemente. Sin embargo, cuando el tiempo de espera parece largo, los clientes tienen la impresión de que la calidad del servicio es deficiente. Los gerentes tratan de cambiar la tasa de llegada de los clientes o de diseñar el sistema para que los largos tiempos de espera parezcan más cortos de lo que realmente son.
4. **Tiempo total en el sistema**. El tiempo total transcurrido desde la entrada al sistema hasta la salida del mismo ofrece indicios sobre problemas con los clientes, eficiencia del servidor o capacidad. Si algunos clientes pasan demasiado tiempo en el sistema del servicio, tal vez sea necesario cambiar la disciplina en materia de prioridades, incrementar la productividad o ajustar de algún modo la capacidad.
5. **Utilización de las instalaciones de servicio**. La utilización colectiva de instalaciones de servicio refleja el porcentaje de tiempo que éstas permanecen ocupadas. El objetivo de la gerencia es mantener altos niveles de utilización y rentabilidad, sin afectar adversamente las demás características de operación.

El mejor método para analizar un problema de filas de espera consiste en relacionar las cinco características de operación y sus respectivas alternativas con su valor monetario. Sin embargo, es difícil asignar un valor económico a ciertas características (como el tiempo de espera de un cliente en un banco). En estos casos, es necesario que un analista compare el costo necesario para aplicar la alternativa en cuestión, frente a una evaluación subjetiva del costo que implicaría el hecho de no hacer dicho cambio.

Presentaremos ahora tres modelos y algunos ejemplos que ilustran la forma en que los modelos de filas de espera ayudan a los gerentes de operaciones en la toma de decisiones. Analizaremos problemas que requieren la utilización de los modelos de un solo servidor, de múltiples servidores y de fuente finita, todos ellos con una sola fase.

Modelo A: de un solo servidor

El modelo de filas de espera más sencillo corresponde a un solo servidor y una sola fila de clientes. Para especificar con más detalle el modelo, haremos las siguientes suposiciones:

1. La población de clientes es infinita y todos los clientes son pacientes.
2. Los clientes llegan de acuerdo con una distribución de Poisson y con una tasa media de llegadas de λ.
3. La distribución del servicio es exponencial, con una tasa media de servicio de μ.
4. A los clientes que llegan primero se les atiende primero.
5. La longitud de la fila de espera es ilimitada.

A partir de ellas, podemos aplicar varias fórmulas para describir las características de operación del sistema:

p = utilización promedio del sistema = $\frac{\lambda}{\mu}$

$P_{(n)}$ = probabilidad de que n clientes estén en el sistema = $(1 - p)\, p^n$

L = número promedio de clientes en el sistema de servicio = $\frac{\lambda}{\mu - \lambda}$

L_q = número promedio de clientes en la fila de espera = pL

W = tiempo promedio transcurrido en el sistema, incluido el servicio = $\frac{1}{\mu - \lambda}$

W_q = tiempo promedio de espera en la fila = pW

Ejemplo del cálculo de las características de operación de un sistema con un solo canal y una sola fase
La gerente de un supermercado está interesada en brindar un buen servicio a las personas de mayor edad que compran en su local. Actualmente, el supermercado cuenta con una caja de salida reservada para los jubilados. Estas personas llegan a la caja a un ritmo promedio de 30 por hora, de acuerdo con una distribución de Poisson, y son atendidos a una tasa promedio de 35 clientes por hora, con tiempos de servicio exponenciales. Calcule los siguientes promedios:

a. Utilización del empleado de la caja de salida.
b. Número de clientes que entran al sistema.
c. Número de clientes formados en la fila.
d. Tiempo transcurrido dentro del sistema.
e. Tiempo de espera en la fila.

Solución. La caja de salida puede representarse como un sistema con un solo canal y una sola fase. Usamos las ecuaciones correspondientes a las características de operación del modelo con un solo servidor para calcular las características promedio:

a. La utilización promedio del empleado de la caja de salida es:

$$p = \frac{\lambda}{\mu} = \frac{30}{35} = 0{,}857 \quad ; \quad \text{o sea, } 85{,}7\%$$

b. El número promedio de clientes que entran al sistema es:

$$L = \frac{\lambda}{\mu - \lambda} = \frac{30}{35 - 30} = 6 \text{ clientes}$$

c. El número promedio de clientes formados en la fila es:

$$L_q = \text{pL} = 0{,}857\ (6) = 5{,}14 \text{ clientes}$$

d. El tiempo promedio transcurrido dentro del sistema es:

$$W = \frac{1}{\mu - \lambda} = \frac{1}{35 - 30} = 0{,}20 \text{ hora} \quad ; \quad \text{o sea, } 12 \text{ minutos}$$

e. El tiempo promedio transcurrido en la fila es:

$$W_q = pW = 0{,}857\ (0{,}20) = 0{,}17 \text{ hora} \quad : \quad \text{o sea, } 10{,}28 \text{ mimnutos}$$

Ejemplo del análisis de las tasas de servicio usando el modelo con un solo servidor La gerente del supermercado mencionado, desea respuestas de las siguientes preguntas:

a. ¿Qué tasa de servicio se requerirá para lograr que los clientes pasaran, en promedio, sólo 8 minutos en el sistema?
b. Con esa tasa de servicio, ¿cuál sería la probabilidad de tener más de cuatro clientes en el sistema?
c. ¿Qué tasa de servicio se requeriría para que fuera de sólo 10% la probabilidad de tener más de cuatro clientes en el sistema?

Solución.

a. Usamos la ecuación correspondiente al tiempo promedio dentro del sistema y resolvemos para μ.

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

$$8 \text{ minutos} = 0{,}133 \text{ horas} = \frac{1}{\mu - 30}$$

$$0{,}133\,\mu - 0{,}133\,(30) = 1$$

$$\mu = 37{,}52 \text{ clientes/hora}$$

b. La probabilidad de que haya más de cuatro clientes en el sistema es igual a 1 menos la probabilidad de que haya cuatro o menos clientes en el sistema.

$$p = 1 - \sum_{n=0}^{4} P_{(n)}$$

$$= 1 - \sum_{n=0}^{4} (1 - p)\,p$$

$$\text{y} \quad p = \frac{30}{37{,}52} = 0{,}80 \quad ; \quad \text{entonces,}$$

$$p = 1 - 0{,}2\,(1 + 0{,}8 + 0{,}8^2 + 0{,}8^3 + 0{,}8^4)$$

$$p = 1 - 0{,}672 = 0{,}328$$

Por lo tanto, existe una probabilidad de casi 33% de que más de cuatro clientes estén en el sistema.

c. Aplicamos la misma lógica que en la parte (b), excepto que μ es ahora una variable de decisión. La forma más fácil de proceder es encontrar primero la utilización promedio correcta y después resolver para la tasa de servicio.

$$p = 1 - (1 - p)\,(1 + p + p^2 + p^3 + p^4)$$

$$= 1 - (1 + p + p^2 + p^3 + p^4) + p\,(1 + p + p^2 + p^3 + p^4)$$

$$= 1 - 1 - p - p^2 - p^3 - p^4 + p + p + p^2 + p^3 + p^4 + p^5$$

$$= p^5$$

$$\text{o bien;} \quad = p^{1/5}$$

$$\text{Si P} = 0{,}10 \quad ; \quad \text{entonces,}$$

$$p = (0{,}10)^{1/5} = 0{,}63$$

En consecuencia, para una tasa de utilización de 63%, la probabilidad de que más de cuatro clientes se encuentren en el sistema es de 10%. Para $\lambda = 30$, la tasa de servicio media deberá ser de:

$$\frac{30}{\mu} = 0{,}63$$

$$\mu = 47{,}62 \text{ clientes/hora}$$

La gerente tiene que encontrar ahora la forma de incrementar la tasa de servicio, de 36 por hora a 48 por hora aproximadamente. Dicha tasa de servicio puede incrementarse en varias formas, que abarcan desde emplear a un trabajador que ayude a empacar la mercancía, hasta instalar equipo electrónico más moderno y veloz en la caja para que lea en menos tiempo los precios de la información impresa en código de barras sobre cada artículo.

El Epcot Center de la fotografía, al igual que Disney World en Orlando, Disneyland en California, EuroDisney cerca de París y Disney Japan cerca de Tokio tienen una característica en común: las largas filas y las esperas que parecen no tener fin. Pero Disney es una de las compañías líderes en el estudio científico de la teoría de colas; analiza los comportamientos de las colas y puede predecir qué juegos atraerán a qué cantidades de multitud. Para mantener contentos a los visitantes, Disney hace tres cosas: (1) hace que las líneas parezcan que avanzan en forma constante; (2) entretiene a la gente mientras espera; y (3) pone señales diciendo a los visitantes a cuántos minutos de lejanía están de cada juego. De esa manera, los padres pueden decidir si una espera de 20 minutos para el Samll World vale más la pena que una espera de 30 minutos para Mr. Frog`s Wild Ride.

Modelo B: de múltiples servidores

En el modelo con múltiples servidores, los clientes forman una sola fila y escogen, entre s servidores, aquel que esté disponible. El sistema de servicio tiene una sola fase. Partiremos de las siguientes suposiciones, además de las que hicimos para el modelo con un solo servidor: tenemos s servidores idénticos, y la distribución del servicio para cada uno de ellos es exponencial, con un tiempo medio de servicio igual a $1/\mu$.

Con estas suposiciones, podemos aplicar varias fórmulas a fin de describir las características de operación del sistema de servicio:

p = utilización promedio del sistema = $\frac{\lambda}{s\mu}$

P_0 = probabilidad de que cero clientes estén en el sistema = $\left[\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}+\frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-p}\right)\right]^{-1}$

P_n = probabilidad de que haya n clientes estén en el sistema = $\begin{cases}\frac{(\lambda/\mu)^n}{n!}P_0 \;; & 0<n<s\\ \frac{(\lambda/\mu)^n}{s!s^{n-s}}P_0 \;; & 0<n<s\end{cases}$

L_q = número promedio de clientes en la fila de espera = $\frac{P_0(\lambda/\mu)^s p}{s!(1-p)^2}$

W_q = tiempo promedio de espera en la fila = $\frac{L_q}{\lambda}$

W = tiempo promedio transcurrido en el sistema, incluido el servicio = $W_q+\frac{1}{\mu}$

L = tiempo promedio de clientes en el sistema de servicio = λW

Ejemplo de estimación del tiempo de ocio y los costos de operación por hora, mediante el modelo con múltiples servidores. La gerencia del correo internacional DHL en la central del barrio de Mataderos, Buenos Aires, está preocupada por la cantidad de tiempo que los camiones de la compañía permanecen ociosos, en espera de ser descargados. Esta terminal de carga funciona con cuatro plataformas de descarga. Cada una de éstas requiere una cuadrilla de dos empleados, y cada cuadrilla cuesta $30 por hora. El costo estimado de un camión ocioso es de $50 por hora. Los camiones llegan a un ritmo promedio de tres por hora, siguiendo una distribución de Poisson. En promedio, una cuadrilla es capaz de descargar un semirremolque en una hora, y los tiempos de servicio son exponenciales. ¿Cuál es el costo total por hora de la operación de este sistema?

Solución. El modelo con múltiples servidores es apropiado. Para encontrar el costo total de mano de obra y de los camiones ociosos, debemos calcular el tiempo promedio de espera en el sistema y el número promedio de camiones en el mismo. Sin embargo, primero tendremos que calcular el número promedio de camiones en la fila y el tiempo promedio de espera en la fila.

La utilización promedio de las cuatro plataformas es:

$$p=\frac{\lambda}{s\mu}=\frac{3}{1\,(4)}=0{,}75 \quad ; \quad \text{o bien} \quad 75\%$$

Para este nivel de utilización, ahora podemos calcular la probabilidad de que no haya ningún camión en el sistema

$$P_0=\left[\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}+\frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-p}\right)\right]^{-1}=\left[\sum_{s=0}^{4-1}\frac{(3/1)^n}{n!}+\frac{(3/1)^s}{4!}\left(\frac{1}{1-0{,}75}\right)\right]^{-1}$$

$$=\frac{1}{1+3+\frac{9}{2}+\frac{27}{6}+\frac{81}{24}+\left(\frac{1}{1-0{,}75}\right)}=0{,}0377$$

El número promedio de camiones en la fila es:

$$L_q = \frac{P_0 (\lambda / \mu)^s p}{s!(1-p)^2} = \frac{0{,}0377\ (3/1)^4\ (0{,}75)}{4!\ (1-0{,}75)^2} = 1{,}53 \text{ camiones}$$

El tiempo promedio de espera en la fila es:

$$W_q = \frac{L_q}{\lambda} = \frac{1{,}53}{3} = 0{,}51 \text{ hora}$$

El tiempo promedio transcurrido en el sistema es:

$$W = W_q + \frac{1}{\mu} = 0{,}51 \quad \frac{1}{1} = 0{,}51 \text{ hora}$$

Por último, el número promedio de camiones en el sistema es:

$$L = \lambda W = 3\ (1{,}51) = 4{,}53 \text{ camiones}$$

Ahora podemos calcular los costos por hora correspondientes a mano de obra y camiones ociosos:

Costo de mano de obra:	\$30 (s) = \$30 (4)	=	\$120,00
Costo de camiones ociosos:	\$50 (L) = \$50 (4,53)	=	\$226,50
		Costo total por hora =	\$346,50

Modelo C: con fuente finita

Consideremos ahora una situación en la que todas las suposiciones del modelo con un solo servidor son apropiadas, excepto una. En este caso, la población de clientes es finita, porque sólo existen N clientes potenciales. Si N es mayor que 30 clientes, resulta adecuado el modelo con un solo servidor, sobre la suposición de que la población de clientes sea infinita. En los demás casos, el modelo con fuente finita es el que más conviene utilizar. Las fórmulas que se usan para calcular las características de operación del sistema de servicio son:

P_0 = probabilidad de que cero clientes estén en el sistema = $\left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$

p = utilización promedio del servidor = $1 - P_0$

L_q = número promedio de clientes en la fila de espera = $N - \frac{(\lambda + \mu)}{\lambda}(1 - P_0)$

L = tiempo promedio de clientes en el sistema = $N - \frac{\mu}{\lambda}(1 - P_0)$

W_q = tiempo promedio de espera en la fila = $L_q\ [(N - L)\ \lambda]^{-1}$

W = tiempo promedio transcurrido en el sistema, incluido el servicio = $L\,[(N-L)\ \lambda]^{-1}$

Ejemplo de análisis de los costos de mantenimiento aplicando el modelo con fuente finita. Hace casi tres años, Gear Tandil SA instaló un conjunto de 10 robots que incrementó considerablemente la productividad de su mano de obra, pero en el último tiempo la atención se ha enfocado en el mantenimiento. La empresa no aplica el mantenimiento preventivo a los robots, en virtud de la gran variabilidad que se observa en la distribución de las averías. Cada máquina tiene una distribución exponencial de averías (o distribución entre llegadas), con un tiempo promedio de 200 horas entre una y otra falla. Cada hora-máquina perdida como tiempo ocioso cuesta $30, lo cual significa que la empresa tiene que reaccionar con rapidez en cuanto falla una máquina. La empresa contrata sólo a una persona de mantenimiento, quien necesita 10 horas de promedio para reparar un robot. Los tiempos de mantenimiento real están distribuidos exponencialmente. La tasa de salarios es de $10 por hora para el encargado de mantenimiento, el cual puede dedicarse productivamente a otras actividades cuando no hay robots que reparar. Calcule el costo diario por concepto de tiempo ocioso de la mano de obra y los robots.

Solución. El modelo con fuente finita es apropiado para este análisis, porque sólo 10 máquinas constituyen la población de clientes y las demás suposiciones se han cumplido. En esta caso, $\lambda = 1/200$, o sea, 0,005 averías por hora, y $\mu = 1/10 = 0{,}10$ robots por hora. Para calcular el costo del tiempo ocioso para la mano de obra y los robots, tenemos que estimar la utilización promedio del empleado de mantenimiento y L, es decir, el número promedio de robots incluidos en el sistema de mantenimiento. Sin embargo, para mostrar cómo se utiliza el modelo con fuente finita, computaremos primero todas las estadísticas de operación.

La probabilidad de que el sistema de mantenimiento esté vacío es:

$$P_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu} \right)^n \right]^{-1} = \frac{1}{\sum_{n=0}^{10} \frac{10!}{(10-n)!} \left(\frac{0{,}005}{0{,}10} \right)^n} = 0{,}538$$

La utilización promedio del empleado de mantenimiento es:

$$p = 1 - P_0 = 1 - 0{,}583 = 0{,}462 \quad ; \quad \text{o sea,} \quad 46\%$$

El número promedio de robots en espera de ser reparados es:

$$L_q = N - \frac{(\lambda + \mu)}{\lambda} (1 - P_0) = 10 - \frac{0{,}005 + 0{,}10}{0{,}005} (1 - 0{,}538) = 0{,}30 \text{ robots}$$

El número promedio de robots que están en la fila y en proceso de reparación es:

$$L = N - \frac{\mu}{\lambda} (1 - P_0) = 10 - \frac{0{,}10}{0{,}005} (1 - 0{,}538) = 0{,}76 \text{ robots}$$

El tiempo promedio de espera de los robots, para ser atendidos por el encargado del mantenimiento es:

$$W_q = L_q [(N - L) \lambda]^{-1} = 0{,}30 \frac{1}{(10 - 0{,}76)(0{,}005)} = 6{,}49 \text{ horas}$$

Finalmente, el tiempo promedio que transcurre desde que un robot averiado empieza a requerir el servicio hasta que se lleva a cabo su reparación es:

$$W = L[(N - L) \lambda]^{-1} = 0{,}76 \frac{1}{(10 - 0{,}76)(0{,}005)} = 16{,}45 \text{ horas}$$

El costo diario por concepto de tiempo ocioso de la mano de obra y los robots es:

Costo de mano de obra:	($10/hora) (8 horas/día) (0,462 de utilización	=	$ 36,96
Costo de camiones ociosos:	(0,76 robot) ($30/robots hora) (8 horas/día)	=	$182,40
	Costo total diario	=	$219,36

ÁREAS DE DECISIÓN PARA LA ADMINISTRACIÓN

Después de analizar un problema de filas de espera, la gerencia es capaz de mejorar el sistema de servicio introduciendo cambios en uno o varios de los siguientes aspectos:

1. **Tasas de llegada**. Es frecuente que la administración tenga la posibilidad de influir en la tasa de llegada de los clientes, λ, ya sea por medio de publicidad, promociones especiales o precios diferenciales. Por ejemplo, una empresa telefónica aplica precios diferenciales para inducir un cambio en los patrones de las llamadas residenciales de larga distancia, de modo que en lugar de que los clientes las hagan durante el día, prefieran hacerlas por la noche.
2. **Número de instalaciones de servicio**. Al aumentar el número de recursos o instalaciones de servicio, como depósitos de herramientas, casetas de peaje o cajeros automáticos en bancos, o bien, al dedicar algunos recursos de una fase determinada a un conjunto de servicios único, la gerencia logra acrecentar la capacidad del sistema.
3. **Número de fases**. Los gerentes pueden optar por asignar tareas de servicio a fases secuenciales, si consideran que dos instalaciones de servicio secuenciales son más eficientes que una sola. Por ejemplo, en un problema típico de las líneas de ensamble, la decisión se refiere al número de fases necesarias dentro de la misma. La determinación del número de trabajadores que se requieren en la línea también implica la asignación de cierto conjunto de elementos de trabajo a cada uno de ellos. Un cambio en la disposición de la instalación suele incrementar la tasa de servicio, μ, de cada recurso y la capacidad de todo el sistema.
4. **Número de servidores por instalación**. Los gerentes influyen en la tasa de servicio al asignar más de una persona a una misma instalación de servicio.
5. **Eficiencia del servidor**. Mediante un ajuste de la razón entre el capital y la mano de obra, ya sea ideando métodos de trabajo mejorados o instituyendo programas de incentivos, la gerencia puede elevar la eficiencia de los servidores asignados a una instalación de servicio. Los cambios de ese tipo se reflejan en μ.
6. **Regla de prioridad.** Los gerentes establecen la regla de prioridad que debe aplicarse, deciden si cada instalación de servicio deberá tener una regla de prioridad diferente y si se permitirá que, por motivos de prioridad, se altere el orden previsto (señalando, en este último caso, en qué condiciones se hará tal cosa). Esas decisiones afectan los tiempos de espera de los clientes y la utilización de los servidores.
7. **Disposición de las filas.** Los gerentes pueden influir en los tiempos de espera de los clientes y en la utilización de servidores, al decidir si habrá una sola fila o si cada instalación tendrá su respectiva fila en el curso de una fase o servicio determinado.

Es obvio que todos estos factores están relacionados entre sí. Es muy posible que un ajuste en la tasa de llegada de clientes, λ, tenga que ir acompañado de un incremento en la tasa de servicio, μ, de una u otra forma. Las decisiones en torno al número de instalaciones, el número de pases y la disposición de las filas de espera también están relacionadas entre sí.

En cada uno de los problemas que analizamos con los modelos de filas de espera, las llegadas mostraron una distribución de Poisson (o sea, tiempos exponenciales entre llegadas), los tiempos de servicio exhibieron una distribución exponencial, las instalaciones de servicio tenían una disposición sencilla y la disciplina prioritaria consistía en que a quien llega primero, se atiende primero. La teoría de filas de espera se ha usado para desarrollar otros modelos en los que estos criterios no se cumplen, pero dichos modelos son muy complejos. Muchas veces, el carácter de la población de clientes, las restricciones impuestas a las filas, la regla de prioridad, la distribución del tiempo de servicio y la disposición de las instalaciones son tan especiales que la teoría de filas de espera ya no resulta útil. En esos casos, se utiliza a menudo la simulación.

PUNTOSRELEVANTES

- Las filas de espera se forman cuando los clientes llegan a un servicio a un ritmo más rápido que la tasa a la cual pueden ser atendidos. Debido a que las tasas de llegada de los clientes varían, es posible que se formen largas filas de espera a pesar de que la tasa de servicio prevista en el diseño del sistema sea apreciablemente más alta que la tasa promedio de llegada de los clientes.
- Cuatro elementos son comunes en todos los problemas de filas de espera: una población de clientes, una fila de espera, un sistema de servicio y una regla de prioridad para determinar a qué cliente se atenderá a continuación.
- Los modelos de filas de espera se han desarrollado para usarse en el análisis de sistemas de servicio. Si las suposiciones formuladas al crear un modelo de filas de espera son congruentes con la situación real, las fórmulas del modelo pueden resolverse para pronosticar el rendimiento del sistema en lo referente a la utilización de servidores, el tiempo promedio de espera para los clientes y el número promedio de clientes que estarán en el sistema.

TÉRMINOSCLAVE

- Disciplina prioritaria
- Fase
- Fila de espera
- Instalación de servicio
- Población de clientes
- Regla de prioridad
- Sistema de servicio
- Tiempos entre llegada

REFERENCIASBIBLIOGRÁFICAS

Cooper, Robert B. *Introduction to Queuing Theory*. 2ª ed. Elsevier-North Holland, New York. 1980.

Hillier, F.S. y Lieberman, G.S. *Introduction to Operations Research*. 2ª ed. Holden-Day. San Francisco. 1975.

More, P.M. *Queues, Inventories and Maintenance*. John Wiley & Sons. New York. 1958.

Saaty, T.L. *Elements of Queuing Theory with Applications*. McGraw-Hill. New York. 1961.