MODELOS PARA LA TOMA DE DECISIONES EN LINEAS DE ESPERA

María Rodríguez de Estofán y Sandra Franco de Berduc
Universidad Nacional de Tucumán- Argentina
mrestofan@tucbbs.com.ar, sandranfranco@hotmail.com

Resumen

Un problema que se presenta con frecuencia en muchos organismos, como ser bancarios, gubernamentales, supermercados, etc., es el de intentar proporcionar los mayores niveles de servicios posibles, sin necesidad de incrementar su capacidad operativa. Nuestro trabajo consiste en analizar los fundamentos de la teoría de colas exponiendo modelos matemáticos que permiten tomar decisiones oportunas optimizando la planificación de un servicio y el uso de sus recursos disponibles. El objetivo esencial de conocer y aplicar la teoría de colas es la minimización de los costos totales, que surgen de dos fuentes: la propia espera y la capacidad del sistema. Por lo tanto el fin último de un directivo o gerente de un organismo es el de encontrar un equilibrio entre el costo de proporcionar un determinado nivel del servicio, con una cierta capacidad, y el costo de la espera de los clientes. Con este artículo intentamos hacer un aporte a la enseñanza de la matemática, mostrando algunos modelos que se sustentan en la teoría de probabilidades, con gran aplicación en múltiples situaciones de la vida real.

Introducción

Es importante mostrar los aportes que hace la matemática en la toma de decisiones de todo tipo de organizaciones: comerciales, industriales y gubernamentales. Para ello es necesario el estudio del comportamiento de personas, equipos, costos y procedimientos, a fin de comprender sus funcionamientos y optimizar su eficiencia y eficacia. Un problema que se suele presentar con frecuencia en muchos organismos es el de intentar proporcionar los mayores niveles de servicios posibles sin necesidad de incrementar su capacidad. Por ejemplo en una fábrica la carencia de los recursos necesarios origina la espera de trabajos a ser procesados, cuyo almacenamiento produce retrasos con aumentos de costos del trabajo e incumplimiento de la entrega final. Otra situación más adversa es aquella en la que el trabajo en espera es una persona, insatisfecha por el tiempo que pierde en una cola. Este caso es bastante habitual cuando los clientes tienden a solicitar servicios en forma aleatoria. Ejemplos comunes de esperas se presentan en la compra de entradas para espectáculos, oficinas de correos, mostradores de facturación en aeropuertos, hospitales, entradas y salidas de estacionamientos, etc. Todas estas situaciones constituyen problemas de colas o de líneas de espera. Con este artículo intentamos hacer un aporte a la enseñanza de la matemática, mostrando algunos modelos que se sustentan en la teoría de probabilidades, con gran aplicación en múltiples situaciones de la vida real. Nuestro trabajo consiste en analizar los fundamentos de la teoría de colas exponiendo modelos matemáticos que permiten tomar decisiones oportunas optimizando la planificación de un servicio y el uso de sus recursos disponibles. Las decisiones se basan en el empleo del método científico y en el uso de herramientas y conocimientos de muchas ciencias, entre ellas, la matemática, la economía, la física y del comportamiento.

Teoría de colas o líneas de espera

En muchas operaciones se forman líneas de espera para la prestación de un servicio, entre ellas, cuando los clientes esperan en fila para liquidar sus compras; las máquinas de una fábrica esperan ser reparadas o los aviones esperan para aterrizar en un

aeropuerto. La característica común de estos ejemplos en apariencia distintos, es que un número de unidades físicas (llegadas) intentan recibir un servicio de un número limitado de instalaciones (los servidores). Como consecuencia, las llegadas deben esperar algunas veces en líneas hasta que llegue el turno de ser atendidas.
La teoría de colas se ocupa del estudio de las colas de espera de todo tipo concebible y la necesidad de reducir los embotellamientos y congestiones. La teoría de colas analiza la capacidad de una unidad operacional para prestar algún servicio. Las unidades o elementos de interés llegan a la instalación donde se presta el servicio deseado y abandonan el sistema. Estos modelos son característicos de las cajas de supermercados, peajes de autopistas, mostradores de ventas, instalaciones de embarque, etc.
El *objetivo* de los estudios realizados basándonos en la teoría de colas es determinar el número óptimo de personas o instalaciones necesarias para atender a los clientes que llegan al azar y minimizar los costos del servicio con el de la espera o congestión.
En general, la teoría de colas ofrece información sobre la probabilidad de que un determinado número de personas, máquinas, etc. tenga que esperar en la cola, durante cuanto tiempo espera hacerlo y el porcentaje de tiempo inactivo de la instalación que presta el servicio. Estas cantidades se usan para determinar si debería aumentarse el tamaño del espacio de espera o la velocidad del servicio. Toda esta información es obtenida de modelos matemáticos, que constan de fórmulas analíticas, basados en las distribuciones de probabilidades Poisson y Exponencial.

Características de la teoría de colas
Todo problema de teoría de colas puede describirse en término de tres características: la *llegada*, la *cola* y el *servidor*.

1) *La llegada*: se describen por su distribución analítica, que pueden especificarse de dos formas: distribución del número de llegadas por unidad de tiempo o distribución del tiempo entre llegadas. Si la distribución de llegadas se especifica en la primera forma, se deberá describir el número de llegadas que pueden ocurrir en cualquier período de tiempo dado. Es posible describir el número de llegadas que ocurren en una hora. Cuando las llegadas ocurren al azar, la información que interesa está dada por la probabilidad de que ocurran n llegadas en un período dado, donde $n = 0,1,2,...$
Si se supone que las llegadas ocurren con una tasa promedio constante y que son independientes entre sí, se presentan acorde con la distribución de probabilidad de Poisson. La probabilidad de n llegadas en el tiempo T está dada por

$$P(n,T) = \frac{e^{-\lambda T}(\lambda T)^n}{n!} \qquad \text{con} \quad n = 0,1,2, \ldots \quad \text{donde:}$$

λ = Tasa promedio de llegadas por unidad de tiempo. T = Intervalo de tiempo.
n = Número de llegadas en el tiempo T.
P(n, T) = Probabilidad de que ocurra n llegadas en el tiempo T

Para valores pequeños de λT existe una alta probabilidad de que ocurran cero llegadas en el tiempo T y que la mayor parte de las probabilidades se concentran alrededor de 0, 1, 2 llegadas. A medida que aumenta el valor de λT la forma de la distribución cambia hacia una forma más simétrica (normal) y la probabilidad del número de llegadas aumenta.

El segundo método de especificación de llegadas está dado por el tiempo que transcurre entre llegada y llegada. En este caso se debe especificar la distribución de probabilidad de una variable aleatoria continua que mida el tiempo transcurrido entre una llegada y otra. Si las llegadas siguen la distribución Poisson, se demuestra matemáticamente que el tiempo entre llegadas seguirá una distribución exponencial.

$P(T \le t) = 1 - e^{-\lambda t}$ con $0 \le t < \infty$ donde:

$P(T \le t)$ es la probabilidad de que el tiempo entre llegadas T sea menor o igual que un valor dado t.
λ tasa media de llegadas por unidad de tiempo.
t un tiempo dado.

A medida que t aumenta, la probabilidad de que haya ocurrido 1 llegad se aproxima a 1.
La distribución exponencial y la de Poisson son equivalentes en cuanto a las suposiciones fundamentales sobre las llegadas. Por lo tanto, cualquiera de las dos puede usarse para especificar las llegadas: todo depende si se desea calcular el tiempo entre llegadas o el número de llegadas que ocurrirán en un tiempo dado.

2) ***La cola***: la naturaleza de la cola también afecta el tipo de modelo de teoría de colas que se formule. Una disciplina de colas es la bien conocida regla de *"quién llega primero se atiende primero"*.
Cuando se describe la teoría de cola, es necesario especificar la longitud de la línea de espera. Una suposición matemática común es que la *línea de espera puede alcanzar una longitud infinita*.
Por último, debe definirse el comportamiento de los clientes en la cola. La conducta de los *clientes* que se presupone en los modelos de teoría de colas simples, es que estos *esperarán hasta recibir el servicio*.
Para propósitos analíticos, las suposiciones más comunes en las teorías de colas son: servir primero a quién llegó primero, que la longitud de la línea es infinita y que las llegadas esperarán en la línea hasta que se les dé el servicio.
3) ***El prestador del servicio***: también existen varias características del prestador del servicio que afectan al problema de teoría de colas. Una de estas, es la distribución del tiempo de servicios. Al igual que el tiempo de llegada, el tiempo de servicio puede variar de un cliente al siguiente. Una presuposición común para la *distribución del tiempo de servicios implica una distribución exponencial*.
La segunda característica del prestador del servicio que debe especificarse es el *número de prestadores* que se encontrarán presentes. En ocasiones, a cada prestador del servicio se le llama *canal*.
El servicio puede proporcionarse en *una sola fase* o en *fases múltiples*. Una situación de fases múltiples es aquella en la que el cliente debe pasar a través de dos o más prestadores en secuencia para terminar el servicio.
La combinación de varios prestadores y varias fases del servicio da lugar a una gran variedad de problemas de teorías de colas.

Formulación de problemas de teorías de colas

Una vez que se han dado las suposiciones sobre las llegadas, la cola y los prestadores del servicio, se desea predecir, el *desempeño* de un sistema de colas específico. El *desempeño* que se predice para el sistema puede describirse mediante el *número promedio de llegadas en la cola*, el *tiempo promedio de espera de una llegada y el porcentaje de tiempo perdido de los prestadores del servicio*. Estas medidas de desempeño se pueden utilizar para decidir la cantidad de prestadores del servicio que deben colocarse, los cambios que se pueden hacer en la velocidad del servicio u otros cambios en el sistema de colas.

Cuando se evalúan las medidas de desempeño de teoría de colas, deben determinarse los *costos totales* siempre que sea posible. Esto se hace añadiendo el *costo del tiempo de espera de la llegada y el costo de los prestadores del servicio*. En el caso de reparación de máquinas, el tiempo de espera de la máquina es función del costo de la producción perdida. En los casos en que las llegadas son los clientes, resulta muy difícil estimar el costo del tiempo de espera. No siempre es posible determinar el costo total de un sistema de colas. En lugar de esto se utilizan objetivos sustitutos. Por ejemplo, un objetivo sustituto es que los clientes no deben esperar más de un promedio de 5 minutos para obtener el servicio. Las medidas y parámetros del desempeño para los modelos de teorías de colas, se especifican mediante la siguiente notación:

λ = Tasa promedio de llegadas (el número de llegadas por unidad de tiempo).

$1/\lambda$ = Tiempo promedio entre las llegadas.

μ = Velocidad media del servicio (el número de unidades a las que se le da servicio por unidad de tiempo cuando el prestador se encuentra trabajando).

$1/\mu$ = Tiempo promedio requerido para el servicio.

ρ = Factor de utilización del prestador del servicio (la proporción del tiempo en que el prestador del servicio trabaja).

P_n = Es la probabilidad de que n unidades (llegadas) se encuentren en el sistema.

L_q = Número promedio de unidades en la cola (longitud promedio de la cola).

L_s = Número promedio de unidades en el sistema.

W_q = Tiempo promedio de espera en la cola.

W_s = Tiempo promedio de espera en el sistema.

En el sistema se refiere a las unidades que pueden encontrarse en la cola o en el servicio. Es decir, W_q se refiere al tiempo de espera de una unidad en la cola antes de que comience el servicio y W_s se refiere al tiempo total de espera más el tiempo necesario para tener el servicio En condiciones uniformes, las condiciones de arranque iniciales no afectan las medidas del desempeño. La condición uniforme se logrará solamente cuando μ sea mayor que λ, la velocidad del servicio debe ser superior a la velocidad de llegadas para que se presente la condición uniforme. Siempre que $\mu \leq \lambda$ el sistema de colas es inestable y la línea se puede acumular potencialmente hasta el infinito debido a que las unidades llegan con mayor rapidez de las que reciben el servicio. Se supondrá entonces que $\mu > \lambda$ a lo largo del trabajo.

Modelo simple de teoría de colas

Se basa en las siguientes suposiciones: a) Un solo prestador de servicio y una fase.

b) Distribución de llegadas Poisson donde λ = tasa promedio de llegadas.

c) Tiempo de servicio exponencial, donde μ = tasa promedio del servicio.

d) Disciplina de colas de servir primero a quién llega primero, todas las llegadas esperan en línea hasta que se les da el servicio y existe la posibilidad de una longitud infinita en la cola.

A partir de estas suposiciones se pueden derivar las siguientes estadísticas de desempeño: $\rho = \frac{\lambda}{\mu}$ $P_0 = 1 - \frac{\lambda}{\mu}$ $P_n = P_0\left(\frac{\lambda}{\mu}\right)^n$ $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$ $L_S = \frac{\lambda}{(\mu-\lambda)}$

$W_q = \frac{\lambda}{\mu(\mu-\lambda)}$ $W_S = \frac{1}{(\mu-\lambda)}$

Modelo con varios prestadores del servicio

El modelo simple con llegadas tipo Poisson y tiempos de servicios exponenciales puede extenderse para incluir sin gran dificultad a varios prestadores del servicio. Si se establece que **S** es igual al *número de los prestadores de los servicios*, las mediciones de desempeño del sistema de colas con varios prestadores del servicio serán:

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

$$P_n = P_0\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \qquad 1 \le n \le s \qquad P_n = P_0\left[\frac{\left(\frac{\lambda}{\mu}\right)^n}{s!(s)^{n-s}}\right] \qquad n \ge s$$

$$L_q = \frac{P_0\left(\frac{\lambda}{\mu}\right)^s \rho}{s!(1-\rho)^2} \qquad L_S = L_q + \frac{\lambda}{\mu} \qquad W_q = \frac{L_q}{\lambda} \qquad W_S = W_q + \frac{1}{\mu}$$

Estas fórmulas se verifican en condiciones de uniformidad y se presupone que las llegadas son de tipo Poisson, el tiempo de servicio exponencial; se aplica la disciplina de colas de atender primero a quién llega primero, todas las llegadas esperan en la cola hasta recibir servicios y la cola tiene longitud infinita.

Aplicación 1: suponga que un cajero bancario puede atender a los clientes a una velocidad promedio de 10 clientes por hora $(\mu = 10)$. Además, suponga que los clientes llegan a la ventanilla de los cajeros a una tasa promedio de 7 por hora $(\lambda = 7)$. Se considera que las llegadas siguen la distribución Poisson y el tiempo de servicio sigue

la distribución exponencial. En la condición uniforme el sistema de cola tendrá las siguientes características de desempeño:

$\rho = \frac{7}{10}$ el prestador del servicio trabajará el 70% del tiempo.

$P_0 = 1 - \frac{7}{10} = \frac{3}{10}$; 30% del tiempo no habrá clientes en el sistema.

$P_n = \frac{3}{10}\left(\frac{7}{10}\right)^n$; esta fórmula determina la posibilidad de que n clientes se encuentren en el sistema en cualquier momento dado, n = 1,2,3,..., $P_1 = 0{,}21$, $P_2 = 0{,}147$, $P_3 = 0{,}1029$; etc.

$L_q = \frac{7^2}{10(10-7)} = 1{,}63$; en promedio 1,63 clientes estarán en la cola.

$L_S = \frac{7}{(10-7)} = 2{,}33$; en promedio 2,33 clientes estarán en el sistema.

$W_q = \frac{7}{10(10-7)} = 0{,}233$; el cliente pasa un promedio de 0,233 horas esperando en la cola.

$W_S = \frac{1}{(10-7)} = 0{,}333$; el cliente pasa un promedio de 0,333 horas en el sistema.

Es posible evaluar el desempeño del sistema de colas. El administrador tendrá que tomar en consideración el "tiempo perdido del prestador del servicio (30%), el tiempo que espera el cliente (0,233 horas) en la cola y la longitud de la línea que se forma (1,63 clientes)". Si este rendimiento es inaceptable, se puede colocar un segundo prestador del servicio o hacer otros cambios en las características de las llegadas, de la cola o del prestador de los servicios.

Aplicación 2: suponga que se coloca un segundo cajero en la aplicación anterior. ¿Qué tanto mejorará el servicio?. Los cálculos de desempeño para S=2 son:

$\rho = \frac{7}{2(10)} = 0{,}35$; los prestadores utilizan el 35% del tiempo

$P_0 = 0{,}4814$; prob. de que no haya clientes. $P_1 = 0{,}3369$; prob de que haya 1 cliente.

$P_2 = 0{,}1179$; prob. de que haya 2 clientes. $P_3 = 0{,}0413$; prob. de que haya 3 clientes.

$P_4 = 0{,}0145$; prob. de que haya 4 clientes. $L_q = 0{,}0977$; un prom de 0,0977 clientes estará en la línea. $L_S = 0{,}7977$; un promedio de 0,7977 clientes estará en el sistema.

$W_q = 0{,}0139$; el cliente pasa un promedio de 0,0139 horas en la cola (menos de un 1').

$W_S = 0{,}1139$; el cliente pasa un promedio de 0,1139 horas en el sistema.

Con dos prestadores del servicio, las estadísticas de los clientes mejoran. Ahora se tiene un promedio de solamente 0,097 clientes en la línea y el cliente espera en promedio solamente 0,0139 horas para recibir el servicio (menos de un minuto). El costo de este servicio es que los prestadores solamente están ocupados durante el 35% de su tiempo. A menos que se desee un servicio extraordinariamente bueno, es probable que el banco no desee incurrir en el gasto del segundo cajero.

Conclusiones

Con este trabajo hemos querido destacar la importancia de la Matemática en la toma de decisiones de optimización de servicios a menores costos. Aunque, hemos mostrado los casos más simples de los modelos de teorías de colas, existen muchos modelos más elaborados, donde varían los supuestos. Uno de ellos, fue considerar que la longitud de la cola es infinita, lo que significa que la llegada de un cliente para ser atendido o el servicio completo, no afecta la probabilidad de futuras llegadas. Sin este supuesto las ecuaciones de los modelos con uno o con varios prestadores de servicio difieren.

También, los modelos cambian si el tiempo de la prestación del servicio es constante que se traduce fundamentalmente en variaciones de la longitud promedio de la cola y del tiempo promedio de espera.

La teoría de colas no ha sido ideada sólo para contestar *cuánto* tiempo de espera *debe* introducirse en un sistema, sino por el contrario, responde dos preguntas importantes: ¿Qué cantidad de tiempo de espera es posible en un sistema?, ¿Cómo cambiará este tiempo de espera luego de alterar las instalaciones? Para solucionar estos problemas los ejecutivos de empresas recurren a los modelos que aporta la teoría de colas.

El objetivo esencial de conocer y aplicar esta teoría es la minimización de los costos totales, que surgen de dos fuentes: la propia espera y la capacidad del sistema. Los relacionados con la espera incluyen los costos de las personas que prestan el servicio, del tiempo de espera de los clientes y los derivados de la posible pérdida de clientes o de ventas por no haber sido atendidos a tiempo. El costo de la capacidad del sistema se refiere al costo de mantener un determinado nivel del servicio. Por lo tanto el fin último del directivo de una empresa es encontrar un equilibrio entre el costo de proporcionar un buen nivel de servicio con una cierta capacidad y el costo de la espera de los clientes.

Bibliografía

Gould, f. J., Eppen, G. D. y Schmidt, C. P. (1992) *Investigación de Operaciones en la Ciencia Administrativa*. Prentice-Hall Hispanoamericana. México.

Kotler, P. (1973) *Mercadotecnia Aplicada*. Interamericana. México.

Sacristán, G., Pérez Gómez, A. (1985) *La Enseñanza: su Teoría y su Práctica* Capítulos 8 y 10. Akal Editor. España

Shamblin, J. E. y Stevens, G. T. (1975) *Investigación de Operaciones.* McGraw-Hill. México.