FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

MODELOS DE TEORÍA DE COLAS

Trabajo de fin de Grado. Grado en Estadística.

Autor: Gabriel Esteban Velázquez

Tutor: D. Antonio Rufián Lizana

Resumen

Hay muchos modelos en la teoría de colas y en este trabajo se explican teórica y prácticamente a resolver algunos de los modelos más generales y que se encuentran con más frecuencias en nuestras vidas.

Los modelos son explicados como procesos de nacimiento y muerte, donde el nacimiento es la entrada de un cliente al sistema y muerte la salida del cliente. Nuestro estudio se basará fundamentalmente en modelos donde el tiempo entre llegadas de los clientes y el tiempo de servicio seguirán una distribución exponencial, ya que es la distribución que más se da en el sistema de colas debido a su propiedad de pérdida de memoria, donde las llegadas solo depende del momento en el que llegan y no del tiempo entre llegadas.

Para poder entender estos modelos previamente debemos de clasificar el sistema de cola según su estructura: Fuente de entrada, disciplina de la cola y mecanismo de servicio. Además las colas tienen unos parámetros que nos permite calcular el número de clientes en el sistema, el tiempo de espera en el sistema, entre otros, que es el objetivo de este estudio.

Abstract

There are many models on the queueing theory and they are explained in this work theoretically and practically as a resolution of the more general models, which meet with more frequencies in our lives.

The models are explained as processes of birth and death, where the birth is the entry of a client into the system and the death is the leaving of the client. Our study will be based on models where the time between customers'arrivals and time service will follow an exponential distribution, since it is the distribution which occurs with more frequency on the queueing system due to its lost-memory property, where arrivals depend only on the time when you arrive and not on the time between customers'arrivals.

To be able understand these models previously we should classified the queueing system according to its structure: Input source, queue discipline and service mechanism. Furthermore, the queues have parameters that allow us to calculate the number of clients in the system, the waiting time in the system, among others, which is the objective of this study.

Índice general

Página

Capítulo 1

Introducción

La teoría de colas aparece a principios del siglo veinte para estudiar los problemas de congestión de tráfico que se presentaban en las comunicaciones telefónicas. Entre 1903 y 1905, Erlang es el primero en tratar el tráfico telefónico de forma científica, y establece la unidad de tráfico telefónico, que recibe su nombre. Posteriormente esta teoría se ha aplicado a multitud de problemas de la vida real, como el tráfico de automóviles, la regulación de semáforos en una ciudad, la determinación de cajeros en los supermercados , o el control de los tiempos de espera de los procesos que acceden al procesador de un ordenador que trabaja en tiempo compartido. El objetivo es el estudio matemático de colas y líneas de espera. De manera más general, la intención es estudiar las causas y la solución de la congestión.

Las colas (líneas de espera) son parte de la vida diaria. Todos esperamos en colas para comprar en un supermercado, hacer un depósito en el banco, enviar un paquete por correo, obtener comida en la cafetería, comprar billetes para el teatro, etc. Nos hemos acostumbrado a una considerable cantidad de esperas, pero todavía nos molesta cuando éstas son demasiado largas. Sin embargo, tener que esperar no sólo es una molestia personal. El tiempo que la población de un país pierde al esperar en las colas es un factor importante tanto de la calidad de vida como de la eficiencia de su economía.

La teoría de colas utiliza los modelos de colas para representar los tipos de sistemas de líneas de espera. Por lo tanto, estos modelos de líneas de espera son muy útiles para determinar cómo operar un sistema de colas de la manera más eficaz. Proporcionar demasiada capacidad de servicio para operar el sistema implica costos excesivos; pero si no se cuenta con suficiente capacidad de servicio surgen esperas excesivas con todas sus desafortunadas consecuencias. Los modelos permiten encontrar un balance adecuado entre el costo de servicio y la cantidad de espera.
A continuación, definiremos algunos conceptos básicos que nos ayudará a entender la teoría de colas, desde una visión matemática.

1.1. Procesos estocásticos

Un fenómeno de espera suele ser modelado como un proceso estocástico, en tiempo continuo, con un número discretos de estados, que evoluciona a saltos cuando aparecen nuevos clientes en el sistema o cuando desaparecen de él.

Un proceso estocástico se define como una colección indexada de variables aleatorias $\{X_t\}$, donde el subíndice t toma valores de un conjunto T dado. Con frecuencia T se toma como el conjunto de enteros no negativos y X representa una característica de interés medible en el tiempo t. Por ejemplo, el proceso estocástico, $X_1, X_2, X_3, ...$ Puede representar la colección de niveles de inventario semanales (o mensuales) de un producto dado, o puede representar la colección de demandas semanales (o mensuales) de este producto, o las llegadas de individuos a una cola.

A continuación daremos algunas característica de los procesos estocástico:

- $\{X_{(t)}; t \in T\}$ es un proceso estocástico si $X(t)$ es una variable aleatoria para cada t, normalmente t indica tiempo.
  Los procesos estocásticos pueden ser:

  1. Continuos: cuando t puede tomar cualquier valor dentro de un intervalo.
  2. Discretos: cuando t toma valores dentro de un conjunto discreto de puntos.

  Denotando $X_{(t)} = x.$ indica que el proceso aleatorio se encuentra en el estado x en el tiempo t.

- Sea un proceso estocástico $\{N_{(t)}; t \in T\}$. Decimos que $\{N_{(t)}; t \in T\}$ es un proceso de conteo si y solo si:

  1 $N(0) = 0$ c.s.

  2 $N(t)$ toma sólo valores enteros y no negativos c.s.

  3 Si $s < t$ entonces $N(s) < N(t)$.

  4 $N_(t) - N_(s)$ = Número de sucesos ocurridos en el intervalo de tiempo $(s, t]$.

- Un proceso $\{X_{(t)}; t \geq 0\}$ se dice que es de Markov si y sólo si:

  $$P\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, ..., X(t_1) = x_1\} = P\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\}$$

1.2. Proceso de Poisson

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto periodo de tiempo. Concrétamente, se especializa en la probabilidad de ocurrencia de sucesos con probabilidades muy pequeñas, o sucesos raros.
En teoría de colas, será importante esta distribución ya que se supondrá en muchas ocasionas

que las llegadas de los clientes al sistema o cola, seguirá una distribución aleatoria de Poisson.

La función de probabilidad de la distribución de Poisson es

$$f(k,\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

donde

- k es el número de ocurrencias del evento o fenómeno (la función nos da la probabilidad de que el evento suceda precisamente k veces)
- λ es un parámetro positivo que representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado.

Su función de distribución vendrá dada por

$$F(X) = \sum_{x=0}^{X} \frac{e^{-\lambda}\lambda^x}{x!}$$

1.2.1. Esperanza

$$E(X) = \sum_k kf(k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$$

Tomamos $y = k-1$ y obtenemos

$$E(X) = \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

1.2.2. Varianza

Por la propiedad de la varianza tenemos que $Var(X) = E(X^2) - [E(X)+^2]$. Además $E(X^2) = E[X(X-1)] + E(X)$. Entonces, sea:

$$E[X(X-1)] = \sum_k k(k-1)f(x) = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}$$

De igual forma para la esperanza, tomamos $y = k-2$, por lo que obtenemos

$$E[X(X-1)] = \lambda^2 e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda^2$$

Así, $E(X^2) = E[X(X-1)] + E(X) = \lambda^2 + \lambda$.

Con lo cual

$$Var(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

1.2.3. Función de momentos

La función generadora de momentos de la distribución de Poisson con valor esperado λ viene dada por:

$$E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk} f(k;\lambda) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{\lambda(e^t-1)}$$

1.2.4. Suma de variables aleatorias de Poisson

Sea $X_i \sim Po(\lambda_i)$, $i = 1, ..., N$

Entonces la suma de N variables aleatorias de Poisson independientes:

$$Y = \sum_{i=1}^{N} X_i \sim Po(\sum_{i+1}^{N} \lambda_i)$$

1.2.5. Coeficiente de asimetría y curtosis relativa

- Coeficiente de asimetría: $\frac{1}{\sqrt{\lambda}}$
- Curtosis relativa: $3 + \frac{1}{\lambda}$

1.3. Distribución exponencial

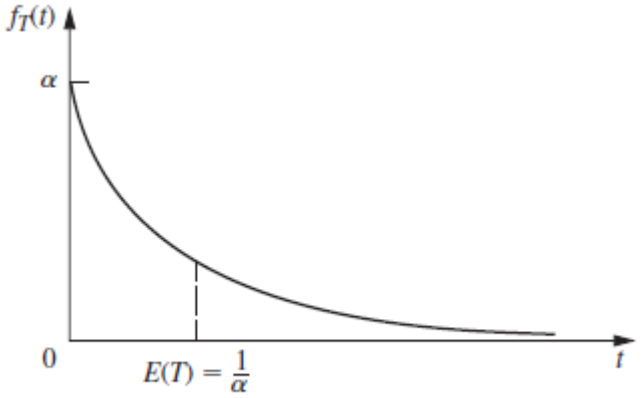
La distribución exponencial es utilizada para determinar la probabilidad de que en cierto tiempo suceda un determinado evento.
Será importante para nuestro estudio, ya que supondremos en muchos casos que el tiempo que ocurre entre la llegada de un cliente y el siguiente, sigue una distribución Exponencial.

Una variable aleatoria X tiene una distribución exponencial si su función de densidad está dada por

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & si \quad x > 0 \\ \\ 0 & c.c. \end{cases}$$

Donde $\lambda > 0$. Algunas de las propiedades más relevantes de la distribución exponencial se exponen a continuación:

1. La $f(t)$ es estrictamente decreciente:



   Se verifica que: $P(0 \leq T \leq \Delta t) > P(\Delta t \leq T \leq 2\Delta t)$

2. Falta de memoria

   En el análisis del comportamiento de las Líneas de Espera (o colas), se reconoce que el proceso de llegada de los clientes al sistema ocurre de forma totalmente aleatoria. Se entiende por aleatorio que la ocurrencia de un evento no se ve afectado por el tiempo transcurrido desde la ocurrencia de un evento anterior.

   $P(T > t + \Delta t \mid T > \Delta t) = P(T > t) = \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda \Delta t}} = e^{-\lambda t}$

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t > 0 \\ 0 & c.c. \end{cases}$$

$$F(t) = P(T \leq t) = \int_0^1 f(x)dx = \int_0^1 \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big]_0^1 = -e^{-\lambda t} + 1$$

$$P(T > t) = 1 - P(T \leq t) = 1 - (-e^{\lambda t} + 1) = e^{-\lambda t}$$

   Solo depende de t y no de Δt.

3. Relación de la distribución exponencial con la distribución de la Poisson.
   Se puede interpretar a la distribución exponencial como el tiempo que transcurre hasta

el primer evento de Poisson. De hecho, las aplicaciones más relevantes de la distribución exponencial son situaciones en donde se aplica el proceso de Poisson.

La relación entre la distribución exponencial y la distribución de Poisson la podemos observar de la siguiente manera. Recordemos que la distribución de Poisson es una distribución con un solo parámetro λ, donde λ representa el número medio de eventos por unidad de tiempo.

Consideremos ahora la variable aleatoria X descrita por el tiempo que se requiere para que ocurra el primer evento. Haciendo uso de la distribución de Poisson, encontramos que la posibilidad de que no ocurra algún evento, en el periodo hasta el tiempo t está dada por

$$p(0, \lambda t) = P(X = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

Podemos ahora utilizar lo anterior y hacer que X sea el tiempo para el primer evento de Poisson. La probabilidad de que la duración del tiempo hasta el primer evento exceda x es la misma que la probabilidad de que no ocurra algún evento de Poisson en x. Esto último, por supuesto, está dado por como resultado,

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

Así la función de distribución acumulada para X está dada por

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

Para reconocer la presencia de la distribución exponencial basta con derivar la función de distribución acumulada anterior para obtener la función de densidad de la distribución exponencial.

1.4. Distribución Gamma

Antes de estudiar la distribución gamma, es pertinente observar y/o examinar algunos detalles de la función a la que debe su nombre, la función gamma.

1.4.1. Función Gamma $\Gamma(\alpha)$

Es una función que extiende el concepto de factorial a los números complejos. Fue presentada, en primera instancia, por Leonard Euler entre los años 1730 y 1731. La función

gamma $\Gamma(\alpha)$ se define,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

, para $\alpha > 0$.

Con el fin de observar algunos resultados o propiedades de esta función, procederemos a integrar por partes. Tomando $u = x^{\alpha-1}$ y $dv = e^{-x}dx$, obtenemos

$$\Gamma(\alpha) = -e^{-x}x^{\alpha-1}|_0^\infty + \int_0^\infty e^{-x}(\alpha-1)x^{\alpha-2}dx = (\alpha-1)\int_0^\infty x^{\alpha-2}e^{-x}dx$$

Para $\alpha > 1$, lo cual ocasiona la fórmula $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$.
Al aplicar reiteradamente la fórmula anterior

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\Gamma(\alpha-2) = (\alpha-1)(\alpha-2)(\alpha-3)\Gamma(\alpha-3)$$

y así sucesivamente. Se evidencia que cuando $\alpha = n$, donde n es un entero positivo,

$$\Gamma(n) = (n-1)(n-2)...\Gamma(1)$$

. Sin embargo, por la definición de $\Gamma(\alpha)$, $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$, y de aquí $\Gamma(n) = (n-1)!$
Algunas propiedades adicionales de $\Gamma(\alpha)$ son:

- $\Gamma(n+1) = n!$ si n es un entero positivo.
- $\Gamma(n+1) = n\Gamma(n)$, $n > 0$.
- $\Gamma(1/2) = \sqrt{\pi}$

1.4.2. Distribución Gamma

Una variable aleatoria X tiene una distribución gamma si su función de densidad viene dada por:

$$f(x,\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & para \quad \alpha,\beta,x > 0 \\ \\ 0 & c.c. \end{cases}$$

Esperanza

$$E(X) = \int_{-\infty}^\infty xf(x)dx = \int_0^\infty x\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}dx = \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty x^\alpha e^{-x/\beta}dx$$

Si hacemos el cambio $u = x/\beta$; $x = u\beta$, con lo que $dx = \beta du$, así

$$E(X) = \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty (u\beta)^\alpha e^{-u}\beta du = \frac{\beta}{\Gamma(\alpha)}\int_0^\infty u^\alpha e^{-u}du = \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha} = \frac{\beta\alpha\Gamma(\alpha}{\Gamma(\alpha)} = \alpha\beta$$

Varianza

De igual forma que para la distribución de Poisson tenemos que $Var(X) = E(X^2) - [E(X)]^2$. Encontremos entonces

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2 \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^{\infty} x^{\alpha+1} e^{-x/\beta} dx$$

De la misma forma, $u = x/\beta$; $x = u\beta$, con lo que $dx = \beta du$, así:

$$E(X^2) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^{\infty} (u\beta)^{\alpha+1} e^{-u} \beta du$$

$$= \frac{\beta^2}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha+1} e^{-u} du = \frac{\beta^2 \Gamma(\alpha+2)}{\Gamma(\alpha} = \frac{\beta^2(\alpha+1)\alpha\Gamma(\alpha}{\Gamma(\alpha)} = (\alpha+1)\alpha\beta^2$$

Con esto,

$$Var(X) = E(X^2) - [E(X)]^2 = (\alpha+1)\alpha\beta^2 - (\alpha\beta)^2 = \alpha^2\beta^2 + \alpha\beta^2 - \alpha^2\beta^2 = \alpha\beta^2$$

1.5. Distribución Erlang

Una variable aleatoria X tiene una distribución gamma si su función de densidad viene dada por:

$$f(x, \alpha, \beta) = \begin{cases} \frac{\lambda^k e^{-\lambda x} x^{k-1}}{(k-1)!} & para \quad x, \lambda > 0 \\ \\ 0 & c.c. \end{cases}$$

Se puede comprobar que la distribución de Erlang es el equivalente de la distribución gamma con el parámetro $\alpha = k = 1, 2, ...$ y $\lambda = 1/\beta$.

La relación entre los modelos de la distribución de Poisson y de Erlang es la siguiente. Si el número de sucesos aleatorios independientes que ocurren en un intervalo de tiempo de longitud t es una variable aleatoria de Poisson de parámetro λt, entonces el tiempo que transcurre hasta la k-ésima ocurrencia (o entre una determinada ocurrencia hasta la k-ésima siguiente) tiene distribución de Erlang de parámetros k y λ.

Para demostrarlo notamos por X_t a la variable aleatoria que representa el número de sucesos que ocurren en un intervalo de tiempo de longitud t que sigue una distribución de Poisson $Poi(\lambda t)$ y por T la variable aleatoria que representa el tiempo hasta la k-ésima ocurrencia (o entre una determinada ocurrencia hasta la k-ésima siguiente). Calculamos la función de distribución de la variable aleatoria T:

$$F_T(t) = P[T \leq t] = P[X_t \geq k] = 1 - P[X_t < k] =$$

$$= 1 - \sum_{x=0}^{k-1} P[X_t = x] = 1 - \sum_{x=0}^{k-1} e^{\lambda t} \frac{(\lambda t)^x}{x!}, t > 0$$

Derivando se obtiene su función de densidad:

$$f_T(t) = -\left[\sum_{x=0}^{k-1} \left\{-\lambda e^{-\lambda t} \frac{(\lambda t)^x}{x!}\right\} + \sum_{x=0}^{k-1} \left\{e^{-\lambda t} \frac{(\lambda t^{x-1})}{x!}\right\}\right] = \lambda e^{-\lambda t} \left[\sum_{x=0}^{k-1} \frac{(\lambda t)^x}{x!} - \sum_{x=0}^{k-1} \frac{\lambda^{x-1}}{(x-1)!}\right], t > 0 \tag{1.1}$$

Haciendo $r = x - 1$ en la segunda suma se obtiene

$$f_T(t) = \lambda e^{-\lambda t} \left[\sum_{x=0}^{k-1} \frac{(\lambda t)^x}{x!} - \sum_{r=0}^{k-2} \frac{\lambda^r}{r!}\right] = \lambda e^{-\lambda t} \frac{((\lambda t)^{k-1}}{(k-1)!} = \frac{\lambda^k}{(k-1)!} t^{k-1} e - \lambda t, t > 0, \tag{1.2}$$

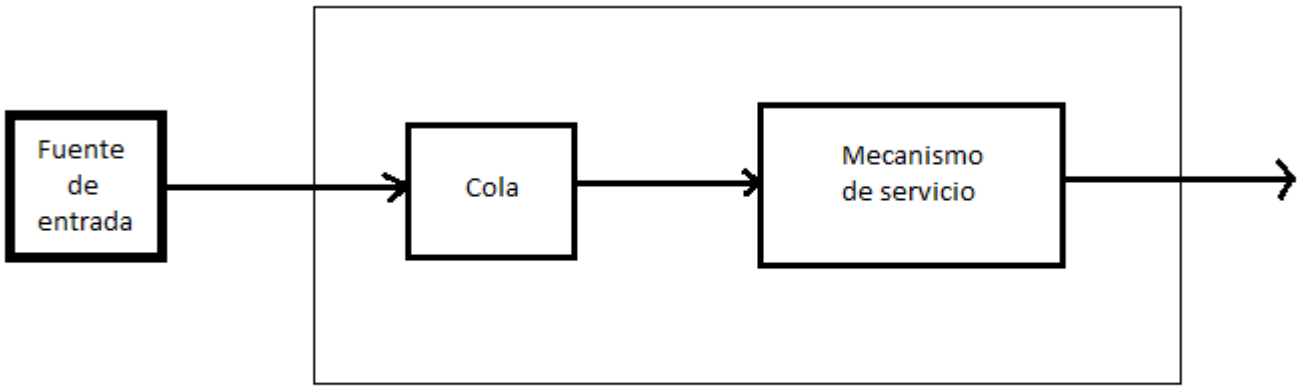
que corresponde a la de una $Gamma(k, \lambda) \equiv Erlang(k, \lambda)$

Capítulo 2

Estructura de un modelo de colas

2.1. Estructua básica de un modelo de colas

Los clientes que requieren un servicio se generan en el tiempo en una fuente de entrada. Luego, entran al sistema y se unen a una cola. En determinado momento se selecciona un miembro de la cola para proporcionarle el servicio mediante alguna regla conocida como disciplina de la cola. Se lleva a cabo el servicio que el cliente requiere mediante un mecanismo de servicio, y después el cliente sale del sistema de colas. Se puede comprobar de una manera más clara en la siguiente figura:



2.2. Fuente de entrada

Su característica es el tamaño. Llamamos tamaño al número total de clientes que pueden requerir servicio en un determinado momento. Podemos suponer que el tamaño es finito o infinito. Se debe especificar el patrón estadístico mediante el cual se generan los clientes a través del tiempo. Como vimos en la introducción, las llegadas de los clientes suele seguir una distribución de Poisson.

2.3. Cola

La cola es el lugar donde los clientes esperan antes de recibir el servicio. Ésta posee dos características principales, en primer lugar la capacidad de la cola, es decir, el número máximo de clientes que puede llegar a soportar. Esta capacidad puede ser finita o infinita, el supuesto de cola infinita es el estándar en la mayoría de modelos ya que poner un límite a la cola puede complicar bastante el análisis, solo será necesario el supuesto contrario cuando el límite de cola sea bastante pequeño y se llegue a él con regularidad. El otro factor determinante de la cola es la disciplina que sigue, que la veremos a continuación.

2.4. Disciplina de la cola

La disciplina de la cola se refiere al orden en el que sus miembros se seleccionan para recibir el servicio.

Los modelos más importantes son los siguientes:

- FIFO (First-In-First-Out): se le da servicio al primero que ha llegado, de forma que la cola está ordenada según el orden de llegada de los usuarios.
- LIFO (Last-In-First-Out): se le da servicio al último que ha llegado, de forma que la cola está ordenada en orden inverso al de llegada de los usuarios.
- SIRO (Service-In-Random-Order): se sortea aleatoriamente cuál de los usuarios en espera accederá al servicio.

No obstantes, otro procedimiento para establecer la disciplina de la cola puede ser el de establecer determinadas prioridades a los diferentes usuarios según algunas de sus características.

En sistemas finitos, en los que el número de usuarios en espera es limitado, es necesario establecer además qué sucede con aquellos usuarios que acceden al sistema cuando la cola de espera está completa. Por último, en los sistemas en que los usuarios son humanos, hay que tener en cuenta otros factores propios del comportamiento humano como el hecho de que hay individuos que no respetan el orden establecido en la cola o bien que hay usuarios que, a la vista de la cola, renuncian a acceder al sistema.

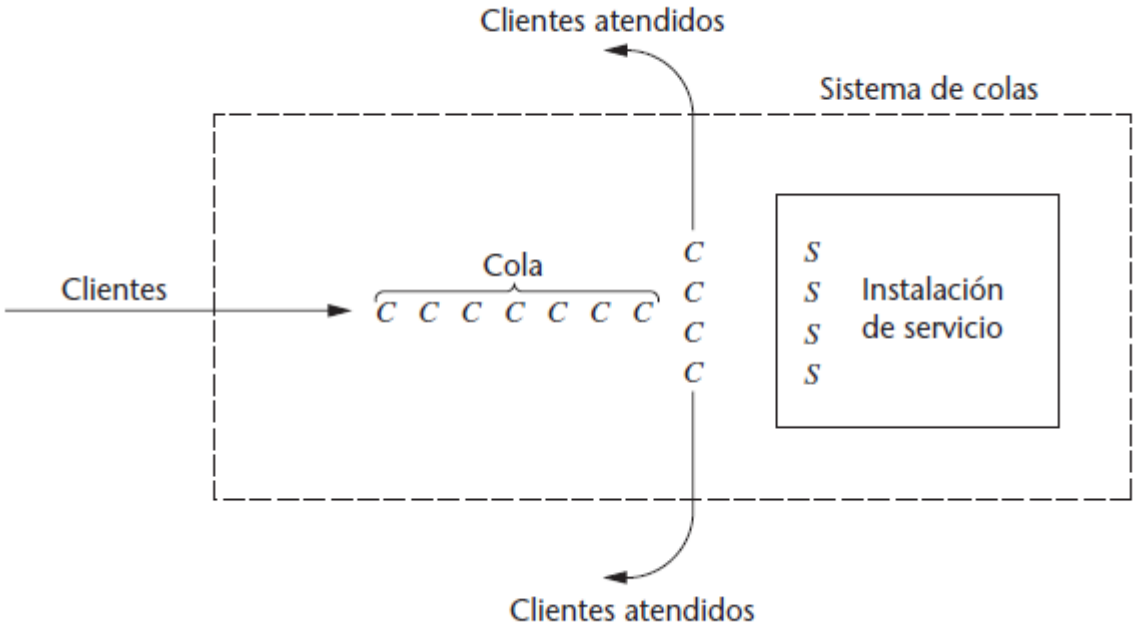
2.5. Mecanismo de servicio

El mecanismo de servicio consiste en una o más estaciones de servicio, cada una de ellas con uno o más servidores o canales de servicio paralelos, llamados **servidores**.

Los modelos de colas deben especificar el número de servidores. Si el tiempo que tardan los usuarios en salir del sistema es mayor que el intervalo entre llegadas, la cola aumentará indefinidamente y el sistema puede llegar a colapsarse. Por tanto, es necesario diseñar el sistema de forma que el tiempo de servicio sea igual o menor que el intervalo entre llegadas. En esta situación es importante saber cuánto tiempo va a estar un servidor inactivo, tiempo que ha de ser mínimo para optimizar el rendimiento del sistema. No obstante, en la mayoría de los sistemas la duración del servicio es también una magnitud aleatoria.

Los modelos más elementales suponen una estación, ya sea con un servidor o con un número finito de servidores. **Se llama tiempo de servicio (o duración del servicio)** al tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación en una estación . Un modelo de un sistema de colas determinado debe especificar la distribución de probabilidad de los tiempos de servicio de cada servidor (y tal vez de los distintos tipos de clientes), aunque es común suponer la misma distribución para todos los servidores. La distribución del tiempo de servicio que más se usa en la práctica por ser más manejable que cualquier otra es la dsitribución *exponencial*. Otras distribuciones de tiempos de servicio importante son la distribución *degenerada* (tiempos de servicio constante) y la distribución *Erlang*, ya explicada en el Capitulo 1.

En la siguiente ilustración podemos ver un sistema de colas. cada cliente se representa por una C y cada servidor por una S.



2.6. Un proceso de colas elemental

La teoría de colas se aplica a muchos tipos diferentes de situaciones. El tipo que más prevalece es el siguiente: una sola línea de espera (que puede estar vacía en ciertos momentos) se forma frente a una instalación de servicio, dentro de la cual se encuentra uno o más servidores. Cada cliente generado por una fuente de entrada recibe el servicio de unos de los servidores, después de esperar un tiempo en la cola (si la cola está vacía el tiempo es 0).

Un servidor no tiene que ser un solo individuo; puede ser un grupo de personas,por ejemplo, una cuadrilla de reparación que combina fuerzas para realizar, de manera simultánea, el servicio que solicita el cliente. Aún más, los servidores ni siquiera tienen que ser personas. En muchos casos puede ser una máquina, un vehículo, un dispositivo electrónico, etc. En esta misma línea de ideas, los clientes que conforman la cola no tienen que ser personas.

Por ejemplo, pueden ser unidades que esperan ser procesadas en cierto tipo de máquina, o automóviles que deben pasar por una caseta de cobro. En realidad, no es necesario que se forme una línea de espera física delante de una estructura material que constituye la estación de servicio. Los miembros de la cola pueden estar dispersos en un área mientras esperan que el servidor venga a ellos, como las máquinas que esperan reparación.

El servidor o grupo de servidores asignados a un área constituyen la estación de servicio de esa área. De todas maneras, la teoría de colas proporciona, entre otros, un número promedio de clientes en espera (el tiempo promedio de espera), puesto que es irrelevante si los clientes esperan en grupo o no.

2.7. Parámetros de colas

Una cola se puede etiquetar como (i, j, c), donde i y j pueden ser varios tipos de distribución. c es un número entero positivo. Su significado son los siguientes:

- i: Distribución tiempo de servicio.
- j: Distribución tiempo entre llegadas.
- c: Número de servidores

Las letras utilizadas para las distribuciones son las siguientes:

- M = distribución exponencial.
- D = distribución degenerada (tiempos constantes).
- E_k = distribución Erlang de parámetro k.
- G = distribución general (permite cualquier distribución arbitraria).

La terminología que utilizaremos, a menos que se establezca otra cosa, será la siguiente:

- $N(t)$ = número de clientes en el sistema de colas en el tiempo t $(t \geq 0)$
- P_n(t) = probabilidad de que n clientes estén en el sistema en el tiempo t
- s = número de servidores en el sistema.
- λ_n = tasa media de llegadas de nuevos clientes cuando hay n clientes en el sistema
- μ_n = tasa media de servicio para todo el sistema cuando hay n clientes.

Cuando λ_n es constante para toda n, se denota por λ. Cuando la tasa media de servicio por servidor ocupado es constante para toda $n \geq 1$, se denotará por μ. En este caso, $\mu_n = s\mu$ cuando $n \geq s$ (cuando los s servidores estén ocupados). En esta circunstancia:

- $1/\lambda$ = tiempo esperado entre llegadas.

- $1/\mu$ = tiempo esperado de servicio.
- ρ = factor de utilización para la instalación de servicio, es decir, la fracción esperada de tiempo que los servidores individuales están ocupados.
- $\rho = \lambda/s\mu \rightarrow$ la fracción esperada de tiempo que los servidores individuales están ocupados es igual a la fracción entre la tasa de llegadas de nuevos clientes al sistema (λ si n es constante) y entre la tasa media de servicios para todo el sistema ($\mu_n = s\mu$ ya que $n \geq s$).

También se requiere cierta notación para describir los resultados de estado estable. Cuando un sistema de colas apenas inicia su operación, el estado del sistema (el número de clientes que esperan en el sistema) se encuentra bastante afectado por el estado inicial y el tiempo que ha pasado desde el inicio. Se dice entonces que el sistema se encuentra en condición transitoria. Sin embargo, una vez que ha pasado suficiente tiempo, el estado del sistema se vuelve, en esencia, independiente del estado inicial y del tiempo transcurrido (excepto en circunstancias no usuales). En este contexto, se puede decir que el sistema ha alcanzado su **condición de estado estable**.

La notación siguiente supone que el sistema se encuentra en la condición de *estado estable*:
P_n =probabilidad de que haya exactamente n clientes en el sistema.
L = número esperado de clientes en el sistema = $\sum_{n=1}^{\infty} nP_n$
L_q= longitud esperada de la cola (excluye los clientes que están en servicio)

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

.

$\mathcal{W}$ = tiempo de espera en el sistema (incluye tiempo de servicio) para cada cliente.
$W = E(\mathcal{W})$
$\mathcal{W}_q$ = tiempo de espera en el sistema (excluye tiempo de servicio) para cada cliente.
$W_q = E(\mathcal{W}_q)$

2.8. Relaciones entre L, W, L_q y W_q

Supongamos que λ_n es una constante λ para toda n. Se ha demostrado que en un proceso de colas en estado estable, $L = \lambda W$.

Dado que John Little proporcionó la primera demostración rigurosa, a veces se le da el nombre de **fórmula de Little**. La demostración prueba que $L_q = \lambda W_q$

Si las λ_n no son iguales, entonces λ se puede sustituir en estas ecuaciones por $\bar{\lambda}$, la tasa promedio entre llegadas a largo plazo.

Si suponemos que el tiempo medio de servicio es una constante $1/\mu$, para toda $n \geq 1$. Se tiene entonces que $W = W_q + \frac{1}{\mu}$

2.9. Ejemplos de sistemas de colas reales

Puede parecer que la descripción de los sistemas de colas es algo abstracta y que sólo es aplicable en situaciones prácticas bastante especiales. Por el contrario, los sistemas de colas

se aplican con sorprendente frecuencia en una amplia variedad de contextos. Para ampliar el horizonte sobre sus aplicaciones, se mencionarán brevemente varios ejemplos reales de sistemas de colas que pertenecen a varias categorías generales.

Una clase importante de sistemas de colas que se encuentra en la vida diaria es el sistema de **servicio comercial**, en donde los clientes externos reciben un servicio de una organización comercial. Muchos de estos sistemas incluyen un servicio de persona a persona en un local fijo, como una peluquería (los peluqueros son los servidores), el servicio de una cajera de banco, las cajas de cobro de un supermercado y una cola en una cafetería (canales de servicio en serie). Sin embargo, muchos otros sistemas son de un tipo diferente, como la reparación de aparatos domésticos (el servidor va hacia el cliente), una máquina de monedas (el servidor es una máquina) y una gasolinera (los clientes son automóviles).

Otra clase importante es la de sistemas de **servicio de transporte**. En algunos de estos sistemas los vehículos son los clientes, como los automóviles que esperan para pasar por una caseta de cobro o un semáforo (el servidor), un camión de carga o un barco que esperan que una cuadrilla les dé el servicio de carga o descarga y un avión que espera aterrizar o despegar en una pista (el servidor). (Un estacionamiento es un ejemplo poco usual de este tipo, en el que los automóviles son los clientes y los espacios son los servidores, pero no existe una cola porque si un estacionamiento está lleno, los clientes se van a otro.) En otros casos, los vehículos son los servidores, como los taxis, los camiones de bomberos y los elevadores.

En años recientes, la teoría de colas se ha aplicado más a los **sistemas de servicio interno** donde los clientes que reciben el servicio son personal interno o parte de la organización. Los ejemplos incluyen sistemas de manejo de materiales, en donde las unidades de manejo de materiales (los servidores) mueven cargas (los clientes); sistemas de mantenimiento, en los cuales las brigadas de mantenimiento (los servidores) reparan máquinas (los clientes) y puestos de inspección en los que los inspectores de control de calidad (los servidores) inspeccionan artículos (los clientes). Las instalaciones para empleados y los departamentos que les prestan servicio también entran en esta categoría. Además, las máquinas se pueden ver como servidores cuyos clientes son los trabajos que están procesando. Un ejemplo relacionado muy importante es un centro de cómputo en el que la computadora se puede ver como el servidor.

Existe un reconocimiento creciente de que la teoría de colas también se puede aplicar a **sistemas de servicio social**. Por ejemplo, un sistema judicial es una red de colas, donde los juzgados son las instalaciones de servicio, los jueces (o los jurados) son los servidores y los casos que esperan el proceso son los clientes. Un sistema legislativo es una red de colas similar, en el cual los clientes son los asuntos que el congreso va a tratar. Algunos sistemas de salud pública son sistemas de colas. Al principio de este capítulo, se vio nuestro ejemplo prototipo (la sala de urgencias de un hospital), pero también las ambulancias, las máquinas de rayos X y las camas del hospital pueden actuar como servidores en sus propios sistemas. En forma parecida, las familias en espera de viviendas de interés social u otros servicios pueden ser clientes de un sistema de colas.

Aun cuando éstas son cuatro clases amplias de sistemas de colas, la lista todavía no se agota. En realidad, la teoría de colas comenzó a principios de siglo con aplicaciones a

ingeniería telefónica (el fundador de la teoría de colas, A. K. Erlang, era empleado de la Danish Telephone Company, en Copenhague), y la ingeniería telefónica constituye todavía una importante aplicación. Lo que es más, cada individuo tiene sus propias líneas de espera personales: tareas, libros que leer, etc. Estos ejemplos son suficientes para sugerir que los sistemas de colas sin duda se presentan con toda frecuencia en muchas áreas de la sociedad.

Capítulo 3

Procesos de Nacimiento y Muerte

La mayor parte de los modelos elementales de colas suponen que las entradas (llegadas de clientes) y las salidas (clietes que se van) del sistema ocurren de acuerdo con un *nacimiento y muerte*.

En el contexto de la teoría de colas, el término **nacimiento** se refiere a la *llegada* de un cliente al sistema de colas, mientras que el término **muerte** se refiere a la *salida* del cliente servido.

El proceso de nacimiento y muerte describe en *términos probabilísticos* cómo cambia $N(t)$ al aumentar t. En general, sostiene que los nacimientos y muertes individuales ocurren de manera aleatoria, y que sus tasas medias de ocurrencia dependen del estado del sistema actual. Los supuestos del proceso de nacimiento y muerte son los siguientes:

Supuesto 1. Dado $N(t) = t$, la distribución de probabilidad actual del tiempo que falta para el próximo nacimiento (llegada) es exponencial con parámetro λ_n $n(n = 0, 1, 2, ...)$.

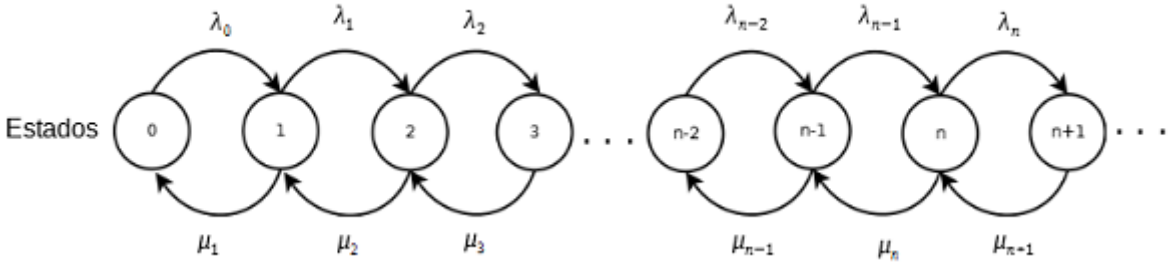
Supuesto 2. Dado $N(t) = n$, la distribución de probabilidad actual del tiempo que falta para la próxima muerte (terminación de servicio) es exponencial con parámetro $\mu_n(n = 1, 2, ...)$.

Supuesto 3. La variable aleatoria del supuesto 1 (el tiempo que falta hasta el próximo nacimiento) y la variable aleatoria del supuesto 2 (el tiempo que falta hasta la siguiente muerte) son mutuamente independientes. La siguiente transición del estado del proceso es:

un solo nacimiento $n \to n + 1$

o una sola muerte $n \to n - 1$

lo que depende de cuál de las dos variables es más pequeña.



En el caso de un sistema de colas, λ_n y μ_n representan, respectivamente, la tasa media de llegada y la tasa media de terminaciones de servicio, cuando hay n clientes en el sistema. En algunos sistemas de colas, los valores de las λ_n serán las mismas para todos los valores de n, y las μ_n también serán las mismas para toda n excepto para aquella n tan pequeña que el servidor esté desocupado.

Excepto en algunos casos especiales, el análisis del proceso de nacimiento y muerte es complicado cuando el sistema se encuentra en condición transitoria. Se han obtenido algunos resultados sobre esta distribución de probabilidad de $N(t)$ pero son demasiado complicados para darles un buen uso práctico. Por otro lado, es bastante directo derivar esta distribución después de que el sistema ha alcanzado la condición de estado estable (en caso de que pueda alcanzarla). Este desarrollo parte del **diagrama de tasas** , como se escribe a continuación.

Considere cualquier estado particular n (n = 0,1,2,...) del sistema. Suponga que en el tiempo 0 se inicia el conteo del número de veces que el sistema entra a este estado y el número de veces que sale de él, como se denota a continuación:

- $E_n(t)$ = número de veces que el proceso entra al estado n hasta el tiempo t.
- $L_n(t)$ = número de veces que el proceso sale del estado n hasta el tiempo t.

Como los dos tipos de eventos (entrar y salir) deben alternarse, estos dos números serán iguales o diferirán en solo 1; es decir, $|E_n(t) - L_n(t)| \leq 1$.

Al dividir ambos lados entre t y después hacer que $t \to \infty$ se obtiene

$$|\tfrac{E_n(t)}{t} - \tfrac{L_n(t)}{t}| \leq \tfrac{1}{t} \text{ entonces } \lim_{t\to\infty} |\tfrac{E_n(t)}{t} - \tfrac{L_n(t)}{t}| = 0$$

Si se dividen $E_n(t)$ y $L_n(t)$ entre t se obtiene la tasa real (número de eventos por unidad de tiempo) a la que ocurren estos dos tipos de eventos, y cuando $t \to \infty$ se obtiene la tasa media (número esperado de eventos por unidad de tiempo):

$\lim_{t\to\infty} \frac{E_n(t)}{t}$ = tasa media a la que el proceso entra al estado n.
$\lim_{t\to\infty} \frac{L_n(t)}{t}$ = tasa media a la que el proceso sale del estado n.

Estos resultados conducen al siguiente principio clave:

Principio de tasa de entrada = tasa de salida. Para cualquier estado $n(n = 0, 1, 2, ...)$ del sistema, la tasa media de entrada = tasa media de salida.

La ecuación que expresa este principio se llama **ecuación de balance** del estado n. Después de construir las ecuaciones de balance de todos los estados en términos de probabilidades P_n desconocidas, se puede resolver este sistema de ecuaciones para calcularlas.

A fin de ilustrar una ecuación de balance, considere el estado 0. El proceso entra a este estado solo desde el estado 1. En consecuencia, la probabilidad de estado estable de encontrarse en el estado 1 (P_1) representa la proporción de tiempo que es posible que el proceso entre al estado 0. Dado que el proceso se encuentra en el estado 1, la tasa media de entrada al estado 0 es μ_1 (para cada unidad acumulada de tiempo que el proceso pasa en el estado 1, el número esperado de veces que lo dejaría para entrar al estado 0 es μ_1). Desde cualquier otro estado, esta tasa media es 0. Por lo tanto, la tasa media global a la que el proceso deja su estado actual para entrar al estado 0 (la tasa media de entrada) es

$$\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1$$

Por el mismo razonamiento, la tasa media de salida debe ser $\lambda_0 P_0$, de manera que la ecuación de balance del estado 0 es

$$\mu_1 P_1 = \lambda_0 P_0$$

En el caso de todos los demás estado, existen dos transiciones posibles, hacia dentro y hacia afuera del estado. Entonces, a cada lado de las ecuaciones de balance de estos estados representa la suma de las tasas medias de las dos transiciones incluidas. Por lo demás, el razonamiento es igual que para el estado 0. Estas ecuaciones de balance se resume en la siguiente tabla:

Estado	**Tasa de entrada = Tasa de salida**
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P3 = (\lambda_2 + \mu_2) P_2$
$\vdots$	$\vdots$
n-1	$\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1 P_{n-1}} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$
$\vdots$	$\vdots$

Observe que la primera ecuación de balance contiene dos variables (P_0 y P_1), las primeras dos ecuaciones contienen tres variables (P_p, P_1 y P_2) y así sucesivamente, de manera que siempre se tiene una variable "adicional". Por lo tanto, el procedimiento para resolver estas ecuaciones es despejar todas las variables en términos de una de ellas, entre las cuales la más conveniente es P_0.

Estado:

0 $P_1 = \frac{\lambda_0}{\mu_1} P_0$

1 $P_2 = \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$

2 $P_3 = \frac{\lambda_2}{\mu_3} P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) = \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$

$\vdots$ $\vdots$

n-1 $P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} + \frac{1}{\mu_n}(\mu_{n-1} P_{n-1} - \lambda_{n-2} P_{n-2}) = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\lambda_{n-1} \lambda_{n-2} \ldots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} P_0$

n $P_{n+1} = \frac{\lambda_n}{\mu_{n+1} P_n} + \frac{1}{\mu n+1}(\mu_n P_n - \lambda_{n-1} P_{n-1)} = \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_n \lambda_{n-1} \ldots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} P_0$

$\vdots$ $\vdots$

La notación se simplifica de la siguiente manera:

$$C_n = \frac{\lambda_{n-1} \lambda_{n-2} \ldots \lambda_0}{\mu_n \mu_{n-1} \ldots \mu_1}$$

Entonces, las probabilidades de estado estable son:

$P_n = C_n P_0$ para $n = 0, 1, 2, \ldots$

El requisito

$\sum_{n=0}^{\infty} P_n = 1$

implica que

$(1 + \sum_{n=0}^{\infty} C_n) P_0 = 1$

de manera que

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} C_n}$$

Dada esta información,

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

También, como el número de servidores s representa el número de clientes que pueden estar en servicio (y no en la cola) al mismo tiempo:

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

Lo que es más, las relaciones dadas anteriormente conducen a:

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

donde $\bar{\lambda}$ es la tasa de llegadas promedio a largo plazo. Como λ es la tasa media de llegadas cuando el sistema se encuentra en el estado $n(n = 0, 1, 2, ...)$ y P_n es la proporción de tiempo que el sistema está en este estado,

$$\bar{\lambda} = \sum_{n=0}^{\infty}\lambda_n P_n$$

Varias de las expresiones que se acaban de presentar incluyen sumas con un número infinito de términos. Por fortuna, estas sumas tienen soluciones analíticas de muchos casos especiales interesantes[1]. En otros casos, se puede aproximar al sumar un número finito de términos en una computadora.

Estos resultados de estado estable se desarrollan bajo supuesto de que los parámetros λ_n y μ_n tienen valores tales que el proceso, en realidad, puede alcanzar la condición de estado estable. Este supuesto siempre se cumple si $\lambda_n = 0$ para algún valor de n mayor que el estado inicial, de forma que sólo son posibles un número finito de estados. También se cumple siempre cuando λ y μ están definidas y $\rho = \lambda/(s\mu) < 1$. No se cumple si $\sum_{n=1}^{\infty} C_n = \infty$.

3.1. Modelos de colas basados en el proceso de nacimiento y muerte(infinitas)

En estos modelos suponemos que todos los tiempos entre llegadas son independientes e idénticamente distribuidos de acuerdo con una distribución exponencial (es decir, el proceso de entrada es de Poisson), que todos los tiempos de servicio son idependientes e idénticamente distribuidos de acuerdo con otra distribución exponencial y que el número de servidores es s (cualquier entero positivo). En consecuencia, estos modelos es un caso especial del proceso de nacimiento y muerte cuando la tasa media de llegadas al sistema de colas y la tasa media de servicio por servidor ocupado son constantes (λ y μ) e independientes del estado del sistema.

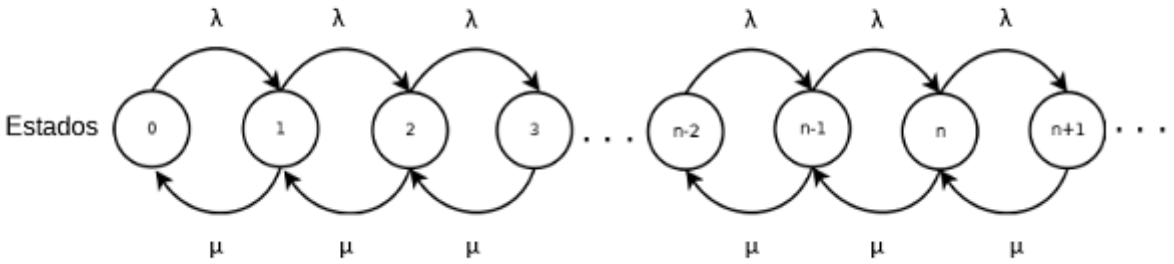
Tasa de servicio de sistema La tasa del servicio del sistema μ_n representa la tasa media de los servicios terminados de todo el sistema de colas cuando existen n clientes en él.

[1]Estas soluciones están basadas en los siguientes resultados conocidos para la suma de cualquier serie geométrica.
$\sum_{n=0}^{N} x^n = \frac{1-x^{N+1}}{1-x}$, para cualquier x
$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$, si $|x| < 1$.

3.2. Modelo M/M/1

Cuando el sistema tiene un sólo servidor (s=1), la implicación es que los parámetros del proceso de nacimiento y muerte son $\lambda_n = \lambda(n = 0, 1, 2, ...)$ y $\mu_n = \mu(n = 1, 2, ...)$ y los factores C_n del proceso de nacimiento y muerte se reducen a

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n \text{ para } n = 0, 1, 2, ...$$



Por lo tanto,

$$P_n = \rho^n P_0, \text{ para } n = 0, 1, 2, ...$$

donde

$$P_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1} = 1 - \rho$$

Así

$$P_n = (1 - \rho)\rho^n$$

En consecuencia,

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

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

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

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

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

De forma similar

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

$$= L - 1(1 - P_0)$$

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

Cuando $\lambda \geq \mu$, esto es, cuando la tasa media de llegadas excede la tasa media de servicio, la solución anterior "no sirve"(puesto que la suma para calcular P_0 diverge). En este caso, la cola crecería sin límite. Aun cuando $\lambda = \mu$, el número esperado de clientes en el sistema crecerá sin límite y con lentitud a través del tiempo, y aunque siempre es posible un regreso temporal a no tener clientes, las probabilidades de tener números grandes de clientes crecen en forma significativa con el tiempo.

Si se supone de nuevo que $\lambda < \mu$, se puede obtener la distribución de probabilidad del tiempo de espera en el sistema $\mathcal{W}$ de una llegada aleatoria cuando la *disciplina* de la cola es primero en entrar, primero en salir. Si esta llegada encuentra n clientes en el sistema, tendrá que esperar $n + 1$ tiempos de servicio exponenciales, inclusive el propio. Por lo tanto, sean $T_1, T_2, ...$ las variables aleatorias independientes de los tiempos de servicio que tienen una distribución exponencial con parámetros μ, y sea

$S_{n+1} = T_1 + T_2 + ... + T_{n+1}$ para $n = 0, 1, 2, ...$

de manera que S_{n+1} representa el tiempo de espera condicional, dado que hay n clientes en el sistema. Se sabe que la distribución de S_{n+1} es Erlang. Como la probabilidad de que una llegada aleatoria encuentre n clientes en el sistema es P_n se concluye que

$$P(\mathcal{W} > t) = \sum_{n=0}^{\infty} P_n P(S_{n+1} > t)$$

Lo que después de una manipulación algebraica considerable se reduce a

$$P(\mathcal{W} > t) = e^{-\mu(1-\rho)t}$$

La conclusión es que $\mathcal{W}$ tiene una distribución exponencial con parámetro igual a $\mu(1-\rho)$. Por lo tanto

$$W = E(\mathcal{W}) = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu - \lambda}$$

Estos resultados incluyen el tiempo de servicio en el tiempo de espera. En algunas situaciones el tiempo de espera más importante es justo hasta que comienza el servicio. Considérese el tiempo de espera en la cola (excluyendo el tiempo de servicio) $\mathcal{W}_q$ para la llegada aleatoria cuando la disciplina de la cola es primero en llegar, primero en salir. Si esta llegada

no encuentra clientes en el sistema, se le sirve de inmediato, de manera que

$$P\left[\mathcal{W}_q = 0\right] = P_0 = 1 - \rho$$

Si encuentra $n > 0$ clientes, entonces tendrá que esperar n tiempos de servicios exponenciales hasta que su propio servicio comienze, de forma que

$$P\left[\mathcal{W}_q > t\right] = \sum_{n=1}^{\infty} P_n P\left[S_n > t\right]$$

$$= \sum_{n=1}^{\infty} (1-\rho)\rho^n P\left[S_n > t\right] = \rho \sum_{n=0}^{\infty} P_n P\left[S_{n+1} > t\right]$$

$$= \rho P\left[\mathcal{W} > t\right] = \rho e^{-\mu(1-\rho)^t}$$

para $t \leq 0$.

Al obtener la media de esta distribución (o aplicar $L_q = \lambda W_q$ o $W_q = W - (1/\mu)$)

$$W_q = E(\mathcal{W}_q) = \frac{\lambda}{\mu(\mu - \lambda)}$$

.

3.2.1. Ejemplo

Una tienda de alimentación es atendida por una persona. Aparentemente, el padrón de llegadas de clientes durante los sábados se comporta siguiendo un proceso de Poisson con una tasa de llegada de 10 personas por hora. A los clientes se les atiende siguiendo un orden de tipo FIFO (primero en entrar, primero en ser servido) y, debido al prestigio de la tienda, una vez que llegan, están dispuestos a esperar el servicio. Se estima que el tiempo que lleva atender a un cliente se distribuye exponencialmente con un tiempo medio de 4 minutos. Determina:

a) La probabilidad de que haya línea de espera.

b) La longitud media de la línea de espera.

c) El tiempo medio que un cliente permanece en la cola.

d) La probabilidad de que un cliente esté menos de 12 minutos en la tienda.

Solución

a) Se trata de un modelo (M/M/1).

λ_n: tasa media de llegada cuando se encuentran n clientes en el sistema.
μ_n: tasa media de servicio cuando se encuentra n clientes en el sistema.
ρ: factor de utilización del sistema: $\frac{\lambda_n}{\mu_n}$.

En este caso $s = 1$, $\lambda_n = \lambda =$ constante, $\mu_n = \mu =$ constante.

$\lambda =$ número de clientes que llegan/unidad de tiempo = (10/60) clientes/minuto.
$\mu =$ número de servicios/unidad de tiempo = (1/4) servicio/minuto.

Así, $\rho = \frac{10/60}{1/4} = \frac{2}{3} < 1$(condición de estabilidad del sistema).

P(haya línea de espera) = 1 - P(no haya línea de espera) = 1- (P(nadie en la tienda) + P(1 cliente en la tienda)=1-($P_0 + P_1$).
Evidentemente, para que exista línea de espera debe de haber más de dos personas ya que solo hay un servidor.

Utilizando los cálculos anteriores:

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

.

La probabilidad de que tengamos n clientes en el sistema viene dada por:
$P_n = (1 - \rho)\rho^n$ para $n = 0, 1, 2, ...$

$P_1 = (1 - \rho)\rho = (1 - \frac{2}{3})\frac{2}{3} = \frac{2}{9}$

Luego:

P(haya una línea de espera) = $1 - (P_0 + P_1) = 1 - (\frac{1}{3} + \frac{2}{9}) = \frac{4}{9}$

b) Longitud esperada (media) de la cola.

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\left(\frac{10}{60}\right)^2}{\frac{1}{4}\left(\frac{1}{4} - \frac{10}{60}\right)} = 1{,}3 \approx 1$$

cliente.

c) El tiempo medio que un cliente espera en la cola se obtiene:

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\frac{10}{60}}{\frac{1}{4}\left(\frac{1}{4}-\frac{10}{60}\right)} = 7{,}98 \approx 8$$

minutos.

d) Sea la variable aleatoria T = tiempo que un cliente pasa en el sistema. T se distribuye como una exponencial de parámetros $(\mu - \lambda)$, la probabilidad de esperar más de 12 minutos es:

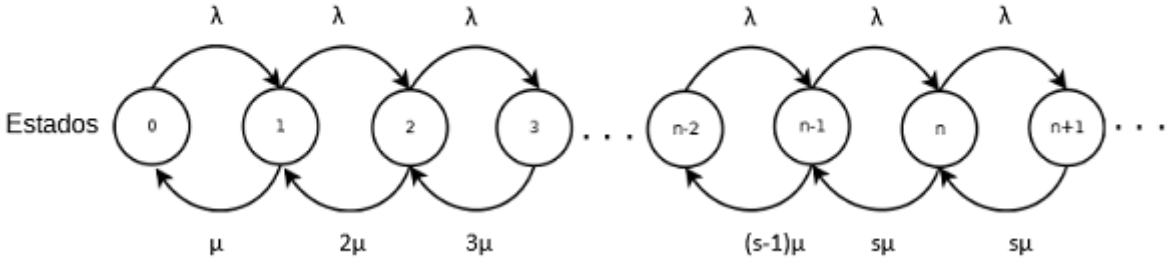
$$P(T \leq 12) = 1 - e^{-(1/4-1/6)12} = 0{,}632$$

.

3.3. Modelo (M/M/S), $S > 1$

Cuando $s > 1$, los factores C_n se convierten en

$$C_n = \begin{cases} \frac{(\lambda\mu)^n}{n!} & para \quad n = 1, 2, ..., s \\ \frac{(\lambda\mu)^s}{s!}\left(\frac{\lambda}{s\mu}\right)^{n-s} = \frac{(\lambda/\mu)^n}{s!s^{n-s}} & para \quad n = s, s+1, ... \end{cases} \tag{3.1}$$



En consecuencia, si $\lambda < s\mu$ (de manera que $\rho = \lambda/(s\mu) < 1$) entonces

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

donde el término para $n = 0$ en la última suma lleva al valor correcto de 1 debido a la convención de que $n! = 1$ cuando $n = 0$. Estos factores C_n dan también

Más aún,

$$L_q = \sum_{n=s}^{\infty}(n-s)P_n = \sum_{j=0}^{\infty} jP_{s+j}$$

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

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

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

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

$W = W_q + \frac{1}{\mu}$

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

El método de un solo servidor para encontrar la distribución de probabilidad de los tiempos de espera se puede extender al caso de varios servidores.Cuando $s - 1 - \lambda/\mu = 0$, $\frac{(1-e^{-\mu t(s-1-\lambda/\mu)})}{s-1-\lambda/\mu}$ debe sustituirse por μt y se obtiene (para $t \geq 0$)

$P\{\mathcal{W} > t\} = e^{-\mu t} \left[\frac{1+P_0(\lambda/\mu)^s}{s!(1-\rho)} \left(\frac{1-e^{-}\mu t(s-1-\lambda/\mu)}{s-1-\lambda/\mu} \right) \right]$

y

$P\{\mathcal{W} > t\} = (1 - P\{\mathcal{W}_q = 0\}) e^{-s\mu(1-\rho)t}$

donde

$P\{\mathcal{W}_q = 0\} = \sum_{n=0}^{s-1} P_n$.

Si $\lambda \geq s\mu$, de forma que si la tasa media de llegadas excede a la tasa media máxima de servicio, la cola crece sin límite y las soluciones de estado estable anteriores no se pueden aplicar.

3.3.1. Ejemplo

Los trabajadores de una fábrica tienen que llevar su trabajo al departamento de control de calidad antes de que el producto llegue al final del proceso de producción. Hay un gran número de empleados y las llegadas son aproximadamente de 20 por hora. El tiempo para inspeccionar una pieza sigue una distribución exponencial de media de 4 minutos. Calcular el número medio de trabajadores en el control de calidad si hay:

a) Dos inspectores.

b) Tres inspectores.

Solución

a) En este apartado $s = 2$.
λ_n: tasa media de llegada cuando se encuentran n clientes en el sistema.
μ_n: tasa media de servicio cuando se encuentran n clientes en el sistema.
ρ: factor de utilización del sistema, $\rho = \frac{\lambda_n}{s\mu_n}$.

λ = número de trabajadores que llegan/unidad de tiempo = 1/3 trabajadores/minutos.

μ = número de servicio/unidad de tiempo = 1/4 servicio/minuto.

Se trata de un modelo para una cola infinita con más de un servidor (M/M/s).

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\frac{1}{1-\lambda(s\mu)}]} = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{(\lambda/\mu)^2}{2!}\frac{1}{1-(\lambda/2\mu)}} = 0{,}2$$

Entonces:

$$L_q = \frac{P_0(\lambda/\mu)^2\rho}{s!(1-\rho)^2} = 1{,}06 \approx 1$$

Así el número medio de trabajadores en el control de calidad es de 1.

b) En este apartado lo que cambia es que s pasa a valer 3, pero λ y μ sigue valiendo lo mismo:

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\frac{1}{1-\lambda(s\mu)}]} = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2!\mu^2} + \frac{(\lambda/\mu)^3}{3!}\frac{1}{1-(\lambda/3\mu)}} = 0{,}25$$

Calculamos a continuación el número medio de trabajadores en el control:

$$Ł_q = P_0\frac{\lambda^s}{s!\mu^s}\frac{1}{\left(1-\frac{\lambda}{s\mu}\right)^2} = 0{,}14 \approx 0$$

De este modo el número medio de trabajadores en el contros de calidad es de 0.

3.4. Ejemplo práctico

La sala de urgencias del Hospital General proporciona cuidados médicos rápidos a los casos de emergencia que llegan en ambulancia o vehículos particulares. En todo momento se cuenta con un médico de guardia. No obstante, debido a la creciente tendencia a usar estas instalaciones para casos de urgencia en lugar de ir a una clínica privada, cada año el hospital experimenta un aumento continuo del número de pacientes que llegan a la sala de emergencias. Como resultado, es bastante común que los pacientes que llegan durante las horas pico (temprano en la tarde) tengan que esperar turno un segundo médico a esta sala durante esas horas pico, para que se puedan atender dos casos de emergencia al mismo tiempo. El hospital ha contratado a un estadístico para que estudie esta opción.

Solución El estadístico ha concluido que los casos de emergencia llegan casi de manera aleatoria (*proceso de entrada Poisson*), por lo que los tiempos entre llegadas tienen una distribución exponencial. También llegó a la conclusión de que el tiempo que necesita el doctor para atender a los pacientes sigue aproximadamente una *distribución exponencial*. Por estos motivos eligió el modelo M/M/s para hacer el estudio.

Al proyectar los datos disponibles para el turno de la tarde al año próximo, estimó que los pacientes llegarán a una tasa promedio de uno cada media hora. Un doctor requiere un promedio de 20 minutos para atender al paciente.

Tomando una hora como unidad de tiempo:

$$\frac{1}{\lambda} = \frac{1}{2} \text{ horas por cliente}$$
$$\frac{1}{\mu} = \frac{1}{3} \text{ horas por cliente,}$$

de manera que

$$\lambda = 2 \text{ clientes por hora}$$
$$\mu = 3 \text{ clientes por hora.}$$

Las dos alternativas bajo consideración son: continuar con un solo doctor durante este turno ($s = 1$) o agregar un segundo doctor ($s = 2$). En ambos casos:

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

de forma que el sistema debe acercarse a la condición de estado estable. (En realidad, como λ varía un poco durante los otros turnos, el sistema nunca alcanzará verdaderamente la condición de estado estable, pero el estadístico administrador piensa que los resultados correspondientes proporcionarán una buena aproximación.) Por lo tanto, usa las ecuaciones anteriores para obtener los resultados que se muestran en la siguiente tabla:

	$s=1$	$s=2$
ρ	$2/3$	$1/3$
P_0	$1/3$	$1/2$
P_1	$2/9$	$1/3$
P_n para $n \geq 2$	$\frac{1}{3}\left(\frac{2}{3}\right)^n$	$\left(\frac{1}{3}\right)^n$
L_q	$4/3$	$1/12$
L	2	$\frac{3}{4}$
W_q	$2/3$	$1/24$ (en horas)
W	1	$3/8$ (en horas)
$P\left[\mathcal{W}_q > 0\right]$	0.667	0.167
$P\left[\mathcal{W}_q > \frac{1}{2}\right]$	0.404	0.022
$P\left[\mathcal{W}_q > 1\right]$	0.245	0.003
$P\left[\mathcal{W}_q > t\right]$	$\frac{2}{3}e^{-t}$	$\frac{1}{6}e^{-4t}$
$P\left[\mathcal{W} > t\right]$	e^{-t}	$\frac{1}{2}e^{-3t}(3-e^{-t})$

Con base en estos resultados, concluye en forma tentativa que para el siguiente año sería inadecuado un solo médico para brindar atención con relativa prontitud, lo que es necesario en la sala de emergencias de un hospital.

Capítulo 4

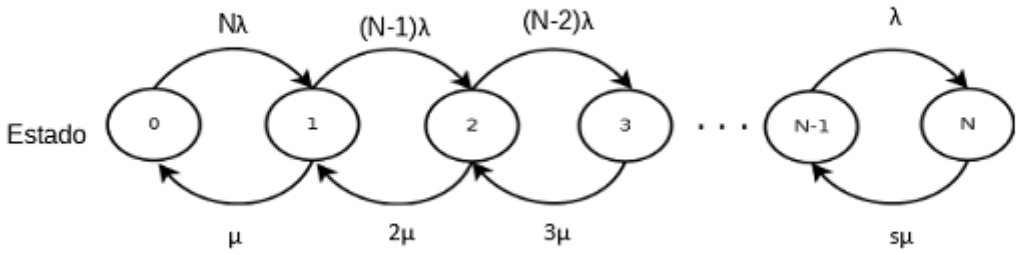
Modelos de colas basados en el proceso de nacimiento y muerte(finitas)

En este caso, no se permite que el número de clientes en el sistema exceda un número especificado (denotado por K), por lo que la capacidad de la cola es K-s. A cualquier cliente que llega cuando la cola está llena se le niega la entrada al sistema y lo deja para siempre (no podría entrar en el sistema nunca). Desde el punto de vista del proceso de nacimiento y muerte, la tasa media de entrada al sistema se hace cero en estos momentos. Por lo mismo, la única modificación necesaria en el modelo M/M/s para introducir la cola finita es cambiar los parámetros λ_n a

$$\lambda_n = \begin{cases} \lambda & para \quad n = 0, 1, 2, ..., K-1 \\ 0 & para \quad n \geq K \end{cases}$$

Como $\lambda_n = 0$ para algunos valores de n, un sistema de colas que se ajuste a este modelo alcanzará en algún momento la condición de estado estable, aun cuando $\rho = \lambda/s\mu \geq 1$.

Por lo general este modelo se etiqueta como M/M/s/K, donde K es finito.



$$\mu_n = \begin{cases} n\mu & para \quad n = 1, 2, ..., s \\ s\mu & para \quad n = s, s+1, ... \end{cases}$$

4.1. Modelo (M/M/1/K)

$$C_n = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n = \rho^n & para \quad n = 0, 1, 2, ..., K-1 \\ 0 & para \qquad n \geq K \end{cases}$$

Si $\rho = 1$ entonces $P_n = 1/(K+1)$ para $n = 0, 1, 2, ..., K$, de forma que $L = \frac{K}{2}$.
A continuación veremos para $\rho \neq 1$

$P_0 = \frac{1}{\sum_{n=0}^{k}(\lambda/\mu)^2} = \frac{1}{\left[\frac{1-(\lambda/\mu)K+1}{1-\lambda/\mu}\right]} = \frac{1-\rho}{1-\rho^{K+1}}\rho^n$, para $n = 0, 1, 2, ..., K$.
Entonces,

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

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

$$= \rho \frac{-(K+1)\rho^K + K\rho^{K+1} + 1}{(1-\rho^{k+1})(1-\rho)} = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

Como es usual (cuando $s = 1$),

$L_q = L(1 - P_0)$

En este resultado no exige que $\lambda < \mu$.

Cuando $\rho < 1$ se puede verificar que el segundo término de la última expresión de L converge hacia 0 cuando $K \to \infty$, por lo que, sin duda, todos los resultados anteriores convergen hacia los resultados correspondientes que se han obtenido previamente cuando K la tomábamos como infinito.

De igual manera que para K infinito se puede obtener los tiempos de espera esperados de los clientes que llegan al sistema:

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

donde

$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{K-1} \lambda P_n = \lambda(1 - P_k)$

4.2. Modelo (M/M/s/k)

De nuevo, lo hacemos para más de un servidor. Debido a que este modelo no permite más de K clientes en el sistema, K es el número máximo de servidores que pueden tenerse. Suponga que $s \leq K$. En este caso, C_n se expresa como

$$C_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} & para \quad n = 0, 1, 2, ..., s \\ \frac{(\lambda/\mu)^s}{s!}\left(\frac{\lambda}{s\mu}\right)^{n-s} = \frac{(\lambda/\mu)^n}{s!s^{n-s}} & para \quad n = s, s+1, ..., K \\ 0 & para \quad n > K \end{cases}$$

Así,

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!}P_0 & para \quad n = 1, 2, ..., s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}}P_0 & para \quad n = s, s+1, ..., K \\ 0 & para \quad n > K \end{cases}$$

donde

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

(Estas fórmulas aplican la convención de que $n! = 1$ cuando $n = 0$).

Si se adapta a este caso la derivación de L_q del modelo M/M/s, se llega a

$$L_q = \frac{P_0(\lambda/\mu)^s\rho}{s!(1-\rho)^2}[1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho)]$$

,

donde $\rho = \lambda/(s\mu)$. En el caso de que $\rho = 1$, es necesario aplicar la regla de L'Hopital dos veces a la expresión de L_q. De otra manera, todos estos resultados se cumplen para toda $\rho > 0$. La razón para que este sistema de colas alcance la condición de estado estable aun cuando $\rho \geq 1$ es que $n \geq K$, de modo que el número de clientes en el sistema no puede seguir creciendo en forma indefinida.

Se puede demostrar que

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$$

.

igual que para un servidor, se obtiene W y W_q a partir de estas cantidades.

4.2.1. Ejemplo

Un pequeño taller mecánico de coches cuenta con una zona acondicionada para realizar reparaciones, además tiene otras dos donde se pueden aparcar los coches en espera de sus reparación. Según un estudio, cada día vienen al taller un promedio de 2 coches a solicitar su reparación según una distribución de Poisson. En el taller trabajan dos mecánicos que trabajan juntos en la reparación de cada coche y tardan una media de 1 día en realizarla según una distribución exponencial. Además se contrata un equipo de reparación móvil (herramientas + mecánico especializado) que tiene una productividad similar a la propia del taller, que interviene desplazándose hasta el taller cuando se agotan las plazas libres. El coste de contratación del equipo móvil es de 20 000 euros en función a la fracción de tiempo que trabaja.
El beneficio medio por cada reparación realizada es de 30 000 euros.

a) Contruye un modelo de colas apropiado para representar la situación.

b) ¿Qué tanto por ciento de tiempo están todos los mecánicos sin trabajar?

c) Cuál es el beneficio medio diario del taller.

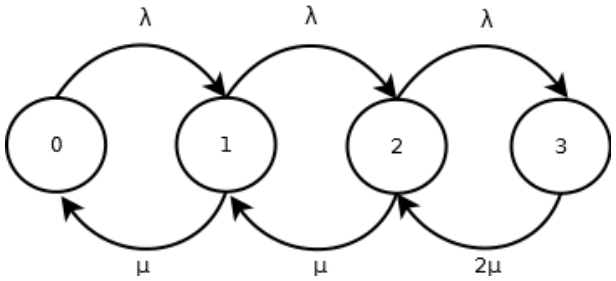
d) La dirección del taller está manejando la posibilidad de habilitar una de las zonas de aparcamiento para realizar separaciones. En tal caso se dejaría de contratar al equipo móvil y cada mecánico trabajaría por separado y se estima que demoren una media de 3 días en reparar un coche según una distribución exponencial. ¿Qué recomedarías a la dirección del taller?

e) En los intentos por mejorar el servicio, se han realizado gestiones con el ayuntamiento para alquilar dos plazas de parking en la calle frente al taller. El alquiler mensual de cada plaza es de 25 000 euros ¿Es conveniente alquilar las plazas? Nota: el equipo móvil sigue siendo contratado cuando hay por lo menos tres coches en el taller.

Solución

a) Modelo para una cola finita con más de un servidor.
λ =número de coches que llegan/unidad de tiempo = 2 coches/día.
μ = número de servicios/unidad de tiempo = 1 reparación por día.
s = 3
Tamaño de la cola= 2.
Tamaño de la población = L.

$\rho = \frac{\lambda}{s\mu} = \frac{2}{3} < 1$ (condición de estabilidad del sistema).



b) $P_0 = \frac{1}{1+\sum_{n=1}^{3} c_n}$, así

$c_1 = \frac{\lambda_0}{\mu_1} = \frac{2}{1} = 2$

$c_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} = \frac{2 \cdot 2}{1 \cdot 1} = 4$

$c_3 = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} = \frac{2 \cdot 2 \cdot 2}{1 \cdot 1 \cdot 2} = 4$

$P_0 = \frac{1}{1+\sum_{n=1}^{3} c_n} = \frac{1}{1+2+4+4} = \frac{1}{11} = 0{,}0909$

Es decir, que el 9.09 % del tiempo están todos los mecánicos sin trabajar.

c) Sea B_{rep} = beneficio por reparación.

Beneficio= $B_{rep} \cdot \bar{\lambda}$−(Coste equipo móvil)$P_3$ siendo $\bar{\lambda} = \lambda(1 - P_M)$ con M=longitud máxima de la cola, es decir M=3.
$P_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n} = P_0 = c_n P_0$, así

$P_3 = c_3 P_0 = 4 \cdot \frac{1}{11} = \frac{4}{11} = 0{,}3636$

$\bar{\lambda} = \lambda(1 - P_M) = 2(1 - \frac{4}{11}) = \frac{14}{11} = 1{,}2727$

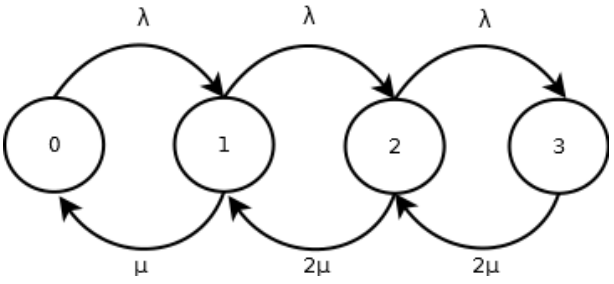
Beneficio $= 30000 \cdot 1{,}2727 - 20000 \cdot 0{,}3636 = 38181 - 7272 = 30909$ euros por día.

d) λ = número de coches que llegan/unidad de tiempo = 2 coches/día.

μ = número de servicios/unidad de tiempo = 1/3 reparación/día.

$P_0 = \frac{1}{1+\sum_{n=1}^{3} c_n}$ con $c_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$, así

$c_1 = \frac{\lambda_0}{\mu_1} = 2/(1/3) = 6$



$c_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2} = \frac{2\cdot 2}{(1/3)(1/3)} = 18$

$c_3 = \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} = \frac{2\cdot 2\cdot 2}{(1/3)\cdot(3/3)\cdot(2/3)} = 54$

$P_0 = \frac{1}{1+\sum_{n=1}^{3} c_n} = \frac{1}{1+6+18+54} = 0{,}012658$

$P_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n} P_0 = c_n P_0$

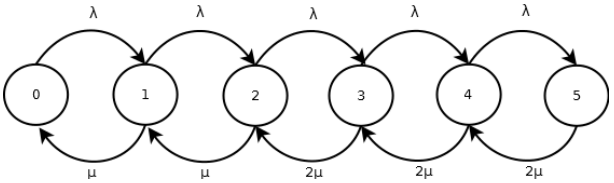
$P_1 = c_1 P_0 = 6 \cdot 0{,}012658 = 0{,}075948$
$P_2 = c_2 P_0 = 18 \cdot 0{,}012658 = 0{,}227844$
$P_3 = c_3 P_0 = 54 \cdot 0{,}012658 = 0{,}683532$

$\bar{\lambda} = \lambda(1 - P_M) = 2(1 - (0{,}683532)) = 0{,}632936$

Beneficio $= B_{rep} \cdot \bar{\lambda} = 30000 \cdot 0{,}632936 = 18988{,}8$ euros/día, de ahí que no convenga hanilitar la nueva zona.

e) λ = número de coches que llegan/unidad de tiempo = 2 coches/día.

μ = número de servicios/unidad de tiempo = 1 reparación/día.



$P_0 = \frac{1}{1+\sum_{n=1}^{3} c_n}$ con $c_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$, así

$c_1 = \frac{\lambda_0}{\mu_1} = 2/1 = 2$

$c_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2} = \frac{2\cdot 2}{(1\cdot 1} = 4$

$c_3 = \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} = \frac{2\cdot 2\cdot 2}{1\cdot 1\cdot 2} = 4$

$c_4 = \frac{\lambda_0\lambda_1\lambda_2\lambda_3}{\mu_1\mu_2\mu_3\mu_4} = \frac{2\cdot 2\cdot 2\cdot 2}{1\cdot 1\cdot 2\cdot 2} = 4$

$c_5 = \frac{\lambda_0\lambda_1\lambda_2\lambda_3\lambda_4}{\mu_1\mu_2\mu_3\mu_4\mu_5} = \frac{2\cdot 2\cdot 2\cdot 2\cdot 2}{1\cdot 1\cdot 2\cdot 2\cdot 2} = 4$

$P_0 = \frac{1}{1+\sum_{n=1}^{5} c_n} = \frac{1}{1+2+4+4+4+4} = 0{,}052631$

$P_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n} P_0 = c_n P_0$

$P_1 = c_1 P_0 = 2 \cdot 0{,}052631 = 0{,}105263$
$P_2 = c_2 P_0 = 4 \cdot 0{,}052631 = 0{,}210526$
$P_3 = c_3 P_0 = 4 \cdot 0{,}052631 = 0{,}210526$
$P_4 = c_4 P_0 = 4 \cdot 0{,}052631 = 0{,}210526$
$P_5 = c_5 P_0 = 4 \cdot 0{,}052631 = 0{,}210526$

Sea B_{rep} = beneficio por reparación.

Beneficio = $B_{rep} \cdot \bar{\lambda}$ − (Coste equipo móvil)$(P_3 + P_4 + P_5)$-Coste diario alquiler, siendo $\bar{\lambda} = \lambda(1 - P_M)$ con M = longitud máxima de la cola, es decir M = 3.

$\bar{\lambda} = \lambda(1 - P_M) = 2(1 - 0{,}210526) = 1{,}578948.$

Beneficio $=30000 \cdot 1{,}578948 - 20000 \cdot (0{,}210526 + 0{,}21056 + 0{,}210526) - 2 \cdot \left(\frac{25000}{30}\right)$
$= 47368{,}44 - 12631{,}56 - 1666{,}7 = 33070{,}18$ euros/día $>$ 30 909 euros/día, por lo tanto conviene alquilar las plazas de parking.

Capítulo 5

Variación de fuente de entrada finita al modelo M/M/s

En este apartado, la única diferencia con el modelo M/M/s es que la fuente de entrada está limitada; es decir, el tamaño de la población potencial es finito. En este caso, sea N el tamaño de esa población, Cuando el número de clientes en el sistema de colas es $n(n = 0, 1, 2, ..., N)$, existen sólo $N - n$ clientes potenciales restantes en la fuente de entrada.

Como $\lambda_n = 0$ para $n = N$, cualquier sistema que se ajuste a este modelo alcanzará en algún momento la condición de estado estable.

5.1. Variación de fuente de entrada finita al modelo M/M/1

Cuando $s = 1$, los factores C_n se reduce para este modelo de la siguiente forma

$$C_n = \begin{cases} N(N-1)...(N-n+1)\left(\frac{\lambda}{\mu}\right)^n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n & para \quad n \leq N \\ 0 & para \quad n > N \end{cases}$$

Entonces, si se usa de nuevo la convención de que $n! = 1$ cuando $n = 0$,

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

$P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0$, si $n = 1, 2, ..., N$

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

que se puede reducir a

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

$L = \sum_{n=0}^{N} nP_n = L_q + 1 - P_0$

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

Por último,

$$W = \frac{L}{\bar{\lambda}} \qquad W_q = \frac{L_q}{\bar{\lambda}} \tag{5.1}$$

donde

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{N} (N-n)\lambda P_n = \lambda(N - L)$$

5.2. Variación de fuente de entrada finita al modelo M/M/s

Para $N \geq s > 1$,

$$C_n = \begin{cases} \frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n & para \quad n = 0, 1, 2, ..., s \\ \frac{N!}{(N-n)!s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n & para \quad n = s, s+1, ..., N \\ 0 & para \quad n > N \end{cases}$$

Entonces

$$P_n = \begin{cases} \frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & si \quad 0 \leq n \leq s \\ \frac{N!}{(N-n)!s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_0 & si \quad s \leq n \leq N \\ 0 & si \quad n > N \end{cases}$$

donde
$P_0 = 1/\left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n\right]$
Por último,
$L_q = \sum n = sN(n-s)P_n$
y
$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$
con lo que después se obtienen W y W_q igual que en el caso de un servidor.

5.2.1. Ejemplo

Una cantera ha contratado los servicios de una excavadora para recoger la grava y cargarla en camiones. El tiempo medio que tarda la excavadora en cargar un camión es de 10

minutos. La cantera dispone de una flota de 4 camiones, cada uno de ellos tarda una media de 15 minutos en transportar la grava a su destino y volver a la cantera. Tanto los tiempos de viaje como los tiempos de carga se suponen distribuidos exponencialmente. El coste de la excavadora es de 20 euros por hora de servicio. Por otro lado se estima que cada hora que un camión pasa en la cantera representa un coste de 12 euros, ya que durante ese tiempo no esta efectuando servicio de transporte.

a) ¿Qué modelo de colas permite representar este sistema?

b) ¿Cuál es el porcentaje de tiempo que la excavadora esta desocupada?

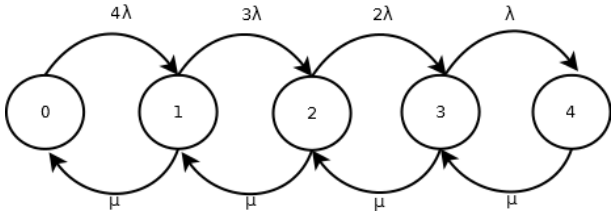
c) ¿Cuál es el número medio de camiones que estarán fuera de la cantera?

d) ¿Cuál es el tiempo medio que un camión pasa en la cantera?

e) La empresa explotadora de la cantera cree que podría minimizar costes si alquila los servicios de otra excavadora que entraría en funcionamiento en caso que hubiese 3 o más camiones en la cantera. La tasa de servicio de esta segunda excavadora sería igual al de la primera pero su coste sería igual al de la primera pero su coste sería de 25 euros la hora de servicio. ¿Resulta aconsejable contratar esta segunda excavadora?

Solución

a) Modelo para una fuente finita con un servidor.
λ = número de camiones que llegan/unidad de tiempo = 4 camiones/hora.
μ = número de servicios/unidad de tiempo = $\frac{1}{10}\frac{servicios}{minutos}\frac{60minutos}{1hora} = \frac{60}{10}\frac{sericio}{hora}$
Tamaño de la cola = 4
Tamaño de la población = 4



b) P(estar desocupada) = P_0

Los datos del sistema de colas son: $s = 1, \mu = cte, \rho = \frac{\lambda}{\mu}$

$P_0 = \frac{1}{\sum_{n=0}^{m} \frac{m!}{(m-n)!}\rho^n}$ con m = tamaño de la fuente finita, en este caso $m = 4$.

$P_0 = \frac{1}{1+\frac{8}{3}+\frac{48}{9}+\frac{192}{2}+\frac{384}{81}} = 0{,}047$, luego el 4.79 % del tiempo está la excavadora desocupada.

c) Nos piden por tanto la diferencia entre el número de camiones total y el número de camiones que están fuera de la cantera.

$$L = m - \frac{\mu}{\lambda}(1 - P_0) = 4 - \frac{6}{4}(1 - 0{,}0479) = 2{,}572$$

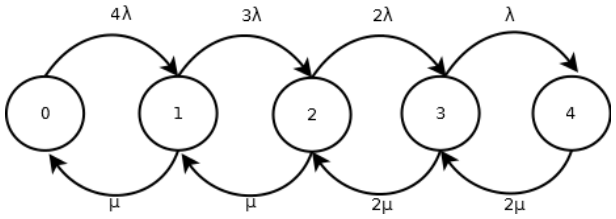
La longitud media del sistema es de 2.572, así el número de camiones que están fuera de la cantera es $m - L = 4 - 2{,}572 = 1{,}428$ camiones.

d) $\bar{\lambda} = \lambda(m - L) = 4(4 - 2{,}572) = 5{,}712$

$W = \frac{L}{\bar{\lambda}} = \frac{2{,}572}{5{,}712} = 0{,}45$ horas.

e) Coste alternativa 1 excavadora
E[Coste total]=E[Coste servicio]+E[Coste espera]
$= 20(1 - P_0) + C_w L = 20(1 - 0{,}0479) + 12(2{,}572) = 49{,}906$ euros/hora.

COste alternativa 2 excavadoras.
Se trata de un modelo general.
λ = número de camiones que llegan/unidad de tiempo = 4 camiones/hora.
μ = número de servicios/unidad de tiempo = 6 servicios/hora.



E[Coste total] = E[Coste servicio]+E[Coste espera]
$= 20(1 - P_0) + 25(P_3 + P_4) + C_w L$

$P_0 = \frac{1}{1+\sum_{n=1}^{4} c_n}$, con $c_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}$, así

$c_1 = \frac{\lambda_0}{\mu_1} = \frac{4\lambda}{\mu} = \frac{16}{6} = \frac{8}{3}$
$c_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} = \frac{4\lambda 3\lambda}{\mu\mu} = \frac{4 \cdot 4 \cdot 3 \cdot 4}{6 \cdot 6} = \frac{16}{3}$
$c_3 = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} = \frac{4\lambda 3\lambda 2\lambda}{\mu\mu\mu} = \frac{64}{18}$
$c_4 = \frac{\lambda_0 \lambda_1 \lambda_2 \lambda_3}{\mu_1 \mu_2 \mu_3 \mu_4} = \frac{4\lambda 3\lambda 2\lambda\lambda}{\mu\mu\mu\mu} = \frac{64}{54}$

$P_0 = \frac{1}{1+\sum_{n=1}^{4} c_n} = \frac{1}{1+\frac{8}{3}+\frac{16}{3}+\frac{64}{18}+\frac{64}{54}} = 0{,}0727$
$P_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n} P_0 = c_n P_0$

$P_1 = c_1 P_0 = (8/3)0{,}0727 = 0{,}1938$
$P_2 = c_2 P_0 = (16/3)0{,}0727 = 0{,}3877$
$P_3 = c_3 P_0 = (64/18)0{,}0727 = 0{,}2584$
$P_4 = c_4 P_0 = (64/54)0{,}0727 = 0{,}0861$
$L = \sum_{n=0}^{4} nP_n = 0 \cdot 0{,}0727 + 1 \cdot 0{,}1938 + 2 \cdot 0{,}3877 + 3 \cdot 0{,}2584 + 4 \cdot 0{,}0861 = 2{,}088$ camiones.
E[Coste total]$= 20(1 - P_0) + 25(P_3 + P_4) + C_w L = 20(1 - 0{,}0727) + 25(0{,}2584 + 0{,}0861) + 12 \cdot 2{,}088 = 52{,}2145$ euros.

Por tanto no resulta aconsejable la segunda excavadora.

Capítulo 6

Modelos de colas con distribuciones no exponenciales

Todos los modelos de toría de colas ya explicados anteriormente se basan en el proceso de nacimiento y muerte, lo que hace necesario que tanto los tiempos entre llegadas como los de servicio tengan distribuciones exponenciales. Este tipo de distribuciones de probabilidad tiene muchas propiedades convenientes para la teoría de colas, pero sólo en cierto tipo de sistemas de colas proporciona un ajuste razonable. En particular, el supuesto de tiempos entre llegadas exponenciales implica que las llegadas ocurren al azar (proceso de entrada de Poisson), lo cual es una aproximación razonable en muchas situaciones pero no cuando las llegadas están programadas o reguladas con todo cuidado.Todavía más, las distribuciones de tiempos de servicio reales con frecuencia se desvían bastante de la forma exponencial, en particular cuando los requerimientos de servicio de los clientes son muy parecidos. Por ello, es importante disponer de otros modelos de colas que usen otras distribuciones de probabilidad.
Se han podido obtener algunos resultados útiles con algunos modelos.

6.1. Modelo M/G/1

Este modelo supone que el sistema de colas tiene un servidor y un proceso de entradas de Poisson (tiempo entre llegadas exponenciales) con una tasa media de llegadas fija λ. Como siempre, se supone que los clientes tienen tiempos de servicio independientes con la misma distribución de probabilidad, pero no se imponen restricciones sobre cuál debe ser esta distribución de tiempos de servicio. En realidad. sólo es necesario conocer o estimar la media $1/\mu$ y la varianza σ^2 de esta distribución.
Cualquier sistema de líneas de espera de este tipo podrá alcanzar, en algún momento, una condición de estado estable si $\rho = \lambda/\mu < 1$. Los resultados de estado estable disponibles de este modelo general son los siguientes:
$P_0 = 1 - \rho$,
$L_q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)}$,
$L = \rho + L_q$,
$W_q = \frac{L_q}{\lambda}$,

$W = W_q + \frac{1}{\mu}$.
Si se toma en cuenta la complejidad que representa el análisis de un modelo que permite cualquier distribución de tiempos de servicio, es notable que se haya podido obtener una fórmula tan sencilla de L_q. Esta fórmula es uno de los resultados más importantes de la teoría de colas gracias a la facilidad con que se aplica y al predominio de los sistemas M/G/1 en la práctica. Esta ecuación de L_q (o su contraparte de W_q) con frecuencia recibe el nombre de fórmula de Pollaczek-Khintchine, en honor de dos pioneros del desarrollo de teoría de colas que dedujeron la fórmula de manera independiente a principios de la década de 1930.
Observe que para cualquier tiempo de servicio esperado fijo $1/\mu$, L_q, L, W_q y W se incrementan cuando σ^2 aumenta. Este resultado es importante porque indica que la congruencia del servidor tiene gran trascendencia en el desempeño de la instalación de servicio, no sólo en su velocidad promedio.
Cuando la distribución de los tiempos de servicio es exponencial, $\sigma^2 = 1/\mu^2$ y los resultados anteriores se reducen a los correspondientes al modelo M/M/1 que se presentó anteriormente.

6.2. Modelo M/D/s

Cuando el servicio consiste básicamente en la misma tarea rutinaria que el servidor realiza para todos los clientes, tiende a haber poca variación en el tiempo de servicio que se requiere. Muchas veces, el modelo M/D/s proporciona una representación razonable de este tipo de situaciones porque supone que todos los tiempos de servicio son iguales a una constante fija (la distribución de tiempos de servicio degenerada) y que tiene un proceso de entradas de Poisson con tasa media de llegadas fija λ.
Cuando sólo se tiene un servidor, el modelo M/D/1 es un caso especial del modelo M/G/1,donde $\sigma^2 = 0$, con lo que la fórmula de Pollaczek-Khintchine se reduce a
$L_q = \frac{\rho^2}{2(1-\rho)}$
donde a partir de este valor de L_q se pueden obtener L, W_q y W como ya se demostró. Observe que el valor de estas L_q y W_q es exactamente igual a la mitad que en el caso de tiempos de servicio exponenciales (el modelo M/M/1) en el que $\sigma^2 = 1/\mu^2$, y entonces al decrecer σ^2 pueden mejorar mucho las medidas de desempeño de un sistema de colas.
En el caso de la versión de más de un servidor de este modelo (M/D/s) se dispone de un método complicado para obtener la distribución de probabilidad de estado estable del número de clientes en el sistema y su media [si se supone que $\rho = \lambda/(s\mu) < 1$]. Existen tabulaciones de estos resultados para muchos casos.

6.3. Modelo $M/E_k/s$

El modelo M/D/s supone una variación cero en los tiempos de servicio ($\sigma = 0$), mientras que la distribución exponencial de tiempos de servicio supone una variación muy grande ($\sigma = 1/\mu$). Entre estos dos casos extremos hay un gran intervalo ($0 < \sigma < 1/\mu$), donde caen la mayor parte de las distribuciones de tiempos de servicio reales. Otro tipo de distribución teórica de tiempos de servicio que concuerda con este espacio intermedio es la distribución de Erlang (llamada así en honor del fundador de la teoría de colas).

$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1} e^{-k\mu t}$
donde μ y k son parámetros estrictamente positivos de la distribución y k está restringido a valores enteros. Su media y desviación estándar son:
Media $= \frac{1}{\mu}$
Desviación estándar $= \frac{1}{\sqrt{k}} \frac{1}{\mu}$.
En este contexto, k es el parámetro que especifica el grado de variabilidad de los tiempos de servicio con relación a la media. Por lo general, se hace referencia a k como el parámetro de forma.
La distribución de Erlang es muy importante en teoría de colas por dos razones. Para describir la primera suponga que $T_1, T_2, ..., T_k$ son k variables aleatorias independientes con una distribución exponencial idéntica, cuya media es $1/(k\mu)$. Entonces, su suma,
$T =_1 + T_2 + ... + T_K$
sigue una distribución de Erlang con parámetros μ y k.
La distribución de Erlang también es útil debido a que es una gran familia (dos parámetros) de distribuciones que permiten sólo valores no negativos. Así, por lo general se puede obtener una aproximación razonable de la distribución empírica de los tiempos de servicio si se usa una distribución de Erlang. En realidad, tanto la distribución exponencial como la degenerada (constante) son casos especiales de distribución de Erlang con $k = 1$ y $k = \infty$, respectivamente. Los valores intermedios de k proporcionan distribuciones intermedias con media $= 1/\mu$, moda $= (k-1)/(k\mu)$ y varianza $= 1/(k\mu^2)$. Por lo tanto, después de estimar la media y la varianza de una distribución de servicio empírica, estas fórmulas de la media y la varianza se pueden usar para elegir el valor de k que se ajuste a estas estimaciones de manera más cercana.
Vamos a considerar el modelo $M/E_k/1$, que es el caso especial del modelo M/G/1 donde los tiempos de servicio tienen una distribución de Erlang con parámetro de forma igual a k. Cuando se aplica la fórmula de Pollaczek-Khintchine con $\sigma^2 = 1/(k\mu^2)$ (y los resultados correspondientes dados por (M/G/1) se obtiene
$L_q = \frac{\lambda^2(k\mu^2)+\rho^2}{2(1-\rho)} = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu-\lambda)}$,
$W_q = \frac{1+k}{2k} \frac{\lambda}{\mu(\mu-\lambda)}$,
$W = W_q + \frac{1}{\mu}$,
$L = \lambda W$.
Con varios servidores (M/Ek/s) se puede aprovechar la relación de la distribución de Erlang con la distribución exponencial que se acaba de describir para formular un proceso de nacimiento y muerte modificado (cadena de Markov de parámetro continuo) en términos de las fases del servicio exponencial individual (k por cliente) y no en términos de los clientes. Sin embargo, no ha sido posible derivar una solución general de estado estable [cuando $\rho = \lambda/(s\mu) < 1$] para la distribución de probabilidad del número de clientes en el sistema. Más bien se necesitaría una teoría avanzada para resolver en forma numérica los casos individuales. Una vez más, estos resultados se han obtenido y tabulado para casos numéricos.

Ejercicio final

Un banco dispone de 3 ventanillas de atención. El tiempo entre llegadas de los clientes son los siguientes:
17.67, 21.10, 32.34, 0.74, 18.22, 0.96, 256.09, 40.32, 38.49, 62.72, 9.06, 17.87, 40.10, 15.88, 3.52, 68.89, 21.39, 75.52, 14.5 y 46.94.

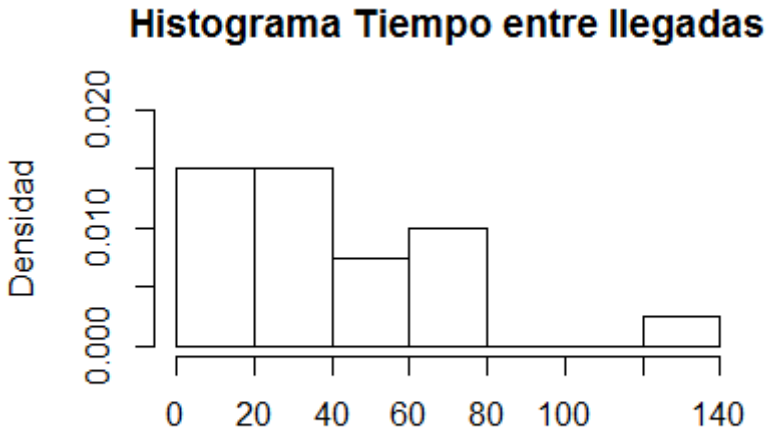
Además se ha recogido el tiempo de servicio por persona y son:

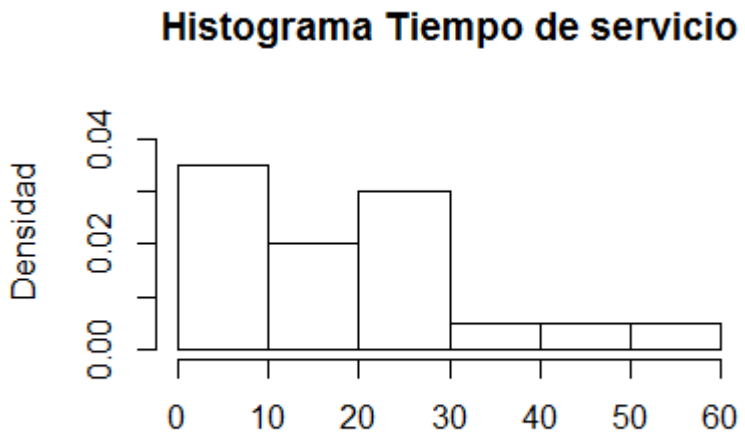
22.79, 8.56, 29.17, 11.29, 24.58, 14.64, 19.56, 56.05, 5.02, 4.37, 34.65, 9.19, 47.30, 17.14, 24.56, 3.04, 8.07, 7.01, 23.70 y 20.73.

El banco se plantea si le conviene aumentar el número de ventanillas para satisfacer mejor a los clientes. El coste que le supone abrir una nueva ventanilla es de 6 euros la hora. El coste horario de espera se ha estimado en 18 euros por cliente.
Los datos de este problema han sido generados aleatoriamente en el programa R, mediante la función rexp(n,rate), donde n es el número de simulaciones y rate es 1/media.

Solución

En primer lugar deberíamos comprobar qué distribución sigue nuestros datos. Para ello vamos a representar los datos en un histograma:





A simple vista es posible que sigan una distribución exponencial. Para asegurarnos del todo vamos a realizar un contraste de hipótesis de bondad de ajuste. Realizaremos el test de Kolmogorov-Smirnov.
La hipótesis a contrastar de este test es:

H_0: Los datos analizados siguen una distribución M.

H_1: Los datos analizados no siguen una distribución M.

Estadístico de contraste es $D = \sup_{1 \leq i \leq n} |\hat{F}_n(x_i) - F_0(x_i)|$

donde:

- x_i es el i-ésimo valor observado en la muestra (cuyos valores se han ordenado previamente de menor a mayor).
- $\hat{F}_n(x_i)$ es un estimador de la probabilidad de observar valores menores o iguales que x_i.
- $F_0(x)$ es la probabilidad de observar valores menores o iguales que x_i cuando H_0 es cierta.

Así pues, D es la mayor diferencia absoluta observada entre la frecuencia acumulada observada $\hat{F}_n(x)$ y la frecuencia acumulada teórica $F_0(x)$, obtenida a partir de la distribución de probabilidad que se especifica como hipótesis nula.

Si los valores observados $\hat{F}_n(x)$ son similares a los esperados $F_0(x)$, el valor de D será pequeño. Cuanto mayor sea la discrepancia entre la distribución empírica $\hat{F}_n(x)$ y la distribución teórica , mayor será el valor de D.

Haremos el test a los datos del tiempo entre llegadas pero para ello nos hará falta el parámetro de la distribución exponencial. Se utilizará la media muestral.
La media de la muestra es 39.5185. En nuestro ejemplo redondearemos la media y usaremos que la media es 40.

Entonces aplicando el contraste de Kolmogorov-Smirnov a los datos nos da un estadístico $D = 0{,}22767$ y un p valor de 0.2154. El p valor es mayor que α, tomando $\alpha = 0{,}05$, por lo que no hay evidencias para rechazar la hipótesis nula. Se puede suponer que el tiempo entre llegadas sigue una distribución exponencial con parámetro 40.

De igual forma con los tiempos entre servicio, calculamos la media muestral, cuyo valor es 19.571. Redondeamos y utilizaremos que la media es 20. Se vuelve a hacer el test de K-S.

El estadístico D es 0.14628,y el p valor = 0.7323. Lo que, de igual manera, no hay evidencias para rechazar la hipótesis nula.

En conclusión, los tiempos entre llegadas, y el tiempo de servicio siguen una exponencial.

Por lo tanto vamos a suponer que los clientes llegan al banco a una tasa de 40 por hora. El tiempo de servicio es de 20 personas por hora.

Nuestro ejercicio se trata de un modelo (M/M/s).

λ =número de clientes que llegan/unidad de tiempo= 40 clientes/hora.

μ = número de servicio/unidad de tiempo = 20 personas/hora.

Para este modelo concreto se utilizará:

- $P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\lambda/(s\mu)}}$
- $L_q = \frac{P_0(\lambda\mu)^s \rho}{s!(1-\rho)^2}$
- $L = L_q + \frac{\lambda}{\mu}$
- Coste total = Coste servicio + Coste de espera
- Coste servicio = s $\cdot$ (euros por servicio/hora) = s $\cdot$ 6
- Coste espera = L $\cdot$(euros por coste en espera /cliente) = 18 $\cdot$ L

	$s = 3$	$s = 4$	$s = 5$
P_0	1/9	3/23	9/67
L_q	8/9	4/23	8/201
L	26/9	50/23	2.0398
Coste servicio	18	24	30
Coste espera	52	39.13	36.72
Coste total	70	63.13	66.72

Por tanto, al banco le interesa abrir solo una ventanilla más.

Capítulo 7

Softwars para aplicar teoría de colas

Existen muchos softwars que pueden solucionar problemas de teorías de colas, realmente casi todos los programas matemáticos son capaces de hacerlo debido al cálculo simple de sus elementos. A continuación he elegido algunos diferentes:

WinQSB es un sistema interactivo de ayuda a la toma de decisiones que contiene herramientas muy útiles para resolver distintos tipos de problemas en el campo de la investigación operativa. El sistema está formado por distintos módulos, uno para cada tipo de modelo o problema. Entre ellos destacaremos el Queuining Analysis que permite el cálculo y resolución de problemas de teoría de colas, dependiendo de su modelo.

POM-QM: es un programa para ciencias de la decisión: métodos cuantitativos, investigación de operaciones, gestión de la producción y las operaciones. De nuevo, este programa va como el anterior por módulos. El módulo que se utilizará para un estudio de teorías de colas será el Waiting Lines.

Además, en R existe un package para calcular y estudia problemas relacionados con colas y líneas de espera. Este paquete se llama Queuing y proporciona herramientas versátiles para el análisis de modelos de colas basados en nacimiento y muerte y redes de cola de formulario de producto único y multiclase.

Capítulo 8

Conclusiones

En la teoráa de colas se usa que los tiempos entre llegadas es exponencial, o familia de la exponencial debido a su propiedad de pérdida de memoria, donde la llegada de un cliente al siguiente solo depende del tiempo de llegada de cada uno y no del incremento del tiempo entre ambos.

Los modelos elementales de colas se pueden suponer y estudiar como un modelo de nacimiento y muerte, donde nacimiento es la llegada de un cliente al sistema y muerte se refiere a la salida del cliente servido.

Si la tasa media de llegadas excede a la tasa media máxima de servicio, la cola crece sin límite por lo que no cumpliría la estabilidad del modelo y no podríamos utilizar los modelos propuestos en este trabajo.

Bibliografía

- Introducción a la Investigación de Operaciones. Frederick S. Hillier, Gerald J. Liberman.
- Operations research. Principles and Practice. Phillips Ravindran Solberg.
- Investigación Operativa. Quintín Martín Martín, M. Teresa Santos Martín, Yanira del Rosario De Paz Santana.
- Fundamentals of queueing theory. Donald Gross, Carl M. Harris.