UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

DESARROLLO MATEMÁTICO DE LOS MODELOS BÁSICOS DE LÍNEAS DE ESPERA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

ANA LILIA DE LA CRUZ PLACENCIA

DIRECTORA DE TESIS:
DRA. IDALIA FLORES DE LA MORA

2004

FACULTAD DE CIENCIAS
SECCION ESCOLAR

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito: "Desarrollo Matemático de los Modelos Básicos de Líneas de Espera"

realizado por Ana Lilia De la Cruz Placencia

con número de cuenta 8918893-8 , quien cubrió los créditos de la carrera de: Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dra. Idalia Flores De la Mota
Propietario	M. en C. José Antonio Flores Díaz
Propietario	M. en A.P. María del Pilar Alonso Reyes
Suplente	M. en I.de O. María del Carmen Hernández Ayuso
Suplente	Act. Jaime Vázquez Alamilla

Consejo Departamental de Matemáticas

M. en C. José Antonio Flores Díaz

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL
DE
MATEMATICAS

A la **Universidad Nacional Autónoma de México** por permitirme hacer una carrera profesional.

Mi reconocimiento y gratitud a la Dra. Idalia Flores de la Mota por el tiempo y apoyo otorgado a esta tesis.

Igualmente mi más sincero agradecimiento a los demás miembros del jurado del examen profesional:

M. en A. P. María del Pilar Alonso Reyes
M. en I. de O. María del Carmen Hernández Ayuso
M. en C. José Antonio Flores Díaz
Act. Jaime Vázquez Alamilla

por la oportunidad que me brindaron al añadir sus conocimientos a este trabajo.

Mi agradecimiento:

A Dios Padre, Hijo y Espíritu Santo por estar siempre a mi lado, escucharme, quererme, ayudarme y darme la vida que me dio.

A mi Madre Caritina Placencia Arellano por su paciencia, su esfuerzo, su sacrificio, su apoyo y sobre todo su amor que, a pesar de mi forma de ser, siempre me ha dado. Gracias porque este trabajo lo realice por ti y para ti Mamá. Gracias por estar conmigo. Te amo.

A mi Padre el Ing. Francisco De la Cruz De la Riva por creer en mi, por haber sido una noble persona y por amarme más que a nadie en el mundo. Te amo y pienso en ti día tras día.

A mis Hijos Gustavo y Emmanuel por ser la razón de mi vida, por su ternura y amor. Los amo y estoy muy orgullosa de ser su madre.

A mi Esposo Gustavo por su esfuerzo y su gran responsabilidad que lo anima a seguir adelante, y por alentarme cuando estoy a punto de caer derrotada.

A mi hermano Francisco Javier por todas sus enseñanzas, por ser un gran músico y mi ejemplo durante tanto tiempo y que junto con mis Primos Raúl y Víctor me brindaron su cariño y me animaron a ser tan independiente.

A mis princesas Vanesa, Alicia, Nelly y Ana Raquel por su ternura. Y a Christian por el cariño que nos da.

A mis tíos Luis y Raúl por su apoyo y suplir de alguna manera a mi Padre.

A mis tías Raquel y Bella por su impresionante cariño, y a mi tía Virginia por su sinceridad.

A mis Primos Luis, Rodolfo, Roberto, Arturo, Aída y Claudia porque de ellos recibí diferentes enseñanzas.

A mi familia de Torreón por hacerme sentir especial.

A todos mis demás familiares que siempre me han apoyado de una u otra forma.

A mis amigos: Víctor Hugo Lemus Pacheco por escucharme y aconsejarme en cualquier momento y que junto con Wilfrido Pastrana Añorve, ejemplo a seguir, me dieron su apoyo cuando más lo necesite además de grandes momentos de risas y estudios.

A mis amigas: Carla Meneses Cleto por estar a mi lado en los grandes momentos de mi vida; Laura González Lozada primero por estar viva y segundo por animarme a que yo lo esté; Natalia Romero Pérez y Claudia Medellín Franco por esa amistad perdurable a pesar de la distancia.

Y por último a todas esas amistades que he tenido a lo largo de la vida que me apoyaron en mis estudios y que se alegran conmigo de que este trabajo está hecho.

INDICE

INTRODUCCIÓN

Las líneas de espera están presentes en la actualidad en cualquier área de una sociedad: existen filas en los bancos para recibir servicio, en los centros comerciales al momento de hacer los pagos de mercancía, en las instituciones educativas cuando se realizan trámites escolares, en las fábricas al momento en que los productos pasan por algún proceso de desarrollo o reparación, en el área de telecomunicaciones al esperar una transmisión, etc.

Cada una de las líneas de espera tiene asociados costos, ya sea por disponer de un servicio excesivo o por la falta de éste, que repercuten en pérdidas económicas dentro de un sistema.

La Teoría de Colas es el estudio matemático de las líneas de espera. Utiliza los modelos de colas para representar así a los diferentes sistemas que se presentan en la realidad. Cada modelo tiene asociadas fórmulas que indican cuál es el desempeño del sistema correspondiente y muestran el número promedio de espera que se presentará bajo ciertas circunstancias.

Las fórmulas mencionadas son de gran utilidad al modelar y explicar matemáticamente un problema pero desgraciadamente no existen libros en español que desarrollen cada fórmula e indiquen su punto de partida. Las traducciones de libros de Investigación de Operaciones, que existen en México, exponen de forma general las fórmulas finales y no se detienen, debido a que no es su propósito, a explicar el origen y desarrollo de ellas.

Así esta Tesis tiene su objetivo en el desarrollo de las fórmulas más frecuentes que se utilizan en la modelación matemática de la Teoría de Colas, ya que pensando en el hecho de que en la carrera de Actuaría se presentan conceptos matemáticos, financieros y estadísticos, que son desarrollados dentro de las diferentes materias para darle un mayor sustento a cada una de ellas; entonces al desarrollar las fórmulas de modelos de espera se sustentará la Teoría de Colas, elemento de la Investigación de Operaciones.

El Capítulo 1 define los sistemas de espera desde lo que es en sí una "cola" o línea de espera, la cantidad de posibles clientes, el tiempo en que éstos son servidos, hasta la forma en que se recibe el servicio.

Incluye también la notación de cada definición para así empezar a trabajar en lo que son las demostraciones de las fórmulas utilizadas en la Teoría de Colas.

Continúa el primer capítulo con el origen de todo proceso de colas que son las definiciones y demostraciones básicas del Proceso de Nacimiento y Muerte, así como la demostración práctica de la Fórmula de Little, que representa el equilibrio estadístico en las líneas de espera.

En el Capítulo 2 se demuestran, una a una, las fórmulas utilizadas en cada uno de los cuatro sistemas básicos y frecuentes de espera. Concluyendo con una breve explicación para poder llevar a cabo el proceso de decisión.

El Capítulo 3 expone 4 ejemplos prácticos de como se utilizan las fórmulas desarrolladas de los sistemas de colas vistas y 1 ejemplo que presenta la manera de realizar un proceso de decisión. En particular se presenta un

ejemplo de: el proceso de revalidación de la inscripción en la Maestría de Ingeniería de Sistemas en servicios escolares, una clínica de oftalmología en donde se realizan exámenes de glaucoma, un salón de belleza, una fábrica donde el proceso de pintar piezas de troncos, que fueron previamente cortados y pulidos, es la cola a estudiar, y por último, un negocio de comida rápida que necesita que se determine cuántos cajeros utilizar en cierto periodo del día.

Al final de este trabajo se presenta como apéndice un breve desarrollo de procesos estocásticos, una presentación de la distribución exponencial, y sus propiedades, y un glosario de conceptos básicos, que son antecedentes útiles para comprender más a fondo las fórmulas y definiciones aquí tratadas.

1. TEORIA DE COLAS

De alguna manera u otra, siempre se ha experimentado el esperar en una "cola" o línea de espera. Desafortunadamente, este fenómeno es más y más frecuente en la creciente sociedad urbanizada y congestionada.

Las personas esperan en una línea para entrar al metro, pagar en el supermercado, ser atendidos en servicios médicos, utilizar cajeros automáticos, comprar boletos en el cine, avanzar en los carros en el tránsito, etc., pero no solo es un fenómeno exclusivo para los humanos, también existen procesos de espera para diferentes materiales y objetos a construir o ensamblar dentro de las fábricas, así como para los aparatos o vehículos que necesitan reparaciones o mantenimiento, es decir, las líneas de espera se presentan de forma infinita en la sociedad.

Esta situación ocurre debido a que la demanda actual de un servicio es mayor a la capacidad actual para proporcionar tal servicio.

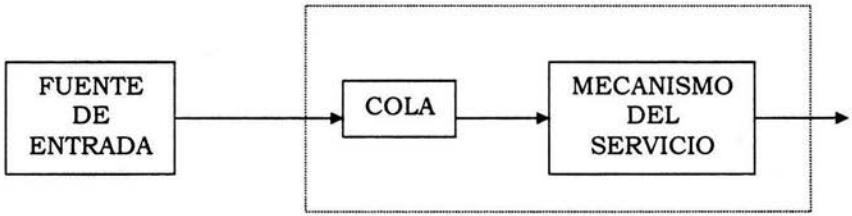
A los clientes, no les gusta esperar y a los gerentes o encargados de los establecimientos tampoco les parece que esto ocurra debido al posible gasto que representa para el negocio, que puede ser un costo social como lo es el de clientes perdidos. Pero proporcionar demasiado servicio también representa desembolsos excesivos de dinero, ya que un sistema con un número muy grande de servidores en el que cada cliente que llega es inmediatamente servido origina que se paguen sueldos al personal que tendrá bastantes tiempos ociosos.

El objetivo del negocio es lograr el equilibrio entre los costos del servicio y el asociado con su espera. Esto representa un proceso difícil ya que con frecuencia es imposible predecir con exactitud cuántas unidades llegarán a solicitar el servicio y cuánto tiempo se requerirá para proporcionarlo. La Teoría de Colas otorga la información vital requerida para tomar una decisión de este tipo.

1.1. ESTRUCTURA BÁSICA DE LAS LÍNEAS DE ESPERA

Un sistema de colas puede ser descrito por clientes que requieren un servicio (proporcionado por un servidor) en un determinado periodo.

Los clientes, que se generan por medio de una fuente de entrada, llegan aleatoriamente al sistema y forman una o varias líneas de espera para ser atendidos. Si el servidor está desocupado, de acuerdo con ciertas reglas preestablecidas, conocidas como disciplina en la cola, se selecciona a uno de los clientes formados para darle servicio. El cual será proporcionado mediante un mecanismo de servicio. El cliente será atendido en un periodo determinado de tiempo, llamado tiempo de servicio, al finalizar el cliente abandona el sistema.



1.1.1 Elementos en el proceso de las Líneas de Espera.

a) *Fuente de entrada.* Indica el número total de clientes potenciales distintos del servicio. Puede ser finita o infinita. En la realidad el tamaño es finito, pero en casos especiales tales como los que se presentan en este trabajo se supondrá sistemas con fuente de entrada infinita.

b) *Tiempo entre llegadas.* Es el lapso transcurrido entre el momento de la llegada de un cliente y el inmediatamente anterior (*inter-arribo*). Puede ser una constante o una variable aleatoria independiente, cuya distribución de probabilidad puede conocerse o no.

c) *Cola.* Se caracteriza el número máximo de clientes que pueden esperar en la cola. Pueden ser finitas o infinitas. En la realidad sólo existen las finitas. En este trabajo se presentan en las secciones 2.1 y 2.2 el caso infinito y en las secciones 2.3 y 2.4 sistemas con cola finita.

d) *Disciplina de la cola.* Se refiere a la forma en la que se seleccionan los miembros de la cola para recibir el servicio. Existen las siguientes políticas a saber: el primero en llegar a la cola es el primero a quien se le proporciona el servicio; prioridad, donde las características del cliente indican en qué orden se proporciona el servicio; "último en llegar, primero en ser servido", o bien, aleatoriedad.

e) *Mecanismo de servicio.* Consiste en la manera como esperan los clientes (en una o varias colas, con o sin opción a cambio de cola). Incluye la estructura de las estaciones de servicio que pueden estar en serie, en paralelo o mixtas.

f) *El tiempo de servicio.* Es el lapso transcurrido desde el momento en que se inicia el servicio de un cliente hasta su término. Este intervalo puede ser una constante o una variable aleatoria, dependiente o independiente cuya distribución de probabilidad se puede o no conocer.

g) *El número de servidores.* El sistema más simple es el de un solo servidor, el cual puede dar servicio a un solo cliente a la vez. Un sistema multi-servidores tiene n servidores idénticos y puede dar servicio a n clientes simultáneamente.

1.1.2 Notación en los modelos de Líneas de Espera

N(t): Número de clientes en el sistema de colas, en el instante t ($t > 0$).

λ : Número promedio (tasa) de llegadas al sistema por unidad de tiempo.

μ : Número promedio (tasa) de servicios por unidad de tiempo.

$1/\lambda$: Tiempo promedio que transcurre entre dos llegadas consecutivas.

$1/\mu$: Tiempo promedio de servicio de un cliente.

$\rho = \lambda/\mu$: Factor de utilización del sistema con un servidor.

c : Número de servidores en el sistema.

$\rho_c = \lambda/(c\mu)$: Factor de utilización de un sistema con c servidores múltiples.

$P_n(t)$: Probabilidad de que haya *n* clientes en el sistema en el tiempo *t*. (n=0,1...)

$P_0(t)$: Probabilidad de que en el momento *t* de arribo a la cola, el sistema se encuentren vacío.

L : Número esperado de clientes que permanecen en el sistema.

L_q : Número esperado de clientes en la cola.

W : Tiempo promedio de espera de un cliente en el sistema.

W_q : Tiempo promedio de espera de un cliente en la cola.

T : Tiempo de espera de un cliente en el sistema.

Estado del sistema: Número de clientes en el sistema de colas.

1.1.3 Notación para describir los sistemas de Líneas de Espera

La notación tiene la siguiente forma:

$$(a/b/c):(d/e/f)$$

donde:

a : Distribución de llegada.
b : Distribución del servicio.
c : Número de servidores en paralelo en el sistema.
d : Disciplina del servicio.
e : Máximo número de clientes que pueden estar en el sistema (esperando y recibiendo servicio).
f: Fuente de generación de clientes.

Los símbolos usados para a y b son:

M: Llegada con distribución de Poisson y servicio distribuido exponencialmente.
D: Llegada o servicio determinístico.
E: Llegada y servicios con distribución de Erlang y gamma respectivamente.
GI: Llegada con una distribución general independiente.
G: Servicios con una distribución general independiente.

Los símbolos usados para *d* son:

FCFS	:	Primero en llegar, primero en ser servido.
LCFS	:	Último en llegar, primero en ser servido.
SIRO	:	Servicio en orden aleatorio.
GD	:	Disciplina general de servicio.
NPRP	:	Servicio prioritario no abortivo.
RPP	:	Servicio prioritario abortivo.

1.2 PROCESO DE NACIMIENTO Y MUERTE

El proceso de nacimiento y muerte es caracterizado por la propiedad de que cuando ocurre una transición de un estado a otro, ésta ocurre sólo hacia un estado vecino, es decir, suponiendo que el espacio de estados es $S = \{0,1,2,...,i,...\}$, entonces la transición, cuando el proceso en el tiempo t se encuentra en el estado *i*, puede ser solamente al estado vecino (i-1) o (i+1). El término nacimiento se refiere a la llegada de un nuevo cliente al sistema de colas y la muerte se refiere a la salida de un cliente servido.

En términos generales se dice que los nacimientos y las muertes individuales ocurren de forma aleatoria en donde sus tasas medias de ocurrencia dependen únicamente del estado actual del sistema.

Considerando que λ_i es la tasa de llegadas al sistema y μ_i es la tasa de servicios o tasa de salida del sistema cuando éste se encuentra en el estado i.

Entonces una cadena de Markov de tiempo continuo N(t), $t \in T$ con espacio de estados $S = \{0,1,2,...\}$ y con tasas q_i y q_{ij},

donde:

q_i es la tasa de transición hacia fuera del estado i por unidad de tiempo que pasa en el estado i

y

q_{ij} es la tasa de transición del estado i al estado j en el sentido de que q_{ij} es el número esperado de veces que el proceso pasa del estado i al estado j por unidad de tiempo que esta en el estado i, definidas de acuerdo a:

$q_{i,i+1} = \lambda_i$ $\quad i=0,1,...,$

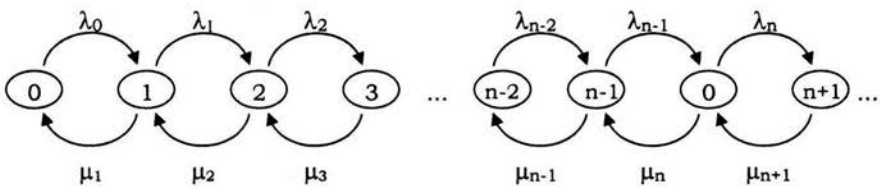
$q_{i,i-1} = \mu_i$ $\quad i=1,2,...,$

$q_{ij} = 0,$ $\quad j = i \pm 1, \quad j \neq i, \quad i=0,1,..., \quad$ y

$q_i = \lambda_i + \mu_i$ $\quad i=0,1,..., \quad \mu_0 = 0,$

Es un proceso de

a) Nacimiento puro si: $\mu_i = 0$ para i=1,2,...,
b) Muerte puro si: $\lambda_i = 0$ para i=0,1,..., y
c) Nacimiento y muerte si alguna de las λ_i y alguna de las μ_i son positivas.



Así, el proceso de nacimiento y muerte describe de manera probabilística como cambia N(t) a medida que incrementa t. Por ejemplo un tablero telefónico donde N(t) es el número de llamadas que ocurren en un intervalo de tamaño t; una epidemia donde el número de muertes que han ocurrido en (0,t) es N(t); una cola de un banco donde N(t) es el número de clientes esperando o en servicio en el tiempo t; o bien, una ciudad cuya población es N(t) en el tiempo t.

Considerando que $p_{ij}(t)$ es la función de probabilidad de transición continua del estado i al estado j (A1.2.1) y está dada por:

$$p_{ij}(t) = Pr\{X(t+u) = j \mid X(u) = i\} \qquad t>0, \quad i,j \in S$$

que es independiente de $u \geq 0$.

Se usa la ecuación Chapman-Kolmogorov hacia adelante (A1.2.11) para i,j=1,2,..., que es:

$$p'_{ij}(t) = \sum_{k \neq i} p_{ik}(t) q_{kj} + q_j p_{ij}(t)$$

y junto a

$$\begin{aligned} &q_{k,j} = 0, j = k \pm 1, && j \neq k, \\ &q_{i,i+1} = \lambda_i && i=0,1,..., \\ &q_{i,i-1} = \mu_I && i=1,2,..., \\ &q_i = \lambda_i + \mu_i && i=0,1,..., \\ &\mu_0 = 0, \end{aligned}$$

se tiene que

$$p'_{ij}(t) = \sum_{k \neq i} p_{ik}(t) q_{kj} + q_j p_{ij}(t) = q_{j-1,j}\, p_{i,j-1}(t) + q_{j+1,j}\, p_{i,j+1}(t) + q_j\, p_{i,j}(t)$$

de donde se llega al las ecuaciones Chapman-Kolmogorov hacia adelante para el proceso de nacimiento y muerte:

$$p'_{ij}(t) = \lambda_{j-1} p_{i,j-1}(t) + \mu_{j+1} p_{i,j+1}(t) - (\lambda_j + \mu_j)\, p_{i,j}(t) \qquad (1.2.1)$$

$$p'_{i0}(t) = -\lambda_0 p_{i,0}(t) + \mu_1 p_{i,1}(t) \qquad (1.2.2)$$

con i, j = 1,2,....

Las condiciones de frontera o iniciales, son:

$$p_{ij}(0+) = \delta_{ij},\ i,j = 0,1,... \qquad (1.2.3)$$

Considerando que 0+ es el tiempo t = 0, en el que el sistema sale del estado i y pasa al estado j, tiempo de salida de i a j.

Sea

$$P_j(t)=Pr\{N(t) = j\}, \qquad j=0,1,...,t>0,$$

Si el tiempo t=0, es cuando el sistema deja el estado i, se tiene

$$P_j(0)=Pr\{N(0) = j\} = \delta_{ij}, \qquad (1.2.4)$$

Entonces

$$P_j(t) = p_{ij}(t)$$

y la ecuación hacia delante se escribe como

$$P'_j(t) = -(\lambda_j + \mu_j)\, P_j(t) + \lambda_{j-1}\, P_{j-1}(t) + \mu_{j+1}\, P_{j+1}(t),\ j= 1,2,..., \qquad (1.2.5)$$

$$P'_0(t)= -\lambda_0 P_0(t)+ \mu_1 P_1(t) \qquad (1.2.6)$$

Si se supone que todas las λ_i y las μ_i son distintas de cero, se tiene que la cadena de Markov es irreducible, puesto que es una cadena no nula persistente y los límites

$$\lim_{t\to\infty} p_{ij}(t)= p_j$$

existen y son independientes del estado inicial i, entonces las ecuaciones (1.2.5) y (1.2.6) se convierten en

$$0 = -(\lambda_j + \mu_j)\, p_j + \lambda_{j-1}\, p_{j-1} + \mu_{j+1}\, p_{j+1}, \qquad j > 1, \tag{1.2.7}$$

$$0 = -\lambda_{0p} p_0 + \mu_1 p_1. \tag{1.2.8}$$

Sea

$$\pi_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j}, \qquad j \geq 1, \text{ y}$$

$$\pi_j = 1; \tag{1.2.9}$$

De donde la solución de lo anterior es obtenida por inducción. De (1.2.8) se tiene que

$$p_1 = \left(\frac{\lambda_0}{\mu_1}\right) p_0 = \pi_1 p_0$$

y si $p_k = \pi_k p_0$, $k = 1,2,\ldots,j$, se tiene de (2.2.7)

$$p_{j+1}\mu_{j+1} = \lambda_i\, \pi_j p_0, \qquad \text{o}$$

$$p_{j+1} = \pi_{j+1} p_0$$

Así, si $\sum_{k=0}^{\infty} \pi_k < \infty$, entonces

$$p_j = \frac{\pi_j}{\sum \pi_k}, \qquad j \geq 0. \tag{1.2.10}$$

Incidentalmente, $\sum_{k=0}^{\infty} \pi_k < \infty$ es una condición suficiente para que el proceso de nacimiento y muerte tenga todos sus estados no nulos persistentes.

1.3. LA FÓRMULA DE LITTLE - RELACIÓN $L = \lambda W$

Una de las relaciones más importantes que tiene la Teoría de Colas es

$$L = \lambda W \qquad (1.3.1)$$

Donde λ es la tasa media de llegadas, L es el número esperado de unidades en el sistema, y W es el tiempo promedio de espera en un estado en equilibrio. Sea el número esperado en la cola y el tiempo promedio de espera en la cola en el estado estable L_q y W_q respectivamente. Existe una relación similar

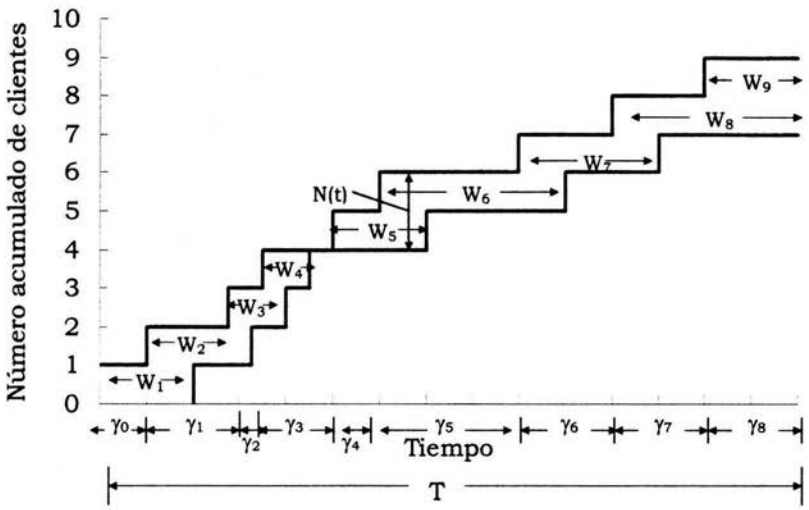
$$L_q = \lambda W_q \qquad (1.3.2)$$

Little en 1961 prueba estas fórmulas de una forma muy rigurosa, Jewell en 1967 la prueba basado en la teoría de renovación. En este momento se explicará la prueba de Eilon´s (1969) que es una prueba generalizada, simple y clara.

La prueba no depende de (i) las distribuciones de tiempo entre llegadas ni del tiempo de servicio, (ii) del número de servidores en el sistema, o (iii) de la disciplina de la cola.

Considerando la Fig. 1.3.1. La línea superior indica el número acumulado de llegadas y la línea inferior el número acumulado de salidas del sistema. La distancia vertical entre dos líneas indica el número de clientes presentes en el sistema en ese instante, mientras que la distancia horizontal denota el tiempo de espera en el sistema (tiempo de espera en la cola más el tiempo de servicio)

Figura 1.3.1



Si se supone que el sistema ha estado en operación por algún tiempo y que se a llegado a una condición de estado estable. Se considera un intervalo de tiempo T que puede o no incluir más de un periodo ocupado.

Dados

A(T) = numero total de llegadas durante el periodo T.

B(T) = área entre las dos líneas

= tiempo de espera total en el sistema de todos los cliente que llegaron durante T.

= $w_1 + w_2 + \cdots$

$\lambda(T)$ = tasa media de llegadas durante T

$$= \frac{A(T)}{T}$$

$W(T)$ = tiempo promedio de espera en el sistema durante T

$$= \frac{B(T)}{A(T)}$$

$L(T)$ = número promedio de clientes en el sistema durante T

$$= \frac{B(T)}{T}$$

Se tiene, entonces, que

$$L(T) = \frac{B(T)}{T} = \frac{B(T)}{A(T)} \frac{A(T)}{T} = W(T)\lambda(T).$$

Si se considera que el límite existe cuando $T \to \infty$ y está dado por

$$\lim_{T\to\infty} \lambda(T) = \lambda$$

y

$$\lim_{T\to\infty} W(T) = W$$

Entonces el límite para $L(T)$ cuando $T \to \infty$ también existe y está dado por

$$L = \lim_{T\to\infty} L(T)$$

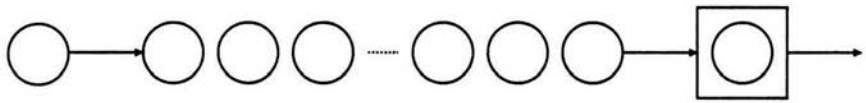
Por lo que

$$L = \lambda W$$

2. SISTEMAS DE COLAS

2.1 UNA COLA - UN SERVIDOR - POBLACIÓN INFINITA. (M/M/1):(FCFS/ ∞/ ∞).

Existen clientes que llegan a un sistema para pedir un servicio. Si el canal de servicios está vacío, la unidad entra y recibe servicio. Si hay ya una o más clientes en el canal, el cliente que llega debe tomar su lugar al final de la cola y esperar su turno para recibir el servicio. La disciplina de la cola es "primero en llegar-primero en ser servido". Existe un sólo servidor.



Persona que llega — Línea de espera — Canal de servicio

Se supone una población infinita, una línea de espera de capacidad ilimitada. El tiempo entre llegadas tiene una distribución Poisson con parámetro λ, esto es, los tiempos inter-arrivo son exponencialmente independientes con media $1/\lambda$ y los tiempos de servicio son independientes y se distribuyen exponencialmente con parámetro μ. El factor de utilización del sistema es $\rho = \lambda/\mu$, λ y μ son las tasas de llegada y servicio respectivamente.

Es necesario que $\rho = \lambda/\mu < 1$, debido a que la distribución de equilibrio no existe cuando la tasa de arribos λ es mayor o igual a la tasa de servicios μ, en este caso el largo de la cola crece sin límite.

Se asume que existe un estado estable y dado:

$$p_n = \lim_{t\to\infty} Pr\{N(t)=n\} \qquad \text{para } n=0,1,...$$

Considerando el caso de un sólo servidor, sea N(t) el número de clientes en el sistema en el tiempo t, en servicio y en la cola. El incremento N(t) en una unidad corresponde al arribo de un cliente, p_n es la parte del tiempo que el sistema está en el estado n.

Considerando el estado 0. El sistema puede salir del estado 0 sólo cuando se da una llegada. Entonces pasa del estado 0 al estado 1. La parte del tiempo que el proceso está en el estado 0 es p_0 y, como λ es la tasa de llegadas, la tasa en la que el sistema deja el estado 0 para pasar al estado 1 es λp_0. Ahora el sistema puede entrar o salir al estado 0 solamente del estado 1 por una partida o después de ser servido completamente. La parte de tiempo que el sistema está en el estado 1 es p_1 y μ es la tasa en la que se deja el estado 1 debido a un servicio completo; la tasa en la que el sistema deja el estado 1 para ir al estado 0 es μp_1. Así la tasa en la que el proceso entra al estado 0 es μp_1. Usando el principio de igualdad, se tiene

$$\lambda p_0 = \mu p_1 \qquad (2.1.1)$$

Considerando el estado 1. El sistema puede salir del estado 1 de dos formas, ya sea por una llegada o por una salida. La parte del tiempo en que el proceso está en el estado 1 es μp_1 y la tasa total en que las llegadas o salidas ocurren, es decir la tasa en que el sistema deja el estado 1 es $\lambda p_1 + \mu p_1$. El sistema puede entrar o salir del estado 1 de dos formas, por una llegada del estado 0 o por una llegada o una salida del estado 2. Así la tasa en que el sistema entra al estado 1 es $\lambda p_0 + \mu p_2$. Por el principio de igualdad se tiene que.

$$\lambda p_1 + \mu p_1 = \lambda p_0 + \mu p_2.$$

Procediendo en forma semejante, se tiene que, para n>1,

$$\lambda p_n + \mu p_n = \lambda p_{n-1} + \mu p_{n+1}. \qquad (2.1.2)$$

Las ecuaciones (2.1.1) y (2.1.2) son llamadas ecuaciones de balance, como dos tasas son balanceadas, así también relación conservadoras de flujo.

Las ecuaciones (2.1.1) y (2.1.2) junto con la relación $\sum_{n=0}^{\infty} p_n = 1$ da la solución completa para p_n. Una forma sencilla para solucionar este sistema es la siguiente. De (2.1.2) se tiene:

$$\begin{aligned} \lambda p_n - \mu p_{n+1} &= \lambda p_{n-1} - \mu p_n. \\ &= \lambda p_{n-2} - \mu p_{n-1}, \qquad \text{sustituyendo n-1 por n ,} \\ &\quad \ldots \quad \ldots \\ &= \lambda p_0 - \mu p_1, = 0, \qquad \text{por (2.1.1)} \end{aligned}$$

Así,

$$p_{n+1} = \frac{\lambda}{\mu} p_n = \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} p_{n-1}\right) = \left(\frac{\lambda}{\mu}\right)^2 p_{n-1}$$

$$= \left(\frac{\lambda}{\mu}\right)^3 p_{n-2} = \cdots = \left(\frac{\lambda}{\mu}\right)^{n+1} p_0, \qquad n = 0, 1, 2,...$$

Usando $\sum_{n=0}^{\infty} p_n = 1$, se tiene

$$p_0\left[1 + \sum_{n=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^n\right] = 1.$$

Si $\lambda/\mu = \rho < 1$, entonces:

$$p_0\left[\frac{1}{1-\rho}\right] = 1, \qquad (2.1.3)$$

es decir,

$$p_0 = 1 - \rho,$$

así que

$$p_n = (1 - \rho)\,\rho^n, \qquad n = 1, 2, \ldots .$$

La distribución es geométrica con pérdida de la memoria.

2.1.1 MEDIDAS DE EFECTIVIDAD

Debido a que N es el número de clientes, se tiene:

$$L = E\{N\} = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \rho(1-\rho)\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho(1-\rho)}{(1-\rho)^2}$$

$$L = \frac{\rho}{1-\rho} \qquad (2.1.4)$$

equivalentemente, se puede escribir:

$$L = \frac{\lambda}{\mu - \lambda} \qquad (2.1.5)$$

Si se denota por N_q al número de clientes en la cola y a su valor esperado por L_q, se tiene:

$$L_q = E\{N_q\} = 0p_0 + \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n$$

$$= L - (1 - p_0) = \frac{\rho}{1-\rho} - \rho$$

Así la media del largo de la cola es:

$$L_q = \frac{\rho^2}{1-\rho} \qquad (2.1.6)$$

o equivalentemente:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \qquad (2.1.7)$$

Se denota por L'_q el tamaño esperado de una cola no vacía, es decir, el tamaño esperado de que haya siempre cola en todos los tiempos. Se escribe:

$$L'_q = E[N_q \mid N_q \neq 0] = \sum_{n=1}^{\infty}(n-1)p'_n = \sum_{n=2}^{\infty}(n-1)p'_n$$

Donde p'_n es la distribución de probabilidad condicional de n en el sistema con cola no vacía; esto es:

$$p'_n = \Pr\{n \text{ en el sistema} \mid n \geq 2\}.$$

Y por las leyes de probabilidad condicional se tiene:

$$p'_n = \frac{\Pr\{n \text{ en el sistema y } n \geq 2\}}{\Pr\{n \geq 2\}}$$

$$p'_n = \frac{p_n}{\sum_{n=2}^{\infty} p_n} \qquad (n \geq 2)$$

$$p'_n = \frac{p_n}{1 - p_0 - p_1}$$

$$p'_n = \frac{p_n}{1 - (1-\rho) - (1-\rho)\rho}$$

$$p'_n = \frac{p_n}{\rho^2}.$$

La distribución de probabilidad de p'_n es la distribución de probabilidad normalizada de p_n cuando los casos n = 0 y n = 1 son omitidos. Así:

$$L'_q = \sum_{n=2}^{\infty}(n-1)\frac{p_n}{\rho^2} = \frac{L - p_1 - (1 - p_0 - p_1)}{\rho^2}.$$

Entonces

$$L'_q = \frac{1}{1-\rho}, \qquad (2.1.8)$$

O equivalentemente

$$L'_q = \frac{\mu}{\mu - \lambda}, \qquad (2.1.9)$$

Es útil probar que para toda n, siendo N el número de clientes en el sistema, que

$$\Pr\{N \geq n\} = \rho^n.$$

Debido a que:

$$\begin{aligned}\Pr\{N \geq n\} &= \sum_{k=n}^{\infty}(1-\rho)\rho^k \\ &= (1-\rho)\rho^n \sum_{k=n}^{\infty}\rho^{k-n} \\ &= \frac{(1-\rho)\rho^n}{1-\rho} \\ &= \rho^n.\end{aligned}$$

Para encontrar el tiempo promedio de espera en el sistema, W, se recurre a (1.3.1), $L = \lambda W$, se obtiene la esperanza y resulta:

$$W = E\{W\}$$

$$= \frac{E\{N\}}{\lambda}$$

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

$$= \frac{1}{\mu(1-\rho)}. \qquad (2.1.10)$$

$$= \frac{1}{\mu - \lambda}. \qquad (2.1.11)$$

Por último para obtener el tiempo promedio de espera en la cola antes de recibir el servicio, W_q, se tiene que tener en cuenta la disciplina de la cola que en este caso es "primero en llegar – primero en ser servido". La variable aleatoria tiene una propiedad interesante ya que es parte discreta y parte continua. El tiempo de espera en su mayoría es una variable aleatoria continua exceptuando que existe la probabilidad de que el retraso sea 0, esto es, que el cliente sea servido inmediatamente después de su llegada.

Sea T_q la variable aleatoria designada para el tiempo de espera en la cola. Entonces:

$$W_q(0) = \Pr\{T_q \leq 0\}$$
$$= \Pr\{W_q = 0\}$$
$$= \Pr\{\text{el sistema esté vacío al ocurrir una llegada}\}$$
$$\equiv q_0.$$

Donde q_n es la probabilidad condicional de que una llegada esté a punto de ocurrir. Esta probabilidad no siempre es igual P_n. Sin embargo si la entrada es Poisson, $q_n = P_n$. Entonces:

$$W_q(0) = P_0 = 1 - \rho. \qquad (2.1.12)$$

Considerando que $W_q(t)$, la probabilidad de que un cliente espere un tiempo menor o igual que t para ser servido. Si hay n clientes que llegaron antes para ser servidos en un tiempo entre 0 y t, todos los clientes deben ser servidos dentro del tiempo t. Debido a que la distribución de servicio tiene pérdida de la memoria, la distribución del tiempo requerido para atender a n clientes es independiente del tiempo en que ocurren las llegadas y es la convolución de n variables aleatorias exponenciales, que es de tipo Erlang-n.

Como las entradas son Poisson (Apéndice 2), los puntos en que éstas ocurren están espaciados uniformemente y por lo tanto la probabilidad que una llegada encuentre n en el sistema es idéntica a la distribución de probabilidad estacionaria del tamaño del sistema. Aclarando:

$$W_q(t) = \Pr\{T_q \leq t\}$$

$$= \sum_{n=1}^{\infty} \left[\Pr\{\text{n servidos} \leq t \mid \text{al llegar existan n en el sistema}\} * P_n\right]$$
$$+ W_q(0)$$

$$= (1-\rho)\sum_{n=1}^{\infty}\rho^n \int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!}e^{-\mu x}dx + (1-\rho)$$

$$= (1-\rho)\rho\int_0^t \mu e^{-\mu x}\sum_{n=1}^{\infty}\frac{(\mu x\rho)^{n-1}}{(n-1)!}dx + (1-\rho)$$

$$= (1-\rho)\rho\int_0^t \mu e^{-\mu x(1-\rho)}dx + (1-\rho)$$

$$= (1-\rho)e^{-\mu(1-\rho)t} \qquad (t>0).$$

Así la distribución del tiempo de espera en la cola es

$$W_q(t) = \begin{cases} 1-\rho & (t=0) \\ (1-\rho)e^{-\mu(1-p)t} & (t>0) \end{cases} \qquad (2.1.13)$$

Entonces el tiempo promedio de espera en la cola es:

$$\begin{aligned} W_q &= E[T_q] \\ &= \int_0^{\infty} t W_q(t)dt \\ &= 0\left(1-\frac{\lambda}{\mu}\right) + \int_0^{\infty} t\frac{\lambda}{\mu}(\mu-\lambda)e^{-(\mu-\lambda)t}dt \\ &= \frac{\lambda}{\mu}\int_0^{\infty} t(\mu-\lambda)e^{-(\mu-\lambda)t}dt. \end{aligned}$$

De donde:

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} \qquad (2.1.14)$$

Con esta expresión se concluye que si un individuo debe esperar en promedio W_q unidades de tiempo antes de recibir un servicio que dura, en promedio, $1/\mu$ unidades de tiempo, entonces W, el tiempo total de espera en el sistema, es:

$$W = W_q + \frac{1}{\mu} \qquad (2.1.15)$$

2.2 UNA COLA - C SERVIDORES - POBLACIÓN INFINITA.

(M/M/c) : (FCFS/∞/∞).

Ahora se vera un modelo de servidores múltiples en el que cada servidor es independiente y tienen idéntico tiempo de servicio, cuya distribución es exponencial. El proceso de llegadas se asume Poisson.

Se considera que hay c servidores Si hay más de c clientes en el sistema, todos los c servidores están ocupados y cada uno atiende con una tasa media μ; la tasa media de salida del sistema es $c\mu$. Cuando hay menos de c clientes en el sistema, $n < c$, solamente n de los c servidores están ocupados y el sistema tiene una tasa media de salida de $n\mu$. Esto es

$$\mu_n = \begin{cases} n\mu & (1 \le n < c) \\ c\mu & (n \ge c) \end{cases} \qquad (2.2.1)$$

Regresando al modelo de nacimiento y muerte de las ecuaciones (1.2.7) y (1.2.8) se llega a:

$$p_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}} p_n - \frac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1} \qquad (n>1)$$

$$p_1 = \frac{\lambda_0}{\mu_1} p_0 \qquad (2.2.2)$$

y

$$p_2 = \frac{\lambda_1 + \mu_1}{\mu_2} p_1 - \frac{\lambda_0}{\mu_2} p_0$$

$$= \frac{\lambda_1 + \mu_1}{\mu_2}\frac{\lambda_0}{\mu_1}p_0 - \frac{\lambda_0}{\mu_2}p_0$$

$$= \frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0,$$

$$p_3 = \frac{\lambda_2 + \mu_2}{\mu_3}p_2 - \frac{\lambda_1}{\mu_3}p_1$$

$$= \frac{\lambda_2 + \mu_2}{\mu_3}\frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0 - \frac{\lambda_1}{\mu_3}\frac{\lambda_0}{\mu_1}p_0$$

$$= \frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}p_0,$$

...

El patrón, si se recuerda (1.2.9), nos lleva a

$$p_n = \frac{\lambda_0\lambda_1\cdots\lambda_{j-1}}{\mu_1\mu_2\cdots\mu_j}p_0$$

$$= \prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i}p_0$$

$$= \pi_n p_0 \qquad (n \geq 1) \qquad (2.2.3)$$

Utilizando (2.2.1) y (2.2.3) y con el hecho de que $\lambda n = \lambda$ para toda n, se obtiene

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n}p_0 & (1 \leq n < c) \\ \dfrac{\lambda^n}{c^{n-c}c!\mu^n}p_0 & n \geq c \end{cases} \qquad (2.2.4)$$

Para encontrar p_0, se utiliza la condición límite

$$\sum_{n=0}^{\infty} p_n(t) = 1,$$

Sustituyendo

$$p_0\left[\sum_{n=0}^{c-1}\frac{\lambda^n}{n!\mu^n}+\sum_{n=c}^{\infty}\frac{\lambda^n}{c^{n-c}c!\mu^n}\right]=1.$$

Sea $r = \lambda/\mu$ y $\rho = r/c = \lambda/c\mu$. Entonces

$$p_0\left[\sum_{n=0}^{c-1}\frac{r^n}{n!}+\sum_{n=c}^{\infty}\frac{r^n}{c^{n-c}c!}\right]=1.$$

Considerando la serie

$$\sum_{n=c}^{\infty}\frac{r^n}{c^{n-c}c!} = \frac{r^c}{c!}\sum_{n=c}^{\infty}\left(\frac{r}{c}\right)^{n-c}$$

$$= \frac{r^c}{c!}\sum_{n=c}^{\infty}\left(\frac{r}{c}\right)^{m}$$

$$= \frac{r^c}{c!}\frac{1}{(1-r/c)}$$

$$= \frac{r^c}{c!}\sum_{n=c}^{\infty}\left(\frac{r}{c}\right)^{n-c} \qquad (r/c = \rho_c < 1). \qquad (2.2.5)$$

Entonces

$$p_0 = \left[\sum_{n=0}^{c-1}\frac{r^n}{n!} + \frac{cr^c}{c!(c-r)}\right]^{-1}$$

$$p_0 = \left[\sum_{n=0}^{c-1}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c\left(\frac{c\mu}{c\mu-\lambda}\right)\right]^{-1} \qquad (r/c=\rho_c<1). \qquad (2.2.6)$$

2.2.1 MEDIDAS DE EFECTIVIDAD

Considerando L_q el largo esperado de la cola, las ecuaciones (2.2.4) y (2.2.6), por definición se tiene:

$$\begin{aligned}
L_q &= \sum_{n=c}^{\infty}(n-c)p_n \\
&= \sum_{n=c}^{\infty}(n-c)\frac{r^n}{c^{n-c}c!}p_0 \\
&= \frac{r^c p_0}{c!}\sum_{n=c}^{\infty}(n-c)\rho_c{}^{n-c} \\
&= \frac{r^c p_0}{c!}\sum_{m=1}^{\infty}m\rho_c{}^{m} \\
&= \frac{r^c}{c!}\rho_c p_0\sum_{m=1}^{\infty}m\rho_c{}^{m-1} \\
&= \frac{r^c}{c!}\rho_c p_0\frac{d}{d\rho}\left[\sum_{m=1}^{\infty}\rho_c{}^{m}\right] \\
&= \frac{r^c}{c!}\rho_c p_0\frac{d}{d\rho}\left[\rho_c\frac{1}{1-\rho_c}\right].
\end{aligned}$$

Así

$$L_q = \frac{r^{c+1}/c}{c!(1-r/c)^2} p_0$$

$$L_q = \left[\frac{(\lambda/\mu)^c \lambda\mu}{(c-1)!(c\mu-\lambda)^2}\right] p_0 \qquad (2.2.7)$$

Para encontrar el tiempo promedio de espera en la cola W_q se utiliza (2.2.7), entonces:

$$W_q = \frac{L_q}{\lambda} = \left[\frac{(\lambda/\mu)^c \mu}{(c-1)!(c\mu-\lambda)^2}\right] p_0 \qquad (2.2.8)$$

Usando la relación entre W y W_q (2.1.15) se tiene el tiempo promedio de espera en el sistema W

$$W = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^c \mu}{(c-1)!(c\mu-\lambda)^2}\right] p_0, \qquad (2.2.9)$$

Finalmente para encontrar el número esperado de clientes en el sistema L se tiene

$$L = \lambda W = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^c \lambda\mu}{(c-1)!(c\mu-\lambda)^2}\right] p_0$$

Sea T_q, la variable aleatoria del "tiempo gastado esperando en la cola" y considerando $W_q(t)$ su función de densidad

$$W_q(0) = \Pr\{T_q = 0\} = \Pr\{\text{c} - 1 \text{ o menos en el sistema}\} = \sum_{n=0}^{c-1} p_n = p_0 \sum_{n=1}^{c-1} \frac{r^n}{n!}.$$

Usando la ecuación (2.2.6) y con

$$\sum_{n=1}^{c-1} \frac{r^n}{n!} = \frac{1}{p_0} - \frac{cr^c}{c!(c-r)}$$

se tiene

$$W_q(0) = p_0\left[\frac{1}{p_0} - \frac{cr^c}{c!(c-r)}\right]$$

$$= 1 - \frac{cr^c p_0}{c!(c-r)}$$

$$= 1 - \frac{c(\lambda/\mu)^c}{c!(c-\lambda/\mu)} p_0 .$$

Para $T_q > 0$ y asumiendo la política de primero en llegar, primero en ser servido se llega a

$W_q(t) = Pr\{T_q \leq t\}$

$$= \sum_{n=c}^{\infty}[Pr\{\ n - c + 1 \text{ servidos } \leq t \mid \text{al llegar se encuentran n en el sistema}\} * p_n] + W_q(0) \qquad (t>0).$$

Si $n \geq c$, la tasa promedio de salida es Poisson con media $c\mu$, entonces el tiempo entre servicios sucesivos es exponencial con media $1/(c\mu)$, y la distribución del tiempo para los n–c–1 servicios es Erlang tipo $(n - c - 1)$.

Entonces

$$W_q(t) = \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c}c!} \int_0^t \frac{\mu c(\mu c x)^{n-c}}{(n-c)!} e^{-\mu c x} dx + W_q(0) \qquad (t > 0)$$

$$= p_0 \frac{r^c}{(c-1)!} \int_0^t \mu e^{-\mu c x} \sum_{n=c}^{\infty} \frac{(\mu r x)^{n-c}}{(n-c)!} dx + W_q(0)$$

$$= p_0 \frac{r^c}{(c-1)!} \int_0^t \mu e^{-\mu c x} e^{\mu r x} dx + W_q(0)$$

$$= p_0 \frac{r^c}{(c-1)!} \int_0^t \mu e^{-\mu x(c-r)} dx + W_q(0)$$

$$= \frac{r^c \left(1 - e^{-\mu(c-r)t}\right)}{(c-1)!(c-r)} p_0 + W_q(0)$$

$$= \frac{(\lambda/\mu)^c \left(1 - e^{-(\mu c-\lambda)t}\right)}{(c-1)!(c-\lambda/\mu)} p_0 + W_q(0) \qquad (t > 0)$$

Resumiendo

$$W_q(t) = \begin{cases} 1 - \dfrac{c(\lambda/\mu)^c}{c!(c-\lambda/\mu)} p_0 & (t = 0) \\ \dfrac{(\lambda/\mu)^c \left(1 - e^{-(\mu c-\lambda)t}\right)}{(c-1)!(c-\lambda/\mu)} p_0 + W_q(0) & (t > 0) \end{cases} \qquad (2.2.10)$$

2.3 UNA COLA - C SERVIDORES EN PARALELO - POBLACIÓN FINITA. (M/M/c):(FCFS/K/∞), c<K.

Considerando un modelo de servidores en paralelo en el que en el sistema tiene una capacidad limitada. Utilizando la Ecuación (2.2.3) junto con

$$\lambda_n = \begin{cases} \lambda & (0 \le n < K) \\ 0 & (n \ge K) \end{cases}$$

y

$$\mu_n = \begin{cases} n\mu & (0 \le n < c) \\ c\mu & (c \le n \le K) \end{cases}$$

se tiene

$$p_n = \begin{cases} \dfrac{\lambda^n}{n\mu(n-1)\mu\cdots(1)\mu} p_0 & (0 \le n < c) \\ \dfrac{\lambda^n}{\underbrace{c\mu c\mu\cdots c\mu c\mu}_{n-c \text{ tér min os}}(c-1)\mu(c-2)\mu\cdots(1)\mu} p_0 & (c \le n \le K) \end{cases}$$

o de forma equivalente

$$p_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n p_0 & (0 \le n < c) \\ \dfrac{1}{c^{n-c}c!}\left(\dfrac{\lambda}{\mu}\right)^n p_0 & (c \le n \le K) \end{cases} \qquad (2.3.1)$$

Utilizando la condición límite $\sum_{n=0}^{K} p_n = 1$ se obtiene p_0, esto es,

$$\sum_{n=0}^{K} p_n = \sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 + \sum_{n=c}^{K} \frac{1}{c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n p_0 = 1;$$

entonces

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{K} \frac{1}{c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}.$$

Si se considera $r = \lambda/\mu$, se tiene que:

$$\sum_{n=c}^{K} \frac{1}{c^{n-c}c!} r^n = \frac{r^c}{c!}\sum_{n=c}^{K}\left(\frac{r}{c}\right)^{n-c} = \begin{cases} \dfrac{r^c}{c!}\dfrac{1-\rho_c^{K-c+1}}{1-\rho_c} & \rho_c = r/c = \lambda/c\mu \neq 1 \\ \dfrac{r^c}{c!}(K-c+1) & \rho_c = 1 \end{cases}$$

$$p_0 = \begin{cases} \left[\displaystyle\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{(\lambda/\mu)^c}{c!}\frac{1-(\lambda/c\mu)^{K-c+1}}{1-\lambda/c\mu}\right]^{-1} & \lambda/c\mu \neq 1 \\ \left[\displaystyle\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{(\lambda/\mu)^c}{c!}(K-c+1)\right]^{-1} & \lambda/c\mu = 1 \end{cases} \qquad (2.3.2)$$

2.3.1 MEDIDAS DE EFECTIVIDAD

El largo esperado de la cola y el tamaño del sistema se obtienen de:

$$L_q = \sum_{n=c}^{K} (n-c)p_n$$

$$= \frac{p_0}{c!} \sum_{n=c}^{K} \frac{(n-c)r^n}{c^{n-c}}$$

$$= \frac{p_0 (c\rho_c)^c \rho_c}{c!} \sum_{n=c}^{K} (n-c)\rho_c^{\,n-c-1}$$

$$= \frac{p_0 (c\rho_c)^c \rho_c}{c!} \sum_{i=1}^{K-c} i\rho_c^{\,i-1} = \frac{p_0 (c\rho_c)^c \rho_c}{c!} \frac{d}{d\rho_c}\left[\frac{1-\rho_c^{\,K-c+1}}{1-\rho_c}\right]$$

$$= \frac{p_0 (c\rho_c)^c \rho_c}{c!} \frac{d}{d\rho_c}\left[\frac{1-\rho_c^{\,K-c+1}}{1-\rho_c}\right],$$

o

$$L_q = \frac{p_0 (c\rho_c)^c \rho_c}{c!(1-\rho_c)^2}\left[1-\rho_c^{\,K-c+1} - (1-\rho_c)(K-c+1)\rho_c^{\,K-c}\right]. \qquad (2.3.3)$$

Para $\rho = 1$, es necesario emplear la doble regla de L'Hôpital.

Para obtener el tamaño esperado del sistema se recurre a:

$$L_q = \sum_{n=c}^{K} (n-c)p_n$$

$$= \sum_{n=c}^{K} np_n - c\sum_{n=c}^{K} p_n$$

$$= \sum_{n=0}^{K} np_n - \sum_{n=0}^{c-1} np_n - c\sum_{n=c}^{K} p_n$$

$$= L - \sum_{n=0}^{c-1} np_n - c\left(1 - \sum_{n=0}^{c-1} p_n\right) = L - \sum_{n=0}^{c-1} (n-c)p_n - c$$

$$= L - \sum_{n=0}^{c-1} (n-c)p_n - c$$

Así

$$L = L_q + c - \sum_{n=0}^{c-1} (c-n)p_n ,$$

O

$$L = L_q + c - p_0 \sum_{n=0}^{c-1} \frac{(c-n)(\rho_c c)^n}{n!} \qquad (2.3.4)$$

Los valores esperados para los tiempos de espera se obtienen usando la formula de Little, es decir,

$$W = \frac{L}{\lambda'} , \qquad \lambda' = \lambda(1-p_K) \qquad (2.3.5)$$

y

$$W_q = W - \frac{1}{\mu} \tag{2.3.6}$$

o

$$W_q = \frac{L_q}{\lambda'} \tag{2.3.7}$$

Donde λ' es la tasa media actual de clientes entrando al sistema. Cuando c es grande, es más fácil calcular (2.3.3), (2.3.7), (2.3.6) y (2.3.5) de forma recurrente para encontrar L que usar (2.3.4).

2.4 UNA COLA - UN SERVIDOR - POBLACIÓN FINITA. (M/M/1):(FCFS/K/∞).

El análisis del sistema de la sección (2.1) supone que el número de clientes que requieren servicio en un periodo de tiempo determinado es infinito. Este caso no corresponde a la realidad ya que una población es, por regla, de tamaño finito. Así, una cola finita, tiene un número de clientes en el sistema no mayor a un número especificado (denotado por K). A cualquier cliente que llega cuando la cola está "llena", se le evita la entrada al sistema y, por tanto, sale para siempre. Desde el punto de vista del proceso de nacimiento y muerte, la tasa media de entrada al sistema se vuelve cero en estos instantes. Entonces la única modificación necesaria en el modelo básico para introducir una cola finita, es cambiar los parámetros λ_n a

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

Como $\lambda_n = 0$ para algunos valores de n, un sistema de colas que se ajusta a este modelo, finalmente alcanzará una condición de estado estacionario.

Usando el principio de igualdad de tasas, la ecuación de balance se escribe de la siguiente forma:

$$\lambda p_0 = \mu p_1$$

$$\lambda p_n + \mu p_n = \lambda p_{n-1} + \mu p_{n+1}, \qquad n = 1, 2,\ldots, K-1 \qquad (2.4.1)$$

$$\lambda p_{K-1} = \mu p_K.$$

Si se resuelven las dos primeras ecuaciones por recursividad, se tiene que

$$p_n = p_0\rho^n, \qquad n = 0, 1, 2,..., K-1.$$

Usando la última ecuación se llega a

$$p_K = \rho p_{K-1} = \rho(p_0\rho^{K-1}) = p_0\rho^K .$$

Así

$$p_n = p_0\rho^n, \qquad n = 0, 1, 2,..., K. \tag{2.4.2}$$

Usando la condición de normalidad $\sum_{n=0}^{K} p_n = 1$, se tiene

$$p_0 \sum_{n=0}^{K} \rho^n = 1$$

entonces

$$p_0 = \begin{cases} \left[\sum_{n=0}^{K} \rho^n\right]^{-1} = \dfrac{1-\rho}{1-\rho^{K+1}} & \rho \neq 1 \\ \dfrac{1}{K+1} & \rho = 1 \end{cases} \tag{2.4.3}$$

Así, para n = 0, 1,..., K, se tiene

$$p_n = \begin{cases} p_0\rho^n = \dfrac{(1-\rho)\rho^n}{1-\rho^{K+1}} & \rho \neq 1 \\ \dfrac{1}{K+1} & \rho = 1 \end{cases} \qquad (2.4.4)$$

Su función de probabilidad está dada por

$$G(s) = \sum_{n=0}^{K} p_n s^n = \frac{1-\rho}{1-\rho^{K+1}}\left[\frac{1-(\rho s)^{K+1}}{1-\rho s}\right] \qquad \rho \neq 1 \qquad (2.4.5)$$

La distribución del número en el sistema es uniforme para $\rho = 1$ y geométrica truncada para $\rho \neq 1$.

2.4.1 MEDIDAS DE EFECTIVIDAD

El número esperado de clientes en el sistema

para $\rho = 1$

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

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

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

Para $\rho \neq 1$

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

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

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

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

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

$$= \frac{\rho-K\rho^{K+1}-\rho^{K+1}+K\rho^{K+2}-\rho^{2}+K\rho^{K+2}+\rho^{K+2}-K\rho^{K+3}}{(1-\rho)^{2}\left(1-\rho^{K+1}\right)}$$

$$= \frac{\rho-\rho^{2}+\rho^{K+2}+\rho^{K+3}-\rho^{K+3}-2\rho^{K+2}+2\rho^{K+2}-K\rho^{K+1}-\rho^{K+1}+2K\rho^{K+2}-K\rho^{K+3}}{(1-\rho)^{2}\left(1-\rho^{K+1}\right)}$$

$$= \frac{\rho-\rho^{2}-\rho^{K+2}+\rho^{K+3}}{(1-\rho)^{2}\left(1-\rho^{K+1}\right)} + \frac{\left(-K-1+2K\rho+2\rho-K\rho^{2}-\rho^{2}\right)\rho^{K+1}}{(1-\rho)^{2}\left(1-\rho^{K+1}\right)}$$

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

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

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

Así

$$L = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \qquad (2.4.6)$$

El largo esperado de la cola se obtiene de (2.3.4) para c = 1 y de (2.4.3)

$$L = L_q + c - p_0 \sum_{n=0}^{c-1} \frac{(c-n)(\rho c)^n}{n!}$$

$$L_q = L - c + p_0 \sum_{n=0}^{c-1} \frac{(c-n)(\rho c)^n}{n!}$$

$$L_q = L - 1 + p_0$$

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

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

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

$$L_q = L - \frac{\rho\left(1-\rho^K\right)}{1-\rho^{K+1}} \qquad (2.4.7)$$

Para el tiempo promedio de espera en la cola se considera (2.3.7) y (2.4.2)

$$W_q = \frac{L_q}{\lambda'} \qquad \lambda' = \lambda(1-p_K)$$

$$W_q = \frac{L_q}{\lambda\left(1-p_K\right)}$$

$$W_q = \frac{L_q}{\lambda\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{\frac{\rho}{\rho}L_q\left(\frac{1-\rho^K}{1-\rho^K}\right)}{\lambda\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{\frac{\lambda}{\mu}L_q\left(\frac{1-\rho^K}{\left(1-\rho^K\right)\rho}\right)}{\lambda\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q\left(\frac{1-\rho^K}{\rho-\rho^{K+1}+1-1}\right)}{\mu\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q\left(\frac{1-\rho^K+\rho^{K+1}-\rho^{K+1}}{1-\rho^{K+1}-(1-\rho)}\right)}{\mu\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q\left(\frac{1-\rho^{K+1}-\left(\rho^K-\rho^{K+1}\right)}{1-\rho^{K+1}}\right)}{\mu\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q\left(\frac{1-\left(\frac{\rho^K-\rho^{K+1}}{1-\rho^{K+1}}\right)}{1-\left(\frac{1-\rho}{1-\rho^{K+1}}\right)}\right)}{\mu\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q\left(\frac{1-\left(\frac{(1-\rho)\rho^K}{1-\rho^{K+1}}\right)}{1-p_0}\right)}{\mu\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q\left(\frac{1-p_0\rho^K}{1-p_0}\right)}{\mu\left(1-p_0\rho^K\right)}$$

$$W_q = \frac{L_q}{\mu(1-p_0)} \qquad (2.4.8)$$

Y el tiempo total de espera en el sistema se tiene de (2.3.6)

$$W = W_q + \frac{1}{\mu} \qquad (2.4.9)$$

2.5 PROCESO DE DECISIÓN.

Como se menciono en el Capítulo 1, es necesario tomar ciertas decisiones en cuestión de determinar el nivel de servicio que minimiza el costo esperado de servicio y el costo esperado de la espera.

El proceso de decisión en las líneas de espera consiste en fijar el nivel de los parámetros:

a) Número requerido de servidores en cada unidad de servicio,

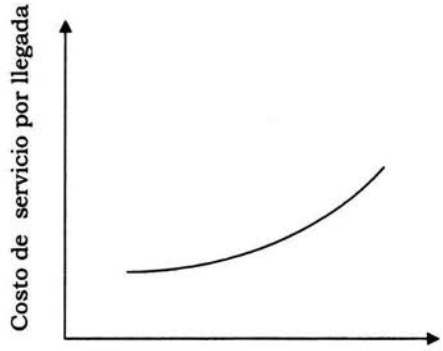
b) Número requerido de unidades de servicio,

c) Eficiencia (rapidez) del servicio.

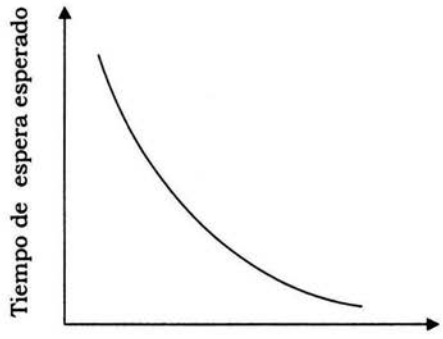
Un ejemplo del punto "a)" es cuando es necesario determinar el tamaño de la brigada de mantenimiento (donde la brigada completa es el servidor), como ejemplos del inciso "b)" están el número de cajeros automáticos a instalar dentro de cierto perímetro o el número de almacenes diferentes y, por ultimo, para "c)" se tiene que es necesario determinar la eficiencia de los servidores cuando se selecciona el tipo de equipo de manejo de materiales (servidores) que se debe comprar para transportar ciertos tipos de carga (clientes).

Para lograr un equilibrio entre el costo de operación del sistema y el costo asociado a la espera, es necesario decidir sobre el nivel de los parámetros mencionados, los cuales se pueden fijar individualmente o combinándolos unos con otros.

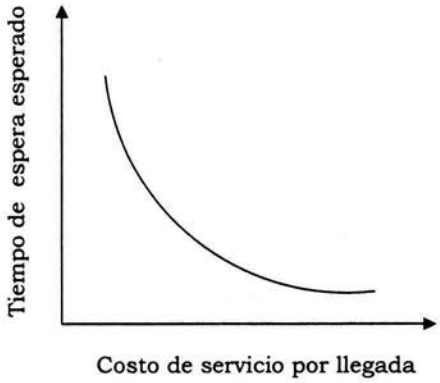
Así se tiene que costo de servicio está en función del nivel de servicio, representativamente es:



Y el tiempo de espera esperado también está en función del nivel de servicio, como se sugiere:



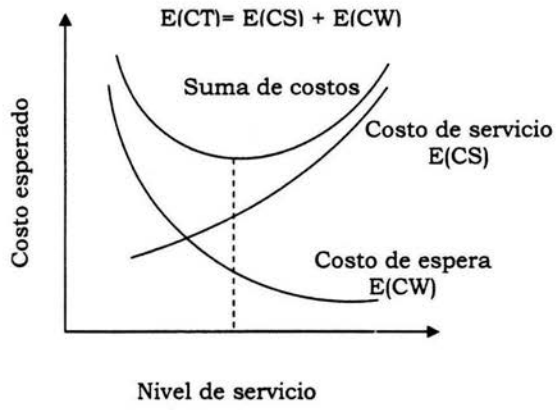
Como se puede observar estos dos factores crean presiones opuestas en quien toma las decisiones, por lo que el objetivo de reducir los costos de servicio sugiere un nivel mínimo del mismo, en oposición a que, para evitar los tiempos de espera largos, es necesario un nivel alto de servicio. Así es necesario encontrar el equilibrio que se puede obtener al relacionar ambas decisiones, como se muestra:



Considerando que se evalúa en forma explicita el costo de espera y tomando en cuenta que el objetivo es minimizar el costo esperado del servicio y el costo esperado de la espera, se tiene la representación matemática del objetivo:

$$\text{Minimizar } E(CT) = E(CS) + E(CW)$$

Donde gráficamente es:



Es decir, el problema se reduce al seleccionar el punto sobre la curva que dé el mejor balance promedio al solicitar un servicio y el costo de proporcionarlo. Para obtener tal balance es necesario responder a preguntas tales como: ¿en qué medida el gasto en el servicio es equivalente (respecto a su efecto negativo) al retraso de un cliente durante una unidad de tiempo?, para lo que hay que adoptar (explicita o implícitamente) una medida común de su impacto. La elección de tal medida es el costo para lo que hay que estimar el correspondiente de espera.

Debido a la gran diversidad de situaciones de líneas de espera, es necesario aplicar el proceso de estimación del costo de espera dependiendo de cada modelo. Tal costo puede darse: en organizaciones lucrativas donde dar un mal servicio implique pérdida de ganancias o clientes perdidos, en servicios no lucrativos que el costo se más bien social o en procesos internos como son los efectuados en la producción de donde el tiempo perdido por retraso en manufactura o maquinaria genere costos en la productividad de la operación.

La forma pasa representar la función de costo espera depende del problema individual. Así que dada la característica de una línea de espera, se debe hablar de valores esperados de costo y no del costo en sí.

Si se considera el caso en que los clientes son internos (por ejemplo producción) la propiedad fundamental del sistema de colas que determina la tasa actual a la que se incurre en los costos de espera es N, el número de clientes en el sistema. Entonces pensando en la función de N, g (N), se construye para la situación en particular estimando g(n), la tasa del costo de espera cuando N = n, con n = 1,2,..., donde g(0) = 0 y calculando las p_n probabilidades de que existan n personas en el sistema en un periodo de tiempo determinado, se calcula el valor esperado de costo de demora, E(g(n)), para el caso discreto, esto es:

$$E(CW) = E(g(N)) = \sum_{n=0}^{\infty} g(n)p_n$$

Si se tiene el caso especial en el que g(N) es una función lineal (la tasa de costo espera es proporcional a N), se tiene

$$g(N) = C_wN$$

donde C_w es el costo de espera por unidad de tiempo para cada cliente. Y, entonces

$$E(CW) = C_w \sum_{n=0}^{\infty} np_n = C_wL$$

Considerando la situación en que los clientes son externos al sistema de colas como sucede en las organizaciones comerciales, de servicio de

transporte y sociales, la magnitud de sus costos tiende a quedar muy afectada por el tiempo que permanecen los clientes. Así la propiedad fundamental del sistema de colas que determina el costo de espera en el que se incurre es W.

Sea h(W) la función de costo espera para estas situaciones, es muy común que no sea lineal. Una manera de construirla es estimar h(w), el costo en el que se incurre cuando un cliente espera un tiempo W = w, para distintos valores de w y después ajustar un polinomio a estos puntos.

La esperanza de esta función de una variable aleatoria continua se define por:

$$E(h(W)) = \int_0^{\infty} h(w) f_W(w) d_w$$

En donde $f_W(w)$ es la función de densidad de probabilidad de W. Sin embargo como E(h(W)) es el costo de espera esperado por cliente y E(CW) el costo de espera esperado por unidad de tiempo, es decir, las cantidades no son los iguales para relacionarlas es necesario multiplicar por λ la tasa media de llegadas

$$E(CW) = \lambda E(h(W)) = \lambda \int_0^{\infty} h(w) f_W(w) d_w$$

3. EJEMPLOS DE SISTEMAS DE COLAS

3.1 REVALIDACIÓN DE INSCRIPCIÓN EN SERVICIOS ESCOLARES

Como ejemplo del modelo de una cola - un servidor – población infinita, se considera el proceso de revalidación de la inscripción en la Maestría de Ingeniería de Sistemas en servicios escolares, donde una sola persona recibe la documentación de los alumnos que forman una línea de espera.

Las llegadas de los alumnos se distribuyen de forma Poisson con λ igual a 35 alumnos por hora, mientras que el número de servicios se distribuye exponencialmente con parámetro μ igual a 40 servicios por hora.

Se tiene presente que no existe prioridad en el servicio y se considera que el mecanismo del servicio es "primero en llegar primero en ser atendido".

Para encontrar L, el número esperado de alumnos en el sistema se recurre a (2.1.5)

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

$$= \frac{35}{40 - 35}$$

$$= 7 \text{ alumnos}$$

El largo promedio de la cola, L_q, está dado por (2.1.7)

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

$$= \frac{35^2}{40(40 - 35)}$$

$$= \frac{1225}{40(5)}$$

$$= 6.125,$$

esto es,

$$L_q = 6 \text{ alumnos}$$

El tiempo promedio de espera en el sistema, W, se obtiene de (2.1.11)

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

$$= \frac{1}{40 - 35}$$

$$= \frac{1}{5}$$

$$= 0.2 \text{ de hora},$$

Es decir 12 minutos.

El tiempo promedio de espera en la cola se encuentra con (2.1.15)

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

$$= 0.2 - \frac{1}{40}$$

$$= 0.2 - 0.025$$

$$= 0.175 \text{ de hora}$$

Esto es aproximadamente 10 minutos 30 segundos.

Al llegar a este sistema en promedio se encuentran 7 personas en él, se espera en una cola de 6 personas durante 10 minutos y medio y se recibe el servicio en minuto y medio más.

En un sistema como este tener esperando en cola 6 personas es aceptable puesto que 10 minutos en la cola no es excesivo si pensamos en el número de gente que esta delante de nosotros y el trámite a realizar, el tiempo de servicio es relativamente pequeño. Aun así, si se deseara tener a menos estudiantes esperando o disminuir el tiempo en la fila lo que queda por hacer es aumentar el número de servidores.

3.2 CLÍNICA OFTALMOLÓGICA

Una clínica de oftalmología ofrece gratis un examen de glaucoma cada miércoles. Existen 3 oftalmólogos en servicio. El examen de glaucoma se realiza en promedio en 20 minutos y se distribuye exponencialmente. Los clientes llegan de acuerdo a un Proceso Poisson con una media de 6 pacientes por hora, bajo la base de primero en llegar - primero en ser servido.

Primero se encuentra p_0 de la ecuación (2.2.6) ya que es un factor que aparece en las medidas de efectividad.

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \left(\frac{c\mu}{c\mu - \lambda}\right)\right]^{-1}$$

$$= \left[\sum_{n=0}^{2} \frac{1}{n!}\left(\frac{6}{3}\right)^n + \frac{1}{3!}\left(\frac{6}{3}\right)^3 \left(\frac{(3)(3)}{(3)(3) - 6}\right)\right]^{-1}$$

$$= \left[1 + 2 + 2 + \frac{2^3(9)}{3!(3)}\right]^{-1}$$

$$= \left[1 + 2 + 2 + \frac{2^3(9)}{3!(3)}\right]^{-1}$$

$$= \frac{1}{9}$$

Para encontrar, L_q, el largo esperado de la cola se considera (2.2.7)

$$L_q = \left[\frac{(\lambda/\mu)^c \lambda\mu}{(c-1)!(c\mu-\lambda)^2}\right] p_0$$

$$= \left[\frac{(6/3)^3(6)(3)}{(3-1)!((3)(3)-6)^2}\right]\left(\frac{1}{9}\right)$$

$$= \left[\frac{2^3(18)}{2!(3^2)}\right]\left(\frac{1}{9}\right)$$

$$= \frac{8}{9}$$

$$= 0.889$$

Para encontrar el tiempo promedio de espera en la cola W_q se utiliza (2.2.8):

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

$$= \frac{0.889}{6}$$

$$= 0.0988 \text{ de hora}$$

Es decir 5 minutos 55 segundos

El tiempo promedio de espera en el sistema W se obtiene de la ecuación (2.2.9)

$$W = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^c \mu}{(c-1)!(c\mu-\lambda)^2}\right] p_0$$

$$= \frac{1}{3} + \left[\frac{(6/3)^3 3}{(3-1)!((3)(3)-6)^2}\right]\frac{1}{9}$$

$$= \frac{1}{3} + \left[\frac{(2)^3 3}{(2!)(3)^2}\right]\frac{1}{9}$$

$$= \frac{1}{3} + \left[\frac{24}{18}\right]\frac{1}{9}$$

$$= \frac{1}{3} + \frac{4}{27}$$

$$= \frac{13}{27}$$

$$= 0.4815$$

Aproximadamente 28 minutos 53 segundos

Finalmente para encontrar el número esperado de clientes en el sistema L se utiliza la Formula de Little

$$L = \lambda W$$

$$= 6(0.4815)$$

$$= 2.889 \text{ clientes}$$

Este sistema muestra lo que es un trabajo organizado, ya que a pesar de que pareciera que 6 clientes por hora como llegada y un tiempo de servicio de 20 minutos en promedio indicaría que estar esperando por el servicio fuera dilatado, se llega a que sólo se esperan 5 minutos por el servicio lo que es un tiempo bastante bueno por lo tanto se concluye que esta clínica trabaja de forma eficiente.

Suponiendo que como el servicio es gratuito se quisiera disminuir el número de oftalmólogos que atienden, debido a las características de este sistema en particular, esto no se puede realizar ya que el problema se dispararía al tener que ρ no es menor que 1.

3.3 SALÓN DE BELLEZA

Si se supone un salón de belleza con 5 estilistas que atienden a una sola persona cada vez, en el lugar pueden esperar sólo 6 personas más (un total de 11 personas en el sistema)

Las llegadas de los clientes se distribuyen de forma Poisson con λ igual a 9 clientes por hora, mientras que el tiempo de servicio se distribuye exponencialmente con parámetro μ igual a 1.5 servicios por hora.

Se tiene presente que no existe prioridad en el servicio y se supone que el mecanismo del servicio es "primero en llegar primero en ser atendido".
Para calcular p_0 se recurre a (2.4.2)

$$p_0 = \left[\sum_{n=0}^{c-1}\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{(\lambda/\mu)^c}{c!}\frac{1-(\lambda/c\mu)^{K-c+1}}{1-\lambda/c\mu}\right]^{-1}$$

$$= \left[\sum_{n=0}^{4}\frac{1}{n!}\left(\frac{9}{1.5}\right)^n + \frac{(9/1.5)^5}{5!}\frac{1-(9/7.5)^7}{1-9/7.5}\right]^{-1}$$

$$= [1+6+18+36+54+836.95]^{-1}$$

$$= [951.95]^{-1}$$

$$= 0.00105$$

El largo esperado de la cola, L_q, se obtiene de (2.3.3)

$$L_q = \frac{p_0(c\rho)^c \rho}{c!(1-\rho)^2}\left[1 - \rho^{K-c+1} - (1-\rho)(K-c+1)\rho^{K-c}\right]$$

$$= \frac{(0.00105)6^5(1.2)}{5!(1-1.2)^2}\left[1-(1.2)^7 - (1-1.2)(7)(1.2)^6\right]$$

$$= \frac{9341.00219}{4.8}[1.5972]$$

$$= 3.26 \text{ clientes}$$

Para obtener, L, el tamaño esperado del sistema se recurre a (2.3.4)

$$L = L_q + c - p_0 \sum_{n=0}^{c-1} \frac{(c-n)(\rho c)^n}{n!}$$

$$= 3.26 + 5 - 0.00105 \sum_{n=0}^{4} \frac{(5-n)(6)^n}{n!}$$

$$= 8.26 - 0.00105(209)$$

$$= 8.26 - 0.22$$

$$= 8.04$$

$$= 8 \text{ clientes}$$

El tiempo esperado en el sistema, W, se obtiene con (2.3.5)

$$W = \frac{L}{\lambda'}, \qquad \lambda' = \lambda(1 - p_K)$$

Se obtiene entonces λ' utilizando (2.3.1),

$$\lambda' = \lambda(1-p_K)$$

$$= 9(1 - (0.00105)(6)^{11}/(5)^6 5!)$$

$$= 9(1 - 0.203)$$

$$= 7.172$$

Entonces

$$W = \frac{L}{\lambda'}$$

$$= \frac{8.04}{7.172}$$

$$= 1.12 \text{ horas}$$

$$= 1 \text{ hora } 7 \text{ minutos}$$

El tiempo esperado en la cola, W_q, se obtiene con (2.3.7)

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

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

$$= \frac{3.26}{7.172}$$

$$= 0.455 \text{ de hora}$$

$$= 27 \text{ minutos } 16 \text{ segundos}$$

En este sistema se espera para ser atendido un poco más de 27 minutos, sólo existen en promedio 3 personas esperando en cola y 5 siendo atendidos. Considerando que no todos los clientes van a permanecer tanto tiempo para recibir el servicio es necesario pensar en tener cambios en el número de estilistas y asociar los costos del servicio y los costos asociados por la espera para asegurar que el sistema trabaja en forma eficiente.

Suponiendo que el pago de cada estilista es $120 por hora de trabajo, si el salón de belleza trabaja 8 horas, entonces el costo por cada servidor es de $960 diarios; se considera un sistema estable cuyas tasas de llegada y servicio son las mismas para todo el tiempo en que funciona al día.

El costo asociado por la espera se supone como el que se tiene por la pérdida de la ganancia de cada cliente que sale del sistema sin ser atendido y es igual a $200, si se sabe que actualmente son 1 de cada 3

personas que abandonan si esperan más de 15 minutos, entonces diariamente se deja de ganar $1600.

Debido a que el problema original conlleva a que 1 cliente abandone el sistema por hora, es decir se pierdan $48,000 mensuales en promedio, se comparan resultados tomando en cuenta que el número de servidores cambia para saber cuál es el número de estilistas necesario para un desempeño más eficiente.

Así considerando diferentes cantidades de servidores se tiene que:

a) si el número de estilistas que atienden son 5, que es el problema original, los costos esperados asociados son:

E(CS) = $144,000

E(CW) = $48,000

E(CT) = E(CS) + E(CS) = $192,000

b) si son 3 los servidores, entonces

K	=	11		
c	=	3		
λ	=	9		
μ	=	1.5		
p_0	=	0.00		
λ'	=	4.50		
L_q	=	7.01 clientes		
L	=	10.01 clientes		
W	=	2.23 de hora	=	2 horas 13 min. 32 seg
W_q	=	1.56 de hora	=	1 hora 33 min. 29 seg

Por lo que el ahorro mensual por no tener 2 estilistas es de $57,600, pero si el número de clientes que abandonan el sistema por hora aumenta a 2 puesto que la espera es de hora y media, el costo esperado por tener a más personas sin servicio aumenta a $112,000. Es decir los costos son

E(CS) = $86,400

E(CW) = $112,000

E(CT) = E(CS) + E(CS) = $198,400

Se tiene un gasto de $6,400 más que en problema original.

c) si se tienen 4 estilistas, se tiene que

K	=	11		
c	=	4		
λ	=	9		
μ	=	1.5		
p_0	=	0.00		
λ'	=	5.95		
L_q	=	5.21 clientes	=	5 clientes
L	=	9.18 clientes	=	9 clientes
W	=	1.54 de hora	=	1 hora 32 min. 36 seg
W_q	=	0.88 de hora	=	0 horas 52 min. 30 seg

Por lo que los costos promedios asociados son

E(CS) = $115,200

E(CW) = $80,000

E(CT) = E(CS) + E(CS) = $195,200

Que son $3,200 más que el problema original.

d) si hay 6 servidores atendiendo, el sistema tiene $\rho = 1$, aplicando la doble regla de L'Hôpital, (2.3.3), se llega a:

K	=	11		
c	=	6		
λ	=	9		
μ	=	1.5		
p_0	=	1		
L_q	=	1.70946 clientes	=	2 clientes
L	=	7.02568 clientes	=	7 clientes
W	=	0.88104 de hora	=	0 horas 52 min. 51 seg
W_q	=	0.21437 de hora	=	0 horas 12 min. 51 seg

El gasto por tener 1 servidor más es de $28,800 pero en esta situación no existen clientes que abandonen el sistema pues el tiempo de espera es pequeño, así que el costo total es el del estilista.

Por lo que los costos esperados son

E(CS) = $172,800

E(CW) = $0

E(CT) = E(CS) + E(CS) = $172,800

Que son $19,200 menos que el problema original.

e) Si el número de estilistas es 7, origina que.

K	=	11				
c	=	7				
λ	=	9				
μ	=	1.5				
p_0	=	0.00				
λ'	=	8.41				
L_q	=	0.78 clientes				
L	=	7.05 clientes				
W	=	0.84 de hora	=	50 min.	19 seg	
W_q	=	0.09 de hora	=	5 min.	34 seg	

Nadie abandona el sistema por lo que al considerar que el gasto por 2 estilistas más es de $57,600. Que es un costo excesivo si consideramos que además solo habrá una persona esperando en el sistema.

Así los costos esperados son

E(CS) = $201,600

E(CW) = $0

E(CT) = E(CS) + E(CS) = $201,600

Que son $9,600 más que el problema original.

Resumiendo

	c = 3	c = 4	c = 5	c = 6	c = 7
E(CS)	$86,400	$115,200	$144,000	$172,800	$201,600
E(CW)	$112,000	$80,000	$48,000	$0	$0
E(CT)	**$198,400**	**$195,200**	**$192,000**	**$172,800**	**$201,600**

Por lo tanto si el salón de belleza decide aumentar a 6 servidores, tendrá una ganancia de $19,200 extra en comparación de la que se tiene originalmente.

3.4 EL PROCESO DE PINTURA EN UNA FÁBRICA DE TRONCOS

Una fábrica tiene entre sus procesos pintar piezas de troncos que fueron previamente cortados y pulidos. Las piezas llegan al departamento de pintura con una tasa de uno cada 15 minutos con distribución Poisson. Pintarlos tiene un proceso exponencial que toma 10 minutos. El total de piezas que se aceptan en el departamento de pintura es de 7 piezas

Se considera un mecanismo de primero en llegar primero en ser servido.

Para encontrar el número esperado en el sistema, L, se recurre a (2.4.6)

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

Sustituyendo se tiene que si

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

$$= \frac{4}{6}$$

$$= 0.667$$

entonces

$$L = \frac{(0.667)}{1-(0.667)} - \frac{(8)(0.667)^8}{1-(0.667)^8}$$

$$= \frac{0.667}{0.333} - \frac{0.312}{0.961}$$

$$= 2 - 0.325$$

$$= 1.675 \text{ troncos}$$

El largo esperado de la cola, L_q, se obtiene de (2.4.7)

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

$$= 1.675 - \frac{0.667\left(1-(0.667)^7\right)}{1-(0.667)^8}$$

$$= 1.675 - \frac{0.667(0.941)}{1-0.039}$$

$$= 1.675 - \frac{0.628}{0.961}$$

$$= 1.675 - 0.653$$

$$= 1.02 \text{ troncos}$$

Para encontrar, W_q, el tiempo promedio de espera en la cola se considera la ecuación (2.4.8)

$$W_q = \frac{L}{\mu(1 - p_0)}$$

Para lo cual es necesario encontrar p_0 con la ecuación (2.4.3)

$$p_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

$$= \frac{1 - (0.667)}{1 - (0.667)^8}$$

$$= \frac{0.333}{0.961}$$

$$= 0.347$$

Entonces

$$W_q = \frac{1.675}{6(1 - 0.347)}$$

$$= \frac{1.675}{3.918}$$

$$= 0.4275 \text{ de hora}$$

Esto es 25 minutos 39 segundos.

El tiempo total de espera en el sistema, W, se calcula con (2.4.9)

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

$$= 0.4275 + \frac{1}{6}$$

$$= 0.4275 + 0.1667$$

$$= 0.5942 \text{ de hora}$$

es decir 35 minutos 39 segundos

En este caso el tiempo que esperan los troncos por ser pintados es de 25 minutos 39 segundos que es bastante grande, puesto que sólo se encuentra 1 tronco en cola y otro que ya está pintado en parte. Se tendría que pensar en un ajuste dentro de la fábrica para agilizar este proceso.

3.5 PROCESO DE DECISIÓN EN UN NEGOCIO DE COMIDA RÁPIDA

El gerente de un negocio de comida rápida dedicada al pollo frito, sabe que proporcionar un servicio rápido es la clave del éxito. Es posible que los clientes que esperan mucho vayan a otro lugar la próxima vez. Estima que cada minuto que un cliente tiene que esperar antes de terminar su servicio le cuesta un promedio de 30 centavos en negocio futuro perdido. Por lo tanto desea estar seguro de que siempre tiene suficientes cajas abiertas para que la espera sea mínima.

Un empleado de tiempo parcial opera cada caja, obtiene la orden del cliente y cobra. El costo total de cada trabajador es de $9 por hora. Y siempre hay por lo menos dos cajas disponibles.

Durante la hora del almuerzo, los clientes llegan según un proceso Poisson con tasa media de 54 por hora. Se estima que el tiempo necesario para servir a un cliente tiene una distribución exponencial con media de 2.5 minutos.

El problema es determinar cuántas cajas debe abrir el gerente durante este tiempo para minimizar el costo total esperado por hora.

Entonces considerando la función de costo espera

$$g(n) = \begin{cases} 0 & \text{si } n = 0, 1, \ldots, c \\ 18(n-c) & \text{si } n = c+1, \ldots \end{cases}$$

para c = 2, 3, 4,..., donde c es el número de cajeros trabajando.

Obteniendo los valores de p_n, de 0 a 25, puesto que $p(n > 25) \approx 0$, se tiene que para c = 2, c = 3, c = 4, c = 5 y c = 6 los costos promedios de espera son:

N = n	c = 2			c = 3			c = 4		
	g(n)	p_n	$g(n)p_n$	g(n)	p_n	$g(n)p_n$	g(n)	p_n	$g(n)p_n$
0	0	0.021	0	0	0.125	0	0	0.143	0
1	0	0.041	0	0	0.239	0	0	0.274	0
2	0	0.039	0	0	0.230	0	0	0.262	0
3	18	0.037	1	0	0.147	0	0	0.167	0
4	36	0.036	1	18	0.094	2	0	0.080	0
5	54	0.034	2	36	0.060	2	18	0.038	1
6	72	0.033	2	54	0.038	2	36	0.018	1
7	90	0.032	3	72	0.024	2	54	0.009	0
8	108	0.030	3	90	0.016	1	72	0.004	0
9	126	0.029	4	108	0.010	1	90	0.002	0
10	144	0.028	4	126	0.006	1	108	0.001	0
11	162	0.027	4	144	0.004	1	126	0.000	0
12	180	0.026	5	162	0.003	0	144	0.000	0
13	198	0.024	5	180	0.002	0	162	1E-04	0
14	216	0.023	5	198	0.001	0	180	5E-05	0
15	234	0.022	5	216	0.001	0	198	2E-05	0
16	252	0.022	5	234	0.000	0	216	1E-05	0
17	270	0.021	6	252	0.000	0	234	6E-06	0
18	288	0.020	6	270	0.000	0	252	3E-06	0
19	306	0.019	6	288	0.000	0	270	1E-06	0
20	324	0.018	6	306	0.000	0	288	6E-07	0
21	342	0.017	6	324	0.000	0	306	3E-07	0
22	360	0.017	6	342	0.000	0	324	1E-07	0
23	378	0.016	6	360	0.000	0	342	7E-08	0
24	396	0.015	6	378	0.000	0	360	3E-08	0
25	414	0.015	6	396	0.000	0	378	2E-08	0
E(CW)			103			13			3

N = n	c = 5			c = 6		
	g(n)	p_n	$g(n)p_n$	g(n)	p_n	$g(n)p_n$
0	0	0.146	0	0	0.147	0
1	0	0.280	0	0	0.282	0
2	0	0.269	0	0	0.270	0
3	0	0.172	0	0	0.172	0
4	0	0.082	0	0	0.083	0
5	0	0.032	0	0	0.032	0
6	18	0.012	0	0	0.010	0
7	36	0.005	0	18	0.003	0
8	54	0.002	0	36	0.001	0
9	72	0.001	0	54	0.000	0
10	90	0.000	0	72	0.000	0
11	108	0.000	0	90	0.000	0
12	126	0.000	0	108	0.000	0
13	144	0.000	0	126	0.000	0
14	162	0.000	0	144	0.000	0
15	180	0.000	0	162	0.000	0
16	198	0.000	0	180	0.000	0
17	216	0.000	0	198	0.000	0
18	234	0.000	0	216	0.000	0
19	252	0.000	0	234	0.000	0
20	270	0.000	0	252	0.000	0
21	288	0.000	0	270	0.000	0
22	306	0.000	0	288	0.000	0
23	324	0.000	0	306	0.000	0
24	342	0.000	0	324	0.000	0
25	360	0.000	0	342	0.000	0
E(CW)			1			0

En resumen, los costos totales esperados para diferentes cantidades de servidores son:

	c = 2	c = 3	c = 4	c = 5	c = 6
E(CS)	18	27	36	45	54
E(CW)	103	13	2	1	0
E(CT)	**121**	**40**	**38**	**46**	**54**

Debido a que es una función cóncava el mínimo costo total esperado se encuentra al disponer de 4 cajeros dentro del negocio, por lo tanto este es el número de empleados que deben estar en servicio para que la espera sea mínima a la hora del almuerzo.

CONCLUSIONES

En este trabajo se revisan los Sistemas de Teoría de Colas

a) Una cola – un servidor – población infinita
b) Una cola – c servidores en paralelo – población infinita
c) Una cola – c servidores en paralelo – población finita
d) Una cola – un servidor – población finita

Desarrollando en cada uno de ellos la demostración de las fórmulas que los representan.

El Capítulo 3 presenta una aplicación práctica de tales fórmulas, su utilidad consiste en que muestra las referencias de cómo plantear modelos reales con características semejantes a los ejemplos, y su respectivo proceso de decisión.

El origen teórico de las líneas de espera es el proceso de nacimiento y muerte el cual a su vez tiene sus bases en la teoría de procesos estocásticos por lo que es útil su presentación en el apéndice A1.

La utilidad del desarrollo de las fórmulas permite conocer no simplemente el resultado final de su aplicación sino todo el sustento teórico que justifica el uso de las mismas, esto además de proporcionar una base sólida de conocimiento ayuda en la modelación de problemas reales a identificar adecuadamente un modelo apropiado para la solución de tales problemas.

El resultado final en la aplicación de los modelos de colas es sumamente sensible a los elementos tasa de llegada y tasa de servicio los cuales en la mayoría de las veces son difíciles de conocer con exactitud porque pueden comportarse como variables aleatorias, sin embargo estos pueden ser estimados adecuadamente mediante las funciones de distribución exponencial, cuyas propiedades son explicadas en el apéndice A2.

En la mayoría de las veces lo que es importante saber es el tiempo que se espera en el sistema puesto que representa un costo tanto para el sistema como para el cliente, este tiempo puede variar desde un mínimo tolerable y aceptable hasta un máximo intolerable que provoque el desacuerdo del cliente por permanecer en el sistema o altos costos relacionados que conlleven a un daño en éste. Siempre es importante tomar decisiones con respecto al comportamiento del sistema, tales decisiones, que podrían llevar a modificar un sistema, se pueden ahora sustentar con la Teoría de Líneas de Espera.

Este trabajo puede ser consultado por personas que estudien una licenciatura en matemáticas, actuaría o ingeniería, que estén interesadas en la Teoría de Colas puesto que se presenta una explicación de las fórmulas usadas en los modelos de sistemas de colas finito e infinito para 1 o varios servidores, así como sus aplicaciones.

El desarrollo de Teoría de Colas es extenso y quedan por desarrollar temas tales como la explicación de sistemas aun más complejos como sistemas con algún tipo de prioridad o los sistemas que mezclan aborto. Otro tema interesante en el que se puede trabajar es el experimentar un modelo de colas a lo largo de cierto tiempo o con diferentes números de servidores, que en términos formales se llama Simulación en Líneas de Espera.

APÉNDICE

A1. PROCESOS ESTOCÁSTICOS

Un proceso estocástico es una colección indexada de variables aleatorias $\{X_t\}$, donde el índice t toma valores de un conjunto T dado. Es frecuente que Y sea considerado como el conjunto de enteros no negativos y X_t represente una característica de interés mesurable en el tiempo t.

El interés de los procesos estocásticos es describir el comportamiento de un sistema durante algunos periodos.

La estructura de los procesos estocásticos de tiempo discreto con espacio infinito se define de la siguiente forma:

El estado actual del sistema puede estar en una de M + 1 categorías mutuamente excluyentes llamadas estados, los cuales se etiquetan como de 1 a M. La variable aleatoria X, representa el estado del sistema en el tiempo t, de manera que sus únicos valores posibles son 0, 1,...,M. El sistema se observa como puntos del tiempo dados, etiquetados t=0,1,2,...Así, los procesos estocásticos $\{X_t\} = \{X_0,X_1,X_2,...\}$ proporciona una representación matemática de cómo evoluciona el estado del sistema físico.

A1.1 CADENAS DE MARKOV

Sea un conjunto discreto de puntos en el tiempo t= 0,1,2,..., en el estado del sistema. Las observaciones de puntos en el tiempo sucesivos definen al conjunto de variables aleatorias X_0, X_1, X_2,...Los valores asumidos por las variables aleatorias X_t son los estados del sistema en el tiempo t. Suponiendo que X_t asume el conjunto de valores finitos 0, 1,...,m; Entonces X_t = i implica que el estado del sistema en el tiempo t es i. La familia de variables aleatorias $\{X_t, t \geq 0\}$ es un proceso estocástico con espacio de parámetros discretos t = 0,1,2,... y espacio del estado discreto S= {0,1,2,...,m}.

Una cadena de Markov es un proceso estocástico $\{X_t, t \geq 0\}$ que para toda $x_i \in S$,

$$Pr\{X_t = x_t \mid X_{t-1} = x_{t-1},..., X_0 = x_0\} = Pr\{X_t = x_t \mid X_{t-1} = x_{t-1}\} \qquad (A1.1.1)$$

La ecuación anterior indica un tipo de dependencia entre la variables aleatorias X_t; intuitivamente, esto implica que dado el estado presente del sistema, el futuro es independiente del pasado. La probabilidad condicional:

$$Pr\{X_t = k \mid X_{t-1} = j\}, \qquad j,k \in S$$

Es la probabilidad de transición del estado j al estado k, denotada por:

$$p_{jk}(t)=Pr\{X_t = k \mid X_{t-1} = j\}. \qquad (A1.1.2)$$

La cadena de Markov es llamada (temporalmente) homogénea si $p_{jk}(t)$ no depende de t, es decir,

$$Pr\{X_t = k \mid X_{t-1} = j\} = Pr\{X_{t+m} = k \mid X_{t+m-1} = j\}$$

Para m = -(t-1), -(t-2),...,0,1,....En estos casos $p_{jk}(t)$ se denota simplemente por p_{jk}. La probabilidad de transición p_{jk}, es conocida como probabilidad de transición de un paso, debido a que el paso t-1 al siguiente paso t (o del paso t+m-1 al paso t+m se da en un solo paso).

La probabilidad de transición

$$Pr\{X_{t+n} = k \mid X_t = j\}$$

Del estado j al estado k en n pasos, (del estado j en el paso t al estado k en el paso t+n) es llamada la probabilidad de transición de n pasos, denotada por

$$p_{jk}^{(n)} = Pr\{X_{t+n} = k \mid X_t = j\} \qquad (A1.1.3)$$

Así las probabilidades de transición de n pasos son simplemente la probabilidad condicional de que el sistema se encuentre en el estado k justo después de n pasos (unidades de tiempo), dado que comenzó en el estado j en cualquier tiempo t. De donde $p_{jk}^{(1)} = p_{jk}$.

Como las $p_{jk}^{(n)}$ son probabilidades condicionales, deben ser no negativas, y como el proceso debe hacer una transición a algún estado, se tiene:

$$p_{jk}^{(n)} \geq 0, \qquad \text{para toda } j,k;\ n=0,1,2,...$$

y

$$\sum_k p_{jk}^{(n)} = 1, \qquad \text{para toda } j \in S$$

A1.2 CADENAS DE MARKOV DE TIEMPO CONTINUO

Dado $\{X_t(t),\ 0 \le t \le \infty\}$ un proceso de Markov con espacio de estados finito $S=\{0,1,2,...\}$. Asumiendo que es una cadena temporalmente homogénea. La función de probabilidad de transición continua está dada por:

$$p_{ij}(t) = Pr\{X(t+u) = j \mid X(u) = i\} \qquad t>0, \quad i,j \in S \qquad (A1.2.1)$$

que es independiente de $u \ge 0$. Se tiene para toda t,

$$0 \le p_{ij}(t) \le 1, \quad \sum_j p_{ij}(t) = 1, \qquad \text{para toda } j \in S.$$

Se supone que:

$$\lim_{t \to 0} p_{ij}(t) = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j. \end{cases}$$

Un proceso estocástico $\{X_t(t),\ 0 \le t \le \infty\}$ de tiempo continuo es una cadena de Markov continua sí:

- Tiene un número finito de estados,
- Tiene probabilidades de transición estacionarias.

Denotando a la matriz de probabilidades de transición por

$$P(t)=[p_{ij}(t)], \qquad i,j \in S$$

Sea $p_{ij}(0)= \delta_{ij}$, la condición inicial puede ser puesta como

$$P(0)=I.$$

Ahora se denota a la probabilidad de que el sistema esté en el estado j en el tiempo t por

$$\pi_j(t) = \Pr\{X(t) = j\};$$

el vector $\pi(t) = \{\pi_1(t), \pi_2(t),...\}$ es el vector de probabilidades del estado del sistema en el tiempo t. $\pi(0)$, es el vector de probabilidad inicial. Entonces

$$\pi_j(t) = \sum_i \Pr\{X(t+u) = j \mid X(u) = i\} \Pr\{X(u) = i\}$$

$$= \sum_i p_{ij}(t) \Pr\{X(0) = i\} \qquad \text{(A1.2.2)}$$

$$= \sum_i p_{ij}(t)\, \pi_1(0).$$

Así, dado el vector de probabilidades inicial $\pi(0)$ y la función de transición $p_{ij}(t)$, las probabilidades del estado pueden ser calculadas y el comportamiento probabilístico del sistema puede ser determinado completamente. La forma de la matriz (1.2.2) es

$$\pi(t) = \pi(0)P(t). \qquad \text{(A1.2.3)}$$

A1.2.1 Tiempo de permanencia

El tiempo de espera para cambiar del estado i a otro estado es una variable aleatoria τ_i, es decir el tiempo de permanencia en el estado i es τ_i, entonces $\Pr\{\tau_i > s + t \mid X(0) = i\} = \Pr\{\tau_i > s + t \mid X(0) = i,\ \tau_i > s\}\ \Pr\{\tau_i > s \mid X(0) = i\}$, $t \geq 0$. (A1.2.4)

Sea $\overline{F}_i(u) = \Pr\{\tau_i > s \mid X(0) = i\}$, $u \geq 0$.

Entonces (A1.2.4) puede ser escrito como

$$\overline{F}_i(t+s) = \overline{F}_i(t)\ \overline{F}_i(s), \qquad s, t \geq 0;$$

Esto es, que la distribución de probabilidad del tiempo que falta para que el proceso haga una transición fuera del estado dado siempre es la misma, independientemente de cuánto tiempo haya pasado el proceso en ese estado. Así, la variable aleatoria no tiene memoria; el proceso olvida su historia. Existe sólo una distribución de probabilidad (continua) que posee esta propiedad, y es la distribución exponencial. Por lo tanto

$$\overline{F}_i(u) = e^{-a_i u}, \qquad u \geq 0, \qquad a_i > 0 \text{ es una constante}$$

Entonces, el tiempo de permanencia τ_i, en el estado i es una exponencial con parámetro a_i. Con tiempos de permanencia τ_i y τ_j independientes.

Se tiene entonces, para $t \geq 0$, $T \geq 0$, la ecuación Chapman-Kolmogorov

$$p_{ij}(T + t) = \sum_k p_{ik}(T)\, p_{kj}(t), \qquad i,j,k \in S \qquad \text{(A1.2.5)}$$

o en forma de matriz

$$P(T+t)=P(T)P(t). \qquad \text{(A1.2.6)}$$

A1.2.2 Matriz de densidad de transición o generador infinitesimal

Sea la derivada en t=0:

$$q_{ij} = \frac{d}{dt} p_{ij}(0) = \lim_{h\to 0} \frac{p_{ij}(h) - p_{ij}(0)}{h} = \lim_{h\to 0} \frac{p_{ij}(h)}{h}, \quad i \neq j, \quad y$$

$$q_{ii} = \frac{d}{dt} p_{ii}(0) = \lim_{h\to 0} \frac{p_{ii}(h) - p_{ii}(0)}{h} = \lim_{h\to 0} \frac{p_{ii}(h) - 1}{h} \qquad \text{(A1.2.7)}$$

Si $-q_{ii} = q_i$. Lo que muestra que q_{ij}, para $i \neq j$ es siempre finito. Mientras que $q_i \geq 0$ existe siempre y es finito mientras S es finito, q_i puede ser infinito cuando S es de numerabilidad infinita.

Sea $Q=(q_{ij})$, entonces de (A1.2.7) la notación de la matriz es:

$$Q = \lim_{t\to 0} \frac{P(h) - 1}{h}$$

De (A1.2.7), para t pequeño,

$$p_{ij}(h) = hq_{ij} + o(t), \qquad i \neq j,$$

$$p_{ii}(h) = hq_i + o(t), \qquad \text{(A1.2.8)}$$

donde o(h) es un símbolo que denota una función de h que tiende a 0 más rápidamente que h, esto es, $o(h)/h \to 0$ como $h \to 0$.

Así

$$\sum_j p_{ij}(h)=1, \qquad \text{ó}$$

$$\sum_{j \neq i} p_{ij}(h) + p_{ii}(h) - 1 = 0;$$

Lo que lleva a

$$\sum_{j \neq i} q_{ij} + q_{ii} = 0 \qquad \text{ó}$$

$$\sum_{j \neq i} q_{ij} = q_i \qquad \text{(A1.2.9)}$$

La matriz $Q=(q_{ij})$ es llamada la matriz de transición o generador infitesimal o simplemente matriz Q. La matriz Q es tal que (i) sus elementos que están en la diagonal son negativos y los elementos fuera de la diagonal positivos, y (ii) cada renglón suma cero. Dado $S=\{0,1,2,...,m\}$ es un conjunto finito, se tiene

$$Q = \begin{bmatrix} -q_0 & q_{01} & \cdots & q_{0m} \\ q_{10} & -q_1 & \cdots & q_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ q_{m0} & q_{m1} & \cdots & -q_m \end{bmatrix}$$

A1.2.3 Comportamiento Límite

Un estado j es accesible del estado i ($i \rightarrow j$) sí, para alguna $t>0$, $p_{ij}(t)>0$. Los estados i y j se comunican sí $i \rightarrow j$ y $j \rightarrow i$. Una cadena de Markov de tiempo continuo es irreducible si todo estado es accesible para todos los otros estados (o sí cada pareja de estados son comunicados)

Sea α_{ij} el primer tiempo de entrada del estado i al estado j sin visitar j en el tiempo medio y sea F(.) su DF, es decir,

$$F_{ij}(t)=Pr\{\alpha_{ij} < t\}, \qquad t>0$$

$$F_{ij}(t)=0, \qquad t \leq 0$$

Un estado i persistente se da sí

$$\lim_{t \rightarrow \infty} F_{ii}(t)=1$$

en otro caso es un estado i transitorio.

En términos de $p_{ij}(t)$ se tiene que e estado i es transitorio sí

$$\int_0^{\infty} P_{ii}(t)dt < \infty.$$

Si el estado i, es nulo persistente, entonces

$$\lim_{t \rightarrow \infty} p_{ii}(t)=0,$$

y si el estado i es no nulo persistente, entonces

$$\lim_{t\to\infty} p_{ii}(t)>0.$$

De la ecuación (A1.2.5) Chapman-Kolmogorov se tiene

$$p_{ij}(h + t) = \sum_{k} p_{ik}(h)\, p_{kj}(t)$$

$$p_{ij}(h + t) = \sum_{k\neq i} p_{ik}(h)\, p_{kj}(t) + p_{ii}(h)\, p_{ij}(t)$$

Entonces

$$\frac{p_{ij}(h+t)-p_{ij}(t)}{h} = \sum_{k\neq i}\frac{p_{ik}(h)}{h}p_{kj}(t)+\left(\frac{p_{ii}(h)-1}{h}\right)p_{ij}(t).$$

Si se toma el límite cuando $h\to 0$ y se asume que el orden de se puede intercambiar las sumas con el límite se tiene que

$$\lim_{h\to 0}\frac{p_{ij}(h+t)-p_{ij}(t)}{h} = \sum_{k\neq i}\left[\lim_{h\to 0}\frac{p_{ik}(h)}{h}\right]p_{kj}(t)+\left[\lim_{h\to 0}\frac{p_{ii}(h)-1}{h}\right]p_{ij}(t) \qquad \text{ó}$$

$$p'_{ij}(t) = \sum_{k\neq i} q_{ik}\, p_{kj}(t) + q_i\, p_{ij}(t), \qquad \text{(A1.2.10)}$$

Esta es otra forma de la ecuación Chapman-Kolmogorov hacia atrás; en términos de elementos de la matriz Q. Entonces

$$P'(t)=QP(t). \qquad \text{(A1.2.10a)}$$

De igual forma se obtiene la ecuación Chapman–Kolmogorov hacia adelante, se parte de la ecuación (A1.2.5)

$$p_{ij}(t + h) = \sum_{k} p_{ik}(t)\, p_{kj}(h)$$

$$= \sum_{k \neq i} p_{ik}(t)\, p_{kj}(h) + p_{ii}(t)\, p_{ij}(h).$$

Se asume que las operaciones de límite y suma son intercambiables y se procede de forma semejante para obtener

$$p'_{ij}(t) = \sum_{k \neq i} p_{ik}(t) q_{kj} + q_j\, p_{ij}(t), \qquad \text{(A1.2.11)}$$

En notación de matriz se tiene

$$P'(t) = P(t)Q. \qquad \text{(A1.2.11a)}$$

Si se usa (A1.2.3), las ecuaciones (A1.2.10a) y (A1.2.11a) en la forma

$$\frac{d}{dt}\{\pi(t)\} = Q\,\pi(t) = \pi(t)Q. \qquad \text{(A1.2.12)}$$

De donde se describe de otra forma una cadena de Markov continua:

1. La variable aleatoria T_i tiene una distribución exponencial con media $1/q_i$.

2. Cuando sale de un estado i, el proceso se mueve a otro estado j, con probabilidad p_{ij}, en donde p_{ij} satisface las condiciones:

   $p_{ij} = 0$ para toda i,

   y

   $\sum_j p_{ij} = 1$ para toda i.

3. El siguiente estado al que se pasa después de i es independiente del tiempo que pasó en el estado i.

   Las intensidades de transición son:

   $$q_i = \frac{d}{dt} p_{ii}(0) = \lim_{t \to 0} \frac{1 - p_{ii}(t)}{t}, \qquad i \in S$$

   y

   $$q_{ij} = \frac{d}{dt} p_{ij}(0) = \lim_{t \to 0} \frac{p_{ij}(t)}{t} = q_i p_{ij}, \qquad \text{para } j \neq i,$$

   donde $p_{ij}(t)$ es la función de probabilidad de transición continua (A1.2.1).

Se interpretan q_i y q_{ij} como las tasas de transición. En particular, q_i es la tasa de transición hacia fuera del estado i por unidad de tiempo que pasa en el estado i. (Así, q_i es el recíproco del tiempo esperado que el proceso pasa en el estado i por cada visita al estado i; es decir $q_i = 1/E[T_i]$). De igual forma, q_{ij} es la tasa de transición del estado i al estado j en el sentido de que q_{ij} es el número esperado de veces que el proceso transita del estado i al estado j por unidad de tiempo que pasa en el estado i.

Entonces:

$$q_i = \sum_{j \neq i} q_{ij}$$

Cada vez que el proceso entra al estado i, la cantidad de tiempo que pasará en el estado i antes de que ocurra una transición al estado j es una variable aleatoria T_{ij}, donde $i,j \in S$ y $j \neq i$. Las T_{ij} son variables aleatorias independientes, donde cada T_{ij} tiene una distribución exponencial con parámetro q_{ij}, de manera que $E[T_{ij}] = 1/ q_{ij}$. El tiempo que pasa en el estado i hasta que ocurre una transición al estado (T_i) es el mínimo (sobre $j \neq i$.) de las T_{ij}. Cuando ocurre la transición, la probabilidad de que sea al estado j es $p_{ij} = q_{ij} / q_i$.

Resumiendo:

Una cadena de Markov de tiempo continuo es un proceso estocástico tal que:

(i) La transición que hace de un estado a otro estado del espacio de estados S es como una cadena de Markov de tiempo discreto, y

(ii) El tiempo que espera en el estado i antes de moverse a otro estado es una variable aleatoria exponencial cuyo parámetro depende de i pero no del estado al que visita.

A2. DISTRIBUCIÓN EXPONENCIAL

Para formular un modelo de Teoría de Colas como una representación del sistema real es necesario determinar las distribuciones de los tiempos de llegada y de servicio que pueden tomar cualquier forma, sin embargo, la que se utiliza es la distribución exponencial, su importancia radica en el hecho de que es suficientemente realista para proporcionar predicciones razonables y al mismo tiempo es suficientemente sencilla para ser matemáticamente manejable.

Sea T una variable aleatoria que representa los tiempos entre llegadas o tiempos de servicio. Entonces T tiene una distribución exponencial con parámetro α si su función de densidad de probabilidad es

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{para } t \geq 0 \\ 0 & \text{en otro caso} \end{cases} \qquad \text{(A2.1)}$$

Donde las probabilidades acumuladas para $t \geq 0$ son

$$P\{T \leq t\} = 1 - e^{-\alpha t} \qquad \text{(A2.2)}$$

$$P\{T > t\} = e^{-\alpha t} \qquad \text{(A2.3)}$$

Y el valor esperado y la varianza de T son

$$E(T) = \frac{1}{\alpha} \tag{A2.4}$$

$$var(T) = \frac{1}{\alpha^2} \tag{A2.5}$$

Consta de las siguientes propiedades:

Propiedad 1: $f_T(t)$ es una función estrictamente decreciente de t ($t \geq 0$).

Esta propiedad es útil considerando que las tareas que realiza un servidor pueden ser distintas con cada cliente, esto es generalmente el tiempo es breve pero en ocasiones se alarga.

Para los tiempos entre llegadas, se descartan las situaciones en las que los clientes que llegan al sistema tienden a posponer su entrada si ven a otro entrar antes que ellos. Además es totalmente consistente puesto que al graficar los tiempos entre llegadas contra el tiempo, a veces se tiene la apariencia de que hay varias aglomeraciones con grandes separaciones entre ellas, esto es debido a la gran probabilidad de que los tiempos entre llegadas sean pequeños y la poca probabilidad de que sean grandes, lo que es la parte justa de la aleatoriedad.

Propiedad 2: Falta de memoria.

Esta propiedad matemáticamente se establece que para t y Δt positivas se tiene

$$P\{T > t + \Delta t \mid T > \Delta t\} = P\{T > t\}$$

Es decir, la probabilidad del tiempo que falta hasta que ocurra el evento siempre es la misma, sin importar cuanto tiempo (Δt) haya pasado.

Fenómeno que ocurre con la distribución exponencial debido a que

$$P\{T > t + \Delta t \mid T > \Delta t\} = \frac{P\{T > \Delta t, T > t + \Delta t\}}{P\{T > \Delta t\}}$$

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

$$= \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}}$$

$$= e^{-\alpha t}$$

$$= P\{T > t\} \qquad \text{(A2.6)}$$

Para los tiempos entre llegadas se describe la situación común en donde el tiempo en que transcurre hasta la siguiente llegada está totalmente influenciada por cuándo ocurrió la última llegada.

Propiedad 3: El mínimo de las variables aleatorias exponenciales independientes tiene una distribución exponencial.

Sean T_1, T_2,..., T_n variables aleatorias exponenciales independientes con parámetros α_1, α_2,...,α_n, respectivamente, sea U la variable aleatoria, descrita por

$$U = \min \{T_1, T_2, \ldots, T_n\}. \tag{A2.7}$$

Si T_i representa el tiempo que pasa hasta que ocurre un tipo especial de evento, entonces U representa el tiempo que pasa hasta que ocurre el primero de los n eventos diferentes.

Como para cualquier $t \geq 0$ se tiene

$$\begin{aligned} P\{U > t\} &= P\{T_1 > t, T_2 > t, \ldots, T_n > t\} \\ &= P\{T_1 > t\}\, P\{T_2 > t\} \cdots P\{T_n > t\} \\ &= e^{-\alpha_1 t} e^{-\alpha_2 t} \cdots e^{-\alpha_n t} \\ &= \exp\left(-\sum_{i=1}^{n} \alpha_i t\right) \end{aligned} \tag{A2.8}$$

De donde se nota que U tiene una distribución exponencial con parámetro

$$\alpha = \sum_{i=1}^{n} \alpha_i \tag{A2.9}$$

Suponiendo que existe varios (n) tipos diferente de clientes, pero que los tiempos entre llegadas de cada tipo (tipo i) tienen distribución exponencial con parámetro α_i ($i = 1,2,\ldots,n$), el tiempo que falta a partir de un instante específico hasta la llegada del siguiente cliente del tipo i tendrá la misma

distribución. Por ello, si T_i es el tiempo restante medido a partir del instante en que llega un cliente de cualquier tipo, esta propiedad 3 se refiere a que U, el tiempo entre llegadas para el sistema de colas completo, tiene distribución exponencial con parámetro α definido por la ecuación (A2.9). Como resultado se puede elegir ignorar la distinción entre los clientes y seguir teniendo tiempos entre llegadas exponenciales para el modelo de colas.

Si se supone que n servidores que prestan servicio en un momento determinado y que todos ellos tienen la misma distribución exponencial con parámetro μ, sea T_i el tiempo que falta para que el servidor i (i = 1,2,..n) complete el servicio, con igual distribución con parámetro $\alpha_i = \mu$. Se concluye que U, el tiempo hasta la siguiente terminación de servicio para cualquier servidor, tiene distribución exponencial con parámetro $\alpha = n\mu$.

Cuando se usa esta propiedad es útil determinar las probabilidades de cuál de las variables aleatorias exponenciales será la que tiene el valor mínimo. En particular, la probabilidad de que T_j para el servidor j resulte ser el menor de las n variables aleatorias es

$$P\{T_j = U\} = \alpha_j / \sum_{i=1}^{n} \alpha_i \qquad \text{para } j = 1,2,\ldots,n.$$

Propiedad 4: Relación con la distribución Poisson.

Suponiendo que el tiempo entre dos ocurrencia consecutivas de un tipo específico de evento (esto es, llegadas o terminación de servicio por un servidor ocupado) tiene una distribución exponencial con parámetro α. La propiedad 4 tiene que ver con la implicación resultante sobre la distribución de probabilidad del número de veces que ocurre este evento en un periodo dado. En particular, sea X(t) el número de ocurrencias en el tiempo t ($t \geq 0$), en donde el tiempo 0 es el instante en que comienza la cuenta. Entonces para n = 0,1,2,...

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{+\alpha t}}{n!} \qquad (A2.10)$$

es decir, X(t) tiene una distribución Poisson con parámetro αt. Así para n=0,

$$P\{X(t) = 0\} = e^{-\alpha t} \qquad (A2.11)$$

que es justo la probabilidad obtenida a partir de la distribución exponencial para que ocurra el primer evento después de un tiempo t.

La media de la distribución Poisson es

$$E\{X(t)\} = \alpha t \qquad (A2.12)$$

De manera que el número esperado de eventos por unidad de tiempo es α. Así, se dice que α es la tasa media a la que ocurren los eventos.

Cuando se cuentan los eventos de manera continua, se dice que el proceso de conteo $\{X(t); t \geq 0\}$ es un proceso Poisson con parámetro α (tasa media).

Cuando los tiempos de servicio tienen una distribución exponencial con parámetro μ, esta propiedad brinda información útil sobre la terminación del servicio, que se obtiene al definir X(t) como el número de servicios completos logrados por un servidor siempre ocupado en un tiempo transcurrido t, en donde $\alpha = \mu$. Para modelos de múltiples servidores, también se puede definir X(t) como el número de terminaciones de servicio logradas por n servidores siempre ocupados en un tiempo transcurrido t, en donde $\alpha = n\mu$.

Esta propiedad es valiosa en particular para describir el comportamiento probabilístico de las llegadas cuando los tiempos entre ellas siguen una distribución exponencial con parámetro λ. En este caso, X(t) representa el número de llegadas en un tiempo transcurrido t, en donde $\alpha = \lambda$ es la tasa media de llegadas. Entonces, las llegadas ocurren de acuerdo con un proceso de entradas Poisson con parámetro λ. Este tipo de modelos se describen también con la suposición de que tienen llegadas Poisson.

Algunas veces se dice que las llegadas ocurren aleatoriamente y esto significa que suceden de acuerdo con un proceso de entradas Poisson. Una interpretación intuitiva de este fenómeno es que cada periodo de longitud fija tiene la misma oportunidad de tener una llegada sin importar cuándo ocurrió la llegada anterior, como lo sugiere la siguiente propiedad.

Propiedad 5: Para todos los valores positivos de t, $P\{T \leq t + \Delta t \mid T > \Delta t\} = \alpha\Delta t$, para Δt pequeño.

Considerando que T es el tiempo que pasa desde el último evento de cierto tipo (llegada o terminación de servicio) hasta el siguiente, y suponiendo que ha transcurrido un tiempo t sin que ocurra uno. Se sabe, por la propiedad 2, que la probabilidad de que ocurra un evento dentro del siguiente intervalo de tiempo, de longitud fija Δt, es una constante, sin importar cuán grande o pequeño sea t.

La propiedad 5 analiza que cuando el valor de Δt es pequeño, esta probabilidad constante se puede aproximar de manera muy cercana por $\alpha\Delta t$. Lo que es más, cuando se consideran distintos valores pequeños de Δt, esta probabilidad es, en esencia, proporcional a Δt, con factor de proporcionalidad igual a α. De hecho, es la tasa media a la cual ocurren los eventos (propiedad 4), por lo que el número esperado de eventos en el intervalo de longitud Δt es exactamente $\alpha\Delta t$. La única razón por la que la probabilidad de que ocurra un evento difiere de este valor es la posibilidad de que ocurra más de un evento, lo cual tienen una probabilidad despreciable cuando Δt es pequeño.

Para ver matemáticamente por qué se cumple la propiedad 5, note que el valor constante de la probabilidad (para un valor fijo de $\Delta t > 0$) es sencillamente, para cualquier $t \geq 0$,

$$P\{T \leq t + \Delta t \mid T > \Delta t\} = P\{T \leq t\}$$
$$= 1 - e^{-\alpha\Delta t} \qquad \text{(A2.13)}$$

Como la expansión de la serie e^x, para cualquier x, es

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!} \qquad \text{(A2.14)}$$

Entonces se tiene que para Δt pequeño

$$P\{T \leq t + \Delta t \mid T > \Delta t\} = 1 - 1 + \alpha\Delta t - \sum_{n=2}^{\infty} \frac{(-\alpha\Delta t)^n}{n!}$$

$$= \alpha\Delta t$$

Esta propiedad proporciona una aproximación conveniente de la probabilidad de que ocurra el evento de interés en el siguiente intervalo de tiempo pequeño (Δt).

Propiedad 6: No afecta agregar o desagregar.

Esta propiedad se usa para verificar que el proceso de entrada es Poisson.

Suponiendo que existen n tipos (tipos i) diferentes de clientes, los que arriban de acuerdo a un proceso de llegadas Poisson con parámetro λ_i (i = 1, 2,...,n). Suponga que se trata de procesos independientes, la propiedad dice que el proceso de entrada agregado (llegada de todos los clientes sin importar de que tipo son) también es Poisson, con parámetro (tasa de llegada) $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$. Esto es, si se tiene un proceso Posson no afecta agregar.

Si se supone que cada cliente que llega tiene una probabilidad fija p_i de pertenecer al tipo i (i = 1, 2,...,n), con

$$\lambda_i = p_i\lambda \quad y \quad \sum_{i=1}^{n} p_i = 1,$$

La propiedad dice que el proceso de entrada para los clientes de tipo i también debe ser Poisson con parámetro λ_i. Es decir, si se tiene un proceso Poisson, no afecta desagregar.

A3. CONCEPTOS BASICOS

Variables aleatorias discretas: dado X una variable aleatoria discreta. Si el número de valoras posibles de X es finito, o una infinito contable, X es llamado una variable aleatoria discreta. Los valores posibles de X pueden ser listados como x_1, x_2,....En el caso finito la lista termina. En el caso infinito contable la lista continua indefinidamente. Por ejemplo: X = número de trabajadores llegando cada semana.

Los posibles valores de X son dados por el espacio de la fila de X, que se denota por Rx.

$$Rx = \{0,1,2,...\}$$

Dado X una variable aleatoria discreta. Con cada resultado posible x_i en Rx, un número $p(x_i)=P(X=x_i)$ es la probabilidad que la variable aleatoria es igual al valor de X. El número $p(x_i)$, i = 1,2,...debe satisfacer las 2 condiciones siguientes:

1.- $p(x_i) \geq 0$ para toda i

2.- $\sum_{j=0}^{\infty} p(x_i) = 1$

Proceso estocástico: es una colección de variables aleatorias X_t, donde el subíndice t es un parámetro que corre sobre un conjunto de índices T. Comúnmente el índice t corresponde a unidades discretas del tiempo, y el conjunto de índices es $T = \{0,1,2,...\}$. Así X_t, podría representar los resultados de los lanzamientos sucesivos de una moneda, el número de personas en una sala cinematográfica como función del tiempo o el número de personas en una cola esperando un servicio en determinado

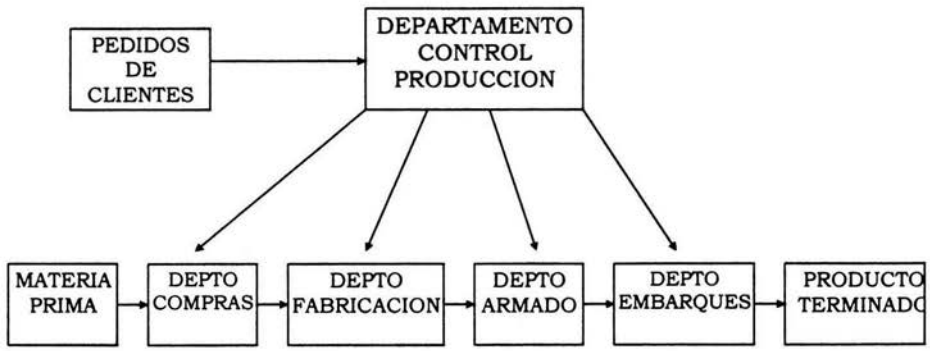
momento. Los procesos estocásticos para los cuales el conjunto T es de la forma $T = [0, \infty)$ son particularmente importantes en sus aplicaciones.

Espacio de estados: es el conjunto de todos los posibles valores asociados a un proceso estocástico. El valor asociado a X_k es el estado del proceso en el tiempo k. Si el espacio de estado contiene valores discretos el proceso estocástico se denomina discreto; de otra manera, el proceso estocástico es continuo.

Transición: es cualquier cambio de estado.

Sistema: es un conjunto de componentes que interactúan entre sí como una unidad para la consecución de un propósito explícito o implícitamente definido.

Por ejemplo tomemos una fábrica que produce y arma componentes para dar un producto. Las partes principales del sistema son el departamento de fabricación que produce las componentes, y el departamento de armado que produce los productos. Hay un departamento de compras que mantiene un suministro de materia prima y un departamento de embarques que despacha productos terminados. Un departamento de control de la producción recibe pedidos y asigna el trabajo a los otros departamentos



Entidad o Componente: denota un objeto de interés en un sistema (puede haber varios componentes).

Atributos: denota las propiedades de cualquier entidad del sistema (una entidad puede tener varios atributos).

Actividad: es todo proceso que provoque cambios en el sistema.

Estado del sistema: indica una descripción de todas las entidades o componentes, atributos y actividades de un sistema, de acuerdo con su existencia en un determinado periodo de tiempo.

En el sistema de la fábrica, las entidades son los departamentos, pedidos, componentes y productos. Las actividades son los procesos de manufactura de los departamentos, y los atributos son factores tales como las cantidades para cada pedido, tipo de componente o número de máquinas en un departamento.

Medio ambiente del sistema: todo sistema se encuentra enmarcado dentro de un sistema mayor que le sirve como marco de referencia, a este sistema mayor se le conoce como medio ambiente del sistema. Los cambios que ocurren dentro del sistema lo afectan con frecuencia sin embargo ciertas actividades del sistema pueden producir cambios que no reaccionan en el mismo sino que ocurren en el medio ambiente del sistema. En la modelación de sistemas se debe establecer un limite entre el sistema y su medio ambiente

Actividades endógenas: son aquellas que ocurren dentro del sistema a considerar.

Actividades exógenas: son aquellas que ocurren en el medio ambiente del sistema.

Actividad determinística: se considera como tal cuando el resultado de una actividad se describe completamente en términos de su entrada.

Actividad estocástica: cuando los efectos de la actividad varían aleatoriamente en distintas salidas.

Sistema continuo: son aquellos en que los cambios son predominantemente suaves.

Sistema discreto: son aquellos en que los cambios son predominantemente discontinuos.

Hay pocos sistemas totalmente continuos o totalmente discretos sin embargo en la mayoría predomina un tipo de cambio, de manera que por lo general se puede clasificar a los sistemas como discretos o continuos.

BIBLIOGRAFIA

1. MEDHÍ, JYOTIPRASAD.,
   Stochastic Models In Queueing Theory,
   Academic Press, Inc.,
   1st. edition,
   1991.

2. GROSS, DONALD Y HARRIS, CARL,
   Fundamentos de la teoría de colas,
   John Wiley & Sons,
   2nd. edition,
   1976.

3. COOPER, ROBERT B.,
   Introduction to queueing theory,
   The Macmillan Company, New York,
   1st. edition,
   1972.

4. FREDERICK S. HILLIER Y GERALD J. LIEBERMAN,
   Investigación de Operaciones,
   McGraw-Hill,
   7ma. Edición,
   2002.

5. PRAWDA WITENBERG, JUAN,
Métodos y modelos de Investigación de Operaciones II,
Limusa,
8va reimpresión,
1994.

6. WOLFF, RONALD,
Stochastic Modeling and Theory of Queues.