TEORÍA DE COLAS

- **Teoría de Colas**
- **Modelo M/M/1**

TEORÍA DE COLAS

Una Cola se presenta con frecuencia cuando se solicita un servicio por parte de una serie de clientes y tanto el servicio como los clientes son de tipo probabilístico.

La primera aplicación de teoría de Colas se debe al matemático danés Erlang sobre conversaciones telefónicas en 1909 para el cálculo de tamaño de las centralitas.

La Teoría de Colas es una disciplina de Investigación Operativa que se encarga de poponer modelos para el manejo eficiente de *Líneas de Espera*.

Una *Línea de Espera* es una hilera formada por uno o varios clientes que aguardan para recibir un servicio. Los clientes pueden ser personas, objetos, máquinas que requieren un mantenimiento, contenedores de mercancías para ser embarcados, elementos de inventario para ser utilizados, etc.

Una *Línea de Espera* se forma por un desequilibrio temporal entre la demanda de un servicio y la capacidad del sistema para gestionarlo.

Los *Modelos de Líneas de Espera* son muy útiles para determinar cómo operar un sistema de colas de la manera más eficaz, permiten encontrar un balance adecuado entre el costo de servicio y la cantidad de espera: Proporcionar demasiada capacidad de servicio para operar el sistema implica costos excesivos. De otra parte, si no se cuenta con suficiente capacidad de servicio surgen esperas excesivas con desafortunadas consecuencias.

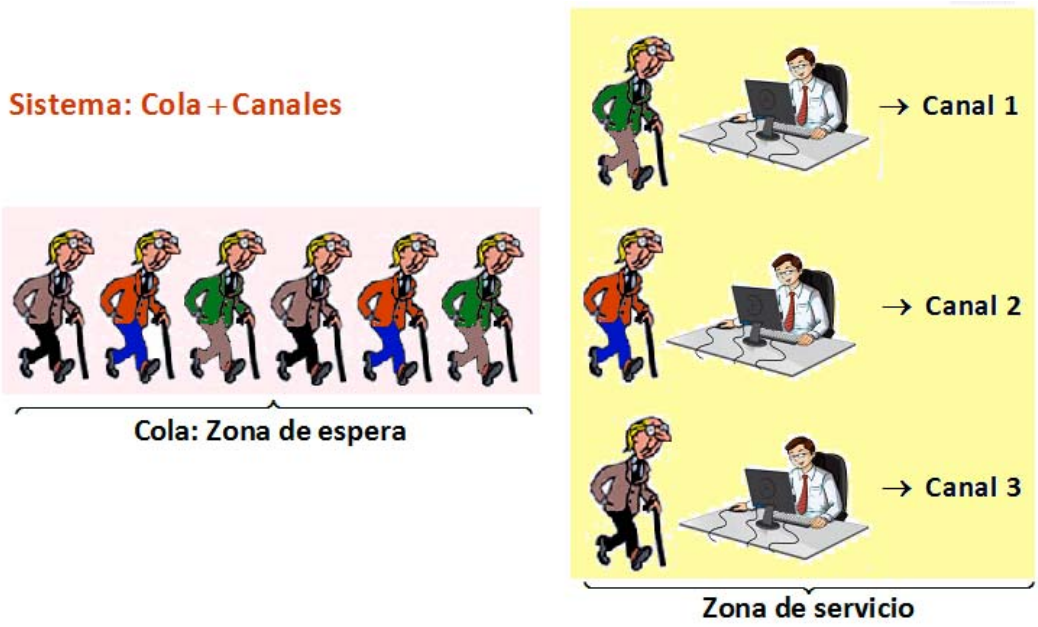
ESTRUCTURA DE LOS PROBLEMAS DE LÍNEAS DE ESPERA: Aunque cada situación específica tiene características diferentes, cuatro elementos son comunes a toda *Línea de Espera*:

- **Una población de clientes que genera clientes potenciales.**
- **Una línea o fila de espera formada por los clientes.**
- **La instalación del servicio, formada por una persona (o un equipo), una máquina (o grupo de máquinas) que se requiere para proveer el servicio que el cliente solicita.**
- **Una regla de prioridad para seleccionar al siguiente cliente que será atendido por la instalación de servicio.**

El término 'cliente' se utiliza en un sentido general, pudiendo ser una persona, piezas esperando su turno para ser procesadas, una lista de trabajo esperando para ser impresas en una impresora de red, etc.

PROCESO BÁSICO DE COLAS

Cola (Zona de espera): En todo sistema los flujos de entrada y salida no están sincronizados. La cola es una acumulación de clientes (personas, productos, objetos) que están a la espera de ser servidos.
Una cola puede evitarse:
(a) Tasa media de llegadas < Capacidad de servicio
(b) Cuando se tiene un control sobre la dispersión de los tiempos de llegadas y la de los tiempos de servicio.



Básicamente la mayoría de los modelos de colas consiste: Los *clientes* que requieren un servicio se generan en el tiempo en una *fuente de llegada*, después entran al *sistema* y se unen a una *cola*.
En determinado momento se selecciona a un cliente 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.

- **En el sistema se puede actuar en las siguientes características:**

(a) Ley que rige las llegadas.
(b) Disciplina de la cola.
(c) Ley que determina el servicio (elección entre tipo y número de canales).

Consideracions sobre los sistemas:

(a) Cuando hay varios canales en paralelo es conveniente mantener una cola única.
(b) Cuando hay tiempos de servicios muy dispares para los diferentes clientes que forman la cola conviene establecer canales separados.
(c) Si la cola aumenta hasta cierto limite conviene aumentar la capacidad de los canales.

- **Con relación a la *disciplina de la cola* hay que considerar:**

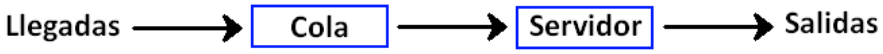
(a) Llegadas individuales o en grupos.
(b) Dependencia del múmero de clientes que pretenden incorporarse al sistema en función del número de clientes que se encuentran en el mismo.
(c) Disuasión de los clientes que pretenden incorporarse al sistema en función de la longitud de la cola (con una cierta probabilidad).

- **En relación con el servicio hay que considerar:**

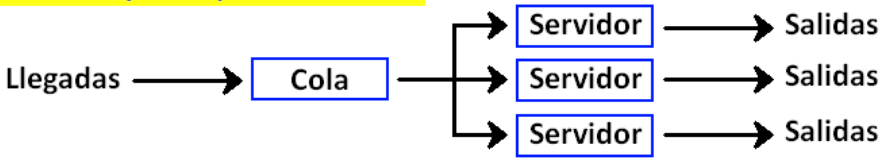
(a) Ley de servicio única a lo largo del tiempo y para todos los clientes.
(b) Ley de servicio variable en función del tipo de cliente y longitud de la cola.

TIPOS DE SISTEMAS DE COLAS

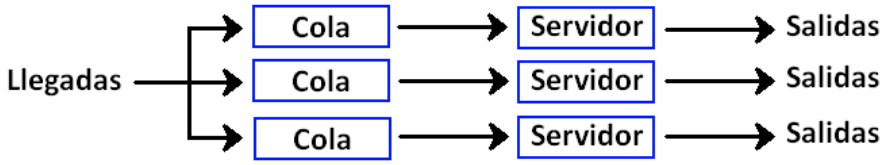
Una Cola y un Servidor



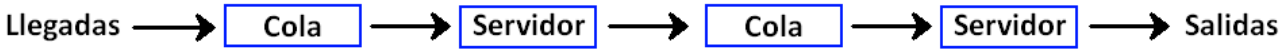
Una Cola y múltiples Servidores



Varias Colas y múltiples Servidores



Una Cola y Servidores secuenciales



DISTRIBUCIONES DE PROBABILIDAD DEL MECANISMO DE SERVICIO

Las fuentes de variación en los problemas de colas o filas de espera provienen del carácter aleatorio de la llegada de clientes y de las variaciones que se registran en los distintos tiempos de servicio. Generalmente cada una de esas fuentes suele describirse mediante una distribución de probabilidad.

La mayor parte de los modelos de colas estocásticas asumen que el tiempo entre diferentes llegadas de clientes siguen una distribución exponencial $\text{Exp}(\lambda)$, o lo que es equivalente, que el ritmo de llegadas sigue una distribución de Poisson.

También es habitual admitir que el ritmo de atención al cliente cuando el servidor está ocupado sigue una distribución de Poisson y la duración de la atención al cliente sigue una distribución exponencial.

La distribución de llegadas: $p_n(t)=\dfrac{(\lambda t)^n}{n!}e^{-\lambda t}$, $n=0,1,2,\cdots$ **clientes**

$p_n(t)\equiv$ **Probabilidad de que *n* clientes estén en el sistema en el tiempo t**

El tiempo entre llegadas, se define como la probabilidad de que no llegue ningún cliente, es decir: $p_0(t)=\dfrac{(\lambda t)^0}{0!}e^{-\lambda t}=e^{-\lambda t}$ **, siendo una distribución exponencial.**

PROCESOS DE POISSON

Si los tiempos entre llegadas/servicios de clientes se distribuyen según una exponencial $\text{Exp}(\lambda)$, el número de llegadas/servicios de clientes hasta un cierto tiempo es un proceso de Poisson.

- **Sea la variable $X\sim\text{Exp}(\lambda)$ la variable aleatoria entre llegadas o tiempo de servicio, su función de densidad** $f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{para } x\geq 0\\ 0 & \text{en otro caso}\end{cases}$ **es estrictamente decreciente.**

La función de distribución: $F(x)=\int_{-\infty}^{x} f(t)\,dt=\begin{cases}1-e^{-\lambda x} & x\geq 0\\ 0 & x<0\end{cases}$

$X\sim\text{Exp}(\lambda)\ \rightarrow\ E(X)=\dfrac{1}{\lambda}\quad V(X)=\dfrac{1}{\lambda^2}$

CARACTERÍSTICAS DE LOS SISTEMAS DE COLAS

Un sistema de colas se describe adecuadamente con seis característica:

- **Fuente de llegada de clientes.**
- **Patrón de servicio de servidores.**
- **Disciplina de cola.**
- **Capacidad del sistema.**
- **Número de canales de servicio.**
- **Número de etapas de servicio.**

Fuente de llegada de clientes: **En situaciones de colas habituales, la llegada de clientes es estocástica, esto es, depende de una variable aleatoria, con lo que se necesita conocer la distribución probabilística entre dos llegadas sucesivas de clientes.**
La fuente de llegada puede variar con el tiempo, cuando se mantiene constante se dice estacionaria, si varía (por ejemplo, con las horas del día) se llama no-estacionaria.
Pueden contemplarse distintas situaciones: Clientes que llegan independiente o simultáneamente (llegan lotes), en este último caso hay que definir su distribución probabilística. Clientes que abandonan la cola por ser demasiado larga o que tras esperar mucho abandonan.
El supuesto normal, para un modelo básico de colas, es que la llegada de clientes hasta un momento específico sigue una distribución de Poisson, aunque no sea la única distribución que puede considerarse.

***Patrón de servicio de servidores*: Pueden presentar un tiempo de servicio variable, en cuyo caso hay que asociar una función de probabilidad. Pueden atender en lotes o de modo individual.**
El tiempo de servicio puede variar con el número de clientes en la cola, trabajando más rápido o más lento, en este caso se conoce como patrones de servicio dependientes. El patrón de servicio puede ser no-estacionario variando con el tiempo transcurrido.

***Disciplina de cola*: Es la regla en el orden que se van a seleccionar los clientes que se encuentran a la espera de ser atendidos en la cola, existen varias reglas, entre las más comunes se pueden encontrar:**

♠ **FIFO (first in first out): Se atiende al cliente en el orden que llegan a la cola, el primero en llegar será el primero en ser atendido. En los modelos básicos de colas se supone como normal la disciplina de primero en entrar, primero en salir, a menos que se establezca de otra manera.**
♠ **LIFO (last in first out): Consiste en atender primero al que ha llegado de último, también se le conoce como 'pila'.**
♠ **RSS (random selection of service): Se selecciona a los clientes de una cola de**

forma aleatoria, con algún procedimiento de prioridad o algún otra preclasificación.

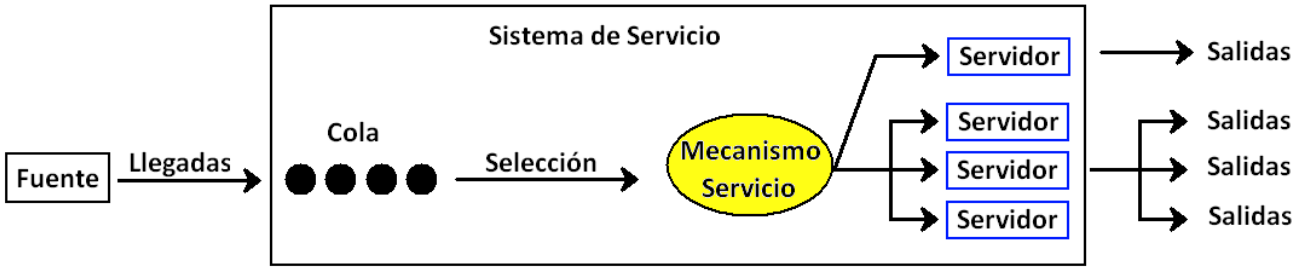
♠ **Processor Sharing: Todos los clientes experimentan con eficacia el mismo retraso, ya que comparten entre todos los clientes de la cola la capacidad del sistema atendiendo a todos por igual.**

***Capacidad del sistema*: Es el número máximo de clientes que pueden estar dentro del sistema haciendo cola antes de ser atendidos para recibir el servicio, al igual que la fuente de llegada este número puede ser finito o infinito.**

***Número de canales de servicio*: Es preferible utilizar sistemas multiservicios con una única línea de espera para todos que con una cola por servidor. Al hablar de canales de servicio paralelo se trata generalmente de una cola que alimenta a varios servidores.**

***Número de etapas de servicio*: Puede ser unietapa o multietapa, en este último el cliente puede pasar por un número de etapas mayor que uno. En algunos sistemas multietapa se admite la vuelta atrás o reciclado, modo habitual en sistemas productivos como controles de calidad y reprocesos.**

MECANISMO DE SERVICIO



Se caracteriza por las *estaciones de servicio* (servidores o dependientes) y por los *canales de servicio* que desembocan en cada uno de los servidores.

Una única cola puede desembocar en varios servidores que van siendo ocupados de acuerdo a una disciplina de selección, el caso habitual de asignación al primer servidor que queda libre.

Puede haber multiplicidad en el número de servidores, es posible encontrar múltiples colas que surtan clientes a un único o a múltiples servidores.

En una determinada estación de servicio, el cliente entra en uno de estos canales y el servidor le presta el servicio completo.

Los modelos de colas deben especificar el número de estaciones de servicio (canales de servicio en serie) y el número de servidores (canales paralelos) en cada una de ellas. Los modelos elementales se componen de una estación, ya sea con un servidor o con un número finito de ellos.

La variable más importante que caracteriza el mecanismo de servicio es el tiempo de servicio es el *tiempo de servicio*. Se denomina *tiempo de servicio* el que transcurre desde el inicio del servicio para el cliente hasta su terminación en una estación.

El modelo de un sistema de colas debe especificar la distribución de probabilidad de los *tiempos de servicio* de cada servidor, siendo habitual suponer la misma distribución para todos los servidores.

La distribución de servicio que más se emplea en la práctica (por ajustarse a un gran número de situaciones como por su simplicidad en el cálculo) es la *distribución exponencial*. Otras distribuciones qie se utilizan son la *distribución degenerada* (para tiempos de servicio constantes) y la *distribución de Erlang* (*Gamma*) para combinaciones de distribuciones exponenciales.

NOTACIÓN DEL MECANISMO DE SERVICIO

$\lambda_n \equiv$ **Tasa media de llegadas cuando hay *n* clientes en el sistema, también número esperado de llegadas de clientes por unidad de tiempo cuando se encuentran *n* clientes en el sistema.**

$\mu_n \equiv$ **Tasa media de servicio en todo el sistema, esto es, número esperado de clientes que son despachados por unidad de tiempo por todos los servidores en su conjunto.**

$s \equiv$ **Número de servidores en el sistema de colas.**

Muchas veces, el número de clientes en el sistema no afecta a la tasa media de llegadas y la tasa media de servicio. En este caso, λ_n **y** μ_n **se denotan por** λ **y** μ**, respectivamente.**

Cuando los servidores se encuentran ocupados se tiene $\mu_n = s\mu$

$u_s = \dfrac{\lambda}{\mu} \equiv$ **Utilización promedio del sistema.**

$\rho = \dfrac{\lambda}{s.\mu}$ **(factor de utilización)** $\equiv$ **Congestión de un sistema.**

El factor de utilización ρ **da una idea de la capacidad del sistema que es utilizada por los clientes entrantes.**

$\rho < 1 \rightarrow$ **Tasa de servicio > Tasa de llegada de clientes**

$\rho > 1 \rightarrow$ **Tasa de llegada de clientes > Tasa de servicio** $\rightarrow$ **La cola crece con el tiempo**

$p_n(t) = (1-\rho)\,\rho^n \equiv$ **Probabilidad de que haya n clientes en el sistema en el instante t con** $\rho = \lambda / \mu$

$N \equiv$ **Número de clientes en el sistema en estado estable.**

$N(t) \equiv$ **Número de clientes en el sistema de colas en el instante t** $(t \geq 0)$**. También, estado del sistema en el instante t.**

$Long\_cola = N(t) - s \equiv$ **Longitud de cola**

$L \equiv$ **Número esperado de clientes en el sistema, es decir, el sumatorio de las probabilidades de cada estado por el número de clientes en su correspondiente estado:** $L = \sum_{n=0}^{\infty} n p_n$

$L_q = pL \equiv$ **Longitud esperada de la cola, se trata de una variable que es medida de los clientes esperando en cola excluidos aquellos que están recibiendo servicio, se expresa por la fórmula:** $L_q = \sum_{n=s}^{\infty}(n-s)p_n$

$W \equiv$ **Tiempo de espera en el sistema incluyendo el tiempo de servicio** $(1/\mu)$ **para cada cliente. En condiciones de estabilidad, se utiliza la esperanza de la variable aleatoria:** $W = E(w) = \frac{L}{\lambda}$

$W_q \equiv$ **Tiempo de espera en la cola excluido el tiempo de servicio** $(1/\mu)$ **para cada cliente. En condiciones de estabilidad se tiene,** $W_q = E(w_q) = \frac{L_q}{\lambda}$

Suponiendo $\lambda_n = cte$, **en un proceso de colas en estado estable, el número de clientes en el sitema independientemente del tiempo transcurrido es igual a la tasa de llegadas por el tiempo de espera medio en el sistema, es decir:** $L = \lambda W$
Deduciéndose que $L_q = \lambda W_q$

Siendo el número de clientes en el sistema igual al número de clientes servidos más el número de clientes esperando en la cola: $L = L_s + L_q$

Suponiendo que el tiempo medio de servicio es una constante $(1/\mu)$ **para** $\forall n \geq 1$, **se tiene entonces que el tiempo en el sistema es igual al tiempo en cola más el tiempo de servicio** $(T_{sis} = T_{cola} + T_{serv})$: $W = W_q + \frac{1}{\mu}$

TERMINOLOGÍA DEL MECANISMO DE SERVICIO

David Kendall introdujo en 1953 una notación que permite describir las colas y mostrar sus características pudiendo clasificar los diferentes tipos de colas por medio de iniciales. De este modo, se tiene:

A / S / c / K / N / D

A **Distribución entre el tiempo de llegadas consecutivas**

- **M** **≡ Tiempos entre llegadas distribuidos de forma exponencial (Proceso de Poisson)**
- **D** **≡ Tiempos entre llegadas deterministas, con tiempo promedio constante**
- **G** **≡ Tiempos entre llegadas generales (cualquier distribución)**
- **E_k** **≡ Existe una distribución tipo Erlang**
- **H_k** **≡ Mezcla de k exponenciales**
- **P_h** **≡ Tipo fase**

S **Patrón de servicio de servidores, es decir, hace referencia a la distribución probabilística de los tiempos de servicio. Puede tomar los mismos valores que A.**

c **Número de servidores (o número de dependientes), también se denota por s.**

K **Capacidad del sistema, es decir, el número máximo de clientes que puede haber en el sistema. Cuando se trata de una cola infinita el parámetro se puede omitir.**

N **Cualquier tipo de disciplina de la cola (FIFO, LIFO, RSS, etc), se puede omitir el parámetro en caso de ser FIFO.**

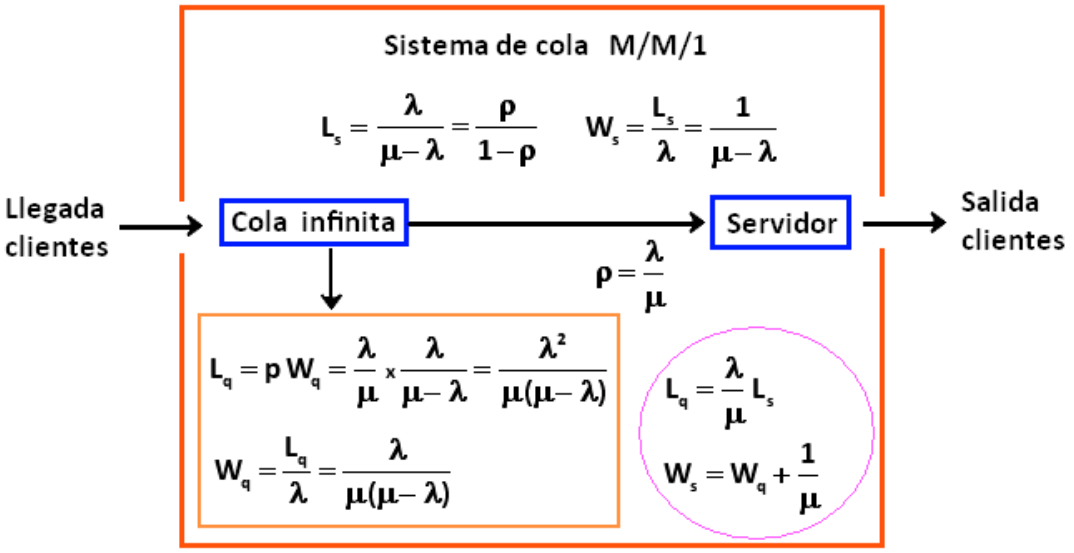
D **Tamaño de la población de entrada, en caso de ser infinita el parámetro se puede omitir.**

MODELOS DE COLAS SIMPLES

MODELO DE COLA M/M/1

El sistema de espera se caracteriza porque los tiempos de llegadas y los tiempos de servicio se distribuyen de manera exponencial y tienen un único servidor.

Según sus características la disciplina de la cola es FIFO y el tamaño de la población de entrada es infinito, es decir, el número de clientes en el sistema no afecta a la tasa de llegadas.



En el modelo M/M/1 se verifica:

El tiempo de llegadas se distribuye según $\text{Exp}(\lambda)$

El tiempo de servicio se distribuye según $\text{Exp}(\mu)$

Un único servidor $s = 1$

El Factor de utilización en el caso de un servidor $\rho = \frac{\lambda}{\mu}$ **coincide con la probabilidad de que un cliente nuevo tenga que esperar para ser servido** $p = \frac{\lambda}{\mu}$

Probabilidad de que ningún cliente se encuentre en el sistema de colas: $p_0 = 1 - \frac{\lambda}{\mu}$

Probabilidad de que *n* clientes se encuentren en el sistema de colas: $p_n = 1 - \frac{\lambda}{\mu}$

Número promedio de clientes en el sistema: $L_s = \lambda W_s = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$

Tiempo promedio de estancia en el sistema: $W_s = \frac{L_s}{\lambda} = \frac{1}{\mu-\lambda} \quad \left(W_s = W_q + \frac{1}{\mu}\right)$

Número promedio de clientes en cola: $L_q = p\,W_q = \frac{\lambda}{\mu} \times \frac{\lambda}{\mu-\lambda} = \frac{\lambda^2}{\mu(\mu-\lambda)} \quad (L_q = pL_s)$

Tiempo promedio de espera en cola: $W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$

Número de clientes servidos: $L = L_s + L_q$

Probabilidad de tiempo de espera nulo en cola: $p_0 = P(W_q = 0) = 1 - \rho$

Probabilidad de tiempo de espera en cola > t: $P(W_q > t) = \rho e^{-\mu(1-\rho)t} \quad t > 0$

Probabilidad de tiempo de estancia en el sistema > t: $P(W_s > t) = e^{-\mu(1-\rho)t} \quad t > 0$

En el mostrador de facturación de una aerolínea llega un promedio de 45 clientes por hora, cuando su capacidad media es de 60 clientes por hora. Si un cliente espera una media de 3 minutos en la cola, se pide:

a) Tiempo medio que un cliente pasa en la facturación.
b) Número medio de clientes en la cola.
c) Número medio de clientes en el sistema en un momento dado.

Solución:

a) La información de la que se dispone es:

Media de llegada de clientes: $\lambda = 45$ **clientes/hora** $= 45 / 60 = 0,75$ **clientes / minutos**

Media de servicio a clientes: $\mu = 60$ **clientes/hora** $= 60 / 60 = 1$ **clientes / minutos**

Tiempo promedio de espera en la cola: $W_q = 3$ **minutos**

El tiempo promedio que un cliente pase en el sistema $W_s = W_q + \frac{1}{\mu}$ **es:**

$W_s = 3 + \frac{1}{1} = 4$ **minutos (3 minutos en la cola + 1 minuto en el servicio)**

b) El número promedio de clientes en la cola L_q **se puede calcular:**

$L_q = \lambda W_q = 0,75 \text{ x } 3 = 2,25$ **clientes**

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{0,75^2}{1(1-0,75)} = 2,25 \text{ clientes}$$

con lo cual, puede haber más de dos clientes en la cola.

c) El número promedio de clientes en el sistema L **es:**

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{0,75}{1-0,75} = 3 \text{ clientes}$$

o también, $L_s = \lambda W_s = 0,75 \ \frac{\text{clientes}}{\text{minutos}} \text{ x } 4 \ \cancel{\text{minutos}} = 3$ **clientes**

Hay un promedio de 3 clientes en el sistema, al haber un sólo mostrador (servidor) sólo un cliente puede estar en servicio, teniendo los demás clientes que estar en la cola, lo que indica que hay 2 clientes en espera.

En un restaurante de carretera llega una media de 90 personas a la hora, cuando tiene disponibilidad de dar servicio a 120 clientes a la hora. Sabiendo que los clientes esperan una media de 2 minutos en la cola, se pide:

a) Probabilidad que el sistema se encentre sin ocupar.

b) Probabilidad que un cliente tenga que esperar al encontrarse el sistema ocupado.

c) Número medio de clientes en la cola.

d) Probabilidad de que hay 4 clientes en la cola.

Solución:

a) La información de la que se dispone es:

Media de llegada de clientes: $\lambda = 90$ **clientes/hora** $= 90 / 60 = 1,5$ **clientes / minutos**

Media de servicio a clientes: $\mu = 120$ **clientes/hora** $= 120 / 60 = 2$ **clientes / minutos**

Tiempo promedio de espera en la cola: $W_q = 2$ **minutos**

La probabilidad de que el sistema se encuentre ocioso es $(1-\rho)$**, siendo** ρ **el factor de utilización del sistema (probabilidad de que el sistema se encuentre ocupado).**

Como hay un único servidor, el factor de utilización coincide con la probabilidad de que un cliente nuevo tenga que esperar en el servicio.

$$\rho = \frac{\lambda}{\mu} = \frac{1,5 \text{ clientes / minutos}}{2 \text{ clientes / minutos}} = 0,75$$ **probabilidad de sistema ocupado**

$1-\rho = 1-0,75 = 0,25$ **probabilidad de sistema sin ocupar**

b) La probabilidad de que un cliente llegue y tenga que esperar se interpreta como que sea el primer cliente en la cola, esto es, $p_n(t) = (1-\rho)\,\rho^n$ **con** $\rho = \lambda / \mu$

Cuando $n = 1 \;\rightarrow\; p_1(t) = P(L_s = 1) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right) = (1-0,75)\,x\,0,75 = 0,1875$

Existe un 18,75% de posibilidad de que haya un cliente en la cola a la espera de ser atendido.

c) El número promedio de clientes en la cola: $L_q = \lambda\, W_q = 1,5\,x\,2 = 3$ **clientes**

d) $p_n(t) = P(L_s = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \;\rightarrow\; p_4(t) = P(L_s = 4) = (1-0,75)(0,75)^4 = 0,079$

En un lavado a presión de coches la tasa media de llegadas es de 12 coches por hora y son atendidos a una tasa promedio de 15 coches por hora, con tiempos de servicios exponenciales. Se pide:

a) Probabilidad de tener 0 clientes en el sistema.
b) Número promedio de clientes que entran en el sistema de lavado.
c) Número promedio de clientes en la cola.
d) Tiempo promedio que un cliente espera en la cola.
e) Probabilidad de tener una cola de más de 2 clientes.
f) Probabilidad de esperar más de 25 minutos en la cola y en el sistema de lavado.

Solución:

a) Es un modelo de cola M/M/1 con la siguiente información:

Media de llegada de clientes: $\lambda = 12$ **clientes/hora** $= 12/60 = 0,2$ **clientes / minutos**

Media de servicio a clientes: $\mu = 15$ **clientes/hora** $= 15/60 = 0,25$ **clientes / minutos**

El factor de utilización $\rho = \dfrac{\lambda}{\mu} = \dfrac{0,2 \text{ clientes / minutos}}{0,25 \text{ clientes / minutos}} = 0,8$

es la probabilidad de que el sistema lavado se encuentre ocupado, que al tener ún unico servidor coincide con con la probabilidad de que un cliente nuevo tenga que esperar en el servicio, es decir, $p = \dfrac{\lambda}{\mu} = 0,8$

Es decir, con probabilidad $0,2$ **el sistema de lavado está vacío, que es la probabilidad de tener 0 clientes en el sistema.**

b) El número promedio de clientes que entran en el sistema es:

$$L_s = \lambda W_s = \frac{\lambda}{\mu - \lambda} = \frac{0,2}{0,25 - 0,2} = 4 \text{ clientes}$$

con lo que el tiempo promedio de estancia en el sistema es:

$$W_s = \frac{L_s}{\lambda} = \frac{4}{0,2} = 20 \text{ minutos} \quad \text{, o bien} \quad W_s = \frac{1}{\mu - \lambda} = \frac{1}{0,25 - 0,20} = 20 \text{ minutos}$$

c) El número promedio de clientes en la cola: $L_q = p\, L_s = 0,8 \times 4 = 3,2$ **clientes**

d) El tiempo promedio que un cliente espera en la cola:

$$W_q = \frac{L_q}{\lambda} = \frac{3,2}{0,2} = 16 \text{ minutos} \quad , \quad W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{0,2}{0,25\,(0,25 - 0,2)} = 16 \text{ minutos}$$

e) Para calcular la probabilidad de tener una cola de más de 2 clientes se necesita saber la probabilidad de que haya 0, 1 y 2 clientes.

donde, $p_n(t) = (1-\rho)\,\rho^n$ **con** $\rho = \lambda / \mu = 0,8$

$$n = 0 \;\rightarrow\; p_0(t) = P(L_s = 0) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^0 = (1-0,8) \times 0,8^0 = 0,2$$

$$n = 1 \;\rightarrow\; p_1(t) = P(L_s = 1) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^1 = (1-0,8) \times 0,8^1 = 0,16$$

$$n = 2 \;\rightarrow\; p_2(t) = P(L_s = 2) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^2 = (1-0,8) \times 0,8^2 = 0,128$$

$$P(L_s > 2) = 1 - P(L_s > 2) = 1 - \left[P(L_s = 0) + P(L_s = 1) + P(L_s = 2)\right] =$$
$$= 1 - \left(0,2 + 0,16 + 0,128\right) = 0,512$$

Existe un 51,2% de posibilidad de encontrar una cola con más de dos cleintes.

e) La probabilidad del tiempo de espera de un cliente en el sistema:

$$P(W_s > t) = e^{-\mu(1-\rho)t} \;\rightarrow\; P(W_s > 25) = e^{-0,25(1-0,8)25} = e^{-1,25} = 0,286$$

La probabilidad del tiempo de espera de un cliente en la cola:

$$P(W_q > t) = \rho\, e^{-\mu(1-\rho)t} \;\rightarrow\; P(W_q > 25) = 0,8\, e^{-0,25(1-0,8)25} = 0,8\, e^{-1,25} = 0,23$$

Una sucural bancaria decide instalar un cajero en un barrio de ciudad que no tiene un servicio semejante. En la investigación inicial se recogen datos diariamente sobre los tiempos de llegadas de los clientes, detectando que el cliente que llega prefiere esperar para poder utilizar el servicio cuando éste se encuentra ocupado. Con los datos recogidos, se estima que las llegadas siguen un proceso de Poisson, el tiempo de servicio es exponencial, el tiempo medio de servicio es de 5 minutos por cliente y el tiempo medio transcurrido entre dos llegadas consecutivas es de 8 minutos. Se pide:

a) Tiempo promedio de espera que debe sufrir cada cliente en cola.

b) Tamaño promedio de la cola y probabilidad de que al acudir al cajero haya alguna persona en la cola.

Solución:

a) Es un modelo de cola M/M/1 con $\lambda = 1/8 = 0{,}125$ y $\mu = 1/5 = 0{,}2$

El tiempo promedio de espera en cola: $W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{0{,}125}{0{,}2\left(0{,}2-0{,}125\right)} = 8{,}33 \text{ minutos}$$

b) El tamaño promedio de la cola se refleja por el número medio de clientes en la cola:

$$W_q = L_q / \lambda \quad \rightarrow \quad L_q = \lambda W_q = 0{,}125 \times 8{,}33 = 1{,}04 \text{ clientes}$$

La probabilidad de que al acudir al cajero haya alguna persona en la cola es $1 - p_0 - p_1$, donde $p_n(t) = (1-\rho)\,\rho^n$

El factor de utilización $\rho = \lambda / \mu = 0{,}125 / 0{,}2 = 0{,}625$ es la probabilidad de que el cajero se encuentre ocupado, que al tener ún unico servidor coincide con con la probabilidad de que un cliente nuevo tenga que esperar en el servicio, es decir, $p = \lambda / \mu = 0{,}625$

$$n = 0 \quad \rightarrow \quad p_0(t) = P(L_s = 0) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^0 = (1-0{,}625) \times 0{,}625^0 = 0{,}375$$

$$n = 1 \quad \rightarrow \quad p_1(t) = P(L_s = 1) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^1 = (1-0{,}625) \times 0{,}625^1 = 0{,}234$$

$$P(L_s > 1) = 1 - P(L_s > 1) = 1 - \left[P(L_s = 0) + P(L_s = 1)\right] = 1 - 0{,}375 - 0{,}234 = 0{,}391$$

Existe un 39,1% de posibilidad de encontrar alguna persona en la cola.

Una base de mantenimiento de aviones tiene recursos para revisar únicamente un motor de avión a la vez. Para devolver los aviones lo antes posible, sigue la política de revisar solo un motor de los cuatro motores de los aviones que llegan a la base según una distribución de Poisson de tasa media uno al día. El tiempo requerido para revisar un motor (una vez que comienza el trabajo) sigue una distribución exponencial de tasa 1/2 al día.
Se ha hecho una propuesta para cambiar la política de revisión, de forma que se revisen los cuatro motores de forma consecutiva cada vez que un avión llegue a la base, que supone cuadriplicar el tiempo esperado de servicio, con una frecuencia de revisión de cada avión cuatro veces menor.
Se pide comparar las dos alternativas aplicando la teoría de colas.

Solución:

En los dos casos se trata de colas M/M/1, puesto que tanto los tiempos entre llegadas como los tiempos de servicio son variables aleatorias con distribución exponencial.

- **En la situación actual, la tasa de llegadas es $\lambda = 1/1 = 1$ aviones al día, y la tasa de servicio es $\mu = 1/(1/2) = 2$ aviones al día.**

Factor de utilización: $\rho = \frac{\lambda}{\mu} = \frac{1}{2} = 0,5$

Número promedio de aviones en el sistema: $L_s = \frac{\lambda}{\mu - \lambda} = \frac{1}{2-1} = 1$ **avión**

Número promedio de aviones en cola: $L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{1}{2(2-1)} = \frac{1}{2}$ **avión**

Tiempo promedio de estancia en el sistema: $W_s = \frac{1}{\mu - \lambda} = \frac{1}{2-1} = 1$ **día**

Tiempo promedio de espera en cola: $W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{1}{2(2-1)} = \frac{1}{2}$ **día**

- **Con la propuesta de cambiar de política de revisión, la tasa de llegada es $\lambda = 1/4 = 0,25$ aviones al día, y la tasa de servicio es $\mu = 1/(4/2) = 0,5$ aviones al día.**

Como el factor de utilización: $\rho = \frac{\lambda}{\mu} = \frac{0,25}{0,5} = 0,5 < 1$ **es el mismo, el estado sigue siendo estacionario.**

Número promedio de aviones en el sistema: $L_s = \frac{\lambda}{\mu - \lambda} = \frac{0,25}{0,5 - 0,25} = 1$ **avión**

Número promedio de aviones en cola: $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{0,25^2}{0,5(0,5-0,25)} = \frac{1}{2}$ **avión**

Tiempo promedio de estancia en el sistema: $W_s = \frac{1}{\mu-\lambda} = \frac{1}{0,5-0,25} = 4$ **días**

Tiempo promedio de espera en cola: $W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{0,25}{0,5(0,5-0,25)} = 2$ **días**

CONFIGURACIONES DE MANTENIMIENTO:

Actual:	$L_s = 1$ **avión**	$L_q = 1/2$ **avión**	$W_s = 1$ **día**	$W_q = 1/2$ **día**
Propuesta:	$L_s = 1$ **avión**	$L_q = 1/2$ **avión**	$W_s = 4$ **días**	$W_q = 2$ **días**

Con la propuesta de cambio se observa que cada vez que un avión vaya a ser revisado pasará en el sistema el cuádruple del tiempo que pasaba con el sistema anterior, pero como cada avión va a ir con frecuencia cuatro veces menor, el tiempo perdido en el taller a largo plazo va a ser igual.

En este caso, la decisión entre las dos configuraciones se toma en función de los costes de operación.