DOCUMENTOS DE INVESTIGACIÓN

Facultad de Administración

No. 138, ISSN: 0124-8219
Mayo de 2012

Aplicación de colas de Poisson en procesos de 'toma de decisiones' en la gestión de servicios médicos

Diego Fernando Cardona Madariaga
Javier Leonardo González Rodríguez
Miller Rivera Lozano
Jesús Andrés Romero Dávila

Universidad del Rosario
Facultad de Administración

Aplicación de colas de Poisson en procesos de 'toma de decisiones' en la gestión de servicios médicos

Documento de investigación No. 138

Diego Fernando Cardona Madariaga
Javier Leonardo González Rodríguez
Miller Rivera Lozano
Jesús Andrés Romero Dávila

Grupo de Investigación en Perdurabilidad Empresarial (GIPE)
Línea de Investigación en Estrategia

Universidad del Rosario
Facultad de Administración
Editorial Universidad del Rosario
Bogotá D.C.
2012

Aplicación de colas de Poisson en procesos de 'toma de decisiones' en la gestión de servicios médicos / Diego Fernando Cardona Madariaga…[et al.].—Bogotá: Editorial Universidad del Rosario, 2012.

46 p. (Documento de Investigación; 138)

ISSN: 0124-8219

Procesos de Poisson / Procesos puntuales / Administración de servicios de salud / Investigación operacional / Líneas de espera / Calidad del servicio / Servicios médicos de urgencias – Administración / I. Cardona Madariaga, Diego Fernando / II. Gonzalez Rodriguez, Javier Leonardo / III. Rivera Lozano, Miller / IV. Romero Dávila, Jesús Andrés / V. Universidad Colegio Mayor de Nuestra Señora del Rosario, Facultad de Administración, Grupo de Investigación en Perdurabilidad Empresarial (GIPE), Línea de Investigación en Estrategia / VI. Título. / VII. Serie.

658.83 SCDD 20

Diego Fernando Cardona Madariaga
Javier Leonardo González Rodríguez
Miller Rivera Lozano
Jesús Andrés Romero Dávila

Corrección de estilo
Lina Morales

Diagramación
Fredy Johan Espitia Ballesteros

Editorial Universidad del Rosario
http://editorial.urosario.edu.co

ISSN: 0124-8219

Primera edición: Mayo de 2012
Hecho en Colombia
Made in Colombia

Contenido

Índice

Gráficos

Tablas

Aplicación de colas de Poisson en procesos de 'toma de decisiones' en la gestión de servicios médicos

Diego Fernando Cardona Madariaga*
Javier Leonardo González Rodríguez**
Miller Rivera Lozano***
Jesús Andrés Romero Dávila****

Resumen

En este documento se revisa teóricamente la distribución de probabilidad de Poisson como función que asigna a cada suceso definido, sobre una variable aleatoria discreta, la probabilidad de ocurrencia en un intervalo de tiempo o región del espacio disjunto. Adicionalmente se revisa la distribución exponencial negativa empleada para modelar el intervalo de tiempo entre eventos consecutivos de Poisson que ocurren de manera independiente; es decir, en los cuales la probabilidad de ocurrencia de los eventos sucedidos en un intervalo de tiempo no depende de los ocurridos en otros intervalos de tiempo, por esta razón se afirma que es una distribución que no tiene memoria. El proceso de Poisson relaciona la función de Poisson, que representa un conjunto de eventos independientes sucedidos en un intervalo de tiempo o región del espacio con los tiempos dados entre la ocurrencia de los eventos según la distribución exponencial negativa. Los anteriores conceptos se usan en la teoría de colas, rama de la investigación de operaciones que describe y brinda soluciones a situaciones en las que un conjunto de individuos o

* Ingeniero Civil, M. Sc. y Ph. D. en Ciencias Administrativas; profesor asociado de la Facultad de Administración de la Universidad del Rosario con funciones de director de investigaciones. Correo electrónico: diego.cardona@urosario.edu.co

** Médico, Especialista en Salud Pública y Ph. D. en Economía y Gestión de la Salud; profesor principal de la Facultad de Administración de la Universidad del Rosario con funciones de director de la Maestría en Administración de la Salud. Correo electrónico: javier.gonzalez@urosario.edu.co

*** Ingeniero de Sistemas, M. Sc. en Administración; coordinador del Laboratorio de Modelamiento y Simulación de la Facultad de Administración de la Universidad del Rosario. Correo electrónico: miller.rivera@urosario.edu.co

**** Estudiante de matemáticas, joven investigador, Universidad Distrital Francisco José de Caldas. Correo electrónico: jesusromerodavila@yahoo.com

elementos forman colas en espera de que se les preste un servicio, por lo cual se presentan ejemplos de aplicación en el ámbito médico.

Abstract

This paper reviews theoretically the Poisson probability distribution function that assigns to each event defined on a discrete random variable, the probability of occurrence in a time interval or region of disjoint space. Additionally, review the Negative Exponential distribution used to model the interval time between consecutive Poisson events occurring independently, or which occurrence probability of the events in a time interval does not depend on those that occurred at other time intervals, for this reason is identified as a distribution without memory. The Poisson process relates the Poisson function that represents a set of independent events occurred in a time interval or space region with the times given between the occurrences of events according to the Negative Exponential distribution. The above concepts are used in queuing theory, a branch of operations research that describes and provides solutions to situations where a set of individuals or elements are at lines waiting to receive a service, which are exemplified with application in the medical field.

1. Introducción

El presente documento es la continuación del titulado “Una aproximación de la variable aleatoria a procesos de toma de decisión que implican condiciones de riesgo e incertidumbre”. Ahora se aborda la aplicación de colas de Poisson en situaciones propias de la teoría de colas, para procesos de prestación de servicios que se caracterizan por la llegada aleatoria de los clientes al sistema, representada a través de una distribución de probabilidad de Poisson, con tasas medias de servicio que obedecen a una distribución de probabilidad exponencial negativa.

La aplicación de las colas de Poisson en la teoría de colas permite abstraer y modelar servicios médicos caracterizados por la llegada de pacientes de forma aleatoria, independientes entre sí, en los que se considera nula la llegada de dos o más pacientes de forma simultánea, con tasas medias de servicio que obedecen a una distribución exponencial negativa, que disponen una disciplina de atención de los pacientes (FIFO, LIFO, SIRO, GD) con uno o más servidores y una cola (demanda del servicio) generalmente considerada como infinita.

El desarrollo de la temática propuesta incluye la caracterización de las distribuciones de probabilidad de Poisson y exponencial negativa, la explicación del proceso de Poisson y la relación de las dos distribuciones en la solución de problemas de la teoría de colas. Adicionalmente se incluye, para la teoría de colas, la descripción de los conceptos básicos, la notación y terminología utilizada, los modelos elementales y su aplicación en procesos de toma decisión en la gestión de servicios médicos.

Las distribuciones de probabilidad de Poisson y exponencial negativa son utilizadas frecuentemente en problemas de la teoría de colas; la distribución de Poisson representa un número de sucesos independientes que ocurren a una velocidad constante dentro de un intervalo de tiempo específico, expresado en unidades que van, por ejemplo, desde segundos hasta años. Permite modelar situaciones tan diversas como el número de llamadas que llegan a una central telefónica, el número de bacterias que se reproducen en una cierta población, el tiempo que pueda durar una vía cerrada a causa de un derrumbe, el número de personas que llegan a un autoservicio o la cantidad de seguros solicitados a una aseguradora; la distribución exponencial negativa representa la longitud de los intervalos de ocurrencia entre los sucesos que se distribuyen según la distribución Poisson (Canavos, 1988).

El proceso de Poisson relaciona las distribuciones de Poisson y exponencial negativa en el tratamiento de problemas propios de la teoría de colas; a continuación se adelanta una revisión general de cada una de estas distribuciones y de los conceptos y modelos incluidos en dicha teoría, con el fin de adelantar la revisión de situaciones específicas de la gestión de salud que permiten su aplicación.

2. Distribución de Poisson

Sea X una variable aleatoria que representa eventos aleatorios independientes que ocurren a una rapidez constante sobre el tiempo o el espacio. Se dice que la variable aleatoria X tiene una distribución de Poisson con la siguiente distribución de probabilidad (Canavos, 1988):

$$P(X = k) = \begin{cases} \dfrac{e^{-(\lambda t)}(\lambda t)^k}{k!}, si\ k = 0,1,\ldots,n,\ldots \\ 0\ en\ otro\ caso \end{cases}$$

Se dice de esta forma que la variable aleatoria X, tiene una distribución de Poisson con parámetro $\lambda > 0$ (Meyer, 1992). Para detalles de la deducción de la fórmula de la distribución de Poisson ver Canavos (1988).

La definición anterior corresponde a una distribución de probabilidad, puesto que satisface los siguientes axiomas de un espacio de probabilidad. La distribución de probabilidad de Poisson tiene como espacio muestral el conjunto de números enteros 0,1,2,3,...; la probabilidad asociada a cada punto muestral se obtiene reemplazando directamente en la fórmula de distribución de probabilidad presentada.

a. La función siempre es positiva, pues cada uno de los tres factores que la definen es también positivo para cualquier valor de k de acuerdo con los siguientes argumentos: una función exponencial siempre es positiva, además, dado que el parámetro $\lambda > 0$, implica que $\lambda^k > 0$ para cualquier k, y por último el factorial de un número entero positivo siempre es mayor o igual a uno, $k! \geq 1$.

b. Utilizando desarrollos en series de potencias y aplicando propiedades de la sumatoria, sobre todos los valores del espacio muestral, se tiene:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!}$$ *, suma de probabilidades sobre el espacio muestral*

$$\sum_{k=0}^{\infty} P(X = k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$ *, propiedades de la sumatoria*

$$\sum_{k=0}^{\infty} P(X = k) = e^{-(\lambda)}\, e^{\lambda}$$

$$\sum_{k=0}^{\infty} P(X = k) = 1$$

$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$, *corresponde a la expansión en series de Taylor de* e^{λ}

Si *X* es una variable aleatoria con distribución de Poisson, de parámetro λ,

$$E(X) = \lambda$$
$$VAR\ (X) = \lambda$$

Donde $E(X)$ denota el valor esperado de la variable aleatoria, este indicador también se conoce con el nombre de promedio, y $VAR\ (X)$ denota la varianza de la variable aleatoria. La distribución de Poisson tiene la propiedad interesante de tener su valor esperado (promedio) igual a su varianza.

3. Distribución exponencial negativa

Una de las distribuciones continuas que ha encontrado aplicación en una amplia clase de fenómenos propios de la teoría de colas y de la teoría de la confiabilidad es la llamada distribución exponencial negativa, caso particular de la distribución gamma y de la distribución de Erlang; esta función queda completamente determinada por un parámetro $\lambda > 0$. La función que define esta importante distribución de probabilidad es (Meyer, 1992):

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \; si \; x > 0 \\ o \; en \; otro \; caso \end{cases}$$

Para esta función también es fácil verificar que se trata de una distribución de probabilidad a partir de las siguientes consideraciones:

a. $f(x) > 0$ *para* $x > 0$, dado que la función exponencial siempre es positiva, al igual que el parámetro λ.

b. La segunda condición pide que $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_0^{\infty} \lambda e^{-\lambda x}\, dx = \lim_{b\to\infty} \int_0^{b} \lambda e^{-\lambda x} dx, \text{ suma sobre los valores de probabilidad}$$

$$\int_0^{\infty} \lambda e^{-\lambda x}\, dx = -\lambda \lim_{b\to\infty} \int_b^{0} \lambda e^{-\lambda x} dx, \text{ propiedad de linealidad de la integral}$$

$$\int_0^{\infty} \lambda e^{-\lambda x}\, dx = -\lambda(-\tfrac{1}{\lambda}) \lim_{b\to\infty} \left[e^{-\lambda x} \Big|_b^0 \right]$$

$$\int_0^{\infty} \lambda e^{-\lambda x}\, dx = 1(1-0) = 1, \text{ se satisface la condición pedida}$$

Por ser la distribución exponencial negativa una función continua para la que existe de forma explícita una antiderivada, permite mostrar respecto a la función de distribución acumulada (FDA) los siguientes resultados:

a. La FDA de una función de variable aleatoria está dada por $P(X \leq x)$, es decir, esta permite calcular la probabilidad que la variable aleatoria X tome un valor menor o igual que cualquier valor de su espacio muestral. Para el caso de la distribución exponencial negativa se tiene:

$$P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \text{ aquí } f \text{ es una variable aparente de integración}$$

$$P(X \leq x) = \int_0^x \lambda e^{-\lambda t} dt$$

$$P(X \leq x) = 1 - e^{-\lambda x} \; si \; x \geq 0$$

b. El valor esperado o el promedio de esta distribución se define y está dado por:

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

c. La varianza de una distribución de probabilidad se encuentra definida de la siguiente manera:

$$VAR(X) = E[X - E(X)^2]^2$$

Un resultado conocido en la teoría de la probabilidad, que permite un cálculo sencillo para la varianza de una variable aleatoria, es:

$$VAR(X) = E(X^2) - E(X)^2$$

Se utiliza esta última expresión para el cálculo de la varianza de la distribución exponencial negativa, donde $E(X^2) = \frac{2}{\lambda^2} \; y \; E(X) = \frac{1}{\lambda}$, por lo tanto:

$$VAR(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2$$

$$VAR(X) = \frac{2-1}{\lambda^2} = \frac{1}{\lambda^2}$$

d. Una importante propiedad de la que goza la distribución exponencial negativa es la falta de memoria; es la única distribución continua que satisface esta propiedad.

Sean *a y t* números reales positivos, considérese la siguiente probabilidad condicional:

$$P(X > a + t/X > a)$$

Sean A y B dos eventos asociados a un espacio muestral Ω, la probabilidad condicional entre los eventos A y B está dada por:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \; con \; P(B) > 0$$

$$P(B/A) = \frac{P(B \cap A)}{P(A)} \; con \; P(A) > 0$$

Por lo tanto

$$P(X > \alpha + t/X > \alpha) = \frac{P[(X > \alpha + t) \wedge (X > \alpha)]}{P(X > \alpha)}$$

$$P(X > \alpha + t/X > \alpha) = \frac{P(X > \alpha + t)}{P(X > \alpha)}$$

$$P(X > \alpha + t/X > \alpha) = \frac{e^{-\lambda(\alpha + t)}}{e^{-\lambda\alpha}}$$

$$P(X > \alpha + t/X > \alpha) = e^{-\lambda t}$$

e. La distribución exponencial negativa es monótona decreciente, esta condición permite establecer la siguiente propiedad:

$P(0 \leq X \leq \Delta t) > P(t \leq X \leq t + \Delta t)$ Si $t > 0$; los dos intervalos anteriores tienen una longitud Δt, y las probabilidades corresponden al área bajo la curva de la función de probabilidad, un argumento para establecer la desigualdad anterior.

4. Proceso de Poisson

Las siguientes características permiten identificar el modelo de Poisson como el idóneo ante una caracterización de un experimento de teoría de colas (Meyer, 1992).

Sea X_t una variable aleatoria (proceso estocástico) que puede tomar los valores 0,1,2,..., durante un intervalo de tiempo $(0,t)$, y denótese la probabilidad de que la variable aleatoria tome el valor n, en el tiempo t, como sigue:

$$P_n(t) = P(X_t = n), n = 0,1,2, \ldots$$

La variable aleatoria debe satisfacer ciertas condiciones a fin de poder concluir que obedece a un modelo de Poisson. Algunas de estas consideraciones deben ser aceptadas como ciertas a fin de poder formular un modelo de Poisson; la rigurosidad matemática podrá ser consultada en textos de probabilidad como los Meyer (1992) y Canavos (1988):

A_1. Los valores que puede tomar la variable aleatoria en intervalos de tiempo no solapados, son considerados variables aleatorias independientes.

Dos o más eventos A y B son independientes si $P(A \cap B) = P(A)P(B)$, resultado que se puede generalizar para dos o más eventos independientes.

A_2. Sea X_t, la variable aleatoria definida como antes, y sea Y_t, otra variable aleatoria que define el número de eventos que se dan en el intervalo $(t_1, t_1 + t)$ para cualquier $t_1 > 0$. Las variables aleatorias X_t y Y_t, están ambas definidas por la misma distribución de probabilidad, esto se conoce mejor como el *valor de probabilidad* y depende solo del tamaño del intervalo en que se considere y no de los puntos extremos del intervalo. Matemáticamente se dice que la función de probabilidad que define a los eventos que se dan en el intervalo de interés tiene incrementos independientes estacionarios.

A_3. Sea $\lambda > 0$ una constante dada, si $p_1(\Delta t)$, es aproximadamente igual a $\lambda\Delta t$, donde Δt es suficientemente pequeño. Lo anterior se expresa así: $p_1(\Delta t) \sim \lambda\Delta t$. Lo anterior significa que si el intervalo es lo suficientemente pequeño, la probabilidad asociada a la ocurrencia de un evento es directamente proporcional a dicha longitud del intervalo.

A_4. La probabilidad de que se dé la realización de dos o más eventos en un instante despreciable de tiempo es prácticamente nula, esto está dado por la expresión $\sum_{k=2}^{\infty} p_k(\Delta t) \sim 0$. Lo anterior implica que $p_k(\Delta t) \rightarrow 0$, *para* $k \geq 2$.

A_5. Se introduce una condición inicial para terminar de caracterizar el modelo; este requiere que $X_0 = 0$, o mejor aún, de manera equivalente, $P_0(0) = 1$.

Un experimento con las características anteriores define un proceso de Poisson (Meyer, 1992).

5. Relación entre las distribuciones de Poisson y exponencial negativa

Las aplicaciones más importantes de la distribución exponencial negativa se presentan en situaciones propias de la teoría de colas (Walpole, 1992). Cuando se habló de la distribución de Poisson, se utilizó para modelar eventos discretos que ocurren a una rapidez constante; también es una variable aleatoria el tiempo que transcurre entre la ocurrencia de estos eventos. A continuación se muestra cómo a partir de la distribución de Poisson se obtiene la distribución de probabilidad de esta segunda variable de interés.

Haciendo uso de la distribución de Poisson, la probabilidad de que no ocurra algún evento en el periodo hasta el tiempo *t*, está dada por:

$$P(k=0) = \frac{e^{-(\lambda t)}(\lambda t)^0}{0!} = e^{-(\lambda t)}$$ (Walpole, 1992)

Denótese por *X*, el tiempo para que se dé el primer evento de Poisson.

La probabilidad de que no ocurra un evento de Poisson es la misma que la duración del tiempo hasta que el primer evento exceda x, luego haciendo uso del resultado anterior es posible obtener la distribución de probabilidad de la variable aleatoria *X*.

$P(X \geq x) = e^{-(\lambda x)}$ Hipótesis anterior

$P(X < x) = 1 - e^{-(\lambda x)}$ Probabilidad del complemento

$F(x) = 1 - e^{-(\lambda x)}$ Definición de función de distribución acumulada

$f(x) = -(-\lambda)e^{-(\lambda x)} = \lambda e^{-(\lambda x)}$ Derivando respecto a *x*, se obtiene la distribución exponencial negativa

Se concluye que la distribución de probabilidad que mide los tiempos de llegadas de eventos discretos de Poisson es una distribución exponencial negativa.

6. Teoría de colas

La teoría de colas es muy común en el quehacer diario, al contrario de lo que se puede pensar (Hillier, 2010). ¿Quién no ha estado en una cola, en espera de ser atendido en situaciones como las siguientes?: pagar en una caja registradora de un supermercado, ser atendido en una central de urgencias de un hospital, comprar una boleta de entrada al cine o realizar una consignación en una central bancaria.

Además de los ejemplos anteriores, en los que se ven involucradas personas, es posible observar otras situaciones en las que las colas son conformadas por objetos: aviones que esperan despegar de un aeropuerto, máquinas averiadas que esperan ser reparadas en un almacén de una fábrica, barcos en espera de zarpar de un puerto o llamadas en espera de ser atendidas por parte de una central telefónica de servicio al cliente (Hillier, 2010). Las colas se encuentran en nuestro entorno y todas causan alguna molestia a la que no quisiéramos estar expuestos, pero se quiera o no se debe hacer parte de ellas en algún momento. Por lo tanto, es relevante idear mecanismos que permitan disminuir los tiempos de espera (Taha, 2004).

Dada una instalación de servicio, el hecho de ser impredecible la llegada de algún cliente o su tiempo de permanencia en ella hace de la situación un experimento aleatorio (fenómeno). El interés principal al estudiar fenómenos de espera consiste en estimar el tiempo que tarda en ser atendido un cliente en la cola, para este caso podría también interesar el porcentaje de tiempo que dura una instalación de servicio en desuso.

De esta forma, la teoría de colas aborda el estudio en distintos escenarios, utilizando modelos para mostrar una abstracción del experimento real; cada modelo lleva consigo algún tipo de fórmula que indica el desempeño del sistema en cuestión y la cantidad promedio de tiempo de espera bajo diversas circunstancias (Hillier, 2010).

6.1. Estructura de los modelos de colas

Los elementos principales de un modelo de colas son el cliente y el servidor, entre quienes se establece una relación con el fin de obtener el tiempo que necesita el cliente para completar su servicio. El modelo básico de un sistema de colas se caracteriza por (Hillier, 2010):

1. Cuando los clientes llegan al sistema y no es posible conocer en qué cantidad lo harán, ni tampoco en qué momento, la fuente de entrada se comporta de forma estocástica (probabilística), sin embargo, pueden existir modelos de colas en los que se conozca el número de llegadas de forma previa.
2. Los clientes pueden ser atendidos al ingresar al sistema si hay algún servidor libre, en caso contrario el cliente se incorpora a una cola de servicio en espera de ser atendido por algún servidor que se desocupe, conforme a un mecanismo de selección establecido para su atención.
3. Los clientes son atendidos y salen del sistema; usualmente no es posible conocer de antemano el tiempo que toma el servicio (Hillier, 2010)

6.2. Fuente de entrada

La fuente de entrada, también conocida como la población potencial que puede hacer parte de la cola, es en esencia la cantidad de clientes que pueden requerir el servicio; así, por ejemplo, los habitantes de un municipio son clientes potenciales de la sala de urgencia del hospital local aun cuando no hagan uso de este servicio. El tamaño de esta población de entrada bien puede ser finito o infinito, de esta forma, la fuente de entrada también lo podrá ser; por planteamientos teóricos suele considerarse una fuente de entrada infinita, a fin de poder hacer más fácil el manejo analítico del modelo; de igual forma, de no hacerse comentario alguno acerca del tamaño de la fuente de entrada ha de suponerse también su tamaño como infinito (Hillier, 2010).

El acceso al sistema por parte de los clientes permite establecer los tiempos entre llegadas, mientras que el tiempo de servicio en el sistema obedece a la suma del tiempo de permanencia en la cola y del tiempo que tomó el servicio en el servidor por el cual fue atendido el cliente (Taha, 2004).

Usualmente, el patrón estadístico que rige la generación de los clientes en el tiempo obedece a un proceso de Poisson. En este caso el conteo de clientes que arriban al sistema está modelado por la distribución de Poisson, y el intervalo de tiempo entre las llegadas sucesivas de clientes se rige por la distribución exponencial negativa.

6.3. Cola

La cola del sistema está formada por los clientes que esperan recibir el servicio cuando los servidores se encuentran ocupados, puede ser finita o infinita. Lo deseable sería que fuese finita, sin embargo en la mayoría de los modelos de colas se considera infinita, pero en caso de tener una cola relativamente pequeña se puede considerar como finita.

Una situación que corresponde a una cola finita puede observarse cuando un cliente que arriba a un supermercado, en vista de la congestión de las filas en las cajas registradoras, decide abandonar el supermercado; este es un sistema de colas limitado. Por el contrario, a las colas de índole infinita suele aplicárseles la disciplina de la cola, es decir, en algún momento el cliente será atendido según el mecanismo de selección implementado (Hillier, 2010).

6.4. Disciplina de la cola

También conocida como disciplina del servicio, representa el orden en el cual se seleccionan los clientes de la cola para recibir el servicio. Existen distintos métodos para seleccionar los miembros de la cola. Usualmente la disciplina más utilizada consiste en que el primero que llega es el primero que se sirve (FIFO, PLPS o FCFS —*first come, first served*—). Otra disciplina usada consiste en que el último que llega es el primero que se sirve (LIFO, LCFS —*last come, first served*—); además se encuentran colas de servicio que atienden en un orden aleatorio (SIRO —*service in ramdom order*—) y las denominadas de disciplina general (GD), que obedecen, como su nombre lo indica, a un orden de atención particular de cada sistema (Taha, 2004).

Una característica de los clientes humanos respecto a una cola es la de poder tomar decisiones, como por ejemplo cambiarse de una cola a otra a fin de reducir su tiempo de espera, como también rehusar a hacerla debido al largo tiempo de permanencia previsible.

6.5. Proceso de servicio

Esta parte del proceso describe la manera como son atendidos los clientes en el sistema, que puede contar con una o más estaciones de servicio (servidores). Cuando hay más de un servidor es posible atender a tantos clientes de forma simultánea como servidores hay en el sistema; un caso de aplicación es el de una central de pago de servicios públicos donde se cuenta con varios cajeros (Taha, 2004).

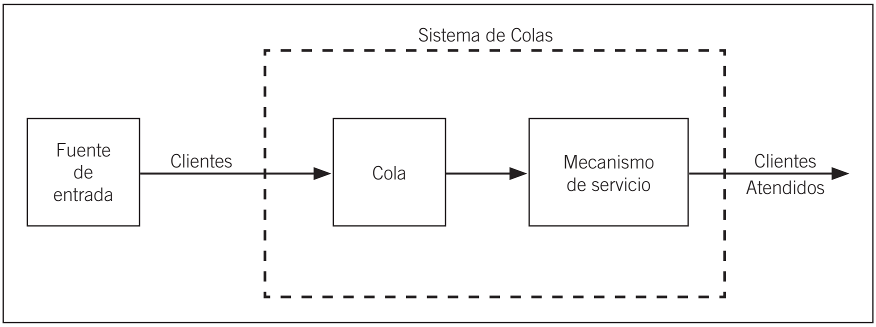
Cuando la instalación tiene varios servidores que ofrecen el mismo servicio se denomina sistema de servidores paralelos; por el contrario, cuando los servidores se disponen en serie, de forma tal que el cliente debe pasar obligatoriamente por cada uno de ellos para completar el servicio, se le conoce como un sistema de cola en serie (Taha, 2004). El modelo más elemental de una cola está conformado por un servidor o un número finito de servidores y una sola cola (Hillier, 2010).

El tiempo transcurrido desde el inicio del servicio a un cliente y su culminación recibe el nombre de tiempo de servicio (también el de duración del servicio). El modelo de colas debe definir con claridad la distribución de probabilidad que modela los tiempos de servicio, tanto para cada servidor como para cada tipo distinto de cliente en caso de ser necesario. Es común que todos los servidores utilicen la misma distribución de probabilidad (Hillier, 2010); en la práctica la distribución de probabilidad que más se utiliza para describir los tiempos de servicio es la distribución exponencial negativa, aunque también son utilizadas otras distribuciones (Hillier, 2010).

Algunas consideraciones generales acerca del servicio son las siguientes:

1. Tanto la llegada al sistema como la prestación del servicio pueden obedecer a una distribución determinística (programada) o estocástica (probabilística); por lo general, en el tratamiento de las colas se asocian estos eventos a una distribución del tipo estocástico.
2. Cada servidor tiene su respectiva cola, o existe una cola única para todos los servidores.
3. El servicio se presta de forma individual o grupal.

Gráfico 1 . Proceso básico de colas



Fuente: Hillier (2010).

6.6. Estado estable

Definido para los sistemas de colas en los que los flujos de atención, con el paso del tiempo, permiten alcanzar un estado que no depende del verificado al inicio de la operación del sistema; de esta forma, el estado estable de una cola es obtenido una vez ha pasado un tiempo suficientemente largo desde el inicio de las operaciones, que no es afectado por el estado de inicio (Hillier, 2010).

6.7. Estado transitorio

El estado de un sistema de colas antes de alcanzar el estado estable (Winston, 2005).

7. Notación usada en la teoría de colas

En 1951 Kendall diseñó una notación que aún se acepta para la representación de un sistema de una cola (Winston, 2004). Antes de continuar, es conveniente caracterizar algunos parámetros establecidos por esta notación:

M = Define una variable aleatoria que se distribuye de forma exponencial, bien sea para los tiempos de llegada o los tiempos de servicio.
D = Los tiempos, ya sean los de llegada o los de servicio, son de tipo determinístico.
E_k = Los tiempos de llegada o de servicio están definidos por una distribución de Erlang de parámetro k.
G = Los tiempos de llegada o de servicio están definidos por alguna distribución general.

Este sistema queda completamente determinado por el siguiente conjunto de seis características:

1. Se ha de especificar la naturaleza del proceso de arribo de los clientes al sistema.
2. Se especifica la distribución de las salidas (naturaleza de los tiempos de servicio).
3. Cantidad de servidores en paralelo.
4. Disciplina de la cola.
5. Esta característica especifica la cantidad máxima (finita o infinita) de clientes en el sistema (incluidos los clientes de la cola y el servicio).
6. Tamaño de la fuente: la población se considera infinita, a menos que los clientes potenciales igualen en número a la cantidad de servidores (Winston, 2004).

Para el tratamiento de procesos de colas se presenta a continuación un conjunto de parámetros aceptados de forma general (Hillier, 2010):

$N(t)$ = Define para el tiempo t, $t \geq 0$ número de clientes en el sistema de colas.

$P_n(t)$ = Probabilidad de que en el tiempo t, existan n clientes en el sistema.
S = Cantidad de servidores en el sistema de colas.
λ_n = Frecuencia media de llegada cuando hay n clientes en el sistema, en caso de ser λ_n constante para toda n, se nota λ.
μ_n = Frecuencia media de salida del servicio cuando hay n clientes en el sistema; de igual forma, cuando esta frecuencia es constante para cualquier $n \geq 1$, se nota μ.

Usualmente la teoría de colas tiende a dedicar gran parte de su análisis a los estados de condición estable, esto debido en gran medida a la dificultad de los estados transitorios. El siguiente conjunto de parámetros se establecen para sistemas de colas que han alcanzado un estado estable (Taha, 2004):

P_n = Probabilidad de que se encuentren exactamente n clientes en el sistema.
$L = \sum_{n=0}^{\infty} nP_n$ Cantidad de clientes esperados en el sistema.

$L_q = \sum_{n=s}^{\infty}(n-s)P_n$ Número promedio de clientes en la cola (se excluyen los clientes que están en servicio).
W = Tiempo esperado en el sistema (incluye el tiempo de servicio) para cada cliente.
W_q = Tiempo esperado en la cola (se excluye el tiempo de servicio) para cada cliente.

John D. Little, en 1961, presentó las siguientes ecuaciones, que relacionan la longitud de la cola con el tiempo de espera:

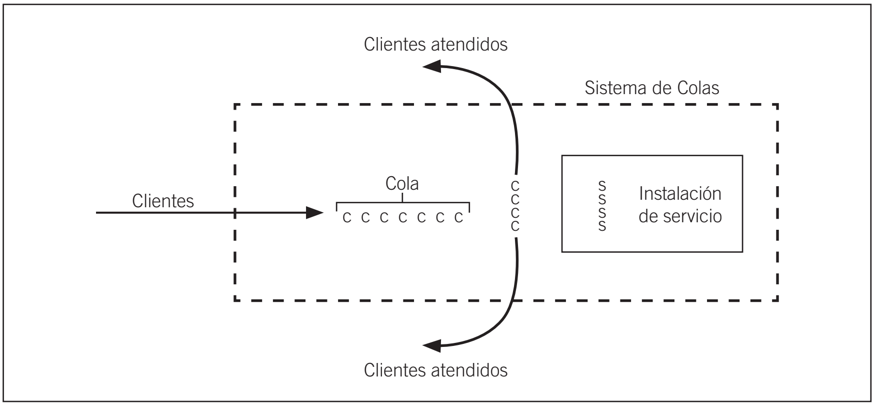
$L = \lambda W$

$L_q = \lambda W_q$, (Hillier, 2010).

7.1. Procesos de nacimiento y muerte

El proceso de nacimiento se define a partir de la llegada de los clientes al sistema; el de muerte se evidencia con la salida del cliente del sistema tras ser servido. La mayor parte de los modelos elementales de colas suponen que las entradas (llegadas de clientes) y las salidas (clientes que se van) del sistema ocurren de acuerdo al proceso de nacimiento y muerte (Hillier, 2010). Un proceso de nacimiento y muerte se encuentra entre los procesos estocásticos de tiempo continuo, donde siempre el estado del sistema está dado por un entero no negativo (Winston, 2004).

Gráfico 2. Sistema elemental de colas



Fuente: Hillier (2010).

Se denota en el tiempo t, $t \geq 0$ por $N(t)$, el número de clientes que se encuentran en el sistema. Este proceso de nacimiento y muerte se da en términos probabilísticos, la variación de $N(t)$ con respecto al tiempo. Siendo más precisos, los nacimientos y muertes ocurren de manera aleatoria y sus tasas de ocurrencias dependen del actual estado del sistema (Hillier, 2010).

Un proceso de nacimiento y muerte está regido por las leyes siguientes (Winston, 2005):

1. Cuando $N(t) = n$, la distribución exponencial negativa de parámetro λ_n, $n = 0,1,2,\ldots$, representa la probabilidad actual del tiempo que falta para el próximo nacimiento.
2. Cuando $N(t) = n$, la distribución exponencial negativa de parámetro μ_n, $n = 1,2,\ldots$, representa la probabilidad actual del tiempo que falta hasta la próxima muerte.
3. Los nacimientos y las muertes son independientes, esto implica que las anteriores variables que definen los procesos nacimiento y muerte también lo son.

Según cuál de las dos variables aleatorias sea la más pequeña se tiene la siguiente transición del proceso (cambios de estado):

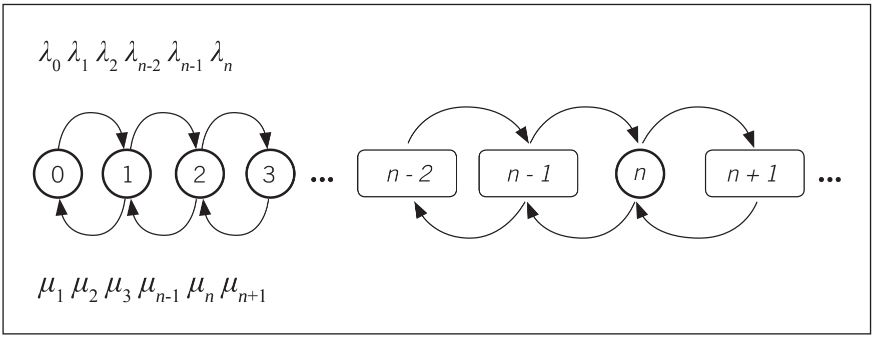
$n \rightarrow n + 1$, cuando se da un nacimiento, ingreso de un cliente al sistema.

$n \rightarrow n - 1$, cuando una muerte, salida de un cliente del sistema.

En la teoría de colas λ_n *y* μ_n son, respectivamente, la tasa media de llegada y la tasa media de terminaciones del servicio, para *n* clientes en el sistema. Lo anterior implica que cualquier proceso de nacimiento y muerte queda completamente determinado al conocerse λ_n *y* μ_n; además, como todos los estados son no negativos, se tiene que $\mu_0 = 0$ (Winston, 2005).

Siendo más específicos, los procesos de nacimiento y muerte encajan dentro de las cadenas de Markov de tiempo continuo, son fundamentales para tener un entendimiento de los procesos de estado estable (Taha, 2004).

Gráfico 3. Diagrama de tasas para el proceso de nacimiento y muerte



Fuente: Taha (2004).

7.1.1. Ecuaciones de balance de un proceso de nacimiento y muerte

Se obtienen a partir del principio tasa de entrada = tasa de salida, lo anterior indica que para un estado *n* del sistema las tasas medias de entrada y salida son iguales; se llama ecuación de balance para el estado *n* a la ecuación que manifiesta este principio (Hillier, 2010).

Tabla 1. Ecuaciones de balance de un proceso de nacimiento y muerte

Estado	Tasa de entrada = tasa de llegada
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$
:	:
n - 1	$\lambda_{n-2} P_{n-2} + \lambda_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \lambda_{n+1} \mu_{n+1} = (\lambda_n + \mu_n) P_n$

Fuente: Hillier (2010).

7.1.2. Probabilidad de estado estable

Es la probabilidad de encontrar el proceso en un estado *j* luego de un número suficientemente grande de transiciones (cambio de estado). Siendo más precisos, cuando *n* tiende a infinito esta probabilidad tiende a π_j (Hillier, 2010).

Utilizando las ecuaciones de balance se obtiene la probabilidad de estado estable de un sistema de colas (Taha, 2004); a continuación se realiza su deducción haciendo uso también de las ecuaciones de Little:

1. De manera recursiva, en las ecuaciones de balance es posible caracterizar las siguientes probabilidades de estado estable:

$P_1 = \frac{\lambda P_0}{\mu}, P_2 = \frac{\lambda^2 P_0}{\mu^2}, \ldots P_j = \frac{\lambda^j P_0}{\mu^j}$, en adelante se conocerá esta expresión como (1).

2. Se define la intensidad de tráfico como:

   $$\rho = \frac{\lambda}{\mu}$$

3. Para simplificar la notación, sea $C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n \mu_{n-1}\cdots\mu_1}$, para $n = 1,2,\ldots,$

4. En la ecuación $\sum_{n=0}^{\infty} P_n = 1$ se sustituye (1) y se obtiene la probabilidad de estado estable, donde $P_n = P_0 c_n$, y se obtiene el siguiente resultado:

   $P_0(1 + p + p^2 + \ldots) = 1$

$$P_0 \sum_{r=0}^{\infty} \rho^r = 1$$

$$\frac{P_0}{1-\rho} = 1 \; si |\rho| < 1$$

$$P_0 = 1 - p$$

5. Ahora el anterior resultado se sustituye en la ecuación de estado estable (1), y se obtiene el siguiente resultado:

$$P_n = \frac{\lambda^n P_0}{\mu^n}$$

$$P_n = P_o c_n$$

$$P_n = (1 - p)p^n$$

Para asegurar la convergencia de la serie geométrica $\sum_{r=0}^{\infty} \rho^r$ es necesario que $0 \leq p < 1$; si $p \geq 1$, la distribución de estados será de tipo estable, esto es $\lambda \geq \mu$, se tendría una tasa de llegada mayor o igual a la tasa de servicio, la cola aumenta sin límite; en otras palabras, no se podría conseguir el estado estable (Winston, 2004).

6. Obtención de la cantidad de clientes esperados en el sistema L. A fin de poder manipular analíticamente los resultados siguientes se supondrá que se alcanzó el estado estable donde $|p| < 1$; en todos los apartados se utilizan las ecuaciones de estado estable.

$$L = \sum_{n=0}^{\infty} nP_n$$

$$L = \sum_{n=0}^{\infty} n\,(1-\rho)\rho^n$$

$$L = (1-\rho)\rho \sum_{n=0}^{\infty} n\rho^{n-1}$$

$$L = (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n)$$

$$L = (1-\rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right)$$

$$L = (1-\rho)\rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$L = (1-\rho)\rho \frac{1}{(1-\rho)^2}$$

$$L = \frac{\rho}{1-\rho}$$

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

$$L = \frac{\frac{\lambda}{\mu}}{\frac{\mu-\lambda}{\mu}}$$

$$L = \frac{\lambda}{\mu-\lambda}$$

7. De forma análoga se deduce el resultado para el número promedio de clientes en el sistema L_q. Se observa que si se encuentran 0 o 1 cliente en el sistema, nadie se encuentra presente en la cola; por el contrario, si n personas se encuentran presentes ($n \geq 1$), hay $n - 1$ en la Cola (Winston, 2004).

   $L_q = \sum_{n=s}^{\infty}(n-s)P_n$ Para nuestro caso $s = 1$, por lo tanto

   $$L_q = \sum_{n=s}^{\infty}(n-1)P_n$$

   $$L_q = \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n$$

   $$L_q = \frac{\lambda}{\mu-\lambda} - (1-P_0)$$

   $$L_q = \frac{\lambda}{\mu-\lambda} - \rho$$

   $$L_q = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu}$$

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

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

8. De acuerdo a las fórmulas de Little se tiene que:

   $W = \frac{L}{\lambda}$ y $W_q = \frac{L_q}{\lambda}$ Por lo tanto

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

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

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

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

Los anteriores resultados y los otros que se deducen o enuncian en otros modelos obedecen al supuesto de haber alcanzado la condición de estado estable.

8. Modelos elementales de colas

8.1. Modelo *M/M/1*

Según la notación presentada anteriormente, define un modelo de colas en el que la distribución de probabilidad de los tiempos de llegada y de salida es de tipo exponencial negativa, con un servidor. Para este modelo se abordarán a continuación dos casos particulares, el primero para una cola infinita y el segundo para una finita.

8.1.1. Modelo *M/M/1/G/∞/∞*

Supone que la tasa de llegada por unidad de tiempo es λ y la de servicio es μ, con un tiempo de servicio que se distribuye exponencialmente para un solo servidor.

En las ecuaciones de balance presentadas anteriormente los parámetros λ_n y μ_n están definidos como se muestra a continuación; las tasas de llegada y de servicio son constantes, es decir, siempre se espera la misma cantidad de clientes de llegada en un intervalo de tiempo específico, como también el número de clientes servidos en algún otro intervalo de tiempo.

$\lambda_n = \lambda,\ n = 0,1,2,3,\ldots,$

$\mu_0 = 0,$

$\mu_n = \mu,\ n = 1,2,3,\ldots,$ es decir se supone a λ_n y μ_n constantes para los estados indicados.

8.1.2. Aplicación del modelo *M/M/1/G/∞/∞* en una central de urgencias

La central de urgencias del puesto de salud de un pequeño municipio opera con un solo médico de turno. Con base en datos históricos de la operación y funcionamiento del sistema de atención se han establecido los siguientes parámetros que permiten adoptar el modelo *M/M/1/G/∞/∞* expuesto anteriormente:

1. La tasa media de llegada es de cuatro pacientes por hora, es decir, $\lambda = 4$.
2. La llegada de los pacientes obedece a una distribución de probabilidad Poisson de acuerdo a los siguientes argumentos:

    - La llegada de los pacientes a la central de urgencias es aleatoria, es decir, no es posible determinar de antemano cuántos pacientes llegarán ni en qué orden.
    - Las llegadas de los pacientes son independientes entre sí. Técnicamente, los motivos por los cuales un paciente visita la central de urgencias no están influenciados por los relacionados con otro paciente; es importante aclarar que en este caso no se consideran situaciones de enfermedades pandémicas.
    - La llegada de dos o más pacientes de forma simultánea es prácticamente nula.

3. La tasa media de los tiempos de servicio es de 10 minutos, $\mu = 10$.
4. Para los tiempos de servicio se ha establecido que obedecen a una distribución exponencial negativa, es decir que la cantidad de pacientes atendidos en un intervalo de tiempo no influye en la cantidad de pacientes atendidos en algún otro intervalo, característica conocida como falta de memoria de la distribución exponencial negativa.
5. La central de urgencias, por disponer de un solo médico para la atención de los pacientes, es un sistema con un único servidor.
6. El mecanismo de selección para la atención de los pacientes obedece a un sistema FIFO, PLPS o FCFS —*first come, first served*—; sin embargo, para los pacientes que requieren atención inmediata por su estado de salud se aplica un mecanismo de selección general.
7. Todos los habitantes del municipio son clientes potenciales, esto permite considerar una fuente de tamaño infinito. Se espera que todo aquel paciente que llega a la central pueda ser atendido, situación que determina un sistema de cola infinita.

De acuerdo a los argumentos revisados y a la notación presentada anteriormente, el modelo que caracteriza a la central de urgencias es del tipo $M/M/1/G/\infty/\infty$, lo que se resume en las siguientes características: los tiempos de llegada, como los de servicio, son de tipo exponencial; existe un solo servidor (médico); el mecanismo de selección para la prestación del servicio consiste en que el primero en llegar es el primero en ser servido, de no presentarse una situación de extrema urgencia; y la fuente de entrada se puede considerar de tamaño infinito (población potencial que puede requerir el servicio), es decir, cola de tamaño infinito.

La solución del modelo propuesto contempla entonces las siguientes actividades:

1. Definir las constantes λ_n y μ_n, con valores expresados en las mismas unidades. En este caso la unidad de tiempo utilizada es de horas, sin embargo podría convertirse también a minutos:

   $\mu = \frac{1}{10\ min} = \frac{1}{10\ min}\frac{60\ min}{1\ h} = \frac{6}{h}$, seis clientes por hora.

2. En una situación como la planteada para la central de urgencias pueden calcularse y son de interés particular algunos parámetros como:

   - L, cantidad esperada de pacientes en la central de urgencias. Se utiliza para su cálculo una expresión que involucra a p, este valor —recuérdese— corresponde al factor de utilización del servicio que esta dado por: $\rho = \frac{\lambda}{\mu}$

     $$L = \frac{\rho}{1-\rho}$$

     $L = \frac{\frac{2}{3}}{1-\frac{2}{3}} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$, por lo tanto el número esperado de clientes en el sistema es de 2 personas.

   - W, tiempo esperado en el sistema. Una vez encontrado el parámetro L, es muy sencillo a partir de una de las fórmulas de Little, $L = \lambda W$, hallar el parámetro $W = \frac{L}{\lambda}$ que define el tiempo esperado en el servicio en el sistema

     $$W = \frac{2}{4} = 0{,}5\ h$$

- W_q, tiempo esperado en la cola. Es importante notar que al transcurrir una cantidad de tiempo el sistema alcanza la condición de estado estable, esto debido a que su factor de utilización $\rho = \frac{\lambda}{\mu} = \frac{4}{6} < 1$, haciendo uso de las ecuaciones de estado estable y de las fórmulas de Little. El cálculo de este parámetro resulta inmediato a partir de la ecuación $W = W_q + \frac{1}{\mu}$

  $W_q = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}h$, en promedio se permanece 33,33 minutos en la cola.

- L_q, cantidad esperada de pacientes en la cola. Por último interesa L_q, el número promedio de clientes en la cola. Haciendo uso de la ecuación de Little se obtiene el resultado pedido.

  $L_q = \lambda W_q$

  $L_q = 4\frac{1}{3} = \frac{4}{3}$, el número esperado de clientes en la cola es 1,33 personas.

- Una cuestión de interés podría ser encontrar la probabilidad de que un paciente sea atendido de inmediato, es decir que encuentre la sala desocupada al llegar un paciente a la central de urgencias. El cálculo de este parámetro se realiza a través de:

  $P_0 = 1 - \rho$

  $P_0 = 1 - \frac{\lambda}{\mu}$

  $P_0 = 1 - \frac{2}{3}$

  $P_0 = \frac{1}{3} = 33{,}33\%$

8.1.3. Modelo *M/M/1/G/K/∞*

Siguiendo con el modelo *M/M/1*, se estudia ahora la situación de una cola limitada, es decir, un modelo del tipo *M/M/1/G/K/∞*, este es un caso particular del modelo *M/M/1/G/∞/∞* con capacidad para K clientes. Este modelo tiene la particularidad de negar el servicio al siguiente cliente $K + 1$ cuando hay K clientes en el sistema; en este caso, el máximo tamaño alcanzado por la cola es $K - 1$ clientes, al suponer que hay un cliente que en el momento es atendido.

Para todos los modelos M/M/1 se supone que la tasa de llegada por unidad de tiempo es λ y la de servicio es μ, con un tiempo de servicio que se distribuye exponencialmente para un solo servidor. De acuerdo a las ecuaciones de balance presentadas anteriormente, los parámetros λ_n y μ_n están definidos como se muestra a continuación. Las tasas de llegada y de servicio son constantes, es decir que siempre se espera la misma cantidad de clientes de llegada en un intervalo de tiempo específico, como también el número de clientes servidos en algún otro intervalo de tiempo. Para este caso específico:

$\lambda_n = \begin{cases} \lambda, si\ n = 0,1,\ldots,K-1 \\ 0, si\ n \geq K \end{cases}$, como el tamaño máximo de clientes en el sistema es K, no se aceptan más clientes (tasa de llegada nula)

$\lambda_k = 0$, la tasa de nacimiento se hace cero cuando se alcanza el estado K

$\mu_0 = 0$

$\mu_n = \mu,\ n = 1,2,\ldots, K$

Las ecuaciones de estado estable para este modelo adoptan la siguiente forma:

- $C_n = \begin{cases} \rho^n\ si\ n = 1,2,\ldots,K \\ 0\ si\ n > K \end{cases}$, como λ_n y μ_n son constantes para los valores anteriormente señalados, y sabiendo que $\rho = \frac{\lambda}{\mu}$, el parámetro C_n adopta la forma propuesta.

- $P_0 = \frac{1-\rho}{1-\rho^{K+1}}$

A continuación se deduce P_0:

Se conoce la ecuación de estado estable $P_0 = (\sum_{n=0}^{\infty} \rho^n)^{-1}$ cuando la cantidad de clientes en el sistema es infinita, para este caso la suma se da hasta el estado K, por lo tanto:

$P_0 = (\sum_{n=0}^{K} \rho^n)^{-1}$, $\sum_{n=0}^{K} \rho^n$ es la suma finita de una progresión geométrica

$$P_0 = \left(\frac{1-\rho^{K+1}}{1-\rho}\right)^{-1}$$

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

P_0 se calculó haciendo uso de una progresión geométrica finita. El resultado obtenido anteriormente es consecuencia de la siguiente formulación, donde $x = p$ y $N = K$:

$$\sum_{n=0}^{N} x^n = \frac{1-x^{N+1}}{1-x}$$

- De igual forma se sabe que:

  $P_n = P_0 p^n$, *si* $n = 0,1,2,\ldots$, en el caso de una cola de tamaño infinito.

Para el modelo actual *M/M/1/G/K/∞*, la ecuación anterior se define como:

$$P_n = \begin{cases} P_0\rho^n \ si\ n = 1,2,\ldots K \\ 0, si\ n = K+1, K+2,\ldots \end{cases}$$

Y adopta la forma siguiente, reemplazando $P_0 = \frac{1-\rho}{1-\rho^{K+1}}$ en la expresión anterior.

$$P_n = \rho^n \left(\frac{1-\rho}{1-\rho^{K+1}}\right) si\ n = 0,1,2,\ldots,K$$

- La ecuación de estado estable para cola infinita *M/M/1/G/∞/∞* es $L = \sum_{n=0}^{\infty} nP_n$; en el modelo actual la suma se hace hasta el estado K, por tanto se tiene:

$$L = \sum_{n=0}^{K} nP_n$$

$$L = \sum_{n=0}^{K} n\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\rho^n$$

$$L = \sum_{n=0}^{K} n\rho\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\rho^{n-1}$$

$$L = \rho\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\sum_{n=0}^{K} n\rho^{n-1}$$

$$L = \rho\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\sum_{n=0}^{K} \frac{d}{d\rho}(\rho^n)$$

$$L = \rho\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\frac{d}{d\rho}\left(\sum_{n=0}^{K} \rho^n\right)$$

$$L = \rho\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\frac{d}{d\rho}\left(\frac{1-\rho^{K+1}}{1-\rho}\right)$$

$$L = \rho\left(\frac{1-\rho}{1-\rho^{K+1}}\right)\left(\frac{-(K+1)\rho^K(1-\rho)+(1-\rho^{K+1})}{(1-\rho)^2}\right)$$

$$L = \frac{\rho\left[1-(K+1)\rho^K+K\rho^{K+1}\right]}{(1-\rho^{K+1})(1-\rho)}$$

Los anteriores resultados son válidos cuando $p \neq 1$, o mejor aún, cuando $\lambda \neq \mu$.

- Cuando $p = 1$, las ecuaciones de estado estable son las siguientes:

  La ecuación $P_n = p^n P_0$, se convierte en $P_n = P_0$ dado que $p^n = 1^n = 1$

- Utilizando la relación $\sum_{n=0}^{K} P_n = 1$, el parámetro P_o se convierte en $P_o = \frac{1}{K+1}$ por:

  $$\sum_{n=0}^{K} P_0 = 1$$

  $$\underbrace{(P_o + P_0 + \cdots + P_0)}_{K+1 \text{ veces}} = 1$$

  $$(K+1)P_o = 1$$

  O de manera equivalente:

  $P_n = \frac{1}{K+1}$, dado que $P_n - P_0$

A continuación se hace la deducción de los valores que adoptan los parámetros L, L_q, W y W_q para el modelo $M/M/1/G/K/\infty$:

1. Obtención de la cantidad de clientes esperados en el sistema L.

Para este modelo se tiene la siguiente ecuación de estado estable $L = \sum_{n=0}^{K} nP_n$ por lo tanto

$$L = \sum_{n=0}^{K} n\left(\frac{1}{K+1}\right)$$

$$L = \frac{1}{K+1}\sum_{n=0}^{K} n$$

$$L = \frac{1}{K+1}\left(\frac{K(K+1)}{2}\right)$$

$$L = \frac{K}{2}$$

Aquí se ha hecho uso de:

$\sum_{x=1}^{N} x = \frac{N(N+1)}{2}$, suma de los primeros n naturales.

2. De forma análoga se deduce el resultado para el número promedio de clientes en el sistema L_q. Se observa que si se encuentran 0 o 1 cliente en el sistema, nadie se encuentra presente en la cola; por el contrario, si K personas se encuentran presentes ($K \geq 1$), hay K - 1 en la cola (Winston, 2004).

Para la deducción de L_q, se procede como sigue:

$$L_q = \sum_{n=1}^{K}(n-1)\,P_n$$

$$L_q = \sum_{n=1}^{K} nP_n - \sum_{n=1}^{K} P_n$$

$$L_q = L - \left[\left(\sum_{n=0}^{K} P_n\right) - P_0\right]$$

$$L_q = L - [1 - P_0]$$

3. En este caso las ecuaciones $L = \lambda W$ y $L_q = \lambda W_q$ no son aplicables por el hecho de no ser λ_n constante para cualquier n (Hillier, 2010). Para subsanar este inconveniente se suele sustituir λ_n por $\bar{\lambda}$, donde:

$$\bar{\lambda} = \sum_{n=0}^{K-1} \lambda P_n$$

$$\bar{\lambda} = \lambda \sum_{n=0}^{K-1} P_n$$

$$\bar{\lambda} = \lambda[(\sum_{n=0}^{K} P_n) - P_K]$$

$$\bar{\lambda} = \lambda(1 - P_K)$$

4. Reemplazando λ_n por $\bar{\lambda}$ en las fórmulas de Little se tiene:

$$W = \frac{L}{\bar{\lambda}}, W_q = \frac{L_q}{\bar{\lambda}}$$

8.2. Modelo *M/M/s*

Según la notación presentada anteriormente, define un modelo de colas donde tanto la distribución de probabilidad de los clientes de llegada como la de los clientes de salida son de tipo exponencial, con s servidores en paralelo, con tasa de llegadas λ y tasa de servicio μ y una sola cola de clientes que esperan ser atendidos en alguno de los s servidores. Si están presentes $n \leq s$ clientes en el sistema, todos los n clientes se encuentran en atención; por el contrario, si hay presentes $n > s$ clientes en el sistema, se encuentran $n - s$ clientes en la cola de servicio (Winston, 2004).

8.2.1. Modelo *M/M/s/G/∞/∞*

Para poder analizar el modelo *M/M/s/G/∞/∞* a través del proceso de nacimiento y muerte, conviene analizar y definir los siguientes parámetros:

$\lambda_n = \lambda$ $si n = 0,1,2, \ldots,$

Ahora bien, si los s servidores se encuentran ocupados, los servicios están dados a una tasa de $s\mu$.

En el caso de un modelo con varios servidores las ecuaciones de estado estable en un proceso de nacimiento y muerte adoptan la siguiente forma:

$$C_n = \begin{cases} \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \; si \; n = 1,2,\dots,s \\ \dfrac{\left(\frac{\lambda}{\mu}\right)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{s!s^{n-s}} \; si \; n = s, s+1, s+2, \dots, \end{cases}$$

Ahora bien, si $\frac{\lambda}{s\mu} = \rho < 1$, las ecuaciones de estado estable son:

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

Retomando la ecuación de estado estable $P_n = C_n P_o$ se tienen los siguientes resultados:

$$P_n = \begin{cases} \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 \; si \; 0 \le n \le s \\ \dfrac{\left(\frac{\lambda}{\mu}\right)^n}{s!s^{n-s}} P_0 \; si \; n \ge s \end{cases}$$

A continuación se hace la deducción de los valores que adoptan los parámetros L, L_a, W y W_a para el modelo M/M/s/G/∞/∞:

1. Encontrar una expresión para L_a:

$L_q = \sum_{n=s}^{\infty}(n-s)P_n$, haciendo $j = n - s$, la anterior expresión adopta la siguiente forma:

$$L_q = \sum_{j=0}^{\infty} jP_{s+j}$$

$$L_q = \sum_{j=0}^{\infty} j \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} \rho^j P_0$$

$$L_q = \rho \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} P_0 \sum_{j=0}^{\infty} j\rho^{j-1}$$

$$L_q = \rho \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} P_0 \sum_{j=0}^{\infty} \frac{d}{d\rho}\rho^j$$

$$L_q = \rho \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} P_0 \frac{d}{d\rho} \sum_{j=0}^{\infty} \rho^j$$

$$L_q = \rho \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} P_0 \frac{d}{d\rho} \left(\frac{1}{1-\rho}\right)$$

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

2. Ahora es muy sencillo obtener una expresión para W_a, a partir de la anteriormente encontrada:

$$W_q = \frac{L_q}{\lambda}$$

3. El resultado que permite obtener un resultado para W, es la conocida fórmula $W = W_q + \frac{1}{\mu}$, y se sabe que $L = \lambda W$, por lo tanto:

$$L = \lambda \left(W_q + \frac{1}{\mu}\right)$$

8.2.2. Aplicación del modelo *M/M/s/G/∞/∞* en una central de urgencias

Una institución prestadora de salud (IPS) ofrece sus servicios de central de urgencias con dos sedes diferentes ubicadas en el mismo sector de influencia geográfica en la ciudad de Bogotá. Con base en la revisión de los registros de atención de urgencia de cada una de las sedes se han establecido los siguientes parámetros de funcionamiento y operación:

1. La tasa media de llegada de cada una de las dos sedes es de ocho pacientes por hora, es decir, $\lambda = 8$.
2. La llegada de los pacientes obedece a una distribución de probabilidad Poisson de acuerdo a los siguientes argumentos:

    - La llegada de los pacientes a cualquiera de las dos sedes es aleatoria, es decir, no es posible determinar de antemano cuántos pacientes llegarán a cada sede y en qué orden lo harán.

- Las llegadas de los pacientes son independientes entre sí. Técnicamente, los motivos por los cuales un paciente visita cualquiera de las dos sedes por lo general no están influenciados por los motivos relacionados con otro paciente; es importante aclarar que en este caso no se consideran situaciones de enfermedades pandémicas o eventos como accidentes de tránsito.
- La llegada de dos o más clientes de forma simultánea es prácticamente nula.

3. La tasa media de los tiempos de servicio estimados por la IPS para cada uno de los médicos es de 5 pacientes por hora, es decir, $\mu = 5$.
4. Para los tiempos de servicio se ha establecido que obedecen a una distribución exponencial negativa, es decir que la cantidad de pacientes atendidos en un intervalo de tiempo no influye en la cantidad de pacientes atendidos en otro intervalo, característica conocida como falta de memoria de la distribución exponencial negativa.
5. La sede A cuenta con dos médicos, mientras que la sede B con tres médicos. Cada uno de los médicos se considera como un servidor.
6. El mecanismo de selección para la atención de los pacientes obedece a un sistema FIFO, PLPS o FCFS —*first come, first served*—; sin embargo, para los pacientes que requieren atención inmediata por su estado de salud se aplica un mecanismo de selección general.
7. Todas las personas adscritas a la IPS son clientes potenciales de cualquiera de las dos sedes; de acuerdo con este presupuesto se puede considerar una fuente de tamaño infinito para cada sede. Se espera que todo paciente que llega cualquiera de las dos sedes pueda ser atendido; en esta situación se configura un sistema de cola infinita.

Desarrollando el ejercicio de acuerdo a parámetros establecidos se encuentra que:

Tabla 2. Parámetros de interés para colas de Poisson, sedes A y B

Parámetros		Sede A			Sede B		
		$\lambda = 8$	$\mu = 5$	$S = 2$	$\lambda = 8$	$\mu = 5$	$S = 3$
L	$L = \lambda\left(W_q + \frac{1}{\mu}\right)$ Cantidad de pacientes en la sede	4,44 Pacientes en la sede			1,9129 Pacientes en la sede		
L_a	$L_q = \frac{\rho\left(\frac{\lambda}{\mu}\right)^s P_0}{s!\,(1-\rho)^2}$ Cantidad de pacientes en la cola	2,844 Pacientes en la cola			0,312 Pacientes en la cola		
W	$W = W_q + \frac{1}{\mu}$ Tiempo de espera en de un paciente en la sede	0,555 horas			0,2391 horas		
W_a	$W_q = \frac{L_q}{\lambda}$ Tiempo de espera de un paciente en la cola	0,355 horas			0,391 horas		

Fuente: elaboración de los autores.

En vista de los resultados obtenidos acerca del funcionamiento y operación de las dos sedes, la IPS estudia unificarlas con el fin de disminuir sus costos operacionales e incrementar sus niveles de productividad, y plantea las dos posibilidades que se exponen en la tabla 3.

Tabla 3. Parámetros de interés para colas de Poisson. Unificación de sedes

Parámetros		Unificación desde					
		Situación I			Situación II		
		$\lambda = 16$	$\mu = 5$	$S = 4$	$\lambda = 16$	$\mu = 5$	$S = 5$
L	$L = \lambda\left(W_q + \frac{1}{\mu}\right)$ Cantidad de pacientes en la sede	5,585 Pacientes en la sede			3,712 Pacientes en la sede		
L_a	$L_q = \frac{\rho\left(\frac{\lambda}{\mu}\right)^s P_0}{s!\,(1-\rho)^2}$ Cantidad de pacientes en la cola	2,3857 Pacientes en la cola			0,51298 Pacientes en la cola		
W	$W = W_q + \frac{1}{\mu}$ Tiempo de espera en de un paciente en la Sede	0,34911 horas			0,23206 horas		
W_a	$W_q = \frac{L_q}{\lambda}$ Tiempo de espera de un paciente en la cola	0,14911 horas			0,003206 horas		

Fuente: Elaboración de los autores.

9. Conclusiones

La ciencia de la matemática se convierte para la teoría de colas en una herramienta que permite construir modelos que ayudan a predecir el comportamiento de las medidas de desempeño, con el fin obtener de manera oportuna información que permita encontrar un equilibrio entre la prestación de un servicio y el costo que este acarrea.

Las medidas de desempeño para los procesos de colas pueden representarse a través de diferentes distribuciones de probabilidad, de acuerdo a la medición de los resultados obtenidos con la realización del experimento.

Los modelos de Poisson permiten abordar situaciones propias de la teoría de colas, caracterizadas por la llegada aleatoria e independiente de los clientes, con tiempos de servicio exponenciales que evidencian la falta de memoria respecto al tiempo empleado por algún cliente anterior en el proceso de atención.

La teoría de colas se hace evidente en múltiples escenarios del quehacer habitual de las personas y en el desarrollo de procesos operativos de instalaciones industriales y tecnológicas.

En la gran mayoría de los casos de la teoría de colas la llegada de los clientes ocurre de manera aleatoria, esto facilita su desarrollo teórico; sin embargo, existen otras situaciones en que la llegada de los clientes al sistema puede ser programada.

La complejidad de la determinación de resultados analíticos a partir de la construcción de modelos admite una distribución de probabilidad diferente a la exponencial negativa, pero exige entonces la implementación de procesos de simulación.

Referencias

Allen, A. O. (2005). *Probability, Statistics and Queueing Theory with Computer Applications,* 2.ª ed. Washington, D. C.: Elsevier.

Canavos, G. C. (1988). *Probabilidad y estadística, aplicaciones y métodos*. México, D. F: McGraw-Hill.

Chung, K. L. (1983). *Teoría elemental de la probabilidad y los procesos estocásticos*. Barcelona: Reverte.

Cooper, R. (1981). *Introduction to Queuing Theory,* 2.ª ed. Nueva York: North Holland.

Eppen, G. D. (2000). *Investigacion de operaciones en la ciencia administrativa*. Mexico, D. F.: Prentice Hall.

Grimmett, G. & Stirzaker, D. (1992). *Probability and Random Processes*. Londres: Oxford University Press.

Gross, D. & Harris, C. M. (1985). *Fundamentals of Queueing Theory*. Nueva York: John Wiley and Sons.

Hillier, F. & Lieberman, G. (2010). *Introducción a la investigación de operaciones,* 9.ª ed. México, D. F.: McGraw-Hill.

Meyer, P. L. (1992). *Probabilidad y aplicaciones estadísticas*, ed. revisada. México, D. F: Adisson Wesley.

Prabhu, N. U. (1997). *Foundations of Queueing Theory*. Boston: Kluwer.

Ross, S. (1970). *Applied Probability Models whith Optimization Applications*. San Francisco: Holden-Day.

Ross, S. (1998). *A Firts Course in Probability*. Washington, D. C.: Prentice Hall.

Saaty, T. (1983). *Elements of Queuing Theory with Aplications*. Nueva York: Dover.

Serfozo, R. (1999). *Introduction to Stochastic Networks*. Nueva York: Springer.

Taha, H. A. (2004). *Investigación de operaciones,* 7.ª ed. México, D. F.: Pearson.

Trivedi, K. S. (2002). *Probability and Statistics with Reliability, Queueing and Computer Science Applications,* 2.ª ed. Nueva York: John Wiley and Sons.

Tudor, C. (1994). *Procesos estocásticos*. México, D. F.: Sociedad Matemática Mexicana.

Walpole, R. E. (1992). *Probabilidad y estadística para ciencias e ingeniería,* 4.ª ed. México, D. F.: McGraw-Hill

Winston, W. L. (2005). *Investigación de operaciones. Aplicaciones y algoritmos.* México, D. F.: Thompson.

Universidad del Rosario
Facultad de Administración