UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE CIENCIAS

SISTEMAS DE LINEAS DE ESPERA

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A

JAVIER IBARRA PIÑA

MEXICO, D. F. 1993

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Indice

0.1 Introducción. 9

1 Revisión de los Fenómenos de Espera. 11
1.1 Algunos Ejemplos. 11
1.2 Objetivos de la Investigación de la Congestión 12
1.3 Medición de la Congestión. 14
1.4 La Investigación de los Sistemas de Espera. 15
1.5 La Modificación de Sistemas de Espera. 18

2 Algunos Procesos Estocásticos de Interés 21
2.1 Procesos Estocásticos 21
2.2 Cadenas de Markov Discretas 22
2.2.1 Matrices de Probabilidad de Transición de Una Cadena de Markov 24
2.2.2 Un Modelo de Líneas de Espera Como Cadena de Markov 26
2.2.3 Comportamiento Límite de las Cadenas de Markov 27
2.2.4 Interpretación de la Distribución Límite 29
2.3 Procesos Poisson 31
2.3.1 Distribución de Poisson 31
2.3.2 Proceso Poisson 34
2.3.3 Distribución Exponencial y Propiedad de Pérdida de la Memoria. . 35
2.3.4 Procesos No-Homogéneos 37
2.3.5 Proceso Poisson Marcado. 38
2.4 Cadenas de Markov Continuas 40
2.4.1 Proceso de Nacimiento Puro 41
2.4.2 Proceso de Yule 46
2.4.3 Proceso Puro de Muerte. 48
2.4.4 Proceso de Nacimiento y Muerte 49
2.5 Cadenas de Markov Continuas con Estado Finito 53
2.6 Procesos de Renovación 59

2.6.1 Dos Ejemplos de Procesos de Renovación 62
2.7 Proceso de Poisson Visto Como Proceso de Renovación 63
2.7.1 Procesos Acumulativos y Relacionados. 65

3 Sistemas de Líneas de Espera 69
3.1 Descripción de una Línea de Espera 69
3.1.1 Población o Fuente 69
3.1.2 Descripción de los Arribos 71
3.1.3 Distribución del Tiempo de Servicios 71
3.1.4 Número de Servidores 71
3.1.5 Disciplina de la Cola 72
3.2 Notación y Estructura Básica. 72
3.3 La Fórmula $L = \lambda W$ 81
3.4 Arribos Poisson y Tiempos de Servicio Exponenciales. 83
3.4.1 El Sistema $M/M/1$. 85
3.4.2 El Sistema $M/M/\infty$. 89
3.4.3 El Sistema $M/M/s$ 90
3.5 El Sistema $M/G/1$ y El Sistema $M/G/\infty$. 92
3.5.1 El Sistema $M/G/1$. 93
3.5.2 Cadenas de Markov Ajustadas. 95
3.5.3 Longitud L Promedio de una Cola en Equilibrio. 96
3.5.4 El Sistema $M/G/\infty$. 98
3.5.5 Apéndice 100
3.6 Variaciones y Extensiones 100
3.6.1 Sistemas con Abortos 101
3.6.2 Tasa de Servicio Variables. 102
3.6.3 Un Sistema con Retroalimentación. 103
3.6.4 Una Cola con Dos Servidores y Sobreflujo 103
3.7 Colas Con Prioridad 106

4 Aplicación de los Modelos de Espera. 113
4.1 Descripción del Sistema I para Obtener el Modelo I. 114
4.1.1 Ambiente del Sistema I. 115
4.1.2 Tratamiento de la Información Para la Obtención del Modelo I. . . . 116
4.1.3 El Modelo I. 124
4.2 Descripción del Sistema II para Obtener el Modelo II 125
4.2.1 Ambiente del Sistema II 126

4.2.2 Tratamiento de la Información Para la Obtención del Modelo II. . . 126
4.2.3 El Modelo II. 132
4.3 Análisis Comparativo Entre el Model I y el Modelo II. 133
4.3.1 Análisis de las Modificaciones al Sistema I. 134
4.3.2 Repercución de las Modificaciones al Sistema. 134

Lista de Tablas

3.1 *Estados y transiciones que ocurren con las tasas dadas* 105
3.2 *Tasas de transición de la cadena de Markov descrita por* $(X(t), Y(t))$. . . . 107
3.3 *Tabla de resultados* 112

4.1 *Tabla modelo para la recopilación de los tiempos de arribo, en intervalos de un minuto.* 116
4.2 *Frecuencia Promedio de Arribos por minuto para el Primer Modelo.* 117
4.3 *Arribos Febrero 1991* 119
4.4 *Datos correpondientes a los servicios del primer modelo* 123
4.5 *Resumen de resultados para el Modelo I.* 125
4.6 *Frecuencia Promedio de Arribos por minuto para el Segundo Modelo.* 127
4.7 *Arribos abril 1993* 128
4.8 *Datos correspondientes a los servicios del segundo modelo* 130
4.9 *Resumen de resultados para el Modelo II.* 133
4.10 *Resumen de resultados de las comparaciones de los Modelo I y II. * Representa el aumento o decremento porcentual de las modificaciones realizadas con respecto al Modelo II.* 135

Lista de Figuras

2.1 *Un proceso de Poisson Marcado. $W_1, W_2, \cdots$ son los los tiempos de los eventos en un proceso de Poisson de tasa λ. Las variables aleatorias $Y_1, Y_2 \cdots$ son las marcas, suponiendo que sean independientes e idénticamente distribuidas, e independientes del proceso Poisson.* 39

2.2 *Un proceso de Poisson marcado.* 41

2.3 *Un modelo gráfico de un proceso de muerte, muestra los tiempos de estancia $S_N, \cdots, S_1$ y los tiempos de espera $W_1, W_2, \cdots, W_N$* 48

2.4 *La relación entre dos tiempos de ocurrencia X_k y el proceso de renovación $N(t)$.* 60

2.5 *El exceso de vida γ_t, la vida corriente δ_t, y la vida total β_t.* 62

2.6 *El exceso de vida γ_t sobrepasa x si y sólo si no hay renovaciones* 64

2.7 *Un proceso de renovación, en el cual una variable aleatoria Y_i asociada representa una porción del i-ésimo intervalo de renovación.* 66

3.1 *Elementos de un sistema de líneas de espera.* 70

3.2 *Diagrama-tiempo de una línea de espera.* 76

3.3 *El número de arribos y servicios ocurridos en un sistema de líneas de espera. Los valores suavizados en (b) simbolizan el comportamiento promedio. La tasa de arribos por unidad de tiempo es λ, el número promedio de clientes en el sistema es L, y el tiempo promedio gastado por un cliente en el sistema es W.* 83

3.4 *Si n clientes están esperando en el sistema de líneas de espera con abortos, la llegada de un cliente produce una entrada al sistema con una probabilidad p_n, y una no entrada o aborto con probabilidad $1 - p_n$.* 84

3.5 *Los períodos ocupados B_k y los desocupados I_k de un sistema de líneas de espera. Cuando los arribos forman un proceso Poisson, entonces $X_k = B_k + I_k$, $k = 1, 2, \cdots$ son variables aleatorias independientes no-negativas e idénticamente distribuidas, y forman un proceso de renovación* 89
3.6 *Un sistema de líneas de espera con s servidores* 91
3.7 *Para el modelo $M/G/\infty$ el número de clientes en el sistema en el tiempo t corresponde al número de parejas ordenadas (W_k, V_k) para los cuales $W_k < t$ y $W_k + V_k > t$. (En esta figura se representa a tres clientes en el sistema en el tiempo t).* 99
3.8 *Una Cola con retroalimentación* 103
3.9 *Un modelo de sobreflujo con dos servidores* 104
3.10 *En equilibrio, la tasa de flujo de entrada en cualquier estado debe ser igual a la tasa de flujo de salida. Aquí se ilustra el estado (m, n) cuando $m \geq 1$ y $n \geq 1$.* 108

4.1 *Histograma de frecuencia de los arribos al sistema I* 118
4.2 *Comparación entre la distribución empírica y la distribución real, para los arribos del Sistema I* 120
4.3 *Histograma de frecuencias de los arribos al Sistema II.* 128
4.4 *Comparación entre la distribución empírica y la distribución real, para los arribos del Sistema II.* 131

0.1 Introducción.

El motivo de este trabajo, es una de las actividades humanas más comunes pero también más desagradables, la espera. Se puede preguntar ¿Qué lleva a los hombres a tratar de entender estos fenómenos?, la respueta es que através del entendimiento se pueden encontrar métodos para reducirla. Piense por un momento cuanto tiempo pasa esperando en diversas colas (o líneas de espera) a lo largo del día: esperando en una calle transitada una luz verde; haciendo cola para entrar a un estacionamiento; esperando para tomar el desayuno; esperando para hacer una llamada telefónica, etcetera. La lista es interminable, pero lo que es seguro es que todas los días se tiene que esperar por algo.

Las colas y los sistemas de líneas de espera han sido objeto de investigación considerable desde la aparición del primer sistema telefónico, además de los modelos originados en biología y genética, que han sido los ejemplos más claros de procesos aleatorios reales de estado de espacio discreto.

En los años posteriores a la Segunda Guerra Mundial, los temas principales de la investigación de operaciones, (inventarios y control de producción), se vieron incrementados por el interés de esta nueva área. Se descubrió rápidamente que los modelos de rehabilitación de sistemas complejos podían ser formulados en términos de colas (descomposturas como arribos y reparaciones como servicios). Además, estos dos aspectos han producido un fuerte incremento en la literatura de problemas de optimización para modelos particulares de colas.

La simulación de sistemas de computo y sistemas de transmisión de datos, abrieron el camino, en los sesentas, a los estudios de colas caracterizadas por disciplinas de servicio complejas y crearon la necesidad de analizar sistemas interconectados. El progreso en esta área fué rápidamente aplicado, (aplicaciones en la industria fueron aceptadas en los setentas). En el presente en la industria de computo, modelos de redes de colas han resultado como paquetes de software para las soluciones automáticas de los problemas que se presentan en el diseño de nuevas computadoras y en la evaluacion e implementación de sistemas existentes.

El objetivo de este trabajo es la elaboración de un texto que permita el entendimiento y las aplicaciones de la teoría de líneas de espera, la disposición de los temas es la siguiente:

En el Capítulo 1 se describe el campo de aplicación de la teoría, y se establece la metodología para la elaboración de los modelos que describen a los sistemas de líneas de espera. En el Capítulo 2 se desarrolla la teoría de procesos estocásticos necesaria para poder abordar la de Sistemas de líneas de espera, se inicia con la defición de procesos estocásticos,

y de las cadenas de Markov discretas, se hace una explicación de los procesos de Poisson y se sigue con las cadenas de Markov continuas, que generan los procesos de nacimiento puro, los procesos puros de muerte y los procesos de nacimiento y muerte; la definición de los procesos de renovación permitirá describir a los fenómenos Poisson como uno de estos.

El Capítulo 3 describe la teoría de líneas de espera que es la que brinda los modelos matemáticos para poder describir las colas. Mientras que en el Capítulo 4 se elabora un ejemplo de aplicación de la teoría. El ejemplo es el fenómeno de espera que se registra en la biblioteca de la Facultad de Ciencias. Este fenómeno sufrió modificaciones a finales de 1992 y se dispuso de información para la descripción del modelo que describía al sistema anterior y al posterior. Así, se describen ambos y se realiza un análisis comparativo.

Capítulo 1

Revisión de los Fenómenos de Espera.

1.1 Algunos Ejemplos.

Se comenzará analizando algunos ejemplos de congestión. Esto servirá para ilustrar el amplio campo de aplicaciones que cubre la teoría de líneas de espera, así como para introducir algunas ideas importantes. La circunstancia común a todos los sistemas que se van a considerar es el flujo de *clientes* que requieren un *servicio* , habiendo cierta restricción en el servicio. Por ejemplo, los clientes pueden ser aviones solicitando permiso para despegar, y la restricción en el *servicio* es que en la pista solamente puede estar un avión. O bien los clientes pueden ser pacientes que llegan a una clínica para consultar a un médico, la restricción en el servicio es el hecho de que sólo puede atenderse a un paciente a la vez. De hecho estos dos ejemplos son casos de lo que se llamará cola, (o línea de espera), con un *sólo servidor*. Un ejemplo de colas con varios *servidores* son las filas para comprar gasolina en una estación de servicio, o también las que se forman delante de las cajas de un supermercado. En estos casos la restricción consiste en que no pueden atenderse más de, m clientes simultáneamente.

Algunos otros ejemplos de sistemas en los que está restringido el número de los clientes que pueden atenderse simultáneamente son los siguientes:

a) Artículos que van sobre una banda transportadora y que deben empacarse en cajas;

b) Máquinas que se paran de cuando en cuando y necesitan ser atendidas por un operador antes de ponerlas de nuevo en marcha, y el operador sólo puede atender una máquina a la vez;

c) Algunos problemas relacionados con centrales telefónicas;

d) Artículos, (que se entenderan como clientes), que llegan a un departamento de prueba, y el número de estos que pueden tratarse a la vez, está limitado por el de los trabajadores que hay en el departamento.

En los ejemplos citados hasta ahora, la restricción en el servicio consiste en que solamente un número limitado de clientes pueden ser atendidos a la vez, y la congestión surge debido a que los clientes no atendidos deben formarse y esperar su turno para recibir el servicio. Algunas veces, sin embargo, la restricción consiste en que el servicio sólo puede prestarse durante algunos períodos limitados, y en estos puede o no existir límite al número de clientes que pueden atenderse a la vez. Por ejemplo, si los clientes son los peatones que esperan poder cruzar una calle transitada, en un crucero sin control, el servicio sólo es dable cuando se produce un claro suficiente entre los automóviles que pasan; cuando esto ocurre, puede cruzar la calle (es decir, recibir el servicio); simultáneamente un gran número de clientes. Si los clientes son automóviles esperando la luz verde en un semáforo, o esperando poder entrar de una calle a una avenida principal, el servicio está restringido tanto por la necesidad de esperar un *período libre* como esperar que los clientes que hayan llegado antes logren pasar. Esto es, se tiene una mezcla de los dos tipos de restricción en el servicio. Otro ejemplo es el de una cola de gente esperando un automóvil de alquiler; en este caso, el que un cliente pueda ser atendido en cuanto llegue depende de que haya o no automóviles desocupados en el sitio.

Un tipo de aplicación, estrechamente relacionado con una cola de un sólo servidor se refiere al tamaño de los almacenes. Supóngase que los clientes de la descripción anterior son artículos colocados de cuando en cuando en un almacén, con la restricción de que sólo puede almacenarse un número limitado de artículos a la vez. El servicio a un cliente consiste en sacar un artículo del almacén para ser usado. En esta aplicación, estaremos interesados en la relación entre el contenido máximo del almacén y aquel en que se corre el riesgo de que este quede vacío en el momento de un nuevo pedido.

1.2 Objetivos de la Investigación de la Congestión

En los ejemplos de la sección anterior se presenta una congestión si hay suficiente irregularidad en el sistema. Por ejemplo, en una cola con un sólo servidor, supóngase ya sea que los clientes lleguen en forma irregular o bien que hay una demora apreciable en el tiempo necesario para atender a un cliente, o ambas cosas. Entonces, de vez en cuando habrá, al mismo tiempo, más de un cliente demandando un servicio; con excepción de uno de ellos, todos deberán formar cola y esperar su turno para el servicio, y entonces se producirá

una congestión. Este sencillo hecho ilustra un importante principio general; *la congestión que ocurre en un sistema depende esencialmente de las irregularidades en el mismo y no solamente de las propiedades en promedio de él.* En capítulos posteriores se especificarán estas irregularidades en terminos matemáticos.

El objetivo práctico en la investigación de un sistema con congestión será generalmente corregirlo; modificándolo de alguna manera. Por ejemplo, la proporción de llegadas de clientes puede ser tan alta que se formen colas grandes, con la consecuencia de que el tiempo de espera por cliente sea excesivo, o bien dicha frecuencia puede ser tan baja que las facilidades o medios de servicio estén sin usarse durante una gran parte de tiempo. En uno u otro caso una modificación del sistema podría ser conveniente. O quizá pueda considerarse una reorganización radical del sistema, tal como una reducción en el tiempo de servicio mediante la automatización. De cualquier manera, es útil poder predecir la cantidad de congestión que es posible que ocurra en el sistema modificado. Esto no solamente es necesario para indicar cuál de las varias modificaciones es tal vez la mejor para su estudio experimental, sino que en algunas aplicaciones, especialmente en la industria, es imposible probar experimentalmente la modificación antes de una decisión relativa a su introducción. En este caso, la decisión debe estar basada ya sea en algunas conjeturas o en una predicción teórica de lo que podría suceder al introducirse dicha modificación.

Por ejemplo, suponga que los clientes son obreros que de vez en cuando necesitan afilar sus herramientas en una máquina afiladora. Se pierde producción mientras los obreros forman cola esperando su turno en dicha máquina. No se necesita un análisis teórico para encontrar el monto de este tiempo perdido; puede determinarse de acuerdo con la frecuencia con que se necesita afilar las herramientas, el tiempo medio requerido para afilar una de ellas y otras propiedades del sistema. Pero suponga que se está considerando la introducción de una segunda máquina de afilar. Una decisión racional sobre su introducción debe depender en parte, de evaluar la utilidad anual proveniente de la producción adicional debida a la reducción del tiempo perdido en las colas y compararla con el costo de compra de la nueva máquina. Por supuesto, la producción adicional no puede medirse experimentalmente antes de decidir sobre la nueva máquina, pero suelen usarse resultados teóricos para predecir lo que pasará, y se comprueba la predicción mediante experimentos en pequeña escala antes de entrar de lleno a la aplicación práctica.

Existen diversas maneras de describir la congestión; por ejemplo, en función de los períodos libres y ocupados del servidor. Con el fin de predecir una o más de estas cantidades debe especificarse el sistema suficientemente, y esto generalmente significa el dar:

a) la población o fuente, esto se refiere a la entidad de donde emanan los clientes.

b) la pauta de llegadas. Esto se refiere tanto a la frecuencia media de llegada de los clientes como al patrón estadístico de las llegadas.

c) el mecanismo de servicio. Esto se refiere a establecer cuándo está disponible el servicio, cuántos clientes pueden atenderse a la vez y cuánto tiempo toma el proporcionar el servicio.

d) la disciplina de la cola. Esto se refiere al método por el cual se selecciona un cliente para darle un servicio de entre todos los que esperan recibirlo.

La descripción detallada de estos supuestos será realizará en el Capítulo 3.

Una vez descrito el sistema, se convierte en un problema matemático el predecir que es lo que sucederá con este sistema. Por supuesto, cualquier descripción del sistema en términos matemáticos inevitablemente simplificará en exceso la situación práctica, en virtud del uso que se hará de conceptos ideales.

1.3 Medición de la Congestión.

Consideremos ahora brevemente algunas propiedades de un sistema de espera que pueden ser de interés práctico y que sería deseable calcular matemáticamente.

Existen tres propiedades de importancia:

a) la esperanza y la distribución de la longitud de tiempo que tiene que esperar un cliente para recibir el servicio;

b) la esperanza y la distribución del número de clientes en el sistema en un instante cualquiera;

c) la esperanza y la distribución de los periodos en que el servidor está desocupado. En colas con un sólo servidor se dice que comienza un periodo de ocupación cuando un cliente llega al sistema estando libre el servidor y dura hasta el siguiente instante en que no hay clientes esperando servicio y el servidor está desocupado.

Estas tres propiedades del sistema de espera están relacionadas de una manera general, en cuanto a que las tres esperanzas tienden a crecer a medida que el sistema está más congestionado, pero en una aplicación práctica particular no estará generalmente interesado en las tres cantidades a la vez.

Se define el *tiempo en cola* de un cliente como el tiempo que pasa desde que entra en el sistema hasta el momento en que inicia su servicio, mientras que el *tiempo de espera*

se define como el tiempo desde que entra en el sistema hasta que termina su servicio. Así, el tiempo de espera es igual al tiempo de cola más el tiempo de servicio. En este trabajo, se usarán resultados en términos de cualquiera de las definiciones anteriores que sea más conveniente, aunque tal vez, por razones prácticas se tenga que trabajar con una sóla propiedad.

El tiempo en cola o el tiempo de espera de un cliente es de interés cuando existe una pérdida económica si se hace esperar al cliente en la cola. Si la pérdida por unidad de retraso es constante, sólo necesitamos considerar la media de los tiempos de cola o de espera. Los casos en que la distribución, y no sólo su esperanza, es de interés aparecen cuando un cliente puede abandonar el sistema si se le hace esperar demasiado tiempo, o cuando hay un castigo si la espera excede a un cierto tiempo crítico. Por ejemplo, si los clientes consisten en lotes de material que han pasado por una etapa de proceso y están esperando una segunda etapa, puede suceder que los retrasos pequeños no tengan consecuencias, pero los grandes retrasos deterioren el producto final. En una situación de este tipo estaremos interesados en la distribución del tiempo de cola.

El estudio del número de individuos que esperan servicio es de particular interés cuando existe dificultad en el acomodo de los clientes mientras esperan. La tercera variable, es decir, la longitud de los períodos de servicio, es probablemente la menos ocupada en la práctica, pero en ocasiones se necesita, por ejemplo, cuando se concentra la atención en el mecanismo de servicio y se desea que los períodos de servicio largos ocurran con poca frecuencia.

En algunos problemas especiales puede necesitarse de otras variables. Así, en un problema en el que los clientes se pierden si llegan en un período de mucha congestión, podría estarse interesado en calcular propiedades del número de clientes perdidos. En general, en una aplicación práctica, el objetivo debe ser la caracterización de la congestión mediante la cantidad que tenga la interpretación práctica más directa, y casi siempre esto se refiere a la cantidad más directamente relacionada con los aspectos económicos del problema.

1.4 La Investigación de los Sistemas de Espera.

En este trabajo interesará específicamente el aspecto matemático de los fenómenos de espera. Aún así, se puede apreciar que el número de combinaciones del mecanismo de llegada, del mecanismo de servicio y la disciplina de la cola es muy grande, y que no es posible en un trabajo de extensión limitada ir más allá de algunas de la técnicas más importantes que se han usado para algunos de los modelos más comunes. Efectivamente, existen disponibles

actualmente soluciones matemáticas sólo para las situaciones relativamente más simples. Si se desean resultados numéricos para un sistema complejo, estas deberán obtenerse mediante experimentos por muestreo.

Resultará provechoso el situar el aspecto matemático de la materia en una perspectiva correcta mediante la revisión de los pasos que pueden intervenir en una investigación práctica de un problema de líneas de espera, estos son:

(i) una investigación preliminar;

(ii) determinación de las posibles modificaciones al sistema (a simple vista);

(iii) la obtención de la mayor cantidad de datos;

(iv) el estudio detallado de los efectos de la modificación del sistema;

(v) formulación de una recomendación práctica;

(vi) una prueba a pequña escala de la modificación;

(vii) la acción práctica completa.

Por supuesto, este listado intenta ser sólo una guía general, y en aplicaciones particulares pueden omitirse algunos pasos o quizá otros necesiten ser repetidos.

Los pasos *(i)–(vii)* se describen brevemente en seguida.

(i) *Investigación preliminar*

Los pasos principales en este punto son

a) en un sistema complejo, la construcción de un diagrama de flujo de los clientes y una lista de los puntos en los que puede haber restricciones en el servicio;

b) la medición aproximada de las frecuencias de llegada, tiempos de servicio y tiempos de cola en los principales puntos de congestión.

Esto dará una idea general de los principales puntos en los que ocurre congestión, o en los que hay servidores ociosos durante una gran proporción del tiempo

(ii) *Determinación a simple vista de las posibles modificaciones.*

En teoría usualmente habrá muchas maneras de modificar el sistema; en la siguiente sección se dará una lista de algunos de dichos cambios que son a la vez practicables y que pueden razonablemente conducir a un aumento en la eficiencia, esto es, se establecen los lineamientos generales de las investigaciones detalladas.

(iii) *Obtención de mayor cantidad de datos.*

En esta etapa se estiman las propiedades estadísticas de las llegadas y el tiempo de servicio con el detalle suficiente para ayudar en la selección de un modelo apropiado y poder asignar valores numéricos a los parámetros del modelo. Así, en una cola simple con un sólo servidor en la que se sospeche que las llegadas son completamente al azar, podría

bastar con obtener datos para estimar la frecuencia de llegadas, comprobar la aleatoriedad de las llegadas, para estimar la distribución de frecuencia de llegadas, y para estimar la frecuencia del tiempo de servicio. Sucede comúnmente que las llegadas son aleatorias en ciertos períodos limitados, pero el número de llegadas cambia sistemáticamente, digamos en el curso del día, en estos casos deberán determinarse los cambios sitemáticos. En sistemas más complejos, con posibilidad de que la frecuencia de llegadas y el tiempo de servicio dependan del número de clientes en el sistema, o donde haya sistemas de prioridad, será quizá necesario llevar un registro detallado del comportamiento del sistema. En todos los casos en que se requieran resultados cuantitativos confiables es esencial el usar los métodos de muestreo más adecuados.

Además de las cantidades que describen las llegadas, el tiempo de servicio, etc., será generalmente conveniente medir la congestión registrando los tiempos de cola, o la distribución del número de clientes en la cola; estos pueden usarse para comprobar la aplicabilidad de cualesquiera fórmulas que se propongan usar en la etapa (iv).

(iv) *Los factores de la modificación*

Aquí se tratará de predecir los efectos del tipo de modificación bajo consideración. Esto puede hacerse de tres maneras: mediante experimentación directa, generalmente usando un diseño factorial; mediante cálculos teóricos , y finalmente, mediante un estudio de simulación. Por lo general, la primera se descartará; por ejemplo, si el objeto de la investigación es decir si se justifica una modificación costosa de un sistema, usualmente se excluirá la experimentación directa. Tanto en un estudio teórico como en uno de simulación, se supondrá que algunos parámetros del sistema, por ejemplo, las frecuencias de llegada, no cambian después de la modificación y, de esa manera, pueden obtenerse predicciones de algunas propiedades relevantes del sistema modificado. Como se sugiere en (iii), frecuentemente será aconsejable si los tiempos teóricos de cola, etc., del sistema original concuerdan bien con la observación.

A menudo resultará de efectividad la combinación del análisis matemático con la simulación, usando esta última, por ejemplo, para comprobar las suposiciones hechas para simplificar el análisis matemático.

(v) *Formulación de recomendaciones prácticas.*

Esto implica el seleccionar de entre las modificaciones practicables en (iv) aquella que satifaga algún criterio de favorabilidad. Este criterio puede estar expresado ya sea solamente en términos de costo, o bien que deba reducirse al mínimo un costo sujeto a alguna condición, como que la probabilidad de que el tiempo de espera de un cliente exceda, digamos w_0, no sea mayor que α_0, por ejemplo. En general, los criterios de optimización,

de cualquier tipo que sean, se sujetan a un examen crítico.

(vi) *Prueba en pequña escala*

En cuanto sea posible, es muy aconsejable una prueba en pequeña escala de la modificaciónes propuestas en (v). Esto tiene por objeto descubrir cualesquiera tropiezos prácticos que hayan sido pasados por alto en el análisis teórico, y el comprobar que las predicciones teóricas hechas en (iv) sean razonablemente precisas. Es conveniente efectuar observaciones suficientemente detalladas para poder interpretar cualquier discrepancia entre, digamos, los tiempos de cola observados y los predichos en (iv). Una discrepancia sería entre el comportamiento teórico y el observado puede deberse bien sea a un análisis teórico o estudio de simulación inadecuado, donde quizá se ha ignorado algún efecto importante, o a un cambio en los parámetros del sistema que se supusieron constantes en el análisis teórico.

Así, puede suceder, por razones ajenas o no a la modificación, que la frecuencia de llegadas durante la prueba en pequeña escala es notablemente diferente de la del período inicial (i) y (iii).

(vii) *Acción práctica completa.*

El paso final es la introducción a toda escala del nuevo sistema. Es aconsejable la observación de vez en cuando del sistema modificado.

1.5 La Modificación de Sistemas de Espera.

En muchas aplicaciones prácticas, los tipos de cambio en el sistema que son factibles y que pueden traer ventajas prácticas serán bastante claros. Sin embargo, a veces es útil el disponer de una lista de las modificaciones más comunes, que se darán a continuación.

En un sistema complejo en el que cada individuo pasa cada vez por varios puntos donde hay una cola, puede ser posible el reorganizar la pauta de flujo. Sin embargo, si se trata de una sola cola, las modificaciones pueden hacerse a las llegadas, al mecanismo de servicio y a la disciplina de la cola.

(i) *Modificaciones a las llegadas.*

a) modificar la frecuencia media total de llegadas, por ejemplo, mediante la exclusión de algunos clientes para el servicio;

b) controlar los tiempos de llegada de los clientes individuales con un sistema de citas, usualmente diseñado para producir llegadas regulares;

c) emparejar las variaciones sistemáticas en el número de llegadas, tratando, de asegurar un flujo más uniforme de clientes en el día, sin controlar los tiempos de llegada de los clientes individuales;

d) arreglar las cosas de manera que los clientes se animen o se desalienten para entrar en la cola, dependiendo del número de clientes que ya estén en ella.

(ii) *Modificación al mecanismo de servicio.*

Se puede:

a) reducir el tiempo medio de servicio;

b) reducir el coeficiente de variación del tiempo de servicio;

c) hacer arreglos para que los tiempos de servicio se reduzcan durante los períodos de congestión mayor que la promedio;

d) cambiar la capacidad del sistema proporcionando más servidores;

e) hacer arreglos para que la capacidad aumente temporalmente, bien sea cuando se observe una congestión alta, o cuando se espera que lleguen más clientes del número promedio;

f) aumentar la disponibilidad de las estaciones de servicio, bien sea en promedio, o haciendo que sea más factible la disponibilidad del servicio cuando hay clientes presentes.

(iii) *Modificación de la disciplina de la cola.*

Intentar:

a) dar prioridad, preponderante o no preponderante, a clientes "importantes", esto es a aquellos para los que el costo de espera por unidad de tiempo es alto;

b) dar prioridad a los clientes cuyos tiempos de servicio se espera que sean cortos;

c) introducir un sistema en el que la disciplina de la cola no es "primero en llegar-primero en ser servido" algún artificio que asegure la reducción de la probabilidad de tiempos largos de cola;

d) en una cola con varios servidores, modificar la regla mediante la cual se asignan los clientes a los servidores. Existen varias maneras de lograr esto.

Capítulo 2

Algunos Procesos Estocásticos de Interés

2.1 Procesos Estocásticos

Un *proceso estocástico* es una familia de variables aleatorias X_t, donde t es un parámetro que corre sobre un conjuto de índices T. Comunmente el índice t corresponde a unidades discretas del tiempo, y el conjunto de índices es $T = \{0, 1, 2, \cdots\}$. En este caso, 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, la posición de una partícula en un espacio específico, también 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. Aquí t frecuentemente representa al tiempo, pero también puede representar cualquier otro parámetro, por ejemplo, t puede representar la distancia a un origen arbitrario y X_t ser el número de artículos defectuosos en el intervalo $(0, T]$ o el número de autos en el mismo intervalo a lo largo de una autopista.

Los *procesos estocásticos* se clasifican por su *espacio de estados*, o el rango de valores posibles que toma la variable aleatoria X_t, en su conjunto de índices T y la relación de dependencia que existe entre las variables aleatorias X_t. Los procesos estocásticos que se estudiarán en este trabajo y sus características serán descritos a lo largo de este capítulo.

2.2 Cadenas de Markov Discretas

Un *poceso de Markov* $\{X_t\}$ es un proceso estocástico con la siguiente propiedad; dado el valor de X_t, los valores de X_s para $s > t$ no están influenciados por los valores X_u para $u < t$. Es decir, la probabilidad de cualquier comportamiento futuro del proceso, cuando el estado presente es conocido, no es alterado por el conocimiento adicional de su comportamiento pasado. Una *cadena de Markov discrta* es un proceso de Markov cuyo espacio de estados es un conjunto finito o contable y cuyo conjunto de índices $T = (0, 1, 2, \cdots)$. En términos formales la propiedad de Markov es:

$$\Pr\{X_{n+1} = j / X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i\} = \Pr\{X_{n+1} = j / X_n = i\} \quad (2.1)$$

para todos los puntos n y todos los estados $i_0, i_1, \ldots, i_{n-1}, i, j$.

Es conveniente etiquetar el espacio de estados de una cadena de Markov como el conjunto de enteros no-negativos $\{0, 1, 2, \cdots\}$ y frecuentemente se dece que X_n se encuentra en el estado i cuando $X_n = i$.

La probabilidad de que X_{n+1} se encuentre en el estado j dado que X_n está en el estado i es llamada *probabilidad de transición de un paso* y es denotada como $P_{ij}^{n,n+1}$. Es decir,

$$P_{ij}^{n,n+1} = \Pr\{X_{n+1} = j / X_n = i\}. \quad (2.2)$$

Esta notación enfatiza el hecho de que en general las probabilidades de transición son funciones que dependen no únicamente de los estados iniciales y finales, sino que también dependen del tiempo de transición. Cuando las probabilidades de transición de un paso son independientes del tiempo n, se dice que la cadena de Markov tiene *probabilidades de transición estacionarias*. Dado que todas las cadenas de Markov que usaremos tienen probabilidades de transición estacionarias, la discusión estará limitada a este caso. Entonces $P_{ij}^{n,n+1} = P_{ij}$ que es independiente de n y P_{ij} es la probabilidad condicional de que el estado vaya de i a j en un paso. Una notación conveniente para representar las probabilidades de transición es la forma matricial.

$$\mathbf{P} = \begin{Vmatrix} P_{00} & P_{01} & P_{02} & P_{03} & \cdots \\ P_{10} & P_{11} & P_{12} & P_{13} & \cdots \\ P_{20} & P_{21} & P_{22} & P_{23} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & P_{i3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{Vmatrix}$$

y para referirse a la matriz de Markov $\mathbf{P} = (P_{ij})$ se hará como *la matriz de probabilidad de transición* del proceso.

El $i+1$-ésimo renglón de $\mathbf{P}$ es la distribución de los valores de X_{n+1} dada la condición $X_n = i$. si el número de estados es finito $\mathbf{P}$ es una matriz cuadrada finita cuyo orden es igual al número de estados. Al ser P_{ij} una probabilidad satisface:

$$P_{ij} \geq 0 \tag{2.3}$$

$$\sum_{i=0}^{\infty} P_{ij}^{n} = 1 \tag{2.4}$$

La condición (2.4) expresa el hecho de que algunas transiciones ocurren en cada evento, (por conveniencia, se dice que una transición ha ocurrido aunque el estado permanezca sin cambio.)

Un proceso de Markov está completamente definido una vez que su matriz de transición de probabilidad y su estado inicial X_0 (o la distribución de probabilidad de X_0) están especificadas.

Sea $\Pr\{X_0 = i\} = p_i$. Es suficiente con mostrar como se calculan la cantidades

$$\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \cdots, X_n = i_n\} \tag{2.5}$$

ya que las probabilidades $X_{j1}, \cdots, X_{jk}, para j_1 < \cdots < j_k$, pueden ser obtenidas, de acuerdo con el axioma de probabilidad total, sumando los términos de la forma (2.5).

Por la definición de probabilidad condicional se tiene,

$$\begin{aligned} &\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \cdots, X_n = i_n\} = \\ &\quad \Pr\{X_0 = i_0, X_1 = i_1, \cdots, X_{n-1} = i_{n-1}\} \\ &\times \Pr\{X_n = i_n / X_0 = i_0, X_1 = i_1, \cdots, X_{n-1} = i_{n-1}\} \end{aligned} \tag{2.6}$$

Por la definición de proceso de Markov,

$$\begin{aligned}&\Pr\{X_n = i_n / X_0 = i_0, X_1 = i_1, \cdots, X_{n-1} = i_{n-1}\}\\ &= \Pr\{X_n = i_n / X_{n-1} = i_{n-1}\} P_{i_{n-1} i_n}.\end{aligned} \tag{2.7}$$

Sustituyendo, (2.7) en (2.6) se tiene,

$$\begin{aligned}&\Pr\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\}\\ &= \Pr\{X_0 = i_0, X_1 = i_1, \cdots, X_{n-1} = i_{n-1}\} P_{i_{n-1} i_n}.\end{aligned}$$

Entonces, por inducción, (2.5) se convierte en:

$$\begin{aligned}&\Pr\{X_0 = i_0, X_1 = i_1, \cdots X_n = i_n\}\\ &= p_{ij} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}} P_{i_{n-1} i_n}.\end{aligned} \tag{2.8}$$

Esto muestra que todas las probabilidades están especificadas una vez que las probabilidades de transición y la distribución inicial son dadas, y en este sentido el proceso está definido por estas cantidades. Los cálculos anteriores muestran que (2.1) es equivalente a:

$$\begin{aligned}&\Pr\{X_{n+1} = j_1, \cdots, X_{n+m} = j_m / X_0 = i_0, \cdots, X_n = i_n\}\\ &= \Pr\{X_{n+1} = j_1, \ldots, X_{n+m} = j_m / X_n = i_n\}\end{aligned} \tag{2.9}$$

para los puntos n, m y todos los estados $i_0, \ldots, i_n, j_1, \ldots, j_m$.

2.2.1 Matrices de Probabilidad de Transición de Una Cadena de Markov

Una cadena de Markov está completamente definida por su matriz de transición de probabilidad de un paso y la distribución del estado del proceso en el tiempo 0. El análisis de una cadena de Markov consiste básicamente en el cálculo de las probabilidades de las posibles relaciones de los procesos. Los cálculos que interesan principalmente son las matrices de probabilidad de transición de n-pasos $\mathbf{P}^{(n)} = (P_{ij}^{(n)})$. Aquí $P_{ij}^{(n)}$ denota la probabilidad del proceso estando en el estado i vaya al estado j en n transiciones. Formalmente, sería:

$$P_{ij}^{(n)} = \Pr\{X_{n+m} = j / X_m = i\}. \tag{2.10}$$

Observe que se está trabajando únicamente con un proceso que tiene probabilidades de transición estacionarias, de otra forma el lado izquierdo de (2.10) dependería también de m.

La propiedad de Markov permite expresar (2.10) en términos de (P_{ij}) como se muestra en el siguiente teorema.

Teorema 2.1 Las probabilidades de transiciones de n-pasos de una cadena de Markov satisfacen

$$P_{ij}^{(n)} = \sum_{k=0}^{\infty} P_{ik} P_{kj}^{(n-1)}, \tag{2.11}$$

donde se define

$$P_{ij}^{(0)} = \begin{cases} 1 & si\ i = j \\ 0 & si\ i \neq j \end{cases}$$

De la teoría de matrices se reconoce la relación (2.11) como la fórmula de multiplicación de matrices, entoces $\mathbf{P}^{(n)} = \mathbf{P} \times \mathbf{P}^{n-1}$. E iterando esta fórmula, se obtiene:

$$\mathbf{P}^{(n)} = \mathbf{P} \times \mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P} = \mathbf{P}^n; \tag{2.12}$$

en otras palabras, las probabilidades de transición de n-pasos $P_{ij}^{(n)}$ son las entradas en la matriz $\mathbf{P}^n$, es decir la n-ésima potencia de $\mathbf{P}$.

Demostración: La demostración procede vía análisis de transición de un paso, seguida de la propiedad de Markov. El evento de ir del estado i al estado j en n transiciones puede ser realizado de formas determinadas, ir a algún estado intermedio k, ($k = 0, 1, 2...$), en la primera transición e ir del estado k al estado j en las restantes $(n-1)$ transiciones. Dada la propiedad de Markov, la probabilidad de la segunda transición es $P_{kj}^{(n-1)}$ y la de la primera P_{ik}. Si se usa la ley de probabilidad total se tiene

$$\begin{aligned} P_{ij}^{(n)} &= \Pr\{X_n = j / X_0 = i\} \\ &= \sum_{k=0}^{\infty} \Pr\{X_n = j, X_1 = k / X_0 = i\} \end{aligned}$$

$$= \sum_{k=0}^{\infty} \Pr\{X_1 = k/X_0 = i\} \Pr\{X_n = j/X_0 = i, X_1 = k\}$$

$$= \sum_{k=0}^{\infty} P_{ik} P_{kj}^{(n-1)}.$$

Si la probabilidad de los procesos estando inicialmente en el estado j es p_j, es decir, la función de distibución de X_0 es $\Pr\{X_0 = j\} = p_j$, entonces la probabilidad del proceso estando en el estado k en el tiempo n ès

$$p_k^{(n)} = \sum_{j=0}^{\infty} p_j P_{jk}^{(n)} = \Pr\{X_n = k\}. \tag{2.13}$$

2.2.2 Un Modelo de Líneas de Espera Como Cadena de Markov

Los clientes llegan a tomar un servicio y toman su lugar en una línea de espera. Durante un determinado período de tiempo , una cliente es servido, suponiendo que al menos un cliente está presente. Si no hay clientes esperando el servicio, entonces, durante este período no hay servicios. Durante el período de servicios nuevos clientes pueden llegar a tomar su lugar en la línea de espera. Supóngase que el número real de clientes que llegan en el n-ésimo período es una variable aleatoria ξ_n cuya distribución depende del período y está dada por:

$$\Pr\{k \text{ clientes llegan en el periodo de servicio}\} = \Pr\{\xi_n = k\} = a_k$$

para $k = 0, 1, \cdots$ donde $a_k \geq 0$ y $\sum_{k=0}^{\infty} a_k = 1$.

También se supone que $\xi_1, \xi_2, \cdots$ son variables aleatorias independientes. El estado del sistema al inicio de cada período se define como el número de clientes en la línea esperando por un servicio. Si el estado presente es i, entonces después de un período el estado es:

$$j = \begin{cases} i - 1 + \xi & \text{si } i \geq 1, \\ \xi & \text{si } i = 0 \end{cases} \tag{2.14}$$

donde ξ es el número de clientes que han llegado mientras un cliente es atendido. En términos de las variables aleatorias del proceso se puede expresar (2.14) formalmente como:

$$X_{n+1} = (X_n - 1)^+ + \xi_n,$$

donde $Y^+ = max\{Y, 0\}$. Dado (2.14), la matriz de probabilidad de transición se calcula, y se obtiene:

$$\mathbf{P} = \begin{Vmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{Vmatrix}.$$

Es intuitivamente claro que si el número esperado de clientes nuevos, $\sum_{k=0}^{\infty} k a_k$, que arriban durante el período de servicio exceda a uno, entonces con el paso del tiempo, la longitud de la línea de espera crece, sin límite. Por otro lado, si $\sum_{k=0}^{\infty} k a_k < 1$, entonces la longitud de la línea de espera alcanza un equilibrio estadístico que es descrito por la distribución límite:

$$\lim_{n\to\infty} \Pr\{X_n = k / X_0 = j\} = \pi_k > 0 \qquad \text{para } k = 0, 1, \cdots$$

donde, $\sum_{k=0}^{\infty} \pi_k = 1$. Las cantidades importantes que tienen que ser determinadas por este modelo, incluyendo a las fracciones de largos períodos de tiempo en que los servidores están desocupados, dada por $\sum_{k=0}^{\infty}(1+k)\pi_k$.

2.2.3 Comportamiento Límite de las Cadenas de Markov

Supongase que una matriz de probabilidad de transición $\mathbf{P} = (P_{ij})$ que tiene un número finito de estados $\{0, 1, \cdots, N\}$ tiene la propiedad de que, cuando es elevada a una cierta potencia k, la matriz $\mathbf{P}^k$ tiene todas las entradas estrictamente positivas. La cadena de Markov que es definida por esta matriz, es llamada *regular*. El hecho más importante de estas cadenas de Markov regulares, es la existencia de una *distribución de probabilidad límite* $\pi = (\pi_0, \pi_1, \cdots, \pi_N)$ donde $\pi_j > 0$ para $j = 0, 1, \cdots, N$ y $\sum_j \pi_j = 1$ y esta distribución es independiente del estado inicial. Formalmente,para una matriz de probabilidad de transición regular $\mathbf{P} = (P_{ij})$ se tiene la siguiente convergencia:

$$\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j > 0 \qquad \text{para } j = 0, 1, \cdots, N,$$

o en términos de la cadena de Markov $\{X_n\}$,

$$\lim_{n\to\infty} \Pr\{X_n = j / X_0 = i\} = \pi_j > 0 \qquad j = 0, 1, \cdots, N.$$

Esta convergencia significa, que cuando $n \to \infty$, la probabilidad de encontrar a la cadena de Markov en el estado j es aproximadamente π_j no importando el estado de la cadena en el tiempo 0.

Teorema 2.2 *Sea* $\mathbf{P}$ *una matriz de transición de probabilidad regular en los estados* $0, 1, \cdots, N$. *Entonces la distribución límite* $\pi = (\pi_0, \pi_1, \cdots, \pi_N)$ *es la única solución no negativa de las ecuaciones:*

$$\pi_j = \sum_{k=0}^{N} \pi_k P_{kj}, \qquad j = 0, 1, \cdots, N. \tag{2.15}$$

$$\sum_{k=0}^{N} \pi_k = 1 \tag{2.16}$$

Demostración Dado que la cadena de Markov es regular, se tiene la distribución límite, $\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$, tal que $\sum_{k=0}^{N} \pi_k = 1$. Se escribe $\mathbf{P}^n$ como el producto de las matrices $\mathbf{P}^{n-1}\mathbf{P}$ en la forma

$$P_{ij}^{(n)} = \sum_{k=0}^{N} P_{ik}^{(n-1)} P_{kj}, \qquad j = 0, 1, \cdots, N, \tag{2.17}$$

y ahora, se hace $n \to \infty$. Entonces, $P_{ik}^{(n)} \to \pi_j$ mientras que, $P_{ik}^{(n-1)} \to \pi_k$ y (2.17) pasa a ser de la forma, $\pi_j = \sum_{k=0}^{N} \pi_k P_{kj}$.

Para demostrar que la solución es única, supongase que $x_0, x_1, \cdots, x_N$ resuelve la ecuación

$$x_j = \sum_{k=0}^{N} x_k P_{kj} \qquad j = 0, 1, \cdots, N \tag{2.18}$$

y

$$\sum_{k=0}^{N} x_k = 1. \tag{2.19}$$

Se Desea demostrar que la probabilidad límite, $x_j = \pi_j$. Se comenzará por multiplicar (2.18) por P_{jl} y se suma sobre j

$$\sum_{j=0}^{N} x_j P_{jl} = \sum_{j=0}^{N} \sum_{k=0}^{N} x_k P_{kj} P_{jl} = \sum_{k=0}^{N} x_k P_{kl}^{(2)}. \tag{2.20}$$

por (2.18) se tiene que, $x_l = \sum_{j=0}^{N} x_j P_{jl}^{(2)}$ y (2.20) se convierte

$$x_l = \sum_{k=0}^{N} x_k P_{kl}^{(n)} \qquad l = 0, 1, \cdots, N.$$

Repitiendo n veces el argumento anterior, se deduce que

$$x_l = \sum_{k=0}^{N} x_k P_{kl}^{(n)} \qquad l = 0, 1, \cdots, N,$$

y tomando el límite y usando el hecho de que $P_{kl}^{(n)} \to \pi_l$ se observa que:

$$x_l = \sum_{k=0}^{N} x_k \pi_l, \qquad l = 0, 1, \cdots, N.$$

Pero por (2.19) se tiene que $\sum_k x_k = 1$, entoces $x_l = \pi_l$.□

2.2.4 Interpretación de la Distribución Límite

Dada una matriz de transición regular P para un proceso de Markov $\{X_n\}$ con $N+1$ estados $0, 1, \cdots, N$ se tienen las siguientes ecuaciones lineales, las cuales deben ser resueltas.

$$\pi_i = \sum_{k=0}^{N} \pi_k P_{ki} \quad \text{para} \quad i = 0, 1, \cdots, N$$

y

$$\pi_0 + \pi_1 + \cdots + \pi_N = 1.$$

la interpretación principal de la solución $(\pi_0, \pi_1, \cdots, \pi_N)$ es la del límite de la distribución

$$\pi_j = \lim_{n\to\infty} P_{ij}^{(n)} = \lim_{n\to\infty} \Pr\{X_n = j / X_0 = i\}.$$

Es decir, después de que el proceso ha estado en operación por un largo período de tiempo, la probabilidad de encontrar el proceso en el estado j es π_j, independientemente del estado inicial.

Existe una segunda interpretación de la distribución límite $\pi = (\pi_0, \pi_1, \cdots, \pi_N)$, que juega un papel más importante en muchos modelos. Se dice que π_j también expresa la fracción media del tiempo en que el proceso $\{X_n\}$ está en el estado j cuando $n \to \infty$.

Así, si cada visita al estado j incurre en un cierto "costo" c_j, entonces el costo medio por unidad de tiempo asociado con está cadena de Markov es:

$$\text{Costo medio por unidad de tiempo} = \sum_{j=0}^{N} \pi_j c_j.$$

Para verificar esta interpretación, suponga que si una sucesión $a_0, a_1, \cdots$ de números reales converge a el límite a, entonces el promedio de estos números también converge

$$\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} a_k = a.$$

Aplicando este resultado a la convergencia de $\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$ para calcular:

$$\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} P_{ij}^{(k)} = \pi_j.$$

Ahora $\frac{1}{m}\sum_{k=0}^{m-1} P_{ij}^{(k)}$ es exactamente la fracción media del tiempo durante los pasos $0, 1, \cdots, m-1$ que el proceso pasa en el estado j. De hecho, la fracción real del tiempo en el estado j es (aleatoria):

$$\frac{1}{m} \sum_{k=0}^{m-1} 1\{X_k = j\}.$$

donde

$$1\{X_k\} \begin{cases} 1 & \text{si } X_k = j \\ 0 & \text{si } X_k \neq j \end{cases}$$

Entonces la fracción *media* de visitas es obtenida tomando los valores esperados de acuerdo con:

$$\begin{aligned} E[\frac{1}{m} \sum_{k=0}^{m-1} 1\{X_k = j\}/X_0 = i] &= \frac{1}{m} \sum_{k=0}^{m-1} E[1\{X_k = j\}/X_0 = i]. \\ &= \frac{1}{m} \sum_{k=0}^{m-1} \Pr\{X_k = j/X_0 = i\}. \\ &= \frac{1}{m} \sum_{k=0}^{m-1} P_{ij}^{(k)}. \end{aligned}$$

Debido a que el $\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$, la fracción media del tiempo en el que el proceso pasa en el estado j es:

$$\lim_{m\to\infty} E\left[\frac{1}{m}\sum_{k=0}^{m-1} 1\{X_k = j\}/X_0 = i\right] = \lim_{m\to\infty}\frac{1}{m}\sum_{k=0}^{m-1} P_{ij}^{(k)} = \pi_j,$$

independientemente del estado inicial i.

2.3 Procesos Poisson

El comportamiento Poisson es un fenómeno prevaleciente en la naturaleza y la distribución Poisson una de las más manejables y útiles.

2.3.1 Distribución de Poisson

La distribución Poisson con parámetro $\mu > 0$ está dada por:

$$p_k = \frac{e^{-\mu}\mu^k}{k!} \qquad \text{para } k = 0, 1, \cdots \tag{2.21}$$

Sea X una variable aleatoria que tiene la distribución Poisson (2.21). Calculamos la media o el primer momento como:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k p_k = \sum_{k=0}^{\infty} \frac{k e^{-\mu}\mu^k}{k!} \\ &= \mu e^{-\mu}\sum_{k=1}^{\infty}\frac{\mu^{k-1}}{(k-1)!} \\ &= \mu. \end{aligned}$$

Para evaluar la varianza,

$$\begin{aligned} E[X(X-1)] &= \sum_{k=2}^{\infty} k(k-1)p_k \\ &= \mu^2 e^{-\mu}\sum_{k=2}^{\infty}\frac{\mu^{(k-2)}}{(k-2)!} \\ &= \mu^2. \end{aligned}$$

Entonces

$$E[X^2] = E[X(X-1)] + E[X] = \mu^2 + \mu,$$

mientras que

$$\begin{aligned} \sigma_X^2 &= Var[X] = E[X^2] - E^2[X] \\ &= \mu^2 + \mu - \mu^2 = \mu. \end{aligned}$$

Como se puede observar la distribución Poisson tiene la extraña característica de que su media y su varianza están dadas por el valor μ.

Dos propiedades fundamentales de la distribución Poisson, las cuales aparecen frecuentemente en diferentes formas, son la suma de dos variables aleatorias independientes con distribución Poisson, y cierta descomposición aleatoria de los fenómenos de Poisson. Se establecerán estas propiedades en los siguientes teoremas.

Teorema 2.3 *Sean X y Y variables aleatorias independientes con distribución Poisson y parámetros μ y ν, respectivamente. Entonces la suma $X+Y$ tiene distribución Poisson con parámetro $\mu+\nu$.*

Demostración Por la ley de probabilidad total se tiene,

$$\begin{aligned} \Pr\{X+Y=n\} &= \sum_{k=0}^{n} \Pr\{X=k, Y=n-k\} \\ &= \sum_{k=0}^{n} \Pr\{X=k\} \Pr\{Y=n-k\} \\ &= \sum_{k=0}^{n} \left\{\frac{\mu^k e^{-\mu}}{k!}\right\} \left\{\frac{\nu^{n-k} e^{-\nu}}{(n-k)!}\right\} \\ &= \frac{e^{-(\mu+\nu)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \mu^k \nu^{(n-k)}. \end{aligned} \tag{2.22}$$

Recordando la expansión binomial de $(\mu+\nu)^n$

$$(\mu+\nu)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \mu^k \nu^{n-k}$$

así (2.22) se simplifica de la forma

$$\Pr\{X+Y=n\} = \frac{e^{-(\mu+\nu)}(\mu+\nu)^n}{n!}, \qquad n = 0, 1, \cdots.$$

que es, la deseada distribución Poisson. □

Para describir el segundo resultado, se considerará a una variable aleatoria Poisson N con parámetro $\mu > 0$. Se escribe a N como la suma de unos en la forma :

$$N = 1 + 1 + \cdots + 1$$

y también, considerando cada uno separada e independientemente, desapareciendo con probabilidad $1-p$ y quedándose, con probabilidad p. ¿Cual es la distribución de la suma M si es de la forma $M = 1 + 0 + 0 + 1 + \cdots + 1$?

El siguiente teorema , establece la respuesta de manera precisa.

Teorema 2.4 *Sea N, una variable aleatoria Poisson con parámetro μ, y condicional sobre N, sea M una variable aleatoria con distribución binomial con parámetros n y p. Entonces, la distribución incondicional de M es Poisson con parámetro μp.*

Demostración la demostración se realiza aplicando la Ley de Probabilidad Total. Entonces,

$$\begin{aligned}
\Pr\{M=k\} &= \sum_{n=0}^{\infty} \Pr\{M=k/N=n\}\Pr\{N=n\} \\
&= \sum_{n=k}^{\infty} \{\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}\} \{\frac{\mu^n e^{-\mu}}{n!}\} \\
&= \frac{e^{-\mu}(\mu p)^k}{k!} \sum_{n=k}^{\infty} \frac{[\mu(1-p)]^{n-k}}{(n-k)!} \\
&= \frac{e^{-\mu}(\mu p)^k}{k!} e^{\mu(1-p)} \\
&= \frac{e^{-\mu p}(\mu p)^k}{k!} \qquad k = 0, 1, \cdots.
\end{aligned}$$

y se llega a la distribución. □

2.3.2 Proceso Poisson

Definición 2.1 *Un proceso Poisson con intensidad o tasa $\lambda > 0$ es un proceso estocástico de valores enteros $\{X(t); t \geq 0\}$ para el cual.*

i) Para cualquier punto del tiempo $t_0 = 0 < t_1 < t_2 < \cdots < t_n$, los incrementos del proceso

$$X(t_1) - X(t_0), X(t_2) - X(t_1), \cdots, X(t_n) - X(t_{n-1})$$

son variables aleatorias independientes.

ii) Para $s > 0$ y $t > 0$, la variable aleatoria $X(s+t) - X(s)$ tiene distribución Poisson.

$$\Pr\{X(s+t) - X(s) = k\} = \frac{(\lambda t)e^{-\lambda t}}{k!} \text{ para } k = 0, 1, \cdots;$$

y

iii) $X(0) = 0$

en particular, se observa que si $X(t)$ es un proceso Poisson con tasa $\lambda > 0$, entonces los momentos son:

$$E[X(t)] = \lambda t \text{ y } Var[X(t)] = \sigma^2_{X(t)} = \lambda t.$$

Ejemplo 2.1 *Las llegadas de los clientes a una cierta tienda, siguen un proceso Poisson con tasa $\lambda = 4$ por hora. Dado que la tienda inicia sus operaciones a las 9 : 00 hrs., ¿Cúal es la probabilidad de que, exactamente un cliente haya llegado exactamente a las 9 : 30 hrs. y que cinco hayan llegado a las 11 : 30 horas?*

Midiendo el tiempo t en horas, de las 9 : 00 hrs., se quiere determinar $\Pr\{X(\frac{1}{2}) = 1; X(\frac{5}{2}) = 5\}$, *usando la interpretación de $X(\frac{5}{2}) - X(\frac{1}{2})$ y $X(\frac{1}{2})$, y readecuando la pregunta como:*

$$\begin{aligned}
\Pr\left\{X\left(\frac{1}{2}\right) = 1, X\left(\frac{5}{2}\right) = 5\right\} &= \Pr\left\{X\left(\frac{1}{2}\right) = 1, X\left(\frac{5}{2}\right) - X\left(\frac{1}{2}\right) = 4\right\} \\
&= \left\{\frac{e^{-4(\frac{1}{2})}4(\frac{1}{2})}{1!}\right\}\left\{\frac{e^{-4(2)}[4(2)]^4}{4!}\right\} \\
&= (2e^{-2})\left(\frac{512}{3}e^{-8}\right) = .0154965.
\end{aligned}$$

2.3.3 Distribución Exponencial y Propiedad de Pérdida de la Memoria.

Una variable aleatoria que se presenta con mucha frecuencia es la variable aleatoria exponencial. Aunque puede ocurrir en muchos casos diferentes, se presentará por medio de un proceso Poisson:

Definición 2.2 *Dado un proceso Poisson con parámetro λ, se designa con cero el momento en que se empieza a observar el proceso. Sea T el tiempo que transcurre hasta que ocurre el primer evento. Se llama variable aleatoria exponencial con parámetro λ.*

Ya que el tiempo se mide en forma continua (y positiva), de inmediato se puede ver que T es una variable aleatoria continua que tiene su rango dentro de los enteros positivos. La forma de la función de distribución y de la función de densidad para T está dada por el siguiente teorema.

Teorema 2.5 *Suponga que T es una variable aleatoria exponencial con parámetro λ. Entonces*

$$F_T(s) = \begin{cases} 0, & si \ \ s < 0 \\ 1 - e^{-\lambda s}, & si \ \ s \geq 0 \end{cases}$$

y

$$f_T(s) = \begin{cases} \lambda e^{-\lambda s}, & si \ \ s > 0 \\ 0, & c.o.c. \end{cases}$$

Demostración: Ya que el tiempo se mide en forma positiva, se ve de inmediato que $\Pr(T \leq s) = 0$; si $s < 0$; por tanto $F_T(s) = 0$. $s < 0$. Sin embargo, sea cierto tiempo $s \geq 0$. El tiempo transcurrido para el primer evento *excede* a s si y solo si no hay eventos en el intervalo $(0, s)$. Pero ya se ha visto que la probabilidad de que no haya eventos en $(0, s)$ es $e^{-\lambda s}$ (ver la ecuación (2.21), con $k = 0$); por lo tanto $\Pr(T > s) = e^{-\lambda s}$. Pero $\Pr(T \leq s) = 1 - \Pr(T > s)$ lo que da $F_T(s) = 1 - e^{-\lambda s}$ para $s \geq 0$. La función de densidad para T está dada por la derivada de $F_T(s)$; ya que

$$\frac{d}{ds} F_T(s) = \begin{cases} \lambda e^{-\lambda s}, & s > 0 \\ 0, & s < 0 \end{cases}$$

se tiene

$$f_T(s) = \begin{cases} \lambda e^{-\lambda s}, & s > 0 \\ 0, & s \leq 0, \end{cases} \quad \square$$

Teorema 2.6 *Si T es una variable aleatoria exponencial, entonces*

$$\Pr(T > a + b / T > a) = \Pr(T > b).$$

Demostración: Primero se aclara la notación que se emplea. El evento de que $T > a$ se denota por A; el evento de que $T > b$, se denota con B, y el evento de que $T > a + b$ se denota con C. Se quiere demostrar que $\Pr(C/A) = \Pr(B)$. Por definición, $\Pr(C/A) = \Pr(C \cap A)/\Pr(A)$. El evento C, en que $T > a + b$, es un subconjunto del evento A, en que $T > a$. Por tanto, $C \cap A = C$ con lo que se tiene $\Pr(C/A) = \Pr(C)/\Pr(A)$. Luego,

$$\Pr(A) = \Pr(T > a) = \int_a^\infty \lambda e^{-\lambda s} ds = e^{-\lambda a}$$

$$\Pr(B) = \Pr(T > b) = \int_b^\infty \lambda e^{-\lambda s} des = e^{-\lambda b}$$

$$\Pr(C) = \Pr(T > a + b) = \int_{a+b}^\infty \lambda e^{-\lambda s} ds = e^{-\lambda(a+b)}$$

con lo que se puede ver que

$$\Pr(C/A) = \frac{\Pr(C)}{\Pr(A)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = \Pr(B). \quad \square$$

Se puede suponer que el tiempo para que falle un foco es una variable aleatoria exponencial. Entonces el Teorema 2.6 muestra que si el foco ha sido usado durante, supongase, 100 horas sin fallar, la probabilidad de que dure otras 50 horas es la misma de que durara 50 horas cuando estaba nuevo. Por lo tanto podemos decir que la variable aleatoria exponencial no tiene memoria; el hecho de que haya alcanzado determinado valor no afecta la probabilidad de que obtenga un valor mayor. Esta propiedad es una consecuencia de las propiedades de independencia que se supucieron: los eventos de Poisson ocurren en forma independiente.

2.3.4 Procesos No-Homogéneos

La tasa λ de un proceso Poisson $X(t)$ es la constante proporcional en la probabilidad de que un evento ocurra durante un intervalo de tiempo muy pequeño (arbitrario). Para explicar esto más precisamente, se tiene:

$$\begin{aligned}\Pr\{X(t+h)-X(t)=1\} &= \frac{(\lambda h)e^{-\lambda h}}{1!}\\ &= (\lambda)(1-\lambda h+\frac{1}{2}\lambda^2h^2\cdots)\\ &= \lambda h+o(h),\end{aligned}$$

donde $o(h)$ denota al sobrante inespecífico de orden más pequeño que h es decir, $[\frac{o(h)}{h} \to 0$ cuando $h \to 0]$.

En muchas aplicaciones, se considera la tasa $\lambda = \lambda(t)$ (que varía con respecto al tiempo). Un proceso es considerado como proceso Poisson *no-homogéneo* o *no-estacionario*, para distinguirlo de los estacionarios u homogéneos que se manejaron anteriormente. Si $X(t)$ es un proceso Poisson no-homogéneo con tasa $\lambda(t)$, entonces el incremento $X(t) - X(s)$, dado el número de eventos en un intervalo $(s,t]$, tiene una distribución Poisson con parámetro $\int_s^t \lambda(u)du$, y los incrementos sobre intervalos diferentes son independientes.

Ejemplo 2.2 *Las demandas de un servicio de primeros auxilios en cierto Hospital, ocurre de acuerdo con un proceso de Poisson no-homogéneo, cuya tasa tiene la siguiente distribución,*

$$\lambda(t) = \begin{cases} 2t & si\ 0 \le t < 1\\ 2 & si\ 1 \le t < 2\\ 4t & si\ 2 \le t < 4\end{cases}$$

donde t se mide en horas. ¿Cuál es la probabilidad de que dos demandas ocurran en las primeras dos horas de operación y dos en las segundas dos horas ?

Dado que la demanda en intervalos diferentes es independiente se puede contestar a estas dos preguntas separadamente.

La media de las primeras dos horas es, $\mu = \int_0^1 2tdt + \int_1^2 2dt = 3$, y por esto,

$$\Pr\{X(2)=2\} = \frac{e^{-3}(3)^2}{2!} = .2240.$$

para las segundas dos horas, $\mu = \int_2^4 (4-t)dt = 2$ *y*

$$\Pr\{X(4) - X(2) = 2\} = \frac{e^{-2}(2)^2}{2!} = .2707.$$

2.3.5 Proceso Poisson Marcado.

Ahora suponga que una variable aleatoria Y_k se asocia con el k-ésimo evento en un proceso de Poisson de tasa λ. Sea $Y_1, Y_2, \cdots$; independientes y que comparten una función de distribución común,

$$G(\gamma) = \Pr\{Y_k \leq \gamma\}.$$

La secuencia de parejas $(W_1, Y_1), (W_2, Y_2), \cdots$; es llamada *proceso de Poisson marcado.*

Se iniciará el análisis de los procesos de Poisson marcados con uno de los casos más simples. Para un valor fijo $p\,(0 < p < 1)$ supongase

$$\Pr\{Y_k = 1\} = p, \qquad \Pr\{Y_k = 0\} = q = 1 - p.$$

Ahora considere separadamente el proceso de puntos marcados con unos y el de los marcados con ceros. En este caso se define al proceso Poisson relevante explicado por

$$X_1(t) = \sum_{k=1}^{X(t)} Y_k \quad y \quad X_0(t) = X(t) - X_1(t).$$

Los incrementos no traslapados en $X_1(t)$ son variables aleatorias independientes, $X_1(0) = 0$, y finalmente aplicando el Teorema 2.4, se obtiene que $X_1(t)$ tiene una distribución Poisson con media λpt. En resumen, $X(t)$ es un proceso Poisson con tasa λp, y con un argumento paralelo se puede mostrar que $X_0(t)$ es un proceso Poisson con tasa $\lambda(1-p)$. *Lo que es más interesante es que* $X_0(t)$ *y* $X_1(t)$ *son procesos independientes.* La propiedad relevante a probar es que $\Pr\{X_0(t) = j \text{ y } X_1(t) = k\} = \Pr\{X_0(t) = j\} \times \Pr\{X_1(t) = k\}$ para $j, k = 0, 1, \cdots$. Se estable esta independencia escribiendo

$$\begin{aligned}
\Pr\{X_0(t) = j, X_1(t) = k\} &= \Pr\{X(t) = j + k, X_1(t) = k\} \\
&= \Pr\{X_1(t) = k / X(t) = j + k\} \Pr\{X(t) = j + k\} \\
&= \frac{(j+k)!}{j!k!} p^k (1-p)^j \frac{(\lambda t)^{j+k} e^{-\lambda t}}{(j+k)!}
\end{aligned}$$

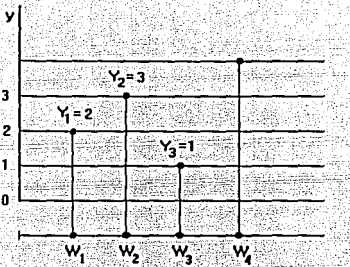

Figura 2.1: *Un proceso de Poisson Marcado. $W_1, W_2, \cdots$ son los los tiempos de los eventos en un proceso de Poisson de tasa λ. Las variables aleatorias $Y_1, Y_2 \cdots$ son las marcas, suponiendo que sean independientes e idénticamente distribuidas, e independientes del proceso Poisson.*

$$= \left[\frac{e^{-\lambda p t}(\lambda p t)^k}{k!}\right]\left[\frac{e^{-\lambda(1-p)t}(\lambda(1-p)t)^j}{j!}\right]$$
$$= \Pr\{X_1(t) = k\}\Pr\{X_0(t) = j\}$$
$$\text{para } j, k = 0, 1, \cdots$$

En la Figura 2.1, los tiempos Poissones originales de los eventos $W_1, W_2, \cdots$ se muestran en el eje inferior. Entonces un punto es un lugar (t, γ) en el plano (W_n, Y_n) para cualquer n. Para cualquier entero $k = 0, 1, 2, \cdots$ uno obtiene un punto de proceso que corresponde a los tiempos W_n para los cuales $Y_n = k$. El mismo razonamiento que en el caso cero–uno se aplica para implicar que cada uno de estos procesos es Poisson, la tasa para el k-ésimo proceso λa_k, y el proceso para valores distintivos de k son independientes.

Para establecer el resultado correspondiente cuando los valores $Y_1, Y_2, \cdots$; son variables aleatorias continuas, se requiere un alto grado de sofisticación, aunque las ideas preponderantes son básicamente las mismas. Para establecer formalmente, definimos lo que llamamos media para un proceso Poisson no homogéneo puntual en el plano, esto, aplicando la teoría de los procesos no homogéneos de la sección anterior. Sea $\theta = \theta(x, y)$ una función no negativa definida sobre la región S en el plano (x, y), para cada subconjunto A del conjunto S,

sea $\mu(A) = \int\int_A \theta(x,y)dxdy$ el volumen sobre $\theta(x,y)$ sobre A. Un proceso puntual no homogéneo Poisson con función de intensidad $\theta(x,y)$ es un proceso puntual $\{N(A); A \subset S\}$ para el cual

- para cada subconjunto A de S, la variable aleatoria $N(A)$ tiene una distribución Poisson con media $\mu(A)$; y
- para conjuntos disconjuntos $A_1, \cdots, A_m$ de S, las variables aleatorias $N(A_1), \cdots, N(A_m)$ son independientes.

Se observa facilmente que un proceso puntual homogeneo Poisson con intensidad λ constante corresponde a la función $\theta(x,y)$, y $\theta(x,y) = \lambda$ para toda x, y.

Con esta definición a la mano, se dará la descomposición apropiada del resultado para procesos de Poisson Marcados generales.

Teorema 2.7 *Sea $(W_1, Y_1); (W_2, Y_2), \cdots$; un proceso de Poisson marcado donde $W_1, W_2, \cdots$; son tiempos de espera en un proceso de Poisson de tasa λ y $Y_1, Y_2, \cdots$; son variables aleatorias independientes, continuas e idénticamente distribuidas con función de densidad de probabilidad $g(y)$. Entonces $(W_1, Y_1), (W_2, Y_2), \cdots$; forma un proceso bidimensional no homogéneo puntual de Poisson en el plano (t, y), donde el número promedio de puntos en la región A está dado por:*

$$\mu(A) = \int\int_A \lambda g(y)dydt \tag{2.23}$$

La Figura 2.2 muestra el diagrama.

El Teorema 2.7 afirma que el número de puntos en intervalos disconjuntos son variables aleatorias independientes. Por ejemplo, el tiempo de espera correspondiente a los valores positivos $Y_1, Y_2, \cdots$; forma un proceso Poisson, en la forma en que asocia los tiempos con valores negativos, y estos dos procesos son independientes.

2.4 Cadenas de Markov Continuas

En esta sección se presentarán algunos aspectos importantes de las cadenas de Markov, de tiempo continuo y estados discretos. Específicamente se tratará con la familia de variables aleatorias $\{X(t); 0 \leq t < \infty\}$ donde los valores posibles de $X(t)$ son los enteros no-negativos. Se debe observar el caso en el que, $\{X(t)\}$ es un proceso de Markov con

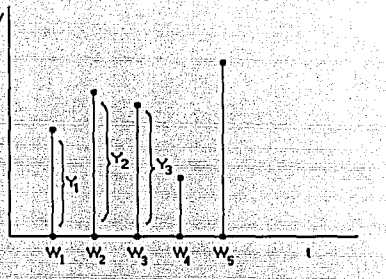

Figura 2.2: *Un proceso de Poisson marcado.*

probabilidades de transición estacionarias. Así, la función transición de probabilidad para $t > 0$ es :

$$P_{ij}(t) = \Pr\{X(t+u) = j / X(u) = i\}, \qquad i, j = 0, 1, 2, \cdots,$$

que es independiente de $u \geq 0$.

Esto usualmente se usa en la investigación de modelos de procesos estocásticos basados en fenómenos físicos. En este caso se postulará la forma de $P_{ij}(h)$ para h pequeña y usando la propiedad de Markov, derivando un sistema de ecuaciones diferenciales que satisfagan $P_{ij}(t)$ para toda $t > 0$. La solución de estas ecuaciones sobre un conjunto acotado de condiciones da $P_{ij}(t)$. Hay que tener presentes los postulados de los fenómenos Poisson para generar el proceso de nacimiento puro.

2.4.1 Proceso de Nacimiento Puro

Una generalización de proceso natural es permitir la oportunidad de que un evento ocurra en un instante dado del tiempo dependiendo de los eventos que han ocurrido. Un ejemplo de este fenómeno es la reproducción de seres vivos, en el cual bajo ciertas condiciones (comida suficiente, no mortalidad, no migración,etc.), la probabilidad infinitesimal de que haya un nacimiento en un instante dado es directamente proporcional al tamaño de la población en

ese instante[1].

Considerese una secuencia de números positivos, $\{\lambda_k\}$. Se define *el proceso de nacimiento puro como proceso de Markov que satisface los siguientes postulados*

i) $\Pr\{X(t+h) - X(t) = 1/X(t) = k\} = \lambda_k h + o_{1k}(h), (h \to 0+)$

ii) $\Pr\{X(t+h) - X(t) = 0/X(t) = k\} = 1 - \lambda_k h + o_{2k}(h)$,

iii) $\Pr\{X(t+h) - X(t) < 0/X(t) = k\} = 0, (k \geq 0)$.

Frecuentemente se postula,

iv) $X(0) = 0$.

Con estos postulados $X(t)$ no denota el tamaño de la población, pero si el número de nacimientos en el intervalo de tiempo $(0, t]$. notese que el lado derecho de (i) y de (ii) son $P_{k,k+1}(h)$ y $P_{k,k}(h)$, respectivamente, así $o_{1k}(h)$ y $o_{2k}(h)$ no dependen de t.

Definimos a $P_n(t) = \Pr\{X(t) = n\}$, suponiendo $X(0) = 0$.

De la misma forma como en el proceso de Poisson, se derivará un sistema de ecuaciones diferenciales,

$$\begin{aligned} P_0'(t) &= -\lambda_0 P_0(t), \\ P_n'(t) &= -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t) \text{ para } n \geq 1, \end{aligned} \tag{2.24}$$

con condiciones iniciales

$$P_0(0) = 1; \qquad P_n(0) = 0, \qquad n > 0.$$

De hecho, si $h > 0$, y $n \geq 1$, entonces por la ley de probabilidad total,la propiedad de Markov y el postulado (iii) se tiene:

$$\begin{aligned} P_n(t+h) &= \sum_{k=0}^{\infty} P_k(t) \Pr\{X(t+h) = n/X(t) = k\} \\ &= \sum_{k=0}^{\infty} P_k(t) \Pr\{X(t+h) - X(t) = n-k/X(t) = k\} \\ &= \sum_{k=0}^{n} P_k(t) \Pr\{X(t+h) - X(t) = n-k/X(t) = k\}. \end{aligned}$$

[1] Este ejemplo es conocido como el proceso de Yule, que será considerado posteriormente.

Ahora, para $k = 0, 1, \cdots, n-2$ se tiene:

$$\begin{aligned}&\Pr\{X(t+h) - X(t) = n-k / X(t) = k\}\\ &\leq \Pr\{X(t+h) - X(t) \geq 2 / X(t) = k\}\\ &= o_{1k}(h) + o_{2k}(h)\end{aligned}$$

ó

$$\Pr\{X(t+h) - X(t) = n-k / X(t) = k\} = o_{3nk}(h)$$

$$k = 0, 1, \cdots, n-2.$$

Así

$$\begin{aligned}&P_n(t+h)\\ &= P_n(t)[1 - \lambda_n h + o_{2n}h] + P_{n-1}(t)[\lambda_{n-1}h + o_{1n-1}(h)] + \sum_{k=0}^{n-2} P_k(t) o_{3nk}(h)\end{aligned}$$

ó

$$\begin{aligned}&P_n(t+h) - P_n(t)\\ &= P_n(t)[-\lambda_n h + o_{2n}(h)] + P_{n-1}(t)[\lambda_{n-1}h + o_{1n-1}(h)] + o_n(h), \qquad (2.25)\end{aligned}$$

donde claramente, $\lim_{h\to 0} \frac{o_n(h)}{h} = 0$ con $t \geq 0$ y $o_n(h)$ esta acotada por la suma $\sum_{k=0}^{n-2} o_{3nk}(h)$ que no depende de t.

Dividiendo por h y tomando el límite cuando $h \to 0$, validamos la relación (2.24) donde, el lado izquierdo se debó haber escrito en forma de derivada. De hecho de (2.25) se observa que una vez que $P_n(t)$ es una función continua de t. Se reemplaza t por $t-h$ en (2.24), dividiendo por h, y tomando el límite cuando $h \to 0$, se encuentra que cada $P_n(t)$ es una derivada y también satisface la ecuación (2.25).

La primera ecuación de (2.24) puede ser resuelta inmediatamente y

$$P_0(t) = e^{-\lambda_0 t} \quad \text{para } t > 0. \qquad (2.26)$$

Define a S_k como el tiempo entre el k y $(k+1)$-ésimo nacimiento, así

$$P_n(t) = \Pr\left\{\sum_{i=0}^{n-1} S_i \leq t \leq \sum_{i=0}^{n} S_i\right\}$$

Las variables aleatorias S_k son llamadas "tiempo de estancia" entre nacimientos y

$$W_k = \sum_{i=0}^{k-1} S_i = \text{el tiempo en el que ocurre el } k\text{-ésimo nacimiento.}$$

Se ha observado que $P_0(t) = e^{-\lambda_0 t}$. Por esto

$$\Pr\{S_0 \leq t\} = 1 - \Pr\{X(t) = 0\} = 1 - e^{\lambda_0 t}.$$

Es decir, S_0 tiene una distribución exponencial con parámetro λ_0.

Debe deducirse de los postulados (i) y (iv) que S_k, $k > 0$, también tiene una distribución exponencial con parámetro λ_k y que las S_i son también independientes.

Esta distribución caracteriza el proceso de nacimiento puro en término de sus tiempos de estancia.

Para resolver recursivamente las ecuaciones diferenciales de (2.24) se introduce:

$$\begin{aligned} Q_n(t) &= e^{\lambda_n t} P_n(t) \text{para } n = 0, 1, \cdots \text{ entonces} \\ Q_n'(t) &= \lambda_n e^{\lambda_n t} P_n(t) + e^{\lambda_n t} P_n'(t). \\ &= e^{\lambda_n t}[\lambda_n P_n(t) + P_n'(t)]. \\ &= e^{\lambda_{n-1} t} P_{n-1}(t); \qquad \text{[usando (2.25)].} \end{aligned}$$

Integrando ambos lados de la ecuación y usando las condiciones de frontera $Q_n(0) = 0$ para $n \geq 1$ se tiene:

$$Q_n(t) = \int_0^t e^{\lambda_{n-1} x} \lambda_{n-1} P_{n-1}(x) dx,$$

ó

$$P_n(t) = \lambda_{n-1} e^{\lambda_n t} \int_0^t e^{\lambda_n x} P_{n-1}(x) dx, \qquad n = 1, 2, \cdots \tag{2.27}$$

ahora es claro que toda $P_k(t) \geq 0$, pero existe la posibilidad de que:

$$\sum_{n=0}^{\infty} P_n(t) < 1.$$

Para asegurar que el proceso sea valido, es decir asegurar que $\sum_{n=0}^{\infty} P_n(t) = 1$ para toda t, se debe restringir el valor de λ_k de acuerdo con la siguiente ecuación

$$\sum_{n=0}^{\infty} P_n(t) = 1 \text{ si y sólo si } \sum_{n=0}^{\infty} \frac{1}{\lambda_n} = \infty. \tag{2.28}$$

El argumento intuitivo para este resultado es que el tiempo S_k entre nacimientos consecutivos se distribuye exponencialmente con un parámetro correspondiente λ_k. Por ello $\sum_n \frac{1}{\lambda_n}$ es igual al tiempo esperado antes de que la población sea infinita. Comparativamente $1 - \sum_{n=0}^{\infty} P_n(t)$ es la probabilidad de que $X(t) = \infty$. Si $\sum_n \frac{1}{\lambda_n} < \infty$ el tiempo esperado para que la población sea infinita es finito. Entonces es posible que para toda $t > 0$ la probabilidad de que $X(t) = \infty$ sea positiva.

Cuando dos de los parámetros $\lambda_0, \lambda_1, \cdots$ no son iguales, la ecuación diferencial (2.27) puede ser resuelta apartir de la siguiente fórmula:

$$\begin{aligned} P_0(t) &= e^{-\lambda_0 t} \\ P_1(t) &= \lambda_0\left(\frac{1}{\lambda_1 - \lambda_0}e^{-\lambda_0 t} + \frac{1}{\lambda_0 - \lambda_1}e^{-\lambda_1 t}\right) \end{aligned} \tag{2.29}$$

y

$$\begin{aligned} P_n(t) &= \Pr\{X(t) = n / X(0) = 0\} \\ &= \lambda_0 \cdots \lambda_{n-1}[B_{0,n}e^{-\lambda_0 t} + \cdots + B_{n,n}e^{-\lambda_n t}] \quad \text{para } n > 1 \end{aligned} \tag{2.30}$$

donde,

$$B_{0,n} = \frac{1}{(\lambda_1 - \lambda_0)\cdots(\lambda_n - \lambda_0)},$$

$$B_{k,n} = \frac{1}{(\lambda_0 - \lambda_k)\cdots(\lambda_{k-1} - \lambda_k)(\lambda_{k+1} - \lambda_k)\cdots(\lambda_n - \lambda_k)}$$

para $0 < k < n$ y

$$B_{n,n} = \frac{1}{(\lambda_0 - \lambda_n)\cdots(\lambda_{n-1} - \lambda_n)}$$

Porque como se supuso $\lambda_j \neq \lambda_k$ cuando $j \neq k$, el denominador de (2.30) no desaparece y $B_{k,n}$ está bien definido.

Ahora se verificará que $P_1(t)$, en la forma en que se da en (2.29), satisface (2.27). La ecuación (2.26) da $P_0(t) = e^{-\lambda_0 t}$. Después al sustituir esto en (2.27) con $n = 1$ se obtiene:

$$P_1(t) = \lambda_0 e^{-\lambda_1 t}\int_0^t e^{\lambda_1 r}e^{-\lambda_0 r}dr$$

$$= \lambda_0 e^{-\lambda_1 t}(\lambda_0 - \lambda_1)^{-1}[1 - e^{-(\lambda_0 - \lambda_1)t}]$$
$$= \lambda_0 \left(\frac{1}{\lambda_1 - \lambda_0} e^{-\lambda_0 t} + \frac{1}{\lambda_0 - \lambda_1} e^{-\lambda_1 t}\right),$$

que concuerda con (2.29).

2.4.2 Proceso de Yule

El proceso de Yule describe el crecimiento de una población en la cual cada miembro tiene una probabilidad $\beta h + o(h)$ de dar vida a un nuevo miembro durante un intervalo de tiempo de longitud h ($\beta > 0$). Suponiendo independencia y no interacción de los miembros de la población, por el teorema del binomio se expresa:

$$\Pr\{X(t+h) - X(t) = 1 / X(t) = n\} = \binom{n}{1}[\beta h + o(h)][1 - \beta h + o(h)]^{n-1}$$
$$= n\beta h + o(h),$$

es decir, para el proceso de Yule los parámetros son $\lambda_k = n\beta$. Lo que significa que la total de tasa de nacimiento es directamente proporcional al tamaño de la población , la constante de proporcionalidad está dada por la tasa individual de nacimiento β. Así, el proceso de Yule forma una analogía estocástica del modelo determinístico de crecimiento de población definido por la ecuación diferencial $\frac{dy}{dt} = \alpha y$. En el modelo determinístico, la tasa $\frac{dy}{dt}$ de crecimiento de la población es directamente proporcional al tamaño de la población y. En el modelo estocástico, el crecimiento infinitesimal determinístico dy es sustituido por la probabilidad de un incremento unitario durante el intervalo de tiempo infinitesimal dt.

El sistema de ecuaciones (2.24) en el caso en que $X(0) = 1$ se convierte

$$P_n'(t) = -\beta[nP_n(t) - (n-1)P_{n-1}(t)], \qquad n = 1, 2, \cdots,$$

sobre las condiciones iniciales

$$P_1(0) = 1, \qquad P_n(0) = 0, \qquad n = 2, 3, \cdots$$

y la solución es

$$P_n(t) = e^{-\beta t}(1 - e^{-\beta t})^{n-1}, \qquad n \geq 1. \tag{2.31}$$

La ecuación (2.31) corresponde a una distribución geométrica con $p = e^{-\beta t}$.

La solución general análoga a (2.30) pero por proceso de nacimiento puro iniciando con $X(0) = 1$ es

$$P_n(t) = \lambda_1 \cdots \lambda_{n-1}[B_{1,n}e^{-\lambda_1 t} + \cdots + B_{n,n}e^{-\lambda_n t}], \quad n > 1 \tag{2.32}$$

Cuando $\lambda_n = \beta n$, se mostrará que (2.32) se reduce a la solución dada (2.31) para un proceso de Yule con parámetro β.

$$\begin{aligned} B_{1,n} &= \frac{1}{(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)\cdots(\lambda_n - \lambda_1)} \\ &= \frac{1}{\beta^{n-1}1(2)\cdots(n-1)} \\ &= \frac{1}{\beta^{n-1}(n-1)!} \\ B_{2,n} &= \frac{1}{(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2)\cdots(\lambda_n - \lambda_2)} \\ &= \frac{1}{\beta^{n-1}(-1)(1)(2)\cdots(n-2)} \\ &= \frac{-1}{\beta^{n-1}(n-2)!} \end{aligned}$$

y

$$\begin{aligned} B_{k,n} &= \frac{1}{(\lambda_1 - \lambda_k)\cdots(\lambda_{k-1} - \lambda_k)(\lambda_{k+1} - \lambda_k)\cdots(\lambda_n - \lambda_k)} \\ &= \frac{(-1)^{k-1}}{\beta^{n-1}(k-1)!(n-k)!} \end{aligned}$$

entonces, de acuerdo con (2.32),

$$\begin{aligned} P_n(t) &= B^{n-1}(n-1)!(B_{1,n}e^{-\beta t} + \cdots + B_{n,n}e^{-n\beta t} \\ &= \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!}(-1)^{k-1}e^{-k\beta t} \\ &= e^{-\beta t}\sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!}(-e^{-\beta t})^j \\ &= e^{-\beta t}(1 - e^{-\beta t})^{n-1} \end{aligned}$$

que establece (2.31).

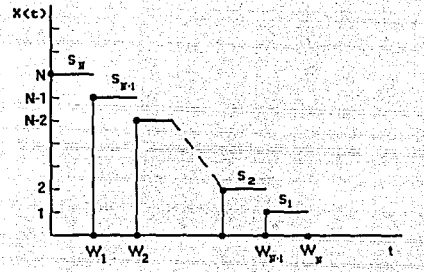

Figura 2.3: *Un modelo gráfico de un proceso de muerte, muestra los tiempos de estancia* $S_N, \cdots, S_1$ *y los tiempos de espera* $W_1, W_2, \cdots, W_N$

2.4.3 Proceso Puro de Muerte.

Complementariamente al proceso creciente de nacimiento puro está el proceso decreciente puro de muerte. Este se mueve sobre los estados $N, N-1, \cdots, 2, 1$ y finalmente es absorvido en el estado 0 (extinción). El proceso es especificado por el parámetro $\mu_k > 0$ para $k = 1, 2, \cdots, N$, donde el tiempo de estancia en el estado k se distribuyen exponencialmente con parámetro μ_k, todos los tiempos de estancia son independientes. Una muestra sencilla de que es a lo que nos referimos a la Figura 2.3

Alternativamente, se tiene el proceso de muerte puro como proceso de Markov $X(t)$ cuyo espacio de estados es $0, 1, \cdots, N$ y para los cuales

(i) $\Pr\{X(t+h) = k-1 / X(t) = k\} = \mu_k h + 0(h), k = 1, \cdots, N$

(ii) $\Pr\{X(t+h) = k / X(t) = k\} = 1 - \mu_k h + o(h), k = 1, \cdots, N$

(iii) $\Pr\{X(t+h) > k / X(t) = k\} = 0. k = 1, \cdots, N$

El parámetro μ_k es la tasa de mortalidad en operación mientras el proceso se encuentra en el estado k. Comunmente se asigna a $\mu_0 = 0$.

Cuando los parámetros $\mu_1, \mu_2, \cdots, \mu_N$ son distintos, es decir que $\mu_j \neq \mu_k$ si $j \neq k$, entonces tenemos las siguientes probabilidades de transición:

$$P_N(t) = e^{-\mu_N t},$$

y para $n < N$,

$$P_n(t) = \Pr\{X(t) = n / X(0) = N\}$$
$$= \mu_{n+1}\mu_{n+2}\cdots\mu_N\{A_{n,n}e^{-\mu_n t} + \cdots + A_{N,n}e^{-\mu_N t}\} \quad (2.33)$$

donde

$$A_{k,n} = \frac{1}{(\mu_N - \mu_k)\cdots(\mu_{k+1} - \mu_k)(\mu_{k-1} - \mu_k)\cdots(\mu_n - \mu_k)}$$

2.4.4 Proceso de Nacimiento y Muerte

La generalización es ahora inminente, teniendo el proceso de nacimiento puro y el proceso de muerte puro, permitiendo que $X(t)$, tenga la propiedad de incrementar la población y de disminuirla también. Así, si el proceso en el tiempo t se encuentra en el estado n, es posible, después de un cierto período de tiempo, ir al estado $n+1$ o $n-1$ y el resultado es un *proceso de nacimiento y muerte*.

El *proceso de nacimiento y muerte* brinda una herramienta fuerte, en la sección anterior se analizó el proceso de Poisson $\{X(t); t \geq 0\}$, el cual contaba el número de ocurrencias de algún tipo de evento, que podía ser interpretado también como las llegadas a un determinado lugar con una tasa promedio λ. Ahora se puede pensar en una llegada como un nacimiento, para un proceso Poisson la probabilidad de un nacimiento en un intervalo pequeño de tiempo h es $\lambda h e^{-\lambda h} = \lambda h + o(h)$ y esta probabilidad es independiente de los nacimientos que haya habido anteriormente, λ puede ser vista como la tasa de nacimiento. Para cualquier sistema es razonable suponer que la tasa de nacimiento depende del número de habitantes en la población presente, es decir, la probabilidad de un nacimiento en un período corto de tiempo h debe ser $\lambda_n h + o(h)$, donde n es el tamaño de la población y la tasa de nacimiento λ_n depende de este número. También podemos pensar en las muertes o decrementos de la población, con una probabilidad de muerte en un intervalo de longitud h igual a $\mu_n h + o(h)$. De esta forma la idea intuitiva de un *proceso de nacimiento y muerte* es que algún tipo de población está simultáneamente ganando y perdiendo miembros a través de nacimientos y muerte, por ejemplo, la población humana de este planeta. La población que se tiene en mente para las aplicaciones del proceso de nacimiento y muerte son los clientes en una línea de espera, las llegadas corresponden a los nacimientos y las salidas (después de recibir un servicio) corresponden a las muertes.

Definición 2.3 *Se considerará un proceso estocástico $\{X(t); t \geq 0\}$ con espacio de estado discreto $0, 1, 2, \cdots$, supongase que este proceso describe un sistema que se encuentra en el*

estado i_n, $n = 0, 1, 2, \cdots$, en el tiempo t, si y sólo si $X(t) = n$ (el sistema tiene una población de n elementos o clientes en el momento t).

Entonces se dice que el sistema está descrito por un proceso de nacimiemto y muerte si existen tasas de nacimiento no negativas $\{\lambda_n, n = 0, 1, 2, \cdots\}$ y tasa de mortalidad no negativas $\{\mu_n, n = 1, 2, 3, \cdots\}$. De acuerdo con los siguientes postulados [2]

i) Los cambios de estado (estando en el estado i_n) sólo pueden ser a los estados i_{n+1} ó i_{n-1} si $n \geq 1$, porque estando en i_0 únicamente se puede ir al estado i_1.

ii) Si en el momento t el sistema se encuentra en el estado i_n, la probabilidad de que entre el tiempo t y $t+h$ la transición del estado i_n al estado i_{n+1} ocurra (es decir $i_n \to i_{n+1}$) es igual a $\lambda_n h + o(h)$, y la probabilidad de que la transición $i_n \to i_{n-1}$ ocurra (si $n \geq 1$) es igual a $\mu_n h + o(h)$.

iii) La probabilidad de que en el intervalo de tiempo $(t, t+h)$ ocurra más de una transición es $o(h)$.

El postulado i) dice que solamente puede ocurrir una muerte o un nacimiento a la vez y que no puede ocurrir una muerte si el estado se encuentra vacío. El postulado ii) da las probabilidades de transición, es decir la probabilidad de un nacimiento o una muerte en un intervalo pequeño de tiempo, cuando la población es n. El último postulado establece que cuando la probabilidad de más de un nacimiento o una muerte en un intervalo de tiempo pequeño es despreciable.

Cuando describimos un sistema de líneas de espera como un proceso de nacimiento y muerte, pensamos en el estado i_n como n clientes en el sistema, ya sea esperando o recibiendo un servicio.

Ahora derivando las ecuaciones diferenciales para $P_n(t) = \Pr\{X(t) = n\}$, la probabilidad de que el sistema se encuentre en el estado i_n en el tiempo t.

Si $n \geq 1$, la probabilidad de que en el momento $t + h$ el sistema se encuentre en el estado i_n ($P_n(t + h)$) tiene cuatro componentes.

1) La probabilidad de que estando en el estado n en el momento t no ocurran transiciones, es decir ni muertes ni nacimientos. Esta probabilidad es el producto de $P_n(t)$, la probabilidad de que la transición $i_n \to i_{n+1}$ no ocurra

[2] Estos postulados son conocidos también como los supuestos de "vecinos más cercanos".

o $(1-\lambda_n h+o(h))$ y la probabilidad de que la transición $i_n \to i_{n-1}$ tampoco ocurra o $(1-\mu_n h+o(h))$, es decir:

$$\begin{aligned} & P_n(t)(1-\lambda_n h+o(h))(1-\mu_n h+o(h)) \\ &= P_n(t)[1-\mu_n h+o(h)-\lambda_n h+\lambda_n\mu_n h^2-\lambda_n h o(h)+o(h)] \qquad (2.34) \\ &= P_n(t)[1-\mu_n h-\lambda_n h+o(h)] = P_n(t)(1-\lambda_n h-\mu_n h)+o(h). \end{aligned}$$

ya que

$$o(h)(1-\mu_n h+o(h)) = o(h)$$

$$\lambda_n\mu_n h^2-\lambda_n h o(h)+o(h) = o(h) \text{ y } P_n(t)o(h)=o(h).$$

2) La probabilidad de que el sistema estando en el estado i_{n-1} al momento t, $(P_{n-1}(t))$ por la probabilidad de que la transición $i_{n-1} \to i_n$ ocurra en el intervalo de tiempo $(t, t+h)$, es decir

$$P_{n-1}(t)(\lambda_{n-1}h+o(h))=P_{n-1}(t)\lambda_{n-1}h+o(h). \qquad (2.35)$$

3) La probabilidad de que el sistema estando en el estado i_{n+1} en el momento t, por la probabilidad de que la transición $i_{n+1} \to i_n$ ocurra en el intervalo $(t, t+h)$, es decir,

$$P_{n+1}(t)\mu_n h+o(h) \qquad (2.36)$$

4) La probabilidad de que dos o más transiciones ocurran entre los tiempos t y $t+h$. Por hipótesis esta probabilidad es $o(h)$.

Dado que los eventos anteriores son componentes mutuamente excluyentes el resultado es:

$$P_n(t+h)=[1-\lambda_n h-\mu_n h]P_n(t)+\lambda_{n-1}hP_{n-1}(t)+\mu_{n+1}hP_{n+1}+o(h). \qquad (2.37)$$

restando $P_n(t)$ y dividiendo por h tenemos:

$$\frac{P_n(t+h)-P_n(t)}{h}=-(\lambda_n+\mu_n)P_n(t)+\lambda_{n-1}P_{n-1}(t)+\mu_{n+1}P_{n+1}(t)+\frac{o(h)}{h}. \qquad (2.38)$$

tomando el límite cuando $h \to 0$ se obtiene la ecuación:

$$\frac{dP_n(t)}{dt}=-(\lambda_n+\mu_n)P_n(t)+\lambda_{n-1}P_{n-1}+\mu_{n+1}P_{n+1}(t). \qquad (2.39)$$

Esta ecuación es válida únicamente para $n \geq 1$, para $n = 0$ tenemos

$$\frac{P_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t), \tag{2.40}$$

si el estado inicial es i, entonces las condiciones iniciales están dadas por:

$$P_i(0) = 1 \text{ y } P_j(0) = 0 \text{ para } j \neq i. \tag{2.41}$$

El proceso de nacimiento y muerte depende de un conjunto infinito de ecuaciones diferenciales de las cuales (2.39) y (2.40) con las condiciones iniciales (2.41).

En términos generales, encontrar las soluciones dependientes del tiempo de un proceso de nacimiento y muerte es muy difícil. De cualquier forma, si $P_n(t)$ se aproxima a un valor constante p_n cuando $t \to \infty$ para cada n,entonces el sistema está en equilibrio estadístico. Ya que no se puede, en general, encontrar las soluciones de las ecuaciones diferenciales (2.39) y (2.40) analíticamente, tomando los límites cuando $t \to \infty$ en ambos lados de las ecuaciones y, usando el hecho de que $\lim_{t\to\infty} P_n(t) = p_n$, obteniendo el conjunto de ecuaciones diferenciales:

$$0 = \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} - (\lambda_n + \mu_n) p_n \; n \geq 1 \tag{2.42}$$

$$0 = \mu_1 p_1 - \lambda_0 p_0, \; n = 0. \tag{2.43}$$

Las ecuaciones anteriores nos llevan a

$$p_1 = (\frac{\lambda_0}{\mu_1}) p_0. \tag{2.44}$$

La ecuación (2.42) puede ser escrita como:

$$\mu_{n+1} p_{n+1} - \lambda_n p_n = \mu_n p_n - \lambda_{n-1} p_{n-1}, \; n \geq 1. \tag{2.45}$$

Si se define $g_n = \mu_n p_n - \lambda_{n-1} p_{n-1}$ para $n = 1, 2, \cdots$ se observa que (2.45) puede ser escrita como:

$$g_{n+1} = g_n, \qquad n \geq 1. \tag{2.46}$$

Dado que g_n es constante y, por (2.43), $g_1 = 0$, entonces $g_n = 0$ para toda n o (suponiendo que $\mu_n > 0$ para toda n)

$$p_{n+1} = \frac{\lambda_n}{\mu_{n+1}} p_n, \qquad n \geq 0. \tag{2.47}$$

Entonces iterativamente

$$p_1 = \frac{\lambda_0}{\mu_1} p_0, \; p_2 = \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0, \; p_3 = \frac{\lambda_2}{\mu_3} p_2 = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} p_0.$$

Por inducción se observa que:

$$p_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} p_0, \qquad n \geq 1. \tag{2.48}$$

Este resultado da la solución en términos de p_0, la probabilidad de que el sistema estando en el estado i_0 (sistema vacío). p_0 es determinado por la condición

$$\sum_{n=0}^{\infty} p_n = p_0 + p_1 + p_2 + \cdots = 1. \tag{2.49}$$

Si sustituimos (2.48) en (2.49) se obtiene:

$$p_0 \left(1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \cdots + \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} + \cdots\right) = 1. \tag{2.50}$$

Entonces, las probabilidades estacionarias (2.48) existen si la serie

$$S = 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \cdots + \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} + \cdots < \infty \tag{2.51}$$

(Se supone que λ_n y μ_n son no negativas). Cuando esto sucede, $p_0 = \frac{1}{S} > 0$, o la probabilidad de que el sistema este vacío es positiva. En el caso de un sistema de líneas de espera esto significa que los servidores atienden a todos los clientes. Por otra parte si S diverge, es un indicativo de que el sistema de líneas de espera es inestable, por que los arribos ocurren más rápido, en promedio, que los servicios. En términos prácticos se supondrá que los procesos de nacimiento y muerte que describen a un sistema de líneas de espera poseen probabilidades de transición estacionarias $\{p_n\}$ si y sólo si la serie S converge y entonces están dadas por (2.48) con $p_0 = \frac{1}{S}$.

2.5 Cadenas de Markov Continuas con Estado Finito

Una cadena de Markov continua en el tiempo $X(t)$ $(t > 0)$ es un proceso de Markov en los estados $0, 1, 2 \cdots$. Supongamos que las probabilidades de transición son estacionarias es decir,

$$P_{ij}(t) = \Pr\{X(t+s) = j / X(s) = i\}. \tag{2.52}$$

En esta sección se considerará sólo el caso donde el estado de espacios S es finito, etiquetado como $\{0, 1, 2, \cdots, N\}$.

La propiedad de Markov que P_{ij} satisface

(a) $P_{ij}(t) \geq 0$

(b) $\sum_{j=0}^{N} P_{ij}(t) = 1 \quad i, j = 0, 1, \cdots, N$ y

(c) $P_{ik}(s+t) = \sum_{j=0}^{N} P_{ij}(s) P_{jk}(t)$ para $t, s \geq 0$[3]

y se postula además que

(d)

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

Si $\mathbf{P}(t)$ denota a la matriz $(P_{ij}(t))_{i,j=0}^{N}$ entonces la propiedad (c) puede ser escrita en forma compacta en notación matricial

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s) \quad t, s \geq 0 \tag{2.53}$$

La propiedad (d) afirma que $\mathbf{P}(t)$ es continua en $t = 0$ dado que $\mathbf{P}(0) = \mathbf{I}$ (la matriz identidad) es empleada por (2.53). Pudiendo obtenerse de (2.53) que $\mathbf{P}(t)$ es continua para toda $t > 0$. De hecho $s = h > 0$ en (2.53) entonces por (d) se tiene:

$$\lim_{h \to 0+} \mathbf{P}(t+h) = \mathbf{P}(t) \lim_{h \to 0+} \mathbf{P}(h) = \mathbf{P}(t)\mathbf{I} = \mathbf{P}(t). \tag{2.54}$$

Por otro lado, para $t > 0$ y $0 < h < t$ se escribe (2.53) en la forma

$$\mathbf{P}(t) = \mathbf{P}(t-h)\mathbf{P}(h). \tag{2.55}$$

Pero $\mathbf{P}(h)$ es cercana a la identidad cuando h es suficientemente pequeña y por eso $\mathbf{P}(h)^{-1}$ (la inversa de $\mathbf{P}(h)$) existe, y también se aproxima a la identidad $\mathbf{I}$. Por ello

$$\mathbf{P}(t) = \mathbf{P}(t) \lim_{h \to 0+} (\mathbf{P}(h))^{-1} = \lim_{h \to 0+} \mathbf{P}(t-h) \tag{2.56}$$

Las relaciones límites (2.54) y (2.56) conjuntamente muestran que $\mathbf{P}(t)$ es continua.

[3]Esta es la relación de Chapman-Kolmogorov

En realidad $\mathbf{P}(t)$ no es solamente continua sino también diferenciable en los límites.

$$\lim_{h\to 0+} \frac{1-P_{ii}(h)}{h} = q_i,$$
$$\lim_{h\to 0+} \frac{P_{ij}(h)}{h} = q_{ij} \quad i \neq j, \tag{2.57}$$

que existen cuando $0 \leq q_{ij} < \infty$ $(i \neq j)$ y $0 \leq q_i < \infty$. Comenzando con la relación

$$1 = P_{ii}(h) + \sum_{j=0;\, j\neq i}^{N} P_{ij}(h),$$

dividiendo por h, y haciendo tender h a cero se tiene:

$$q_i = \sum_{j=0;\, j\neq i}^{N} q_{ij}$$

Las tasa q_i y q_{ij} forman una descripción infinitesimal del proceso con

$$\Pr\{X(t+h) = j / X(t) = i\} = q_{ij}h + o(h) \text{ para } i \neq j$$
$$\Pr\{X(t+h) = i / X(t) = i\} = 1 - q_i h + o(h).$$

En contraste con la descripción infinitesimal, la descripción del proceso se efectua como sigue:

Empezar en el estado i, el proceso trabaja con una duración que está exponencialmente distribuida con parámetro q_i. Entonces el proceso brinca al estado $i \neq j$ con probabilidad $p_{ij} = q_{ij}/q_i$; el tiempo en el estado j es exponencialmente distribuido con parámetro q_i, y así sucesivamente la secuencia de los estados visitados por el proceso, denotados por $\xi_0, \xi_1, \cdots$, forman una cadena de Markov con parámetros discretos, llamada *cadena de Markov acoplada*.

Condicionados sobre la secuencia de estados $\xi_0, \xi_1, \cdots$, los tiempos sucesivos $S_0, S_1, \cdots$ son variables aleatorias independientes, exponencialmente distribuidas con parámetros $q_{\xi_0}, q_{\xi_1}, \cdots$, respectivamente.

Suponiendo que (2.57) ha sido verificado y derivando una expresión explícita para

$P_{ij}(t)$ en términos de la matriz infinitesimal

$$\mathbf{A} = \begin{pmatrix} -q_0 & q_{01} & \cdots & q_{0N} \\ q_{10} & -q_1 & \cdots & q_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N0} & q_{N1} & \cdots & -q_N \end{pmatrix}$$

Las relaciones límite (2.57) pueden ser expresadas en la forma matricial

$$\lim_{h \to 0+} \frac{\mathbf{P}(h) - \mathbf{I}}{h} = \mathbf{A}, \tag{2.58}$$

que muestra que A es la matriz derivada de $\mathbf{P}(t)$ en $t = 0$. Formalmente, $\mathbf{A} = \mathbf{P}'(0)$. Con ayuda de (2.58) y refiriendonos a(2.53) tenemos

$$\frac{\mathbf{P}(t+h) - \mathbf{P}(t)}{h} = \frac{\mathbf{P}(t)[\mathbf{P}(h) - \mathbf{I}]}{h} = \frac{\mathbf{P}(h) - \mathbf{I}}{h}\mathbf{P}(t). \tag{2.59}$$

El límite del lado derecho existe y conduce a la matriz de ecuaciones diferenciales

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{A} = \mathbf{A}\mathbf{P}(t). \tag{2.60}$$

donde $\mathbf{P}'(t)$ denota a la matriz cuyos elementos son $P'_{ij}(t) = dP_{ij}(t)/dt$, la existencia de $P'_{ij}(t)$ es una consecuencia obvia de (2.58) y (2.59).

Ejemplo 2.3 *Cadena de Markov de dos estados.*

Considerese una cadena de Markov $\{X(t)\}$ con estados $\{0, 1\}$ cuya matriz infinitesimal es

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{Vmatrix} -\lambda & \mu \\ \lambda & -\mu \end{Vmatrix} \end{array}$$

El proceso alterna entre los estados 0 y 1. los tiempos en el estado 0 son variables aleatorias exponencialmente distribuidas e independientes con parámetro λ. Los tiempos en el estado 1 son, también, variables aleatorias distribuidas exponencialmente e independientes, pero con parámetro μ. En este caso especial, la matriz de ecuaciones diferenciales (2.60) se convierte en

$$\begin{Vmatrix} P'_{00}(t) & P'_{01}(t) \\ P'_{10}(t) & P'_{11}(t) \end{Vmatrix} = \begin{Vmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{Vmatrix} \times \begin{Vmatrix} -\lambda & \lambda \\ \mu & -\mu \end{Vmatrix}.$$

cuyo primer elemento es

$$P'_{00}(t) = -\lambda P_{00}(t) + \mu P_{01}(t). \tag{2.61}$$

Ahora $P_{01}(t) = 1 - P_{00}(t)$, cuya sustitución en (2.61) da

$$P'_{00}(t) = \mu - (\lambda + \mu) P_{00}(t).$$

Sea $Q_{00}(t) = e^{(\lambda+\mu)t} P_{00}(t)$. Entonces

$$\begin{aligned} \frac{dQ_{00}(t)}{dt} &= e^{(\lambda+\mu)t} P'_{00}(t) + (\lambda + \mu) e^{(\lambda+\mu)t} P_{00}(t) \\ &= e^{(\lambda+\mu)t} [P'_{00}(t) + (\lambda + \mu) P_{00}(t)] \\ &= \mu e^{(\lambda+\mu)t} \end{aligned}$$

que puede ser integrado como

$$\begin{aligned} Q_{00}(t) &= \mu \int e^{(\lambda+\mu)t} dt + C \\ &= \left(\frac{\mu}{\lambda + \mu}\right) e^{(\lambda+\mu)t} + C. \end{aligned}$$

La condición inicial $Q_{00}(0) = 1$ determina la constante de integración $C = \lambda/(\lambda + \mu)$. Entonces

$$Q_{00}(t) = e^{(\lambda+\mu)t} P_{00}(t) = \left(\frac{\mu}{\lambda + \mu}\right) e^{(\lambda+\mu)t} + \left(\frac{\lambda}{\lambda + \mu}\right)$$

y

$$P_{00}(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t}. \tag{2.62}$$

Dado que $P_{01}(t) = 1 - P_{00}(t)$, se tiene:

$$P_{01}(t) = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t}, \tag{2.63}$$

y, por simetría,

$$P_{11}(t) = \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda+\mu)t}. \tag{2.64}$$

$$P_{10} = \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu} e^{-(\lambda+\mu)t} \tag{2.65}$$

Regresando a la cadena de Markov general en los estados $\{0, 1, \cdots, N\}$, las ecuaciones diferenciales (2.60) sobre la condición inicial $\mathbf{P}(0) = \mathbf{I}$ pueden ser resueltas por metodos estándar para dar la fórmula

$$\mathbf{P}(t)) = e^{\mathbf{A}t} = \mathbf{I} + \sum_{n=1}^{\infty} \frac{\mathbf{A}^n t^n}{n!} \tag{2.66}$$

Cuando la cadena de Markov es reversible (todos los estados se comunican) entonces $P_{ij}(t) > 0$ para $i, j = 0, 1, \cdots, N$ y el $\lim_{t\to\infty} P_{ij}(t) = \pi_j > 0$ existen independientemente del estado inicial i. La distribución límite debe ser encontrada pasando por el límite (2.60) es decir $\lim_{t\to\infty} \mathbf{P}'(t) = 0$. Las ecuaciones resultantes para $\pi = (\pi_0, \pi_1, \cdots, \pi_N)$ son

$$0 = \pi\mathbf{A} = (\pi_0, \pi_1, \cdots, \pi_N) \begin{Vmatrix} -q_0 & q_{01} & \cdots & q_{0N} \\ q_{10} & -q_1 & \cdots & q_{1N} \\ \vdots & vdots & \ddots & vdots \\ q_{N0} & q_{N1} & \cdots & -q_N \end{Vmatrix},$$

que es lo mismo que

$$\pi_j q_j = \sum_{i\neq j} \pi_i q_{ij} \quad j = 0, 1, \cdots, N. \tag{2.67}$$

La ecuación (2.67) conjuntamente con

$$\pi_0 + \pi_1 + \cdots + \pi_N = 1 \tag{2.68}$$

determinan la distribución límite.

La ecuación (2.67) tiene una interpretación de masa, que ayuda al entendimiento de ella misma. El lado izquierdo $\pi_j q_j$ representa la tasa a futuro o estable en la cual ejecuciones particulares del proceso de Markov dejan el estado j. Esta tasa debe ser igual a la tasa a futuro o estable en las ejecuciones particulares de arribos al estado j, el equilibrio es primordial.

Dichos arribos deben venir de algún estado $i \neq j$. Por ello, el lado derecho $\sum_{i\neq j} \pi_i q_{ij}$ representa la tasa total de arribos.

2.6 Procesos de Renovación

La teoría de renovación comienza con el estudio de sistemas estocásticos cuya evolución a través del tiempo está intercalada con renovaciones o regeneraciones, en un sentido estadístico, el proceso comienza nuevamente. Hoy en día, la materia es visualizada como el estudio de funciones generales de variables aleatorias independientes, idénticamente distribuidas y no-negativas, que representan los intervalos sucesivos entre las renovaciones. Los resultados son aplicables en una amplia variedad de modelos probabilísticos, tanto teóricos como prácticos.

Un *proceso de renovación* $\{N(t), t \geq 0\}$ es un proceso estocástico no-negativo valuado entero que registra las ocurrencias sucesivas de un evento durante el intervalo de tiempo $(0, t]$, donde las duraciones del tiempo entre eventos sucesivos son variables aleatorias *positivas, independientes, e idénticamente distribuidas*. Sea $\{X_k\}_{k=1}^{\infty}$, los tiempos entre las ocurrencias sucesivas de eventos (que representan frecuentemente la duración de algunas unidades sucesivamente puestas en servicio), así X_i es el lapso de tiempo desde el $(i-1)$-ésimo evento hasta la ocurrencia del evento hasta la ocurrencia del evento i-ésimo. Escribiendo,

$$F(x) = \Pr\{X_k = x\}, \qquad k = 1, 2, 3, \cdots,$$

para las distribuciones de probabilidad comunes de $X_1, X_2, \cdots$ Un supuesto básico para los procesos de renovación es que $F(0) = 0$, queriendo decir con esto, que $X_1, X_2, \cdots$, son variables aleatorias positivas. Hay referencia a

$$W_n = X_1 + X_2 + \cdots + X_n, \quad n \geq 1 \ (W_0 = 0, \text{por convención}) \tag{2.69}$$

Como a el *tiempo de espera* hasta la ocurrencia del n-ésimo evento.

La relación entre los tiempos de ocurrencia $\{X_k\}$ y los procesos de renovación $\{N(t), t \geq 0\}$ es mostrada en la Figura 2.4.

Nótese que

$$N(t) = \text{número de índices } n \text{ para los cuales } 0 < W_n \leq t \tag{2.70}$$

El modelo prototipo de renovación envuelve el remplazamiento sucesivo de focos. Un foco es instalado para servir en el tiempo W_0, se funde (o falla) en el momento $W_1 = X_1$ y es sustituido por un foco nuevo. El segundo foco falla en el momento $W_2 = X_1 + X_2$ y es reemplazado por un tercer foco. En general, el n-ésimo foco se funde en el momento $W_n = X_1 + X_2 + \cdots + X_n$ y es inmediatamente sustituido, y el proceso continua. Es

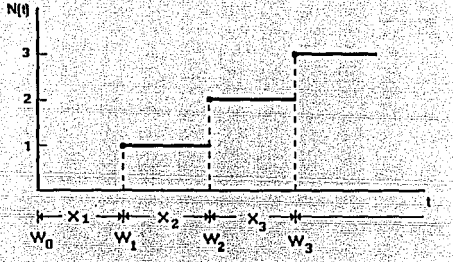

Figura 2.4: *La relación entre dos tiempos de ocurrencia X_k y el proceso de renovación $N(t)$.*

natural suponer que los tiempos de vida sucesivos son estadísticamente independientes, con características probabilistas idénticas, por eso

$$\Pr\{X_k \leq x\} = F(x) \quad \text{para} \quad k = 1, 2, \cdots.$$

En este proceso $N(t)$ guarda el número de focos sustituidos hasta el momento t.

El objetivo principal de la teoría de renovación, es derivar propiedades de ciertas variables aleatorias asociadas con $\{N(t)\}$ y $\{W_n\}$ del conocimiento de la distribución de interocurrencias F. Por ejemplo, es significativo y relevante calcular el número esperado de renovaciones para el tiempo de duración $(0, t]$:

$$E[N(t)] = M(t),$$

que es llamada *función de renovación*. Para este fin, diversas relaciones y fórmulas pertinentes son usadas, la ley de probabilidad de $W_n = X_1 + X_2 + \cdots + X_n$, puede ser obtenida, de acuerdo con las fórmulas de convoluciónde variables aleatorias.

$$\Pr\{W_n \leq x\} = F_n(x)$$

donde $F_1(x) = F(x)$ se supone conocida o descrita, y entonces

$$F_n(x) = \int_0^\infty F_{n-1}(x - \gamma) dF(\gamma) = \int_0^x F_{n-1}(x - \gamma) dF(\gamma).$$

La ligadura de unión fundamental entre el proceso de tiempo de espera $\{W_n\}$ y el proceso de renovación $\{N(t)\}$ es la observación de que

$$N(t) \geq k \quad \text{si sólo si} \quad W_k \leq t \tag{2.71}$$

En palabras, la ecuación (2.71) asienta que el número de renovaciones al tiempo t es cuando menos k, si sólo si la k-ésima renovación ocurre en o antes del momento t.

De (2.71) se sigue que

$$\begin{aligned} \Pr\{N(t) \geq k\} &= \Pr\{W_k \leq t\}. \\ &= F_k(t), \qquad t \geq 0, \quad k = 1, 2, \cdots \end{aligned} \tag{2.72}$$

y consecuentemente,

$$\begin{aligned} \Pr\{N(t) = k\} &= \Pr\{N(t) \geq t\} - \Pr\{N(t) \geq k+1\} \\ &= F_k(t) - F_{k+1}(t), \qquad t \geq 0, \quad k = 1, 2, \cdots. \end{aligned} \tag{2.73}$$

Para la función de renovación $M(t) = E[N(t)]$ sumando las probabilidades finales en la forma $E[N(t)] = \sum_{k=1}^{\infty} \Pr\{N(t) \geq k\}$ y usando la ecuación (2.72) para obtener que:

$$\begin{aligned} M(t) = E[N(t)] &= \sum_{k=1}^{\infty} \Pr\{N(t) \geq k\} \\ &= \sum_{k=1}^{\infty} \Pr\{W_k \leq t\} = \sum_{k=1}^{\infty} F_k(t). \end{aligned} \tag{2.74}$$

Existen algunas variables aleatorias de interes en la teoría de renovación. Tres de estas son: el *exceso de vida* (también llamada variable aleatoria de exceso), la *vida corriente* (también llamada la edad de la variable aleatoria), y la *vida total*, definidas respectivamente, como

$\gamma_t = W_{N(t)+1} - t$ (exceso o vida residual)
$\delta_t = t - W_{N(t)}$ (vida corriente o edad de la variable aleatoria)
$\beta_t = \gamma_t + \delta_t$ (vida total).

Una descripción gráfica de estas variables aleatorias se muestra en la Figura 2.5.

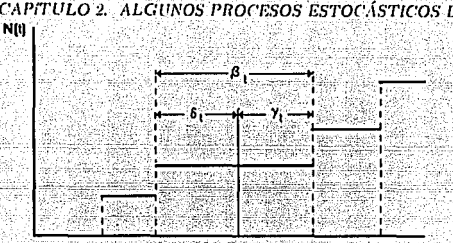

Figura 2.5: *El exceso de vida γ_t, la vida corriente δ_t, y la vida total β_t.*

2.6.1 Dos Ejemplos de Procesos de Renovación

Los ejemplos siguientes nos muestran los diversos contextos en los cuales los procesos de renovación aparecen.

Proceso Poisson

Un proceso Poisson $\{N(t), t \geq 0\}$ con parámetro λ es un proceso de renovación que tiene distribución exponencial interocurrencias.

$$F(x) = 1 - e^{-\lambda x} \quad \text{si} \quad x \geq 0.$$

Este proceso particular de renovación se ampliará en la sección 2.7.

Proceso de Renovación Asociado con Líneas de Espera.

En un sistema con un sólo servidor, están inmiscuidos algunos procesos de renovación en forma natural. Citemos dos ejemplos:

- Si los tiempos de arribo de los clientes forman un proceso de renovación, entonces los momentos de períodos sucesivos ocupados genera un segundo proceso de renovación.
- Por la situación en la cual el proceso de entrada es Poisson. los momentos sucesivos en que el servidor pasa de un estado ocupado a uno libre, determina en un proceso de renovación.

2.7 Proceso de Poisson Visto Como Proceso de Renovación

Como se mencionó anteriormente, el proceso de Poisson con parámetro λ es un proceso de renovación cuyo tiempo de interocurrencias tiene distribución exponencial $F(x) = 1 - e^{-\lambda x}, x \geq 0$. La propiedad de "pérdida de la memoria" de la distribución exponencial (vista en 2.3.3) permite numerosas propiedades del proceso Poisson visto como proceso de renovación.

La Función de Renovación.

Dado que $N(t)$ tiene una distribución Poisson, entonces

$$\Pr\{N(t) = k\} = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \qquad k = 0, 1, \cdots$$

y

$$M(t) = E[N(t)] = \lambda t.$$

Exceso de Vida.

Obsérvese que el exceso de vida en el tiempo t sobrepasa a x si sólo si no hay renovaciones en el intervalo $(t, t + x]$ (Figura 2.6). Este evento tiene la misma probabilidad que una "no renovación" en el intervalo $(0, x]$, dado que un proceso Poisson tiene incrementos estacionarios independientes.

En términos formales se tiene:

$$\Pr\{\gamma_t > x\} = \Pr\{N(t + x) - N(t) = 0\} \tag{2.75}$$
$$= \Pr\{N(x) = 0\} = e^{-\lambda x}.$$

Por esto, en un proceso de Poisson, el excedente de vida, posee la misma distribución exponencial.

$$\Pr\{\gamma_t \leq x\} = 1 - e^{-\lambda x}, \qquad x \geq 0, \tag{2.76}$$

como cualquier vida, otra manifestación de la propiedad de pérdida de la memoria de la distribución exponencial es;

Vida Corriente.

La vida corriente δ_t, no puede exceder t, mientras $X < t$ la vida corriente excede x si sólo si no hay renovaciones en $(t - x, t]$, que nuevamente tiene probabilidad $e^{-\lambda x}$. Por esto la vida corriente sigue la distribución exponencial truncada

$$\Pr\{\delta_t \leq x\} = \begin{cases} 1 - e^{-\lambda x} & \text{para} \quad 0 \leq x < t, \\ 1 & \text{para} \quad t \leq x. \end{cases} \tag{2.77}$$

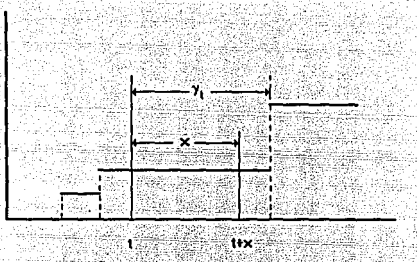

Figura 2.6: *El exceso de vida γ_t sobrepasa x si y sólo si no hay renovaciones*

Vida Total Media.
Usando las propiedades de la media para variables aleatorias no negativas; se tiene

$$\begin{aligned} E[\beta_t] &= E[\gamma_t] + E[\delta_t] \\ &= \frac{1}{\lambda} + \int_0^t \Pr\{\delta_t > x\}dx \\ &= \frac{1}{\lambda} + \int_0^t e^{-\lambda x}dx \\ &= \frac{1}{\lambda} + \frac{1}{\lambda}(1 - e^{-\lambda t}). \end{aligned}$$

Observe que la vida total media es significartivamente más larga que la vida media $1/\lambda = E[X_k]$ de cualquier intervalo de renovación particular. Una expansión más directa de estos fenómenos se muestra cuando t es grande, donde el proceso ha estado en operación durante un tiempo largo. Entonces la vida total media $E[\beta_t]$ es aproximadamente dos veces la vida media, este hecho parece paradogico.

Se examinará la definición de la vida total β_t, para tratar de explicarlo con bases intuitivas. Primero, un tiempo puntual arbitrario t arbitrario es fijado. Entonces β_t, mide la longitud del intervalo de renovación. El fenómeno es conocido como muestra de "longitud-sesgada" y sucede, en numerosas situaciones muestrales.

Distribución Conjunta de γ_t y δ_t.

La distribución conjunta de γ_t y δ_t es determinada de la misma forma que las marginales. De hecho, para cualquier $x > 0$ y $0 < \gamma < t$, el evento $\{\gamma_t > x; \delta_t > \gamma\}$ sucede si solo si no hay renovaciones en el intervalo $(t-\gamma, t+x]$, que tiene una probabilidad $e^{-\lambda(x+\gamma)}$. Por esto

$$\Pr\{\gamma_t > x; \delta_t > \gamma\} = \begin{cases} e^{-\lambda(x+\gamma)} & \text{si} \quad x > 0, \quad 0 < \gamma < t, \\ 0 & \text{si} \quad \gamma \geq t. \end{cases} \tag{2.78}$$

Para el proceso de Poisson, se observa que γ_t y δ_t son independientes, dado que su distribución conjunta es el producto de sus distribuciones marginales.

2.7.1 Procesos Acumulativos y Relacionados.

Sea Y_i una variable aleatoria con la i-ésima unidad o tiempo de vida intervalar ($\{Y_i\}$ identicamente distribuidas), además para el tiempo de vida X_i. Se supondrán X_i y Y_i independientes, pero también, que las parejas $(X_1, Y_1), (X_2, Y_2), \cdots$, son independientes. Se usará la notación $F(x) = \Pr\{X_i \leq x\}$, $G(\gamma) = \Pr\{Y_i \leq \gamma\}$, $\mu = E[X_i]$, y $\nu = E[Y_i]$.

Cierto número de problemas prácticos y teóricos de interes tienen una fórmula natural en estos términos.

Procesos de Renovación que envuelven dos componentes para cada intervalo de Renovación.

Suponga que Y_i representa una porción de la duración X_i. la Figura 2.7 ilustra el modelo. Entonces se tiene la porción Y_1 que ocurre al comienzo del intervalo, pero esta suposición no es escencial para el resultado siguiente.

Sea $P(t)$ la probabilidad de que t caiga en una porción Y_i de algún intervalo de renovación. Cuando $X_1, X_2, \cdots$, son variables aleatorias continuas, el teorema de renovación implica la siguiente evaluación

$$\lim_{t\to\infty} P(t) = \frac{E[Y_1]}{E[X_1]} \tag{2.79}$$

Aquí mencionaremos dos ejemplos concretos.

Ejemplo 2.4 Un Modelo de Cola: *Un proceso de cola, es un proceso en el cual los clientes arriban a un determinado lugar donde reciben un servicio. Si supone que los tiempos entre arribos, y el tiempo que es gastado en efectuar un servicio para un determinado cliente, están gobernados por las leyes de probabilidad.*

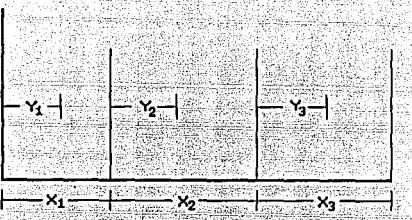

Figura 2.7: *Un proceso de renovación, en el cual una variable aleatoria Y_i asociada representa una porción del i-ésimo intervalo de renovación.*

Si los arribos a una cola siguen un proceso Poisson de intensidad λ, entonces los tiempos sucesivos X_k desde el comienzo del k-ésimo período ocupado, al comienzo del sigiente período ocupado, forma un proceso de renovación. (Un período ocupado es un período ininterrumpido durante el cual la cola nunca está vacía.) Cada X_k está compuesta por una porción ocupada Z_k y una porción desocupada Y_k. Entonces $P(t)$, la probabilidad de que la cola esté vacía en el tiempo t, converge a $E[Y_1]/E[X_1]$. Este ejemplo es tratado más profundamente en el sigiente Capítulo.

Ejemplo 2.5 El Principio de Piter: *El "Principio de Piter" afirma que un trabajador será ascendido hasta alcanzar una posición en la que es incompetente. Cuando esto pasa, la persona permanece en ese puesto hasta que se retira. Considere el siguiente modelo del principio de piter. Una persona es seleccionada en forma aleatoria de una población y se le da el trabajo. Si la persona es competente, permanecerá en el puesto durante un tiempo aleatorio que tiene distribución F y media μ y es ascendido. Si es incompetente , la persona permanecerá durante un tiempo también aleatorio que tiene una distribución G y media $\nu > \mu$ y se retira. Una vez que el puesto está vacante, otra persona es seleccionada en forma aleatoria y el proceso se repite. Supongamos que la población infinita, contiene la fracción p de personas competentes y $q = 1 - p$ de personas incompetentes.*

En el comportamiento a futuro. ¿Que fracción de tiempo es la posición más conve-

niente para una persona incompetente?

Ocurre una renovación cada vez que el puesto es ocupado, y por ello la duración media de un ciclo de renovación es:

$$E[X_k] = p\mu + (1-p)\nu.$$

Para responder a la pregunta, sea $Y_k = X_k$ si la k-ésima persona es competente. Entonces la fracción de tiempo en la cual la posición es más conveniente para una persona incompetente es

$$\frac{E[Y_1]}{E[X_1]} = \frac{(1-p)\nu}{p\mu + (1-p)\nu}.$$

Supongamos que $p = 1/2$ de la población es competente, y $\nu = 10$ mientra $\mu = 1$, entonces

$$\frac{E[Y_1]}{E[X_1]} = \frac{(1/2)(10)}{(1/2)(10) + (1/2)(1)} = \frac{10}{11} = .91$$

Por ello, mientras la mitad de las personas en la población son competentes, el trabajo es llenado por personas competentes sólo el 9 porciento de las veces.

Capítulo 3

Sistemas de Líneas de Espera

3.1 Descripción de una Línea de Espera

La Figura 3.1 representa los elementos de un sistema de líneas de espera, los clientes provienen de una población o fuente, entran al sistema a recibir algún tipo de servicio. El término "clientes" es usado en términos generales, puede ser una persona, una llamada telefónica, un automovil, etc. El tipo de servicio depende del número de servidores, puede ser uno o más. Un servidor es una entidad capaz de realizar un servicio requerido por un cliente. Si todos los servidores están ocupados (o dando un servicio requerido por un cliente) cuando un cliente entra al sistema se tiene que formar en una cola hasta que pueda recibir su servicio.

Para poder describir analíticamente un sistema de líneas de espera, los elementos del sistema deben ser conocidos, de esta manera se definen los más importantes.

3.1.1 Población o Fuente

La población o fuente de clientes potenciales puede ser finita o infinita. Un sistema con fuente infinita es más fácil de describir matemáticamente que un sistema con fuente finita. La razón de esto es que un sitema de fuente finita, el número de clientes en el sistema afecta a la tasa de llegadas de los mismos, es decir, si un alto porcentaje de la población está en el sistema, la tasa de llegadas debe aproximarse a cero. Si la fuente de clientes es finita pero "grande", algunas veces se supondrá que se trata de una población infinita para simplificar los modelos.

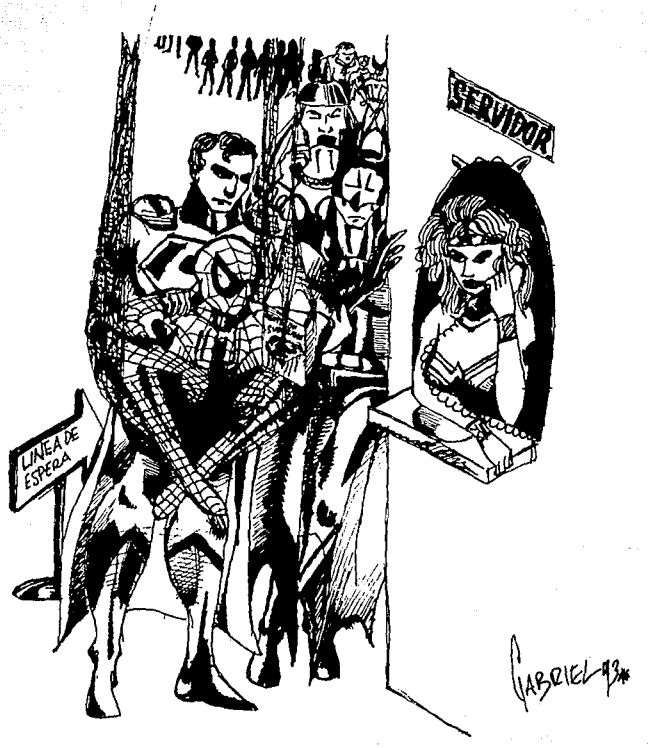

Figura 3.1: *Elementos de un sistema de líneas de espera.*

3.1.2 Descripción de los Arribos

La propiedad de un sistema de líneas de espera de proporcionar servicios a los clientes, no depende únicamente de la tasa promedio de llegadas (λ), sino también del "modo"en que los arribos ocurran. Así, si los arribos de los clientes son espaciados, digamos cado h unidades de tiempo, los servidores podrán brindar mejores servicios que si los clientes llegaran en períodos cortos de tiempo. Se supondrá que los arribos de los clientes serán a los tiempos:

$$0 \leq t_0 \leq t_1 \leq \cdots \leq t_n \cdots$$

(siempre se supondrá que las observaciones de las líneas de espera se inician en $t = 0$). Ahora debemos identificar el proceso estocástico que describe el proceso de llegadas. En términos generales este proceso es descrito en términos de la función de distribución de *tiempos entre llegadas* de los clientes y se denota como $A(t)$, donde:

$$A(t) = \Pr[\text{tiempo entre arribos} \leq t]. \tag{3.1}$$

El supuesto general es que estos intervalos son variables aleatorias independientes e idénticamente distribuidas.

3.1.3 Distribución del Tiempo de Servicios

La operación de un sistema de líneas de espera es dar servicios, la capacidad de este depende del número de servicios que pueda brindar en un determinado período de tiempo, esta capacidad va a ser a representar como la tasa promedio de servicios (μ). La función de distribución del tiempo de servicios se denota como $B(X)$, donde

$$B(x) = \Pr[\text{tiempo de servicio} \leq x] \tag{3.2}$$

Aquí el tiempo de servicio se refiere al tiempo que un cliente se tarda recibiendo un servicio.

3.1.4 Número de Servidores

El sistema de líneas de espera más simple, en este sentido, es el *sistema de un sólo servidor*, el cual puede dar servicio a un sólo cliente a la vez. Un *sistema multi-servidores* tiene c servidores identicos y puede dar servicio a c clientes simultaneamente. En un *sistema de servidores infinitos*, cada cliente que llega es inmediatamente servido.

3.1.5 Disciplina de la Cola

También es conocida como *disciplina de servicio*, esta es la regla para seleccionar los clientes a ser servidos. La disciplina más común es *primero en llegar–primero en ser servido.* Otras disciplinas de la cola podrían ser *último en llegar–primero en ser servido*, y selección aleatoria para servicio, lo cual significa que cada cliente tiene la misma probabilidad de ser seleccionado para recibir un servicio.

Existe una notación que ha sido creada para describir los sistemas de líneas de espera y tiene la forma $A/B/c$. A describe la distribución entre llegadas, B describe la distribución de los tiempos de servicio y c el número de servidores. Los símbolos comunmente usados para A y B son:

GI	Arribos generales interdependientes.
G	Distribución general de tiempos de servicio.
H_k	Distribución de tiempo entre llegadas o tiempo de servicio hiperexponencial con k estados.
E_k	Distribución Erlang-k de tiempos entre arribos o de servicio.
M	Distribución exponencial para tiempos entre llegadas o tiempos de servicio.
D	Distribucion determinística de tiempo entre llegadas o de servicio.

3.2 Notación y Estructura Básica.

El objetivo de esta sección es definir alguna notación y posteriormente introducir uno de los procesos estocásticos más importantes que se encontrará en el desarrollo de la teoría de los sistemas de líneas de espera. Además se derivarán algunos resultados simples pero importantes, que mostrarán diversos aspectos de estos sistemas.

El sistema de líneas de espera que se considera más general es el $G/G/m$; este es el sistema cuya distribución de *tiempo entre llegadas* $A(t)$ es completamente arbitraria y cuya distribución de *tiempos de servicio* $B(x)$ es también completamente arbitrario. (*Todos los tiempos entre llegadas y tiempos de servicio se suponen independientes unos de otros.*).

El sistema tiene m servidores y un determinado orden de servicio que también es arbitrario (en particular, no tiene que ser *primero en llegar–primero en ser servido.*).

Será importante este estudio tanto el arribo de los clientes como el avance de los mismos; así se define a C_n como:

$$C_n \triangleq \textit{el n-ésimo cliente que entra en el sistema.}^{1} \tag{3.3}$$

Definiendo inmediatamente algunos procesos de interés. Por ejemplo, $X(t)$ donde:

$$X(t) \triangleq \textit{número de clientes en el sistema al tiempo } t \tag{3.4}$$

Otro proceso estocástico de interés es el trabajo no terminado o *tiempo de espera requerido para vaciar el sistema, atendiendo a todos los clientes presentes al tiempo t*, $U(t)$, es decir:

$$U(t) \triangleq \textit{el trabajo no terminado al tiempo } t. \tag{3.5}$$

Cuando $U(t) > 0$ se dice que el sistema está ocupado y sólo cuando $U(t) = 0$ el sistema está desocupado.

Para entender los detalles de estos procesos estocásticos se debe antes definir las siguientes variables. Se define el tiempo de arribo al sistema del n-ésimo cliente como:

$$\tau_n \triangleq \textit{el tiempo de arribo de } C_n. \tag{3.6}$$

Entonces se define el tiempo entre arribos del C_{n-1} y C_n como:

$$t_n \triangleq \textit{el tiempo entre llegadas de } C_{n-1} \textit{ y } C_n.$$

$$t_n = \tau_n - \tau_{n-1}. \tag{3.7}$$

Si se supone que todos los tiempos entre llegadas son descritos por la distribución $A(t)$, se tiene que:

$$\Pr[t_n \leq t] = A(t). \tag{3.8}$$

que es totalmente independiente de n. De modo similar el tiempo de servicio para C_n como:

$$X_n \triangleq \textit{el tiempo de servicio para } C_n. \tag{3.9}$$

Se puede suponer también que:

$$\Pr[X_n \leq x] = B(x). \tag{3.10}$$

[1] $\triangleq$ debe ser leído como: *"se define como"*

Las series $\{t_n\}$ y $\{X_n\}$ pueden ser vistas como las variables de entrada del sistema; la manera en que el sistema maneja a los clientes produce aumentos en el tamaño de la cola y en los tiempos de espera que se deben definir. Entonces, se define el tiempo de espera (*tiempo que un cliente permanece en la línea de espera*) como:

$$w_n \triangleq \textit{el tiempo de espera (en cola) de } C_n. \tag{3.11}$$

El tiempo total gastado por C_n en el sistema es la suma de su tiempo de espera (w_n) y de su tiempo de servicio (X_n), que se denota como:

$$S_n \triangleq \textit{el tiempo en el sistema para } C_n.$$

$$S_n = w_n + X_n. \tag{3.12}$$

Así se ha definido el tiempo de arribo para el n-ésimo cliente, el tiempo entre los arribos de dos clientes, el tiempo de servicio para el C_n, el tiempo de espera y el tiempo que gasta en el sistema. Considérese ahora el tiempo entre arribos (o *tiempo entre llegadas*) t_n una vez más. Se denotará ocasionalmente a $\bar{t}$ como el límite de la variable aleatoria t_n, definido como:

$$\bar{t} = \lim_{n\to\infty} t_n. \tag{3.13}$$

que también se denota como $t_n \to \bar{t}$. (*Se requiere que el tiempo entre llegadas t_n tenga una distribución independiente de n, pero esto no será necesario en el caso de muchas otras variables aleatorias de interés.*). La notación para la función de distribución de probabilidad (F.D.P.) será:

$$\Pr[t_n \le t] = A_n(t). \tag{3.14}$$

Y para el límite de F.D.P.:

$$\Pr[\bar{t} \le t] = A(t). \tag{3.15}$$

Entonces $A_n(t) \to A(t)$; por supuesto, para los tiempos entre llegadas se supone que $A_n(t) = A(t)$. De modo similar la función de densidad de probabilidad (f.d.p.) para t_n y $\bar{t}_n$ será $a_n(t)$ y $a(t)$ respectivamente, y $a_n(t) \to a(t)$. Los momentos de los tiempos entre llegadas son también importantes y se denotan como:

$$E[t_n] = \bar{t}. \tag{3.16}$$

De acuerdo con esta notación, la esperanza del tiempo entre llegadas para el límite de la variable aleatoria será dado por $\bar{t}$ en el sentido de que $\bar{t}_n \to \bar{t}$. Como $\bar{t}$, que es el promedio del tiempo entre llegadas de los clientes, es usado frecuentemente en las ecuaciones se da una notación especial que es:

$$\bar{t} = \frac{1}{\lambda}. \tag{3.17}$$

Así λ representa la tasa promedio de llegadas de los clientes a nuestro sistema . Los momentos de orden k del tiempo entre llegadas son también de interés y así se define el k-ésimo momento como:

$$E[t^k] = \bar{t} = a_k \qquad k = 0, 1, 2, \ldots \tag{3.18}$$

Se ha definido a_k como el k-ésimo momento del tiempo entre llegadas t; esto es justamente la notación estandar, y se observa inmediatamente que:

$$\bar{t} = \frac{1}{\lambda} = a_1 = a. \tag{3.19}$$

Esto es, existen tres notaciones especiales para la esperanza del tiempo entre llegadas; en particular, el uso del simbolo "a" es muy común y algunas de estas formas serán utilizadas a través del texto. Recopilando la información observada para el tiempo entre llegadas , se tiene el siguiente glosario:

t_n = El tiempo entre las llegadas de C_n y C_{n-1}.

$$t_n \to t, \qquad A_n(t) \to A(t), \qquad a_n(t) \to a(t)$$

$$\bar{t}_n \to \bar{t} = \frac{1}{\lambda} = a_1 = a, \qquad t_n^k \to t^k = a_k. \tag{3.20}$$

De manera similar se identifica la notación asociada con X_n, w_n y S_n como:

X_n = El tiempo de servicio de C_n.

$$X_n \to X, \qquad B_n(x) \to B(x), \qquad b_n(t) \to b(t)$$

$$\bar{X}_n \to \bar{X} = \frac{1}{\mu} = b_1 = b, \qquad \bar{X}_n^k \to \bar{X}^k = b_k. \tag{3.21}$$

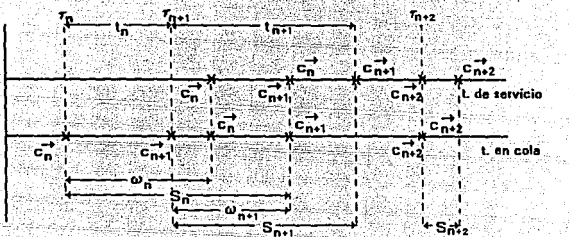

Figura 3.2: *Diagrama-tiempo de una línea de espera.*

w_n = El tiempo de espera para C_n.

$$w_n \to w, \qquad W_n(y) \to W_0(y), \qquad w_n(y) \to w(y)$$
$$\bar{w_n} \to \bar{w} = W_0, \qquad \bar{w_n^k} \to \bar{w^k}. \tag{3.22}$$

S_n = El tiempo en el sistema para C_n.

$$S_n \to S, \qquad S_n(y) \to S(y), \qquad s_n(y) \to s(y)$$
$$\bar{S_n} \to \bar{S} = W, \qquad \bar{S_n^k} \to \bar{S^k}. \tag{3.23}$$

Toda esta notación es evidente excepto tal vez por los símbolos especiales utilizados para el primer momento y también para los momentos de orden mayor de las variables aleatorias usadas (es decir el uso de los símbolos λ, a, μ, b, W_0, y W.).

Dada la notación anterior ahora se sugiere una gráfica para las líneas de espera, que permita una visión clara de la dinámica del sistema de líneas de espera.

El diagrama se muestra en la Figura 3.2. Esta gráfica particular representa un sistema de líneas de espera con una disciplina de servicio *"primero en llegar - primero en ser*

servido", pero se puede observar que la Figura 3.2 podría ser modificada para representar cualquier otra disciplina. En este diagrama la línea inferior de tiempo, representa la cola y la superior el servicio; además el diagrama que se muestra es para el caso de un sólo servidor, esto se puede generalizar también. Las flechas que están "apuntando" a la línea de la cola (o a la de servicio) indican que una llegada a la cola ha ocurrido (o se ha iniciado un servicio). Mientras que las flechas que "salen" de las líneas indican que el cliente ha dejado la cola (o ha terminado su servicio). Se observa también que el C_{n+1} llega antes que C_n inicie su servicio; sólo cuando C_n deja o termina su servicio C_{n+1} puede iniciar el suyo y, por supuesto, estos dos eventos ocurren simultáneamente. Notese que cuando C_{n+2} entra al sistema lo encuentra vacío, e inmediatamente se dirige a recibir el servicio. Se describe también en este diagrama el tiempo de espera (w_n) y el tiempo en el sistema (S_n) para C_n (nótese que $w_{n+2} = 0$). De este modo, conforme el tiempo transcurre se puede identificar el número de clientes en el sistema $X(t)$, el trabajo no terminado al tiempo t, $U(t)$, y también los períodos desocupados y períodos ocupados.

En general en un sistema de líneas de espera uno espera que cuando el número de clientes es muy grande el tiempo de espera también lo sea. Una manifestación inmediata de este hecho es una relación simple entre los valores esperados del sistema, la tasa promedio de llegadas de los clientes al sistema y el promedio de tiempo de estancia en el sistema para los clientes. El propósito de este tema es obtener esta relación y así familiarizarse un poco con el comportamiento del sistema. Con referencia a la Figura 3.1 considerense las llegadas al sistema y contará cuantos clientes entran en determinado tiempo.

Se denotará esto como $a(t)$ donde:

$$a(t) \triangleq \textit{el número de arribos en el intervalo } (0, t). \tag{3.24}$$

También se pueden considerar las salidas del sistema, y contar el número de clientes atendidos que dejan el sistema; es:

$$\delta(t) \triangleq \textit{el número de clientes atendidos en el intervalo } (0, t). \tag{3.25}$$

Claramente $X(t)$, el número de clientes en el sistema al tiempo t, debe ser explicado por la relación:

$$X(t) = a(t) - \delta(t).$$

Por otra parte el área total entre estas dos curvas para un punto determinado t, representa el tiempo total que todos los clientes han estado en el sistema (y las unidades

son "cliente por unidad de tiempo") en el intervalo $(0, t)$; denoteses, ahora a esta área como $\gamma(t)$, además, definiendo a λt como la tasa promedio de llegadas (clientes por unidad de tiempo) en el intervalo $(0, t)$; es decir:

$$\lambda t = \frac{\alpha(t)}{t} \tag{3.26}$$

Se Puede definir a W_t como el tiempo promedio en el sistema por cliente sobre todos los clientes en el intervalo $(0, t)$; como $\gamma(t)$ representa los clientes por unidad de tiempo acumulados hasta el tiempo t, se puede dividir por el número de arribos hasta el tiempo t y obtener:

$$W_t = \frac{\gamma(t)}{\alpha(t)}.$$

Por último, debemos definir $\bar{X}_t$ como el promedio de clientes en el sistema en el intervalo de tiempo $(0, t)$; esto se obtiene mediante la división de clientes por unidad de tiempo acumulados entre la longitud del intervalo t, así:

$$\bar{X}_t = \frac{\gamma(t)}{t}. \tag{3.27}$$

De estas tres últimas ecuaciones se observa la relación:

$$\bar{X}_t = \lambda_t W_t.$$

Supóngase ahora que el sistema de líneas de espera es tal que los siguientes límites existen cuando $t \to \infty$.

$$\lambda = \lim_{t \to \infty} \lambda_t$$

$$W = \lim_{t \to \infty} W_t$$

Note que se han usado las definiciones anteriores para λ y W que representan la tasa promedio de clientes que llegan y el tiempo promedio que gasta cada cliente en el sistema, respectivamente. Si estos dos últimos límites existen, entonces también existirá el límite para $\bar{X}_t$, que se denotará $\bar{X}$ y que representa el número promedio de clientes en el sistema entonces:

$$\bar{N} = \lambda W.$$

Esto establece que *el número promedio de clientes en un sistema de líneas de espera es igual a la tasa promedio de arribo de los clientes a ese sistema, por el tiempo promedio gastado por el cliente en el mismo.*[2]

La justificación al resultado anterior no depende de ningún supuesto específico sobre la distribución de las llegadas $A(t)$ o la distribución del tiempo de servicio $B(x)$; ni tampoco depende del número de servidores del sistema o alguna disciplina particular de servicio. Es importante observar que no se tiene completamente definido aún lo que es un sistema de líneas de espera. Por ejemplo la Figura 3.1 podría representar la entrada a un sistema compuesto de una cola y un servido, en cuyo caso $\bar{N}$ y T como están definidas expresan cantidades para las entradas al sistema; por otro lado, se pudo haber considerado un sistema que únicamente contuviera a las colas, en cuyo caso la relación podría haber sido:

$$\bar{N}_q = \lambda W_0, \tag{3.28}$$

donde $\bar{N}_q$ representa el número promedio de clientes en la cola y como se definió W_0 que representa el tiempo promedio de espera en la cola. Como tercera posibilidad, el sistema de colas definido podría haber considerado únicamente al servidor; en este caso la ecuación se reduciría a:

$$\bar{N}_s = \lambda \bar{x}, \tag{3.29}$$

donde, $\bar{N}_s$ representa el número promedio de clientes en el servicio y $\bar{x}$, por supuesto, representa el tiempo promedio gastado en el servicio. Note que, siempre sucede

$$W = \bar{x} + W_0. \tag{3.30}$$

Los sistemas de líneas de espera pueden representar una clase específica de clientes, tal vez basado en prioridades o algunos otros atributos de esta clase, en cuyo caso la misma relación se aplicaría. En otras palabras, la tasa promedio de llegadas de los clientes a un sistema de líneas de espera, es multiplicada por el tiempo promedio de estancia de los clientes en el sistema, es igual al número promedio de clientes en el sistema, no obstante como se haya definido el sistema.

[2]Este resultado es conocido como *el resultado de Little.*

Ahora, se discutirá un párametro básico "ρ", se conoce como, *factor de utilización*. El factor de utilización es la razón de la tasa de llegadas al sistema y la tasa de servicios, entonces en el caso de un sólo servidor, la definición para ρ es :

$\rho \triangleq$ (tasa promedio de los clientes) $\times$ (el tiempo promedio de servicio)

$$\rho = \lambda \bar{x}. \tag{3.31}$$

Esto es cierto, debido a que los sistemas con un sólo servidor, tienen una cierta capacidad máxima, por ejemplo, un cliente por segundo y el arribo de cada cliente trae consigo $\bar{x}$ segundos de trabajo: como en promedio λ clientes llegan por segundo, entonces $\lambda\bar{x}$ segundos de trabajo, son traidos por cliente cada segundo que pasa en promedio. En el caso, de servidores múltiples (m servidores) la definición es la misma; expresada en términos de parámetros de sistema, se tiene:

$$\rho = \frac{\lambda \bar{x}}{m}. \tag{3.32}$$

Las dos ecuaciones anteriores se aplican cuando, la tasa máxima de servicio es independiente del estado del sistema; si este no es el caso entonces se debe proveer de una definición más cuidadosa. La tasa a la cual el trabajo entra al sistema es algunas veces referida como *la intensidad de tráfico* del sistema y es expresada comunmente en *Erlangs;* para el caso de un sistema con un servidor , el factor de utilización es igual a la intensidad del tráfico ya que para (m) servidores múltiples, la intensidad del tráfico es igual a $m\rho$. Además $0 \leq \rho < 1$, entonces ρ puede ser interpretado como:

$$\rho = E(\text{fracción de servidores ocupados}). \tag{3.33}$$

En el caso de un número infinito de servidores, el factor de utilización ρ no juega un papel importante y estando interesados en el número de servidores ocupados (y su esperanza).

Para que el sistema $G/G/1$ sea estable, debe suceder que $0 \leq \rho < 1$. Ocasionalmente, se permite el caso en que $\rho = 1$ que representa el estado de estabilidad (en particular para el sistema $D/D/1$). La estabilidad aquí una vez más se refiere al hecho de que el límite de la distribución de todas la variables aleatorias de interés exista y que todos los clientes sean, eventualmente atendidos. En cuyo caso debe ocuparse del cálculo siguiente. Sea r el largo de un intervalo de tiempo arbitrario; durante este intervalo se espera (por la ley de los grandes números) que con probabilidad 1 el número de arribos sea muy cercano o

igual a λr. Además, definiendo a ρ_0 como la probabilidad de que esté desocupado en algún momento seleccionado aleatoriamente. Se debe decir que durante el intervalo r, el servidor está ocupado $r - r\rho_0$ segundo y con probabilidad 1, el número de clientes servidos durante el intervalo r es muy cercano a $\frac{r-r\rho_0}{\bar{x}}$. Ahora igualando el número de servidores durante este intervalo, da, para la magnitud de r,

$$\lambda_r \cong \frac{r - r\rho_0}{\bar{x}}.$$

Además, como $r \to \infty$ se tiene que $\lambda\bar{x} = 1 - \rho_0$; usando la definición (3.31) se tiene finalmente una ecuación importante para $G/G/1$

$$\rho = 1 - \rho_0 \tag{3.34}$$

Aquí la interpretación es que ρ es la fracción de tiempo que el servidor está ocupado, esto sirve de soporte para la ecuación (3.29) en la cual $\lambda\bar{x} = \rho$ que se mostró como igual al número promedio de clientes que tienen acceso al servicio.

Existe una gran variedad de procesos de *entrada*, dos tipos simples y de frecuente ocurrencia son matemáticamente descritos y derivan algunos casos complejos. El primero, en el que las entradas de los clientes son en los intervalos de tiempo fijos: $T, 2T, 3T, \cdots$.

El segundo modelo más común, es en el que los procesos de llegadas son "completamente aleatorios" donde los tiempos de llegada forman un proceso Poisson. Comprendiendo el desarrollo axiomático de los procesos Poisson (Subsección 2.3.2) se podrá probar la validez de los supuestos de Poisson en cualquier aplicación. Existen muchos resultados teóricos cuando los tiempos de llegada de los clientes forman un proceso de renovación (Sección 2.6). Los intervalos de tiempo distribuidos exponencialmente corresponden a un proceso de Poisson de arribos como caso especial. Supondrá que la duración de los servicios para clientes individuales son variables aleatorias independientes, no-negativas, idénticamente distribuidas, e independientes del proceso de llegada.

3.3 La Fórmula $L = \lambda W$

Considere un sistema que ha funcionado eficientemente hasta alcanzar aproximádamente un estado estacionario, o una posición de equilibrio estadístico.

Sea:

$L =$ *El número de clientes en el sistema,*

λ = La tasa de arribos de los clientes al sistema; y

W = El tiempo promedio gastado por un cliente en el sistema.

La ecuación $L = \lambda W$ es válida sobre una gran variedad de sistemas, y es de importancia básica en la teoría de colas, ya que refleja dos de las más importantes medidas del comportamiento del sistema, el tamaño medio de la cola y el tiempo promedio de espera en el estado estacionario, es decir, el tamaño medio de la cola y el tiempo promedio de espera evaluado con respecto al límite o la distribución estacionaria del proceso.

La validez de $L = \lambda W$ no disminuile en los detalles de ningún modelo particular, depende únicamente del comportamiento masivo del flujo del sistema. Para respaldar este razonamiento, considere al tiempo T lo suficientemente largo, entonces las fluctuaciones de las estadísticas quedan fuera del promedio. Entonces el número total de clientes que entran al sistama es λT , el número total de clientes que han sido atendidos es $\lambda(T - W)$, y el número neto de clientes que permanecen en el sistema L debe ser la diferencia .

$$L = \lambda T - [\lambda(T - W)] = \lambda W.$$

La figura 3.3 describe la relación $L = \lambda W$.

Por supuesto lo que hemos hecho no necesita una prueba, y de hecho, no se debe dar una prueba. Se debe, de cualquier forma, mostrar algunos ejemplos de verificación de la relación $L = \lambda W$, donde L es la media de la distribución estacionaria de los clientes en el sistema, W es el tiempo medio del cliente en el sistema determinado por la distribución estacionaria, y λ es la tasa de arribos en un proceso de arribos Poisson.

Sea L_0 el número promedio de clientes esperando en el sistema, que aún no son servidos, y W_0 el tiempo promedio de espera en el sistema excluyendo el tiempo de servicio. Análogamente a $L = \lambda W$, se tiene:

$$L_0 = \lambda W_0 \tag{3.35}$$

El tiempo total de espera en el sistema es la suma del tiempo de espera antes de servicio, más el tiempo de servicio. En términos de medias se tiene:

$$W = W_0 + \text{ Tiempo promedio de servicio.} \tag{3.36}$$

En lo que resta de este Capítulo, se estudiarán algunas variedades de los sistemas de líneas de espera. Planteando también cuestiones un poco más elaboradas, los "abortos"por ejemplo, que son clientes que se rehusan a entrar al sistema si la línea de espera es demasiado larga. Más generalmente, un sistema de líneas de espera con abortos, es un sistema en el

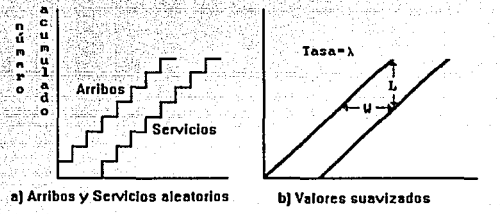

Figura 3.3: *El número de arribos y servicios ocurridos en un sistema de líneas de espera. Los valores suavizados en (b) simbolizan el comportamiento promedio. La tasa de arribos por unidad de tiempo es λ, el número promedio de clientes en el sistema es L, y el tiempo promedio gastado por un cliente en el sistema es W.*

cual los clientes entran con cierta probabilidad, dependiendo del tamaño la cola. Aquí es conveniente diferenciar el *proceso de arribo* y el *proceso de entrada* como se muestra en la figura 3.4.

Existen casos especiales en las colas con sobreflujo en las cuales, los clientes entran si y solo si hay cuando menos un servidor libre, para iniciar el servicio inmediatamente.

3.4 Arribos Poisson y Tiempos de Servicio Exponenciales.

El modelo más simple y también el más extensamente estudiado es aquel que tiene arribos Poisson y tiempos de servicio con distribución exponencial. En este caso el largo de la cola, forma un proceso de nacimiento y muerte (revisado en la Subsección 2.4.1) y la distribución estacionaria correspondiente es conocida.

Sea λ la intensidad o la tasa de arribos del proceso Poisson y supongase que la distribución del tiempo de servicio es exponencial con parámetro μ.

La función de densidad correspondiente es:

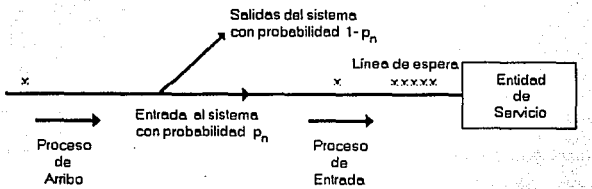

Figura 3.4: *Si n clientes están esperando en el sistema de líneas de espera con abortos, la llegada de un cliente produce una entrada al sistema con una probabilidad p_n, y una no entrada o aborto con probabilidad $1 - p_n$.*

$$g(x) = \mu e^{-\mu x} \quad \text{para } x > 0 \tag{3.37}$$

Para el proceso de arribos Poisson se tiene que :

$$\Pr\{\text{Un arribo en } [t, t+h)\} = \lambda h + o(h) \tag{3.38}$$

y

$$\Pr\{\text{No arribos en } [t, t+h)\} = 1 - \lambda h + o(h) \tag{3.39}$$

De forma similar, la propiedad de pérdida de la memoria de la distribución exponencial (Subsección 2.3.3) que se expresa por su tasa, implica

$$\Pr\{\text{Se complete un servicio en } [t, t+h) / \text{El servicio se inició en el momento } t\} = \mu h + o(h) \tag{3.40}$$

y

$$\Pr\{\text{El servicio no se complete en } [t, t+h) / \text{el servicio se inició en } t\} = 1 - \mu h + o(h). \tag{3.41}$$

La tasa de servicio μ se aplica a un servidor particular. Si k servidores están simultaneamente operando, la probabilidad de que uno de ellos complete un servicio en un intervalo de tiempo de duración h es $(k\mu)h + o(h)$ por eso la tasa del sistema de servicio es $k\mu$.

El principio utilizado aquí es el mismo usado en la derivación de los parámetros infinitesimales del proceso de Yule (Subsección 2.4.2).

Sea $X(t)$ el número de clientes en el sistema en el momento t, contando a los clientes que están en el servicio como un cliente esperando un servicio.

La independencia de los arribos en intervalos de tiempo disjuntos, y la propiedad de pérdida de la memoria de los servicios exponenciales, implican que $X(t)$ es una cadena de Markov homogénea en el tiempo, en particular, un proceso de nacimiento y muerte.

3.4.1 El Sistema $M/M/1$.

Considerese el caso de un sólo servidor. sea $X(t)$ el número de clientes en el sistema en el momento t. El incremento de $X(t)$ en una unidad corresponde al arribo de un cliente, y tomando en cuenta la ecuación (3.38) y (3.41) y los postulados de independencia en los tiempos de servicio y en los arribos, se tiene:

$$\Pr\{X(t+h) = k+1 / X(t) = k\} = [\lambda h + o(h)] \times [1 - \mu h + o(h)] = \lambda h + o(h) \text{ para } k = 0, 1, \cdots.$$

De forma similar el decrecimiento en una unidad de $X(t)$ corresponde a la completación de un servicio, entonces

$$\Pr\{X(t+h) = k-1 / X(t) = k\} = \mu h + o(h) \text{ para } k = 1, 2, \cdots.$$

Entonces $X(t)$ es un proceso de nacimiento y muerte con parámetros de nacimiento

$$\lambda_k = \lambda \text{ para } k = 0, 1, 2, \cdots,$$

y parámetros de muerte

$$\mu_k = \mu \text{ para } k = 1, 2, \cdots.$$

No se puede contemplar ningún servicio si la cola está vacía; en ese caso $\mu_0 = 0$.

Sea

$$\pi_k = \lim_{t \to \infty} \Pr\{X(t) = k\} \text{ para } k = 0, 1, \cdots.$$

la distribución límite o de equilibrio del largo de la cola. El procedimiento para determinar la distribución límite de π_k apartir de los parámetros de nacimiento y muerte λ_k y μ_k, consiste, en obtener primeramente las cantidades intermedias 0_j definidas como:

$$0_0 = 1 \text{ y } 0_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \quad \text{para } j \geq 1 \tag{3.42}$$

y entonces

$$\pi_0 = \frac{1}{\sum_{j=0}^{\infty} 0_j} \text{ y } \pi_k = 0_k \pi_0 = \frac{0_k}{\sum_{j=0}^{\infty} 0_j} \tag{3.43}$$

Cuando $\sum_{j=0}^{\infty} 0_j = \infty$, entonces el $\lim_{t \to \infty} \Pr\{X(t) = k\} = 0$, y el largo de la cola crece sin límites en el tiempo.

Para la cola $M/M/1$ se usará $0_0 = 1$ y $0_j = (\frac{\lambda}{\mu})^j$ para $j = 1, 2, \cdots$. Entonces

$$\sum_{j=0}^{\infty} \pi_j = \sum_{j=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^j = \frac{1}{1 - \lambda/\mu} \text{ si } \lambda < \mu$$

$$= \infty \text{ si } \lambda \geq \mu$$

Esto implica que la distribución de equilibrio no existe cuando la tasa de arribos λ es igual o mayor que la tasa de servicios μ, en este caso el largo de la cola crece sin límite.

Cuando $\lambda < \mu$ existe una distribución límite dada por

$$\pi_0 = \frac{1}{\sum 0_j} = 1 - \frac{\lambda}{\mu} \tag{3.44}$$

y

$$\pi_k = \pi_0 0_k = \left(1 - \frac{1}{\frac{\lambda}{\mu}}\right)\left(\frac{\lambda}{\mu}\right)^k \quad \text{para } k = 0, 1, \cdots \tag{3.45}$$

La distribución de equilibrio (3.45) brinda la respuesta a muchas preguntas que involucran el comportamiento límite del sistema. Tomando la forma de la ecuación (3.45) como la de la distribución geométrica, y de aquí podemos tomar el tamaño esperado de la cola en equilibrio, que estaría dado por

$$L = \frac{\lambda}{\mu - \lambda} \tag{3.46}$$

La razón $\rho = \frac{\lambda}{\mu}$ es llamada *intensidad de tráfico*

$$\rho = \frac{\text{Tasa de arribos}}{\text{Tasa de servicios}} = \frac{\lambda}{\mu} \tag{3.47}$$

Conforme la intensidad de tráfico se aproxima a uno, el largo promedio de la cola $L = \rho(1-\rho)$ tiende a infinito.

Nuevamente usando (3.42) la probabilidad de ser servido inmediatamente es:

$$\pi_0 = 1 - \frac{\lambda}{\mu},$$

la probabilidad, en el comportamiento a futuro, de encontrar un servidor desocupado. El período de ocupación, de un servidor, es $1 - \pi_0 = \frac{\lambda}{\mu}$.

Se puede calcular también la distribución del tiempo de espera en el caso estacionario, cuando $\lambda < \mu$. Si un cliente que arriba al sistema encuentra n personas enfrente de él, su tiempo total de espera T, incluyendo el tiempo de servicio, es la suma de los tiempos de servicio de él y de las demás personas de enfrente de él, todos distribuyéndose exponencialmente con parámetro μ, y dado que los tiempos de servicio son independientes del tamaño de la cola, W tiene una distribución gamma de orden $n+1$ con parámetro μ.

$$\Pr\{T \le t/n \text{ enfrente}\} = \int_0^t \frac{\mu^{n+1}\tau^n e^{-\mu\tau}}{\Gamma(n+1)} d\tau \tag{3.48}$$

Por las leyes de probabilidad se tiene:

$$\Pr\{T \le t\} = \sum_{n=0}^{\infty} \Pr\{T \le t/n \text{ enfrente}\} \times \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

dado que $(\frac{\lambda}{\mu})^n(1-\frac{\lambda}{\mu})$ es la probabilidad en el caso estacionario, de que un cliente al arribar encuentre n clientes en la línea. Ahora sustituyendo (3.48) se tiene.

$$\begin{aligned}\Pr[T \le t] &= \sum_{n=0}^{\infty} \int_0^t \frac{\mu^{n+1}\tau^n e^{-\mu\tau}}{\Gamma(n+1)} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) d\tau \\ &= \int_0^t \mu e^{-\mu\tau} \left(1 - \frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \frac{\tau^n \lambda^n}{\Gamma(n+1)} d\tau\end{aligned}$$

$$= \int_0^t \left(1 - \frac{\lambda}{\mu}\right) \mu e^{-\tau\mu(1-\frac{\lambda}{\mu})} d\tau$$

$$= 1 - e^{-t(\mu-\lambda)}$$

que es también una distribución exponencial.

La media de la distribución exponencial del tiempo de espera, es el recíproco del parámetro de la exponencial; es decir,

$$W = \frac{1}{\mu - \lambda} \tag{3.49}$$

Si se observan un poco las ecuaciones (3.46) y (3.49) se verifica la fórmula

$$L = \lambda W.$$

Un sistema de líneas de espera es un proceso alternante entre períodos ocupados y períodos desocupados, los períodos desocupados se inician en el instante en que el último cliente sale del sistema y termina cuando arriba el siguiente cliente. Cuando los procesos de llegada son Poisson con tasa λ, los períodos desocupados se distribuyen exponencialmente con media

$$E[I_1] = \frac{1}{\lambda}.$$

Un período ocupado es un período ininterrumpido durante el cual el sistema no está vacío. Cuando los arribos a la cola siguen un proceso Poisson, la duración sucesiva X_k, del inicio del k-ésimo período ocupado hasta el comienzo del siguiente período ocupado forma un proceso de renovación (Figura 3.5). Cada X_k se compone de un período ocupado B_k, y un período desocupado I_k. Entonces el teorema de renovación (ver "un modelo de cola" Ejemplo de la Sección 2.4) se aplica y nos dice que $P_0(t)$, la probabilidad de que el sistema esté vacío en el tiempo t, converge a

$$\lim_{t\to\infty} p_0(t) = \pi_0 = \frac{E[I_1]}{E[I_1] + E[B_1]}.$$

Sustituyendo las cantidades conocidas $\pi_0 = 1 - \frac{\lambda}{\mu}$ y $E[I_1] = \frac{1}{\lambda}$ para obtener

$$1 - \frac{\lambda}{\mu} = \frac{1/\lambda}{1/\lambda + E[B_1]}$$

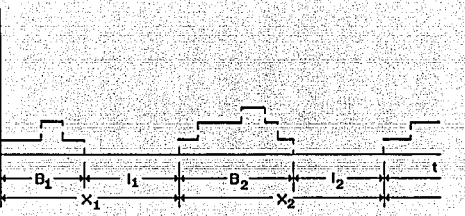

Figura 3.5: *Los períodos ocupados B_k y los desocupados I_k de un sistema de líneas de espera. Cuando los arribos forman un proceso Poisson, entonces $X_k = B_k + I_k$, $k = 1, 2, \cdots$ son variables aleatorias independientes no-negativas e idénticamente distribuidas, y forman un proceso de renovación*

que se resuelve para dar

$$E[B_1] = \frac{1}{\mu - \lambda}$$

que es la media de la duración del período ocupado

En la sección 3.5 en el estudio del modelo $M/G/1$ se resolverá este razonamiento, calculando la media del período ocupado directamente, y usando la teoría de renovación para determinar la fracción de desocupación del servidor π_0.

3.4.2 El Sistema $M/M/\infty$.

Cuando un número ilimitado de servidores están siempre disponibles, los clientes en el sistema son servidos en cualquier instante. La tasa de servicio de un cliente es μ, la tasa de servicios de k clientes es $k\mu$, y obtenemos los parámetros del proceso de nacimiento y muerte

$$\lambda_k = \lambda \quad y \quad \mu_k = k\mu \quad \text{para} \quad k = 1, 2, \cdots.$$

Las cantidades auxiliares de (3.42) son

$$\theta_k = \frac{\lambda_0\lambda_1\cdots\lambda k-1}{\mu_1\mu_2\cdots\mu_k} = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \quad \text{para} \quad k = 1, 2, \cdots,$$

la cual sumada

$$\sum_{k=0}^{\infty}\theta_k = \sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k = e^{\lambda/\mu}$$

entonces

$$\pi_0 = \frac{1}{\sum_{k=0}^{\infty}\theta_k} = e^{-\lambda/\mu}$$

y

$$\pi_k = \theta_k\pi_0 = \frac{(\lambda/\mu)^k e^{-\lambda/\mu}}{k!} \quad \text{para} \quad k = 0, 1, \cdots, \tag{3.50}$$

una distribución Poisson cuya media coincide con la longitud de la cola

$$L = \lambda W$$

Dado que un cliente en este sistema inicia sus servicios en cuanto arriba, el tiempo de espera de los clientes, consiste únicamente en la distribución exponencial del tiempo de servicio, y la media del tiempo de espera es $W = 1/\mu$. Nuevamente, la fórmula $L = \lambda W$ se cumple.

3.4.3 El Sistema $M/M/s$

Cuando se tiene un número s, fijo de servidores y suponiendo que un sevidor nunca esta desocupado si hay un cliente esperando, entonces los parámetros para el proceso de nacimiento y muerte son aproximados a

$$\begin{aligned} \lambda_k &= \lambda \quad \text{para} \quad k = 1, 2, \cdots \\ \mu_k &= \begin{cases} k\mu & k = 0, 1, \cdots, s \\ s\mu & \text{para } k > s \end{cases} \end{aligned}$$

Si $X(t)$ es el número de clientes en el sistema en el tiempo t, entonces el número de clientes que están en servicio es el $\min\{X(t), s\}$ y el número de clientes esperando es el $\max\{X(t) - s, 0\}$. El sistema se ilustra en la Figura 3.6

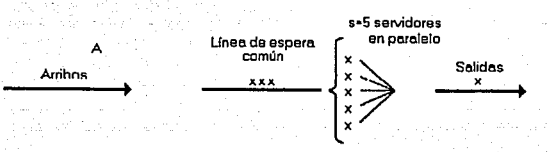

Figura 3.6: *Un sistema de líneas de espera con s servidores*

Las cantidades auxiliares se dan como:

$$\theta_k = \frac{\lambda_0\lambda_1,\cdots,\lambda_{k-1}}{\mu_1\mu_2\cdots,\mu_k} = \begin{cases} \frac{1}{k!}(\frac{\lambda}{\mu})^k & \text{para } k = 0, 1, \cdots, s \\ \frac{1}{s!}(\frac{\lambda}{\mu})^s(\frac{\lambda}{s\mu})^{k-s} & \text{para } k > s \end{cases}$$

y cuando $\lambda < s\mu$, se tiene:

$$\sum_{j=0}^{\infty}\theta_j = \sum_{j=0}^{s-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j + \sum_{j=s}^{\infty}\frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s\left(\frac{\lambda}{s\mu}\right)^{j-s} \tag{3.51}$$

$$= \sum_{j=0}^{s-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j + \frac{(\lambda/\mu)^s}{s!(1-\lambda/s\mu)} \quad \text{para } \lambda < s\mu.$$

La intensidad de tráfico de un sistema $M/M/s$ es $\rho = \lambda/s\mu$. Nuevamente cuando la intensidad de tráfico se aproxima a uno, la longitud de la cola se dispara. Cuando $\lambda < s\mu$, entonces de (3.43) y (3.52)

$$\pi_0 = \left\{\sum_{j=0}^{s-1}\frac{1}{j!}(\frac{\lambda}{\mu})^j + \frac{(\lambda/\mu)^s}{s!(1-\lambda/s\mu)}\right\}^{-1}$$

y

$$\pi_k = \begin{cases} \frac{1}{k!}(\frac{\lambda}{\mu})^k \pi_0 & \text{para } k = 0, 1, \cdots, s \\ \frac{1}{s!}(\frac{\lambda}{\mu})^s (\frac{\lambda}{s\mu})^{k-s} \pi_0 & \text{para } k \geq s \end{cases} \tag{3.52}$$

Evaluando L_0 (el número esperado de clientes en el sistema esperando un servicio).

$$\begin{aligned} L_0 &= \sum_{j=s}^{\infty} (j-s)\pi_j = \sum_{k=0}^{\infty} k\pi_{k+s} \\ &= \pi_0 \sum_{k=0}^{\infty} k \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^k \\ &= \frac{\pi_0}{s!} \left(\frac{\lambda}{\mu}\right)^s \frac{(\lambda/s\mu)}{(1-\lambda/s\mu)^2} \end{aligned} \tag{3.53}$$

Entonces

$$W_0 = \frac{L_0}{\lambda},$$

$$W = W_0 + \frac{1}{\mu}$$

y

$$L = \lambda W = \lambda \left(W_0 + \frac{1}{\mu}\right) = L_0 + \frac{\lambda}{\mu}.$$

3.5 El Sistema $M/G/1$ y El Sistema $M/G/\infty$.

Continuando en el supuesto de que los arribos siguen un proceso de Poisson de tasa λ. Los tiempos de servicio sucesivos $Y_1, Y_2, \cdots$; ahora seguirán una distribución arbitraria $G(\gamma) = \Pr\{Y_k \leq \gamma\}$ con tiempo medio de servicio finito $\nu = E[Y_k]$. La tasa de servicio a futuro es $\mu = \frac{1}{\nu}$, los tiempos de servicio determinísticos de duración fija son un caso especial importante.

3.5.1 El Sistema $M/G/1$.

Si los arribos a una cola siguen un proceso Poisson, entonces la duración sucesiva X_k, del comienzo del k-ésimo período ocupado hasta el inicio del siguiente período ocupado, forma un proceso de renovación. (Un período ininterrumpido que dura mientras la cola no está vacía ver Figura 3.5). Cada X_k está compuesta de un período ocupado B_k, y de un período desocupado I_k. Entonces $p_o(t)$, la probabilidad de que el sistema esté vacío en el tiempo t, converge a

$$\lim_{t\to\infty} p_0(t) = \pi_0 = \frac{E[I_1]}{E[X_1]} = \frac{E[I_1]}{E[I_1] + E[B_1]} \tag{3.54}$$

por el teorema de renovación (ver "un modelo de cola" Ejemplo de la Sección 2.4).

El tiempo "desocupado" es la duración que va de la completación del servicio del último cliente en el sistema hasta el instante del siguiente arribo. Dada la propiedad de "pérdida de la memoria" que caracteriza los tiempos entre arribos en proceso Poisson, cada período desocupado se distribuye exponencialmente con media $E[I_1] = 1/\lambda$.

El período ocupado comprende desde el inicio del primer arribo Y_1, más el tiempo de ocupación que se requiere para dar servicio a todos los clientes que llegan durante este primer tiempo de servicio.

Sea A, que denota este número aleatorio de nuevos arribos. Evaluando la media condicional del período ocupado dado por $A = n$ y $Y_1 = \gamma$. Primero

$$E[B_1/A = 0, Y_1 = \gamma] = \gamma$$

porque no hay arribos, el período ocupado comprende únicamente el tiempo de servicio del primer cliente. Considere el caso en que $A = 1$ y sea B' la duración desde el comienzo del servicio del primer cliente, hasta el momento que la cola está vacía. Entonces

$$\begin{aligned} E[B_1/A = 1, Y_1 = \gamma] &= \gamma + E[B'] \\ &= \gamma + E[B_1], \end{aligned}$$

porque después la completación del servicio del cliente inicial, un arribo inicializa un período ocupado B', que es estadísticamente idéntico al primero, por eso $E[B'] = E[B_1]$. Y de este modo, se deduce:

$$E[B_1/A = n, Y_1 = \gamma] = \gamma + nE[B_1]$$

y usando la ley de probabilidades totales, tenemos que

$$\begin{aligned} E[B_1/Y_1 = \gamma] &= \sum_{n=0}^{\infty} E[B_1/A = n, Y_1 = \gamma] \Pr\{A = n/Y_1 = \gamma\} \\ &= \sum_{n=0}^{\infty} \{\gamma + nE[B_1]\} \frac{(\lambda\gamma)^n e^{-\lambda\gamma}}{n!} \\ &= \gamma + \lambda\gamma E[B_1]. \end{aligned}$$

Finalmente

$$\begin{aligned} E[B_1] &= \int_0^{\infty} E[B_1/Y_1 = \gamma] dG(\gamma) \\ &= \int_0^{\infty} \{\gamma + \lambda\gamma E[B_1]\} dG(\gamma) \qquad (3.55) \\ &= \nu\{1 + \lambda E[B_1]\}. \end{aligned}$$

Ya que $E[B_1]$ aparece en ambos lados de la ecuación (3.55) se resuelve para obtener

$$E[B_1] = \frac{\nu}{1 - \lambda\nu} \quad \text{para } \lambda\nu < 1. \qquad (3.56)$$

Para calcular la fracción del comportamiento a futuro del tiempo desocupado se usa (3.55) y

$$\begin{aligned} \pi_0 &= \frac{E[I_1]}{E[I_1] + E[B_1]} \\ &= \frac{1/\lambda}{1/\lambda + \nu/(1 - \lambda\nu)} \qquad (3.57) \\ &= 1 - \lambda\nu \quad \text{si } \lambda\nu < 1. \end{aligned}$$

Nótese que (3.58) coincide, con la expresión (3.44) obtenida para el sistema $M/M/1$ donde $\nu = 1/\mu$. Por ejemplo; si los arribos ocurren con la tasa $\lambda = 2$ y el tiempo medio

de servicio es de 20 minutos o $\nu = \frac{1}{3}$, entonces en el comportamiento a futuro, el servidor está ocupado $1 - 2(\frac{1}{3}) = \frac{1}{3}$ del tiempo.

3.5.2 Cadenas de Markov Ajustadas.

$X(t)$, el número de clientes en el sistema en el tiempo t, no define un proceso de Markov para el sistema $M/G/1$, porque si este está para predecir el comportamiento futuro del sistema, este debe predecir además el tiempo gastado en el servicio por el cliente que lo está recibiendo. (La propiedad de "pérdida" de la memoria del tiempo de servicio exponencial que brinda esta información adicional hace innecesaria esta pregunta en el sistema $M/M/1$).

Sea X_n el número de clientes en el sistema inmediatamente después de la salida del n–ésimo cliente. Entonces $\{X_n\}$ es una cadena de Markov, de hecho:

$$\begin{aligned} X_n &= \begin{cases} X_{n-1} - 1 + A_n & \text{si } X_{n-1} > 0 \\ A_n & \text{si } X_{n-1} = 0 \end{cases} \\ &= (X_{n-1} - 1)^+ + A_n, \end{aligned} \tag{3.58}$$

donde A_n es el número de clientes que arriban durante el servicio de n–ésimo cliente y donde $x^+ = \max\{x, 0\}$. Dado que el proceso de arribo es Poisson, el número de clientes A_n que arriban durante el servicio del N–ésimo cliente es independiente de los arribos anteriores, la propiedad Markoviana aparece de inmediato. Calculando

$$\begin{aligned} \Pr\{A_n = k\} &= \int_0^\infty \Pr\{A_n = k / Y_n = \gamma\} g G(\gamma) \\ &= \int_0^\infty \frac{(\lambda\gamma)^k e^{\lambda\gamma}}{k!} dG(\gamma). \end{aligned} \tag{3.59}$$

y para $j = 0, 1, \cdots$,

$$\begin{aligned} P_{ij} &= \Pr\{X_n = j / X_{n-1} = i\} = \Pr\{A_n = j - (i+1)^+\} \\ &= \begin{cases} \alpha_{j-i+1} & \text{para } i \geq 1, j \geq i+1 \\ \alpha_j & \text{para } i = 0 \end{cases} \end{aligned} \tag{3.60}$$

3.5.3 Longitud L Promedio de una Cola en Equilibrio.

La cadena de Markov ajustada es de especial interes en el sistema $M/G/1$ porque particularmente en este caso, la distribución estacionaria $\{\pi_j\}$ para la cadena de Markov $\{X_n\}$ iguala a la distribución límite para el proceso de longitud de la cola $\{X(t)\}$. Es decir, $\lim_{t\to\infty} \Pr\{X(t) = j\} = \lim_{n\to\infty} \Pr\{X_n = j\}$. Se utiliza este hecho auxiliar para evaluar la longitud de la cola.

La equivalencia entre la distribución estacionaria de una cadena de Markov $\{X_n\}$ y aquella para el proceso no Markoviano $\{\{X(t)\}\}$ es muy sutil. No es consecuencia de un principio general y no se debe suponer en otras circunstancias sin una justificación cuidadosa. La equivalencia en este caso está bosquejada en el apendice de esta sección

Se calculará la longitud esperada L, de la cola en equilibrio $L = \lim_{t\to\infty} E[X(t)]$ calculando las cantidades correspondientes en la cadena de Markov ajustada, $L = \lim_{n\to\infty} E[X_n]$. Si $X = X_\infty$ es el número de clientes en el sistema después de la salida de un cliente y X' es el número despues de la siguiente salida, entonces por la ecuación (3.59)

$$X' = X - \delta + N \tag{3.61}$$

donde N es el número de arribos durante el período de servicios y

$$\delta = \begin{cases} 1 & \text{si } X > 0 \\ 0 & \text{si } X = 0 \end{cases}$$

en equilibrio, X tiene la misma distribución que X' y, en particular

$$L = E[X] = E[X'], \tag{3.62}$$

y sacando la esperanza en (3.61) da

$$E[X'] = E[X] - E[\delta] + E[N],$$

y por (3.58) y (3.62), se tiene:

$$E[N] = E[\delta] = 1 - \pi_0 = \lambda\nu \tag{3.63}$$

elevando al cuadrado (3.61) se tiene:

$$(X')^2 = X^2 + \delta^2 + N^2 - 2\delta X + 2N(X - \delta)$$

y, dado que $\delta^2 = -\delta$ y $X\delta = X$, entonces

$$(X')^2 = X^2 + \delta + N^2 - 2X + 2N(X - \delta). \tag{3.64}$$

Ahora N, el número de clientes que arriban durante un periodo de servicio, es independiente de X, y por ello también de δ así que:

$$E[N(X - \delta)] = E[N]E[X - \delta] \tag{3.65}$$

y dado que X y X' tienen la misma distribución, entonces

$$E[(X')^2] = E[X^2]. \tag{3.66}$$

Tomando las esperanzas en la ecuación (3.64) se deduce

$$E[(X')^2] = E[X^2] + E[\delta] + E[N^2] - 2E[X] + 2E[N]E[X - \delta]$$

y sustituyendo (3.63) y (3.66) se tiene

$$0 = \lambda\nu + E[N^2] - 2L + 2\lambda\nu\{L - \lambda\nu\}$$

ó

$$L = \frac{\lambda\nu + E[N^2] - 2(\lambda\nu)^2}{2(1 - \lambda\nu)}. \tag{3.67}$$

Para evaluar $E[N^2]$, donde N es en número de arribos durante un tiempo de servicio Y. Condicionando $Y = \gamma$, la variable aleatoria N tiene una distribución Poisson con media igual a $\lambda\gamma$ [ver (3.59)], por ello $E[N^2/Y = \gamma] = \lambda\gamma + (\lambda\gamma)^2$. Utilizando la ley de probabilidades totales se tiene

$$\begin{aligned} E[N^2] &= \int_0^\infty E[N^2/Y = \gamma]dG(\gamma) \\ &= \lambda\int_0^\infty \gamma dG(\gamma) + \lambda^2\int_0^\infty \gamma^2 dG(\gamma) \\ &= \lambda\nu + \lambda 2(\tau^2 + \nu^2), \end{aligned} \tag{3.68}$$

donde τ es la varianza de la distribución del tiempo de servicio $G(\gamma)$. Sustituyendo (3.69) y (3.67) se obtiene:

$$L = \frac{2\lambda\nu + \lambda^2\tau^2 - (\lambda\nu)^2}{2(1 - \lambda\nu)}$$

$$= \rho + \frac{\lambda^2\tau^2 + \rho^2}{2(1-\rho)} \tag{3.69}$$

donde $\rho = \lambda\nu$ es la intensidad de tráfico.

Finalmente, $W = L/\lambda$, que se simplifica como

$$W = \nu + \frac{\lambda(\tau^2 + \nu^2)}{2(1-\rho)}. \tag{3.70}$$

Los resultados (3.69) y (3.70) expresan de alguna forma hechos interesantes. Dicen que para una tasa promedio de arribos dada λ y un tiempo de servicio promedio ν, se puede disminuir el tamaño esperado de la cola L, y también disminuir el tiempo de espera W disminuyendo la varianza del tiempo de servicio. Claramente se observa que el mejor caso posible en este aspecto corresponde a el caso en el que el tiempo de servicio es constante, en cuyo caso $\nu^2 = 0$.

3.5.4 El Sistema $M/G/\infty$.

Existen resultados completos en el caso en que cada cliente inicia su servicio inmediatamente y además los arribos son independientes sobre los otros clientes del sistema.

Dicha situación puede ocurrir cuando se modelan sistemas de auto servicio. Sean $W_1, W_2, \cdots$; tiempos sucesivos de arribos de los clientes, y sean $V_1, V_2, \cdots$; los tiempos de servicio correspondientes. En esta notación el k–ésimo cliente está en el sistema en el tiempo t si y sólo si $W_k \leq t$ (el cliente arribó antes del momento t) y $W_k + V_k > t$ (el servicio va más allá de t).

La secuencia de parejas ordenadas $(W_1, V_1), (W_2, V_2), \cdots$; forman un proceso de Poisson Marcado (ver subsección 2.3.5) y se debe usar la teoría correspondiente para obtener resultados rapidamente en este modelo. La figura 3.7 muestra el proceso de Poisson Marcado.

$X(t)$,el número de clientes en el sitema en el tiempo t,es también el número de puntos (W_k, V_k) para los cuales $W_k \leq t$ y $W_k + V_K > t$.

Esto es el número de puntos (W_k, V_k) en el trapezoide no acotado, descrito por

$$A_t = \{(w; v); 0 \leq w \leq t \text{ y } v > t - w\}.$$

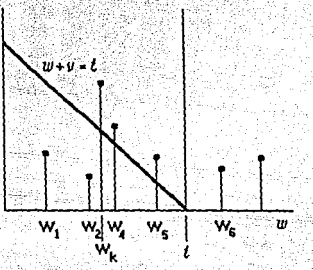

Figura 3.7: *Para el modelo $M/G/\infty$ el número de clientes en el sistema en el tiempo t corresponde al número de parejas ordenadas (W_k, V_k) para los cuales $W_k < t$ y $W_k + V_k > t$. (En esta figura se representa a tres clientes en el sistema en el tiempo t).*

El número de puntos en A_t sigue una distribución Poisson con media

$$\begin{aligned} \mu(A_t) &= \int\int_{A_t} \lambda(dw)dG(v) \\ &= \lambda \int_0^t \left\{ \int_{t-w}^{\infty} dG(v) \right\} dw \\ &= \lambda \int_0^t [1 - G(t-w)]dw \\ &= \lambda \int_0^t [1 - G(x)]dx. \end{aligned} \tag{3.71}$$

en resumen

$$\begin{aligned} p_k(t) &= \Pr\{X(t) = k\} \\ &= \frac{\mu(A_t)^k e^{-\mu(A_t)}}{k!} \quad \text{para } k = 0, 1, \cdots \end{aligned}$$

donde $\mu(A_t)$ está dada por (3.72). Como $t \to \infty$ se tiene

$$\lim_{t\to\infty} \mu(A_t) = \lambda \int_0^{\infty} [1 - G(x)]dx = \lambda\nu$$

donde ν es el tiempo medio de servicio. Así se obtiene la distribución límite

$$\pi_k = \frac{(\lambda\nu)^k e^{-\lambda\nu}}{k!} \quad \text{para} \quad k = 0, 1, \cdots.$$

3.5.5 Apéndice

Se esbozará una prueba de la equivalencia entre la distribución límite del tamaño de la cola y la distribución límite de una cadena de Markov Acoplada en un modelo $M/G/1$. Primero, iniciando en $t = 0$ sea η_n que denota los instantes en los que el tamaño de la cola $X(t)$ se incremente en uno, (un arribo), y sea ξ_n que denota los instantes en los que el tamaño de la cola $X(t)$ decrece en uno, (una salida). Sea $Y_n = X(\eta_n-)$ que denota el largo de la cola inmediatamente antes de una arribo. Y sea $X_n = (\xi_n+)$ que denota el tamaño de la cola inmediatamente despues de una salida. Para cualquier largo de la cola i y en cualquier tiempo t el número de visitas de Y_n a i sobre el tiempo t difiere del número de visitas de X_n a i cuando más por una unidad. Por ello, en el comportamiento a futuro, el promedio de visitas por unidad de tiempo de Y_n a i debe igualar el promedio de visitas de X_n a i, el cual es π_i, la distribución estacionaria de la cadena de Markov $\{X_n\}$. Por ello únicamente se necesita mostrar que la distribución límite de $\{X_n\}$ es la misma que la de $\{Y_n\}$, $X(t)$ es anterior sólo por un arribo. Pero dado que los arribos son Poisson, y estos arribos en intervalos disjuntos de tiempo son independientes, debe suceder que $X(t)$ es independiente de una arribo que ocurra en el tiempo t. Se sigue que $\{X(t)\}$ y $\{Y_n\}$ tienen la misma distribución límite, y por ello $\{X(t)\}$ y la cadena de Markov acoplada $\{X_n\}$ tienen la misma distribución límite.

3.6 Variaciones y Extensiones

En esta sección se conciderarán algunas variaciones sobre los modelos de líneas de espera. Estos ejemplos no agotan todas las posibilidades pero sirven para sugerir la vastedad del area.

Restringiéndose a los arribos Poisson y tiempos de servicio exponencialmente distribuidos.

3.6.1 Sistemas con Abortos

Suponga que un cliente que arriba cuando existen n clientes en el sistema, entra con una probabilidad p_n y no entra con una probabilidad $q_n = 1 - p_n$. Si las líneas largas desaniman a los clientes, entonces, p_n será una función decreciente dependiente de n. Como caso especial, si se tiene un cuarto de espera con capcidad finita C, se debe suponer que:

$$p_n = \begin{cases} 1 & \text{para } n < C \\ 0 & \text{para } n \geq C \end{cases}$$

que indica que una vez que el cuarto de espera está lleno, ningún cliente más puede entrar al sistema.

Sea $X(t)$, el número de clientes en el sistema en el tiempo t. Si el proceso de arribo es Poisson con tasa λ, y un cliente arriba cuando hay n clientes en el sistema, entra con probabilidad p_n, entonces los parámetros apropiados de nacimiento son

$$\lambda_n = \lambda p_n \quad \text{para} \quad n = 0, 1, \cdots$$

En el caso de un sólo servidor, entonces $\mu_n = \mu$ para $n = 1, 2, \cdots$, y se debe evaluar la distribución estacionaria π_k del largo de la cola por el método usual.

En un sistema con abortos,no todos los clientes que arriban entran en el sistema, y algunos se pierden. la *tasa de entrada* es la tasa la cual los clientes realmente entran en el sistema en el estado estacionario y está dada por:

$$\lambda_I = \lambda \sum_{n=0}^{\infty} \pi_n p_n.$$

La tasa a la cual los clientes son perdidos es $\lambda \sum_{n=0}^{\infty} \pi_n q_n$ y la fracción de clientes perdidos en el comportamiento a futuro es

$$\text{Fracción Perdida} = \sum_{n=0}^{\infty} \pi_n q_n.$$

Examinando ahora en detalle el caso de un sistema $M/M/s$ en el cual los clientes que arriban entran en el sistema si y sólo si un servidor está desocupado. Entonces:

$$\lambda_k = \begin{cases} \lambda & \text{para } k = 0, 1, \cdots, s-1 \\ 0 & \text{para } k = s \end{cases}$$

y

$$\mu_k = k\mu \quad \text{para} \quad k = 0, 1, \cdots, s.$$

Para determinar la distribución límite, se tiene:

$$0_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \quad \text{para} \quad k = 0, 1, \cdots, s$$

y entonces:

$$\mu_k = \frac{\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}{\sum_{j=0}^{s}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j} \quad \text{para} \quad k = 0, 1, \cdots, s. \tag{3.72}$$

La fracción a futuro de clientes perdidos es $\pi_s q_s = \pi_s$ dado que $q_s = 1$ en este caso.

3.6.2 Tasa de Servicio Variables.

De modo similar, uno puede considerar un sistema cuya tasa de servicio depende del número de clientes en el sistema, en cualquier momento que la línea exceda un punto crítico de longitud ξ.

Si los arribos son Poisson y las tasas de servicio tienen la propiedad de "pérdida de la memoria," entonces los parámetros de nacimiento y muerte son:

$$\lambda_k = \lambda \quad \text{para} \quad k = 0, 1, \cdots \quad \text{y} \quad \mu_k = \begin{cases} \mu & \text{para} \quad k \leq \xi \\ 2\mu & \text{para} \quad k > \xi \end{cases}$$

Más general, considere arribos Poisson $\lambda_k = \lambda$ para $k = 0, 1, \cdots$; y tasas arbitrarias de servicio μ_k para $k = 1, 2, \cdots$ La distribución estacionaria en este caso está dada por

$$\pi_k = \frac{\pi_0 \lambda^k}{\mu_1 \mu_2, \cdots, \mu_k} \quad \text{para} \quad k \geq 1 \tag{3.73}$$

donde

$$\pi_0 = \left\{1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{\mu_1 \mu_2, \cdots, \mu_k}\right\}^{-1} \tag{3.74}$$

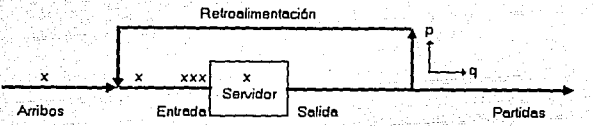

Figura 3.8: *Una Cola con retroalimentación*

3.6.3 Un Sistema con Retroalimentación.

Considere un sistema con un sólo servidor, con arribos Poisson y tiempos de servicio distribuidos exponencialmente, pero suponga que algunos clientes, después de haber salido del servicio, regresan al final de la cola por un servicio adicional, con probabilidad $p = 1 - q$. Suponiendo ahora, que todas las decisiones son estadísticamente independientes, y que el retorno de los clientes a demandar un servicio son estadísticamente los mismos que aquellos clientes que llegan de fuera del sistema. Dejemos que la tasa de arribo sea λ, y la tasa de servicio sea μ. El sistema es descrito en la figura 3.8.

Sea $X(t)$, que denota el número de clientes en el sistema en el tiempo t, entonces $X(t)$ es un proceso de nacimiento y muerte con parámetros $\lambda_n = \lambda$ para $n = 0, 1, \cdots$, y $\mu_n = q\mu$ para $n = 1, 2, \cdots$. Se puede deducir que la distribución en en el caso en que $\lambda < q\mu$ es

$$\pi_k = \left(1 - \frac{\lambda}{q\mu}\right)\left(\frac{\lambda}{q\mu}\right)^k \quad \text{para} \quad k = 0, 1, \cdots \tag{3.75}$$

3.6.4 Una Cola con Dos Servidores y Sobreflujo

Conciderando un sistema con dos servidores, donde el servidor i, tiene una tasa de servicio μ_i para $i = 1, 2$,. Los arribos al sistema siguen un proceso de Poisson con tasa λ. Un cliente que llega cuando el sistema está vacío va al primer servidor. Un cliente que arriba

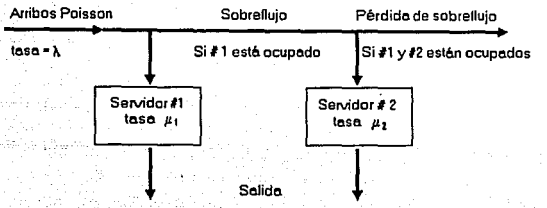

Figura 3.9: *Un modelo de sobreflujo con dos servidores*

cuando el primer servidor está ocupado va con el segundo servidor. Si ambos servidores están ocupados, el cliente se pierde. El flujo se especifica en la figura 3.9.

El estado del sistema es descrito por la pareja $(X(t), Y(t))$ donde

$$X(t) = \begin{cases} 1 & \text{si el servidor \#1 está ocupado} \\ 0 & \text{si el servidor \#1 está desocupado} \end{cases}$$

y

$$Y(t) = \begin{cases} 1 & \text{si el servidor \#2 está ocupado} \\ 0 & \text{si el servidor \#2 está desocupado} \end{cases}$$

Los cuatro estados del sistema son $\{(0,0);(1,0);(0,1);(1,1)\}$, y las transformaciones entre estos estados ocurren con las tasa dadas en la siguiente Tabla 3.1

El proceso $(X(t), Y(t))$ es una cadena de markov continua en el tiempo con estado finito (ver Sección 2.4) y las tasas de transicion en las tablas forman una matriz infinitesimal de la cadena de Markov:

$$A = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \left\| \begin{array}{cccc} -\lambda & 0 & \lambda & 0 \\ \mu_2 & -(\lambda+\mu_2) & 0 & \lambda \\ \mu_1 & 0 & -(\lambda+\mu_1) & \lambda \\ 0 & \mu_1 & \mu_2 & -(\mu_1+\mu_2) \end{array} \right\| \end{array}.$$

Del Estado	Al Estado	Tasa de Transición	Descripción
(0,0)	(1,0)	λ	Arribo cuando el sistema está vacío
(1,0)	(0,0)	μ_1	Servicio completado por #1 cuando #2 está libre
(1,0)	(1,1)	λ	Arribo cuando #1 está ocupado
(1,1)	(1,0)	μ_2	Servicio completado por #2 cuando #1 está ocupado
(1,1)	(0,1)	μ_1	Servicio completado por #1 cuando #2 está ocupado
(0,1)	(1,1)	λ	Arribo cuando #2 está ocupado y #1 está libre
(0,1)	(0,0)	μ_2	Servicio completado por #2 cuando #1 está libre

Tabla 3.1: *Estados y transiciones que ocurren con las tasas dadas*

Encontrando la distribución estacionaria $\pi = (\pi_{(0,0)}, \pi_{(0,1)}, \pi_{(1,0)}, \pi_{(1,1)})$, resolviendo $\pi A = 0$, ó

$$
\begin{array}{rrrrr}
-\lambda\pi_{(0,0)} & +\mu_2\pi_{(0,1)} & +\mu_1\pi_{(1,1)} & & = 0 \\
 & -(\lambda+\mu_2)\pi_{(0,1)} & & +\mu_1\pi_{(1,1)} & = 0 \\
\lambda\pi_{(0,0)} & & -(\lambda+\mu_1)\pi_{(1,0)} & +\mu_2\pi_{(1,1)} & = 0 \\
 & \lambda\pi_{(0,1)} & +\lambda\pi(1,0) & -(\mu_1+\mu_2)\pi_{(1,1)} & = 0
\end{array}
$$

conjuntamente con

$$\pi_{(0,0)} + \pi_{(0,1)} + \pi_{(1,0)} + \pi(1,1) = 1$$

que conduce a la solución

$$
\begin{aligned}
\pi_{(0,0)} &= \frac{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)}{D} \\
\pi_{(0,1)} &= \frac{\lambda^2\mu_1}{D} \\
\pi_{(1,0)} &= \frac{\lambda\mu_2(\lambda+\mu_1+\mu_2)}{D} \\
\pi_{(1,1)} &= \frac{\lambda^2(\lambda+\mu_2)}{D}
\end{aligned}
\tag{3.76}
$$

donde

$$D = \mu_1\mu_2(2\lambda+\mu_1+\mu_2) + \lambda^2\mu_1 + \lambda\mu_2(\lambda+\mu_1+\lambda\mu_2) + \lambda^2(\lambda+\mu_2).$$

La fracción de clientes que son perdidos, a futuro, es la misma que la fracción de tiempo en que ambos servidores están ocupados, $\pi_{(1,1)} = \lambda^2(\lambda + \mu_2)/D$.

3.7 Colas Con Prioridad

Considerando un sistema con un sólo sevidor que tiene dos clases de clientes, "*prioritarios*" y "*no-prioritarios*" que forman arribos independientes que siguen un proceso Poisson con tasas α y β respectivamente. Los tiempos de servicio son independientes y exponencialmente distribuidos con parámetros γ y δ, respectivamente. Junto con estas clases hay una primera llegada, una disciplina de la cola y el servicio de un cliente prioritario nunca es interrumpido. Si un cliente prioritario arriba durante el servicio de uno no-prioritario, entonces el servicio del no-prioritario es inmediatamente interrumpido para atender al prioritario. El servicio interrumpido es concluido cuando no hay clientes prioritarios en el sistema. Se introduce un poco de notación. La tasa de arribo del sistema es $\lambda = \alpha + \beta$, la fracción $p = \alpha/\lambda$ de los clientes prioritarios, y $q = \delta/\lambda$ de los clientes no-prioritarios. El tiempo promedio de servicio del sistema está dado por la propiedad de promedios pesados $1/\gamma$ y $1/\delta$ de los clientes prioritarios y no-prioritarios respectivamente, o

$$\frac{1}{\mu} = p\left(\frac{1}{\gamma}\right) + q\left(\frac{1}{\delta}\right) = \frac{1}{\lambda}\left(\frac{\alpha}{\gamma} + \frac{\beta}{\delta}\right), \tag{3.77}$$

donde μ es la tasa de servicio del sistema. Finalmente introduciendo la intensidad de tráfico $\rho = \lambda/\mu$, y $\sigma = \lambda/\gamma$, y $\tau = \beta/\delta$ para los clientes prioritarios y no-prioritarios respectivamente. Por la ecuación (3.77) se observa que $\rho = \sigma + \tau$.

El estado del sistema es descito por la pareja $(X(t), Y(t))$ donde $X(t)$ es el número de clientes con prioridad en el sistema y $Y(t)$ es el número de clientes no-prioritarios en el sistema. Se hace la observación que desde el punto de vista de los clientes prioritarios el sistema luce como el sistema $M/M/1$. De acuerdo con la distribución límite dada en la ecuación (3.45) la distribución límite es

$$\lim_{t\to\infty} \Pr\{X(t) = m\} = (1 - \sigma)\sigma^m \quad \text{para} \quad m = 0, 1, \cdots \tag{3.78}$$

con $\sigma = \alpha/\gamma < 1$

Utilizando las ecuaciones (3.46) y (3.49), tenemos que el tamaño medio de la cola

Del Estado	Al Estado	Tasa de Transición	Descripción
(m,n)	$(m+1,n)$	α	Arribo de un cliente prioritario
(m,n)	$(m,n+1)$	β	Arribo de un cliente no-prioritario
$(0,n)$ $n \geq 1$	$(0,n-1)$	δ	Completación de un servicio no-prioritario
(m,n) $m \geq 1$	$(m-1,n)$	γ	Completación de un servicios prioritario

Tabla 3.2: *Tasas de transición de la cadena de Markov descrita por* $(X(t), Y(t))$.

para los clientes con prioridad es

$$L_p = \frac{\alpha}{\gamma - \alpha} = \frac{\sigma}{1-\sigma} \tag{3.79}$$

y el tiempo medio para los clientes prioritarios

$$W_p = \frac{1}{\gamma - \alpha} \tag{3.80}$$

El obtener información para los clientes no-prioritarios no es tan fácil, dado que estos arribos son fuertemente afectados por los clientes prioritarios. Aunque, $(X(t), Y(t))$ forma un estado discreto, las cadena de Markov continua, y las técnicas de la Sección 2.5, permiten describir la distribución límite, cuando existe. Las tasas de transición de $(X(t), Y(t))$ son descritas el la Tabla 3.2

Sea

$$\pi_{m,n} = \lim_{t \to \infty} \Pr\{X(t) = m, Y(t) = n\}$$

la distribición límite del proceso. Razonando análogamente a (2.67) y (2.68) del Capítulo 2 (donde la teoría fue descrita), se obtienen las siguientes ecuaciones para la distribución

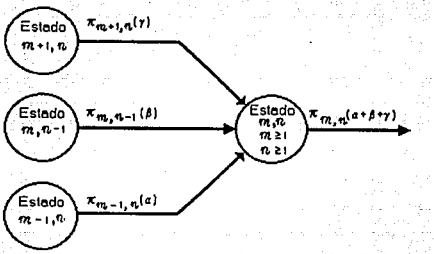

Figura 3.10: *En equilibrio, la tasa de flujo de entrada en cualquier estado debe ser igual a la tasa de flujo de salida. Aquí se ilustra el estado (m, n) cuando $m \geq 1$ y $n \geq 1$.*

estacionaria

$$
\begin{array}{llll}
(\alpha+\beta)\pi_{0,0} & = \gamma\pi_{1,0} & +\delta\pi_{0,1} & \\
(\alpha+\beta+\gamma)\pi_{m,0} & = \gamma\pi_{m+1,0} & & +\alpha\pi_{m-1,0}, \\
 & & & m \geq 1 \\
(\alpha+\beta+\delta)\pi_{0,n} & = \gamma\pi_{1,n} & +\delta\pi_{0,n+1} + \beta\pi_{0,n-1} & \\
 & & & n \geq 1 \\
(\alpha+\beta+\gamma)\pi_{m,n} & = \gamma\pi_{m+1,n} & +\beta\pi_{m,n-1} & +\alpha\pi_{m-1,n}, \\
 & & & m, n \geq 1
\end{array}
\tag{3.81}
$$

Las tasas de transición que nos llevan a la última ecuación del arreglo (3.81) se muestra en la Figura 3.10.

En principio, las ecuaciones del arreglo (3.81), aunadas a la condición $\sum_m \sum_n \pi_{m,n} = 1$, pueden ser resueltas para la distribución estacionaria, cuando esta existe. Solo se determina el número promedio L_n de los clientes no prioritarios en el sistema en estado estacionario, dado por

$$
L_n = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} n\pi_{m,n} \tag{3.82}
$$

Introduciendo ahora la notación

$$M_m = \sum_{n=0}^{\infty} n\pi_{m,n} = \sum_{n=1}^{\infty} n\pi_{m,n} \tag{3.83}$$

así que

$$L_n = M_0 + M_1 + \cdots \tag{3.84}$$

Usando la ecuación (3.78), sea

$$p_m = \Pr\{N(t) = m\} = \sum_{n=0}^{\infty} \pi_{m,n} = (1-\sigma)\sigma^m \tag{3.85}$$

y

$$\pi_n = \Pr\{Y(t) = n\} = \sum_{m=0}^{\infty} \pi_{m,n}. \tag{3.86}$$

Empezando con sumar ambos lados de las dos primeras ecuaciones del arreglo (3.81) para $m = 0, 1, \cdots$ para obtener

$$(\alpha+\beta)\pi_0 + \gamma\sum_{m=1}^{\infty}\pi_{m,0} = \gamma\sum_{m=1}^{\infty}\pi_{m,0} + \delta\pi_{0,1} + \alpha\pi_0,$$

que se simplifica como

$$\beta\pi_0 = \delta\pi_{0,1} \tag{3.87}$$

Posteriormente sumando las dos últimas ecuaciones del arreglo (3.81) sobre $m = 0, 1, \cdots$ para obtener

$$(\alpha+\beta)\pi_n + \delta\pi_{0,n} + \gamma\sum_{m=1}^{\infty}\pi_{m,n} = \gamma\sum_{m=1}^{\infty}\pi_{m,n} + \delta\pi_{0,n+1} + \beta\pi_{n-1} + \alpha\pi_n$$

que se simplifica como

$$\beta\pi_n + \delta\pi_{0,n} = \beta\pi_{n-1} + \delta\pi_{0,n+1}$$

e inductivamente con (3.87), se obtiene

$$\beta\pi_n = \delta\pi_{0,n+1} \quad \text{para} \quad n = 0, 1, \cdots \tag{3.88}$$

Sumando (3.88) sobre $n = 0, 1, \cdots$ y usando $\sum \pi_n = 1$ se obtiene

$$\beta = \delta \sum_{n=0}^{\infty} \pi_{0,n+1} = \delta \Pr\{X(t) = 0, Y(t) > 0\},$$

ó

$$\Pr\{X(t) = 0, Y(t) > 0\} = \sum_{n=1}^{\infty} \pi_{0,n} = \frac{\beta}{\delta} = \tau. \tag{3.89}$$

Dado que (3.85) acienta que $\Pr\{X(t) = 0\} = 1 - \frac{\alpha}{\gamma} = 1 - \sigma$, se tiene

$$\begin{aligned} \pi_{0,0} = \Pr\{X(t) = 0, Y(t) = 0\} &= \Pr\{X(t) = 0\} - \Pr\{X(t) = 0, Y(t) > 0\} \\ &= 1 - \frac{\alpha}{\gamma} - \frac{\beta}{\delta} = 1 - \sigma - \tau \\ &\quad \text{cuando } \sigma + \tau < 1. \end{aligned} \tag{3.90}$$

Con estos resultados preliminares a la mano, se puede determinar $M_n = \sum_{n=1}^{\infty} \pi_{n,n}$. Multiplicando la tercera ecuación del arreglo (3.81) por n y sumando, se deriva:

$$\begin{aligned} (\alpha + \beta + \delta) &= \gamma M_1 + \delta \sum_{n=1}^{\infty} n\pi_{0,n+1} + \beta \sum_{n=1}^{\infty} n\pi_{0,n-1} \\ &= \gamma M_1 + \delta M_0 - \delta \sum_{n=0}^{\infty} \pi_{0,n+1} + \beta M_0 + \beta \sum_{n=1}^{\infty} \pi_{0,n-1} \\ &= \gamma M_1 + \delta M_0 - \delta \left(\frac{\beta}{\delta}\right) + \beta M_0 + \beta(1 - \sigma), \end{aligned}$$

donde la última línea resulta de (3.85) y (3.89). Despues de la simplificación, el resultado es

$$M_1 = \sigma M_0 + \frac{\beta}{\gamma}\sigma. \tag{3.91}$$

Posteriormente se multiplica la última ecuación del arreglo (3.81) por n y sumando para obtener

$$\begin{aligned} (\alpha + \beta + \gamma) M_m &= \gamma M_{m+1} + \beta \sum_{n=1}^{\infty} n\pi_{m,n-1} + \alpha M_{m-1} \\ &= \gamma M_{m+1} + \beta M_m + \beta \sum_{n=1}^{\infty} \pi_{m,n-1} + \alpha M_{m-1}. \end{aligned}$$

Nuevamente, usando (3.85) y simplificando, vemos que

$$(\alpha+\gamma)M_m = \gamma M_{m+1} + \alpha M_{m-1} + \beta(1-\sigma)\sigma^m \quad \text{para} \quad m = 1,2,\cdots. \tag{3.92}$$

Las ecuaciones (3.91) y (3.92) pueden ser resueltas inductivamente para obtener

$$M_m = M_0\sigma^m + \frac{\beta}{\gamma} m\sigma^m \quad \text{para} \quad m = 0,1,\cdots$$

que se suma para obtener

$$L_n = \sum_{m=0}^{\infty} M_m = \frac{1}{1-\sigma}\left[M_0 + \frac{\beta}{\gamma}\frac{\sigma}{1-\sigma}\right]. \tag{3.93}$$

Esto determina a L_n en términos de M_0. Para obtener una segunda relación, multiplicando (3.88) por n y se suma para obtener

$$\begin{aligned}\beta L_n &= \delta\sum_{n=0}^{\infty} n\pi_{0,n+1} = \delta M_0 - \delta\sum_{n=0}^{\infty}\pi_{0,n+1}\\ &= \delta M_0 - \delta\left(\frac{\beta}{\delta}\right)\end{aligned}$$

ó

$$M_0 = \frac{\beta}{\delta}(L_n+1) = \tau(L_n+1). \tag{3.94}$$

Sustituyendo (3.94) en (3.93) y simplificando

$$L_n = \frac{1}{1-\sigma}\left[\tau(L_n+1) + \frac{\beta}{\gamma}\frac{\sigma}{1-\sigma}\right]$$

$$\left(1-\frac{\tau}{1-\sigma}\right)L_n = \frac{1}{1-\sigma}\left[\tau + \frac{\beta}{\gamma}\frac{\sigma}{1-\sigma}\right]$$

y, finalmente

$$L_n = \left(\frac{\tau}{1-\sigma-\tau}\right)\left[1 + \left(\frac{\delta}{\gamma}\right)\frac{\sigma}{1-\sigma}\right] \tag{3.95}$$

ρ	L	L_p	L_n
.6	1.50	.43	1.07
.8	4.00	.67	3.33
.9	9.00	.82	8.18
.95	19.00	.90	18.10

Tabla 3.3: *Tabla de resultados*

La condición de que L_n sea finito (y que la distribución estacionaria exista) es que

$$\rho = \sigma + \tau < 1.$$

Es decir, la intensidad de tráfico ρ en el sistema debe ser menor que uno.

Dado que la tasa de arribos para los clientes no-prioritarios es β, tenemos que el tiempo medio de espera para un cliente no-prioritario está dado por $W_n = L_n/\beta$.

Algunos estudios numericos simples de (3.79) y (3.95) llevan a resultados interesantes adicionando prioridad a un sistema existente. Considerando primero un sistema, digamos, el sistema $M/M/1$ con intensidad de tráfico ρ, cuya longitud promedio de cola está dada por la ecuación (3.46) que es $L = \rho/(1-\rho)$. Modificando el sistema, en cuyo caso la fracción $\rho = 1/2$ de los clientes que tienen prioridad. Supongase que esa prioridad es independiente del tiempo de servicio. Esta suposición lleva a los valores $\alpha = \beta = 1/2$ y $\gamma = \delta = \mu$, donde $\sigma = \tau = \rho/2$. Entonces la longitud media de la cola para los clientes prioritarios y no-prioritarios está dada por

$$L_p = \frac{\sigma}{1-\sigma} = \frac{\rho/2}{1-(\rho/2)} = \frac{\rho}{2-\rho}$$

y

$$L_n = \left(\frac{\rho/2}{1-\rho}\right)\left[1 + \frac{\rho/2}{1-(\rho/2)}\right] = \frac{\rho}{(2-\rho)(1-\rho)}.$$

Las longitudes medias de las colas L, L_p y L_n fueron determinadas por diversos valores de la intensidad de tráfico ρ. Los resultados se muestran en la Tabla 3.3

Esto muestra que tanto el incremento del largo de la cola, como la intensidad de tráfico, se cargan casi exclusivamente en los clientes no-prioritarios.

Capítulo 4

Aplicación de los Modelos de Espera.

Este Capítulo pretende presentar un modelo de aplicación a la teoría de líneas de espera.

En todo sistema existen dos grandes clases de costos, el referente al tiempo de espera de un servicio y el asociado al consumo de los recursos que requiere ese servicio. Por otra parte el objetivo del análisis de los sistemas de líneas de espera consiste en obtener un equilibrio entre los costos de espera y los costos de servicio.

Para ilustrar esta situación se conciderará un fenómeno típico de línea de espera.

El sistema que se describirá es el que se presenta en la biblioteca de la Facultad de Ciencias, se dispone de algunos datos del sistema que existía en ésta, en Febrero de 1991, pero dado que dicho sistema sufrió algunas modificaciones a principios de 1992, se realizaron observaciones del nuevo sistema en Abril de 1993, para poder hacer un análisis comparativo de los dos sistemas.

Estos datos fueron tomados en la hora de mayor movimiento en la biblioteca, que en Febrero de 1991 y en Abril de 1993 era la misma; de las 12:00 hrs. a las 13:00hrs., (esto es debido, fundamentalmente a la disposición de los horarios en la Facultad de Ciencias. Definitivamente el horario con más alumnado es el matutino, la biblioteca da servicio de las 9:00 hrs. a las 19:00 hrs.; el momento en que mayor número de alumnos del turno matutino salen de clases y disponen de tiempo para ir a la biblioteca es aproximadamente de las 12:00 hrs. a 13:00 hrs.). En Febrero de 1991 se toma la semana del 18 al 22; y en Abril de 1993 la semana del 26 al 30, que si bien forman un conjunto reducido de datos, el comportamiento de los sistemas permite muy buenas descripciones.

A lo largo de este Capítiulo se describirán estos dos sistemas y se encontraran sus modelos respectivos, para poder realizar un análisis crítico de las modificaciones realizadas

a dichos sistemas.

La biblioteca de la Facultad de Ciencias se encarga básicamente de ofrecer su acervo en dos modalidades, las cuales son:

- **Consulta Interna**
  Para este tipo de consulta la biblioteca cuenta con una sala especial, a la cual, las personas que requieren de este tipo de servicio, únicamente tienen que entrar a esta sala y buscar la información que requieren. El material existente en la sala, nunca puede salir de ella. Si la persona necesita sacar algún libro tiene que acudir a pedir el servicio que ofrese la biblioteca en su segunda modalidad. (consulta externa).
- **Consulta Externa**
  La consulta externa; permite a las personas que requieren de algún material de la biblioteca, extraerlo de esta. Con opcion de tenerlo bajo su cuidado a lo más por tres días.

  Es aquí donde se forma la línea que se desea estudiar. también, es aquí donde el sistema sufrió las modificaciones, así que las describiremos independientemente por sistema.

4.1 Descripción del Sistema I para Obtener el Modelo I.

Un concepto fundamental en el análisis de los sistemas de líneas de espera es el modelo del sistema. En este se da la descripción de dicho sistema que proporciona una base suficiente para poder predecir su comportamiento.

El análisis se hará describiendo el procedimiento del servicio así como el ambiente y recursos con que contaba la biblioteca de la Facultad de Ciencias en Febrero de 1991 en su modalidad de consulta externa.

Para este tipo de consultas los usuarios tienen que ubicar los libros que requieren en los ficheros, y llenar una forma de préstamo; posteriormente tinen que formar una línea de espera para ser atendidos. Esta línea es la que se desea modelar. En ésta también estarán los usuarios que van a devolver material.

Después de esperar para obtener su turno de servicio, el usuario es atendido por personal de la biblioteca, si únicamente devuelve material, los servidores buscan su ficha de préstamo y la liberan; si solicitán material, las personas que dan servicio en la facultad

lo buscan en los anaqueles, registran el préstamo y así concluyen el servicio, la demanda de servicio puede ser combinada, es decir, se pueden entregar libros al mismo tiempo que se solicitan nuevos.

4.1.1 Ambiente del Sistema I.

En las secciones 1.2 y 3.1 se dió la metodología para la descripción específica de un sistema de líneas de espera, con base a ésto se dará para especificar el Modelo I, la descripción de:

- la población o fuente;
- la pauta de llegadas o descripción de los arribos;
- el mecanismo de servicio, y
- la disciplina de la cola.

Población o Fuente.

Básicamente, la población de este sistema es el alumnado de la Facultad de Ciencias que está compuesto por los estudiantes de las carreras de Actuaría, Biología, Física y Matemáticas, en su mayor parte de nivel de licenciatura, y en menor grado los estudiantes de posgrado de las mismas carreras, así como los casos de préstamo interbibliotecario. Dado que los alumnos constantemente hacen uso de este servicio, se dará por hecho que la población es infinita.

Pauta de Llegadas o Descripción de los Arribos.

En términos generales, las llegadas son completamente aleatorias, por ello se tendrá que empezar suponiendo que, los arribos describen un proceso Poisson, (ésto en vías de construir el modelo más sencillo pero también el más frecuente, el $M/M/s$). Para ello se utilizaron los datos y las pruebas correspondientes que se especificarán en la subsección 4.1.2.

Mecanismo de Servicio.

Se pudo observar, (cuando se realizaron las observaciones), que los servicios también son completamente aleatorios , y por esta misma razón se supondrá que los tiempos de servicio siguen una distribución exponencial, (también en vías de model $M/M/s$). Esto será completamente detallado en la subsección 4.1.2. Cabe señalar que el Sistema I cuenta con tres servidores.

Disciplina de la Cola.

Se trata de un caso de cola única, y la disciplina es la más común; primero en llegar, primero en ser servido.

FRECUENCIA	0	0	0	0	0	0	0	0	0	0	0	0
INERVALO	01	02	03	04	05	06	07	08	09	10	11	12
FRECUENCIA	0	0	0	0	0	0	0	0	0	0	0	0
INTERVALO	13	14	15	16	17	18	19	20	21	22	23	24
FRECUENCIA	0	0	0	0	0	0	0	0	0	0	0	0
INTERVALO	25	26	27	28	29	30	31	32	33	34	35	36
FRECUENCIA	0	0	0	0	0	0	0	0	0	0	0	0
INTERVALO	37	38	39	40	41	42	43	44	45	46	47	48
FRECUENCIA	0	0	0	0	0	0	0	0	0	0	0	0
INTERVALO	49	50	51	52	53	54	55	56	57	58	59	60

Tabla 4.1: *Tabla modelo para la recopilación de los tiempos de arribo, en intervalos de un minuto.*

4.1.2 Tratamiento de la Información Para la Obtención del Modelo I.

Como se señaló en la subsección anterior, el primer paso es suponer un modelo que describa al Sistema I, habitualmente se comenzará por el modelo más sencillo para poder utilizar todas las herramientas descritas en el Capítulo 3, por tal motivo se tratará de demostrar que los arribos siguen una distribución Poisson y que los tiempos de servicio están exponencialmente distribuidos para así tener el modelo $M/M/3$.

Para capturar la información, se tiene que recordar que los modelos descritos en el Capítulo 3 suponen tiempos de servicio y tiempos de arribo completamente aleatorios e independientes, así únicamente se necesita cronometrar los tiempos de arribo y los tiempos de servicio por cliente.

Análisis de los Arribos al Sistema I.

Toda la información que se captura con respecto a los arribos, debe ser tratada de la sigiente forma.

Para cada grupo de datos capturados se elaborará una tabla similar a la Tabla 4.1 donde los valores contenidos (los ceros,) en cada cuadro corresponden al número de arribos ocurridos en los intervalos de tiempo en los que se ha decidido dividir el período de observación.

La longitud de estos intervalos depende de la frecuencia de los arribos, regularmente

FRECUENCIA	1	1	0	0	2	0	0	0	0	0	3	0
INTERVALO	01	02	03	04	05	06	07	08	09	10	11	12
FRECUENCIA	1	0	1	1	0	0	2	0	3	0	1	1
INTERVALO	13	14	15	16	17	18	19	20	21	22	23	24
FRECUENCIA	2	2	1	1	1	1	2	2	1	2	2	1
INTERVALO	25	26	27	28	29	30	31	32	33	34	35	36
FRECUENCIA	2	0	1	0	1	1	1	2	0	1	1	2
INTERVALO	37	38	39	40	41	42	43	44	45	46	47	48
FRECUENCIA	1	0	2	1	0	1	2	0	0	1	1	1
INTERVALO	49	50	51	52	53	54	55	56	57	58	59	60

Tabla 4.2: *Frecuencia Promedio de Arribos por minuto para el Primer Modelo.*

se adecúan de tal forma en que en la mayor parte de ellos sólo se registre un arribo. (Para cumplir los postulados del proceso Poisson descritos en la subsección 2.3.2).

En este caso para modelar el Sistema I se tomaron intervalos de un minuto.

Posteriormente se obtiene la frecuencia promedio por intervalo de las observaciones realizadas para descripción de los arribos del Sistema I.

Apartir de los datos de la Tabla 4.2 se calcula la media de los tiempos de llegadas.

$$\bar{X} = \frac{f_X}{n} = \frac{55}{60} = .95$$

y también con los datos de la tabla 4.2 se realiza el histograma de frecuencias (Figura 4.1).

Ajuste de la Distribución Muestral a la Distribución Poisson.

Teniendo en mente, que los tiempos de llegadas siguen una distribución Poisson, se tiene que realizar una prueba que permita aceptar o rechazar esta suposición.

Se tiene que

$$\bar{X} = .95 = \lambda_1$$

Para $X = x$: es decir x llegadas en un intervalo de un minuto; de acuerdo con la función de distribución Poisson, se tiene

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

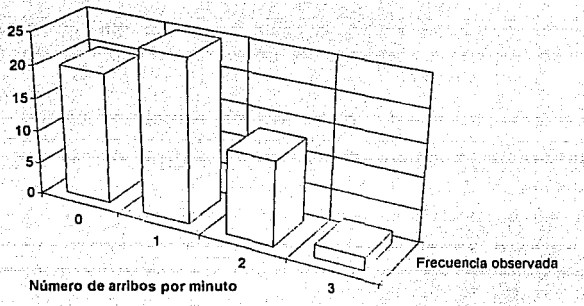

Figura 4.1: *Histograma de frecuencia de los arribos al sistema I*

En la tercera columna de la Tabla 4.3 aparecen las probabilidades de 0,1,2,3 arribos en intervalos de un minuto; en la cuarta columna se da el valor esperado de frecuencias por intervalo, que se obtiene multiplicando la probabilidad correspondiente por el número total de intervalos: en este caso 60.

Con objeto de comparación, esta Tabla presenta en la segunda columna los arribos observados.

Ji-cuadrada Bondad de Ajuste.

Como se puede ver en la Tabla 4.3, los resultados obtenidos de la muestra no concuerdan con los resultados teóricos esperados, por ello se desea ver si las frecuencias observadas difieren significativamente de las frecuencias esperadas.

La idea fundamental de la prueba χ^2 (ji-cuadrada) es la de comparar la función de distribución de las llegadas $\bar{F}_X$, ($\bar{F}_X$ es la función de distribución empírica) con la función F_X que en este caso representa a la distribución Poisson.

Para esto se tiene que saber que tanto difiere $\bar{F}_X$ de F_X, para tomar en cuenta si la distribución empírica sigue la distribución Poisson, y así poder aplicar las derivaciones matemáticas de la teoría de colas.

Se tomarán n intervalos de acuerdo a la distribución muestral considerada, y se calculará para cada valor i

No. de Arribos en Intervalos de 1 min.	Frecuencia Observada O_i	Probabilidad P_i	No. de Arribos Esperados E_i	$\frac{(O_i-E_i)^2}{E_i}$
0	20	0.3867410	23.20446	0.442525
1	25	0.3674039	22.04423	0.396320
2	13	0.1745168	10.47101	0.610807
3	02	0.0552636	03.31581	0.522155
			$\sum_{i=0}^{3} \frac{(O_i-E_i)^2}{E_i} =$	1.971808

Tabla 4.3: *Arribos Febrero 1991*

$$\frac{(O_i - E_i)^2}{E_i}$$

Donde O_i son las llegadas observadas y E_i son las llegadas Esperadas.

Sea $m = n - 1$, se calcula

$$t = \sum_{i=0}^{m} \frac{(O_i - E_i)^2}{E_i}.$$

Se tomarán los niveles de significancia de $\alpha = 5\%$ y $\alpha = 10\%$.

Para la prueba donde

H_0: F_X tiene distribución Poisson.

vs.

H_a: F_X No tiene distribución Poisson.

En la Tabla 4.3 se calcula t

$$t = 1.971808$$

Escogiendo el nivel de significancia del 10% para el tercer grado de libertad se tiene (usando las tablas de χ^2 de [11]).

$$\Pr(t \leq \chi^2_{3,90}) = 1 - \alpha = .90 \quad \Rightarrow \quad \chi^2_{3,90} = 6.251$$

y escogiendo el nivel de significancia del 5% para el tercer grado de libertad se tiene

$$\Pr(t \leq \chi^2_{3,95}) = 1 - \alpha = .95 \quad \Rightarrow \quad \chi^2_{3,95} = 7.815$$

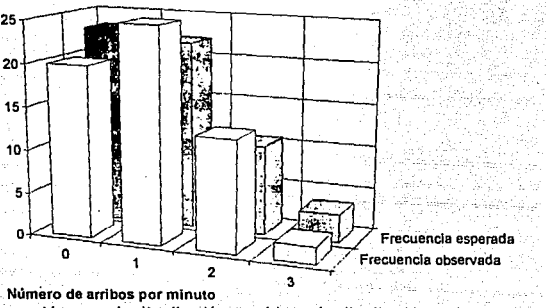

Figura 4.2: *Comparación entre la distribución empírica y la distribución real, para los arribos del Sistema I*

$t < \chi^2_{3,.95}$ y $t < \chi^2_{3,.90}$ por lo tanto no se rechaza la hipótesis.

Con lo que se puede concluir que la función de distribución $\bar{F}_N$ es una buena aproximación de F_N; una distribución Poisson, con parámetro $\lambda_1 = .95$.

Esto se puede observar claramente en la Figura 4.2.

Tiempos de Servicio del Sistema I

Como restricción se tiene que los tiempos de servicio deben seguir una distribución exponencial.

Primeramente se hará el tratamiento de la información capturada en la muestra para ajustar los tiempos de llegadas a una distribución exponencial.

Los tiempos de servicio fueron tomados desde el momento en que el cliente pide su servicio, hasta el momento en que abandona el mostrador de servicio.

Como los tiempos de llegadas del primer sistema, tienen una variabilidad grande, se ha decidido distribuirlos en intervalos de 30 segundos.

El primer paso a seguir es el vaciado de las frecuencias de duración de los servicios por período.

Posteriormente se elaborá la Tabla 4.4, en donde la segunda columna muestra la frecuencia promedio de tiempos de servicio, la primera columna muestra el rango de agru-

pación para los tiempos de servicio: en la tercera aparece la frecuencia relativa , (es decir la frecuencia de clientes dividido entre la frecuencia total de la muestra.), y en la cuarta columna la frecuencia relativa acumulada, que nos define a $S(x)$ como la distribución empírica.

Ajuste de Tiempos de Servicio a Una Distribución Exponencial.

Calculando el estimador de máximo verosimil. Si $f_X(x) = e^{-\mu x}$

$$\begin{aligned} L &= e^{-\mu x_1} e^{-\mu x_2} \cdots e^{-\mu x_n} \\ &= \mu^n e^{-n\mu(x_1+x_2+\cdots+x_n)} \\ &= \mu^n e^{-n\mu \bar{X}} \\ ln(L) &= n ln(\mu) - n\mu \bar{X} \\ \frac{dln(L)}{d\mu} &= \frac{n}{\mu} - n\bar{X} \\ \frac{n}{\mu} - n\bar{X} &= 0 \\ \frac{n}{\mu} &= n\bar{X} \\ n &= n\bar{X}\mu \quad \Rightarrow \\ \hat{\mu} &= \frac{1}{\bar{X}}. \end{aligned}$$

La media de los valores observados es

$$\bar{X} = 178.816602 \ \frac{seg.}{servicios}$$

y

$$\frac{1}{\bar{X}} = 0.005593 \ \frac{servicios}{seg.},$$

o lo que es lo mismo

$$\frac{1}{\bar{X}} = 0.33553 \ \frac{servicios}{min.} = \mu_1.$$

En la Tabla 4.1 se muestra también en la quinta columna los valores que toma $F_X(x) = 1 - e^{-\mu x}$.

Intervalos en Segundos	Frecuencia Promedio	Frecuencia Relativa	Frecuencia Acumulada $S(x)$	$F_X(x)$	$\lvert S(x) - F_X(x)\rvert$
0-29	11.0	0.105769	0.105769	0.149700	0.04391
30-59	16.6	0.159615	0.265384	0.281038	0.01565
60-89	13.4	0.128846	0.394230	0.392082	0.00214
90-119	12.6	0.121153	0.515384	0.485976	0.02910
120-149	10.6	0.101923	0.617307	0.565368	0.05193*
150-179	06.4	0.061538	0.678846	0.632497	0.01631
180-209	05.2	0.050000	0.728846	0.689258	0.03958
210-239	05.4	0.051923	0.780769	0.737253	0.04351
240-269	02.8	0.026923	0.807692	0.777831	0.02985
270-299	02.4	0.023076	0.830769	0.812118	0.01862
300-329	01.8	0.017307	0.848076	0.841162	0.00691
330-359	01.8	0.017307	0.865384	0.865695	0.00031
360-389	02.0	0.019230	0.884615	0.886138	0.00182
390-419	02.2	0.021153	0.905769	0.903978	0.00179
420-449	01.4	0.013461	0.919230	0.918809	0.00042
450-479	00.8	0.007692	0.926923	0.931349	0.00442
480-509	00.8	0.007692	0.934615	0.941952	0.00733
510-539	01.2	0.011538	0.946153	0.950918	0.00176
540-569	01.0	0.009615	0.955769	0.958198	0.00272
570-599	01.0	0.009615	0.965384	0.964908	0.00047
600-629	00.0	0.000000	0.965384	0.970328	0.00494
630-659	00.4	0.003846	0.969230	0.974911	0.00568
660-689	00.6	0.005769	0.975000	0.978786	0.00378
690-719	00.4	0.003846	0.978846	0.982062	0.00321
720-749	00.4	0.003846	0.982692	0.984833	0.00211
750-779	00.0	0.000000	0.982692	0.987175	0.00118
780-809	00.2	0.001923	0.984615	0.989156	0.00154
810-839	00.0	0.000000	0.984615	0.990831	0.00621
840-869	00.2	0.001923	0.986538	0.992247	0.00570
870-899	00.4	0.003846	0.990384	0.993444	0.00306
900-929	00.0	0.000000	0.990384	0.994457	0.00407
930-959	00.0	0.000000	0.990384	0.995313	0.00192
960-989	00.0	0.000000	0.990384	0.996037	0.00565
990-1019	00.0	0.000000	0.990384	0.996619	0.00626

Intervalos en Segundos	Frecuencia Promedio	Frecuencia Relativa	Frecuencia Acumulada $S(x)$	$F_X(x)$	$\lvert S(x) - F_X(x)\rvert$
1020-1049	00.2	0.001923	0.992307	0.997166	0.00485
1050-1079	00.2	0.001923	0.994230	0.997604	0.00337
1080-1109	00.0	0.000000	0.994230	0.997974	0.00374
1110-1139	00.2	0.001923	0.996153	0.998287	0.00213
1140-1169	00.2	0.001923	0.998076	0.998551	0.00047
1170-1199	00.2	0.001923	1.000000	0.998775	0.00122

Tabla 4.4: *Datos correpondientes a los servicios del primer modelo*

Kolmogorov–Smirnov Bondad de Ajuste.

La prueba de Kolmogorov–Smirnov es apropiada para distribuciones continuas.

La hipótesis a probar es que cierta función de distrubución $F_X(x)$ es la función de la cual se tomó una población muestral $x_1, x_2, \cdots, x_n$. El procedimiento a segira es el siguiente

- Se calculan los valores de la función de distribución $S(x)$ de la muestra $x_1, x_2, \cdots, x_n$ (estos valores se encuentran en la Tabla 4.4).
- Se calculan los valores de la función de distribución $F_X(x)$.
- Se determina la distancia máxima entre $S(x)$ y $F_X(x)$.

$$t = \max |S(x) - F_X(x)|$$

- Se escoge el nivel de significancia de $\alpha = 5\%$ y $\alpha = 10\%$.
- Se plantea la hipótesis

  H_0: La función de distribución muestral $S(x)$ sigue una distribución exponencial.

  vs.

  H_a: La función de distribución muestral $S(x)$ no sigue una distribución exponencial.

Los valores comparados se muestran en la columna seis de Tabla 4.4, en este caso el valor máximo es

$$t = 0.05193$$

En la tabla de Kolmogorov-Smirnov tomada de [11] se tiene que

Para $\alpha = 0.05$ el valor en tablas es 0.05360
Para $\alpha = 0.01$ el valor en tablas es 0.06678

Dado que $t < 0.05360$ y $t < 0.06678$ no se rechaza la hipotesis.

Por lo tanto podemos suponer que los tiempos de servicio se distribuyen exponencialmente con parámetro $\mu_1 = 0.3355$.

4.1.3 El Modelo I.

Según el estudio realizado anteriormente, las llegadas y los tiempos de servicio, siguen una distribución Poisson y exponencial respectivamente, y se conoce que el sistema tiene tres servidores así, el Modelo I es un $M/M/3$.

Se explorá este modelo, para poder realizar algunas comparaciones con el Modelo II, para ésto se utilizarán los resultados de la subsección 3.4.3.

Se calculará lo sigiente.

L_0 El número promedio de clientes en cola.
W_0 El tiempo promedio gastado por un cliente en la cola.
L El número promedio de clientes en el sistema.
W El tiempo promedio gastado por un cliente en el sistema.

Con base a la información revelada por las observaciones se tiene que

$\lambda_1 = 0.95$ Arribos por minuto
$\mu_1 = 0.3355$ Servicios por minuto

Calculando el coeficiente de intensidad de tráfico

$$\rho_1 = \frac{\lambda_1}{s\mu_1} = \frac{.95}{3(0.3355)} = \frac{.95}{1.0065} = 0.9438.$$

Lo cual satisface $\rho_1 < 1$.

Para el desarrollo de este problema, el primer paso lógico es obtener el valor de π_0, ya que este valor es fundamental para la resolución del problema.

$$\pi_0 = \left[\left\{\sum_{j=0}^{s-1}\frac{1}{j!}\left(\frac{\lambda_1}{\mu_1}\right)^j\right\} + \frac{((\lambda_1/\mu_1)^s}{s!(1-\lambda_1/s\mu)}\right]^{-1}$$

$$\pi_0 = [75.2479]^{-1}$$

$$\pi_0 = 0.01328.$$

π_0	L_0	W_0	L	W
0.01328	0.8444	0.8893	3.6767	3.8701

Tabla 4.5: *Resumen de resultados para el Modelo I.*

es decir la probabilidad de llegar a la biblioteca y encontrar vacío el sistema es 0.01328.

Una vez calculado el valor de π_0, pueden definirse fácilmente los valores L_0, W_0, L, W.

De esta forma L_0 o el número promedio de clientes que esperan en la cola sin incluir a las personas que están siendo atendidas es, por la ecuación (3.53)

$$L_0 = \frac{\pi_0}{s!}\left(\frac{\lambda_1}{\mu_1}\right)^s \left(\frac{(\lambda_1/s\mu)}{(1-\lambda_1/s\mu_1)^2}\right) = 0.8449 \text{ clientes}$$

W_0 El tiempo promedio gastado por un cliente en la cola; es decir el tiempo que tiene que esperar en promedio un cliente para iniciar su servicio, está dado por la ecuación (3.35))

$$W_0 = \frac{L_0}{\lambda_1} = \frac{0.844919}{.95} = 0.88948 \text{ minutos.}$$

$$W_0 = 53.36 \text{ segundos.}$$

L El número promedio de clientes en el sistema es

$$L = L_0 + \frac{\lambda_1}{\mu_1} = 0.844919 + 2.8315 = 3.6761$$

W El tiempo promedio gastado por un cliente en el sistema es

$$W = W_0 + \frac{1}{\mu_1} = 0.88938 + 2.8315 = 3.7209 \text{ minutos.}$$

La Tabla 4.5 resume esta información.

4.2 Descripción del Sistema II para Obtener el Modelo II

Ahora el análisis se hará describiendo el procedimiento así como el ambiente y recursos con los que contaba la biblioteca de la Facultad de Ciencia en Abril de 1993 en su modalidad de consulta externa.

Para este tipo de consulta, los usuarios tienen que ubicar los libros que requieren en los ficheros; posteriormente ellos mismos tiene que buscar en los anaqueles el material que necesitan; inmediatamente tienen que formar una línea frente a un mostrador para que les autoricen el préstamo. Es esta línea es la que se modelará para representar al Sistema II.

Después de esperar para obtener su turno de servicio, el usuario es atendido por personal de la biblioteca. Ahora se dispon de una computadora, que solamente lee el codigo de barras del libro y registra el préstamo. esta línea es únicamente para préstamo de libros de la biblioteca.

4.2.1 Ambiente del Sistema II

Análogamente a lo realizado en la Subsección 4.1.1, se esbozará el ambiente del Sistema II.

Población o Fuente

la población es la misma población que la del Sistema I.

Pauta de Llegadas o Descripción de los Arribos.

En términos generales las llegadas son completamente aleatorias, y como en el Sistema I se tratará de demostrar que siguen una distribución Poisson.

Mecanismo de Servicio.

Nuevamente se encontró que los tiempos de servicio son aleatorios y se tratará de demostrar que siguen una distribución exponencial.

Disciplina de la Cola.

La disciplina de la cola es, primero en llegar, primero en ser servido. Se trata de una cola única que es atendida por un solo servidor. (Una computadora).

4.2.2 Tratamiento de la Información Para la Obtención del Modelo II.

Nuevamente se tendrán que hacer una serie de suposiciones distribucionales, y se relizarán pruebas de bondad de ajuste para los valores observados, en esta ocasión se buscará ajustar el modelo $M/M/1$.

FRECUENCIA	0	1	0	0	3	0	1	2	2	3	1	1
INTERVALO	01	02	03	04	05	06	07	08	09	10	11	12
FRECUENCIA	1	1	2	3	2	1	1	1	1	1	0	2
INTERVALO	13	14	15	16	17	18	19	20	21	22	23	24
FRECUENCIA	2	2	1	2	1	2	2	2	1	1	1	1
INTERVALO	25	26	27	28	29	30	31	32	33	34	35	36
FRECUENCIA	0	1	1	2	1	1	0	0	2	2	1	0
INTERVALO	37	38	39	40	41	42	43	44	45	46	47	48
FRECUENCIA	0	3	2	0	0	4	3	2	1	1	2	4
INTERVALO	49	50	51	52	53	54	55	56	57	58	59	60

Tabla 4.6: *Frecuencia Promedio de Arribos por minuto para el Segundo Modelo.*

Análisis de los Arribos al Sistema II.

La información se trata de la misma forma que en la sección anterior, la Tabla 4.6 muestra la frecuencia promedio de arribos por minuto para este sistema.

Los datos de la Tabla 4.6 permiten elaborar el histograma de frecuencias que se presenta en la Figura 4.3. El primer paso es obtener el promedio de arribos, es decir $\bar{X}$.

$$\bar{X} = \frac{fx}{n} = \frac{81}{60} = 1.35$$

Ajuste de la Distribución Muestral a la Poisson.

Se desea que los datos obtenidos hayan sido tomados de una función de distribución Poisson, se tendrá que realizar la prueba correspondiente.

Se tiene que:

$$\bar{X} = 1.35 = \lambda_2$$

En la Tabla 4.7 aparecen las probabilidades de 0,1,2,3 y 4 arribos en intervalos de un minuto; en la cuarta columna se da el valor esperado de frecuencias por intervalo, que se obtiene multiplicando la probabilidad correspondiente por el número total de intervalos; en este caso 60. Para poder comparar, esta tabla también contiene en la segunda columna el número de arribos observados en cada intervalo.

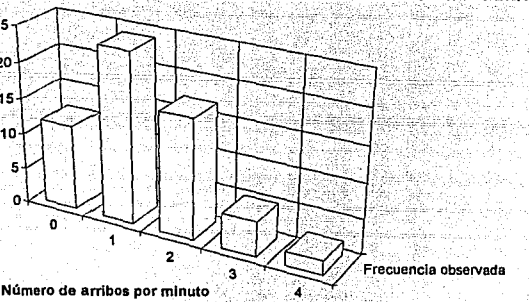

Figura 4.3: *Histograma de frecuencias de los arribos al Sistema II.*

No. de Arribos en Intervalos de 1 min.	Frecuencia Observada O_i	Probabilidad P_i	No. de Arribos Esperados E_i	$\frac{(O_i-E_i)^2}{E_i}$
0	12	0.2592402606	15.554415	0.812236
1	24	0.3499743519	20.998461	1.189830
2	17	0.2362326875	14.173961	0.563462
3	05	0.1063047094	06.378282	0.297832
4	02	0.0358778394	02.152670	0.010827
			$\sum_{i=0}^{4} \frac{(O_i-E_i)^2}{E_i} =$	2.874190

Tabla 4.7: *Arribos abril 1993*

Ji-cuadrada Bondad de Ajuste.

Como se puede ver en la Talba 4.7. los resultados obtenidos de la muestra, no concuerdan con los resultados teóricos esperados, por ello se desea ver si las frecuencias observadas difieren significativamente, análogamente a lo que se hizo en la sección anterior, se tiene que

$$l = \sum_{i=0}^{m} \frac{(O_i - E_i)^2}{E_i} = 2.8711190.$$

Se tomarán los niveles de significancia de $\alpha = 5\%$ y $\alpha = 10\%$, para la prueba donde

H_0: F_X tiene distibución Poisson.

vs.

H_a: F_X no tiene distibución Poisson.

Tomando el nivel de significancia $\alpha = 10\%$ para el cuarto grado de libertad se tiene

$$\Pr(l \leq \chi^2_{4,.90}) = 1 - \alpha = .90 \quad \Rightarrow \chi^2_{4,.90} = 7.779$$

y tomando el nivel de significancia $\alpha = 5\%$ para el cuarto grado de libertad se tiene. (Esto valores son tomados de las tablas correspondientes de [11].).

$$\Pr(l \leq \chi^2_{4,.95}) = 1 - \alpha = .95 \quad \Rightarrow \chi^2_{4,.95} = 9.488$$

$l < \chi^2_{4,.95}$ y $l < \chi^2_{4,.95}$, por lo tanto no se rechaza la hipótesis.

Con lo que podemos concluir que la función de distribución F_X es una buena aproximación de F_X (una distribución Poisson con parámetro $\lambda_2 = 1.35$). Esto se puede observar claramente en la Figura 4.4.

Tiempos de Servicio del Sistema II

Se tiene la misma restricción que para el Sistema I la Tabla 4.2.2 se realzó con la misma metodología que la tabla correspondiente para el Sistema I.

Los tiempos de servicio del Sistema II, tienen una variabilidad menor que la de los del Sistema I, por ello se han distrubuido en intervalos de 10 segundos.

Intervalos en Segundos	Frecuencia Promedio	Frecuencia Relativa	Frecuencia Acumulada $S(x)$	$F_X(x)$	$\lvert S(x) - F_X(x)\rvert$
0-9	18.0	0.194384	0.194384	0.291000	0.09966*
10-19	29.8	0.321811	0.516198	0.520613	0.00111
20-29	16.8	0.181425	0.697624	0.674443	0.02318
30-39	09.2	0.099352	0.796976	0.778911	0.01806
40-49	07.4	0.079913	0.876889	0.849856	0.02703
50-59	04.2	0.045356	0.922246	0.898036	0.02121
60-69	01.2	0.012958	0.935205	0.930755	0.00114
70-79	01.0	0.010799	0.946004	0.952975	0.00697
80-89	01.4	0.015118	0.961123	0.968065	0.00694
90-99	01.0	0.010799	0.971922	0.978312	0.00638
100-109	00.4	0.004319	0.976241	0.985271	0.00902
110-119	00.4	0.004319	0.980561	0.989998	0.00943
120-129	01.0	0.010799	0.991360	0.993207	0.00184
130-139	00.4	0.004319	0.995680	0.995387	0.00029
140-149	00.0	0.000000	0.995680	0.996867	0.00118
150-159	00.2	0.002159	0.997840	0.997872	0.00003
160-169	00.0	0.000000	0.997840	0.998555	0.00071
170-179	00.0	0.000000	0.997840	0.999018	0.00117
180-189	00.2	0.002159	1.000000	0.999333	0.00066

Tabla 4.8: *Datos correspondientes a los servicios del segundo modelo*

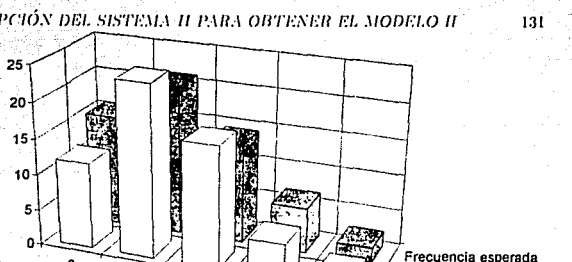

Figura 4.4: *Comparación entre la distribución empírica y la distribución real, para los arribos del Sistema II.*

Ajuste de Tiempos de Servicio a Una distribución Exponencial.

La media de los tiempos de servicio para el Sistema II es

$$\bar{X} = 25.841648 \frac{seg.}{servicio},$$

y

$$\frac{1}{\bar{X}} = 0.038697 \frac{servicios}{seg.},$$

o lo que es lo mismo

$$\frac{1}{\bar{X}} = 2.321833 \frac{servicios}{min.} = \mu_2$$

Kolmogorov-Smirnov Bondad de Ajuste.

En este caso como se muestra en la Tabla 4.8 el estadístico de Kolmogorov-Smirnov es

$$t = 0.09966.$$

En la tabla de Kolmogorov-Smirnov de [11] se tiene que para los niveles de significancia $\alpha = 5\%$ y $\alpha = 1\%$ sucede

Para $\alpha = 0.05$ el valor en tablas es 0.493747.
Para $\alpha = 0.01$ el valor en tablas es 0.611516.

Dado que $I < 0.493747$ y $I < 0.611516$ no se rechaza la hipótesis.

Por lo tanto se puede suponer que los tiempos de servicio se distribuyen exponencialmente con parámetro $\mu_2 = 2.321833$.

4.2.3 El Modelo II.

Según el estudio realizado anteriormente, las llegadas y los tiempos de servicio, siguen una distribución Poisson y exponencial respectivamente, y se conoce que el sistema tiene un servidor, así, el Modelo II es un $M/M/1$.

Se explorará este modelo, para poder realizar algunas comparaciones con el Modelo I, para ésto se utilizarán los resultados de la subsección 3.4.1

Se calculará lo siguiente

L_0 El número promedio de clientes en cola.
W_0 El tiempo promedio gastado por un cliente en cola.
L El número promedio de clientes en el sistema.
W El tiempo promedio gastado por un cliente en el sistema.

Con base a la información revelada por las observaciones, se tiene que

$\lambda_2 = 1.35$ Arribos por minuto.
$\mu_2 = 2.3218$ Servicios por minuto.

Calculando el coeficiente de intensidad de tráfico

$$\rho_2 = \frac{\lambda_2}{\mu_2} = \frac{1.35}{2.321833} = 0.581437.$$

Lo cual satisface $\rho_2 < 1$.

Para el desarrollo de este problema, el primer paso es obtener el valor de π_0. Por la ecuación (3.44) se tiene

$$\pi_0 = 1 - \frac{\lambda_2}{\mu_2}$$
$$\pi_0 = 1 - 0.581437$$
$$\pi_0 = 0.418562,$$

es decir, la probabilidad de llegar a la biblioteca y encontrar vacío el sistema es 0.41856.

π_0	L_0	W_0	L	W
0.41856	0.8075	0.5982	1.3891	1.0290

Tabla 4.9: *Resumen de resultados para el Modelo II.*

L El número promedio de clientes en el sistema es por la ecuación (3.46)

$$L = \frac{\lambda_2}{\mu_2 - \lambda_2} = \frac{1.35}{2.3218 - 1.35} = 1.3891 \; clientes.$$

W El tiempo promedio gastado por un cliente en el sistema es por la ecuación (3.49)

$$W = \frac{1}{\mu_2 - \lambda_2} = \frac{1}{2.3218 - 1.35} = 1.0290 \; minutos.$$

W_0 El tiempo gastado por un cliente en la cola; es decir el tiempo que tiene que esperar en promedio un cliente para iniciar su servicio

$$W_0 = W - \text{tiempo promedio de servicio}$$

$$W_0 = W - \frac{1}{\mu_2}$$

$$W_0 = W - 0.4306 \text{ min.}$$

$$W_0 = 1.0290 - 0.4306 = 0.5982 \text{ min.}$$

$$W_0 = 35892 \text{ segundos.}$$

L_0 El número promedio de personas que esperan ser atendidas. Recordando que $L_0 = \lambda W_0$ se tiene

$$L_0 = \lambda_2 W_0 = 1.35(0.5982) = 0.8075.clientes$$

La Tabla 4.9 resume esta información.

4.3 Análisis Comparativo Entre el Model I y el Modelo II.

Se ha mencionado que el sistema que se forma en la biblioteca de la Facultad de Ciencias sufrió modificaciones a principios de 1992. Ahora se realizará un análisis de las modificaciones realizadas y de las repercuciones de éstas.

4.3.1. Análisis de las Modificaciones al Sistema I.

Estas modificaciones se realizaron en el servicio, es decir, únicamente se alteró el mecanismo de servicio; pero como en todo sistema, la modificación de uno de los elementos de este produce alteraciones en los demás elementos.

La alteración al servicio fue, que en el primer sistema se tenían tres servidores, cuyo objetivo, era la busqueda del material solicitado por los usuarios, y el registro de salida de dicho material. En el nuevo sistema se modificó la disposición del material, y se implanto un sistema de cómputo, así los usuarios buscan su propio material y el servidor, únicamente registra en su computadora la salida de este.

4.3.2 Repercución de las Modificaciones al Sistema.

Como se mencionó anteriormente, las modificaciones a uno de los elementos de un sistema produce alteraciones a los demás elementos. En este caso se modificó el mecanismo de servicio y produjo modificaciones en los arribos; como se puede ver si se compara la tasa de arribos al Sistema I, $\lambda_1 = .95 arribos/minuto$, con la tasa de arribos al Sistema II, $\lambda_2 = 1.35 arribos/minuto$, los arribos se vieron modificados; llegan 70% más usuarios al nuevo sistema que al anterior.

Muy probablemente este aumento en los arribos, sea una consecuencia directa de la reducción del tiempo promedio de espera (W), que se vió fuertemente alterada; en el Sistema I era de 3.8701 min. mientras que, para el Sistema II es de 1.0290 min.; es decir, se redujo en promedio el tiempo de espera en el sistema en en 376.1%.

Además hablando de la intesidad de tráfico, el Sistema I tenía una intensidad de $\rho_1 = 0.9438$, mientras que el actual tiene $\rho_2 = 0.581437$; es decir, la intensidad de tráfico se redujo en un 61.6%. Esto quiere decir que el conflicto que se presenta en la relación demanda–servicio es 61.6% más relajada en el Sistema II.

Se puede hablar también de la longitud promedio de la cola (L) que en el primer sistema resultó de 3.6767 y en el segundo sistema es de 1.3891; lo que representa un 264% de decrecimiento de la longitud promedio de la cola, o de la probabilidad de llegar al sistema y encontrarlo vacío (π_0); que en el Sistema I era 0.01328 y para el Sistema II es 0.418562, es decir ahora es más probable llegar al sistema y encontrarlo vacío.

Estos resultados se resumen en la Tabla 4.10.

En términos generales se puede concluir que el nuevo sistema resulta mucho más eficiente que el anterior.

	λ	ρ_i	π_0	L	W
Modelo I	0.95	0.9138	0.0123	3.6767	3.8701
Modelo II	1.35	0.5814	0.4185	1.3891	1.0290
*	+70%	-61.6%	+315.1%	-264.6%	+376.1%

Tabla 4.10: *Resumen de resultados de las comparaciones de los Modelo I y II. * Representa el aumento o decremento porcentual de las modificaciones realizadas con respecto al Modelo II.*

Bibliografía

[1] Kleinrock, *Queueing Systems.* Vol. 1, sl Theory. Vol. 2 *Computer Applications.* New York: John Wiley & Sons, Inc., 1976.

[2] A.O. Allen. *Probability, Statistics and Queueing Theory.* Orlando: Academic Press, Inc., 1978.

[3] N.U. Prabhu. *Stochastic Storage Processes: Queues, Insurance, Risk and Dams.* Los Angeles: Springer-Verlang, 1980.

[4] Cox-Smith, *Estudio Matemático de las Colas.* México: U.T.E.H.A., 1961.

[5] S. Karlin & H.M. Taylor, *A First Course in Stochastic Process.* New York: Academic Press, Inc., 1975.

[6] H.M. Taylor & S. Karlin, *An Introduction to Stochastic Modeling.* California: Academic Press, Inc., 1984.

[7] E. Parzen, *Stochastic Processes.* San Francisco: John Wiley & Sons, Inc., 1962.

[8] A.M. Mood & F.A. Graybill, *Introduction to the Theory of Statistics.* Singapore: McGraw Hill, Series in Probability and Statistics. 1974.

[9] H.J. Larson, *Introduction to Probability Theory and Statistical Inference.* New York: John Wiley & Sons, Inc., 1978.

[10] E. Parzen, *Modern Probability Theory and Its Applications.* New York: John Wiley & Sons, Inc., 1960.

[11] W.W. Daniel. *Applied Nonparametric Statistics.* Boston: PWS-Kent, The Duxbury Anvanced Series in Statistics and Decision Sciences. 1978.