LA TEORÍA DE COLAS

VÍCTOR MANUEL QUESADA IBARGÜEN
Profesor
Facultad de Ciencias Económicas
Universidad de Cartagena

INTRODUCCIÓN

La vida de todos los habitantes de este planeta es una permanente cola o línea de espera. Se hace cola para hacer efectivo un cheque, retirar dinero de su cuenta o consignar , para entrar al teatro, para entrar a los estadios, a los sitios donde se presentan espectáculos, en las iglesias para tomar la comunión, en el restaurante, en el aeropuerto, en el supermercado, para tomar el bus o el taxi, hacemos cola, de igual manera, frente a los semáforos cuando conducimos.

Son pues tantas las situaciones en donde las colas o líneas de espera están presentes en la vida del hombre, que se ha dedicado un considerable esfuerzo científico a su estudio.

Y es que, en apariencia, la cola que hago hoy para lograr un determinado servicio no tiene mayor importancia en cuanto al tiempo invertido; pero examinemos las siguientes situaciones:

Supongamos una persona que sólo a partir de los 20 años ha empezado a entender que la vida es corta y por tanto hay que aprovecharla de la mejor forma. Este joven individuo realiza normalmente las siguientes actividades: Asiste una vez al banco a realizar alguna transacción invirtiendo en promedio 30 minutos; los martes, cuando el cine es más barato, acude a un teatro y, en promedio, gasta 60 minutos para adquirir boleto.

La espera para tomar el bus diariamente es de unos 20 minutos, completando un promedio de 100 minutos en los 5 días hábiles de la semana y para no complicar más al muchacho, supongamos por último que tiene la responsabilidad de cancelar los consumos por servicios públicos de su residencia, que todos se pagarán al tiempo en un mismo sitio, de modo que los 60 minutos que gasta en esta actividad se lo cargamos a razón de 15 minutos/semana. En promedio, entonces, nuestro amigo gasta 205 minutos semanales haciendo colas, esperando por un servicio, esto equivale a 820 minutos al mes, 9.840 minutos al año, 442.800 minutos en los 45 años adicionales de vida probable que le quedan, es decir, casi dos años (teniendo en cuenta días de doce horas) de su vida.

Si esta es la situación para una persona de bajo perfil, ya nos podemos imaginar los resultados para la gente activa.

En el presente artículo se hará un recorrido por la teoría de colas señalando sus orígenes, su evolución, fundamentación matemática y un ejemplo de aplicación práctica.

NATURALEZA Y ORIGEN

Cuando la demanda por un servicio excede ampliamente la capacidad de prestación del mismo, se forma una cola.

Por ejemplo, si un cajero puede atender a treinta clientes por hora y se presentan 45 clientes a solicitar atención, entonces se formará una cola frente a la ventanilla de este servidor. Si los clientes van a retirar dinero y esta es su única opción, seguramente se someterán a la espera.

En la caja registradora del supermercado, una situación similar podrá implicar pérdida de clientes causando con ello una baja en los beneficios. El dueño del negocio puede entonces pensar que vale la pena invertir en una caja adicional, porque sus costos son cubiertos por los beneficios proporcionados por los clientes impacientes, muchos de los cuales esperarán a ser servidos.

De esta manera se introduce el concepto de optimización en la teoría de colas. En general, a diferencia de la teoría de optimización, cuyo principal objetivo es maximizar o minimizar una función sujeta a restricciones, la teoría de colas es una teoría matemática descriptiva, intenta formular, interpretar y predecir, con

el propósito de un mejor conocimiento de las colas y con el fin de introducir las modificaciones oportunas.

El origen de esta teoría se encuentra en los problemas de congestión de redes telefónicas. En 1905 Erlang (Ingeniero Danés) realizó un trabajo original sobre líneas de espera con el fin de determinar el efecto de la fluctuación de la demanda de servicio en la utilización del teléfono ya que se presentaba un problema de congestión. Erlang se propuso calcular las demoras que sufrían los usuarios.

Los trabajos de Erlang estimularon a otros en este campo y así surgieron trabajos importantes tales como las de Thornton D.Fry en 1928 "The Theory of probability as applied to problems of congestion".

F. Pollaezek desarrolló la fórmula para un solo canal de servicios con ingreso (llegadas) Poisson y tiempo de duración de los servicios con distribución arbitraria.

Borrer estudió el problema de clientes impacientes que después de esperar una cantidad fija de tiempo (impaciencia determinista) la abandonan.

En 1958 H. Glazer examinó el problema de clientes que se trasladan de una a otra cola, de varias formadas ante una instalación.

Aunque los trabajos iniciales se refieren a las inquietudes de Erlang respecto a las congestiones telefónicas, las aplicaciones de las teorías de las colas han sido bastante variadas. Se ha aplicado al tráfico de transporte (aéreo, terrestre, marítimo), colas para servicios, teatros, hospitales, clínicas , inventarios y procesos industriales (Mantenimiento, líneas de montaje, interferencias de máquinas), procesos fisicos (operaciones de una cuadrilla de un puerto, movimiento de partículas hacia un desagüe), procesos epidémicos en biología, crecimiento de población etc.

ELEMENTOS BÁSICOS DE UN SISTEMA DE COLAS

LLEGADAS. Los clientes llegan al sistema en busca de un servicio y pueden ser personas, máquinas que requieren reparación, llamadas telefónicas que deben ser contestadas etc. Los clientes pueden llegar individualmente o por lotes; a intervalos regulares o con un patrón aleatorio; pueden venir de una población infinita (muy grande) o pueden provenir de un conjunto finito ($n<30$).

LOS SERVICIOS. El tiempo que se requiere para concluir el servicio es el segundo elemento de importancia y puede ser el mismo para cada cliente o variar considerablemente en forma aleatoria.

Número de canales (puntos) de servicio. Puede existir un solo canal de servicio o varios (Multicanal).

DISCIPLINA. Cuando los clientes esperan por los servicios puede haber una sola cola o varias para cada servidor o una para varios servidores. El espacio de la cola puede ser limitado y los clientes que llegan cuando la cola está llena pueden retirarse (rechazado). El orden de los servicios se puede basar en la regla FIFO o PEPS (Primero en entrar primero en salir), pero también es posible que exista un servicio rápido o prioritario para algunos clientes.

MEDIDAS DEL RENDIMIENTO.

Existen varias formas para evaluar el rendimiento de un sistema de colas Algunas de las medidas son el tiempo que permanece un cliente en la línea (cola) antes de ser atendido, el tiempo que permanece ocioso un servidor, el número de clientes que en promedio se encuentran en el sistema.

FÓRMULAS APLICABLES A UN SISTEMA DE COLAS

NOTACIÓN.

λ = Número promedio de clientes que llegan en una unidad de tiempo.
u = Número promedio de clientes al cual puede dar servicio la instalación en una unidad de tiempo, suponiendo que no hay escasez de clientes.

L = Número esperado de unidades que se atienden y/o esperan en el sistema.

L_q = Número esperado en cola (no incluye los clientes que se atienden).
P_n = Probabilidad de tener n unidades en el sistema.

W_q = Tiempo probable de espera en cola, de una llegada.

W = Tiempo probable de permanencia en el sistema (tanto en cola como en servicio).

ρ = Tasa de utilización del sistema.

Hay varias relaciones que son de especial interés, por ejemplo:

$\rho = \frac{\lambda}{\mu}$ Tasa de utilización del sistema o probabilidad de que el sistema esté ocupado.

$P_0 = 1 - \frac{\lambda}{\mu}$ = Probabilidad de que el sistema esté vacío.

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

$$L = \frac{\lambda}{\mu\,(\mu - \lambda)}$$

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

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

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

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

P(W>t) = Probabilidad de que el tiempo en el sistema sea mayor que un tiempo t específico.

$P(W>t) = e^{-t/w}$

$P(W_q>t) = \rho\, e^{-t/w}$ igual que la anterior pero referida a la cola.

A partir de estas formulas, bajo supuesto que las llegadas siguen una distribución de Poisson y la duración de los servicios es exponencial, se derivan las medidas de efectividad y servirán de guía para la toma de decisiones respecto al sistema.

LA APLICACIÓN PRÁCTICA (SUGERENCIAS AL INVESTIGADOR).

En la práctica, la aplicación de la teoría de colas requiere que el investigador establezca las tasas de llegada y de servicios e identifique el patrón estadístico que siguen.

A pesar de que la gran mayoría de los sistemas obedecen a distribuciones de Poisson en las llegadas y exponencial en cuanto a la duración de los servicios, es necesario que el investigador, una vez obtenida la información mediante técnicas de muestreo, realice las pruebas de hipótesis necesarias.

EL ESTADO ESTABLE

Si en un sistema se realiza un estudio de colas sin tener en cuenta el momento en que deben ser tomadas las muestras, es muy probable que la información resulte sesgada. En un banco, por ejemplo, no podría el investigador

seleccionar un día que coincida con el pago a los asalariados porque se encontraría con una gran congestión que podría pesar mucho en los resultados.

Se recomienda entonces determinar el estado estable, o sea, aquel período en que el flujo de clientes es más o menos constante. Una vez determinado se procede a muestrear las llegadas y los servicios. En cada caso hay que establecer el tamaño de la muestra usando las técnicas estadísticas. El intervalo de tiempo a utilizar para registrar las llegadas se fija teniendo en cuenta el flujo observado; si es alto, el investigador pudiera registrar llegadas en intervalos de 5 minutos; si es muy bajo, las registraría cada media hora por ejemplo. Esto con el fin de no registrar cantidades de intervalos con ausencia de llegadas.

Ejemplo: Las llegadas a un banco se dan de acuerdo con la siguiente distribución (previamente se estableció N= 537 intervalos de 1 minuto).

n	F_N	P_o(Probabilidad observada= Frecuencia relativa)
0	92	0.1713
1	173	0.3221
2	139	0.2588
3	79	0.1471
4	36	0.0670
5	13	0.0242
6	3	0.0056 X 10-3
7	2	0.0037 X 10-3
TOTAL	537	1.000

Cálculo de la media (λ observado)

$$X = \Sigma_0^7 \frac{n F_n}{N} = \frac{1}{537}(0X92 + 1X173 + 2X139 + 3X79 + 4X36 + 5X13 + 6X3 + 7X2)$$

$$X = \frac{929}{537} \approx 1.73 = \lambda_o = 1.73$$ clientes/minuto

Cálculo de λ Teórico

Si se supone que el comportamiento de las llegadas es Poisson, se pueden calcular las Probabilidades Teóricas así:

$$P_n = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$; t = 1min, con $e^{-\lambda t}$ = 0.1773

Frecuencia Teórica $F_t = N \times P_t$

$P_o = (1.73)^0/0! * e^{-1.73} = 0.1773$	X537	95
$P_1 = (1.73)^1/1! * e^{-1.73} = 0.3067$	X537	165
$P_2 = (1.73)^2/2! * e^{-1.73} = 0.2653$	X537	143
$P_3 = (1.73)^3/3! * e^{-1.73} = 0.1530$	X537	82
$P_4 = (1.73)^4/4! * e^{-1.73} = 0.0662$	X537	36
$P_5 = (1.73)^5/5! * e^{-1.73} = 0.0229$	X537	12
$P_6 = (1.73)^6/6! * e^{-1.73} = 0.0066$	X537	3
$P_7 = (1.73)^7/7! * e^{-1.73} = 0.0016$	X537	1
TOTAL	0.9996 ≈ 1.00	

Para aplicar la prueba Ji-cuadrado y tratando de obviar algunos inconvenientes que se presentan cuando las frecuencias están por debajo de 5 observaciones, se

agruparán las tres últimas categorías así:

n	F_0	F_t	$\Delta = \lvert F_0 - F_t \rvert$	Δ^2	Δ^2/F_t
0	92	95	3	9	0.0947
1	173	165	8	64	0.3879
2	139	143	4	16	0.1119
3	79	82	3	9	0.1098
4	36	36	0	0	0.00
5 ó más	18	16	2	4	0.25
	537			Σ	0.9543

El número de grados de libertad V se calcula como: V = K - 1 -P donde K= # de grupos y P = # parámetros de la distribución, que en la Poisson es 1.

V = 6 - 1 - 1 = 4

De la tabla Ji-cuadrado, con 4 grados de libertad y un nivel de significancia del 5% se obtiene un valor crítico de 9.49, mucho mayor que Ji-cuadrado calculado, por lo tanto teóricamente puede decirse con confianza que la distribución de Poisson es adecuada.

Se establece que $\lambda = 1.73\ min^{-1}$.

LOS SERVICIOS

Estos se muestrean por su duración; a cada cliente que llega a la ventanilla se le toma el tiempo hasta que sale despachado.

En el ejemplo tamaremos 414 observaciones de servicios de consignación en un banco. Se asume que los tiempos son exponenciales

DURACIÓN(Min)	F_n	Probabilidad observada (P_0)
0.0--0.50	90	0.217
0.50--1.00	140	0.338
1.00--1.50	73	0.176
1.50--2.00	48	0.116
2.00--2.50	24	0.058
2.50--3.00	17	0.0341
3.00--3.50	7	0.017
3.50--4.00	5	0.012
4.00--4.50	5	0.012
4.50--5.00	4	$9.66X10^{-3}$
5.00--5.50	1	$2.42X10^{-3}$
TOTAL	414	1.00

Tomando las marcas de clase de la distribución se calcula la media.

$$X = \frac{1}{414}(0.25X90+0.75X140+1.25X73+1.75X48+2.25X24+2.75X17+3.25X7+3.75X5+4.25X5+475X4+5.25X1)$$

X = 1.18

La duración media del servicio es 1.18 minutos por lo tanto la tasa de servicios $u = \frac{1}{1.18} \approx 0.90\,min^{-1}$

LA PRUEBA DE HIPÓTESIS:

La distribución exponencial cuya función de distribución es $e^{-\mu t}$, será la distribución a constatar:

DURACIÓN SERVICIO	$e^{-\mu t}$	P. TEÓRICA P_t	FRECUENCIA TEÓRICA $P_t X N$
P_0	1.0		
P(0.25)	0.798	0.2014	83
P(0.75)	0.509	0.2894	120
P(1.25)	0.325	0.1840	76
P(1.75)	0.207	0.1176	49
P(2.25)	0.132	0.075	31
P(2.75)	0.084	0.0478	20
P(3.25)	0.053	0.0304	13
P(3.75)	0.034	0.0194	8
P(4.25)	0.022	0.0124	5
P(4.75)	0.014	0.0079	3
P(5.25)	0.008	0.0050	2

Obsérvese que en la segunda columna se está calculando la probabilidad de que el tiempo de servicio "t" sea mayor que un tiempo específico P(t>0.25)=0.798, la diferencia entre dos probabilidades consecutivas da la probabilidad puntual (columna 3).

DURACIÓN	F_N	F_T	Δ	Δ^2	Δ^2/F_T
0.25	90	83	7	49	0.59
0.75	140	120	20	400	3.33
1.25	73	76	3	9	0.118
1.75	48	49	1	1	0.020
2.25	24	31	7	49	1.581
2.75	17	20	3	9	0.45
3.25	7	13	6	36	2.77
3.75	5	8	3	9	1.13
4.25	5	5	0	0	0
4.75	5	9	4	16	1.78

Σ 11.76

$X^2_{CALC} = 11.76$

La prueba Ji-cuadrado con 8 grados de libertad al 5% produce un valor crítico de 15.51 el cual es mayor que el Ji-cuadrado calculado, por tanto aceptamos que los servicios siguen una ley exponencial con tasa $\mu = 0.90\ min^{-1}$.

En resumen, el sistema de nuestro ejemplo tiene:

Tasa de llegadas (λ) = 1.73 min^{-1} (clientes por minuto)
Tasa de servicios (μ) = 0.90 min^{-1} (clientes por minuto)

Siendo la tasa de llegadas mayor que la de servicios se requiere de, por lo menos, dos unidades (canales) de servicio (S).

Tomando S = 2 se procede a calcular las medidas de efectividad del sistema:

$$P_0 = \left[\frac{S^s \rho^{s+1}}{S!(1-\rho)} + \sum \frac{(S\rho)^N}{N!} \right]^{-1};$$

$$P_n = \left\{ \frac{(S\rho)^n}{N!} P_0 \ (N = 1, S) \right.$$

$$\left\{ \frac{S^s \rho^N}{S!} P_0 \ (n = S+1.S+2,...) \right.$$

$$L_q = \frac{S^s \rho^{s+1}}{S!(1-\rho)^2} P_0$$

$$\rho = \frac{\lambda}{S\mu} = \frac{1.73}{2 \text{ X } 0.90} = 0.9611$$

$$P_0 = \left[\frac{2^2 (0.96)^3}{2!(1-0.96)} + \Sigma \frac{(2 \text{ X } 0.96)^N}{N!} \right]^{-1}$$

$$= [44.24 + 1 + 1.92 + 1.84]^{-1}$$

$$P_0 = 0.0204 \Rightarrow P_0 \approx 0.02$$

$$P_N = \frac{(S \rho)^N}{N!} P_0$$

$$P_1 = \frac{(2 \text{ X } 0.96)^1}{1} \text{ X } 0.02 = 0.038$$

$$P_2 = \frac{(2 \text{ X } 0.96)^2}{2} \text{ X } 0.02 = 0.037$$

$$P_3 = \frac{2^2 (0.96)^3}{2!} \text{ x } 0.02 = 0.0354$$

(Aquí $P_n = \frac{S^s \rho^n}{S!} P_0$

y así sucesivamente

$$Lq = \frac{2^2 (0.96)^3}{2! (1-0.96)^2} 0.02 = 0.88$$

$$Wq = \frac{Lq}{\lambda} = \frac{0.88}{1.73} = 0.5087 \text{ min}$$

$$W = Wq + \frac{1}{\mu} = 050.87 + \frac{1}{0.90} =$$

1.6198 min

$$L = \lambda W = 2.80$$

De lo anterior se puede colegir que:

- La probabilidad de que los dos servidores se encuentren desocupados es bien baja = 0.02.

La probabilidad de que un cliente que llega tenga que esperar es 0.037 (probabilidad de encontrar el sistema totalmente ocupado).

- El número de personas (clientes, esperados en cola es 0.88 (obsérvese que como promedio, no tiene que ser entero).

- El total de clientes esperando y/o recibiendo servicios es, en promedio, 2.8

- El tiempo que un cliente debe esperar para ser atendido es, en promedio, 0.5 minutos.

- El tiempo que un cliente permanece en el sistema entre espera y atención es, en promedio, 1.62 minutos.

CONCLUSIÓN

Es importante para el administrador de un sistema, poder establecer los patrones que gobiernan una línea de

espera, a fin de poder tomar decisiones anticipadas para brindar un servicio de óptima calidad a sus clientes a unos costos que resulten razonables.

Aunque, como ya se anotara, el propósito de la teoría de colas no es la optimización persé, su ayuda es valiosa al momento de determinar el comportamiento de los clientes que acuden a los sistemas de colas en busca de un servicio.

La poca utilización que algunos administradores le encuentran a herramientas como la descrita, obedece en gran medida a un cierto temor de confrontar la teoría con la práctica. Pareciera como si estuviéramos de acuerdo en que entre las dos no puede existir acuerdo.

Con un poco de manejo estadístico y el estudio de las bases de esta teoría es posible hacer aplicaciones valiosas.

BIBLIOGRAFÍA

BETTIN, Flórez Maritza del Rosario; Montes Vergara Manuel Antonio, "Análisis de Línea de Espera en el Banco Industrial Colombiano". Corporación Tecnológica de Bolívar, Facultad de Ingeniería, Departamento de Ingeniería Industrial. Bajo la dirección de Víctor Manuel Quesada Ibagüen. 1.983

BIERMAN-Bonini-Hausman, "Análisis cuantitativo para la toma de Decisiones", Edit. Irwin. 1996

KAMLESH, Mather, Salow Daniel, "Investigación de Operaciones", El arte de la toma de decisiones", Edit. Prentice Hall, Hispanoamericana.1.996

PANICO, Joseph A,, "Teoría de las Colas", Centro Regional de Ayuda Técnica. 1973

SAATY, Thomas; "Elementos de la Teoría de las Colas". Edit. Iberoamérica.1.967

WAYNE, Winston, "Investigación de Operaciones, Aplicaciones y Algoritmos". Edit. Iberoamérica.1.994

BRONSON, Richard, "Teoría y Problemas de Investigación de Operaciones", McGraw-Hill 1.995.

MONTGOMERY, DOUGLAS C. Y RUNGER GEORGE. "Probabilidad y Estadística, Aplicadas a la Ingeniaría", Edit. McGraw-Hill .1996.