INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE FÍSICA Y MATEMÁTICAS

"INTRODUCCIÓN A PROCESOS ESTOCÁSTICOS Y SISTEMAS DE LÍNEAS DE ESPERA"

TESIS
QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN FÍSICA Y MATEMÁTICAS

PRESENTA

JULIA AGUEDA ROSETE LIMA

Director de Tesis: Dr. Roberto S. Acosta Abreu

México D. F **Marzo de 2009**

ÍNDICE

INTRODUCCIÓN

ANTECEDENTES

La teoría de colas incluye el estudio matemático de colas o líneas de espera y provee un gran número de modelos matemáticos para describirlas.

Generalmente el administrador tiene que tomar decisiones entre

- Asumir los costos derivados de prestar un buen servicio
- Asumir los costos derivados de tener largas colas

Para esto debe lograr un balance económico entre el costo del servicio y el costo asociado a la espera por el servicio.

La teoría de colas en sí no resuelve este problema, pero proporciona la información necesaria para poder tomar decisiones.

PROBLEMÁTICA

El problema fundamental en casi todas las líneas de espera tiene que ver con el equilibrio.

El administrador debe sopesar el costo adicional de proporcionar un servicio más rápido (más carriles de tránsito, pistas de aterrizaje adicionales, más mostradores de registro de salidas) contra el costo inherente a la espera.

Con frecuencia, el costo de esta decisión es directo. Por ejemplo, si encontramos que el tiempo total que pasan los empleados en una fila para poder utilizar una copiadora puede dedicarse a actividades más productivas, compararíamos el costo de instalar una máquina adicional contra el valor del tiempo que se ahorran los empleados. Después de esto la decisión se reduce al costo en pesos, lo cual facilita la elección.

Por otra parte, supongamos que los problema de la línea de espera es la demanda de camas en un hospital, en este caso no podemos simplemente calcular el costo de las camas adicionales sumando los costos de construcción del edificio, equipo adicional requerido y el incremento en el mantenimiento, ya que, de hacerlo así, ¿qué pondríamos del otro lado de la balanza? En este caso se enfrenta el problema de tratar de asignarle una cifra en pesos a la necesidad del paciente de una cama de hospital que no está disponible. Aun cuando podemos estimar los ingresos perdidos para el hospital, ¿qué hay sobre el humano que sufre por esta falta de atención adecuada en el hospital?

Los problemas de decisión de las líneas de espera pueden ser de 3 tipos:

1. Dada una función de costo de espera, una función de costo de servicio del sistema, los parámetros λ y μ, se desea encontrar un número óptimo de servidores *s* que minimiza el costo total esperado
2. Dada una función marginal de servicio por unidad de tiempo para μ fija, el valor de λ y un rango permisible de variación de μ, se desea encontrar el número de servidores *s,* y el valor de μ que minimizan el costo total
3. Dado el costo total de servicio por unidad de tiempo, un costo de servicio fijo por unidad de servicio por unidad de tiempo, un valor λ y μ, se desea encontrar un número de estaciones de servicio y el número de servidores por estación que minimiza el costo total

OBJETIVOS

El objetivo de la teoría de colas consiste en responder a intereses administrativos pertenecientes al diseño y a la operación de un sistema de colas.

Loa objetivos particulares del presente trabajo son:

a. Presentar el desarrollo de la fundamentación de los modelos matemáticos de líneas de espera. Esto se hace en base en la de procesos de Markov.

b. Desarrollar los principales modelos de las líneas de espera, su uso y clasificación fundamentado en las necesidades clientes-servidor(es) y la toma de decisiones basada en los costos de espera y servicio

c. Presentar ejemplos y aplicaciones de los modelos de líneas de espera.

DESCRIPCIÓN DEL TRABAJO

A continuación describiremos brevemente la estructura el trabajo

En el capítulo I se presentan los fundamentos de procesos estocásticos considerando los procesos de Markov y en particular los procesos de nacimiento y muerte que son fundamentales para el presente trabajo.
En el capítulo II se da la fundamentación de los sistemas de líneas de espera se describe la escritura y las medidas de rendimiento para evaluar un sistema de líneas de espera.
En el capítulo III se considera el uso de diferentes modelos de líneas de espera tomando en cuenta diferentes factores como el número de servidores, la modalidad del servicio y la capacidad del sistema.
En el capítulo IV se introduce el costo de los sistemas de líneas de espera. Se consideran el costo de espera, costo del servicio y se ve un sistema de costo mínimo.
Finalmente en el capítulo V se presentan diversos ejemplos y aplicaciones de sistemas de espera y se dan las conclusiones del trabajo

CAPÍTULO I

FUNDAMENTOS DE PROCESOS ESTOCÁSTICOS

La teoría de los procesos estocásticos se centra en el estudio y modelización de sistemas que evolucionan a lo largo del tiempo, o del espacio, de acuerdo a unas leyes no determinísticas, esto es, de carácter aleatorio.
La forma habitual de describir la evolución del sistema es mediante sucesiones o colecciones de variables aleatorias. De esta manera, se puede estudiar cómo evoluciona una variable aleatoria a lo largo del tiempo. Por ejemplo, el número de personas que espera ante una ventanilla de un banco en un instante *t* de tiempo; el precio de las acciones de una empresa a lo largo de un año.

La primera idea básica es identificar un proceso estocástico con una sucesión de variables aleatorias $\{X_t : n \in \mathbb{R}\}$, €donde el subíndice indica el instante de tiempo (o espacio) correspondiente.
Esta idea inicial se puede generalizar fácilmente, permitiendo que los instantes de tiempo en los que se definen las variables aleatorias sean continuos. Así, se podrá hablar de una colección o familia de variables aleatorias $\{X_t : t \in \mathbb{R}\}$, que da una idea más exacta de lo que es un proceso estocástico.

1.1 CONCEPTOS Y CLASIFICACIÓN

Definición: Un *proceso estocástico* esta definido por una colección de variables aleatorias $X_t : t \in T$ donde:

$T \subseteq \mathbf{R}$

t: es el parámetro que se asocia al tiempo

X_t: representa el estado de proceso en el instante *t*.

Ejemplo:

- Si X_t representa la distancia entre dos puntos que se mueven aleatoriamente sobre una recta, entonces $T = \mathbb{R}^+$ y $X_t \mathbb{R}^+$
- Si X_t representa el piso en el que se encuentra un ascensor después de la t-enésima parada, entonces, $T = \mathbb{Z}^+$ y $X_t \mathbb{Z}^+$
- Si X_t representa el número de llamadas a un número telefónico hasta el instante de tiempo t, entonces $T = \mathbb{R}^+$ y $X_t \mathbb{R}^+$

Para que un proceso estocástico esté completamente definido hay que determinar todas las variables aleatorias, es decir, determinar e identificar la distribución de probabilidad asociada a cada una y la distribución conjunta de todas ellas.
Los procesos estocásticos pueden ser clasificados:

- Por el número de valores (o de estados) posibles para la función aleatoria, X_t en un instante determinado. Este número puede ser
  - Finito
  - infinito numerable
  - infinito no numerable.

  (Incluso, el número puede ser de distinta categoría entre dos épocas consideradas).

- Por las épocas t_t en que puede cambiar de estado el sistema.
  - El proceso es *discreto* cuando la serie t_t finita o infinita numerable, está fijada de antemano (la serie no es aleatoria).
  - El proceso es *permanente* cuando el sistema puede cambiar de .estado, en cualquier época.
  - El proceso se llama *discontinuo o continuo* según los cambios de estado tengan lugar por saltos o de modo continuo (*T* es un intervalo).

- El punto de vista más importante por el que se pueden clasificar los procesos es por su ley de evolución en el tiempo.

  En la vida real se producen distintas relaciones entre las variables aleatorias que constituyen un proceso estocástico. Por ejemplo, la ganancia monetaria obtenida en la tirada n-ésima dependerá de la ganancia obtenida en la tirada (n−1).

  Las propiedades probabilísticas de las variables aleatorias son importantes al momento de identificar y clasificar un proceso estocástico que puede ser:
  - Procesos *Markovianos.*
  - Procesos estacionarios.
  - Procesos de incrementos independientes.

  - Proceso de *Markov*. Si la evolución del sistema depende sólo de su estado en el instante *t*, las aplicaciones son muy amplias, en especial debido al hecho de que muchos procesos no markovianos pueden considerarse como tales, mediante una definición conveniente de los estados posibles (cuando el proceso de *Markov* es discreto y discontinuo suele denominarse *cadena de Markov*)
    La característica principal de los procesos estocásticos markovianos es que la distribución de X_{n+1} sólo depende de la distribución de X_n y no de las anteriores $X_{n-1}, X_{n-2,} \dots$ Se puede resumir diciendo que el estado futuro del proceso, sólo depende del estado presente, y no del resto de estados pasados.
    Formalmente se expresa como:

$$\forall n \in n \; y \forall t_1 < \cdots < t_n$$
$$P\left(X_{t_n} \leq X_n \middle| X_{t_1} \leq X_1, \dots X_{t_{n-1}} \leq X_{n-1}\right) = P\left(X_{t_n} \leq X_n \middle| X_{t_{n-1}} \leq X_{n-1}\right)$$

Cuando el espacio de estados E es discreto, entonces se puede escribir

$$P\ \ X_{t_n} = x_n \left| X_{t_1} = x_1,\ldots, X_{t_{n-1}} = x_{n-1} \right. = P\ \ X_{t_n} = x_n \left| X_{t_{n-1}} = x_{n-1} \right.$$

- Procesos estacionarios. La ley temporal del proceso, o incluso sólo, ciertas características del proceso son independientes de una traslación cualquiera del eje tiempo (un proceso estacionario puede ser también un *proceso de Markov*).

Definición: (Función de autocovarianzas de un proceso estocástico)
Dado $X_t : t \in T$ se llama función de autocovarianzas a la función

$\gamma\ r,s = Cov\ \ X_r, X_s = E\left[\ X_r - E\ \ X_r \quad X_s - E\ \ X_s\ \right]$ donde $r,s \in T$.

Existen dos clases de procesos estacionarios:

1. estacionario débil
2. estacionario estricto.

Definición. (Proceso estacionario débil).
Un proceso $X_t : t \in T$, tal que $E\ \ X_t^2 < \infty \forall t \in T$, es un proceso estacionario débil si:

$1. - E\ \ X_t = m \forall t \in T$

$2. - \gamma\ r,s = \gamma\ r+t, s+t\ \ , \forall r,s,t \in T$

Esto implica, también, que *V* ar X_t es constante para todo $t \in T$.

Observaciones:
— Debe existir el momento de orden dos de las variables aleatorias.
— Todas las variables aleatorias tienen la misma media.
— El hecho de que $\gamma\ r,s = Cov\ \ X_r, X_s = Cov\ \ x_{r+t}, X_{s+t}$ significa que la función de autocovarianzas toma el mismo valor para dos variables aleatorias que estén separadas por un retardo *t* en el proceso, independientemente de dónde se encuentren situadas estas variables aleatorias en el tiempo.
— Si se considera t = −s, entonces $\gamma\ r,s = \gamma\ r - s, 0$ es decir, es una función de (r−s). Esta cantidad es la distancia de separación entre las dos variables X_r y X_s. Así, la función de autocovarianzas de un proceso estacionario débil sólo es función de una variable que es la separación (r − s) entre las variables en consideración.
— Si se toma r = s, entonces:

$\gamma\ r,r = Cov\ \ X_r, X_r = Var\ \ x_r \overset{*}{=} Var\ \ X_{r+h}$ para toda $h \in T$

ya que $\gamma\ r+h, r+h = Cov\ \ X_{r+h}, X_{r+h} = \gamma\ r,r$ por(2) (*)

Definición (Proceso estacionario estricto).

Un proceso $X_t : t \in T$, tal que $E\ X_t^2 < \infty \forall t \in T$, es un proceso estacionario estricto si $\forall \in \mathbb{N}, \forall h \in y\ \forall \{t_1, t_2, \ldots t_n t\} \in T$

$\forall n \in \square, \forall h \in T$ y $\forall\ t_1, t_2 \ldots, t_n \in T$ las variables aleatorias $X_{t_1}, X_{t_2} \ldots X_{t_n}$ tienen la misma distribución conjunta que $X_{t_{1+h}}, X_{t_{2+h}} \ldots X_{t_{n+h}}$

P. E. estricto $\Rightarrow$ P. E. débil, pero no al contrario.

Observación: Si $X_t : t \in T$ es un proceso estacionario estricto:

Cuando n=1

Se obtiene que todas las variables aleatorias tienen la misma distribución y, si se supone la existencia de momentos de orden dos, tienen la misma media, con lo que se cumple la condición (1).

Cuando n = 2

(Xr, Xs) se distribuye igual que X_{r+h}, X_{s+h} , $\forall h \in T$. De este modo, $\gamma\ r,s = \gamma\ X_{r+h}, X_{s+h}$ y sólo depende de cuál sea la diferencia (r−s), quedando, así, demostrado (2).

Observación: En el caso de la distribución normal, la estacionalidad débil implica estacionalidad estricta dado que, en este caso, la distribución conjunta n-dimensional queda determinada por las marginales y condicionadas.

Se define un proceso *gaussiano* como aquel que cumple la propiedad de que $\forall t_1, t_2 \ldots t_n \in T$ la distribución de $X_{t_1}, X_{t_2} \ldots X_{t_n}$ es normal n-dimensional.

- Procesos aditivos o de incrementos independientes.

Si la evolución en un intervalo cualquiera $(t, t+dt)$ es independiente del pasado incluido el instante *t*

Se dice que un proceso $X_t : t \in T$ es de incrementos independientes si, $\forall n \in \mathbb{N}\ \forall t_1, \ldots t_n \in T$, con $t_1 < \ldots < t_n$ las variables aleatorias

$$y_1 = X_{t_3} - X_{t_2}$$
$$y_1 = X_{t_2} - X_{t_1}$$
$$\vdots$$
$$y_1 = X_{t_n} - X_{t_{n-1}}$$

son independientes.

Proposición Todo proceso de incrementos ortogonales es un proceso *markoviano*

Demostración:

Suponemos un proceso X_1, X_2, X_3

Para ver que es *Markoviano* mostraremos que

$$P\ \ X_3 = x_3 \,|\, X_2 = x_2, X_1 = x_1\ \ = P\ \ X_3 = x_3 \,|\, X_2 = x_2\ \ .$$

Así, dado que se conocen $X_2 = x_2, X_1 = x_1$

$$P\ \ X_3 = x_3 \,|\, X_2 = x_2, X_1 = x_1\ \ =$$

$$P\ \ X_3 - X_2 = x_3 - x_2 \,|\, X_2 = x_2, X_2 - X_1 = x_2 - x_1\ \ \overset{*}{=}$$

$$P\ \ X_3 - X_2 = x_3 - x_2 \,|\, X_2 = x_2\ \ =$$

$$P\ \ X_3 = x_3 \,|\, X_2 = x_2$$

(*) por hipótesis, los incrementos son independientes.

1.2 DISTRIBUCIONES DE PROBABILIDAD

Las fuentes de la variación en los problemas de líneas de espera se deben a las llegadas aleatorias de los clientes y a las variaciones en los tiempos de servicio. Cada una de estas fuentes se describe con una distribución de probabilidad.

Un modelo de sistema de colas debe especificar la distribución de probabilidad de los tiempos de servicio para cada servidor.

La distribución más usada para los tiempos de servicio es la *exponencial*, aunque es común encontrar la distribución *degenerada o determinística* (tiempos de servicio constantes) o la distribución *Erlang* (Gamma)

Las distribuciones que utilizaremos son:

- M: Distribución exponencial (*markoviana*)
- D: Distribución degenerada (tiempos constantes)
- Ek: Distribución *Erlang*
- G: Distribución general

Considere en la que el número de llegadas y salidas, durante un intervalo de tiempo es controlado por las condiciones siguientes:

1. La probabilidad de que la entrada ó salida ocurra entre los tiempos *t* y $t + \Delta t$ dependen únicamente de Δt por lo tanto la probabilidad no depende de el número de eventos que ocurren en el tiempo *t* ni de el valor específico del periodo *(0,t)* (la función de probabilidad tiene incrementos independientes estocásticos)

2. La probabilidad de que ocurra un evento durante un intervalo de tiempo muy pequeño $0 < \Delta t < 1$ es $\lambda \Delta t + o\ \ \Delta t$

3. en un intervalo de tiempo Δt a lo mas puede ocurrir un evento

1.2.1 DISTRIBUCIÓN EXPONENCIAL

La distribución exponencial describe la probabilidad de que el tiempo de servicio del cliente en una instalación particular no sea mayor de *T* períodos de tiempo.
Por la condición 1 la probabilidad de que no ocurra ningún evento en el tiempo *t+Δt* es: $p_0\ t+\Delta t\ = p_0\ t\ + p_0\ \Delta t\ $ para *Δt>0* suficientemente pequeña.

Por la condición 2 $0 < p_0(\Delta t) < 1$, por lo tanto $p_0(t) = e^{-\mu t}$
Donde μ es una constante positiva y $t \geq 0$

A continuación veremos que para el proceso descrito por $p_n\ t\ $, el intervalo de tiempo entre eventos sucesivos es exponencial

Sean

f(t): (fdp) del intervalo de tiempo *t* entre la ocurrencia de eventos sucesivos $t \geq 0$

T: intervalo de tiempo desde la ocurrencia del último evento.

Entonces:

P el tiempo entre evento exede a T = no ocurren eventos durante T

Esto se puede expresar:

$$\int_T^\infty f\ t\ dt = p_0\ T\ = e^{-\mu T} \qquad T>0$$

O bien

$$\int_0^T f\ t\ dt = 1 - e^{-\mu T} \qquad T>0$$

La probabilidad se calcula usando la fórmula:

$$P(t \leq T) = 1 - e^{-\mu T}$$

Donde: $\mu =$ número promedio de clientes que terminan el servicio por período.

$t =$ tiempo de servicio del cliente.

$T =$ tiempo de servicio objetivo.

$1/\mu$ = media de la distribución del tiempo de servicio

$(1/\mu)^2$ = varianza

A medida que aumenta T, la probabilidad que el tiempo de servicio del cliente sea menor de *T* se aproxima a 1.0. Por su sencillez, veremos un arreglo de un solo canal, una sola fase.

EJEMPLO 1

Un gerente de una tienda departamental, debe determinar si es necesario más entrenamiento para el empleado de servicio al cliente. El cual puede atender un promedio de tres clientes por hora. ¿Cuál es la probabilidad que el empleado atienda a un cliente en menos de 10 minutos?

SOLUCIÓN

En este caso $\mu = 3$ clientes por *hora*

$T = 10$ minutos $= 10/60$ hora $= 0.167$ hora. Entonces

$$P(t \le T) = 1 - e^{-\mu T}$$

$$P(t \le 0.167 \text{ hora}) = 1 - e^{-(3)(0.167)} = 1 - 0.61 = 0.39$$

Por lo tanto. La probabilidad de que el empleado requiera solo de 10 minutos o menos no es muy alta, esto indica la posibilidad de que los clientes tengan que esperar un poco. El gerente debe considerar el entrenamiento adicional del empleado para que éste reduzca el tiempo que toma atender a un cliente.

Algunas características de la distribución exponencial no siempre conforman a una situación real. El modelo de la distribución exponencial se basa en el supuesto que cada tiempo de servicio es independiente de aquellos que lo precedieron. Sin embargo en, la vida real, puede mejorar la productividad a medida que los servidores humanos aprenden su trabajo. Otra suposición de este modelo es que son posibles tiempos de servicio muy pequeños, así como muy grandes. Sin embargo, las situaciones de la vida real requieren con frecuencia tiempos de servicio casi constantes.

1.2.2 DISTRIBUCIÓN DE POISSON

La distribución de *Poisson* describe una variable aleatoria discreta. Esta distribución se usa en modelos de líneas de espera para describir el número de eventos (llegadas o salidas) en un periodo dado.

Los clientes llegan en forma aleatoria a las instalaciones de servicio. La variabilidad en la llegada de los clientes se describe con frecuencia por una *distribución de Poisson*, que especifica la probabilidad que *n* clientes lleguen en *t* períodos de tiempo:

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad \text{para } n = 0, 1, 2, \ldots$$

Donde: $P_n(t) =$ probabilidad de *n* llegadas en *t* períodos de tiempo.

$\lambda =$ número promedio de clientes que llegan por período.

λt = media de la *distribución de Poisson* = varianza

El proceso de *Poisson* es un proceso completamente aleatorio pues el intervalo de tiempo que transcurre hasta que se presenta el próximo evento no depende del tiempo que transcurre desde que ocurrió el evento anterior.
Esta propiedad se le conoce como olvido, falta de memoria ó propiedad *markoviana* de la distribución exponencial

$$P\ \ t > T + S \,|\, t > S\ \ = P\ \ t > T$$

Donde *S* es el intervalo de tiempo desde que ocurrió el último evento.

Como t es exponencial, tenemos:

$$P\ \ t > T+S \,|\, t > S\ \ = \frac{P\ \ t > T*S \,|\, t > S}{P\ \ t > S}$$

$$= \frac{P\ \ t > T+S}{P\ \ t > S}$$

$$= \frac{e^{-\lambda\ T+S}}{e^{-\lambda S}} = e^{-\lambda t}$$

$$= P\ \ t > T$$

Otra característica de la *distribución de Poisson* es que es la única distribución cuya media y varianza son iguales

La *distribución de Poisson* es una distribución discreta; es decir, las probabilidades son para un número específico de llegadas por unidad de tiempo.

EJEMPLO 2

El gerente está rediseñando el proceso de servicio al cliente en una tienda departamental. Es importante atender a cuatro clientes. Los clientes llegan al mostrador a una tasa de dos clientes por hora. ¿Cuál es la probabilidad que cuatro clientes lleguen en cualquier hora?

SOLUCIÓN

En este caso: $\lambda = 2$ clientes por hora,
$t = 1$ hora
$n = 4$ clientes

La probabilidad de que cuatro clientes lleguen en una hora es

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} = \frac{2(1)^4}{4!} e^{-2} = \frac{16}{24} e^{-2} = 0.090$$

Por lo tanto. El gerente de servicio al cliente puede usar esta información para determinar los requerimientos de espacio para el mostrador y el área de espera. Hay una probabilidad relativamente pequeña de que cuatro clientes lleguen en una hora. Por consiguiente, la capacidad de asientos para dos o tres clientes es más que adecuada a menos que el tiempo para atender a cada cliente sea largo.

Otra manera de especificar la distribución de la llegada es hacerlo en términos de los **tiempos entre llegadas** (el tiempo entre las llegadas de los clientes). Si la población de clientes genera clientes según una *distribución de Poisson*,

1.3 PROCESOS DE MARKOV

1.3.1 PROCESO DE *MARKOV* TIEMPO DISCRETO

Un *proceso de Markov* se define como un proceso estocástico, en donde las llegadas son descritas por un *proceso de Poisson* con tasa λ, y el tiempo entre llegadas es una variable aleatoria con distribución exponencial y media $1/\lambda$.

Sea una *cadena de Markov* discreta descrita por el conjunto de estados X_n, $n = 0,1,2,3,\ldots$ cuya probabilidad de transición del estado *i* al estado *j* se denota como p_{ij}. El conjunto de probabilidades de transición de estados se puede representar por la matriz

$$P = \left[p_{ij} \right] \tag{1.1}$$

Los *procesos Markovianos* pueden describir un estado determinado en un instante de tiempo. Para ello se cuenta con las siguientes variables:

$$P_{ij} = P\ \ X_{n+1} = j \,\middle|\, X_n = i$$

Probabilidad de que el sistema pase del estado *i* al estado *j* en m instantes:

$$p_{ij}(m) = P\ \ X_{m+n} = j \,\middle|\, X_n = i$$

Nótese que es una probabilidad condicional de transición de estado probabilidad de que el sistema está en el estado *i* en el instante $m: p(i;m) = P\ \ X_m = i$ y que es un valor de probabilidad.

1.3.2 ECUACIÓN DE CHAPMAN

La matriz dada en la ecuación (1.1) puede ser descrita para un instante m como:

$$\left[P_{ij}(m) \right] = \left[P_{ij} \right] * \left[P_{ij} \right] * \ldots * \left[P_{ij} \right] = \left[P_{ij} \right]^m$$

Esta ecuación se denomina ecuación de *Chapman-Kolmogorov*, o ecuación *CK.*

Se puede entonces, conocidas las probabilidades del sistema en el estado *i* y el instante n, $p(i;n)$, determinar las probabilidades del sistema en el estado *i*, en *m* instantes posteriores, $p(i;n+m)$ con:

$$p(i;n+m) = \sum_i P\ \ X_{m+n} = j / X_n = i\ \ P\ \ X_n = i\ \ = \sum_i p_{ij}(m)\, p(i;n)$$

EJEMPLO 3

Una *cadena de Markov* con dos estados n = 1, 2, tiene la siguiente matriz de entrada:

$$\mathrm{P} = \begin{vmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{vmatrix}$$

SOLUCIÓN

$$P^4 = \begin{vmatrix} 0.61 & 0.4251 \\ 0.5668 & 0.4332 \end{vmatrix} \Rightarrow \begin{matrix} p_{12} \ \ 4 = 0.4251 \\ p_{22} \ \ 4 = 0.4332 \end{matrix}$$

Por lo tanto $p(1;4) = p_{11}(4)\,p(1;0) + p_{21}(4)\,p(2;0) = (0.4)(0.5749) + (0.6)(0.5688) = 0.57$

1.3.3 PROCESO DE *MARKOV* TIEMPO CONTÍNUO

Aplicando la probabilidad de transición de estados para el caso que la distancia entre los instantes m $\Delta m \to 0$, los instantes se cambian del espacio discreto al contínuo, con lo que el proceso se vuelve estocástico. Luego:

$$P_{ij} \ \ t = \lim_{\Delta m \to 0} P_{ij}^{m} = P \ \ N \ \ t+s \ = j \,|\, N(s) = i$$

Con lo que el sistema, descrito por un *proceso contínuo de Markov*, cumple con las siguientes condiciones:

- El tiempo de permanencia en el estado *i,* T_i es una variable aleatoria con distribución exponencial y media $1/v_i$.
- Cuando el proceso sale del estado *i*, pasa al estado *j* con una probabilidad P_{ij} .

Lo que satisface que $\sum jP_{ij} = 1$

> *nota:* Si esta probabilidad no depende de *s*, se dice que el sistema tiene trancisiones estacionarias.

La tasa con la que el sistema pasa del estado *i* al estado *j* es:

$$q_{ij} = \lim_{\Delta m \to 0} \frac{P_{ij}(\Delta m)}{\Delta m} \qquad (1.2)$$

Donde: $P_{ij}(\Delta m)$ es la probabilidad de que el sistema deje el estado *i* antes del periodo Δm y cambie al estado *j*. Dicho valor, utilizando la desigualdad de Markov, es igual a:

$$P_{ij}(\Delta m) = P \ \ T_i < \Delta m \ \ = \ 1 - e^{-v_i \Delta m} \ \ P_{ij} \approx \Delta m v_i P_{ij} \qquad \text{Luego}$$

$$q_{ij} = \frac{\Delta m v_i P_{ij}}{\Delta m} = v_i P_{ij}$$

donde q_{ij} es la tasa instantánea de transición del estado *i* al estado *j*.

La tasa con la que el sistema va a realizar la transición, estando en el estado *i* es:

$$v_i = \lim_{\Delta m \to 0} \frac{1 - P_{ii}(\Delta m)}{\Delta m} \qquad (1.3)$$

$P_{ii}(\Delta m)$ es la probabilidad de que el sistema continúe en el estado *i* después del periodo Δm. Dicho valor, utilizando la *desigualdad de Markov*, es igual a:

$$P_{ii} = \Delta m = P\ T_i > \Delta m = P_i e^{-v_i \Delta m} = 1 - v_i \Delta m\ e^{-v_i \Delta m}$$

v_i es la tasa de transición de estado cuando el sistema está en el estado *i*. Nótese que:

$$\frac{dP_{ij}\ t}{dt} = \lim_{\Delta m \to 0} \frac{P_{ij}(\Delta m + t) - P_{ij}(t)}{\Delta m} \tag{1.4}$$

entonces el equivalente contínuo de la ecuación *CK*

$$p_{ij}\ t+s = \sum_k p_{ik}(t)\, p_{kj}(s) \tag{1.5}$$

Que en forma matricial queda $\left[p_{ij}(t+s)\right] = \left[P_{ij}(t)\right] * \left[p_{ij}(s)\right]$

Luego, reemplazando (1.5) en (1.4)

$$P_{ij}\ t+\Delta m\ - P_{ij}(t) = \sum_k P_{ik}(t) P_{kj}(\Delta m) - P_{ij}(t)$$
$$= \sum_{k \neq j} P_{ik}(t) P_{kj}(\Delta m) - (1 - P_{ii}(\Delta m)) P_{ij}(t)$$

Dividiendo la ecuación por ⊿m y evaluando el límite cuando $\Delta m \to 0$ se tiene en las ecuaciones (1.4) y (1.2)

$$\frac{dP_{ij}(t)}{dt} = \lim_{\Delta m \to 0} \frac{P_{ij}\ \Delta m + t\ - P_{ij}(t)}{\Delta m}$$
$$= \left[\sum_{k \neq j} P_{ik}\ t\ \frac{P_{kj}\ \Delta m}{\Delta m} - \frac{1 - P_{ii}\ \Delta m}{\Delta m} P_{ij}\ t \right]$$

Reemplazando (1.2) y (1.3)

$$\frac{dP_{ij}(t)}{dt} = \sum_{k \neq j} P_{ik}\ t \left[\lim_{\Delta m \to 0} \frac{P_{kj}\ \Delta m}{\Delta m} \right] - P_{ij}\ t \left[\lim_{\Delta m \to 0} \right] \frac{1 - P_{ii}\ \Delta m}{\Delta m}$$
$$= \sum_{k \neq j} q_{kj} P_{ik}\ t\ - v_j P_{ij}\ t$$

Esta ecuación se denomina ecuación hacia atrás de *Kolmogorov*

1.3.4 PROCESO DE *MARKOV* PARA PROCESOS ESTABLES

Es conveniente saber cómo se comporta el sistema, Después de cierto tiempo.
Para el caso discreto, se utiliza el vector fila de probabilidades
En el ejemplo 3 se tiene:

$$p(1;4) = p_{11}(4)\, p(1;0) + p_{21}(4)\, p(2;0) = (0.4)(0.5749) + (0.6)(0.5688) = 0.57$$

Por lo tanto: $\hat{\Pi}(m+n) = \hat{\Pi}(n)[p]^m$

El cual define una relación de recurrencia que permite conocer la evolución del vector de probabilidad de estado en el instante *m*, conociendo el vector de probabilidad inicial, haciendo *n* = 0 de la siguiente forma:

$$\widehat{\Pi}^m = \widehat{\Pi}(0)[p]^m = \cdots \widehat{\Pi}(m-2)[p]^2 = \widehat{\Pi}(m-1)[p] \qquad (1.6)$$

independiente del vector de probabilidad inicial. Por lo tanto, cuando el sistema llega a un estado estable *j*, la probabilidad en estado estable llega a ser:

$$\prod_j = \lim_{m\to\infty}\left[P_{ij}\right]^m$$

Luego el vector de probabilidades en estado estable está dado por:

$$\widehat{\Pi} = \Pi_1, \Pi_2, \Pi_3 \ldots$$

Usando (1.6) $\widehat{\Pi}(m) - \widehat{\Pi}$(m-1)= $\widehat{\Pi}$(m-1){$[P][I]$}

Donde:

$$m \to \infty$$
$$m - 1 \approx m$$
$$\widehat{\Pi}(m) \to \widehat{\Pi}$$

Con lo que la ecuación queda: $\widehat{\Pi} = \widehat{\Pi}[P]$

Donde se cumple la condición de probabilidad $\sum \prod_j = 1$

Y así determinar el vector de probabilidades de estado, en estado estable.

1.4 PROCESO DE NACIMIENTO Y MUERTE

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

- Nacimiento: Llegada de un nuevo cliente al sistema de colas
- Muerte: Salida del cliente servido

El sistema se encuentra en el estado E_n cuando el número de elementos que compone dicho estado es igual a *n*

El proceso es de **nacimiento** puro cuando sólo es posible una transición del estado E_n al estado E_{n+1}

El proceso se considera de **muerte** cuando solo es posible la transición de E_n a E_{n-1}.

El proceso se llama de **nacimiento y muerte** cuando son posibles las transiciones tanto de E_n a E_{n+1} como E_n a E_{n-1}

N(t) es el número de clientes que hay en el sistema en el tiempo *t*. El proceso de nacimiento y muerte describe en términos probabilísticos como cambia *N(t)* al aumentar *t*.

Suposición 1

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

Suposición 2

Dado $N(t) = n$, la distribución de probabilidad actual del tiempo que falta para la próxima muerte (terminación del servicio) es exponencial con parámetro $\mu_n \quad n = 0,1,2\ldots$

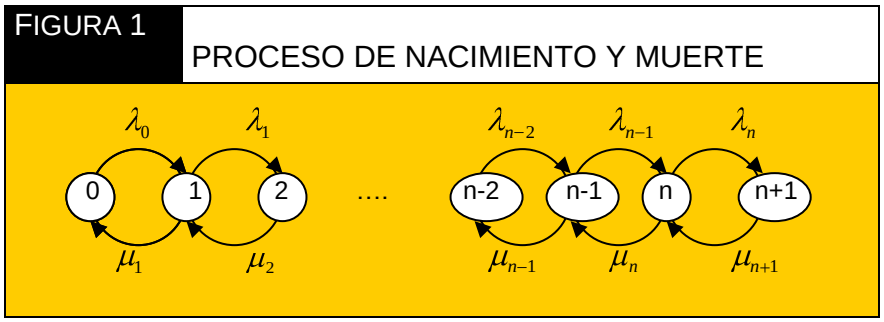
Suposición 3

Las variables aleatorias de los tiempos que faltan para la próxima llegada y para la terminación del servicio son mutuamente independientes

Transición en el estado del proceso $n \to n+1$ o $n \to n-1$

El proceso de nacimiento y muerte es un tipo especial de cadenas de Markov de tiempo continuo.

FIGURA 1 PROCESO DE NACIMIENTO Y MUERTE



Donde:

λ_n: Tasa media de llegadas cuando el sistema está en el estado n (del n al $n+1$)

μ_n: Tasa media de salidas cuando el sistema está en el estado n (del n al $n-1$)

Supongamos que en el tiempo cero se inicia el conteo del número de veces que el sistema entra en cualquier estado *n* y el número de veces que sale del mismo.

Entonces $\left|E_n(t) - L_n(t)\right| \le 1$

Donde:

$E_n(t)$: Número de veces que el sistema entra al estado *n* hasta el tiempo *t*

$L_n(t)$: Número de veces que el sistema sale del estado *n* hasta el tiempo *t*

Como los dos tipos de eventos deben alternarse la diferencia será a lo sumo 1

$$\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| \leq \frac{1}{t}$$

$$\lim_{t\to\infty}\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| = 0$$

Tasa media a la que el proceso entra al estado *n* $\lim_{t\to\infty}\frac{E_n(t)}{t}$

Tasa media a la que el proceso sale del estado *n* $\lim_{t\to\infty}\frac{L_n(t)}{t}$

Para cualquier estado n (n=0,1,...) del sistema, la tasa media de entrada es igual a la tasa media de salida

1.4.1 ECUACIONES DE BALANCE

Para el cálculo de la ecuación de un proceso de nacimiento y muerte, consideramos intervalos de tiempo suficientemente pequeños para que en ellos solo se pueda producir uno o ningún cambio,

Para calcular la ecuación general de un proceso de nacimiento y muerte consideremos la probabilidad de que el sistema se encuentre en el estado E_n en el instante $t+\Delta t$ si se divide el intervalo $(0, t+\Delta t)$ en $(0, t)$ y $(t, t+\Delta t)$ podemos considerar tres casos mutuamente excluyentes que pueden dar lugar a que el sistema tenga tamaño *n* en el instante $t+\Delta t$

1) que el sistema tenga *n* elementos en el instante *t* y no exista cambio entre *t* y $t+\Delta t$ cuya probabilidad será

$$P_n \; 1-\lambda_n-\mu_n$$

2) que el sistema tenga *n*-1 clientes en el instante *t* y se produzca un nacimiento entre el instante *t y* $t+\Delta t$ cuya probabilidad será

$$P_{n-1}\lambda_{n-1}$$

3) que el sistema tenga *n*+1 clientes en el instante *t* y se produzca una muerte entre el instante *t y* $t+\Delta t$ cuya probabilidad será

$$P_{n+1}\mu_{n+1}$$

Si el sistema se encuentran en $E_{n+2}, E_{n+3}, \ldots$ ó $E_{n-2}E_{n-3}\ldots$ en el instante *t* no se podría llegar al estado E_n ya que se ha supuesto que en el intervalo de amplitud solo puede producirse un cambio por lo que la probabilidad buscada cuando $\lim \Delta t \to 0$ será:

$$P_n^{'} = \lambda_{n-1}P_{n-1} - \left(\lambda_n + \mu_n\right) P_n + \mu_{n+1}P_{n+1}$$

Esta ecuación solo es válida cuado $n \geq 1$, si *n*=0 solo es posible una transacción de E_0 a E_1 y la ecuación se reduce a

$$P_n^{'} = \mu_1 P_1 - (\lambda_0 P_0)$$

Después de construir las ecuaciones de balance para todos los estados en término de las probabilidades P_n desconocidas, se puede resolver este sistema de ecuaciones (más una ecuación que establezca que la suma de las Pn debe ser 1).

Estado 0

$$\left\{\begin{matrix}\text{Tasa media global de} \\ \text{entradas al estado 0}\end{matrix}\right\} = \left\{\begin{matrix}\text{tasa media global de} \\ \text{salidas del estado 0}\end{matrix}\right\}$$

$$\mu_1 P_1 = \lambda_0 P_0$$

Donde:

P_n: Probabilidades de estado estable de encontrarse en el estado *n*.

P_1: Representa la proporción de tiempo posible cuando el proceso se encuentra en el estado cero

Nota: como $\mu_0 = 0$ el sistema está en el estado 0 por lo tanto no puede haber muertes.

Estado 1

$$\left\{\begin{matrix}\text{Tasa media global de} \\ \text{entradas al estado 1}\end{matrix}\right\} = \left\{\begin{matrix}\text{tasa media global de} \\ \text{salidas del estado 1}\end{matrix}\right\}$$

$$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$$

Se continúa con esta metodología y se debe construir para todos los estados.

Nota: Recordemos que la $\sum P_n = 1$

Entonces:

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

Estado	
0	$P_1 = (\lambda_0 / \mu_1)P_0$
1	$P_2 = (\lambda_1 / \mu_2)P_1 + (1/\mu_2)(\mu_1 P_1 - \lambda_0 P_0)$
2	$P_3 = (\lambda_2 / \mu_3)P_2 + (1/\mu_3)(\mu_2 P_2 - \lambda_1 P_1)$
⋮	⋮
n - 1	$P_n = (\lambda_{n-1} / \mu_n)P_{n-1} + (1/\mu_n)(\mu_{n-1} P_{n-1} - \lambda_{n-2} P_{n-2})$
n	$P_{n+1} = (\lambda_n / \mu_{n+1})P_n + (1/\mu_{n+1})(\mu_n P_n - \lambda_{n-1} P_{n-1})$
⋮	⋮

Estado	
0	$P_1 = (\lambda_0 / \mu_1)P_0$
1	$P_2 = (\lambda_1 / \mu_2)P_1$
2	$P_3 = (\lambda_2 / \mu_3)P_2$
⋮	⋮
n - 1	$P_n = (\lambda_{n-1} / \mu_n)P_{n-1}$
n	$P_{n+1} = (\lambda_n / \mu_{n+1})P_n$
⋮	⋮

Estado	
0	$P_1 = (\lambda_0 / \mu_1)P_0$
1	$P_2 = (\lambda_0 \lambda_1 / \mu_2 \mu_1)P_0$
2	$P_3 = (\lambda_2 \lambda_1 \lambda_0 / \mu_3 \mu_2 \mu_1)P_0$
⋮	⋮
n - 1	$P_n = (\lambda_{n-1} \lambda_{n-2} \dots \lambda_0 / \mu_n \mu_{n-1} \dots \mu_1)P_0$
n	$P_{n+1} = (\lambda_n \lambda_{n-1} \dots \lambda_0 / \mu_{n+1} \mu_{n+1} \dots \mu_1)P_0$
⋮	⋮

Simplificando la notación $C_n = \begin{cases} \dfrac{\lambda_{n-1}\lambda_{n-2}\dots\lambda_0}{\mu_n\mu_{n-1}\dots\mu_1} & n = 1, 2, \dots \\ 1 & n = 0 \end{cases}$

Entonces el requisito $\sum_{n=0}^{\infty} P_n = 1 \Rightarrow \left\{ \sum_{n=0}^{\infty} C_n \right\} P_0 = 1$

De esta forma $P_0 = \left\{ \sum_{n=0}^{\infty} C_n \right\}^{-1}$

El número de clientes en el sistema es $L = \sum_{n=0}^{\infty} n P_n$

La longitud de espera en la cola es: $L_q = \sum_{n=0}^{\infty} (n-s)P_n$

Donde *s*: número de servidores (representa el número de clientes que pueden estar en servicio y no los que estas en la cola)

De estas relaciones tenemos:

$W = \frac{L}{\bar{\lambda}}$ $\quad$ $W_q = \frac{L_q}{\bar{\lambda}}$ $\quad$ ($\bar{\lambda}$ = la tasa de llegadas promedio)

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

Nota:

- Varias de las expresiones tienen un número infinito de términos. Para muchos casos especiales estas sumas tienen solución analítica o pueden aproximarse por métodos numéricos.
- Estos resultados de estado estable se desarrollaron bajo la suposición de que los parámetros λ_n y μ_n tiene valores tales que el proceso, de hecho puede alcanzar la condición de estado estable.

se cumple si $\begin{cases} \lambda_n = 0 \text{ para algun } n \text{ mayor que el estado inicial} \\ \rho = \frac{\lambda}{s\mu} < 1 \end{cases}$

No se cumple cuando $\sum_{n=1}^{\infty} C_n = \infty$

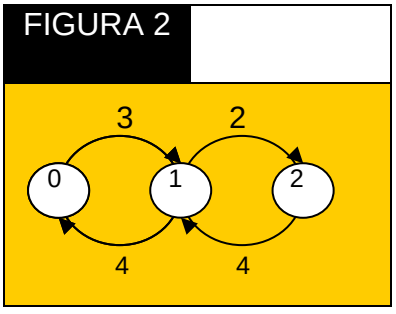
EJEMPLO 4

La estación de gasolina de una pequeña población tiene capacidad para 2 automóviles. Cuando la estación está desocupada llegan 3 automóviles por hora, pero cuando en la estación hay un automóvil la tasa de llegadas disminuye a 2 automóviles por hora.
La tasa a la cual el servidor puede atender a los automóviles que llegan es de 4 por hora.

SOLUCIÓN

Se deben encontrar L, L_q, W, W_q y las probabilidades de estado estable

FIGURA 2



$$\text{Estado } 0 \quad \mu_1 P_1 = \lambda_0 P_0 \Rightarrow 4P_1 = 3P_0$$

$$\text{Estado } 1 \quad \lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$$

$$\mu_2 P_2 = \lambda_1 P_1 \Rightarrow 4P_2 = 2P_1$$

$$P_0 + P_1 + P_2 = 1$$

Resolviendo estas tres ecuaciones obtenemos

$$\left.\begin{array}{l} 4P_1 = 3P_0 \\ 4P_2 = 3P_1 \\ P_0 + P_1 + P_2 = 1 \end{array}\right\} \Rightarrow \begin{cases} P_0 = 0.47 \\ P_1 = 0.353 \\ P_2 = 0.177 \end{cases}$$

Por lo tanto

$$L = \sum n P_n = 0 * 0.47 + 1 * 0.353 + 2 * 0.177 = 0.707$$

$$L_q = \sum (n - s) P_n = 0 * 0.353 + 1 * 0.177 = 0.177$$

$$\bar{\lambda} = \sum \lambda_n P_n = 3 * 0.47 + 2 * 0.353 + 0 * 0.177 = 2.116$$

$$W = \frac{0.707}{2.116} = 0.334 \text{ horas}$$

$$W_q = \frac{L_q}{\lambda} = \frac{0.177}{2.116} = 0.083 \text{ horas}$$

1.4.2 CASOS PARTICULARES

Considerando diferentes valores de las probabilidades de nacimiento y muerte se obtienen casos particulares de procesos estocásticos, cuyas ecuaciones pueden deducirse a partir de la fórmula general.

Una clase importante de procesos son los llamados homogéneos para los cuales las probabilidades son independientes del tiempo.

A continuación se dan las formulas para algunos casos particulares

A. proceso de muerte homogéneo $(\lambda_n = 0)$

$$P_n' = \mu_n P_n + \mu_{n+1} P_{n+1} \cdots\cdots \quad n \geq 1$$

$$P_0' = \mu_1 P_1 \cdots\cdots\cdots\cdots\cdots \quad n = 0$$

B. proceso de nacimiento homogéneo $(\mu_n = 0)$

$$P_n' = \lambda_{n-1} P_{n-1} - \lambda_n P_n \cdots\cdots \quad n \geq 1$$

$$P_0' = \lambda_0 P_0 \cdots\cdots\cdots\cdots\cdots \quad n = 0$$

I. proceso de Poisson $\lambda_n = L = cte$

$$P_n^{'} = L\ P_{n-1} - P_n \ \cdots\cdots\ n \geq 1$$

$$P_0^{'} = -LP_0 \cdots\cdots\cdots\cdots\cdots\ n = 0$$

II. proceso de Yule $L_n = nL$

$$P_n^{'} = \ n-1\ LP_{n-1} - nLP_n \cdots\cdots\ n > 1$$

$$P_0^{'} = -LP_1 \cdots\cdots\cdots\cdots\cdots\cdots\ n = 1$$

Otros procesos de nacimiento particulares son:

I *Proceso de Berniulli*

Para $\lambda_n = \ s-n\ L$ siendo *s* un número finito

Proceso de contagio.

Para $\lambda_n = a + b_n$

Los procesos de Bernoulli y de Poisson son casos Particulares del de contagio cuando *a=c, b=0* y *a=sL; b=-L*

Proceso no homogéneo de Polya,

para $\lambda_n = {L\ 1-bn}\ \big/\ {1+blt}$

C. proceso lineal homogéneo de nacimiento y muerte $\lambda_n = nL,\ \mu_n = nM$

$$P_n^{'} = \ n-1\ LP_{n-1} - (nL+nM)P_n + \ n+1\ MP_{n+1} \cdots\cdots\ n \geq 1$$

$$P_0^{'} = MP_1 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\ n = 0$$

a. proceso de nacimiento y muerte con coeficientes constantes $\lambda_n = L,\ \mu_n = M$

$$P_n^{'} = LP_{n-1} - (L+M)P_n + MP_{n+1} \cdots\cdots\ n \geq 1$$

$$P_0^{'} = -LP_0 + MP_1 \cdots\cdots\cdots\cdots\cdots\ n = 0$$

Este proceso es fundamental en la teoría de colas cuando se considera una sola fuente de llegada

ii. proceso de Erlang $\lambda_n = L,\ \mu_n = nM$

$$P_n^{'} = LP_{n-1} - (L+nM)P_n + (n+1)MP_{n+1} \cdots\cdots\ n \geq 1$$

$$P_0^{'} = -LP_0 + MP_1 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\ n = 0$$

Este proceso aparece en la teoría de colas cuando se considera el caso en el que existen varias fuentes de llegada y estas se producen con la ley de *Poisson*

CAPÍTULO II

FUNDAMENTOS DE SISTEMAS DE LÍNEAS DE ESPERA

Las líneas de espera son parte de nuestra vida cotidiana. Todos esperamos en la fila para comprar un boleto para el cine, efectuar un depósito bancario, pagar los víveres, enviar un paquete por correo, obtener comida en la cafetería, comenzar un recorrido en un parque de diversiones, etcétera. Nos hemos acostumbrado a cantidades notables de espera, pero aún nos molestamos cuando estas son prolongadas.

Sin embargo, tener que esperar no es sólo una pequeña molestia personal. La cantidad de tiempo que la población de un país pasa esperando en filas es un factor primordial tanto en la calidad de vida como en la eficiencia de la economía del país.

Un claro ejemplo, es Estados Unidos, se estima que sus ciudadanos gastan 37'000,000,000 horas anuales esperando en filas. Tiempo que podría usarse en forma productiva, (equivaldría a 20 millones de años-persona de trabajo útil cada año)

Quien haya tenido que esperar en un semáforo, en las cajas de un centro comercial, al abordar el Metro o esperar a registrar la salida en el trabajo ha experimentado lo que es una línea de espera. Quizás uno de los mejores ejemplos de la administración efectiva de las líneas de espera es el de *Walt Disney World*. Un día lo pueden visitar 25,000 personas, pero otro día pueden ser 90,000. Un análisis cuidadoso del proceso de flujo, de la tecnología para movilizar a la gente (manejo de materiales), transporte del equipo, capacidad y distribución de las instalaciones hacen que los tiempos de espera para las atracciones sean de un nivel aceptable.

El análisis de líneas de espera es de interés para los administradores porque son los que efectúan el diseño, planean la capacidad y distribución, controlan el inventario, y efectúan la programación de las actividades en la empresa donde trabajan.

2.1 ¿PORQUÉ SE CREAN LAS LÍNEAS DE ESPERA?

Las líneas de espera se producen cuando la demanda excede la capacidad de servicio

En una línea de espera uno o más "clientes" esperan por un servicio.
Los clientes pueden ser personas u objetos inanimados como las máquinas que requieren de mantenimiento, las órdenes de ventas que esperan a despacharse, o

artículos del inventario que esperan para ser utilizados. Las formas de la línea de espera se deben a un desequilibrio temporal entre la demanda por el servicio y la capacidad del sistema para proporcionar el servicio. En la mayoría de los problemas de líneas de espera en la vida real, la tasa de la demanda varía; es decir, los clientes llegan a intervalos imprevisibles.

Con frecuencia, la tasa para producir el servicio también varía y depende de las necesidades del cliente

EJEMPLO 5

a) Suponga que los clientes del banco llegan a una tasa media de 15 por hora a lo largo del día y que el banco puede atender un promedio de 20 clientes por hora. Siempre se forma una línea de espera porque la tasa de llegada del cliente varía durante el día y el tiempo requerido para atender un cliente también varía.

   Las líneas de espera pueden crecer aun cuando el tiempo para atender a un cliente sea constante.

b) En el servicio de transporte Metro se controla el tiempo del recorrido entre estaciones a lo largo de toda la ruta. Cada tren se programa para llegar a una terminal, digamos, cada 5 minutos. Incluso con el tiempo de servicio constante, se crean líneas de espera. Las personas esperan el próximo tren o no pueden abordar un tren debido a la muchedumbre en un momento del día cuando la demanda es mayor. Por consiguiente, en este caso la variabilidad en la tasa de demanda determina el tamaño de la línea de espera.

Nota: Si no hay variabilidad en la demanda o en las tasas de servicio y se tiene suficiente capacidad en el sistema de servicio, no puede crearse una línea de espera.

2.2 EJEMPLOS DE SISTEMAS DE LÍNEAS DE ESPERA

La teoría de líneas de espera se aplica tanto en los servicios como en las empresas manufactureras, relacionan la llegada de clientes y las características del proceso de producción en el sistema de servicio.
Se define el término **servicio** como el acto de trabajar para un cliente.

a. Servicios comerciales.
Los clientes externos reciben servicios de organizaciones comerciales.
La mayoría de estos ejemplos involucran a clientes que acuden al servidor en un lugar fijo, donde se forma una línea de espera física si los clientes necesitan esperar para que comience el servicio. Sin embargo, para los ejemplos de servicios de plomería y de techado, el servidor va a los clientes, de modo que los clientes en la línea de espera están geográficamente dispersos. En algunos casos,

el servicio se proporciona por teléfono, quizá después de algunos clientes que están en espera (en la línea).

EJEMPLOS DE SISTEMAS DE SERVICIOS COMERCIALES		
Tipo de sistema	*Clientes*	*Servidor* (*es*)
Peluquería	Personas	Peluquero
Cajero automático	Personas	Cajero automático
Caja en tienda	Personas	Cajero
Servicios de plomería	Cañerías tapadas	Plomero
Ventanilla de boletos en un cine	Personas	Cajero
Servicio de corretaje	Personas	Corredor de valores
Gasolinera	Autos	Bomba
Atención para soporte técnico	Personas	Representante técnico
Servicios dentales	Personas	Dentista

b. Servicio interno

En algunos casos, los clientes son empleados de las organizaciones. En otros ejemplos, los clientes son cargas que deben moverse, máquinas a ser reparadas, artículos para inspección, etcétera.

EJEMPLOS DE SISTEMAS DE SERVICIOS INTERNOS		
Tipo de sistema	*Clientes*	*Servidor*(*es*)
Servicios secretariales	Empleados	Secretaria
Servicios de copiado	Empleados	Máquina copiadora
Computadoras grandes	Empleados	Computadora
Servicios de fax	Empleados	Máquina de fax
Sistema de mantenimiento	Máquinas	Cuadrilla de reparación
Estación de inspección	Artículos	Inspector
Sistema de producción	Trabajos	Máquina
Máquinas semiautomáticas	Máquinas	Operador
Depósito de herramientas	Operadores de máquinas	Empleado

c. Servicio de transporte

En ocasiones, los vehículos involucrados son los clientes y en otros casos, cada vehículo es un servidor.

En particular, el servicio de aerolínea y el de elevador involucran a un servidor que sirve a un grupo de clientes en forma simultánea en lugar de uno a la vez. La línea de espera en el ejemplo del estacionamiento tiene capacidad cero porque los autos que llegan (clientes) van a cualquier otro lugar a estacionarse si todos los espacios del estacionamiento (servidores) están ocupados.

EJEMPLOS DE ESTACIONES DE SERVICIOS DE TRANSPORTE		
Tipo de sistema	*Clientes*	*Servidor*(*es*)
Caseta de cobro en autopista	Autos	Cajero
Muelle de carga de camiones	Camiones	Cuadrilla de carga
Área de descarga portuaria	Barcos	Cuadrilla de descarga
Aviones que esperan despegar	Aviones	Pista
Aviones que esperan aterrizar	Aviones	Pista
Servicio de aerolínea	Personas	Avión
Servicio de taxis	Personas	Taxi
Servicio de elevador	Personas	Elevador
Bomberos	Incendios	Carro de bomberos
Estacionamiento	Autos	Espacios de estacionamiento
Servicio de ambulancia	Personas	Ambulancia

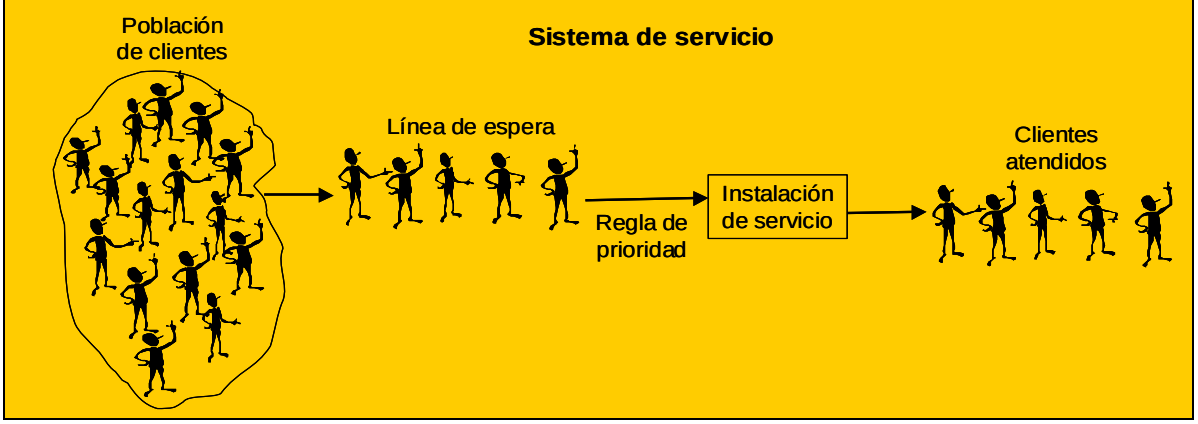
Existen muchos ejemplos adicionales de sistemas de líneas de espera importantes que pueden no estar dentro de las categorías mencionadas. Por ejemplo, un sistema judicial es una red de líneas de espera donde los juzgados son los locales de servicio, los jueces (o paneles de jueces) son los servidores y los casos en espera de juicio son los clientes.

Los sistemas de servicios de salud, como las salas de urgencias de los hospitales, también son sistemas de líneas de espera. Por ejemplo, las máquinas de rayos X y las camas de hospital se pueden ver como servidores en sus propios sistemas de líneas de espera.

Las aplicaciones iniciales de la teoría de líneas de espera (gracias a *A. K. Erlang* en la compañía telefónica de *Copenhague*) fueron la ingeniería telefónica, y el área general de las telecomunicaciones continúa siendo un área de aplicación muy importante. Más aún, todos tenemos nuestras propias líneas de espera personales: asignación de tareas, libros que leer, etcétera. Sin duda, los sistemas de líneas de espera prevalecen en muchas áreas de la sociedad.

2.3 ESTRUCTURA DE LOS PROBLEMAS DE LÍNEAS DE ESPERA

FIGURA 3 ELEMENTOS BÁSICOS DE MODELOS DE LÍNEAS DE ESPERA



El **sistema de servicio** describe el número de líneas y la distribución de las instalaciones.

Después de realizado el servicio, los clientes atendidos salen del sistema.
El análisis de los problemas de líneas de espera inicia con una descripción de los elementos básicos de la situación. Cada situación específica tendrá características diferentes, pero tienen elementos que son comunes a todas ellas (la **Figura 3** muestra estos elementos).

1. POBLACIÓN DE CLIENTES O FUENTE DE ENTRADA

Es un conjunto de individuos (no necesariamente seres vivos) que pueden llegar a solicitar el servicio en cuestión.

Los clientes que entran al sistema se generan a través del tiempo en una fuente de entrada.

En los sistemas de líneas de espera es común que los tiempos entre llegadas varíen, por lo que no se puede predecir la llegada del siguiente cliente al sistema sin embarbo es posible manejar dos casos:

1) Estimar el número esperado de llegadas por unidad de tiempo
2) Estimar la forma de distribución de probabilidad de los tiempos entre llegadas.

- *Tamaño de la Población*: Es el número total de clientes que pueden requerir servicio en determinado momento, es decir el número total de clientes potenciales distintos (el tamaño puede ser infinito o finito).
  - Población finita: Cuando el número de nuevos clientes para el sistema de servicio se ve reducido considerablemente por el número de clientes que ya están en el sistema

    Una *población finita* se refiere a un conjunto reducido de clientes que utilizarán el servicio y que, en ocasiones, deben formarse en una fila. La razón por la que es importante clasificarla como finita es que cuando un cliente sale de su posición como miembro de la población (por ejemplo, una máquina se descompone y requiere servicio), el tamaño del grupo de usuarios se reduce en uno, lo que a su vez reduce la probabilidad de que se vuelva a requerir el servicio. A la inversa, cuando se le ofrece el servicio a un cliente y regresa al grupo de usuarios, la población se incrementa al igual que la probabilidad de que un usuario requiera un servicio

    Por ejemplo, considere un grupo de seis máquinas a las que un encargado de reparaciones da mantenimiento. Cuando una máquina se descompone, la población fuente se reduce a cinco máquinas y la probabilidad de que una de las cinco restantes se descomponga y necesite una reparación es, menor que cuando había seis máquinas operando. Si hay dos máquinas descompuestas y sólo cuatro están operando, la probabilidad de que otra se descomponga cambia de nuevo. A la inversa, cuando una máquina se repara y vuelve a estar en servicio, la población de máquinas se incrementa, aumentando en consecuencia la probabilidad de una descompostura.

  - *Población infinita***.** Es aquella en la que el número de clientes en el sistema no afecta la tasa a la que la población genera nuevos clientes.

    Una *población infinita* es muy grande en relación con el sistema de servicio, de manera que el tamaño de la población, que es consecuencia de las

restas o sumas a la población (un cliente que necesita un servicio o un cliente que recibió el servicio y regresa a la población), no afecta de manera significativa las probabilidades del sistema.

Si en el caso anterior hubiera 100 máquinas en vez de seis, y si una o dos máquinas se descompusieran, la probabilidad de la descompostura de otra no sería muy diferente y podría suponerse, sin un error significativo, que la población (para todos los fines prácticos) es infinita.

Las fórmulas para los problemas de líneas de espera "infinitas" tampoco causarían un error significativo si se aplican a un médico con mil pacientes o a una tienda departamental con diez mil clientes.

- *Forma de las llegadas*: Patrón estadístico mediante el cual se generan los clientes a través del tiempo.

  La suposición normal es que los clientes se generan de acuerdo con un *proceso poisson,* esto equivale a decir que el tiempo entre dos llegadas consecutivas tiene una distribución de probabilidad exponencial.

2. CLIENTE

Es todo individuo de la población que solicita servicio. Suponiendo que los tiempos de llegada de clientes consecutivos son $0 < t_1 < t_2 < ...,$ será importante conocer el patrón de probabilidad según el cual la fuente de entrada genera clientes. Lo más habitual es tomar como referencia los tiempos entre las llegadas de dos clientes consecutivos: $T\ \ k\ = t_k - t_{k-1}$ fijando su distribución de probabilidad. Normalmente, cuando la población potencial es infinita se supone que la distribución de probabilidad de los $T\ \ k$ (que será la llamada distribución de los tiempos entre llegadas) no depende del número de clientes que estén en espera de completar su servicio, mientras que en el caso de que la fuente de entrada sea finita, la distribución de los $T\ \ k$ variará según el número de clientes en proceso de ser atendidos.

3. LÍNEAS DE ESPERA

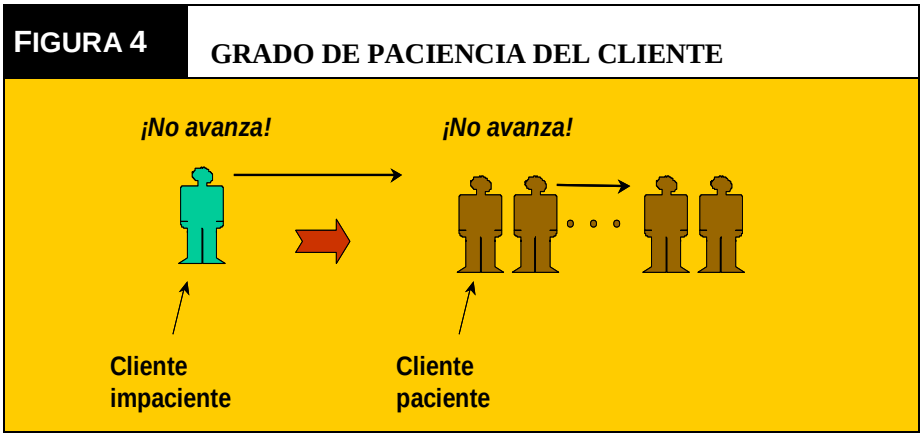
Línea de espera: es el conjunto de clientes que esperan, es decir los clientes que ya han solicitado el servicio pero que aún no han pasado al mecanismo de servicio.
Se caracteriza por el número máximo de clientes que se pueden admitir (antes de dar inicio el servicio)

- *Tamaño de la cola*: puede ser finita o infinita (el estándar es infinita).
  Lo más sencillo, a efectos de simplicidad en los cálculos, es suponerla infinita. Aunque es obvio que en la mayor parte de los casos reales la capacidad de la cola es finita, no es una gran restricción el suponerla infinita si es

extremadamente improbable que no puedan entrar clientes a la cola por haber llegado al número límite en la misma.

- *Regla de prioridad o Disciplina de la cola*:
  Se refiere al orden en el que se seleccionan los clientes para recibir el servicio. Las disciplinas mas usadas son:
  - FIFO (first in first out), también llamada FCFS (first come first served): según la cual se atiende primero al cliente que antes haya llegado (primero en llegar primero en salir.
  - LIFO (last in first out), también conocida como LCFS (last come first served) o pila: que consiste en atender primero al cliente que ha llegado el último.
  - RSS (random selection of service), o SIRO (service in random order), que selecciona a los clientes de forma aleatoria.
  - D*isciplina con derecho preferente* es una regla que permite al cliente de mayor prioridad interrumpir el servicio de otro cliente. Por ejemplo, en una sala de emergencia del hospital, los pacientes con las lesiones graves reciben tratamiento primero, sin importar su orden de llegada.
- *Grado de paciencia*
  Los clientes en las líneas de espera pueden ser *pacientes* o *impacientes*.
  - *Cliente paciente*: es el que entra en el sistema y permanece allí hasta que es atendido
  - *Cliente impaciente*: Es el que decide no entrar en el sistema o lo abandona antes de ser atendido.

FIGURA 4 **GRADO DE PACIENCIA DEL CLIENTE**



4. MECANISMO DE SERVICIO

Es el procedimiento por el cual se da servicio a los clientes que lo solicitan.

Para determinarlo se debe conocer el número de servidores de dicho mecanismo y la distribución de probabilidad del tiempo que le lleve a cada servidor dar un

servicio. Cuando los servidores tienen diferente destreza para dar el servicio, se debe especificar la distribución del tiempo de servicio para cada uno.

El mecanismo de servicio consiste en una o más instalaciones de servicio.

- *Instalación de servicio*: consistente de una persona, cuadrilla, máquina, grupo de máquinas, o ambas y que son necesarias para realizar el servicio para el cliente.

- *El sistema de servicio* queda definido por el número de líneas y los arreglos de las instalaciones.
  El canal Hace referencia al número de servidores que hay en el sistema.

- Tiempo de servicio: Es el tiempo que transcurre desde el inicio del servicio para un cliente hasta su termino.

  Cualquiera que sea el proceso de servicio, es necesario tener una idea de cuánto tiempo se requiere para llevar a cabo el servicio. Esta cantidad es importante debido a que cuanto más dure el servicio, más tendrán que esperar los clientes que llegan. Como en el caso del proceso de llegada, este tiempo pude ser determinístico o probabilístico.

  1. Servicio *determinístico*: cada cliente requiere precisamente de la misma cantidad conocida de tiempo para ser atendido.

  2. Servicio *probabilístico*: cada cliente requiere una cantidad distinta e incierta de tiempo de servicio. Los tiempos de servicio probabilísticos se describen matemáticamente mediante una distribución de probabilidad. En la práctica resulta difícil determinar cuál es la distribución real, sin embargo, una distribución que ha resultado confiable en muchas aplicaciones, es la exponencial. En este caso, su función de densidad depende de un parámetro μ y esta dada por y esta dada por:

     $$s\ t\ =\ {}^{1}\!\!\diagup\!\!_{\mu}\ e^{-\mu\tau}\ ,\qquad t\geq 0$$

     Donde:

     μ = número promedio de clientes atendidos por unidad de tiempo

     $1/\mu$ = tiempo promedio invertido en atender a un cliente

     El tiempo de servicio puede seguir cualquier distribución, pero, antes de analizar el sistema, se necesita identificar dicha distribución.

- Tiempo entre llegadas:

  1. Determinístico: Los clientes llegan en intervalos de tiempo fijo y conocido. Un ejemplo clásico es el de una línea de ensamble, en donde los artículos

llegan a una estación en intervalos invariables de tiempo (conocido como ciclos de tiempo)

2. Probabilístico: El tiempo entre llegadas es incierto y variable. Los tiempos entre llegadas probabilísticos se describen mediante una distribución de probabilidad.
   En el caso probabilístico, la determinación de la distribución real, a menudo, resulta difícil. Sin embargo, la distribución exponencial, ha probado ser confiable en muchos de los problemas prácticos. La función de densidad, para una distribución exponencial depende de un parámetro, digamos λ, y está dada por:

$$f\ t\ =\ 1/\lambda\ e^{-\lambda\tau},\qquad t\geq 0$$

   Donde λ es número promedio de llegadas en una unidad de tiempo.
   Con una cantidad, *t*, de tiempo se puede hacer uso de la función de densidad para calcular la probabilidad de que el siguiente cliente llegue dentro de las siguientes *T* unidades a partir de la llegada anterior, de manera que

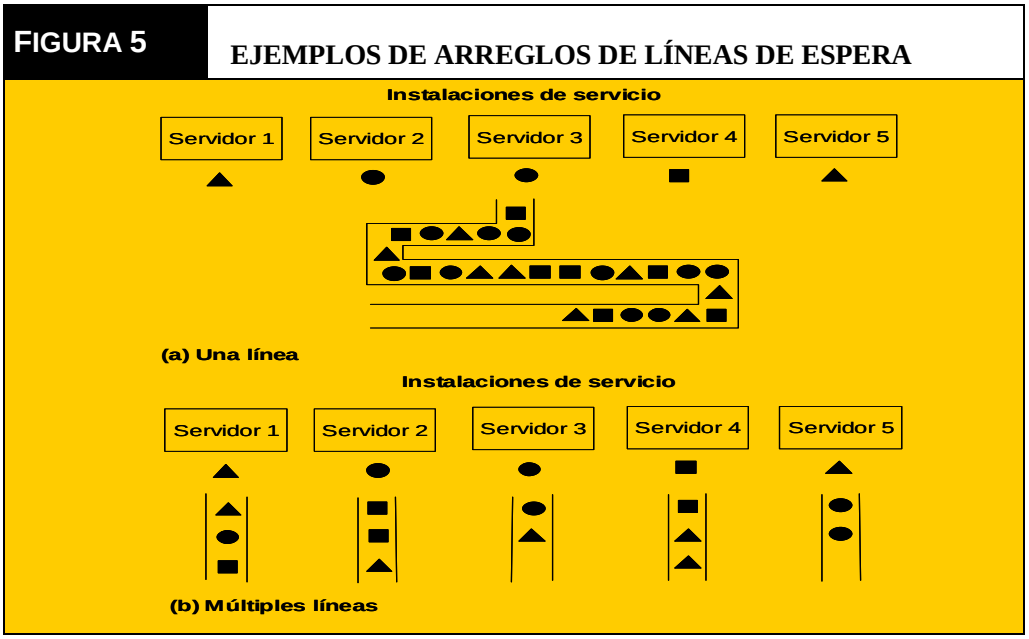
$$\text{P(tiempo entre llegadas} \leq \text{T)} = 1 - e^{-\lambda L}\ \ t \geq 0$$

- *Sistema de línea de espera*: es el conjunto formado por la línea de espera y el mecanismo de servicio, junto con la disciplina de la cola, que es lo que nos indica el criterio de qué cliente de la cola elegir para pasar al mecanismo de servicio.

2.4 NÚMERO DE LÍNEAS DE ESPERA

Las líneas de espera pueden diseñarse para que sean de una *sola línea* o *líneas múltiples*. La **Figura 5** muestra un ejemplo de cada arreglo.

FIGURA 5 EJEMPLOS DE ARREGLOS DE LÍNEAS DE ESPERA



Una línea se utiliza en los mostradores de las aerolíneas, dentro de los bancos, y en algunos restaurantes de comida rápida.
Cuando están disponibles múltiples servidores y cada uno puede manejar todas las transacciones, el arreglo de una sola línea mantiene a los servidores constantemente ocupados y da a los clientes un sentido de imparcialidad. Los clientes creen que son atendidos de acuerdo a como fueron llegando, no tienen que adivinar que línea de espera es la que avanzará más rápido.

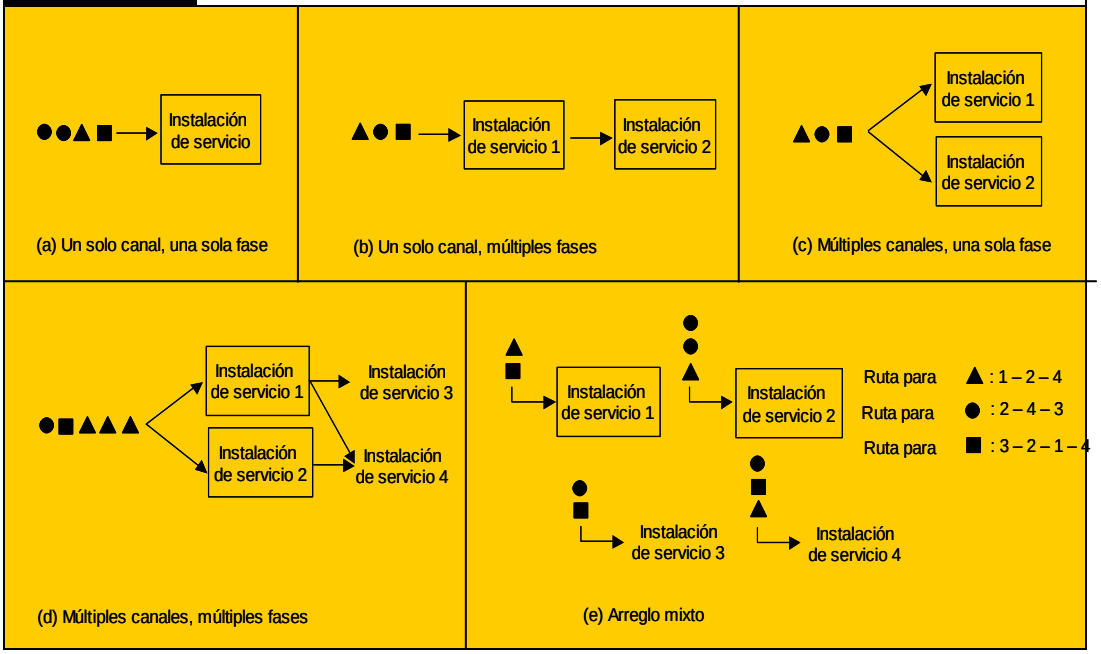
Las *líneas múltiples* se utilizan en los centros comerciales, en las casetas de las autopistas y en las tiendas de descuento.
El diseño de múltiples líneas optimiza cuando algunos de los servidores proporcionan servicios fijos limitados. En este arreglo, los clientes seleccionan los servicios que necesitan y esperan en la línea donde se proporciona ese servicio, como en los centros comerciales que hay líneas especiales para los clientes que pagan en efectivo o compran menos de 10 artículos.

2.5 ARREGLOS DE LAS INSTALACIONES DE SERVICIO

Las instalaciones de servicio consisten en el personal y equipo necesario para realizar el servicio requerido por el cliente. La **Figura 6** muestra cinco ejemplos de los tipos básicos de arreglos de la instalación de servicio.

Los gerentes deben seleccionar un arreglo considerando el número de clientes y la naturaleza de los servicios a realizar, algunos de los servicios requieren de un solo paso, también llamado **fase**, mientras que otros requieren de una secuencia de pasos.

FIGURA 6 ARREGLOS DE INSTALACIONES DE SERVICIO



a) Arreglo de **un canal-una fase**
   Los servicios requeridos por un cliente pueden ser realizados por una sola instalación de servicio. Los clientes forman una sola línea y pasan por la instalación, uno a la vez. Por ejemplos el manejo del automóvil a través del lavado automático, una peluquería atendida por una sola persona.

b) Arreglo de **un canal-múltiples fases**
   Se usa cuando los servicios se realizan mejor en serie por más de una instalación, aunque la cantidad de clientes u otras restricciones limitan el diseño de un canal. Los clientes forman una sola línea y se atienden en forma secuencial de una instalación de servicio a la siguiente. Un ejemplo de este arreglo es el servicio en el automóvil en *McDonald's*, donde la primera instalación toma la orden, las segunda realiza el cobro, y la tercera entrega la comida.

c) Arreglo de **múltiples canales-sola fase**
   Se usa cuando la demanda es muy grande para garantizar el ofrecimiento del mismo servicio en más de una instalación o cuando los servicios ofrecidos por las instalaciones son diferentes. Los clientes forman una o más líneas, dependiendo del diseño. En el diseño de una sola línea, los clientes son atendidos por el primer servidor disponible como en la sala de espera de un banco. Si cada canal tiene su propia línea de espera, los clientes esperan hasta que el servidor que corresponde a su línea pueda atenderlos, como en una instalación de servicio de un banco que atiende a los clientes en el automóvil.

d) Arreglo de **múltiples canales-múltiples fases**
   Ocurre cuando los clientes son atendidos por una de las instalaciones de la primera fase pero entonces requieren del servicio de una instalación de la segunda fase, y así sucesivamente. En algunos casos, los clientes no pueden cambiar de canal después de haber empezado el servicio; en otros pueden hacerlo. Un ejemplo de este arreglo es el de una lavandería automática. Las máquinas de lavado son las instalaciones de la primera fase, y las secadoras son de la segunda

e) Arreglo **mixto**
   Las líneas de espera se forman delante de cada instalación, como en un taller, donde cada trabajo personalizado requiere del uso de varias máquinas y asignaciones de ruta diferentes.

El problema de líneas de espera más complejo involucra a los clientes que tienen una sola secuencia de los servicios requeridos; por consiguiente, el servicio no puede describirse en fases. En dicho caso se usa un arreglo mixto

FIGURA 7 **EJEMPLOS DE ARREGLOS DE INSTALACIONES DE SERVICIO**

	Una sola fase	Múltiples fases
Un solo canal	Una persona "peluquería"	Lavado de autos
Múltiples canales	Ventanillas de los cajeros en un banco	Ingreso a un hospital

2.6 SUPOSICIONES GENERALES DE UN MODELO BÁSICO DE LÍNEAS DE ESPERA

1. Los tiempos de llegadas son independientes e idénticamente distribuidos de acuerdo a una distribución de probabilidad especificada.

2. Todos los clientes que llegan entran al sistema de colas y permanecen ahí hasta que termina el servicio.

3. El sistema de líneas de espera tiene una sola línea infinita, de modo que en la línea puede haber un número ilimitado de clientes (para propósito práctico).

4. La disciplina de la línea de espera es primero en entrar, primero en servir.

5. El sistema de líneas de espera tiene un número especificado de servidores, donde cada uno es capaz de servir a cualquiera de los clientes.

6. Cada cliente es atendido en forma individual por cualquiera de los servidores.

7. Los tiempos de servicio son independientes e idénticamente distribuidos de acuerdo con la distribución de probabilidad especificada.

2.7 MEDIDAS DE RENDIMIENTO PARA EVALUAR UN SISTEMA DE LÍNEAS DE ESPERA

Existen muchas medidas de rendimiento diferentes que se utilizan para evaluar un sistema de líneas de espera en estado estable.

Para diseñar y poner en operación un sistema de líneas de espera, por lo general, los administradores se preocupan por el nivel de servicio que recibe un cliente, así como el uso apropiado de las instalaciones de servicio de la empresa. Algunas de las medidas que se utilizan para evaluar el rendimiento surgen de hacerse las siguientes preguntas:

Preguntas relacionadas con el tiempo, centradas en el cliente

a. ¿Cuál es el tiempo promedio que un cliente recién llegado tiene que esperar en la fila antes de ser atendido?
La medida de rendimiento asociada es el tiempo promedio de espera, representado con W_q

b. ¿Cuál es el tiempo que un cliente invierte en el sistema entero, incluyendo el tiempo de espera y el de servicio?
La medida de rendimiento asociada es el tiempo promedio en el sistema, denotado con W

Preguntas cuantitativas relacionadas al número de clientes

a. En promedio ¿cuántos clientes están esperando en la línea de espera para ser atendidos?
La medida de rendimiento asociada es la longitud media de la línea de espera, representada con L_q

b. ¿Cuál es el número promedio de clientes en el sistema?
La medida de rendimiento asociada es el número medio en el sistema, representado con L

Preguntas probabilísticas que implican tanto a los clientes como a los servidores, por ejemplo:

a. ¿Cuál es la probabilidad de que un cliente tenga que esperar a ser atendido?
La medida de rendimiento asociada es la probabilidad de bloqueo, que se representa por, PW

b. En cualquier tiempo particular, ¿cuál es la probabilidad de que un servidor esté ocupado?
La medida de rendimiento asociada es la utilización. Esta medida indica también la fracción de tiempo que un servidor esta ocupado.

c. ¿Cuál es la probabilidad de que existan n clientes en el sistema?
La medida de rendimiento asociada se obtiene calculando la probabilidad P_0 de que no haya clientes en el sistema, la probabilidad P_1 de que haya un cliente en el sistema, y así sucesivamente. Esto tiene como resultado la distribución de probabilidad de estado, representada por $P_n, n = 0,1,2,...$

d. Si el espacio de espera es finito, ¿Cuál es la probabilidad de que la cola esté llena y que un cliente que llega no sea atendido?
La medida de rendimiento asociada es la probabilidad de negación del servicio, representada por Pd

Preguntas relacionadas con los costos

a. ¿Cuál es el costo por unidad de tiempo por operar el sistema?

b. ¿Cuántas estaciones de trabajo se necesitan para lograr mayor efectividad en los costos?

El cálculo específico de estas medidas de rendimiento depende de la clase de sistema de colas. Algunas de estas medidas están relacionadas entre sí. Conocer el valor de una medida le permite encontrar el valor de una medida relacionada.

2.8 FORMULA DE *LITTLE*

Indicaremos algunas relaciones útiles entre los valores esperados en estado estable de L, W, L_q, W_q

El cálculo de muchas de las medidas de rendimiento depende de los procesos de llegadas y de servicio del sistema de colas en específico. Estos procesos son descritos matemáticamente mediante distribuciones de llegada y de servicio. Incluso sin conocer la distribución especifica, las relaciones entre algunas de las medidas de rendimiento pueden obtenerse para ciertos sistemas de colas, únicamente mediante el uso de los siguientes parámetros de los procesos de llegada y de servicio.

λ=número promedio de llegadas por unidad de tiempo

μ=número promedio de clientes atendidos por unidad de tiempo en una sección

Supongamos una población de clientes infinita y una cantidad limitada de espacio de espera en la fila. El tiempo total que un cliente invierte en el sistema es la cantidad de tiempo invertido en la fila más el tiempo durante el cual es atendido:

$$\begin{Bmatrix}\text{Tiempo promedio}\\ \text{en el sistema}\end{Bmatrix} = \begin{Bmatrix}\text{Tiempo promedio}\\ \text{de espera}\end{Bmatrix} + \begin{Bmatrix}\text{Tiempo promedio}\\ \text{de servicio}\end{Bmatrix}$$

El tiempo promedio en el sistema y el tiempo promedio de espera están representados por las cantidades W y W_q, respectivamente. El tiempo promedio de servicio puede expresarse en términos de parámetros de

$$\rho = \lambda / s\mu$$

Donde *s*=número de servidores

Suponga que un cliente se une a la línea de espera en estado estable. En el momento que va a terminar el servicio, voltea a ver a los clientes que llegaron después que él. Habrá en promedio L_s clientes en el sistema. La cantidad esperada de tiempo que ha transcurrido desde que se unió a la línea de espera es, por definición W_s. Como los clientes llegan con una frecuencia constante λ, durante un tiempo W_s habrán llegado, en promedio, λW_s clientes, y de esto resulta

$$L_s = \lambda W_s.$$

EJEMPLO 6

Consideremos la relación entre el número promedio de clientes en el sistema y el tiempo promedio que cada cliente pasa en el. Imaginemos que un cliente acaba de llegar y se espera que permanezca en el sistema un promedio de media de hora. Durante esta media hora, otros clientes llegan a una tasa, digamos doce por hora. Cuando el cliente en cuestión abandona el sistema, después de media hora, ¿Cuál es el promedio de clientes nuevos que queda?

SOLUCIÓN

Deja tras de sí un promedio de (1/2)*12 = 6 clientes nuevos.

Es decir, en promedio, existen seis clientes en el sistema en cualquier tiempo dado.

Entonces

$$\begin{Bmatrix}\text{Tiempo promedio de}\\ \text{clientes en el sistema}\end{Bmatrix}=\begin{Bmatrix}\text{Numero promedio de}\\ \text{llegadas por unidad de tiempo}\end{Bmatrix}*\begin{Bmatrix}\text{Tiempo promedio}\\ \text{en el sistema}\end{Bmatrix}$$

Es decir $L=\lambda W$

En este caso, el argumento es esencialmente igual, excepto que el cliente voltea al momento de entrar en el servicio, y no después de ser atendido.

Utilizando una lógica parecida se obtiene la relación entre el número promedio de clientes que esperan en la línea y el tiempo promedio de espera en la fila:

$$\begin{Bmatrix}\text{Numero promedio de}\\ \text{clientes en el sistema}\end{Bmatrix}=\begin{Bmatrix}\text{Numero promedio de}\\ \text{llegadas por unidad de tiempo}\end{Bmatrix}*\begin{Bmatrix}\text{Tiempo promedio}\\ \text{en la cola}\end{Bmatrix}$$

De manera que $L_q=\lambda W_q$

Suponga que el tiempo medio de servicio es una constante $1/\mu$ para todo $1\le n$

$$W=W_q+1/\mu \qquad L=L_q+\rho$$

(Si la tasa media de servicio es μ, por consiguiente el tiempo promedio de servicio será $1/\mu$). La *fórmula de Little*, nombrada así en honor de *John D. C. Little* del *M.I.T.*, demostró ser válida bajo condiciones muy generales, es una relación sencilla, pero muy útil, entre las *L* y las *W.*

CAPÍTULO III

USO DE LOS MODELOS DE LÍNEAS DE ESPERA

Los gerentes de operaciones usan los modelos de líneas de espera para equilibrar las ganancias que pueden obtenerse si se aumenta la eficiencia del sistema de servicio contra los costos que esto involucra.
Las largas líneas de espera pueden causar que los clientes abandonen o estén inconformes con el servicio. Por lo que deben considerarse las siguientes operaciones características del sistema.

1. *Longitud de la línea.* El número de clientes en la línea de espera refleja:
   - Las líneas cortas significan que el servicio al cliente es bueno o tiene demasiada capacidad.
   - Las líneas largas indican poca eficiencia del servidor o la necesidad de aumentar la capacidad.

2. *Número de clientes en el sistema.* El número de clientes en la línea de espera y los que son atendidos están relacionados con la eficiencia y capacidad del servicio. Un número grande de clientes en el sistema ocasiona una congestión y puede producir el descontento del cliente, a menos que se aumente la capacidad.

3. *Tiempo de espera en la línea.* Las líneas largas no siempre significan tiempos largos de espera. Si la tasa de servicio es rápida, una línea larga se atiende con eficiencia. Sin embargo, cuando el tiempo de espera parece largo, los clientes perciben una mala calidad del servicio. Los gerentes pueden intentar cambiar la tasa de llegada de los clientes o diseñar el sistema de modo que el tiempo de espera parezca más corto de lo que realmente es. Por ejemplo, en *Walt Disney World* los clientes en línea para una atracción son entretenidos por videos y también son informados acerca del tiempo que les falta por esperar, lo que parece ayudarlos para que la espera no sea desesperante.

4. *Tiempo total en el sistema.* El tiempo total de espera en el sistema desde la entrada hasta la salida del mismo nos indica los problemas con los clientes, la eficiencia del servidor, o la capacidad. Si algunos clientes pasan demasiado tiempo en el sistema de servicio, puede haber la necesidad de cambiar la disciplina de prioridad, aumentar la productividad, o ajustar capacidad de alguna manera.

5. *Uso de la instalación de servicio*. El uso colectivo de las instalaciones refleja el porcentaje de tiempo que están ocupadas. La meta de la administración es mantener uso y rentabilidad alta sin afectar de manera adversa las demás características.

El mejor método para analizar un problema de líneas de espera es relacionar estas características de operación y sus alternativas en unidades monetarias. Sin

embargo, es difícil asignar una cifra de una unidad monetaria en ciertas características (como el tiempo de espera de un comprador en una tienda de alimentos). En tales casos, un analista puede ponderar el costo de implementar la alternativa bajo consideración contra una evaluación subjetiva del costo de *no* hacer el cambio.

3.1 CLASIFICACIÓN DE LOS MODELOS DE LÍNEAS DE ESPERA

Hay muchos modelos de líneas de espera posibles. Por ejemplo, si el tiempo que existe entre las llegadas en el modelo básico M/M/1 se le diera una distribución diferente (no la exponencial) tendríamos un modelo diferente. Para facilitar la comunicación entre aquellos que trabajan con modelos de líneas de espera, *D. G. Kendall* propuso una clasificación o taxonomía con base en la siguiente notación:

A/B/s

Donde

A = distribución de las llegadas.
B = distribución del servicio.
s = número de servidores

Se utilizan diferentes letras para designar ciertas distribuciones colocándolas en la posición A o B.

M = distribución exponencial.
D = número determinístico.
G = cualquier distribución (general) de tiempos de servicio.
GI = cualquier distribución (general) de tiempos de llegada.

MEDIDAS DE DESEMPEÑO

Existen varias maneras de juzgar la calidad del servicio en un sistema de procesamiento. Los resultados pueden evaluarse para un período corto una vez que el sistema abre, o por los resultados a largo plazo o de equilibrio. Por lo general, el tiempo en que los trabajos están en espera es importante y puede observarse el tiempo de espera promedio o una medida como la del porcentaje de los trabajos que esperan más. Una medida relacionada es el tiempo de rendimiento para un trabajo (tiempo de espera más tiempo de servicio); otra es la longitud de la línea de espera. Éstas son medidas de la calidad del desempeño del sistema, desde el punto de vista del cliente.

Otras medidas se relacionan con el costo de operación del sistema, cuyo factor de carga o de uso de la capacidad mide la capacidad del sistema para manejar la carga que llega. La gerencia tiene la opción de agregar más capacidad.

Un sistema de procesamiento dado puede tener cualquier combinación de los elementos descritos hasta ahora. Por consiguiente, existe un número muy grande de posibles sistemas, y ningún modelo matemático puede describirlos todos, por lo que se tratan algunos modelos simples de amplia aplicación y que dan una perspectiva acerca del comportamiento del sistema de líneas de espera, en general.

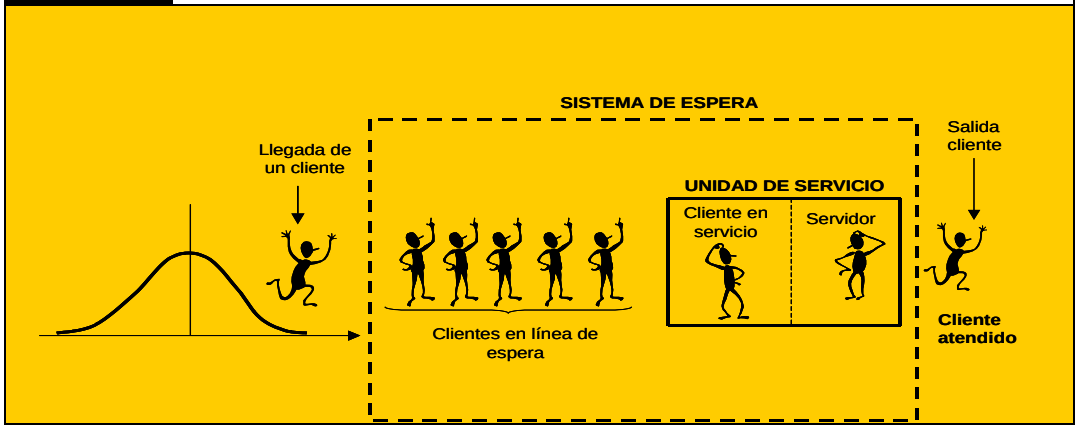
3.2 MODELO DE UN SOLO SERVIDOR M/M/1

3.2.1 MODELO DE UN SOLO SERVIDOR M/M/1 (0<ρ<1). POBLACIÓN INFINITA

El modelo más simple de líneas de espera involucra a un servidor y a una sola línea de clientes. Para especificar un poco más el modelo, se harán las siguientes suposiciones:

1. La población de clientes es infinita y todos los clientes son pacientes.
2. Los clientes llegan conforme a una distribución de *Poisson*, con una tasa media de llegada igual a λ.
3. La distribución de servicio es exponencial, con una tasa media de servicio iguala a μ.
4. Los clientes se atienden sobre una base primero en llegar, primero en ser atiende
5. La longitud de la línea de espera es ilimitada.

FIGURA 8 SISTEMA DE ESPERA



MEDIDAS DE DESEMPEÑO DEL MODELO M/M/1, POBLACIÓN INFINITA

Uso promedio del sistema o factor de carga del sistema $\rho = \frac{\lambda}{\mu}$

El requisito matemático de que $\rho < 1$ es necesario para garantizar la convergencia de la serie geométrica $(1+\rho+\rho^2+\cdots)$ que conduce a un argumento intuitivo. Fundamentalmente, $\rho < 1$ significa que $\lambda < \mu$ lo que establece que la tasa de llegadas debe ser estrictamente menor que la tasa de servicio, para que el sistema alcance estabilidad (condiciones de estado estable). Esto tiene sentido pues bajo otras condiciones, el tamaño de la línea de espera crecería indefinidamente.

Probabilidad de exactamente n clientes en el sistema

$$P_n = P_0\rho^n = P_0\left(\frac{\lambda}{\mu}\right)^n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n$$

Número promedio de clientes en el sistema de servicio

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{\mu\,\rho}{\mu - \mu\,\rho} = \frac{\mu\,\rho}{\mu(1-\rho)} = \frac{\rho}{1-\rho}$$

Numero promedio de clientes en la línea de espera

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

$$= \lambda W_q = \lambda\left(\frac{\lambda}{\mu(\mu - \lambda)}\right) = \frac{\lambda^2}{\mu(\mu - \lambda)} = L_s - \rho = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = \frac{\lambda\mu - \lambda(\mu - \lambda)}{\mu(\mu - \lambda)} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Número promedio de clientes en la línea de espera para un sistema ocupado

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

Tiempo promedio de espera en el sistema, incluye el servicio

$$W_s = \frac{L_s}{\lambda} = \frac{\frac{\lambda}{\mu - \lambda}}{\lambda} = \frac{\lambda}{\lambda(\mu - \lambda)} = \frac{1}{\mu - \lambda}$$

Tiempo promedio de espera en la línea

$$W_q = W_s - \frac{1}{\mu} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\mu - (\mu - \lambda)}{\mu(\mu - \lambda)} = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$W_q = \rho W_s = \frac{\rho}{\mu - \lambda} = \frac{\lambda}{\mu}\left(\frac{1}{\mu - \lambda}\right) = \frac{\lambda}{\mu(\mu - \lambda)}$$

Tiempo promedio que un cliente permanece en la línea de espera en un sistema ocupado

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

Probabilidad de que no haya clientes en el sistema

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right) = (1 - \rho)$$

Probabilidad de que haya n ó más clientes (unidades) en el sistema

$$P(n > k) = 1 - (P_0 + P_1 + P_2 + \ldots + P_{k-1} + P_k) = \rho^{k+1} = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

Probabilidad de que llegue un cliente (k) y tenga que esperar

$$P(n \geq k) = \rho^k$$

Probabilidad de que el tiempo de espera en el sistema exceda alguna cantidad de tiempo t

$$P(W_s > t) = e^{-\mu(1-\rho)t} \qquad \text{para } t \geq 0$$

Probabilidad de que el tiempo de permanencia en la línea de espera exceda a t

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

Relaciones básicas entre estos parámetros

Ley de Little:

$$L_s = \lambda W_s \qquad L_q = \lambda W_q .$$

$$W_s = W_q + 1/\mu \qquad L_s = L_q + \lambda/\mu .$$

EJEMPLO 7

Un fotógrafo de la embajada de los Estados Unidos toma las fotografías para los pasaportes a una tasa promedio de 20 por hora. El fotógrafo debe esperar hasta que el cliente deje de parpadear y hacer gestos, así que el tiempo para tomar una fotografía se distribuye exponencialmente. Los clientes llegan a una tasa promedio de acuerdo a una distribución de *Poisson* de 19 clientes por hora.

a. ¿Cuál es la utilización promedio del fotógrafo?
b. ¿Cuánto tiempo promedio permanece el cliente en el estudio del fotógrafo?

SOLUCIÓN

a. Las suposiciones en el enunciado del problema son consistentes con el modelo de un solo servidor. El factor de utilización del servidor es:

$$\rho = \frac{\lambda}{\mu} = \frac{19}{20} = 0.95 .$$

b. El tiempo promedio que el cliente permanece en el estudio del fotógrafo es

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{20 - 19} = 1 hora$$

EJEMPLO 8

El gerente de una tienda de abarrotes en una comunidad de jubilados está interesado en proporcionar buen servicio a los ciudadanos de la tercera edad que compran en su tienda. Actualmente, la tienda tiene una caja de cobro exclusiva para los ciudadanos de la tercera edad. En promedio, 30 ciudadanos de la tercera edad llegan por hora a la caja, de acuerdo a una *distribución de Poisson*, y se atienden a una tasa promedio de 35 clientes por hora, con tiempos de servicio exponencial. Determine las siguientes características de operación:

a. Probabilidad que no haya clientes en el sistema, o bien, la probabilidad de tener a todos los servidores desocupados.
b. Utilización promedio de la cajera.
c. Número promedio de clientes en el sistema.
d. Número promedio de clientes en la línea.
e. Tiempo promedio de espera en el sistema.
f. Tiempo promedio de espera en la línea.

SOLUCIÓN

a. $P_0 = \left(1 - \frac{\lambda}{\mu}\right) = \left(1 - \frac{30}{35}\right) = 0.1429$.

b. $\rho = \frac{\lambda}{\mu} = \frac{30}{35} = 0.8751.$

c. $L_s = \frac{\lambda}{\mu - \lambda} = \frac{30}{35 - 30} = 6$.

d. $L_q = \rho L_S = \left(\frac{\lambda}{\mu}\right) L_S = \left(\frac{30}{35}\right)(6) = 5.1429$

e. $W_s = \frac{L_s}{\lambda} = \frac{6}{30} = 0.2$

f. $W_q = \rho W_s = \left(\frac{\lambda}{\mu}\right) W_s = \left(\frac{30}{35}\right)(0.2) = 0.1714$

3.2.2 MODELO DE UN SOLO SERVIDOR M/M/1. POBLACIÓN FINITA

Consideraremos el caso en la que todas excepto una de las suposiciones del modelo de un solo servidor son apropiadas. En este caso, la población de clientes es finita, y tiene *N* clientes potenciales. Si *N* >30, el modelo de un solo servidor con la suposición de una población de clientes infinita es adecuado. De no ser así, el modelo de fuente finita es el único que debe usarse.
Las fórmulas utilizadas para calcular las características de operación de este sistema son las que se presenta a continuación.

MEDIDAS DE DESEMPEÑO

Considere

λ=tasa de llegadas promedio
μ =tasa de servicio promedio
N= tamaño de la población

Las medidas de desempeño del modelo de población finita con un único canal o servidor de turno son las siguientes:

Probabilidad que ningún cliente esté en el sistema o que el sistema esté vacío

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

Uso promedio del servidor $\rho = 1 - P_0$

Número promedio de clientes en la línea de espera

$$L_q = \sum_{n=1}^{N} (n-1) P_n = N - \left(\frac{\lambda + m}{\lambda} \right) (1 - P_0)$$

Número promedio de clientes (unidades) dentro del sistema

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

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

Tiempo de espera promedio en la línea

$$W_q = \frac{L_q}{\lambda_{ef}}$$

Tiempo de permanencia promedio dentro del sistema

$$W_s = \frac{L_s}{\lambda_{ef}} \qquad \text{donde } \lambda_{ef} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{N} (N-n) \lambda P_n = \lambda (N - L_s)$$

Número promedio de clientes en la línea en un sistema ocupado

$$L_b = \frac{L_q}{1 - P_0}$$

Tiempo promedio que el cliente permanece en la línea en un sistema ocupado.

$$W_b = \frac{W_q}{1 - P_0}$$

Probabilidad de que al llegar un cliente tenga que esperar o el sistema esté ocupado

$$P(n > 0) = 1 - P_0$$

Probabilidad de n clientes en el sistema

$$P_n = \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu} \right)^n P_0 \qquad \text{para } n=1,2,\ldots N$$

EJEMPLO 9

La *Worthington Gear Company* instaló un banco de 10 robots hace tres años. Desde entonces los robots aumentaron en gran medida la productividad de la empresa. Sin embargo, recientemente la atención se ha enfocado en su mantenimiento. La empresa no hace mantenimiento preventivo en los robots debido a la variabilidad que hay en la distribución de las descomposturas. Cada robot tiene una distribución de descompostura exponencial (o entre llegadas) con un tiempo medio entre fallas de 200 horas. Cada hora perdida por estar el robot descompuesto tiene un costo por interrupción de la producción de $30, lo que significa que la empresa tiene que reaccionar con rapidez ante las fallas de los robots. La empresa cuenta con una persona de mantenimiento, que necesita 10 horas en promedio para arreglar un robot. Los tiempos actuales de mantenimiento se distribuyen exponencialmente. La tasa de salario es de $10 por hora para la persona de mantenimiento quien trabaja productivamente si no hay robots descompuestos. Determine el costo diario por mano de obra y tiempo fuera de servicio del robot.

SOLUCIÓN

El *modelo de fuente finita* es apropiado para este análisis porque hay 10 máquinas en la población de clientes y se cumplen los otros supuestos. En este sentido $\lambda = 1/200$, ó 0.005 descomposturas por hora, y $\mu = 1/10 = 0.10$ robot por hora. Para calcular el costo de mano de obra y del tiempo fuera de servicio del robot, necesitamos calcular la utilización promedio de la persona de mantenimiento y L_s, el número promedio de robots en el sistema de mantenimiento.

El enunciado del problema describe un modelo de fuente finita, con N = 10. La probabilidad de que ningún robot esté en el sistema es

$$P_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1} = \left[\sum_{n=0}^{10} \frac{10!}{(10-n)!}\left(\frac{0.005}{0.10}\right)^n\right]^{-1}$$

$$= \left[\frac{10!}{10!}\left(\frac{0.005}{0.10}\right)^0 + \frac{10!}{9!}\left(\frac{0.005}{0.10}\right)^1 + \frac{10!}{8!}\left(\frac{0.005}{0.10}\right)^2 + \frac{10!}{7!}\left(\frac{0.005}{0.10}\right)^3 + \frac{10!}{6!}\left(\frac{0.005}{0.10}\right)^4 + \frac{10!}{5!}\left(\frac{0.005}{0.10}\right)^5\right.$$

$$\left. + \frac{10!}{4!}\left(\frac{0.005}{0.10}\right)^6 + \frac{10!}{3!}\left(\frac{0.005}{0.10}\right)^7 + \frac{10!}{2!!}\left(\frac{0.005}{0.10}\right)^8 + \frac{10!}{1!}\left(\frac{0.005}{0.10}\right)^9 + \frac{10!}{0!}\left(\frac{0.005}{0.10}\right)^{10}\right]^{-1}$$

$$= [1 + 0.5 + 0.225 + 0.09 + 0.0315 + 0.00945 + 0.0023625 + 0.0004725$$

$$+ 0.000070875 + 0.0000070875 + 0.00000035]^{-1} = 0.5380$$

Uso promedio de la persona de mantenimiento

$$\rho = 1 - P_0 = 1 - 0.5380 = 0.4620 = 46.20\%$$

Número promedio de robots en el sistema de mantenimiento

$$L_s = N - \frac{\mu}{\lambda}(1 - P_0) = 10 - \frac{0.10}{0.005}(1 - 0.5380) = 0.76$$

Tiempo promedio de espera en el sistema

$$W_s = L_s\left[(N - L_s)\lambda\right]^{-1} = 0.76\left[(10 - 0.76)(0.005)\right]^{-1} = 16.43 \text{ horas}$$

Número promedio de robots en la línea

$$L_q = N - \frac{\lambda + \mu}{\lambda}(1 - P_0) = 10 - \frac{0.005 + 0.10}{0.005}(1 - 0.5380) = 0.298$$

Tiempo promedio de espera en la línea

$$W_q = L_q\left[(N - L_s)\lambda\right]^{-1} = (0.298)\left[(10 - 0.76)(0.005)\right]^{-1} = 6.43$$

Los resultados muestran que la persona de mantenimiento utiliza sólo 46.2 por ciento del tiempo y el número promedio de robots que están esperando en la línea o reparándose es 0.76 de robot. Sin embargo, un robot espera un promedio de 16.43 horas en el sistema de reparación, de las cuales 6.43 horas espera en la línea por el servicio.
El costo diario por mano de obra y tiempo de inactividad del robot es:

Costo de mano de obra: ($10/hora) (8 horas/día) (0.462 utilización) = $ 36.96

Costo del robot averiado: (0.76 robot) ($30/robot hora) (8 horas/día) = 182.40

Costo total diario = $219.36

Número promedio de robots en la línea de espera en un sistema ocupado

$$L_b = \frac{L_q}{1 - P_0} = \frac{0.298}{1 - 0.5380} = 0.645 \text{ robot}$$

Tiempo promedio que un robot permanece en la línea de espera en un sistema ocupado

$$W_b = \frac{W_q}{1 - P_0} = \frac{6.43}{1 - 0.538} = 13.92 \text{ horas}$$

Probabilidad de que al llegar un robot tenga que esperar o el sistema esté ocupado.

$$P(n > 0) = 1 - P_0 = 1 - 0.5380 = 0.462 \text{ ó } 46.2\%$$

3.2.3 MODELO DE UN SOLO SERVIDOR CON CAPACIDAD FINITA

Un caso especial de línea de espera *M/M/1* con servicio y tasas de llegadas dependientes del servicio es aquel en que hay un área finita de espera. Si las llegadas se presentan cuando el área de espera está llena, se van. Los problemas de este tipo son comunes en sistemas de servicio como restaurantes, cines, salas de concierto, etcétera. También se pueden presentar en sistemas de manufactura, en que los amortiguadores entre los centros de trabajo tienen una capacidad finita. Supongamos que la cantidad máxima de clientes que permite el sistema es *K*.

MEDIDAS DE DESEMPEÑO

Probabilidad que ningún cliente esté en el sistema o que el sistema esté vacío:

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

Probabilidad de n clientes en el sistema:

$$P_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} \qquad \text{para } n = 0, 1, 2, \ldots, K$$

Para el caso de un área finita de espera no es necesario que $\rho < 1$. De hecho, P_n tiene ese valor para todos los valores de $\rho \neq 1$. Cuando $\rho = 1$ resulta que todos los estados son igualmente posibles, así que

$$P_n = \frac{1}{K+1} \qquad \text{para } 0 \leq n \leq K \text{ (sólo cuando } \rho = 1).$$

La *fórmula de Little* se sigue aplicando pero debemos usar un valor modificado de la tasa de llegadas, porque no a todos los clientes que llegan se les permite entrar al sistema. Cuando hay *K* o más en el sistema, la tasa de llegadas es 0, y entonces la tasa general de llegadas es menor que λ. La tasa efectiva de llegada, λ_{ef} se calcula como sigue:

$\lambda_{ef} = \lambda$P (Cantidad en el sistema < K)+0P(Cantidad en el sistema = *K*)

$= \lambda \left(1\text{-P}\{\text{cantidad en el sistema } 0K\}\right) = \lambda(1\text{-P}_n)$

Las medidas de desempeño se obtienen a partir de L_s, la cantidad esperada en el sistema, en estado estable. L_s se calcula como sigue:

Cantidad promedio de clientes (unidades) dentro del sistema

$$L_s = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \qquad \text{si } \rho = 1 \text{ entonces } L_s = \frac{K}{2}$$

Cantidad promedio de clientes (unidades) en la línea de espera

$$L_q = L_s - (1-P_0)$$

Tiempo promedio de espera en el sistema, incluye el servicio $W_s = \frac{L_s}{\lambda_{ef}}$

Tiempo promedio de espera de los clientes en la línea $W_q = \frac{L_q}{\lambda_{ef}}$

Factor de utilización del servicio $\rho = \frac{\lambda}{\mu}$

EJEMPLO 10

Una atracción conocida en la playa de *Acapulco* es un artista que pinta una caricatura en unos 5 minutos. Sin embargo, como los tiempos que requiere cada dibujo cambian considerablemente, se pueden describir bastante bien con una distribución exponencial. Las personas desean esperar su turno, pero cuando hay más de 10, optan por irse y se les sugiere regresen después. En las horas pico cabe esperar que haya hasta 20 clientes por hora. Suponga que los clientes llegan totalmente al azar en las horas pico.

a. ¿Qué proporción del tiempo tiene la línea de espera su máxima capacidad?
b. En promedio, ¿cuántos clientes optan por irse?
c. Determine las medidas de desempeño para este sistema de líneas de espera.
d. Si el área de espera duplicara su tamaño, ¿cómo afectaría a las respuestas de los incisos *a* y *b*?

SOLUCIÓN

$\lambda = 20$ por hora (Tasa de llegada)
$\mu = 60/5 = 12$ por hora (Tasa de servicio)
$\rho = \lambda/\mu = 20/12 = 5/3 = 1.667$.
K = 11 (Cantidad máxima en el sistema 10 en la línea de espera más el cliente que se está atendiendo).

a. Probabilidad de *n* clientes en el sistema $P_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}$

si el sistema está lleno $n = k$ y se puede expresar como:

$$P_K = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} = \frac{(1-1.667)(1.667)^{11}}{1-(1.667)^{12}} = \frac{-184.25}{-459.5} = 0.40.$$

b. Como la tasa de llegada es 20 por hora y el sistema está lleno el 40% del tiempo, durante las horas pico hay (20) (0.40) = 8 clientes por hora que se retiran. Esto da como resultado $\lambda_{ef} = 20 - 8 = 12$ por hora.

c. Medidas de desempeño.
Cantidad promedio de clientes dentro del sistema

$$L_s = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} = \frac{1.667}{1-1.667} - \frac{(11+1)(1.667)^{11+1}}{1-(1.667)^{11+1}}$$

$$= \frac{1.667}{-0.667} - \frac{5{,}525.97}{-459.5} = -2.5 + 12.03 = 9.53$$

Probabilidad de que ningún cliente esté en el sistema o que el sistema esté vacío

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}} = \frac{1-1.667}{1-(1.667)^{11+1}} = \frac{-0.667}{-459.5} = 0.00145 .$$

El valor (pequeño) de P_0 indica que el sistema casi nunca está vacío. En particular, el artista está inactivo ¡sólo el 0.145% del tiempo!

Cantidad promedio de clientes en la línea de espera

$$L_q = L_s - (1 - P_0) = 9.53 - (1 - 0.00145) = 8.53$$

Anteriormente se indicó que $\lambda_{ef} = 12$, de modo que el tiempo promedio de espera dentro del sistema es $W_s = \frac{L_s}{\lambda_{ef}} = \frac{9.53}{12} = 0.7942$ hora (unos 48 minutos)

y el tiempo promedio de espera de los clientes en la línea, W_q es:

$$W_q = \frac{L_q}{\lambda_{ef}} = \frac{8.53}{12} = 0.7108 \text{ hora (unos 43 minutos)}$$

d. Si el tamaño del área de espera duplicara su tamaño, entonces *K* = 21. En ese caso, P_K se determina con

$$P_{21} = \frac{(1-\rho)\rho^{21}}{1-\rho^{21+1}} = \frac{(1-1.667)(1.667)^{21}}{1-(1.667)^{22}} = \frac{-30{,}533.28}{-76{,}309.31} = 0.40$$

Es interesante que al duplicar la capacidad de la línea de espera no haya diferencia en relación con la probabilidad de que el sistema esté lleno. La razón es que como la tasa de llegada es mayor que la de servicio, el sistema llena su capacidad con rapidez en ambos casos. En estos, la tasa efectiva de llegada, λ_{ef} es aproximadamente igual a la tasa de servicio (aunque siempre se cumple que $\lambda_{ef} < \mu$). Aun cuando los sistemas sean mucho mayores, en este caso P_K es 0.4.

3.3 MODELO DE MÚLTIPLES SERVIDORES *M/M/S*

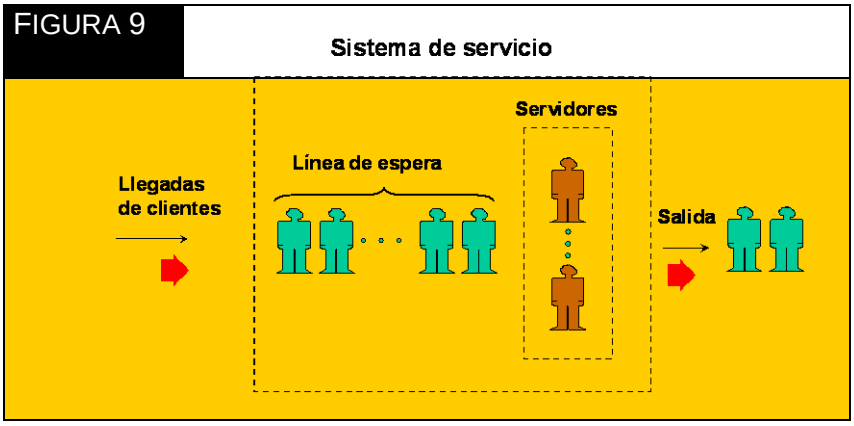
3.3.1 MODELO DE MÚLTIPLES SERVIDORES M/M/S (0<ρ<*S*). POBLACIÓN INFINITA

Con el modelo de múltiples servidores, los clientes forman una sola línea y los atiende uno de los *s* servidores cuando uno está disponible. Recuerde que el tercer símbolo de la etiqueta de un modelo de líneas de espera indica el número de servidores.

Para analizar un sistema de colas *M/M/s* consiste en lo siguiente:

1. una población de clientes infinita
2. un proceso de llegada en el que los clientes se presentan de acuerdo a un proceso de *poisson* con una tasa promedio de λ clientes por unidad de tiempo
3. un proceso de colas que consiste en una sola fila de espera de capacidad finita y disciplina primero en entrar primero en salir.
4. un proceso de servicio de *s* servidores idénticos, que atiende a los clientes con distribución exponencial, con una cantidad promedio μ de clientes por unidad de tiempo

FIGURA 9



Podemos observar que este sistema es distinto al *M/M/1* únicamente en el paso 4, que nos permite tener s servidores.

Recuerde también que $\rho = \lambda / \mu$ es el símbolo para el *factor de utilización* del servidor en un sistema de líneas de espera con un servidor. Para varios servidores, la fórmula cambia a $\rho = \lambda / s\mu$.

MEDIDAS DE DESEMPEÑO

Uso promedio del sistema: $$\rho = \frac{\lambda}{s\mu}$$

Probabilidad que cero clientes estén en el sistema (para $s\mu > \lambda$ *)*

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

Probabilidad que n clientes estén en el sistema

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0, & 0 < n < s \\ \frac{(\lambda/\mu)^n}{s! s^{n-s}} P_0, & n \geq s \end{cases}$$

Probabilidad de que haya s servidores y se tenga que esperar

$$P(n \geq s) = \frac{(\lambda/\mu)^s s\mu}{s!(s\mu - \lambda)} P_0$$

Número promedio de clientes en la línea de espera

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

Número promedio de clientes en el sistema de servicio

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

Número promedio de clientes en la línea para un sistema ocupado

$$L_b = \frac{L_q}{P(n \geq s)}$$

Tiempo promedio de espera de los clientes en la línea

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

Tiempo promedio de espera en el sistema, incluye el servicio

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

Tiempo promedio que permanece el cliente en la línea para un sistema ocupado

$$W_b = \frac{W_q}{P(n \geq s)}$$

Probabilidad de que el tiempo de espera en el sistema exceda a alguna cantidad de tiempo t

$$P\{W_s > t\} = e^{-\mu t}\left[1 + \frac{P_0\left(\frac{\lambda}{\mu}\right)^s}{s!(1-\rho)}\left(\frac{1 - e^{-\mu t\left(s-1-\frac{\lambda}{\mu}\right)}}{s-1-\frac{\lambda}{\mu}}\right)\right] \quad \text{para } t \geq 0$$

Cuando $s-1-\lambda/\mu = 0$, debe sustituirse $\dfrac{1-e^{-\mu t\left(s-1-\frac{\lambda}{\mu}\right)}}{s-1-\frac{\lambda}{\mu}}$ por μt.

Probabilidad de que el tiempo de espera en la línea exceda a alguna cantidad de tiempo t

$$P\{W_q > t\} = \left(1 - P\{W_q = 0\}\right)e^{-s\mu(1-\rho)t} \qquad \text{donde } P\{W_q = 0\} = \sum_{n=0}^{s-1} P_n$$

EJEMPLO 11

Una pequeña escuela de negocios ha asignado una secretaria a cada uno de sus departamentos: contabilidad, finanzas, mercadotecnia y dirección. Cada secretaria procesa material de clase y correspondencia únicamente para su propio departamento. El director de la escuela recibió muchas quejas, especialmente, del departamento de contabilidad acerca de los retrasos en los trabajos secretariales. El asistente del director reúne la información acerca de las tasas de llegada de los trabajos y tiempos de servicio. Después de un análisis detallado encuentra que las solicitudes de trabajo siguen una distribución de *Poisson* con $\lambda = 2$ solicitudes por hora, para todos los departamentos, excepto contabilidad, que tiene $\lambda = 3$ solicitudes por hora. El tiempo promedio de terminación de un trabajo (o servicio) es de 15 minutos independientemente del departamento que lo solicite, estos tiempos se distribuyen exponencialmente.

Debido a restricciones en el presupuesto ninguna secretaria adicional puede ser contratada. El director considera que el servicio puede mejorarse si se agrupa a las secretarias y se les ubica en un solo lugar, usando la política *primero que entra*, *primero que se le da el servicio*. Antes de proponer el nuevo plan el director pide a su asistente que analice el desempeño correspondiente y lo compare con el sistema actual.

SOLUCIÓN

El sistema actual consiste esencialmente de cuatro sistemas **M/M/1** independientes con una tasa de servicio $\mu = 4$ requerimientos por hora. La diferencia en los departamentos es la tasa de llegada. En todos los departamentos $\lambda = 2$ solicitudes por hora, excepto contabilidad. En estos casos ($\rho = \lambda / \mu$):

En el caso de los departamentos de **finanzas, mercadotecnia** y **dirección**

$L_s = \frac{\lambda}{\mu - \lambda} = \frac{2}{4-2} = 1$ trabajo.

$L_q = \frac{\rho\lambda}{\mu - \lambda} = \rho L = \left(\frac{2}{4}\right)(1) = \frac{1}{2} = 0.5$ trabajos

$W_s = \frac{1}{\mu - \lambda} = \frac{1}{4-2} = \frac{1}{2} = 0.5$ de hora ó 30 minutos.

$W_q = \frac{\rho}{\mu - \lambda} = \frac{(1/2)}{4-2} = \frac{1}{4}$ de hora ó 15 minutos.

$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{2}{4} = 0.5$

El número promedio de solicitudes de trabajo en la línea de espera para un sistema ocupado es $L_b = \frac{\lambda}{\mu - \lambda} = \frac{2}{4-2} = 1$ solicitud de trabajo

El tiempo promedio que una solicitud de trabajo permanece en la línea de espera en un sistema ocupado es $W_b = \frac{1}{\mu - \lambda} = \frac{1}{4-2} = 0.5$ horas

Probabilidad de que una solicitud que llegue tenga que esperar, o bien, el sistema esté ocupado $P(n \geq k) = \rho^k$, si $k = 1$ solicitud

$$P(n \geq 1) = \rho^1 = \frac{\lambda}{\mu} = \frac{2}{4} = 0.5 \text{ ó } 50\%.$$

En el caso del departamento de contabilidad estos valores son $\lambda = 3$ y $\mu = 4$

$L_s = \frac{\lambda}{\mu - \lambda} = \frac{3}{4-3} = 3$ trabajos.

$L_q = \frac{\rho\lambda}{\mu - \lambda} = \rho L = \left(\frac{3}{4}\right)(3) = \frac{9}{4} = 2.25$ trabajo

$W_s = \frac{1}{\mu - \lambda} = \frac{1}{4-3} = 1$ hora.

$W_q = \frac{\rho}{\mu - \lambda} = \frac{(3/4)}{4-3} = \frac{3}{4}$ de hora = 45 minutos.

$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{3}{4} = 0.25$

El número promedio de solicitudes de trabajo en la línea de espera para un sistema ocupado es

$$L_b = \frac{\lambda}{\mu - \lambda} = \frac{3}{4-3} = 3 \text{ solicitudes de trabajo}$$

El tiempo promedio que una solicitud de trabajo permanece en la línea de espera en un sistema ocupado es

$$W_b = \frac{1}{\mu - \lambda} = \frac{1}{4-3} = 1 \text{ hora}$$

Probabilidad de que una solicitud que llegue tenga que esperar, o bien, el sistema esté ocupado $P(n \geq k) = \rho^k$, si $k = 1$ solicitud

$$P(n \geq 1) = \rho^1 = \frac{\lambda}{\mu} = \frac{3}{4} = 0.75 \text{ ó } 75\%.$$

La propuesta para reunir al personal de secretarias, crea un sistema de múltiples canales, una sola línea de espera, o un sistema *M/M/4* en este caso. La tasa de llegada son las llegadas combinadas (2 + 2 + 2 + 3) en todos los departamentos, o bien, $\lambda = 2+2+2+3 = 9$ solicitudes de trabajo por hora, $\mu = 4$ solicitudes de trabajo por hora y $s = 4$ servidores.

El factor de utilización promedio del sistema es

$$\rho = \frac{\lambda}{s\mu} = \frac{9}{4(4)} = \frac{9}{16}$$

Probabilidad de que cero clientes estén en el sistema o que el sistema esté desocupado

$$P_0 = \left[\left(\sum_{n=0}^{s-1} \frac{(l/\mu)^n}{n!} \right) + \frac{(\lambda/\mu)^s}{s!(-\rho)} \right]^{-1} = \left[\left(\frac{(9/4)^0}{0!} + \frac{(9/4)^1}{1!} + \frac{(9/4)^2}{2!} + \frac{(9/4)^3}{3!} \right) + \frac{(9/4)^4}{4!(1-9/16)} \right]^{-1} = \left[\frac{27,204}{2,688} \right]^{-1} = 0.0988$$

Número promedio de clientes en el la línea de espera

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2} = \frac{\frac{224}{2267}\left(\frac{9}{4}\right)^4\left(\frac{9}{16}\right)}{4!\left(1-\frac{9}{16}\right)^2} = \frac{1.424456}{4.59375} = 0.31 \text{ solicitudes de trabajo}$$

Número promedio de clientes en el sistema

$$L_s = L_q + \frac{\lambda}{\mu} = 0.31 + \frac{9}{4} = 0.31 + 2.25 = 2.56 \text{ solicitudes de trabajo}$$

Tiempo promedio de espera de las solicitudes de trabajo en el sistema

$$W_s = \frac{L_s}{\lambda} = \frac{2.56}{9} = 0.28\,44 \text{ horas ó 17 minutos}$$

Tiempo promedio de espera de las solicitudes de trabajo en el la línea

$$W_q = \frac{L_q}{\lambda} = \frac{0.31}{9} = 0.034 \text{ horas} = 2.04 \text{ minutos.}$$

Probabilidad de que haya $s = 4$ servidores y se tenga que esperar.

$$P(n \geq s) = \frac{(\lambda/\mu)^s s\mu}{s!(s\mu - \lambda)} P_0 = \frac{\left(\frac{9}{4}\right)^4 (4)(4)}{4!(4 \times 4 - 9)}(0.09881) = \frac{410.0625}{168}(0.09881) = 24.1155\%$$

Número promedio de solicitudes de trabajo en la línea de espera en un sistema ocupado

$$L_b = \frac{L_q}{P(n \geq s)} = \frac{0.31}{0.241155} = 1.2855 \text{ solicitudes de trabajo}$$

Tiempo promedio que permanece el cliente en la línea de espera en un sistema ocupado

$$W_b = \frac{W_q}{P(n \geq s)} = \frac{0.034}{0.241155} = 0.141 \text{ horas u } 8.48 \text{ minutos.}$$

En resumen:

Estado	L_s	L_q	W_s	W_q	P_0
Actual (contabilidad)	3	2.25	1	0.75	0.25
Actual (finanzas, mercadotecnia y dirección)	1	0.5	0.5	0.25	0.50
Secretarias agrupadas	2.56	0.31	0.28	0.034	0.098

3.3.2 MODELO DE MÚLTIPLES SERVIDORES M/M/S (N≥s>1) POBLACIÓN FINITA

En los modelos de colas anteriores se ha supuesto que la población de clientes es infinita, aunque esto nunca es verdadero, para muchas situaciones practicas la suposición es razonable, sin embargo en algunos modelos esta suposición no es apropiada como en el caso del personal de mantenimiento en un laboratorio de computación con 50 computadoras. Como la población es muy limitada en tamaño, obtener medidas de rendimiento utilizando la suposición de que una población infinita puede producir resultados no validos

En este caso, la tasa de llegadas disminuye conforme aumenta el número de clientes en el sistema, porque existen menos clientes restantes que aun no llegan.

Los procesos de llegada para una población finita no se pueden describir de manera matemática mediante una tasa de llegada fija, ya que la tasa de llegada cambia de acuerdo al número de clientes que se encuentran en el sistema (entre más clientes haya en el sistema, menor será la tasa de llegada de clientes).

MEDIDAS DE DESEMPEÑO

Considere:

$\lambda =$ tasa de llegadas promedio

$\mu =$ tasa de servicio promedio,

$N =$ tamaño de la población

Las medidas de desempeño de este modelo de población finita para varios canales o servidores de turno son las siguientes:

Probabilidad que ningún cliente esté en el sistema o que el sistema esté vacío

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

Utilización promedio del servidor

$$\rho = 1 - P_0$$

Número promedio de clientes en la línea de espera

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

Número promedio de clientes (unidades) dentro del sistema

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

Tiempo de espera promedio en la línea $\qquad W_q = \frac{L_q}{\lambda_{ef}}$

Tiempo de permanencia promedio dentro del sistema

$$W_s = \frac{L_s}{\lambda_{ef}} \qquad \text{donde: } \lambda_{ef} = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{N} (N-n)\lambda P_n = \lambda(N - L_s)$$

Número promedio de clientes en la línea en un sistema ocupado

$$L_b = \frac{L_q}{1 - P_0}$$

Tiempo promedio que el cliente permanece en la línea en un sistema ocupado.

$$W_b = \frac{W_q}{1 - P_0}$$

Probabilidad de que al llegar un cliente tenga que esperar o el sistema esté ocupado

$$P(n > 0) = 1 - P_0$$

Probabilidad de n clientes en el sistema

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

3.3.3 MODELO DE MÚLTIPLES SERVIDORES *M/M/s*, CAPACIDAD FINITA

En este modelo no hay un área de espera o la capacidad de ésta es limitada. Como en el caso de reservaciones por teléfono que puede tener un número limitado de llamadas.

Cuando se llena el área de espera los clientes que llegan son rechazados y podrían o no regresar por lo que las medidas de rendimiento que se obtienen suponiendo un área de espera infinita pueden no ser válidas. Por lo que las formulas para calcular las medidas de rendimiento se modifican tomando en cuenta el espacio de espera limitado para obtener resultados validos

Resultados para el caso de varios servidores (s>1).
Como este modelo no permite más de K clientes en el sistema, K es el número máximo de servidores que pueden tenerse. Suponga que $s \le K$.

MEDIDAS DE DESEMPEÑO

Probabilidad que ningún cliente esté en el sistema o que el sistema esté vacío:

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

Probabilidad de n clientes en el sistema:

$$P_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & n = 1,2,\ldots,s \\ \dfrac{1}{s!s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & n = s,s+1,\ldots,K \\ 0 & n > K, \end{cases}$$

Cantidad promedio de clientes (unidades) en la línea de espera:

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

Cantidad promedio de clientes (unidades) dentro del sistema:

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

Tasa efectiva de llegada

$$\lambda_{ef} = \lambda P\left(\text{cantidad en el sistema<K}\right) + OP\left(\text{cantidad en el sistema=K}\right)$$
$$= \lambda\left(1 - P\left\{\text{cantidad en el sistema=K}\right\}\right) = \lambda\left(1 - P_k\right)$$

Tiempo promedio de espera en el sistema, incluye el servicio $W_s = \dfrac{L_s}{\lambda_{ef}}$.

Tiempo promedio de espera de los clientes en la línea $W_q = \dfrac{L_q}{\lambda_{ef}}$.

Factor de utilización del servicio: $\rho = \dfrac{\lambda}{s\mu}$

3.4 M/G/...

3.4.1 MODELO *M/G/1* (σ^2 = varianza del tiempo de servicio), POBLACIÓN INFINITA

Un supuesto que se requiere en algunos modelos es que la distribución del tiempo de servicio sea exponencial. Hay muchos casos en que este supuesto no tiene base. Cabría esperar que los tiempos de servicio fueran raramente exponenciales porque la distribución exponencial tiene la propiedad de amnesia: la cantidad de tiempo que queda en el servicio debería ser independiente del tiempo que ya pasó en ese servicio. Cabría esperar que una distribución modal, como la *normal* o la de *Erlang*, fueran un modelo más fiel de los tiempos de servicio en la mayoría de las circunstancias.

Por esta razón es muy interesante el modelo *M/G/1*. En este caso, G quiere decir "*general*" y significa que puede usarse cualquier distribución del tiempo de servicio con media $E(t)$ y varianza $V(t)$. La condición que $\rho < 1$ todavía se aplica para el estado estable ($\sigma = \lambda / \mu < 1$, donde ρ es igual a $\lambda E(t)$). Excepto para la generalización de la distribución del tiempo de servicio, se aplican las suposiciones del el modelo M/M/1 estándar. Las medidas de eficiencia o desempeño se deducen con la fórmula *Pollaczek-Khintchine* (*P-K*) en honor de dos pioneros del desarrollo de la teoría de líneas de espera que derivaron la fórmula de manera independiente a principios de los años 30 del siglo pasado.

El modelo M/G/1 supone que el modelo de líneas de espera tiene *un servidor* y un *proceso de entradas Poisson* (tiempos entre llegadas exponenciales) con una tasa media de llegadas *fija* λ.

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. De hecho, sólo es necesario conocer (o estimar) la media $1/\mu$ y la varianza σ^2 de esta distribución.

La flexibilidad total en cuanto a la distribución de los tiempos de servicio que proporciona este modelo es en extremo útil, por lo que es lamentable que no haya tenido éxito en el desarrollo de resultados análogos para el caso de varios servidores.

Desafortunadamente, no existe una ecuación para determinar las probabilidades del estado del sistema; sin embargo, se tienen las ecuaciones para L_s, L_q, W_s y W_q. Las medidas de desempeño para este modelo general son las siguientes.

MEDIDAS DE DESEMPEÑO

Factor de utilización del servicio $\rho = \dfrac{\lambda}{\mu}$

Probabilidad que ningún cliente esté en el sistema o que el sistema esté vacío

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

Cantidad promedio de clientes (unidades) dentro del sistema

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

Cantidad promedio de clientes (unidades) en la línea de espera

$$L_q = \frac{\lambda^2\{Var(T) + E(T)^2\}}{2(1-\lambda E(T))}$$

Usando la notación anterior $E(T) = 1/\mu$ y $\lambda E(T) = \lambda/\mu = \rho$;

igualando $Var(T) = \sigma^2$, esta fórmula se transforma en:

$$L_q = \frac{\lambda^2\{\sigma^2 + (1/\mu)^2\}}{2[1-(\lambda/\mu)]} = \frac{\lambda^2\sigma^2 + (\lambda/\mu)^2}{2[1-(\lambda/\mu)]} = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)} \qquad (+)$$

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 para L_q. Esta fórmula es uno de los resultados más importantes de la teoría de líneas de espera gracias a la facilidad con que se aplica y al predominio de los sistemas M/G/1 en la práctica. Esta ecuación para L_q (o su

contraparte para W_q) con frecuencia recibe el nombre de **fórmula de *Pollaczek-Khintchine.***

Tiempo promedio de espera de los clientes en la línea

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

Tiempo promedio de espera dentro del sistema, incluye el servicio

$$W_s = \frac{L_s}{\lambda} = \frac{L_q + \rho}{\lambda} = \frac{\lambda W_q + \frac{\lambda}{\mu}}{\lambda} = \frac{\frac{\lambda\mu W_q + \lambda}{\mu}}{\lambda} = \frac{\lambda(\mu W_q + 1)}{\mu\,\lambda} = \frac{\mu W_q + 1}{\mu} = W_q + \frac{1}{\mu}$$

Tiempo promedio que un cliente pasa en la línea de espera en un sistema que está ocupado

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

Note que para cualquier tiempo de servicio esperado fijo $1/\mu$, L_q, L_s, W_q y W_s se incrementan cuando σ^2 aumenta. Este resultado es importante porque indica que la consistencia del servidor tiene gran trascendencia en el desempeño de la instalación de servicio, no sólo su velocidad promedio. Por consiguiente, pueden mejorarse las medidas del desempeño de un sistema si se disminuye la varianza del tiempo de servicio, aunque no cambie el tiempo promedio de este.
El término σ^2 la varianza del tiempo de servicio en la ecuación (✢) proporciona algunas ideas interesantes. Evidentemente, el número estimado de clientes que esperan por el servicio (L_q) está directamente relacionado con la variabilidad de los tiempos de servicio (σ^2). Por ejemplo. El limitado menú de los restaurantes de comida rápida contribuye a su éxito, porque tal reducción en la variedad de alimentos que ofrecen permite la estandarización del servicio.

La varianza de la distribución exponencial es $1/\mu^2$, observe que sustituyendo este valor por σ^2 en la ecuación (✢), resulta:

$$L_q = \frac{\rho^2 + \lambda^2(1/\mu^2)}{2(1-\rho)} = \frac{\rho^2 + \lambda^2\left(\frac{1}{\mu^2}\right)}{2(1-\rho)} = \frac{\rho^2 + \left(\frac{\lambda}{\mu}\right)^2}{2(1-\rho)}$$

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

Que es equivalente a la ecuación de L_q para el modelo estándar *M/M/1.* Ahora considere el modelo *M/D/1*, con un tiempo de servicio determinista y varianza

cero. Nuevamente, de acuerdo a la ecuación (✈), cuando $\sigma^2 = 0$, entonces:

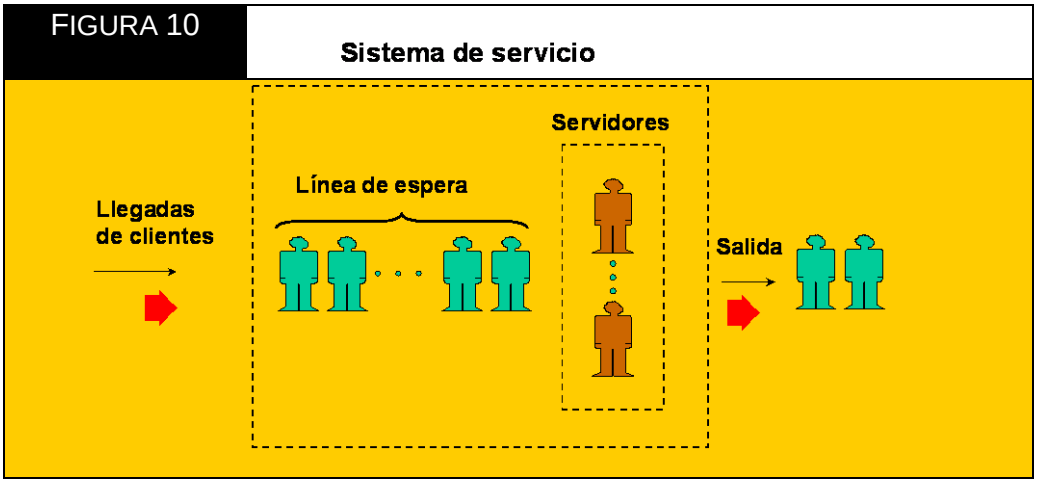
$$L_q = \frac{\rho^2}{2(1-\rho)}$$

De esta manera, se explica la reducción de la medida del congestionamiento L_q a la mitad por la variación en los tiempos de servicio. Esto implica que la variabilidad en el tiempo entre las llegadas cuenta para la congestión restante. Por lo tanto, existe un considerable potencial para la reducción de la congestión al utilizar simplemente citas o reservaciones para controlar la variabilidad en las llegadas. La congestión en un sistema de líneas de espera es igualmente ocasionada por la variabilidad en los tiempos de servicio y los tiempos entre llegadas; por consiguiente, las estrategias para controlar la congestión deben considerar ambos aspectos.

3.4.2 MODELO DE MÚLTIPLES SERVIDORES *M/G/s*, POBLACIÓN FINITA

El problema de cuántas líneas telefónicas se necesitan en un conmutador se ataca típicamente utilizando el modelo *M/G/s*, "eliminando a los clientes bloqueados". Este modelo es una línea de espera multi-canal con *s* servidores (*s* líneas), tiempos entre llegadas exponenciales para las llamadas y una distribución general para el tiempo de servicio, que en este caso es la duración de cada llamada. La frase "eliminación de los clientes bloqueados" Significa que *cuando una llamada encuentra todos los servidores ocupados* (*todas las líneas ocupadas*), *el usuario no ingresa en la línea de espera, sino que simplemente se va.* Esta frase describe claramente el comportamiento del conmutador telefónico tradicional. Los sistemas más sofisticados de ahora tienen una línea de espera para un número finito de clientes, en algunos casos, incluso proporcionando al cliente la oportunidad de disfrutar de una buena melodía.

FIGURA 10



Probabilidad de *j* servidores ocupados.

El problema de seleccionar la cantidad apropiada de líneas (servidores) se ataca calculando la probabilidad en estado estable de que exactamente *j* líneas estén ocupadas. Esto, a su vez, será utilizado para calcular la probabilidad de estado estable de que todas las *s* líneas estén ocupadas. Claramente, si se tienen *s* líneas y todas están ocupadas, la siguiente persona que llame no será capaz de lograr la llamada.

La probabilidad de estado estable de que haya exactamente *j* servidores ocupados, dado que s líneas (servidores) están disponibles, está dada por la expresión

$$P_j = \frac{(\lambda/\mu)^j / j!}{\sum_{k=0}^{s} (\lambda/\mu)^k / k!}$$

donde $\lambda =$ tasa de llegadas (la velocidad a que llegan las llamadas)

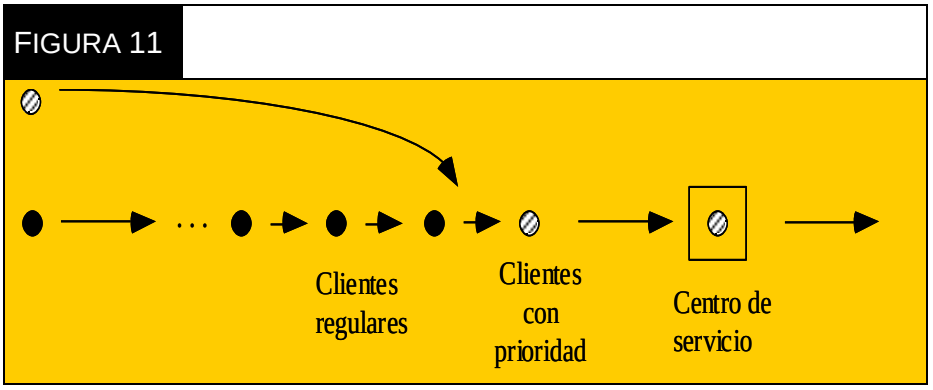
$\frac{1}{\mu} =$ tiempo medio de servicio (duración promedio de una conversación)

$s =$ número de servidores (líneas)

Esta expresión se conoce como *distribución de Poisson truncada o distribución de pérdida de Erlang*. Es válido observar que aunque estamos considerando una distribución general del tiempo de servicio, el valor P_j definido por la ecuación anterior depende solamente de la media de esta distribución.

3.5 MODELOS DE LÍNEAS DE ESPERA CON PRIORIDADES

FIGURA 11



Todos los modelos de líneas de espera presentados hasta ahora suponen que los clientes se atienden según primero en entrar, primero en salir. No todos los sistemas de líneas de espera funcionan así. En algunos se atiende primero a los clientes importantes, antes que otros que han esperado más. La administración puede querer que ciertos clientes especiales tengan prioridad sobre otros. En algunos casos, los clientes en los sistemas de líneas de espera pueden ser trabajos a realizar y las diferentes fechas de entrega para estos trabajos dicta el orden en que se sirve a estos clientes. Los trabajos urgentes necesitan hacerse antes que los rutinarios.

Una *sala de urgencias de un hospital* es un ejemplo de un sistema de líneas de espera donde se usan en forma automática las prioridades. Es natural que el paciente que llega en condición crítica, sea atendido antes que un paciente de rutina quien ya se hallaba esperando.

Los modelos para esos sistemas de líneas de espera hacen las siguientes suposiciones generales:

1. Hay dos o más categorías de clientes. Cada categoría se asigna a una **clase de prioridad**. Los clientes en la clase con prioridad 1 tienen preferencia para recibir el servicio antes que los clientes en la clase con prioridad 2. Si hay más de dos clases de prioridad, entonces, los clientes en la clase con prioridad 2 tienen preferencia sobre los clientes en la clase con prioridad 3, etcétera.

2. Después de clasificar a los clientes de prioridad más alta, los clientes en cada clase de prioridad se sirven sobre la base de primero en entrar, primero en salir. Así, dentro de cada clase de prioridad, la preferencia para recibir servicio se basa en el tiempo que han esperado en el sistema de líneas de espera.

Como se describe en seguida, en realidad hay dos tipos de prioridades.

1. Prioridades sin interrupción**:** una vez que un servidor ha comenzado a atender a un cliente, el servicio tiene que terminar sin interrupción aunque llegue un cliente de prioridad más alta mientras este servicio está en proceso. Una vez que termina el servicio, si hay clientes en la línea de espera, se aplican las prioridades para seleccionar a quien servir. En particular, el seleccionado es el miembro de prioridad más *alta* representada en la línea de espera quien más ha esperado.

2. Prioridades con interrupción: el cliente con menor prioridad que se está sirviendo se *interrumpe* (y regresa a la línea de espera) cada vez que entra al sistema un cliente con prioridad más alta. De este modo, se libera a un servidor para comenzar a servir de inmediato a la nueva llegada. Cuando un servidor logra *terminar* un servicio, se elige el siguiente cliente para comenzar el servicio justo como se describió para las prioridades *sin interrupción.* (El cliente excluido se convierte en el cliente de su clase de prioridad en la línea de espera que ha esperado más, con suerte, volverá a ser atendido pronto y, quizá después de otras interrupciones, terminará).

3.5.1 MODELO DE LÍNEAS DE ESPERA CON PRIORIDADES CON INTERRUPCIÓN

Junto con las suposiciones generales dadas para las prioridades, este modelo hace las siguientes suposiciones.

Suposiciones adicionales

1. Las prioridades con interrupción se usan como se acaba de describir. (Sea n el número de clases de prioridad).
2. Para la clase de prioridad i ($i = 1, 2, \ldots, n$), *los tiempos entre llegadas* de los

clientes en esa clase tienen una distribución *exponencial* con media $1/\lambda_i$.

3. Todos los *tiempos de servicio* tienen una distribución *exponencial* con media $1/\mu$, sin importar la clase de prioridad involucrada.
4. El sistema de líneas de espera tiene un solo servidor.

Así, excepto por la complicación de usar prioridades con interrupción, las suposiciones son las mismas que para el modelo *M/M/l.* Puesto que λ_i es la tasa media de llegadas para clientes en la clase de prioridad i $(i = 1, 2, \ldots, n)$, $\lambda = (\lambda_1 + \lambda_2 + \ldots + \lambda_n)$ es la tasa media de llegadas global para todos los clientes. Entonces, el *factor de utilización* del servidor es

$$\rho = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_n}{\mu}$$

Como en los modelos anteriores, se requiere que $\rho < 1$ para que el sistema de líneas de espera alcance una condición de estado estable para todas las clases de prioridad.

La razón para usar prioridades es *disminuir* los tiempos de espera de los clientes de alta prioridad. Esto se logra a expensas de *aumentar* los tiempos de espera de los clientes de baja prioridad.

Suponiendo que $\rho < 1$, hay fórmulas disponibles para calcular las medidas principales de desempeño (L_s, W_s, L_q y W_q) para *cada una* de las clases de prioridad.

3.5.2 MODELO DE LÍNEAS DE ESPERA CON PRIORIDADES SIN INTERRUPCIÓN

Junto con las suposiciones generales dadas para las prioridades, las de este modelo son las que se muestran enseguida.

Suposiciones adicionales

1. Se usan las prioridades sin interrupción como se describió anteriormente. (De nuevo, sea *n* el número de clases de prioridad).

2 y 3. Las mismas que para el modelo con prioridades con interrupción.

4. El sistema de líneas de espera puede tener cualquier número de servidores.

Excepto por el uso de prioridades no preferentes, estas suposiciones son las mismas que para el modelo *M/M/s*. El factor de utilización de los servidores es

$$\rho = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_n}{s\mu}$$

De nuevo, se requiere que $\rho < 1$ para que el sistema de líneas de espera alcance una condición de estado estable para todas las clases de prioridad.

Como antes, suponiendo que $\rho < 1$, hay fórmulas disponibles para calcular las medidas principales de desempeño (L_s, W_s, L_q y W_q) para *cada una* de las clases de prioridad.

Resultados para el modelo con prioridades sin interrupción:

Sea W_k el tiempo estimado de espera en el sistema en estado estable (incluyendo el tiempo de servicio) para un miembro de la clase de prioridad k. Entonces

$$W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}, \quad \text{para } k = 1, 2, \ldots, N,$$

Donde $A = s! \frac{s\mu - \lambda}{r^s} \sum_{j=0}^{s-1} \frac{r^j}{j!} + s\mu,$

$B_0 = 1,$

$B_k = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{s\mu},$

$s =$ número de servidores,

$\mu =$ tasa media de servicio por servidor ocupado,

$\lambda_i =$ tasa media de llegadas para la clase de prioridad i,

$\lambda = \sum_{i=1}^{N} \lambda_i$ y $r = \frac{\lambda}{\mu}$.

Estos resultados suponen que $\sum_{i=1}^{k} \lambda_i < s\mu$, de manera que la clase de prioridad k puede alcanzar una condición de estado estable.

La *fórmula de Little* se aplica a las clases individuales de prioridad, por lo que L_k, el número esperado de miembros de la clase de prioridad k en el sistema de líneas de espera (incluso que están en servicio), es

$$L_k = \lambda_k W_k, \quad \text{para } k = 1, 2, \ldots, N.$$

Para determinar el tiempo estimado en la línea de espera (sin incluir el tiempo de servicio) para la clase de prioridad k, sencillamente se resta $1/\mu$ de W_k; la longitud de la línea de espera correspondiente se obtiene de nuevo si se multiplica por λ_k.

Para el caso especial en el que $s = 1$, la expresión para A se reduce a $A = \mu^2 / \lambda$.

Variación de un servidor para el modelo de prioridades sin interrupción

La suposición anterior de que el tiempo de servicio esperado $1/\mu$ es el mismo para todas las clases prioritarias es bastante restrictiva. En la práctica, a veces se viola esta suposición debido a diferencias en los requerimientos de servicio entre las clases de prioridad.

Por fortuna, para el caso especial de un servidor, es posible admitir tiempos esperados de servicio distintos y de todas formas obtener resultados útiles.
Sea $1/\mu_k$ la media de la distribución exponencial del tiempo de servicio para la clase de prioridad *k,* entonces

$\mu_k =$ tiempo medio de servicio para la clase de prioridad k, para $k = 1, 2, \ldots, N.$

Así, el tiempo esperado de estado estable en el sistema para un miembro de prioridad k es

$$W_k = \frac{a_k}{b_{k-1}b_k} + \frac{1}{\mu_k}, \quad \text{para} \quad k = 1, 2, \ldots, N,$$

Donde: $a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2},$

$b_0 = 1,$

$b_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}$

Este resultado es válido cuando $\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} < 1,$ que permite a la clase de prioridad *k* alcanzar una condición de estado estable.

La *fórmula de Little* se puede usar como se describió para obtener las otras medidas de desempeño para cada clase prioritaria.

Resultados para el modelo de prioridades con interrupción

Para el modelo de prioridades con interrupción, se necesita retomar la suposición de que el tiempo esperado de servicio es el mismo para todas las clases de prioridad. Con la misma notación que para el modelo sin interrupción, el hecho de poder interrumpir cambia el tiempo esperado *total* en el sistema (incluyendo el tiempo total de servicio) a

$$W_k = \frac{1/\mu}{B_{k-1}B_k}, \quad \text{para} \quad k = 1, 2, \ldots, N,$$

para el caso de *un servidor* $(s = 1)$. Cuando $s > 1$, la W_k se puede calcular mediante un proceso iterativo que se ilustrará con el ejemplo del Hospital General. La L_k que se acaba de definir todavía satisface la relación

$$L_k = \lambda_k W_k, \quad \text{para} \quad k = 1, 2, \ldots, N.$$

Los resultados correspondientes para la línea de espera (excluyendo los clientes en servicio) también se pueden obtener a partir de W_k y L_k igual que para el caso de prioridades sin interrupción. Debido a la propiedad de falta de memoria de la distribución exponencial, las interrupciones no afectan el proceso de servicio (ocurrencias de terminaciones de servicio) de ninguna manera. El tiempo total esperado de servicio para cualquier cliente es $1/\mu$

EJEMPLO 12

Prioridades con interrupción. La sala de urgencias del HOSPITAL GENERAL proporciona cuidados médicos rápidos a los casos urgentes que llegan en ambulancia o vehículos particulares. En cualquier momento se cuenta con un doctor de guardia. No obstante, debido a la creciente tendencia a usar estas instalaciones para casos de emergencia en lugar de ir a una clínica privada, el hospital experimenta un aumento continuo en el número de pacientes anuales que llegan a la sala de urgencias. Como resultado, es bastante común que los pacientes que llegan durante las horas pico (temprano en la tarde) tengan que esperar turno para recibir el tratamiento del doctor. Por esto, se ha hecho una propuesta para asignar un segundo doctor a la sala de emergencia durante esas horas pico, para que se puedan atender dos casos de emergencia al mismo tiempo. Se ha pedido al ingeniero administrador del hospital que estudie esta posibilidad.

El ingeniero administrador comenzó por reunir los datos históricos pertinentes y hacer una proyección de estos datos al siguiente año. Reconoció que la sala de urgencias es un sistema de líneas de espera y aplicó varios modelos de teoría de líneas de espera para predecir las características de la espera en el sistema con uno y dos doctores. El ingeniero 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 con una tasa promedio de media hora También llegó a la conclusión de que el tiempo que necesita el doctor para atender a los pacientes sigue aproximadamente una *distribución exponencial*. Un doctor requiere un promedio de 20 minutos para atender al paciente. Los pacientes llegan a una.

El ingeniero administrador observó que los pacientes no se atienden sobre la base de primero en llegar, primero en salir. En su lugar, parece que la enfermera que realiza las admisiones divide a los pacientes en tres grandes categorías: 1) casos *críticos,* para los que el tratamiento inmediato es vital para la supervivencia; 2) casos *serios,* para los que un tratamiento rápido es importante para prevenir mayor daño, y 3) casos *estables,* en los que el tratamiento puede retrasarse sin consecuencias médicas adversas. Entonces, se atiende a los pacientes en este orden de prioridad, en donde los pacientes de la misma categoría por lo general se atienden según la regla de primero en llegar, primero en salir. Un doctor interrumpe el tratamiento de un paciente si llega un caso nuevo de una categoría de prioridad más alta. Aproximadamente 10% de los pacientes caen en la primera categoría, 30% en la segunda y 60% en la tercera. Como los casos más serios se internan en el hospital después de recibir el tratamiento urgente, el tiempo promedio de tratamiento por un doctor en la sala de urgencias en realidad no difiere mucho entre estas categorías.

El ingeniero ha decidido emplear el modelo de líneas de espera con disciplina de prioridades como una representación razonable de este sistema de líneas de

espera, en el que las tres categorías de pacientes constituyen las tres clases de prioridad del modelo. Como el tratamiento se interrumpe por la llegada de un caso de prioridad más alta, la versión de prioridades con interrupción es la apropiada. Con los datos que se obtuvieron $(\mu = 3$ y $\lambda = 2)$, los porcentajes anteriores conducen a $\lambda_1 = (0.10)(2) = 0.2$, $\lambda_2 = (0.30)(2) = 0.6$, $\lambda_3 = (0.60)(2) = 1.2$.

Obtención de resultados de prioridades con interrupción, s = 2
Estos resultados de prioridades con interrupción para $s = 2$ se obtuvieron como sigue. Ya que los tiempos de espera para los clientes de la clase de prioridad 1 no se ven afectados por la presencia de clientes en clases de prioridad más baja, W_1 será la misma para cualesquiera otros valores de λ_2 y λ_3, incluyendo $\lambda_2 = 0$ y $\lambda_3 = 0$. Entonces, W_1 debe ser igual a W para el modelo correspondiente de *una clase* (modelo M/M/s). Con $s = 2$, $\mu = 3$ y $\lambda = \lambda_1 = 0.2$, lleva a que $W_1 = W = 0.33$

3.6 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 $(0 < \sigma < 1/\mu)$, en el que caen la mayor parte de las distribuciones de tiempos de servicio *reales*. Otra distribución teórica de tiempos de servicio que concuerda con este espacio intermedio es la *distribución Erlang* (llamada así en honor del fundador de la teoría de líneas de espera).

La función de densidad de probabilidad para la distribución de *Erlang* es

$$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1} e^{-k\mu t}, \qquad \text{para } t \geq 0,$$

donde μ y k son parámetros estrictamente positivos de la distribución y k está restringido a valores enteros. (Excepto por esta restricción de entero y la definición de los parámetros, esta distribución es *idéntica a la distribución gama*). Su media y desviación estándar son:

$$\text{media} = \frac{1}{\mu} \qquad \text{y} \qquad \text{desviación estándar, } \sigma = \frac{1}{\sqrt{k}} \frac{1}{\mu}.$$

Así, 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 la teoría de líneas de espera por dos razones.

Para describir la primera suponga que $T_1, T_2, \ldots, T_k$ son k variables aleatorias independientes con una distribución exponencial idéntica, cuya media es $1/(k\mu)$.

Entonces se suma, $T = T_1 + T_2 + \ldots + T_k$, tiene una distribución *Erlang* con parámetros μ y k. El tiempo requerido para realizar cierto tipo tareas podría tener una distribución exponencial. Sin embargo, el servicio total requerido por un cliente, puede incluir una secuencia k de tareas, y no sólo una, realizadas por el servidor. Si las tareas respectivas tienen una distribución exponencial idéntica para su duración, el tiempo total de servicio tendrá una distribución *Erlang*; éste sería el caso, por ejemplo, si el servidor debiera realizar la *misma* tarea exponencial k veces para el cliente.

La distribución *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*. De hecho, tanto la distribución *exponencial* como la *degenerada* (constante) son casos especiales de la *distribución 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 para la media y la varianza se puede usar para elegir el valor de k que se ajuste a estas estimaciones de manera más cercana.

Cuando los tiempos de servicio tienen alguna variabilidad, pero menos que para la distribución exponencial, el modelo *M/E$_k$/s* proporciona un punto intermedio entre los modelos *M/M/s* y *M/D/s*.

Ahora considere el modelo *M/E$_k$/1*, que es justo el caso especial del modelo *M/G/1* donde los tiempos de servicio tienen una *distribución Erlang* con parámetro de forma $= k$. Para la distribución de *Erlang*, con parámetro de forma k $(k = 1, 2, \ldots)$ las medidas de desempeño son:

Factor de utilización del servicio $\rho = \lambda / \mu$.

Cantidad promedio de clientes (unidades) en la línea de espera:

$$L_q = \frac{\frac{\rho^2}{k} + \rho^2}{2(1-\rho)} = \frac{\frac{\lambda^2}{k\mu^2} + \left(\frac{\lambda}{\mu}\right)^2}{2\left(1-\frac{\lambda}{\mu}\right)} = \frac{\frac{\lambda^2}{k\mu^2} + \frac{\lambda^2}{\mu^2}}{2\left(1-\frac{\lambda}{\mu}\right)} = \frac{\frac{\lambda^2 + k\lambda^2}{k\mu^2}}{2\left(\frac{\mu-\lambda}{\mu}\right)} = \frac{\frac{\lambda^2(k+1)}{k\mu^2}}{2\left(\frac{\mu-\lambda}{\mu}\right)} = \frac{\mu\,\lambda^2(k+1)}{2k\mu^2(\mu-\lambda)} = \frac{\lambda^2(k+1)}{2k\mu(\mu-\lambda)} = \left(\frac{k+1}{2k}\right)\left(\frac{\lambda^2}{\mu(\mu-\lambda)}\right)$$

Si $\rho = \lambda / \mu$, entonces $\mu = \lambda/\rho$, de donde

$$L_q = \left(\frac{k+1}{2k}\right)\left(\frac{\lambda^2}{\lambda/\rho\,(\lambda/\rho - \lambda)}\right) = \left(\frac{k+1}{2k}\right)\left(\frac{\lambda^2}{\lambda/\rho(\lambda/\rho - \lambda)}\right) = \left(\frac{k+1}{2k}\right)\left(\frac{\lambda^2}{\lambda^2(1/\rho^2 - 1/\rho)}\right) = \left(\frac{k+1}{2k}\right)\left(\frac{1}{(1-\rho)/\rho^2}\right) = \left(\frac{k+1}{2k}\right)\left(\frac{\rho^2}{1-\rho}\right)$$

o bien $$L_q = \frac{\frac{\rho^2}{k} + \rho^2}{2(1-\rho)} = \frac{\frac{\rho^2 + k\rho^2}{k}}{2(1-\rho)} = \frac{(k+1)\rho^2}{2k(1-\rho)} = \left(\frac{k+1}{2k}\right)\left(\frac{\rho^2}{1-\rho}\right)$$

Cantidad promedio de clientes (unidades) dentro del sistema:

$$L_s = L_q + \rho = \lambda W_s$$

Tiempo promedio de espera de los clientes (unidades) en la línea:

$$W_q = \frac{L_q}{\lambda} = \frac{\left(\frac{k+1}{2k}\right)\left(\frac{\lambda^2}{\mu(\mu-\lambda)}\right)}{\lambda} = \left(\frac{k+1}{2k}\right)\left(\frac{\lambda}{\mu(\mu-\lambda)}\right)$$

Tiempo promedio de espera de los clientes (unidades) en el sistema, incluye el servicio:

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

EJEMPLO 13

La base de mantenimiento de *Friendly Skies Airline* cuenta con instalaciones para reparar sólo un motor de avión a la vez. Por lo tanto, devuelven los aviones a servicio lo más pronto posible, la política ha sido alternar la reparación de los cuatro motores de cada avión. En otras palabras, sólo se repara un motor cada vez que entra un avión al hangar. Con esta política, los aviones llegan en forma aleatoria a una tasa media de uno diario. El tiempo requerido para la reparación de un motor (una vez comenzado el trabajo) tiene una distribución exponencial con media de 1/2 día.

Se ha hecho una propuesta para cambiar la política a fin de que los cuatro motores se reparen de manera consecutiva cada vez que un avión entra al hangar. Esto significaría que cada avión necesitaría venir a la base de mantenimiento un cuarto de las veces. Puesto que el tiempo requerido para la reparación de un motor tiene una distribución exponencial, la teoría estadística indica que el tiempo requerido para la reparación de cuatro motores tiene una distribución *Erlang* con media cuatro veces mayor y parámetro de forma $k = 4$.

Ahora, la administración necesita decidir si continua trabajando como lo ha venido haciendo o adopta la propuesta. El objetivo es minimizar el tiempo promedio de vuelo de la flota completa perdido por día debido a las reparaciones de los motores.

a. Compare las dos alternativas el promedio de tiempo de vuelo perdido por un avión cada vez que llega a la base de mantenimiento.

b. Compare las dos alternativas respecto al número promedio de aviones que pierden tiempo de vuelo debido a la permanencia en la base de mantenimiento.

c. ¿Cuál de estas dos comparaciones es la apropiada para tomar decisiones administrativas? Explique.

SOLUCIÓN

a. y *b*. **Política actual**. Modelo *M/M/1*.
Observe que media $=1/\mu=1/2$, por lo tanto
$\mu=2$ aviones diarios,
$\lambda=1$ avión diario.

$$\rho=\frac{\lambda}{\mu}=\frac{1}{2}=0.5 \text{ ó } 50\%$$

$$L_s=\frac{\lambda}{\mu-\lambda}=\frac{1}{2-1}=1$$

$$L_q=\frac{\lambda^2}{\mu(\mu-\lambda)}=\frac{1^2}{2(2-1)}=\frac{1}{2}=0.5$$

$$W_s=\frac{L_s}{\lambda}=\frac{1}{1}=1$$

$$W_q=\rho W_s=\left(\frac{1}{2}\right)(1)=\frac{1}{2}=0.5$$

Propuesta: Modelo $M/E_k/1$.
Observe que: media $=1/\mu=(1/2)(4)=2$, por lo tanto
$\mu=1/2=0.5$ aviones diarios,
$\lambda=1/4=0.25$ de avión diario,
$k=4$,
$s=1$.

$$\rho=\frac{\lambda}{\mu}=\frac{1/4}{1/2}=\frac{1}{2}$$

$$L_q=\frac{k+1}{2k}\frac{\rho^2}{1-\rho}=\left(\frac{4+1}{2(4)}\right)\left(\frac{(1/2)^2}{1-1/2}\right)=\left(\frac{5}{8}\right)\left(\frac{1/4}{1/2}\right)=\left(\frac{5}{8}\right)\left(\frac{1}{2}\right)=\frac{5}{16}=0.3125$$

$$L_s=L_q+\rho=\frac{5}{16}+\frac{1}{2}=\frac{13}{16}=0.8125$$

$$W_q=\frac{L_q}{\lambda}=\frac{5/16}{1/4}=\frac{5}{4}=1.25$$

$$W_s=\frac{L_s}{\lambda}=\frac{13/16}{1/4}=\frac{13}{4}=3.25$$

Con la política actual un avión pierde 1 día de tiempo de vuelo, en cambio pierde 3.25 días con la política propuesta.

Con la política actual 1 avión está perdiendo tiempo de vuelo por día, en cambio 0.8125 de avión están perdiendo tiempo de vuelo con la política propuesta.

c. La comparación en la parte *b*. es la apropiada para tomar la decisión pues hay que tomar en cuenta que los aviones no pueden ir a servicio muy frecuentemente.

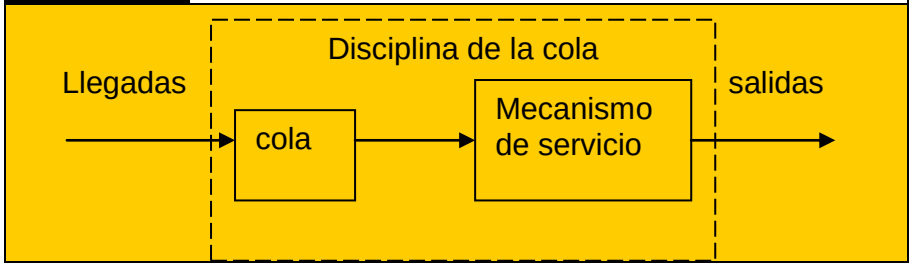
CAPÍTULO IV

COSTOS DE LOS SISTEMAS DE LÍNEAS DE ESPERA

4.1 COSTO DE LA ESPERA Y DEL SERVICIO

Un sistema de colas puede dividirse en sus dos componentes de mayor importancia, la cola y la instalación de servicio. Las *llegadas* son las unidades que entran en el sistema para recibir el servicio. Una vez que se completa el servicio, las llegadas se convierten en *salidas.*
Ambas componentes del sistema tienen costos asociados que deben de considerarse.

FIGURA 12 SISTEMAS DE COLAS



Aunque la teoría de colas proporciona información importante no es una herramienta que por sí sola permita llevar a cabo la toma de decisiones óptima

Las situaciones del tipo de sistema de colas que requieren la toma de decisiones surgen en muy diversas áreas, por lo que no es posible presentar un procedimiento general aplicable a todas las situaciones.

Una situación que en la práctica se presenta con frecuencia es decidir ¿Cuál es el número de servidores que se debe tener en cada instalación?
Para tomar las decisiones respecto a la capacidad de servicio que se ha de proporcionar se basan principalmente en:

- Costos derivados de prestar un buen servicio
- Costos derivados de tener largas cola

Para reducir los costos de servicio $\Rightarrow$ Bajo nivel de servicio
Para reducir los tiempos de espera $\Rightarrow$ Alto nivel de servicio

Se debe encontrar un balance entre el retraso promedio al solicitar un servicio y el costo de proporcionarlo

Para comparar los costos de servicio y los tiempos de espera, es necesario adoptar una medida común de su impacto

Se debe estimar el costo de espera
Se clasifica a los clientes en dos grupos:

- Clientes externos a la organización
  Costo de espera relacionado con la pérdida de ganancias por negocio perdido o costos sociales (Difícil de estimar)
- Clientes internos de la organización
  Costo de espera relacionado con la pérdida de ganancias debida a la productividad perdida.

FIGURA 13

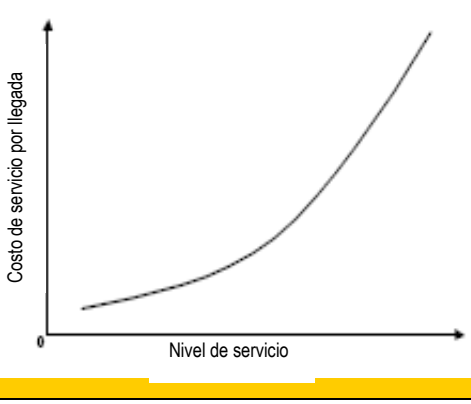
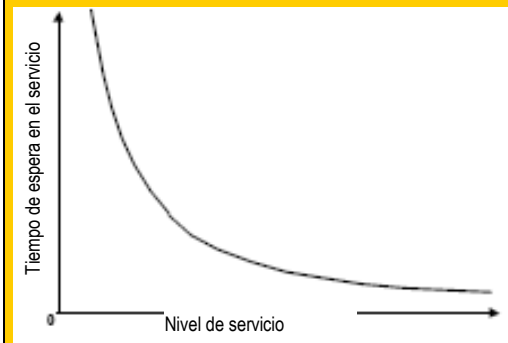


FIGURA 14



COSTO DE ESPERA

Esperar significa desperdicio de algún recurso activo que bien se puede aprovechar en otra cosa y esta dado por:

$$\text{Costo total de espera} = CW \quad L$$

Donde $CW =$ costo de espera por hora por llegada por unidad de tiempo
$L =$ longitud promedio de la línea.

COSTO DE SERVICIO

Este en la mayoría de ocasiones se trata de comprar varias instalaciones de servicio, en estos casos solo se ocupan los costos comparativos o diferenciales.

4.2 SISTEMA DE COSTO MÍNIMO

Aquí hay que tomar en cuenta que para tasas bajas de servicio, se experimenta largas colas y costos de espera muy altos.

Conforme aumenta la capacidad de servicio hay una reducción el numero de clientes en la cola y sus tiempos de espera por lo que disminuyen los costos de la línea de espera, aumenta el costo de servicio y por lo tanto el costo total disminuye, sin embargo, finalmente se llega a un punto de disminución en el rendimiento. Entonces el propósito es encontrar el balance adecuado para que el costo total sea el mínimo.

Las decisiones más comunes en las líneas de espera son fijar los parámetros:

a) Número requerido de servidores en una unidad de servicio

b) Número requerido de unidades de servicio

c) Eficiencia del servicio

de tal forma que se logre un equilibrio entre el costo de operación del sistema y el costo asociado a la espera. el nivel se puede fijar individualmente o combinando unos con otros

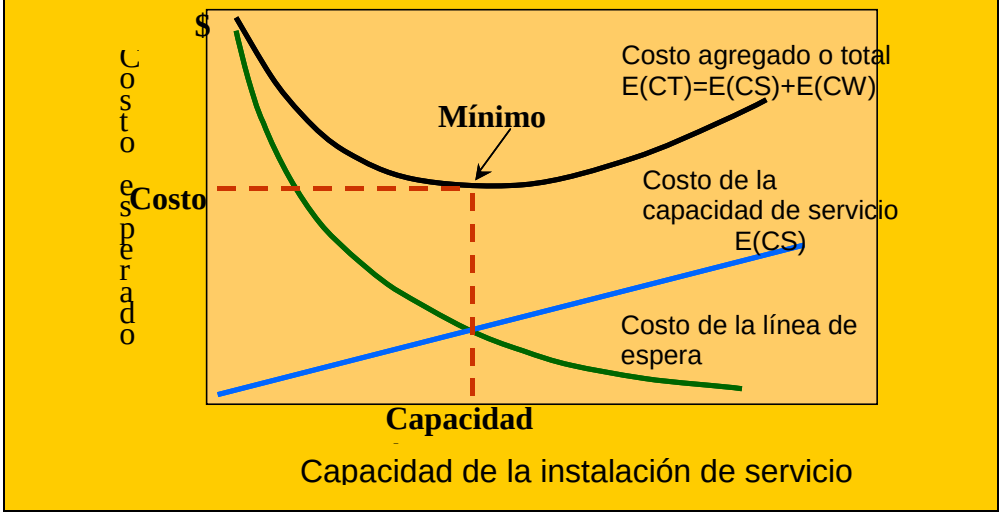
Una vez expresados el costo de espera y el costo del servicio se debe minimizar la suma de estas dos cantidades.

$$Minimizar E(CT) = E(CS) + E(CW)$$

E (CS): Costo esperado del servicio
E (CW): Costo esperado de la espera
E (CT): Costo esperado total

En la **Figura 15** se muestra la relación esencial del equilibrio en condiciones típicas (estado estacionario) de tránsito de clientes.

FIGURA 15 **CAPACIDAD DEL SERVICIO VERSUS LÍNEA DE ESPERA**



La variación en esta función suele estar representada por una curva exponencial negativa.

- El costo de la capacidad de servicio se muestra de una manera sencilla como una función lineal, más que como una función escalonada.
- El costo agregado o total se muestra como una curva en forma de U, que es una aproximación común en estos problemas de equilibrio.

- El costo óptimo idealizado se encuentra en el punto donde se cruzan las curvas de la capacidad de servicio y de la fila de espera.

Para expresar $E\ CW$ (función de costo espera) Como varía en realidad el costo en el que se incurre respecto al comportamiento real del sistema de colas
La forma de esta función depende del contexto del problema pero se puede modelar principalmente de dos maneras.

Si *g(n)* representa el costo esperado de demora de *n* clientes y P_n la probabilidad de que existan *n* personas en el sistema en un periodo de tiempo determinado

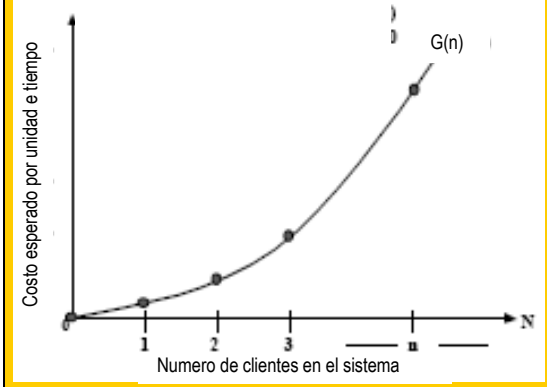
La forma *g(n)* puede ser lineal o no lineal (exponencial, etc.).

Caso 1. Forma *g(N)*

Clientes internos
Costo de espera relacionado con la pérdida de ganancias debida a la productividad perdida.
La propiedad que determina la tasa actual a la que se incurre en costos de espera es *N (el número de clientes en el sistema)*

g(N) se construye estimando *g(n)* cuando *N=n* para *n*=1,2,… y *g*(0)=0

FIGURA 16



Después de calcular las probabilidades P_n para un sistema de colas sepuede calcular $E\ CW = E\ g\ N$

Como *N* es una variable aleatoria, en este cálculo se emplea la expresión para el valor esperado de una función de una variable aleatoria discreta

$$E\ CW = \sum_{n=0}^{\infty} g\ n\ P_n$$

Cuando *g(N)* es una función lineal (es decir, la tasa de costo de espera es proporcional a N) $g\ \ N\ \ = C_w N$

Donde: C_w = Coto espera por unidad de tiempo para cada cliente

Entonces: $$E\ \ CW\ \ = C_w \sum_{n=0}^{\infty} nP_n = C_w L$$

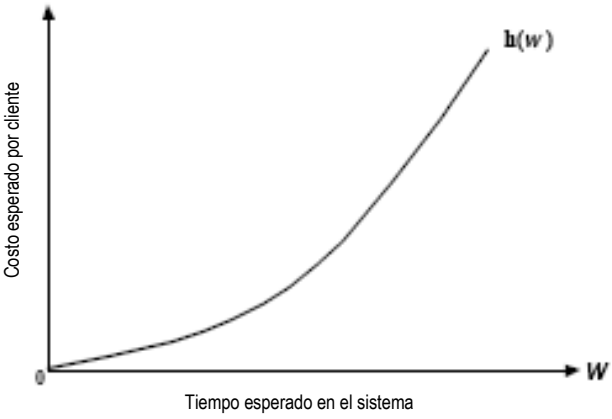
Caso 2. Forma h (*W*)

Clientes externos

Costo de espera relacionado con la pérdida de ganancias por negocio perdido o costos sociales

La propiedad que determina la tasa actual a la que se incurre en costos de espera es *W (el tiempo de espera en el sistema de colas para los clientes individuales)*

FIGURA 17



Una manera de construir *h(W)* es estimar *h(w)* para distintos valores de *w* y después ajustar un polinomio donde *h(w)* es el costo de espera en el que se incurre cuando un cliente espera en un tiempo $W = w$

$$E\ \ h\ \ W\ \ \ = \int h\ \ W\ \ \ f_W\ \ W\ \ dW$$

Donde: f_W = función de densidad de probabilidad de W

$$E\ \ h\ \ W\ \ \neq E\ \ CW$$

costo esperado de espera por cliente $\neq$ costo esperado de espera por unidad de tiempo

$$f_W = E\ \ CW\ \ = \lambda E\ \ h\ \ W\ \ \ = \lambda \int_0^{\infty} h\ \ W\ \ \ f_W\ \ W\ \ dW$$

EJEMPLO 14

Una fábrica cuenta con 10 máquinas para el ensamblaje de dispositivos para computadoras. Sin embargo la empresa sólo cuenta con 8 operadores, así que hay 2 máquinas de reserva para descomposturas eventuales.
Siempre que no haya más de 2 máquinas esperando reparación habrá 8 máquinas funcionando, pero este número se reduce en 1 por cada máquina que espera ser reparada por encima de 2

El tiempo que transcurre hasta que una máquina en operación se descompone (Distribución exponencial con media de 20 días)

El tiempo que se necesita para reparar una máquina
(Distribución exponencial con media de 2 días)

Actualmente la compañía cuenta con un mecánico para la reparación de estas máquinas pero está considerando la contratación de otro más.
Cada mecánico le cuesta a la compañía $280 por día
La pérdida de ganancias estimadas por tener menos de 8 máquinas en operación es $400 por día por cada máquina descompuesta.
¿Debe la compañía contratar un mecánico más?

SOLUCIÓN

N = 10 maquinas
λ = 1/120 clientes por día
μ = ½ clientes por día

Para S = 1

Parametros del modelo

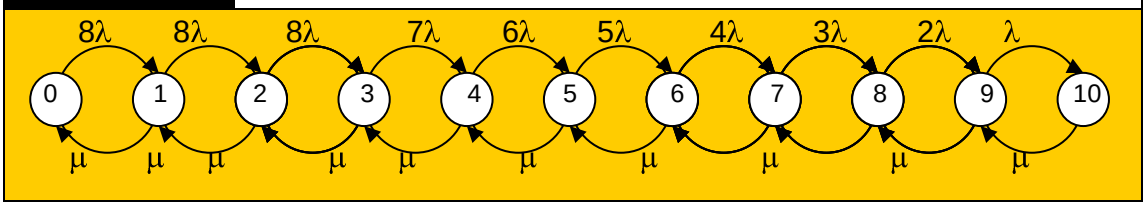
$$\lambda_n = \begin{cases} 8\lambda & \text{para n = 0,1,2} \\ 10\text{-n}\ \lambda & \text{para n = 3,4,...,10} \\ 0 & \text{para n > 10} \end{cases}$$

$\mu_n = \mu$ para n =1,2,...,N

Ecuacuon de balance	
estado 0	$\frac{1}{2}P_1 = \frac{8}{20}P_0$
estado 1	$\frac{8}{20}P_0 + \frac{1}{2}P_2 = \frac{8}{20} + \frac{1}{2}\ P_1$
estado 2	$\frac{8}{20}P_1 + \frac{1}{2}P_3 = \frac{8}{20} + \frac{1}{2}\ P_2$
estado 3	$\frac{8}{20}P_2 + \frac{1}{2}P_4 = \frac{7}{20} + \frac{1}{2}\ P_3$
estado 4	$\frac{7}{20}P_3 + \frac{1}{2}P_5 = \frac{3}{10} + \frac{1}{2}\ P_4$
estado 5	$\frac{3}{10}P_4 + \frac{1}{2}P_6 = \frac{1}{4} + \frac{1}{2}\ P_5$
estado 6	$\frac{1}{4}P_5 + \frac{1}{2}P_7 = \frac{1}{5} + \frac{1}{2}\ P_6$
estado 7	$\frac{1}{5}P_6 + \frac{1}{2}P_8 = \left(\frac{3}{20} + \frac{1}{2}\right) P_7$
estado 8	$\frac{3}{20}P_7 + \frac{1}{2}P_9 = \left(\frac{1}{10} + \frac{1}{2}\right) P_8$
estado 9	$\frac{1}{10}P_8 + \frac{1}{2}P_{10} = \left(\frac{1}{20} + \frac{1}{2}\right) P_9$
estado 10	$\frac{1}{20}P_9 = \frac{1}{2}P_{10}$

$$\sum_{n=0}^{10} P_n = 1$$

FIGURA 18



Para *s* = *2*

Parametros del modelo

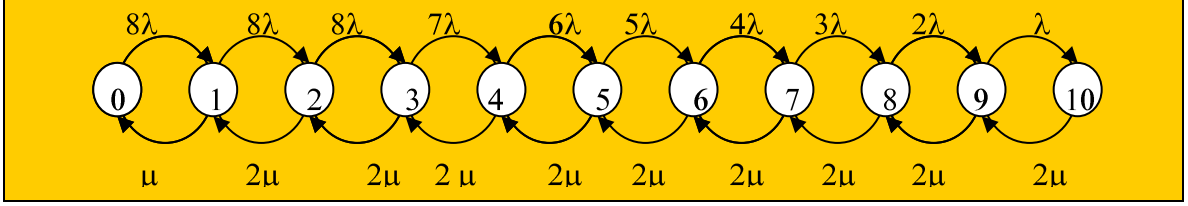
$$\lambda_n = \begin{cases} 8\lambda & \text{para n = 0,1,2} \\ 10\text{-n}\ \lambda & \text{para n = 3,4,...,10} \\ 0 & \text{para n > 10} \end{cases}$$

$$\mu_n = \begin{cases} n\mu & \text{para n = 1,2} \\ s\mu & \text{para s = 3,4,...} \end{cases}$$

Ecuacion de balance	
estado 0	$\frac{1}{2}P_1 = \frac{8}{20}P_0$
estado 1	$\frac{8}{20}P_0 + P_2 = \frac{8}{20} + \frac{1}{2}\ P_1$
estado 2	$\frac{8}{20}P_1 + P_3 = \frac{8}{20} + 1\ P_2$
estado 3	$\frac{8}{20}P_2 + P_4 = \frac{7}{20} + 1\ P_3$
estado 4	$\frac{7}{20}P_3 + P_5 = \frac{3}{10} + 1\ P_4$
estado 5	$\frac{3}{10}P_4 + P_6 = \frac{1}{4} + 1\ P_5$
estado 6	$\frac{1}{4}P_5 + P_7 = \frac{1}{5} + 1\ P_6$
estado 7	$\frac{1}{5}P_6 + P_8 = \frac{3}{20} + 1\ P_7$
estado 8	$\frac{3}{20}P_7 + P_9 = \left(\frac{1}{10} + 1\right) P_8$
estado 9	$\frac{1}{10}P_8 + P_{10} = \left(\frac{1}{20} + 1\right) P_9$
estado 10	$\frac{1}{20}P_9 = \frac{1}{2}P_{10}$

$$\sum_{n=0}^{10} P_n = 1$$

FIGURA 19



Las probabilidades P_n son:

N=n	s=1	s=2
0	0.271	0.433
1	0.217	0.346
2	0.173	0.139
3	0.139	0.055
4	0.097	0.019
5	0.058	0.006
6	0.029	0.001
7	0.012	3×10^{-4}
8	0.030	4×10^{-5}
9	7×10^{-4}	4×10^{-6}
10	7×10^{-5}	2×10^{-7}

Para decidir si la empresa debe contratar un mecánico adicional se debe observar la esperanza de costo total $E(CT)$

E(CW) la podemos obtener formulando una función *g(n)* , ya que en este caso las máquinas se pueden considerar clientes internos.

Recordemos que la pérdida de ganancias estimadas por tener menos de 8 máquinas en operación es $400 por día por cada máquina descompuesta.

Entonces: $g\ \ n\ = \begin{cases} 0 & \text{para n = 0,1,2} \\ 400\ \ \text{n-2} & \text{para n =3,4,...,10} \end{cases}$

Por lo tanto $E(CW) = \sum_{0}^{10} g\ \ n\ \ P_n$

De donde se tiene

		s=1		s=2	
N=n	g(n)	P_n	g (n) P_n	P_n	g(n) P_n
0	0	0.271	0	0.433	0
1	0	0.217	0	0.346	0
2	0	0.173	0	0.139	0
3	400	0.139	56	0.055	24
4	800	0.097	78	0.019	16
5	1200	0.058	70	0.006	8
6	1600	0.029	46	0.001	0
7	2000	0.012	24	3×10^{-4}	0
8	2400	0.030	7	4×10^{-5}	0
9	2800	7×10^{-4}	0	4×10^{-6}	0
10	3200	7×10^{-5}	0	2×10^{-7}	0
E=(CW)		$281 por día		$48 por día	

Como cada mecánico le cuesta a la compañía $280 por día, cuando *s* es conocida

$$E(CS) = sC_n \Rightarrow s\ \ \$280 \text{ por dia}$$

Entonces

S	sC	E(CW)	E(CT)
1	280	281	$561 por día
2	560	48	$680 por día
≥3	≥ 840	≥ 0	≥ 840 por día

Por lo tanto como el mínimo es $561 por día la compañía debe continuar con un mecánico para minimizar los costos totales derivados de la espera y el servicio.

Las llegadas van a la instalación de servicio de acuerdo con la *disciplina de la cola*, es decir, de acuerdo con la regla para decidir cuál de las llegadas se sirve

después. El primero en llegar primero en ser servido es una regla común, pero podría servir con prioridades o siguiendo alguna otra regla. Una vez que se completa el servicio, las llegadas se convierten en *salidas.*

Ambas componentes del sistema tienen costos asociados que deben de considerarse.

CONCLUSIONES

La teoría de colas es el estudio matemático de las colas o líneas de espera. La formación de colas es, por supuesto, un fenómeno común que ocurre siempre que la demanda efectiva de un servicio excede a la oferta efectiva.

Con frecuencia, las empresas deben tomar decisiones respecto al caudal de servicios que debe estar preparada para ofrecer. Sin embargo, muchas veces es imposible predecir con exactitud cuándo llegarán los clientes que demandan el servicio y/o cuanto tiempo será necesario para dar ese servicio; es por eso que esas decisiones implican dilemas que hay que resolver con información escasa. Estar preparados para ofrecer todo servicio que se nos solicite en cualquier momento puede implicar mantener recursos ociosos y costos excesivos. Pero, por otro lado, carecer de la capacidad de servicio suficiente causa colas excesivamente largas en ciertos momentos.

Cuando los clientes tienen que esperar en una cola para recibir el servicio, están pagando un costo, en tiempo, más alto del que esperaban.

Las líneas de espera largas también son costosas por tanto para la empresa ya que producen pérdida de prestigio y pérdida de clientes. La teoría de las colas en si no resuelve directamente el problema, pero contribuye con la información vital que se requiere para tomar las decisiones concernientes prediciendo algunas características sobre la línea de espera: probabilidad de que se formen, el tiempo de espera promedio.

Pero si utilizamos el concepto de "clientes internos" en la organización de la empresa, asociándolo a la teoría de las colas, nos estaremos aproximando al modelo de organización empresarial "just in time" en el que se trata de minimizar el costo asociado a la ociosidad de recursos en la cadena productiva.

SUGERENCIAS PARA ADMINISTRAR LAS LÍNEAS DE ESPERA

1. **Determinar un tiempo de espera aceptable para los clientes**.
   ¿Cuánto tiempo creen los clientes que deberán esperar? Establecer objetivos operacionales con base en lo que es aceptable.

2. **Tratar de desviar la atención de los clientes cuando esperan**.
   Si se proporciona música, un video u alguna otra forma de entretenimiento, puede ayudar a distraer a los clientes del hecho de que se les hace esperar.

3. **Informar a los clientes qué tiempo deben esperar**.
   Esto es especialmente importante cuando el tiempo de espera es más largo de lo normal. Informar por qué el tiempo de espera se prolonga más de lo normal y qué es lo que se esta haciendo para aligerar la espera.

4. **Mantener fuera de la vista de los clientes a los empleados que no los están atendiendo**.
   Nada es más frustrante para quien espera en una línea que ver a los empleados que potencialmente podrían estar atendiéndolos trabajando en otras actividades.

5. **Segmentar a los clientes.**
   Si un grupo de clientes necesita algo que puede hacerse con mayor rapidez, crear una línea especial, de manera que no tengan que esperar a causa de los clientes más lentos.

6. **Capacitar a sus servidores para que sean cordiales**.
   Saludar al cliente por su nombre, o bien proporcionarle alguna otra atención especial, puede para vencer los sentimientos negativos de una larga espera. (*Sugerencia*: en vez de decir a los servidores simplemente que sean cordiales, los psicólogos sugieren que se les diga cuándo deben recurrir a acciones cordiales específicas, como sonreír cuando saludan al cliente, cuando toman pedidos y al dar el cambio en una tienda. Las pruebas que se han hecho utilizando esas conductas específicas demostraron incrementos significativos en la percepción del cliente respecto de la actitud amistosa de los servidores.)

7. **Animar a los clientes para acudir durante periodos de poca actividad**. Informar a los clientes cuáles son los horarios en los que por lo común no tienen que esperar. También indicarles cuáles son los períodos pico; esto puede ayudar a mitigar la carga.

8. **Tener la perspectiva a largo plazo de deshacerse de las líneas de espera**.
   Desarrollar planes para formas alternativas de atención a sus clientes. Cuando sea apropiado, desarrollar planes para automatizar o acelerar de alguna manera el proceso. Esto no quiere decir que usted deba eliminar la atención personal, pues para algunos clientes ésta es deseable.

DEFINICIÓN DE SÍMBOLOS

n = número de clientes en el sistema.

λ = [*lambda*] tasa promedio de llegadas (por ejemplo, clientes que llegan por hora)

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

μ = [*mu*] tasa promedio de servicio para un servidor continuamente ocupado (por ejemplo, capacidad de servicio en clientes por hora).

$1/\mu$ = tiempo esperado de un servicio.

ρ = [*rho*] (λ/μ) utilización promedio del sistema o factor de carga del sistema.

N = número máximo de clientes permitidos en el sistema.

s = número de servidores.

P_n = probabilidad de exactamente n clientes en el sistema.

L_s = número promedio de clientes en el sistema.

L_q = número promedio de clientes en la línea de espera.

L_b = número promedio de clientes en la línea de espera para un sistema ocupado.

W_s = tiempo promedio que un cliente espera en el sistema.

W_q = tiempo promedio que un cliente permanece en la línea de espera.

W_b = tiempo promedio que un cliente permanece en la línea de espera en un sistema ocupado.

k = número de clientes.

Bibliografía

[1] Bonini Ch., Hausman W., Bierman H. *"Análisis cuantitativo para los negocios"*. Novena edición. Irwin-McGraw-Hill, 2000.

[2] Eppen G. D., F, J. Gould, C. P. Schmidt, J. H. Moore, L. R. Weatherford. *Investigación de operaciones en la ciencia administrativa*, quinta edición. Prentice Hall, 2000.

[3] Hamdy A. Taha *Investigación de operaciones. Quinta edición* Pearson, 1995

[4] Hillier, F. S., Hillier, M. S.; Lieberman, G. J. *"Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets"*. Segunda edición. Irwin-McGraw-Hill, 2002.

[5] Nahmias, Steven. *Análisis de la producción y las operaciones, tercera edición*. CECSA, 2004.

[6] Prawda Witenberg, J. *Métodos y modelos de la investigación de operaciones,* segunda edición. Limusa, México, 1981

[7] Ross S.M., A First Course in Probability, Prentice-Hall, New Jersey, 2008.

[8] Ross S.M., Introduction to Probability Models, Academic Press, New York, 1997.

[9] Saaty T. L., Elementos de la Teoría de Colas, Aguilar, Madrid, 1961.

[10] Yih-Long Chang**.** *WinQSB version 2.0*. Prentice Hall. 2006.