**CARRERA INGENIERÍA INDUSTRIAL**

**ASIGNATURA: INVESTIGACIÓN DE OPERACIONES II**

**Tema 3. Sistemas de Servicios.**
**3.1. Introducción.**
**3.2. Conceptos básicos de los sistemas de servicios.**
**3.3. Modelo de una fila y un servidor con tasa de llegada según distribución de Poisson y tasa de servicio según distribución exponencial.**
**3.3.1. Tamaño óptimo de la brigada.**
**3.4. Modelo de un servidor con tasa de servicio constante.**
**3.4.1. Comparación entre los sistemas con tasa de servicio exponencial t tasa de servicio constante**
**3.5 Modelo con múltiples servidores**

**Objetivos: Que el alumno se capaz de:**

- **Resolver problemas de los diferentes modelos de sistemas de servicios aplicando las ecuaciones matemáticas e interpretando sus resultados.**

**Método: Elaboración conjunta.**

**Medios: Tiza y pizarra, Libro de texto, MDD, Pc.**

**Tipo de clase: Clase Encuentro**

**Bibliografía: introducción a la Investigación de Operaciones.**

**Materiales Didácticos Digitales entregados por el profesor.**

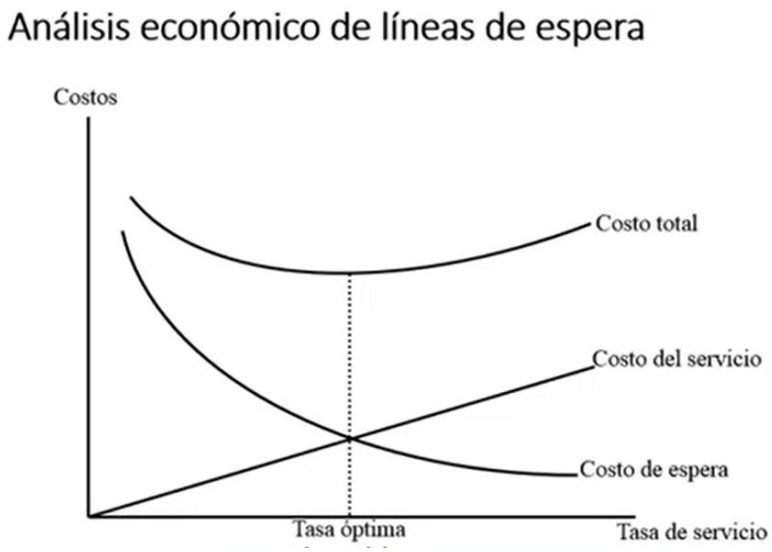
**DESARROLLO**

## 3.1. INTRODUCCIÓN.

Los sistemas de servicios tambien conocidos como sistemas de linea de espera (colas), son sistemas aleatorios o probabilisticos, su analisis se fundamente en la Teoria de las colas, en la que agrupa una coleccion de modelos matemáticos que describen el comportamiento de la linea de espera.

Todos indendientes del sistema economico social nos hemos encontrado en algun momento a una cola, y es conocida que normalmente genera una experiencia que se puede considerar desagradable, al punto que muchas personas desisten de entra a la fila o se retiran de ella por el costo de espera que la fila entraña, lo que trae consigo que los encargados de ofertar el servicio ven una perdida de oprtunidad de incrementar sus ganacias.

Esta es una razon por la que los administradores de los sistemas de servicio se empeñan en reducir estas perdidas, equilibrando el costo de espera con el costo que implica el servicio que ofrecen.




### 3.2. Conceptos básicos de los sistemas de servicios.

**Costo de espera**

Esperar es estar inútil. Es desperdicio. Significa que algún recurso está inactivo cuando podría usarse en forma más productiva (o agradable) en otra parte. De hecho, representa un costo de *oportunidad.* Cuando los camiones están esperando inútiles en una línea de un muelle de carga y descarga, se pierde su productividad; es dinero "que se va por el caño" y que no puede recuperarse.

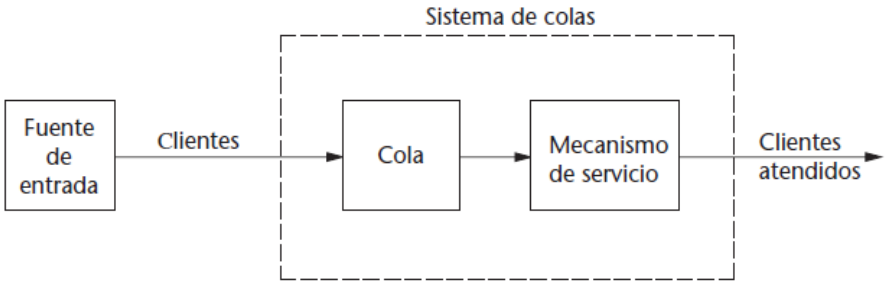
Costo de espera = $C_W * L$

en donde:
Cw = costo de espera en dólares por llegada por unidad de tiempo y
L = longitud promedio de la línea.
Si, por ejemplo, el costo de un camión que espera en la línea es de $20 por hora (incluso el operador) y en promedio hay cuatro camiones esperando, entonces el costo de espera total es $80 por hora.

**Costo de servicio**

Determinar el costo de servicio es más sencillo, en concepto, que determinar el costo de espera. En la mayoría de las aplicaciones se tratará de comparar varias instalaciones de servicio: dos cajeras en un banco contra tres; una brigada de cuatro contra una de cinco; una caja en una tienda contra dos. En estos casos, sólo se necesitan los costos comparativos o diferenciales. Por ejemplo, si se quiere saber cuántas cajas de autobanco deben tener personal, sólo se necesitan los costos de personal. Por otra parte, si la pregunta es cuántas de estas cajas se deben construir, entonces se necesitan los costos de construcción y los de operación de cada ventanilla. Casi siempre los datos necesarios son evidentes para una situación dada.

Un sistema de servicio puede dividirse en sus dos componentes de mayor importancia: la cola o línea de espera y la instalación de servicio como se muestra en la figura siguiente




Entre sus conceptos básicos esta:

**Tamaño de la población** de donde se alimentará el sistema que puede ser finita (ejemplo el mantenimiento que se dará a 25 computadoras de un laboratorio) o infinita, (los autos que llegan a un puesto de peaje). Se asumirá en este tema, para el modelaje, que esta población será infinita

**Proceso de llegada,** las formas en que van llegando los clientes para recibir el servicio, puede ser determinísticas (se conocen con exactitud la cantidad de clientes por unidad tiempo que van a llegar, por ejemplo en las fábricas de ensamblaje se conoce el tiempo en pasar un vehículo de un punto de ensamblaje a otro punto). Otra forma de llegada es probabilística (no se conoce el tiempo que trascurre en la llegada de un cliente a otro a un restaurante, son eventos independientes entre si), en este tema se asumirán que los modelos siguen una distribución de probabilidades de Poisson.

La tasa de llegada o velocidad de llegada (λ): son la cantidad de llegada promedio por unidad de tiempo.

**Longitud de la cola,** numero máximo permisible de clientes que puede admitir. Las colas pueden ser *finitas* o *infinitas*, según si dicho numero es finito o infinito. El supuesto de una *cola infinita* es el estándar de la mayoría de los modelos.

**Disciplina de la cola** se refiere al orden en el que sus miembros se seleccionan para recibir el servicio, por ejemplo, puede ser:

- primero en entrar, primero en salir;
- primero en llegar último en salir;
- de acuerdo con algún procedimiento de prioridad;
- aleatoria; o con algún otro orden.

En los modelos de colas se supone como normal a la disciplina de primero en entrar, primero en salir, a menos que se establezca de otra manera

**Comportamiento de llegada**, los modelos suponen que los clientes son infinitamente pacientes, no hay rechazo a la fila y no renuncian o no se retiran de la cola.

**El proceso de servicio** consiste en una o mas *estaciones de servicio*, cada una de ellas con uno o mas *canales de servicio paralelos*, llamados **servidores**. Si existe mas de una estación de servicio, el cliente puede recibirlo de una secuencia de ellas (*canales de servicio en serie*). En una estación dada, el cliente entra en uno de estos canales y el servidor le presta el servicio completo. Los modelos de colas deben especificar el arreglo de las estaciones y el numero de servidores (canales paralelos) en cada una de ellas. Los modelos

más elementales suponen una estación, ya sea con un servidor o con un numero finito de servidores.

**Tiempo de servicio**) (o *duración del servicio*) es el tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación en una estación. El tiempo de servicio puede ser determinístico (por ejemplo un grupo de clientes entran a una tanda de cine y se conoce el tiempo que esta durará el servicio de entretenimiento).
También puede ser probabilístico, la distribución del tiempo de servicio que más se usa en la práctica (por ser más manejable que cualquier otra) es la distribución *exponencial* todos los modelos de este capítulo serán de este tipo.
La tasa de servicio o velocidad de servicio (μ) es la cantidad de clientes que se atienden en la unidad de tiempo

El modelo supone que la tasa de servicio será siempre superior a la tasa de llegada

$$\mu > \lambda$$

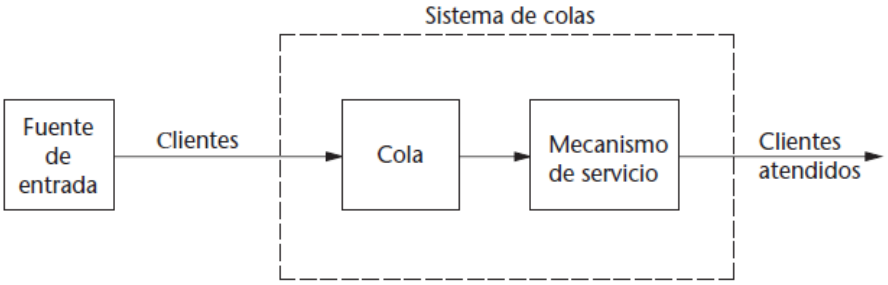
Número de servidores son la cantidad de servidores que pueden atender a los clientes y no tienen límites.

**La Configuración** es el número de fases de servicio, por ejemplo una sola fase o varias fases

Ejemplo los pacientes llegan al hospital son atendidos por el médico y pasan a la fila de la farmacia para adquirir el medicamento.




### 3.3. Modelo de una fila y un servidor con tasa de llegada según distribución de Poisson y tasa de servicio según distribución exponencial.



Número promedio de unidades en la línea de espera

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

Tiempo promedio que la unidad pasa en la línea de espera

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

Número promedio de unidades en el sistema

$$L_s = L_q + \frac{\lambda}{\mu}$$

Tiempo promedio que una unidad pasa en el sistema

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

Probabilidad de que una unidad que llega no tenga que esperar a ser atendida

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

Probabilidad de que no haya unidades en el sistema

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

Probabilidad de que haya n unidades en el sistema

$$P_n = \left(\frac{\lambda}{\mu}\right)^2 P_o$$

**Ejemplo 1. Un supermercado**

Imagínese un supermercado grande con muchas cajas de salida. Supóngase que los clientes llegan para que les marquen su cuenta con una tasa de 90 por hora y que hay 10 cajas en operación. (Nótese que una familia junta de compras se trata como un cliente.) Si hay poco intercambio entre las líneas, puede tratarse este problema como 10 sistemas separados de una sola línea, cada uno con una llegada de 9 clientes por hora. Para una tasa de servicio de 12 por hora:

**Datos: $\lambda$ = 9 clientes por hora**
**$\mu$ = 12 clientes por hora**

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{(9)^2}{12\,(12 - 9)} = 2.25 \text{ clientes}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{9}{12(12 - 9)} = 0.25 \text{ horas o } 15 \text{ minutos}$$

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{9}{12 - 9} = 3 \text{ clientes}$$

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{12 - 9} = 0.33 \text{ horas o } 20 \text{ minutos}$$

$$U = \frac{\lambda}{\mu} = \frac{9}{12} = 0.75 \text{ o } 75\%$$

$$P(L_s > 3) = \left(\frac{A}{\mu}\right)^{n+1} = \left(\frac{9}{12}\right)^{3+1} = 0.32$$

Entonces, para este ejemplo, el cliente promedio espera 15 minutos antes de ser servido (algunos esperan más, otros menos). En promedio, hay un poco más de dos clientes en la línea o tres en el sistema. El proceso completo lleva un promedio de 20 minutos. La caja está ocupada el 75% del tiempo (así, estará desocupada o inútil el 25% del tiempo). Y finalmente, el 32% del tiempo habrá cuatro personas o más en el sistema (o tres o más esperando en la cola)

### 3.3.1. Tamaño óptimo de la brigada.

**Ejemplo 2.** Se está estudiando un muelle de carga y descarga de camiones para aprender cómo debe formarse una brigada. El muelle tiene espacio sólo para un camión, así es un sistema de un servidor. Pero el tiempo de carga o descarga puede reducirse aumentando el tamaño de la brigada.
Supóngase que puede aplicarse el modelo de un servidor y una cola (llegadas Poisson, tiempos de servicio exponenciales) y que la tasa promedio de servicio es un camión por hora para un cargador. Los cargadores adicionales aumentan la tasa de servicio proporcionalmente. Además, supóngase que los camiones llegan con un tasa de dos por hora en promedio y que el costo de espera es de $20 por hora por camión. Si se le paga $5 por hora a cada miembro de la brigada, ¿cuál debe ser el mejor tamaño de esta?

[2] Usando cálculo, es fácil demostrar que la tasa de servicio óptima está dada por:

$$S = A + \frac{AC_w}{C_s}$$

en donde $C_w$ = costo unitario de espera y $C_s$ = costo unitario de servicio.

λ = A = 2 camiones por hora
μ = S = 1 camión por persona por hora
*Cw* = costo de espera = $20 por hora por camión
*Cs* = costo de servicio = $5 por hora por persona
Ahora sea *k* = número de personas en la brigada. Se busca *k* tal que la suma de los costos de espera y servicio se minimicen: Costo total = *CWLs + kCs.*
***Para esto se emplea el proceder de ensayo error o sea tanteando hasta encontra el punto de valor mínimo****.*

Nótese que se usa la longitud de la línea del *sistema,* porque el camión está inútil tanto si espera como si está siendo servido.
Las pruebas se deben iniciar con tres miembros en la brigada, ya que uno o dos no podrían compensar la tasa de llegadas de dos camiones por hora. Para una brigada de tres, la tasa de servicio es de tres camiones por hora y puede encontrarse L, con la ecuación

$$L_s = \frac{\lambda'}{\mu - \lambda'} \qquad \frac{2}{3 - 2} = 2.0 \text{ camiones}$$

$$\begin{aligned} \text{Costo total (3)} &= C_w L_s + kC_s \\ &= (20)\,(2.0) + (3)\,(5) \\ &= \$55 \end{aligned}$$

De igual manera para una brigada de 4

$$L_s = \frac{\lambda'}{\mu - \lambda'} \qquad \frac{2}{4 - 2} = 1.0 \text{ camiones}$$

$$\begin{aligned} \text{Costo total (4)} &= C_w L_s + kC_s \\ &= (20)(1.0) + (4)(5) \\ &= \$40 \end{aligned}$$

El costo total continua disminuyendo por tanto probaremos con 5 hombres

$$L_s = \frac{\lambda'}{\mu - \lambda'} \qquad \frac{2}{5\text{-}2} = 0.67 \text{ camiones}$$

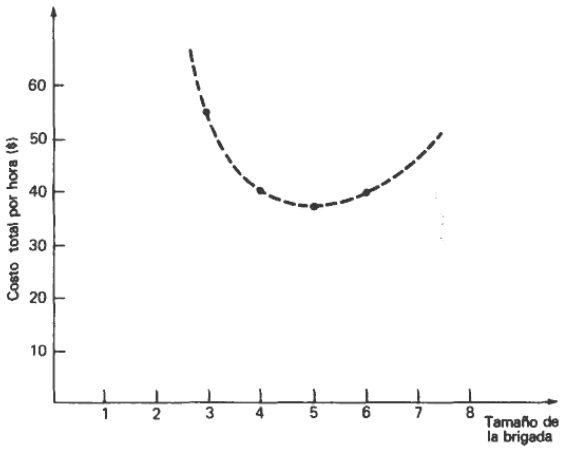
$$\begin{aligned} \text{Costo total (5)} &= C_w L_s + kC_s \\ &= (20)(0.67) + (5)(5) \\ &= \$38.33 \end{aligned}$$

El costo total continúa disminuyendo se prueba con 6 hombres

$$L_s = \frac{\lambda'}{\mu - \lambda'} \qquad \frac{2}{6 - 2} = 0.5 \text{ camiones}$$

$$\begin{aligned} \text{Costo total (6)} &= C_w L_s + kC_s \\ &= (20)(0.5) + (6)(5) \\ &= \$40 \end{aligned}$$

Como este costo es mayor que el de la brigada de cinco, se rebasó el límite inferior de la curva de costo; el tamaño óptimo de la brigada es cinco personas. Los resultados se muestran en la figura




**Ejemplo 3. Restaurante de comida rápida**

Considérese un restaurante de comida rápida con un menú limitado. El restaurante se está diseñando para que todos los clientes se unan a una sola línea para ser servidos. (Igual que en el supermercado, una *orden* es una llegada sin importar cuántas personas comparten esa orden.) Una persona tomará la orden y la servirá. Con sus limitaciones, la tasa de servicio puede aumentarse agregando más personal para preparar la comida y servir las órdenes.
Esto constituye un sistema de un servidor y una línea. Si las llegadas y las salidas son aleatorias, puede aplicarse el modelo de una cola. Supóngase que la administración quiere que el cliente promedio no espere más de dos minutos antes que se tome su orden. Esto se expresa como: Wq = 2 minutos
Supóngase también que la tasa máxima de llegadas es 30 órdenes por hora.

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$ despejando se obtiene $$\mu = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{W_q}}$$

Para este ejemplo, se supuso:
λ = 30 órdenes por hora b
Wq = 2 minutos o 0.033 horas

$$S = \frac{30}{2} + \sqrt{\frac{(30)^2}{4} + \frac{30}{0.033}}$$
$$= 15 + 33.5 = 48.5 \text{ órdenes por hora}$$

Para cumplir los requerimientos, se necesita una tasa de servicio de casi 50 órdenes por hora. Si, por ejemplo, una brigada de cinco puede manejar 45 órdenes por hora y una de seis puede procesar 50 por hora, entonces sería necesario tener la brigada de seis.

### 3.4. Modelo de un servidor con tasa de servicio constante.

Este modelo es el mismo, excepto que se supone que el tiempo de servicio es exactamente el mismo para cada llegada en lugar de ser aleatorio.
Todavía se tiene una sola línea, tamaño de la cola infinito, disciplina de la cola como primero en llegar primero en ser servido y llegadas Poisson.
Las aplicaciones típicas de este modelo pueden incluir un autolavado automático, una estación de trabajo en una pequeña fábrica o una estación de diagnóstico de mantenimiento preventivo. En general, el servicio lo proporciona una máquina.
Las características de operación están dadas por:

Número promedio de unidades en la línea de espera

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

El resto de las expresiones son las mismas a partir de esta $\mathbf{L_q}$ calculada

### Ejemplo 4. Lavado automático de autos

Imagínese un lavado automático de autos con una línea de remolque, de manera que los autos se mueven a través de la instalación de lavado como en una línea de ensamble. Una instalación de este tipo tiene dos tiempos de servicio diferentes:
el tiempo entre autos y el tiempo para completar un auto. Desde el punto de vista de teoría de colas, el tiempo entre autos establece el tiempo de servicio del sistema. Un auto cada cinco minutos da una tasa de 12 autos por hora. Sin embargo, el tiempo para procesar un auto es el tiempo que se debe esperar para entregar un auto limpio. La teoría de colas no considera este tiempo.
Supóngase que el lavado de autos puede aceptar un auto cada cinco minutos y que la tasa promedio de llegadas es de nueve autos por hora (con distribución Poisson). Sustituyendo en las ecuaciones:
$\lambda$ = 9 autos por hora
5 minutos o $\mu$ = 12 autos por hora lad de tiempo)

$$L_q = \frac{\lambda^2}{2\mu(\mu-\lambda)} = \frac{(9)^2}{2(12)(12-9)} = 1.125 \text{ autos}$$

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

1.125 / 9 = 0.125 horas igual a 7.5 min (0.125 h * 60min/h)

$$L_s = L_q + \frac{\lambda}{\mu}$$

1.125 + 9/12 = 1.875 autos

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

0.125 h + 1/9 =

### 3.4.1. Comparación entre los sistemas con tasa de servicio exponencial t tasa de servicio constante.

Comparación entre el modelo con tiempo de servicio exponencial y el tiempo de servicio constante.

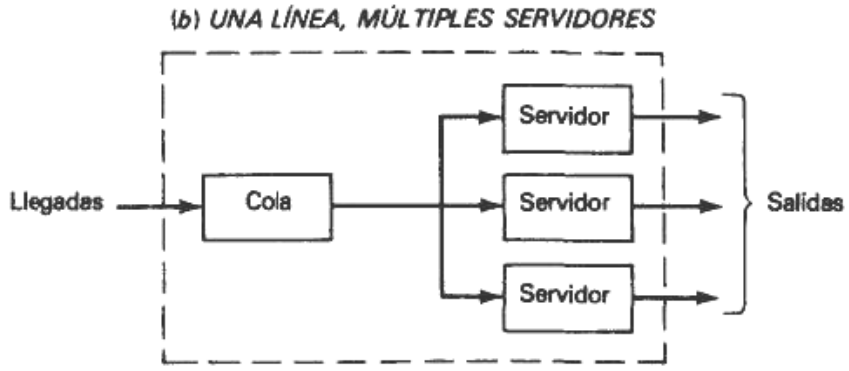
| | *Tiempos de servicio exponenciales* | *Tiempos de servicio constantes* |
|---|---|---|
| *Cola* | | |
| Longitud promedio de la línea | 2.25 | 1.125 |
| Tiempo promedio de espera | 15 min | 7.5 min |
| *Sistema* | | |
| Longitud promedio de la línea | 3 | 1.875 |
| Tiempo promedio de espera | 20 min | 12.5 min |
| Utilización | 75% | 75% |

Nótese que el sistema de tiempos de servicio constantes tiene sólo la mitad de la longitud de cola y del tiempo de espera de lo que tiene el sistema de tiempos de servicio variables. La

longitud de la línea y el tiempo de espera para todo el sistema también son menores; sólo la utilización es la misma para ambos.
¿Por qué un sistema con tiempos de servicio constantes da mucho mejores resultados que el de tiempos de servicio variables? Porque se ha eliminado parte de la aleatoriedad del sistema y con ella la posibilidad de tiempos de servicio muy largos.

### 3.5 Modelo con múltiples servidores

En muchas situaciones reales habrá más de un servidor disponible para atender las llegadas. Esto es típico de los supermercados y los bancos.
Existe una solución general para un sistema de múltiples servidores que tiene una sola línea. Esta situación se muestra en la figura




$N$ = número de servidores
A = tasa promedio de llegadas (llegadas por unidad de tiempo) ($\lambda$)
S = tasa promedio de servicio por cada servidor (llegadas por unidad de tiempo) ($\mu$)

entonces

$$L_q = \frac{(A/S)^{N+1} P_o}{(N-1)!\left(N - \frac{A}{S}\right)^2}$$

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

$$L_s = L_q + \frac{A}{S}$$

$$W_s = \frac{L_s}{A}$$

$$U = \frac{A}{NS}$$

$$P_o = \frac{1}{\frac{1}{N!}\left(\frac{A}{S}\right)^N \left(\frac{NS}{NS - A}\right) + \sum_{i=0}^{N-1} \frac{1}{i!}\left(\frac{A}{S}\right)^i}$$

## CONCLUSIONES

- En el tema se estudiaron tres modelos con tamaño de la fila infinita: llegadas probabilísticas según la distribución de Poisson: con tiempo de servicio probabilístico según la distribución exponencial; el supuesto de primero en llegar, primero en salir.
- Los modelos estudiados son:
  - Una fila y un servidos con llegas según Poisson y tiempo de servicio según la distribución exponencial.
  - Una fila y un servidos con llegas según Poisson y tiempo de servicio constante.
  - Una fila y varios servidos con llegas según Poisson y tiempo de servicio según la distribución exponencial.
- Los sistema con tiempos de servicio constantes da mucho mejores resultados que el d

**Modelo 1 servidor y llegada probabilística**

Probabilidad de que no haya unidades en el sistema

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

Número promedio de unidades en la línea de espera

Para tasa de servicio exponencial

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

Para tasa de servicio constante

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

Número promedio de unidades en el sistema

$$L = L_q + \frac{\lambda}{\mu}$$

Tiempo promedio que la unidad pasa en la línea de espera

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

Tiempo promedio que una unidad pasa en el sistema

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

Probabilidad de que una unidad que llega no tenga que esperar a ser atendida

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

Probabilidad de que haya n unidades en el sistema

$$P_n = \left(\frac{\lambda}{\mu}\right)^2 P_o$$

**Modelo múltiples servidores y llegada probabilística**

$N$ = número de servidores

A = tasa promedio de llegadas (llegadas por unidad de tiempo) ($\lambda$)

S = tasa promedio de servicio por cada servidor (llegadas por unidad de tiempo) ($\mu$)

entonces

$$L_q = \frac{(A/S)^{N+1} P_o}{(N-1)!\left(N - \frac{A}{S}\right)^2}$$

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

$$L_s = L_q + \frac{A}{S}$$

$$W_s = \frac{L_s}{A}$$

$$U = \frac{A}{NS}$$

$$P_o = \frac{1}{\frac{1}{N!}\left(\frac{A}{S}\right)^N\left(\frac{NS}{NS \cdot A}\right) + \sum_{i=0}^{N-1}\frac{1}{i!}\left(\frac{A}{S}\right)^i}$$

Activar Windows

e tiempos de servicio variables, porque se ha eliminado parte de la aleatoriedad del sistema y con ella la posibilidad de tiempos de servicio muy largos.

**ESTUDIO INDEPENDIENTE**

En una oficina la tasa media de llegada de clientes es de 3 cada 30 minutos, además se sabe que atiende 2 personas cada 15 minutos.

Determine Ls, Ws, Lq, Wq