LA CALIDAD DEL SERVICIO, LA GESTION DE FLUJOS Y LA TEORIA DE COLAS

RAMIRO ORREGO POSADA

RESUMEN

El tener que esperar en una cola es una experiencia considerada como desagradable, especialmente si se tiene que esperar de pie. Los períodos largos de espera irritan a las personas y las invitan a desertar, a irse a otra parte e incluso a no regresar. Esta situación que afecta los beneficios potenciales puede gestionarse en forma eficiente utilizando la Teoría de Colas.

Este artículo trata sobre la gestión de la rapidez del servicio y su relación con la teoría de colas. Se describe el fenómeno de espera en forma simple, con la intensión de mostrar el funcionamiento del sistema y el uso de los modelos de colas como una herramienta importante para gestionar los flujos internos en un sistema de prestación del servicio.

1. INTRODUCCION

Uno de los elementos importantes dentro del ciclo de prestación del servicio, tiene relación con las líneas de espera más comúnmente llamadas colas. El tiempo de espera es tal vez uno de los componentes de este ciclo peor gestionado, ya sea por falta de conocimiento para su tratamiento o simple desconocimiento de su importancia, lo cual es grave en esta época de competencia generalizada.

El tiempo de espera, debe ser incluido dentro del concepto de la Calidad del Servicio, ya que él es en gran parte una medida de percepción individual de la calidad del servicio, ya que a nadie le gusta esperar por largos períodos de tiempo para ser atendido. La rapidez de la atención debe ser parte importante de la estrategia de servicio, especialmente si aceptamos la existencia de clientes impacientes que abandonan y se van a otra parte, llevándose consigo un potencial de utilidades.

La rapidez de la atención debe ser parte importante de la estrategia de servicio, especialmente si aceptamos la existencia de clientes impacientes que abandonan y se van a otra parte, llevándose consigo un potencial de utilidades.

La situación es aún peor cuando ésta se perpetúa en el tiempo. Hoy escuchamos sobre deserciones cero, y como ésta tiene un gran impacto sobre el

RAMIRO ORREGO POSADA. Profesor área Métodos Cuantitativos, Departamento de Informática y Sistemas. Universidad Eafit.

negocio, las utilidades aumentan si la relación con el cliente se prolonga en el tiempo.

Para lograr perpetuar esta relación, atraer nuevos clientes se debe gestionar la rapidez del servicio y aún más, se debe mantener, porque de lo contrario se desvanece el atractivo para el cliente.

Ofrecer un servicio rápido no es sólo cuestión de calidad, costos y beneficios están involucrados y los cuales son argumentos competitivos para una compañía.

En un proceso de prestación del servicio se puede diseñar un sistema más rápido que equilibre los costos asociados con el mejoramiento del servicio versus los beneficios perdidos asociados, con la espera, puesto que no sería factible o económico diseñar un sistema de líneas de espera donde nadie tenga que esperar.

En este artículo se presentan los conceptos básicos de un sistema de colas y algunas consideraciones que contribuyen a la comprensión del fenómeno para encausar un estudio de mejoramiento, teniendo en cuenta factores económicos o de criterio.

> El tiempo de espera es tal vez uno de los componentes de este ciclo peor gestionado, ya sea por falta de conocimiento para su tratamiento o simple desconocimiento de su importancia, lo cual es grave en esta época de competencia generalizada.

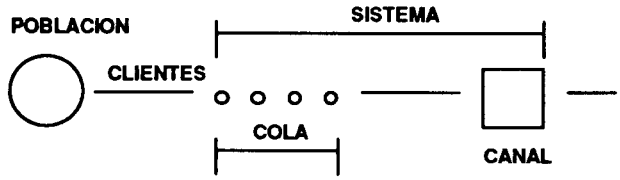
2. FLUJOS

Se refiere a las llegadas del clientes a un sistema en diversos momentos de tiempo. Acá es necesario identificar el ciclo o variación del flujo en el tiempo, porque las acciones de gestión dependerán de la identificación de períodos pico o de máxima demanda que exceden la capacidad física del sistema de prestación del servicio, y de otros períodos donde la demanda es menor que dicha capacidad.

3. SISTEMA DE COLAS

Son dos los elementos que componen un sistema de colas: La cola o línea de espera y el canal -o canales- de servicio, según se ilustra en la **figura 1.**

FIGURA 1
Sistema Básico de Colas



Se ve que existe un elemento externo al sistema, el cual es la población potencial generadora de nuevos clientes que llegan a solicitar servicio. Estos pueden llegar en forma independiente o en lotes y en forma aleatoria o también a intervalos constantes, aunque esta última condición no es común en sistemas de atención a personas.

3.1 La cola

Se refiere a las unidades que están en espera de servicios y que serán atendidos según una disciplina de servicio que establece el canal. La unidades pueden estar o no físicamente frente al canal de servicio, puesto que una cola puede ser formada por clientes que solicitan un servicio por teléfono y quedan en lista de espera.

La cola puede ser finita o infinita, dependiendo de si existe o no un tope superior para el máximo número de clientes admitidos en ella.

3.2 Canal

Se refiere a la persona, proceso o máquina que presta el servicio y que establece la disciplina de servicio, como por ejemplo el primero en llegar el primero en ser atendido, prioridades o cualquier otra.

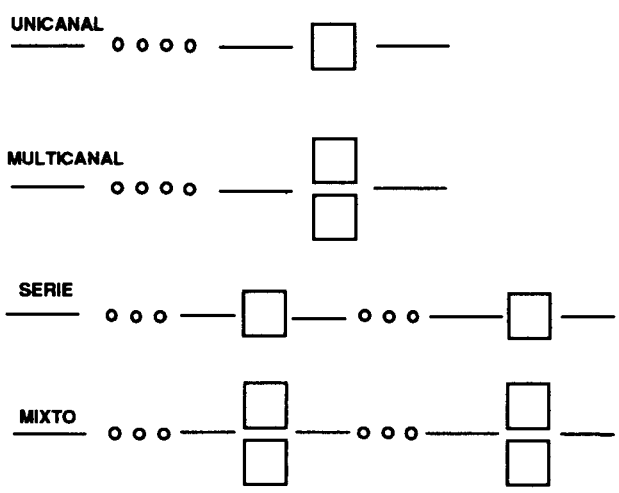
Una característica fundamental en relación con el canal de servicio es el tiempo de servicio, o sea el tiempo necesario para atender a un cliente y el cual una vez conocido, permite calcular la tasa de servicio o número de clientes atendidos por unidad de tiempo.

> Ofrecer un servicio rápido no es sólo cuestión de calidad, costos y beneficios están involucrados y los cuales son argumentos competitivos para una compañía.

3.3 Clasificación de los sistemas de colas

Según el número y disposición de los canales de servicio los sistemas de colas se pueden clasificar como se aprecia en la **figura 2.**

FIGURA 2
Sistemas de Colas



Estos son lo que se consideran los sistemas básicos de colas.

3.4 Sistema

Está conformado por la cola y el canal -canales- de servicio, como puede observarse en la **figura 1.**

3.5 Capacidad del sistema

Está determinada por la cantidad de clientes que se pueden atender -por todos los canales- por unidad de tiempo. También se puede referir a ella como el número de clientes que puede físicamente alojar o manejar en un momento dado. La capacidad del sistema determina entonces el nivel de servicio y es por tanto la primera decisión a tomar, la cual por sus implicaciones en el monto de la inversión y los beneficios esperados es básicamente una decisión estratégica, que afecta la competitividad de la empresa.

Esta decisión compromete el tamaño físico de la instalación y el número necesario de servidores en relación con la variable demanda.

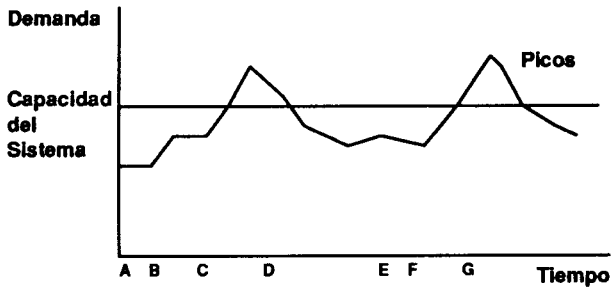
La capacidad por supuesto es limitada, porque no es lógico y mucho menos económico diseñar un sistema donde nadie espere y todos los clientes puedan ser atendidos al mismo tiempo; se debe entonces hallar la capacidad óptima teniendo en cuenta la variación de la demanda, tal que se minimice, no necesariamente a corto plazo, el costo total del sistema.

El conocimiento de la variación de la demanda se debe apoyar en un estudio estadístico, sin embargo otros enfoques indican que muchas empresas ajustan la capacidad acorde al funcionamiento y experiencias acumuladas debido a que la determinación inicial de dicha capacidad es crítica y una decisión que la sobredimensione genera inversiones ociosas y costos operacionales altos.

> Las acciones de gestión dependerán de la identificación de períodos pico o de máxima demanda que exceden la capacidad física del sistema de prestación del servicio, y de otros períodos donde la demanda es menor que dicha capacidad.

Para una capacidad dada, normalmente se presentan picos de demanda, en diferentes horas, días, semanas o meses, que la exceden, tal como se ilustra en la **figura 3.**

FIGURA 3
Capacidad y Variación de la Demanda en el Tiempo



Según se aprecia en la figura hay dos situaciones que se deben gestionar:

- Picos de la demanda: Es una demanda que no puede ser satisfecha en un período dado, conlleva a deserciones y por tanto pérdida de beneficios por insatisfacción de los clientes, que incluso pueden no regresar.

Las acciones para manejar esta situación apuntan a descabezar dichos picos, tratando de trasladar la demanda hacia períodos de baja, ya sea desestimulándola o aplicando diversas acciones relacionadas con políticas de precios, oferta de servicios complementarios y otras que son bien conocidas en marketing.

- Gestión de flujo para cuando la demanda es menor que la capacidad. En este caso se trata de diseñar el sistema de tal forma que permita un servicio rápido a los clientes, optimizando los tiempos de espera o el tamaño de la cola. Es acá donde entra la teoría de colas como una técnica que permite gestionar el flujo rápido de los clientes.

> La capacidad del sistema determina entonces el nivel de servicio y es por tanto la primera decisión a tomar, la cual por sus implicaciones en el monto de la inversión y los beneficios esperados es básicamente una decisión estratégica, que afecta la competitividad de la empresa.

¿Por qué si la demanda es menor que la capacidad, se forma una línea de espera?

La razón por la cual se forma una línea de espera se debe a la aleatoriedad existente entre los tiempos entre llegadas de nuevos clientes y los tiempos de servicio.

3.6 Estadísticas de congestión. Variación de parámetros

Los modelos de la teoría de colas permiten describir el funcionamiento de un sistema de espera, proporcionando una estimación de estadísticas de congestión, entre los cuales se pueden mencionar: tiempo medio esperado en el sistema -W-, tiempo medio esperado en la cola -Wq-; número esperado de unidades en el sistema -L-, número esperado de unidades en la cola -Lq-, utilización del sistema -ρ- y porcentaje de ocio del canal o canales, como las estadísticas básicas. Al usuario le corresponde obtener valores óptimos de funcionamiento ya sea rediseñando el sistema o variando diferentes parámetros del sistema como: la tasa de servicio -μ- y el número de canales -k-.

3.6.1 Variación de la tasa de servicio

Incrementar la tasa de servicio conlleva a un incremento en el nivel de servicio, lo cual puede lograrse mediante análisis y mejoramiento del método de trabajo del servidor -estandarización de la rutina e información-, introducción de medios mecánicos o electrónicos y adición de auxiliares trabajando en equipo.

Un ejemplo bastante sencillo ilustrará, los efectos que tienen sobre las estadísticas de congestión en un sistema de colas, el incremento de la tasa de servicio.

Supóngase que los clientes llegan a una caja de una tienda de abarrotes, según una distribución exponencial con tiempo medio de cuatro minutos entre clientes. La cajera atiende clientes de acuerdo con una distribución de Poisson con tasa μ igual a 20 clientes por hora. Con estos datos, veamos la situación actual:

W = 0.2 horas = 12 minutos. En promedio un cliente espera 12 minutos en el sistema.

Wq = 0.15 horas = 9 minutos promedio de espera en la cola.

L = 3 clientes promedio en un momento dado.

Lq = 2.25 clientes promedio en la cola en un momento dado.

Al agregar un empacador, la tasa de servicio se eleva de 20 a 30 clientes por hora, bajo supuesto que el nuevo tiempo de servicio nuevamente tiene distribución exponencial. En esta condición se tiene:

W = 0.0666 horas = 4 minutos.

Wq = 0.0333 horas = 2 minutos.

L = 1 persona.

Lq = 0.5 personas.

> La determinación inicial de dicha capacidad es crítica y una decisión que la sobredimensione genera inversiones ociosas y costos operacionales altos.

Según puede observarse el tiempo medio de espera en la cola -Wq- tiene una reducción muy significativa, de 9 minutos pasa a 2 minutos. Igualmente el número esperado de clientes en la cola se reduce de 2.25 a 0.5 clientes promedio.

Esto por supuesto tiene un efecto positivo sobre el sistema, mirado desde el punto de vista de posibles deserciones por clientes impacientes, que se llevan utilidades potenciales a otro lugar.

3.6.2 Variación del número de canales

Incrementando el número de servidores o canales de servicio, por supuesto que también se incrementará el nivel de servicio en el sistema.

Si en el ejemplo ilustrativo de la tienda de abarrotes, en lugar de adicionar un empacador como ayudante, se adiciona una nueva cajera con igual tasa de servicio, se pueden manejar dos tipos de modelos de colas:

- En el caso en que al frente de cada cajera se tenga una línea de espera, se tendrían dos sistemas unicanal. Para hallar las estadísticas de congestión basta con analizar un sólo sistema, pero considerando que la tasa de entrada para cada cola se reduce a la mitad -principio de descomposición de la tasa de entrada- debido a que es igualmente probable que un cliente seleccione una u otra caja cuando las líneas tienen igual longitud, o seleccione la caja con menor número de clientes de espera.

Las estadísticas de congestión serían:

W = 0.08 horas = 4.8 minutos promedio.

Wq = 0.03 horas = 1.8 minutos promedio.

L = 0.6 clientes promedio.

Lq = 0.225 clientes promedio.

Estos resultados serían iguales para ambas cajas.

- El segundo caso sería un modelo multicanal, en el cual se forma una sola fila de espera frente a las dos cajeras y donde un cliente en la cola es atendido por la primera cajera que esté desocupada. Para este caso se tendría:

W = 0.0614 = 3.7 minutos.

Wq = 0.7 minutos = 42 segundos.

L = 0.92 clientes.

Lq = 0.17 clientes.

Para este ejemplo ilustrativo la **tabla 1** muestra un resumen de las estadísticas de congestión.

Si el tiempo de espera representa un valor en términos de pérdida de utilidades debido a deserciones de los clientes insatisfechos, se puede observar que un modelo M/M/2 -multicanal- sería la mejor alternativa pues genera el menor tiempo de espera y menor número promedio de unidades en la cola.

TABLA 1
Comparación Estadísticas de Congestión

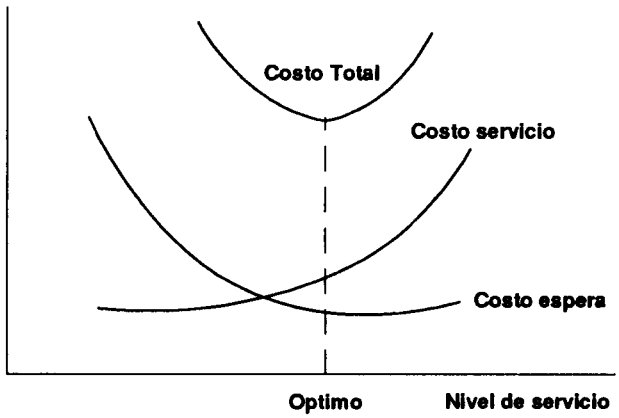
	CONDICION INIC. MODELO M/M/1 CON $\mu = 20$ $\lambda = 15$	MODELO M/M/1 TASA MEJORADA $\mu = 30$ $\lambda = 15'$	2 MODELOS M/M/1 CON $\mu = 20$ $\lambda = 7.5$	MODELO M/M/2 $\mu = 20$
W	12 Min.	4 Min.	4.8 Min.	3.7 Min.
Wq	9 Min.	2 Min.	1.8 Min.	0.7 Min.
L	3 Clientes	1 Cliente	0.6 Cliente	0.92 Clientes
Lq	2.25 Clientes	0.5 Clientes	0.22 Clientes	0.17 Clientes

4. COSTOS ASOCIADOS

Sin embargo, el problema se debe analizar considerando los costos involucrados.

El costo en un sistema de colas tiene dos componentes: costo del servicio y costo asociado con la espera y se desea diseñar el sistema de tal forma que produzca el menor costo total. En la **figura 4** se puede observar la relación de dichos costos con el nivel de servicio y su incidencia en el costo final.

FIGURA 4
Relación Costos del Sistema



Para bajos niveles de servicio se experimentan largas colas y por tanto costos de espera altos. Conforme se incrementa el nivel de servicio se incrementan los costos del mismo, pero disminuyen los costos de espera. El costo total del sistema disminuye, pero a partir de cierto nivel los ahorros en el costo de espera no compensan los incrementos del costo de servicio.

El costo del servicio, visto bajo la óptica de la calidad puede considerarse como inversiones requeridas para mejorar el servicio. El costo de espera se refiere a un beneficio perdido, y es indirecto, ya que es cierto que no se hace ningún pago cuando por ejemplo en un banco, un cliente disgustado se va porque la cola es demasiado larga. El retirarse, el no regresar causa una pérdida de oportunidad, de beneficios potenciales.

5. ALCANCES Y LIMITACIONES DE LA TEORIA DE COLAS

En la vida corriente, determinar los costos asociados con el servicio es sencillo, pero no sucede igual con el costo de espera. Existen muchas situaciones en las que es difícil o muy complejo dar un valor al costo de espera. Por ejemplo ¿cuál sería el costo de espera del cliente para un supermercado?, ¿Para un banco?, ¿Para un servicio de fotocopiado?

El costo en un sistema de colas tiene dos componentes: costo del servicio y costo asociado con la espera y se desea diseñar el sistema de tal forma que produzca el menor costo total.

La percepciones y el comportamiento de cada cliente tiene un rango amplio de variación, algunos son más pacientes que otros e incluso aún la misma persona se comporta diferente en una situación que en otra. Es diferente esperar sentado en un restaurante a esperar de pie frente a un cajero.

Los modelos de colas, establecen unas condiciones ideales para operarlos, sin embargo las condiciones de desempeño de algunos sistemas reales son tan complejas o variables que no permiten el uso de modelo alguno. Demasiados supuestos para su aplicación desvirtúan el modelo y por tanto los resultados.

A pesar de estas limitaciones en la aplicación de los modelos de colas, existen formas alternas para manejar estas situaciones.

1. Por apreciación de expertos dar estimativos para el costo de espera, dependiendo de los factores sicológicos y competitivos de la situación.

2. Evaluar la situación mediante simulación, la cual es una técnica muy eficiente, que permite estimar mediante valores muestrales, estadísticas de congestión bastante confiables y que es aplicable a cualquier situación compleja de líneas de espera.

3. Por muestreo, se puede estimar la cantidad de tiempo que los clientes esperan para ser atendidos y el número de personas impacientes que se van, como base para generar una función de pérdida que oriente la toma de decisiones sobre dónde y cuándo sería rentable

disponer de servicio adicional. Las pérdidas por clientes que se marchan deben darse por apreciación.

4. Diseño del sistema según criterio gerencial.

El último ítem es muy aplicado por organizaciones de servicio, los cuales por criterio establecen según condiciones competitivas, un tiempo máximo de espera para un cliente en la cola o un número máximo de clientes en espera, lo que permite calcular el nivel óptimo de servicio.

Es común en supermercados y tiendas americanas, la existencia de la estrategia de servicio establecidos por criterio como el caso del supermercado Lucky Market en California, donde existen avisos que dicen: "tres son multitud", que incluso el mismo cliente lo puede "gritar", para que inmediatamente otra persona corra a abrir una nueva caja.

Igualmente y gracias a las cajas registradoras inteligentes, grandes cadenas de nuestro medio, evalúan hora por hora los clientes atendidos y número de artículos comprados, información que alimenta a un sistema central. Establecido por criterio un tiempo de espera máximo considerado como óptimo, se determina continuamente el número de cajas necesarias según la afluencia de clientes. Por supuesto esta proyección requiere de un estudio estadístico que tiene en cuenta la temporada comercial, previamente estudiada. Determinado el número óptimo de cajas que deben funcionar por hora, se puede programar el recurso humano necesario utilizando técnicas de programación lineal -problema de asignación-, para lograr una distribución óptima de menor costo.

> La percepciones y el comportamiento de cada cliente tiene un rango amplio de variación, algunos son más pacientes que otros e incluso aún la misma persona se comporta diferente en una situación que en otra. Es diferente esperar sentado en un restaurante a esperar de pie frente a un cajero.

6. CONCLUSIONES

El tiempo de espera es un componente del ciclo de servicio y constituye un momento de verdad que si no es gestionado eficientemente, puede destruir toda la buena imagen que sobre el servicio ofrecido, tenga el cliente. Al igual que el caso de los famosos últimos diez metros, donde la compañía se esmera por su publicidad, ofrece productos o servicios de buena calidad y logra atraer los clientes, pero en los últimos diez metros, ya propiamente en el almacén desilusiona al cliente con una mala atención, sólo que en el caso de las colas el desencanto se produce debido a un alto tiempo de espera.

La teoría de colas es una herramienta útil que ayuda a comprender y analizar el problema de congestión, y si bien pueden existir diversos limitantes para su implementación en algunas situaciones, es posible por formas alternas obtener resultados que orienten la toma de decisiones.

Los clientes que abandonan, los que no regresan, representan utilidades potenciales perdidas, por esto la gerencia debe gestionar la rapidez del servicio para lograr un balance entre las inversiones que incrementan su nivel y las utilidades potenciales perdidas.

¿Cuántos beneficios se pierden a largo plazo por un cliente que no regresará?

BIBLIOGRAFIA

Eigleir, Pierre y Langcard, Eric. Servucción. El marketing de Servicios. México: Mc. Graw Hill. 1989.

Hiller, Frederick S. y Lieberman, Gerald J. Introducción a la investigación de operaciones. 5a. edición, México: Mc. Graw Hill. 1991. 833 p.

Horovitz, Jacques. La calidad del servicio: A la conquista del cliente. Madrid: Mc. Graw Hill. 1991. 105 p.

Reichheld, Frederick y Sascer, W. Earl. Deserción cero: La calidad en las empresas de servicio. Tomado de Weckly. Fax-Colombian Editions.

Schonberger, Richard J. Cómo crear la cadena cliente-proveedor. Hacia una compañía de categoría mundial. Bogotá: Editorial Norma. 1993. 396 p.

Winston, Wayne L. Operations Research: Applications and algorithms. Boston: PWS-Kent Publishing Company. Boston, 1987. 1025 p.