34

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES

ACATLAN

SIMULACION DE UN SISTEMA DE LINEAS DE ESPERA CON SERVICIO EN MASA

T E S I S

QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN MATEMATICAS APLICADAS Y COMPUTACION

P R E S E N T A :

CARLOS JAVIER SOSA PAZ

ASESOR: ACT. MARIA DEL CARMEN GONZALEZ VIDEGARAY

UNAM
CAMPUS ACATLÁN

JUNIO - 2000

DEDICATORIA

A mis padres Javier y Bertha.

A mis hermanos Gabriel y Sandra

A Lolita y Sofía.

A mis compañeros y profesores de la
Universidad Nacional Autónoma de México
y a los del Instituto Politécnico Naciones.

Y a las personas que de alguna
forma este trabajo les pueda ayudar.

AGRADECIMIENTOS

Al Doctor Juan de la Cruz,
por su paciencia en el desarrollo de este trabajo.

A la maestra María del Carmen González Videgaray,
por su apoyo en la dirección de este trabajo.

A Bety Trueba,
por su ayuda en la revisión y conformación del trabajo.

INDICE

ANTECEDENTES

Uno de los problemas cotidianos al que nos enfrentamos todos los seres humanos son las líneas de espera, ya sea al ir al banco, el transporte público, el trafico, en el supermercado, etc.. El propósito de este trabajo es analizar y proponer una solución alternativa al modelo analítico del modelo de líneas de espera con servicio en masa, como podrían ser el servicio de transporte metropolitano, un elevador, un semáforo, etc..

La teoría de líneas de espera se originó en los trabajos de A. K. Erlang que principiaron en 1909.[1] Experimentó con un problema relacionado con la congestión del tráfico telefónico. Durante los períodos ocupados, los que pretendían hacer llamadas sufrían algunas demoras, porque las operadoras eran incapaces de atender las llamadas con la rapidez con que se hacían. El problema original que trató Erlang fue el cálculo de esa demora para una operadora, y en 1917 los resultados se extendieron al caso de varias operadoras. En ese año Erlang publicó su obra, "Solutions of Some Problems in Theory of Probabilities of Significance in Automatic Telephone Exchanges". Los adelantos en el campo del tráfico telefónico continuaron generalmente en el sentido iniciado por Erlang.

Las publicaciones principales fueron las de Molina en 1927 y de Thornton D. Fry en 1928.[2], pero sólo fue hasta el fin de la Segunda Guerra Mundial cuando esos trabajos se extendieron a otros problemas relacionados con líneas de espera.

En la década de los 50 Bailey, Downton, Kendall, y otras personas realizan publicaciones relacionadas con las líneas de espera con servicio en grupo o en masa, introduciendo en ellos los modelos Markovianos de procesos estocásticos.

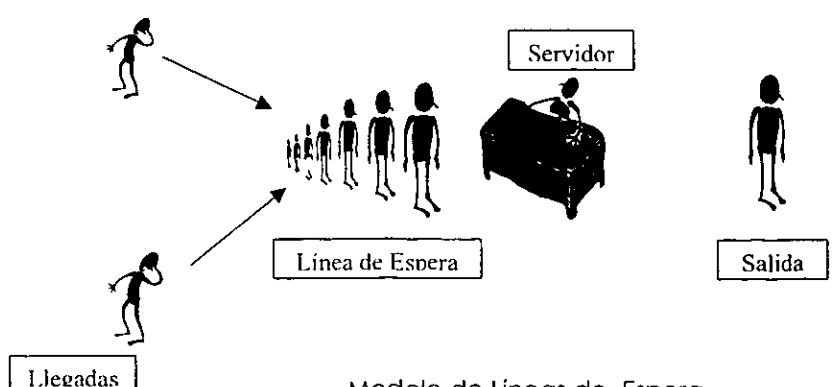
Una cola, o línea de espera, implica unidades que llegan y esperan a ser servidas en las instalaciones que proporcionan el servicio que solicitan. Supongamos que la instalación es una caja registradora en una tienda de abarrotes. Si la cola que se forma es larga, los clientes pueden llegar a impacientarse y marcharse, causando así una baja en los beneficios.

El dueño de la instalación puede decidir que vale la pena hacer una inversión en la compra de otra caja registradora, porque su costo es cubierto por los beneficios proporcionados por los clientes impacientes, muchos de los cuales esperarán entonces a ser servidos. De esta manera se introduce una optimización en la teoría de colas.

La investigación matemática que se ha realizado acerca del proceso de colas simple, en el cual los clientes llegan en forma aleatoria e ingresan a una cola con un orden establecido, por lo general este orden es FIFO por sus siglas en inglés (first in, first out) o PEPS (primero en llegar, primero en salir) los cuales son servidos individualmente: el problema de uno o varios servidores para una o varias líneas de espera ha sido estudiado anteriormente, obteniéndose grandes resultados.

Los modelos que han tenido mayor dificultad para resolverse, son aquellos en los que el servicio es proporcionado en masa o en grupo, y precisamente éste tipo de modelos son los que dan la pauta para el desarrollo de éste trabajo de tesis; convirtiéndose en el objetivo específico de la simulación.

Algunos tipos generales de fenómenos de colas, son los de tráfico de comunicaciones (teléfono, telégrafo, correos), tráfico de transporte (aéreo, terrestre, marítimo), colas para servicios (teatros, restaurantes, autobuses, hospitales y clínicas), inventarios y procesos industriales (mantenimiento, líneas de montaje, interferencias de máquinas), procesos físicos (operaciones de una cuadrilla de un puerto, movimiento de partículas hacia el desagüe), procesos epidémicos en biología, crecimiento de población, incluso selección de artículos para su publicación, consecuencias psicológicas de impulsos nerviosos, dentro de los modelos en masa encontramos el servicio del metro, en un elevador, en los restaurantes, etc.



Modelo de Líneas de Espera

INTRODUCCION

La simulación es un área de estudio que forma parte de la Investigación de Operaciones (IO), la cual es usada prácticamente en todas las áreas de estudio conocidas. Una de sus principales características es que permite estudiar un sistema sin tener que realizar experimentación sobre el sistema real. Sin embargo, ésta no es la única forma de estudiar un sistema; otra posibilidad es construir un modelo analítico conformado por un conjunto de ecuaciones (generalmente diferenciales) que representan al sistema para luego resolverlo para diferentes situaciones, o bien plantear un modelo de optimización que pretende proporcionar la mejor estrategia que el sistema debe adoptar para funcionar mejor de acuerdo con alguna medida de rendimiento establecida en la "función objetivo" y satisfaciendo las diversas condiciones del problema, establecidas en "las restricciones". Los modelos que se obtienen como un conjunto de ecuaciones se denominan con frecuencia modelos analíticos, es decir modelos de ecuaciones diferenciales o de optimización. Por cierto que, los estudiosos de las ecuaciones diferenciales afirman con orgullo que todos los modelos analíticos son de ecuaciones diferenciales, ya que incluso una simple ecuación algebraica es una ecuación diferencial de orden cero.

¿QUÉ ES SIMULACIÓN?

Simulación es una palabra que es familiar a los profesionales de todas las disciplinas e incluso para aquellos que no han estudiado una carrera profesional. De esta manera el significado de la palabra *Simulación* se explica casi por sí misma. Entre los significados que podemos obtener de la gente común y corriente para la palabra "Simular", se encuentran los siguientes: "Imitar la realidad", "emular un sistema", "dar la apariencia o efecto de un sistema o situación real".

Algunas aplicaciones de Simulación que podemos citar son los siguientes:

- Operaciones de mantenimiento
- Simulación del Tráfico de un sistema (Teleproceso, Tráfico aéreo y terrestre, telecomunicaciones. telefonía....).
- Cambios en la configuración de un sistema.
- Simulación económica.
- Estrategias militares.
- Líneas de espera
- Control de inventarios.
- Líneas de producción.

Este trabajo se compone de 4 Capítulos en los que se realiza una descripción de cómo trabajan los modelos de líneas de espera sus aplicaciones, las diferentes distribuciones de probabilidad para las llegadas y para los tiempos de servicio, así como el comportamiento general de los clientes y las disciplinas que pueden existir en diferentes sistemas; también se mencionan algunos valores característicos de las líneas de espera, como los estadísticos más relevantes de éstas.

Posteriormente nos presenta las diferencias que existen entre los modelos de líneas de espera determinísticas y las aleatorias.

Se hace una descripción de los procesos de Poisson, muerte y nacimiento pura, las formulas de Pollaczek y Khintehine. el modelo de Erlang. Clasificación de Kendall y Lee.

Y se describe la parte fundamental de este trabajo, el de las líneas de espera con servicio en masa. La forma en como se ha intentado resolver este sistema en forma analítica sin tener grandes resultados, con excepción de los obtenidos por Dawton, quien sólo los pudo encontrar para el estado estable del sistema.

En el tercer capítulo se presenta una propuesta alternativa a la solución de los modelos de líneas espera con servicio en masa alternativo al analítico, que es precisamente el de la simulación.

Finalmente en el último capítulo se dan algunas recomendaciones y conclusiones del trabajo realizado.

Capitulo 1

CAPITULO I

1 MARCO TEÓRICO

1.1 ELEMENTOS BÁSICOS DEL MODELO DE LÍNEAS DE ESPERA

Una operación de colas se divide en cuatro partes: el ingreso, la línea de espera, la prestación del servicio, y la salida. Con cada una de ellas se asocia una serie de hipótesis alternativas de estudio, algunas de las cuales han sido materia de investigación en este campo.

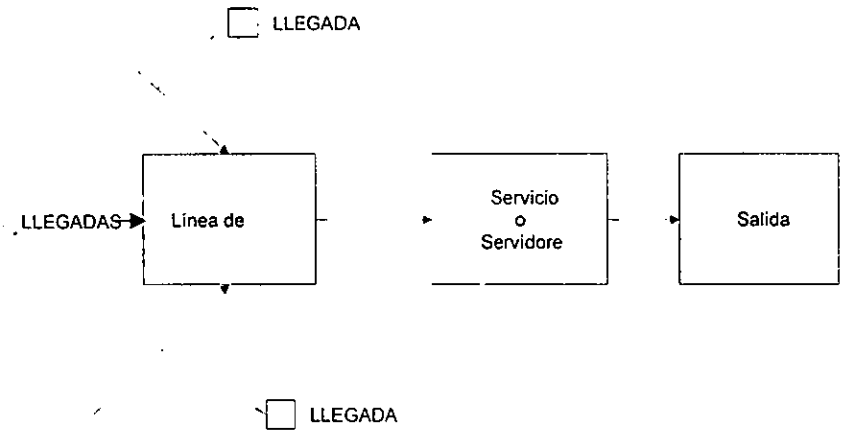


Diagrama I

1- Tipo de distribuciones de llegada y tiempos de servicio.

La forma en que llegan las unidades es ***aleatoria***, si no puede predecirse exactamente cuándo llegará cierta unidad. El tiempo de llegada es una variable aleatoria que puede describirse matemáticamente con una distribución de probabilidad. La que vemos, depende de la forma de las llegadas, como lo muestran los datos observados y la naturaleza de las operaciones. Una de las distribuciones que se encuentra más comúnmente en los problemas de líneas de espera, es la distribución Poisson (por sus características que se verán más adelante), que se emplea en problemas de líneas de espera de un solo canal para llegadas aleatorias, en las que el tiempo de servicio que se proporciona disminuye en forma exponencial.

Las llegadas a una cola (que pueden tener un número inicial de unidades esperando antes que empiece el servicio) se producen conforme a cierta distribución de frecuencias, sobre la que cabe hacer hipótesis, y así se generan los intervalos entre llegadas. Estos intervalos pueden estar distribuidos independientemente en muchos casos prácticos o pueden ser dependientes, como por ejemplo, en el caso de una corriente de automóviles que dejan un semáforo. Las mismas observaciones se aplican a los tiempos de entrada en servicio y a los tiempos de servicio.

2- Variedades del ingreso inicial.

El número inicial de unidades en un sistema cuando empieza una operación puede estar dado por una distribución ya que es diferente para cada ciclo completo de la operación. El ingreso a una cola puede provenir de una población limitada o de una ilimitada, que pueden también constar de varias categorías (poblaciones) de clientes, cada una de las cuales pueden tener las llegadas por distribuciones distintas, de uno en uno o en tandas, y pudiendo ponerse en cola las unidades según un orden prescrito. La distribución de ingreso puede depender de la salida, como en un hospital, donde los pacientes son admitidos si hay camas vacantes.

3- Comportamiento de los clientes.

a) Rebelión: El comportamiento del cliente puede variar. Los clientes que llegan pueden rebelarse (esto es, no unirse a la cola), por la longitud de la cola existente o simplemente por no querer esperar y, en consecuencia, son clientes que se pierden. Algunas veces se pierden porque no tienen ocasión de esperar, como en el caso de señal de ocupado de una llamada telefónica (llamadas perdidas), aunque puede volverse a llamar. Es también posible mantener tal llamada hasta que una línea quede libre. Puede haber una única línea para esperar antes de entrar en servicio o pueden existir varias líneas, como en los Bancos o en los supermercados. Una unidad puede unirse a la línea más cercana independientemente de su longitud. Si está planeado que las llegadas ocurran a intervalos constantes, ellas pueden por ejemplo, presentarse más pronto o más tarde, según una distribución con media en el punto previsto de llegada.

b) Influencia de una información incompleta: En muchos problemas hay que tomar una decisión, como por ejemplo, a qué cola unirse en una operación de múltiples colas, cuando la información concerniente al sistema disponible es escasa, es un caso de información incompleta. En un tráfico congestionado la falta de conocimiento sobre cuál es la mejor ruta a seguir, sin comprobarlas todas, es también "información incompleta".

c) Adaptación del cliente a las condiciones de la cola: Basándose en la experiencia, los viajeros pueden saber cómo tienen que desplazarse, sí más rápido o más despacio, para evitar una cola intolerable. y tales medidas. cuando son adecuadamente estudiadas, pueden incluso evitar una congestión. Una experiencia corriente es que una unidad se una a una línea de espera larga en horas de cierre por miedo a que otra más corta que encuentre sea clausurada más pronto. Hay situaciones en las que, aunque una unidad llegue antes que otra, debe ingresar primero la que le sigue. Hay casos en los que cada instalación de servicio tiene su propia especialidad y, por tanto, su propia cola. como una ventanilla para venta de sellos e impresos de giros o para correspondencia certificada en una estafeta de correos.

d) Colusión, traslados. y renuncias: Varios clientes pueden estar en colusión por lo que una sola persona espera en línea mientras las restantes están libres o atienden a otros asuntos. Algunas, incluso, pueden dedicarse a pasear durante la espera. Las unidades pueden trasladarse de una línea a otra, como en un Banco. Un cliente puede perder la paciencia y abandonar la línea, es decir renunciar.

4- Variedades de colas y canales:

a) Disponibilidad plena o limitada: Los canales de servicio pueden estar disponibles para cualquier unidad que espera en un sistema (disponibilidad plena) o pueden estarlo solamente para algunas de las unidades que esperan. Otras unidades están bloqueadas y tienen que esperar hasta que un canal que puede prestarles el servicio requerido esté disponible. En los sistemas de enlace telefónico, el que se obtenga una conexión libre depende de que una entrada libre en la siguiente línea de espera pueda ser combinada con una salida libre. La idea, por sí misma, muestra la necesidad de economizar enlaces al establecer las combinaciones posibles. Esto es particularmente aplicable en algunas llamadas interurbanas que tienen que pasar a través de más de un centro apartado.

b) Procedimientos o disciplinas de servicio: Si bien los clientes en línea pueden ser escogidos para el servicio por asignación a los canales de una forma ordenada primero en llegar primero en salir o al azar, se les puede asignar prioridades con errores cometidos cuando no está claro qué prioridad asignar o la prioridad que se asigna cambia con el tiempo. Las prioridades pueden ser absolutas o pueden reconocerse al terminar un servicio. Finalmente. las unidades pueden ser elegidas para el servicio de la forma último en llegar. primero en ser servido.

c) Colas en común: Hay varias formas de colas en común. Algunas dan por resultado una espera media más corta, en particular cuando la dispersión del tiempo de servicio por un servidor, ante el que se formó una cola separada, es alta.

d) Instalaciones en serie y en paralelo: Una instalación de servicios puede constar de varios canales en paralelo, algunos de los cuales pueden estar en serie con otros canales o varios canales en paralelo pueden conducir a uno o más canales en serie. En el caso de canales en serie, puede formarse o no una cola ante cada canal. En un supermercado, un cliente puede servirse él mismo en cuanto llega, y así, el número de canales de servicio (aunque no el de cajas registradoras) varía con el número de llegadas. Todos los clientes hacen cola ante las cajas registradoras para un segundo servicio. Un canal de servicio puede tener varias distribuciones de servicio según las diferentes categorías de clientes. Los servidores ociosos pueden ser dedicados a otras tareas cuando no hay cola. Por ejemplo: en reparación de máquinas, encargarse de trabajos auxiliares. Esto depende con frecuencia de la cantidad de servicio solicitado, de la capacidad del servidor y de la extensión de la cola. El canal puede ser móvil como una cinta transportadora en línea de producción con un servidor que ejecuta el trabajo sobre unidades espaciadas a lo largo de la cinta.

e) Canales de servicio especiales: algunos canales de servicio pueden ser especiales mientras que otros son corrientes, como en ciertas instalaciones de tráfico aéreo, en las que algunos mostradores están para servir a los pasajeros cuya hora de partida está dentro de un cierto intervalo. Así, la demanda de servicios especiales de los clientes que llegan varía, dependiendo de la longitud del intervalo entre su llegada y la hora de salida del avión. Los canales en paralelo pueden cooperar en su totalidad para servir las distintas necesidades de cada cliente. Es posible que los clientes tengan que reincorporarse a una línea de espera para servicios adicionales.

f) Interferencia de colas: Dos colas pueden interferirse una con otra, como en el caso de una única senda para el tráfico en una carretera en obra, que tenga que ser utilizada por los automóviles que marchan en ambas direcciones. Los automóviles de una de las direcciones deben de esperar si pasa un automóvil que marcha en la otra dirección. Si llegan más automóviles en esa dirección antes que el otro acabe de pasar, generalmente también intentarán pasar ellos, y muy probablemente la circulación se hace en tandas o en grupos.

5- La salida de una cola:

La salida de una cola puede también ser importante, en especial cuando constituye el ingreso en otra cola en serie con la primera. Las distribuciones de llegadas y servicios pueden depender una de otra. Así por ejemplo: existe correlación si el servicio influye en las llegadas, y recíprocamente. Así en una serie de empresas idénticas que compiten entre sí, aquella que ofrezca un servicio especializado y quizá más rápido, obtendrá una mayor afluencia de clientes.

1.2 ANÁLISIS DESCRIPTIVO ELEMENTAL

Si se consideraran las llegadas y los servicios a intervalos de tiempo constante, sería evidente que es necesario para este caso obtener el número de unidades que esperan y su tiempo de espera.

Sin embargo, en general, no existe en las colas tal regularidad. Los clientes pueden llegar al azar y, por tanto, no es posible, de antemano, estar seguros del instante exacto en que se producirá la llegada. Y lo mismo es aplicable a los tiempos de servicio. Como los clientes esperan, existe una probabilidad de que en cualquier instante haya un cierto número de clientes en línea, pues los tiempos de llegadas y de servicio son dados ahora probabilísticamente; por consiguiente, sólo puede lograrse una evaluación general de la operación a través de un análisis teórico.

Evidentemente, con cada cliente se asocia una serie de probabilidades y de variables aleatorias. Así por ejemplo para el cliente n-ésimo se tiene una variable aleatoria que describe el instante de su llegada, otra que describe su tiempo de espera, otra su tiempo de servicio, etc. Con un conjunto de clientes, se tiene una familia de variables aleatorias para cada una de las situaciones descritas. Cada variable aleatoria puede tener diferente distribución de probabilidad (característica del cliente n-ésimo) para cada instante. Por su dependencia del tiempo y porque puede haber, al menos en teoría, un número infinito de ellas, estas familias de variables aleatorias constituyen procesos estocásticos. Un proceso estocástico es una familia de variables aleatorias que dependen de ciertos parámetros. Para este análisis se tiene un único parámetro que es el tiempo.

El estudio de la teoría de colas está ligado con el problema de encontrar las distribuciones de varias combinaciones (sumas y diferencias) de dichas variables aleatorias cuando se tiene suficiente información sobre ellas. El principal problema es entonces:

1) Relacionar correctamente las variables para describir un problema de colas.
2) Deducir las distribuciones asociadas y determinar su forma real por medición estadística.
3) Emplear las distribuciones para deducir medidas útiles.
4) Aplicar estas medidas para mejorar una operación.

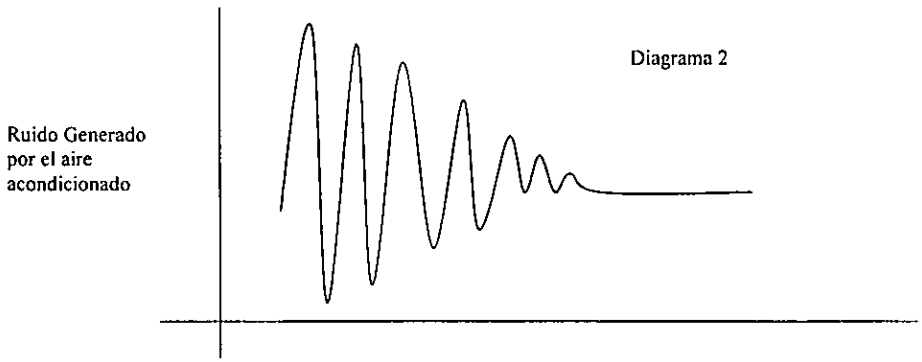
Un fenómeno interesante de la teoría de colas es el que, las diferentes cantidades implicadas dependen del instante en que se examina el sistema. Así por ejemplo el número de clientes en la cola ante la caja de un supermercado será diferente a los diez minutos de ser abierto el establecimiento que dos horas después. En cada instante hay una distribución de probabilidad para la longitud de la cola o para que el tiempo de espera tome un valor prefijado. Por tanto, la distribución de probabilidad resultante variará con el tiempo.

El sistema varía en el tiempo a partir del número inicial de clientes en el instante cero, el cual condiciona los patrones (distribuciones) de la cola. Después el sistema puede verse libre de este efecto (o del efecto de que no hubiera nadie en la línea al principio) y operar con las entradas y salidas de clientes que establecen, finalmente, patrones que son independientes del número inicial. Cuando el efecto inicial se ha "desgastado" se puede esperar encontrar el sistema con el mismo patrón general de probabilidades en un instante dado que en cualquier otro instante. Así, el patrón por sí mismo deja de cambiar con el tiempo y entonces las probabilidades adquieren un equilibrio o estructura de estado estacionario.

Es posible que el equilibrio jamás se alcance por ejemplo, si el tiempo de servicio es mayor que el intervalo entre llegadas. Por tanto, en las colas se tiene siempre el problema de saber si este equilibrio en las probabilidades existe realmente, es decir, bajo qué condiciones las distribuciones asociadas a la cola llegan a ser independientes del tiempo. Es un problema que será esencial considerarlo en algunos casos. En la práctica, muchas operaciones alcanzan, aproximadamente, ese estado de equilibrio y, por consiguiente, se estudian soluciones de equilibrio, por ejemplo de las distribuciones del tiempo de espera y del número de unidades en el sistema. Obsérvese que el equilibrio significa que las probabilidades son independientes del tiempo, pero no que el sistema llega a ser determinista. La cola continúa fluctuando, pero las distribuciones que la describen son fijas en el tiempo. Si se encuentra el equilibrio, entonces podemos encontrar medidas de tendencia central.

Puede darse una imagen de la idea de estado estacionario por medio de una unidad de aire acondicionado similar al que se utiliza durante el verano en estados con clima caluroso. Este aparato es muy viejo y, por tanto, muy reacio a funcionar. La corriente inicial es para él un choque. Se pone en marcha con ruidos estrepitosos y vibraciones. Después de un rato da la impresión de que se recupera, pero existe una resistencia a abandonar su estado inicial de inercia. Después de un largo rato de marcha, los ruidos se van acallando y las vibraciones amortiguándose.

Es posible que el aire acondicionado no alcance el estado operacional estacionario, sino que puede continuar en un estado de transición bajo la influencia de las condiciones iniciales. En tiempo caluroso, cuando más duramente tiene que trabajar, esto es lo que mejor puede esperarse de él. Tal es el caso de algunas colas; al depender de sus distribuciones de ingreso y servicio y de su estado inicial, es posible que nunca se alcance un estado estacionario.



Frecuentemente es más fácil y más práctico suponer que el comportamiento de una cola es independiente del tiempo y el problema se estudia entonces en estado estacionario. Las soluciones estacionarias no siempre son adecuadas y es a menudo deseable el tener soluciones dependientes del tiempo.

Debido a las probables semejanzas entre los clientes, se supone, frecuentemente, que sus distribuciones de llegada y de servicio son idénticas y no depende de si un cliente llega en el instante ***n*** o en el ***m***. En muchos problemas se supone que no es probable que llegue más de un cliente en un intervalo pequeño de tiempo. Estas distintas hipótesis se hacen según como se ajusten a la situación real y como faciliten encontrar una solución.

Como ilustración que relaciona algunas de las variables, observemos que el tiempo de espera, cuando hay cola, del cliente $(n+1)$ - ésimo, w_{n+1}, es igual al tiempo de espera del cliente n-ésimo, w_n, más su tiempo de servicio s_n, menos el intervalo de tiempo de llegada del cliente $(n+1)$ - ésimo, t_{n+1}. El tiempo de llegada del primer cliente no importa ya que entra en servicio inmediatamente.

Se tiene:

$$w_{n+1} = w_n + s_n - t_n$$

Cada s_n y t_n están dadas por una distribución de probabilidad, así como w_n y w_{n+1}. Hay que determinar estas distribuciones desconocidas. Es, por tanto, esencial conocer algunos de los métodos elementales de la teoría de probabilidades para aplicarlos aquí.

Nótese que, si no hay un sitio para esperar, se pierden los clientes que al llegar no pueden obtener servicio. Obsérvese, también, que los canales de servicio pueden tener una estructura complicada, tal como una combinación de canales en serie y en paralelo. El problema se hace aún más interesante y útil.

Un ejemplo de cantidad que es útil calcular es el tiempo de espera. Ahora bien: si la disciplina del servicio es primero en llegar primero en salir, el tiempo de espera en la línea del k-ésimo cliente en ella, es la suma de los tiempos de servicio de las $k-1$ unidades que hay delante de él y del tiempo de servicio que le queda a la unidad que está sirviéndose. Cada uno de ellos es una variable aleatoria que toma valores según unas distribuciones de probabilidad. El problema es hallar la probabilidad de que la suma de estas variables tome cierto valor.

La distribución del tiempo de espera se obtiene tomando la probabilidad de que haya n unidades en la cola y multiplicándola por la distribución del tiempo de espera de la unidad n-ésima, sumando para todos los posibles valores de n. El tiempo medio de espera puede obtenerse de la distribución del tiempo de espera, hallando su media. Obsérvese que si los clientes son elegidos al azar, la distribución de la espera de la unidad que llega después de otra $n-1$ que estaban en la cola será diferente de la que se dijo anteriormente para una cola ordenada.

En el caso de que se encuentren soluciones analíticas, queda todavía el problema de cómo aplicar los resultados a una situación práctica. Así, por ejemplo un problema puede ser estimar las tasas de llegadas y de servicio para una situación práctica, utilizando los valores más verosímiles de estos parámetros en las ecuaciones. ¿ Cómo se harían las determinaciones con un error mínimo?. Si las ecuaciones, cuando se dispone de ellas, no son resolubles analíticamente, se deben ensayar soluciones numéricas.

Hasta ahora se ha expuesto someramente el tratamiento analítico del problema, es decir, cuando es posible describir analíticamente una situación de colas y obtener, también por métodos analíticos, expresiones útiles, tales como el tiempo medio de espera y la longitud media de la cola, de las que pueden deducirse otras muchas determinaciones de utilidad. Con frecuencia sucede que un problema de colas es complicado y no puede describirse fácilmente por ecuaciones analíticas que puedan proporcionar soluciones útiles. El método alternativo es simular el sistema.

El problema de ajustar un modelo teórico a una operación práctica de colas, o a alguna parte de ellas (si por ejemplo puede dividirse en partes, cada una de las cuales satisface condiciones de equilibrio), está principalmente determinado por cuál medida se necesita tomar.

Para muchos casos prácticos se requieren índices que permitan hacer comparaciones. Una de tales comparaciones es el efecto sobre la distribución del tiempo de espera del tipo de servicio utilizado. Otra es el efecto de la disciplina de la cola sobre el tiempo de espera.

Cualquiera que sea la técnica utilizada para estudiar el problema, debe decidirse qué medidas pueden aplicarse a la operación en orden a tomar una decisión. Así por ejemplo el director de una instalación puede decidir entre incrementar el espacio dedicado a esperar o el número de las instalaciones de servicio, comparando el costo de ampliar la instalación con la pérdida de negocio debida a los clientes que se van. El tiempo medio de espera de los clientes a los que les es necesario esperar puede ser demasiado grande y, en consecuencia, el tiempo medio de espera de todos los clientes es a veces una medida inapropiada. El dueño de una empresa puede usar estas cantidades para determinar el espacio necesario para los clientes que esperan y cómo hacer agradable su espera y quizá más corta incrementando el número de canales, instalando servicios más rápidos o haciendo otras cosas.

Una medida de efectividad puede ser útil a un cliente para decidir abandonar o no unirse a una cola, si su longitud media o su tiempo de espera es demasiado grande.

Algunas parámetros importantes para el estudio de un sistema en equilibrio (es decir, cuando el tiempo no afecta ya a la operación, o en estado de transición, o sea dependiente del tiempo) son:

1. La tasa de ingreso.
2. La tasa de servicio.
3. La intensidad de tráfico. esto es, la razón de la media del tiempo de servicio a la media de las longitudes de los intervalos entre llegadas sucesivas.
4. La probabilidad de que el número de unidades en la cola o en el sistema sea un número dado n.
5. El número medio de unidades en cola (o en el sistema, si se incluye el número medio en servicio).
6. La distribución del tiempo de espera, su media y varianza.
7. Cuando intervienen en el proceso de espera mercancías perecederas y cuando existen clientes impacientes, es importante la probabilidad de no esperar más de un período dado.
8. Las probabilidades de no esperar, o sea la proporción de unidades que son servidas sin demora (la proporción de unidades que sufren demoras se obtiene restando la anterior de uno).
9. La probabilidad de que haya alguien esperando.
10. El número medio de individuos esperando, para aquellos que han sido demorados y tiene que esperar.
11. La longitud media de un período de actividad y el número medio de unidades que están siendo servidas. Particularmente importante para un cliente que requiere un pequeño servicio, es la razón de la media del tiempo en cola a la medida de servicio.
12. Es también útil calcular la probabilidad de que exactamente $0 \leq n \leq c$ canales estén ocupados.
13. El número medio de canales ociosos; el coeficiente de pérdida, por estar ociosos, para los canales.
14. La eficiencia operativa, es decir, la proporción de tiempo que un canal de servicio está activo sirviendo a clientes.
15. El número medio de unidades que están en el sistema de cola, en el caso de una población finita; la probabilidad de una llamada pérdida, etc.

El análisis de los costos son esenciales para muchas decisiones en los problemas de teoría de colas.

Puede ser necesario conocer el costo de la operación total; el de un canal adicional, comparado con el de incrementar el servicio con un canal; el costo de la disciplina de cola elegida; el gran número de canales comparando con el de usar el espacio para otros fines; el costo de mantener diferentes longitudes de colas y diferentes extensiones en las salas de espera; el costo de controlar la distribución de ingreso (por ejemplo espaciar en el tiempo la llegada de aviones y comparar el costo de operar con una distribución dada de servicio constante). Puede también ser deseable adoptar medidas de utilización de la instalación y medidas de eficiencia de operación o de salida por unidad de tiempo. El costo de la pérdida de clientes y su tiempo de espera, son también importantes.

El servicio puede verse interrumpido por muchas causas. Así por ejemplo una gran influencia en el ingreso puede crear tal presión en el sistema que haya que interrumpir la operación; o en sistema telefónico, un temporal puede interrumpir por algún tiempo el servicio. Un cambio radical en el servicio puede desanimar a los clientes por completo y verse suspendida así la operación entera. Un accidente inesperado en un Banco (por ejemplo un bandido que asalta a un empleado) puede causar la clausura de la operación total. Una congestión al final del servicio; bien porque los clientes servidos no se marchen y por otras razones que pueden paralizar la operación.

Puede lograse mejores rendimientos mediante la disminución del tiempo de servicio así como en la reducción de la variación en el tiempo de servicio, por disponer de un servicio adicional en períodos de gran afluencia y por el control de la distribución de llegadas. Es ventajoso, por ejemplo, controlar el ingreso para hacerlo regular en vez de aleatorio que conduce a un tiempo en cola más largo para un tiempo de servicio dado. Es evidente que esto sólo puede hacerse en parte. Los clientes continuarán llegando con variación alrededor del intervalo regular. Es interesante suavizar las irregularidades que persisten en la llegada.

Es importante que las llegadas sean controladas de modo que la intensidad de tráfico (razón de la tasa de llegada sobre la tasa de servicio) permanezca menor que la unidad. El escalonamiento de las horas de trabajo, para regularizar el ingreso y la utilización de las instalaciones en un período extenso, podría tender a aliviar graves congestiones.

Hay casos en los que no se encuentra una solución realmente satisfactoria, como no sea la de sustituir la operación por otra más eficiente. Un ejemplo de una solución en algunas instalaciones es que el cliente se sirva él mismo. Las carreteras y autopistas con pasos elevados alivian la congestión del tráfico, que puede continuar presentando dificultades, aún cuando algunos "cuellos de botella" estén bien controlados. Un teléfono puede ser utilizado asignando turnos y reservas para hablar, eliminando así el tiempo malgastado en esperas.

Queremos hacer la observación de que si se quiere solucionar un problema de colas incrementando el número de operarios para prestar el servicio, puede crearse una nueva cola: la de los mismos operarios. que ahora tienen que esperar ociosos la llegada de clientes que demanden su servicio. En este sentido, se crea una nueva congestión, la de canales ociosos, quizá menos deseables que la que se quería remediar. Asociado a esto tendríamos que realizar un estudio relacionado con costos:

El objetivo de un modelo de costos de la espera es el de determinar el nivel de servicio (la tasa de servicio o el número de servidores) que "equilibre" los dos siguientes costos en conflicto:

1. Costo de ofrecer el servicio.

2. Costo que resulta de la demora en el ofrecimiento del servicio.

El primer costo está relacionado con la operación de la instalación y el segundo representa el costo de demora de los clientes. En forma intuitiva, vemos que un incremento en el nivel de servicio debe reducir el tiempo de espera del cliente y viceversa. Esto significa que cuando aumenta (disminuye) el costo de operación de la instalación debido al incremento (disminución) del nivel de servicio, el costo de espera debe disminuir (aumentar). En el diagrama 2 presenta un resumen de los costos, como función del nivel de servicio. El nivel óptimo de servicio se escoge para minimizar la suma de los dos costos. Observamos que ambos costos se definen por unidad de tiempo para mantener el modelo dimensionalmente correcto.

De los dos costos citados, el costo de espera es el más difícil de estimar. Esto sucede particularmente cuando el cliente es un ser humano cuyos intereses quizá no estén en armonía con los del servidor.

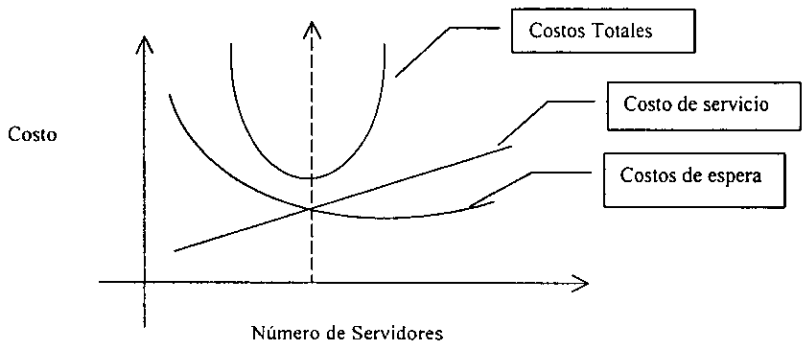


Diagrama 3

1.3 LINEAS DE ESPERA

Comenzaremos nuestro estudio con un problema de una cola determinista, que no ha recibido suficiente atención, por su simplicidad y puede ayudar a aclarar un gran número de conceptos encontrados en las líneas de espera.

Después realizaremos una transición del aspecto determinista al probabilístico de las líneas de espera. Con esto podremos realizar una pequeña discusión del proceso de Poisson, que tiene propiedades muy útiles para las líneas de espera. Basado en el proceso de Poisson y la distribución exponencial ligada a éste, se deduce un segundo modelo de líneas de espera, en el caso de dependencia del tiempo.

Se aplica un análisis similar para estudiar la solución en equilibrio, es decir, independiente del tiempo, de una línea de canal único, con ingreso de Poisson ordenado y de tiempo de servicio exponencial.

1.3.1 Una cola determinista

Comenzaremos estudiando un problema elemental de colas. En el ejemplo siguiente, se toman como fijos los tiempos de llegadas y los de servicio de los clientes. Es posible estudiar entonces la evolución de la línea de espera. Nuestro objetivo es encontrar cuántos tienen que esperar en la cola y cuál es el tiempo total de espera.

Suponemos que una operación de colas se organiza de tal forma que las llegadas ocurren a intervalos regulares y son servidas regularmente. El problema es determinista y su tratamiento elemental.

Si los clientes llegan a un único canal, a intervalos regulares de tiempo de longitud a (es decir, la tasa de llegadas es $1/a$), y son servidos a intervalos regulares de tiempo de longitud b (tasa de servicio $1/b$), entonces si $b < a$, es decir, si $b/a < 1$ no habrá espera. Si $b/a > 1$ el número de clientes esperando crecerá indefinidamente. Si $b = a$, no habrá espera si la operación comienza con una línea vacía.

Si no es así, la cola se mantendrá con longitud constante.

Supongamos que $b/a < 1$ y la operación empieza con una línea de i clientes (obsérvese que $i \geq 2$, pues si $i = 1$ este cliente terminará de ser servido antes de la primera llegada y no habrá espera). La totalidad de los i clientes, serán servidos al final de un intervalo de tiempo de longitud ib. Sin embargo, entonces $[ib/a]$ clientes habrán llegado y tendrán que esperar. Los corchetes indican el mayor entero que es menor que ib/a, ya que los clientes no llegan en fracciones. El último cliente en llegar espera ante la instalación de servicio hasta ser servido.

El tiempo de servicio de estos clientes que, por supuesto, esperarán la línea, es $b[ib/a]$ y de nuevo durante este tiempo $[ib/a + (ib/a)(b/a - 1)]$ clientes habrán llegado. Obsérvese que este número será menor que $[ib/a]$, ya que $b/a < 1$, etc., hasta que se llega a un momento a partir del cual los clientes que llegan no necesitan esperar.

Como ejemplo supongamos que el tiempo de servicio b es la unidad de tiempo; sea a igual a tres unidades de tiempo. Supongamos que $i=50$. El tiempo de servicio del *i-ésimo* cliente que espera es 50 unidades de tiempo. Durante este tiempo han llegado $[50/3] = 16$ unidades que requieren 16 unidades también de tiempo para ser servidas, durante las cuales llegarán seis unidades. Durante el tiempo que tarda en ser servida la última, llegarán dos unidades. Una unidad llegará al término del servicio de la última de estas dos unidades y a partir de entonces no habrá que esperar. El número total de unidades que han esperado en la línea es 74. Se incluye al primero de los *i-clientes* iniciales que inmediatamente entra para servirse al comenzar la operación.

Si denotamos por $P_n(t)$ la probabilidad de que haya n unidades esperando en el instante t, entonces, evidentemente:

$$P_n(0) = \begin{cases} 1 & \text{si} \quad n = i \\ 0 & \text{si} \quad n \neq i \end{cases} \qquad 1.3.1.1$$

Estas dos posibilidades se resumen con el símbolo $P_n(0)=\delta_{in}$ que se conoce como delta de Kronecker. Es igual a 1 si $n= i$ y 0 si $n \neq i$. En este caso podemos saber con certeza que número es. Todo está determinado exactamente; para cualquier valor del tiempo $P_n(t)$ es 0 ó 1.

Pero este no es el caso para modelos probabilísticos que trataremos más adelante.

Es evidente que la longitud de la cola se hace cero después de un espacio de tiempo en el que cada unidad que ha esperado ha sido también servida. Sean A las nuevas llegadas que han de esperar en la línea. Entonces la llegada $(A + 1)$ ocurre después del tiempo de servicio de $i + A$ clientes; es decir: $(i + A)\, b \leq a\,(A + 1)$ o $ib - a \leq A\,(a - b)$.

El número A es el menor entero tal que $A = [(ib - a) / (a - b)] + 1$. Por tanto, la cantidad de tiempo requerida para servir a todas las unidades que han esperado es:

$$b\left(i + \left[\frac{ib-a}{a-b}\right] + 1\right) = b\left[\frac{ia-b}{a-b}\right] \qquad 1.3.1.2$$

De nuevo se incluye al primero de los i clientes iniciales.

Obsérvese que cuanto más próximo sea b a a, más clientes tendrán que esperar hasta que, después de un espacio de tiempo mayor desaparece la línea de espera. Considerando una situación ligeramente diferente podremos escribir expresiones más útiles y fácilmente calculables. Para el resto de esta discusión propondremos las condiciones de que b divide a a y a divide al tiempo.

Supongamos que empezamos la operación admitiendo en el servicio a uno de los clientes que espera inicialmente; si él termina el servicio y no hay ninguna llegada, otro de los clientes iniciales es admitido. Si un cliente llega durante su servicio o el de uno de los siguientes clientes iniciales, la unidad será admitida en el servicio en cuanto la instalación quede vacante. Durante el tiempo de este servicio es posible que llegue otra unidad. Esto depende de la magnitud de $a - b$. El cliente que llegue entrará en el servicio en cuanto quede vacante y todos los clientes que puedan haber llegado son también servidos. Cuando no haya clientes esperando, será admitido en el servicio otro de los i clientes iniciales; las llegadas adicionales serán de nuevo servidas en cuanto quede vacía la instalación, y así hasta que todos los i clientes han entrado en el servicio. Cuando el último de los clientes iniciales entra en el servicio, no habrá ninguno esperando. Si alguno llega, durante este tiempo de servicio, espera y es servido, etc., hasta que los clientes que lleguen no tengan que esperar.

Hemos considerado la operación en esta forma para hacer evidente el siguiente análisis: Consideremos los incrementos de tiempo de servicio ganado por el hecho de que $a>b$. Así, hay una ganancia de $a - b$ unidades de tiempo para cada uno de los clientes que llegan. Teóricamente, puesto que el tiempo de servicio es b, $b / (a - b)$ clientes deben llegar para contribuir a diferir suficientemente el tiempo de servicio a uno de los $i - 1$ clientes que esperan y que se encontrarían allí después del comienzo de la operación. Obsérvese que el primero de ellos ingresó en el servicio inmediatamente. Así, en total, $(i - 1)b/(a - b)$ llegadas son necesarias hasta vaciar la línea.

Por tanto, el número total de clientes que han esperado en la línea desde que comienza la operación es:

$$i-1+\frac{(i-1)b}{a-b}=\frac{(i-1)a}{a-b} \qquad 1.3.1.3$$

y si incluimos el cliente inicial en el servicio, se tiene:

$$\frac{(i-1)a}{a-b}+1=\frac{ia-b}{a-b} \qquad 1.3.1.4$$

Por tanto, el tiempo total hasta que la instalación de servicio está ociosa por primera vez es:

$$T\equiv\left(\frac{ia-b}{a-b}\right)b \qquad 1.3.1.5$$

Luego si un cliente llega en el instante $t < T$, entonces t/a clientes habrían llegado antes que él, elevando el total de aquellos que han esperado en línea a $t/a + i - 1$ y t/b clientes, habrían sido servidos durante el tiempo t. Nótese que no hay pérdida de generalidad escribiendo t como si fuese continuo, pero recordemos que t, en el que ocurre una llegada, debe ser un entero, múltiplo de a. Por tanto, en el instante t habrá:

$$\frac{t}{a}+(i-1)-\frac{t}{b}=t\left(\frac{1}{a}-\frac{1}{b}\right)+(i-1) \qquad 1.3.1.6$$

clientes en línea delante de él. (Obsérvese que si $b > a$ este número será creciente en proporción a t). Sin embargo, habrá $t\ (1/a - 1/b) + i$ clientes en el sistema delante de él.

El tiempo gastado esperando en la línea por un cliente que llega en el primer múltiplo de a después de t es:

$$W(t)=\begin{cases} 0 & T-a\le t \\ \left(t\dfrac{b-a}{ab}+i\right)b-a & 0<t\le T-a \\ (k-1)b & t=0 \end{cases} \qquad 1.3.1.7$$

donde $W(O)$ es el tiempo en la línea para el *k-ésimo* miembro del grupo inicial de i clientes. Esta solución en el estado de transición, que depende del tiempo y del número inicial, alcanza el estado estacionario después del tiempo T, pues entonces el sistema adquiere características independientes del tiempo y del número inicial i.

El tiempo total gastado dentro del sistema por un cliente que llega en el primer múltiplo de a después de t es:

$$W(t) + b \qquad 1.3.1.8$$

donde se incluye el tiempo gastado en ser servido.

Supongamos ahora que el tiempo de servicio y los tiempos de llegadas se alteran durante el período T, siendo las nuevas cantidades b_1 y a_1, respectivamente. Sea $t_0 < T$ el instante en que ocurre esto. La unidad que está siendo servida terminará de serlo con el régimen definido por el tiempo b. Entonces el número de unidades en línea y la unidad de servicio se convierten en el ingreso inicial de una nueva línea. Se establece un nuevo T, que designaremos por T_1 y se tiene una nueva expresión del tiempo de espera. El valor de T, dependerá de los valores de a_1, y b_1; de aquí que pueda tenerse $T_1 < T$. Como se ve examinando la forma de T, $T_1 = T$ si y solamente si $a_1 = a$ y $b_1 = b$.

Si existen c canales de servicio y el número inicial de unidades es $i > c$, con el mismo tiempo de servicio para todos los canales, suponiendo $b/ca < 1$, para vaciar esta línea se requieren incrementos de longitud $a - b/c$, etc. Se pueden asignar prioridades a clientes que llegan con tiempos de llegada diferentes y estudiar el problema utilizando dos sistemas de prioridades, uno con prioridades absolutas (la unidad con prioridad inferior vuelve a la línea si llega otra con prioridad superior) y otro donde no haya prioridad absoluta. Así, aunque la estructura de este problema de colas es simple, pueden contrastarse varios conceptos.

1.3.2 Modelos de probabilidad.

Normalmente se carece de información previa sobre el tiempo de llegada de un cliente y la longitud del servicio. Para efectuar un análisis del funcionamiento de la línea, se hacen hipótesis basadas sobre un estudio de los hábitos y necesidades de los clientes. El hecho de que el empleo de razonamientos probabilísticos haga esto posible, es una indicación de la importancia del concepto de probabilidad.

En el ejemplo determinista, se sabía que los clientes llegarían exactamente en instantes conocidos y recibirían el servicio en tiempos exactos. Consideremos ahora los tiempos de servicio. En general no se sabe cuál pueda ser la duración del servicio a un cliente que llega. Pero a partir de su anterior comportamiento es posible observar la frecuencia relativa con que requiere servicio de una duración dada, es decir, la razón del número de veces que requiere esta duración del servicio al número total de veces que ha requerido algún servicio. O, hablando de otra manera, existe una probabilidad de que su tiempo de servicio no supere una duración dada. Esta probabilidad es la suma de todas las frecuencias de duraciones que varían hasta alcanzar la duración fijada. Así, por ejemplo, si los tiempos de servicio son discretos y sus duraciones son *1, 2, 3*.... unidades de tiempo y sus probabilidades (frecuencias relativas) son, respectivamente, p_1, p_2,... entonces si P_n es la probabilidad de que el cliente requiera un servicio de duración no mayor que n, se tiene:

$$P_n = p_1 + p_2 + \ldots + p_n$$

y de aquí : 1.3.2.1

$$p_n = P_n - P_{n-1}$$

Entre las cantidades útiles que se calculan a partir de estas probabilidades, están la media y la varianza. La media se obtiene multiplicando cada valor de *n* por su probabilidad y sumando para todo *n*. Así, por ejemplo, para un dado sin sesgos, la probabilidad de cada lado es *1/6* y el valor medio está dado por:

$$1*\frac{1}{6}+2*\frac{1}{6}+3*\frac{1}{6}+4*\frac{1}{6}+5*\frac{1}{6}+6*\frac{1}{6}=3.5$$

Por tanto, para un gran número de tiradas el valor medio de los números encontrados está en torno a *3.5*. En el caso de una línea de espera, un cálculo análogo da la extensión esperada de la línea. Además, se puede calcular la dispersión alrededor de este valor como una medida de la fluctuación en la extensión de línea. En el caso del dado, para obtener la dispersión calcularemos primero la varianza. La raíz cuadrada de la varianza es la desviación típica, que es una medida de la dispersión; es decir, cuando el número de tiradas se incrementa, se puede calcular la probabilidad de que el valor medio obtenido permanezca dentro de una desviación típica de 3.5. Se tiene para la varianza:

$$(1-3.5)^2 * 1_6 + (2-3.5)^2 * 1_6 + (3-3.5)^2 * 1/6 + (4-3.5)^2 * 1/6 + (5-3.5)^2 * 1/6 + (6-3.5)^2 * 1/6$$

que es aproximadamente igual a *2.92* y la desviación típica es $\sqrt{2.92}$. Las diferencias están elevadas al cuadrado porque lo que importa es la fluctuación alrededor de la media, prescindiendo de si los números caen por encima o por debajo del valor medio. Cada diferencia se pondera con la probabilidad de que se presente el valor considerado.

La media y la varianza son, respectivamente, el primer momento respecto al origen y el segundo momento respecto a la media. Pueden calcularse otros momentos respecto al origen y a la media. Así, por ejemplo el tercer momento respecto al origen que es la suma de los cubos de los valores, multiplicados cada uno por su probabilidad. Volviendo a p_n, debe tenerse $\sum_{n=0}^{\infty} p_n = 1$, ya que eso da la suma de las probabilidades de que ocurra cualquiera de las duraciones.

Si el tiempo es continuo, utilizaremos *F(x)*, en lugar de P_n, para designar la probabilidad de que la duración no sea mayor que un tiempo *x*. Esto es, *F(x) =prob. (X ≤ x)*, donde *X* es una variable aleatoria que puede tomar cualquiera de los valores posibles de la duración del tiempo de servicio. Si *f(x) dx* es la correspondiente función de frecuencia (es decir, la probabilidad de que se presente algún valor de *x* perteneciente a un intervalo de longitud *dx*), por ser *x* continua puede escribirse, análogamente al caso discreto, la distribución de probabilidad acumulativa:

$$F(x) = \int_0^x f(y)dy \qquad 1.3.2.2$$

y si se puede diferenciar $F(x)$, puede escribirse:

$$f(x) = \frac{dF(x)}{dx} \qquad 1.3.2.3$$

Si cada uno de los n clientes tienen diferentes costumbres, se tendrán varias variables aleatorias $X_1, X_2, \ldots, X_n$ con las correspondientes distribuciones acumulativas $F_1(x_1), \ldots, F_n(X_n)$. Si sus costumbres son independientes, en el sentido que se definirá con precisión más adelante, se puede escribir como función de distribución de probabilidad de su comportamiento conjunto.

$$F(x_1, \ldots x_n) = F_1(x_1)F_2(x_2)\ldots F_n(x_n) = \prod_{i=1}^{n} F_i(x_i) \qquad 1.3.2.4$$

Una vez conocida la función de distribución conjunta, pueden definirse, bajo este aspecto descrito por la distribución, todas las demás cosas.

Puede ser que las costumbres de un cliente sean distintas en diferentes instantes. Así, por ejemplo, puede solicitar, como en un restaurante, servicios más largos a mediodía que por la noche. En cada instante, la probabilidad de duración del servicio está descrita por una distribución de probabilidad que está ahora en función del tiempo. Las propias duraciones están descritas por una variable aleatoria X_t que depende del tiempo. Aquí se tiene:

$$Prob\ (X_t \leq x) = F(x; t) \qquad 1.3.2.5$$

como función de distribución de probabilidad en el instante t. Para cada valor de t, X_t tiene una distribución de probabilidad. Así, para un conjunto de valores de t, se tiene una familia de variables aleatorias. Como ya se ha indicado, una familia de variables aleatorias que dependen de un parámetro, define un proceso estocástico. Si no hay razón para discriminar entre las costumbres de diferentes

clientes en el mismo instante, el mismo proceso estocástico puede utilizarse para describir los tiempos de todos aquellos que demandan servicio. Por el contrario, cada cliente puede tener un proceso estocástico $X(k, t)$, que describe su identidad (k-ésimo para entrar en el servicio) y el instante de entrada al servicio.

Si X fuese, por ejemplo, una variable aleatoria que representa distintas condiciones meteorológicas (con una probabilidad de que se presente cada valor), en diferentes instantes y lugares, tendríamos una variable estocástica dependiente de cuatro parámetros: el tiempo y las tres coordenadas en el espacio. En la teoría de colas, el tiempo es, corrientemente, el único parámetro. Un ejemplo de proceso estocástico es el proceso de Poisson. Que está dado por:

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \qquad 1.3.2.6$$

donde, según veremos más adelante que $P_n(t)$ es la probabilidad de que lleguen n clientes en un intervalo de longitud t. Investigaremos varias propiedades de este útil proceso.

1.3.3 El Proceso de Poisson.

El modelo que aquí se desarrolla está relacionado con la distribución de llegadas bajo hipótesis que se dan con frecuencia en la práctica. El desarrollo examina propiedades del proceso de Poisson. Haciendo hipótesis explícitas que cumplen esas propiedades, se deduce el proceso de Poisson como una distribución de llegada.

A partir de $(\lambda t)^n e^{-\lambda t} / n!$, se observa que la probabilidad de que no haya ninguna llegada, durante el intervalo de tiempo t, es $e^{-\lambda t}$ y la de que haya una sola llegada es $\lambda e^{-\lambda t}$; por tanto, la probabilidad de más de una llegada es:

$$1-\left(e^{-\lambda t}+\lambda e^{-\lambda t}\right)=1-\left\{\left(1-\lambda t+\frac{(\lambda t)^2}{2!}-\cdots\right)+\lambda t\left(1-\lambda t+\frac{(\lambda t)^2}{2!}-\cdots\right)\right\}=\frac{(\lambda t)^2}{2}+\cdots=O\left(t^2\right) \quad 1.3.3.1$$

función que tiene el mismo orden que t^2

Si t es pequeño, los términos en t^2 son despreciables comparados con los que no contienen t o con la Primera potencia de t. Por tanto para t pequeño, la probabilidad de más de una llegada es despreciable. Esta es una propiedad muy deseable en muchas aplicaciones prácticas, que hace útil el proceso de Poisson. La probabilidad de al menos una llegada durante t está dada por:

$$1-e^{-\lambda t}=\lambda t+O\left(t^2\right) \quad 1.3.3.2$$

La probabilidad de que no haya llegadas es $e^{-\lambda t}=1-\lambda t+O\left(t^2\right)$. En ambas expresiones se puede despreciar el último término de la derecha si solo se consideran valores pequeños de t.

Aun a riesgo de ser repetitivo, pero de acuerdo con los métodos clásicos, supongamos que se verifican estas propiedades, o sea que la probabilidad de una única llegada durante un intervalo pequeño de tiempo Δt es $\lambda\Delta t$, y la de más de una llegada durante Δt es despreciable; podemos deducir entonces la distribución de Poisson que, naturalmente, tiene estas propiedades.

Sea $P_n(t)$ la probabilidad de que hayan llegado n unidades en el tiempo t. Obsérvese en primer lugar que $0 \le P_n(t) \le 1$ puesto que las $P_n(t)$ son probabilidades. Además $\sum_{n=0}^{\infty} P_n(t) = 1$, pues ninguna o alguna unidad debe haber llegado durante el tiempo t. Para ver qué ocurre durante el siguiente intervalo pequeño de tiempo Δt, escribimos:

$$\begin{aligned} P_0(t+\Delta t) &= P_0(t)(1-\lambda\Delta t) \\ P_n(t+\Delta t) &= P_n(t)(1-\lambda\Delta t) + P_{n-1}(t)\lambda\Delta t, \qquad n \ge 1 \end{aligned} \qquad 1.3.3.3$$

La primera ecuación da la probabilidad de que no haya ninguna llegada en el tiempo $t + \Delta t$. Esta probabilidad puede relacionarse con el estado del sistema en el instante t. Así, por la ley de las probabilidades compuestas de dos sucesos que se presentan independientemente (realizar el producto de las probabilidades de estos sucesos) es, igual a la probabilidad de que no ocurra ninguna llegada durante el tiempo t multiplicada por la probabilidad de que no haya ninguna llegada durante Δt. Para los casos $n \ge 1$ esta propiedad, tener el mismo número de llegadas durante t y ninguna durante Δt, también permanece, pero además puede haber habido $n-1$ llegadas durante el tiempo t seguidas de una llegada adicional durante Δt. El producto de estas cantidades engendra el segundo término de la derecha. No se ha mencionado la posibilidad de más de una llegada durante el pequeño intervalo Δt por ser despreciable y puede prescindirse de ella en lo que sigue.

Multiplicando, pasando $P_n(t)$ al primer miembro y dividiendo por Δt, el sistema de ecuaciones se transforma en:

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t), \qquad n \ge 1 \qquad 1.3.3.4$$

Si se toman límites cuando $\Delta t \rightarrow 0$, el primer miembro es, por definición, la derivada $P_n'(t) = dP_n(t)/dt$ y las ecuaciones son:

$$\begin{aligned} P_0'(t) &= -\lambda P_0(t) \\ P_n'(t) &= -\lambda P_n(t) + P_{n-1}(t), \qquad n \geq 1 \end{aligned} \qquad 1.3.3.5$$

Son ecuaciones diferenciales lineales con respecto a t y ecuaciones en diferencias de primer orden lineales con respecto a n, generalmente llamadas ecuaciones diferenciales en diferencias.

Se pueden resolver convenientemente estas ecuaciones utilizando una función generatriz. Se define tal función por:

$$P(z,t) \equiv \sum_{n=0}^{\infty} P_n(t) z^n = P_0(t) + P_1(t) z + P_2(t) z^2 + \cdots \qquad 1.3.3.6$$

Puede verse que las $P_n(t)$ se obtienen derivando, $P(z, t)$ n veces con respecto a z, dividiendo entonces por $n!$ y haciendo $z = 0$. Así, si $P(z,t)$ es conocida, $P_n(t)$ puede determinarse fácilmente de esta forma.

El origen de tiempos en un estudio específico puede tornarse en cualquier punto, aun después de que haya habido algunas llegadas. Por tanto, puede ser que para $t=0$, i unidades hayan llegado. En este caso, $P_n(0)$ es cero si $n \neq i$ y la unidad si $n = i$. Así,

$$P(z,0) = \sum_{n=0}^{\infty} P_n(0) z^n = P_i(0) z^i = z^i \qquad 1.3.3.7$$

Obsérvese también que $P(1, t) = 1$ y

$$\frac{\partial P(z,t)}{\partial t} = \frac{\partial}{\partial t} \sum_{n=0}^{\infty} P_n(t) z^n = \sum_{n=0}^{\infty} P_n'(t) z^n \qquad 1.3.3.8$$

Se utilizan derivadas parciales. por tenerse dos variables z y t.

Ahora si multiplicamos el sistema de ecuaciones diferenciales en diferencias para $n \geq 1$ por z^n y la primera ecuación por z^0 y sumamos para todo n se tiene como suma de los primeros miembros $\partial P(z,t)/\partial t$ y como suma de los primeros términos de los segundos miembros $-\lambda P(z,t)$.

Los segundos términos de la derecha sumados para todo n dan:

$$\sum_{n=1}^{\infty} \lambda P_{n-1}(t) z^n = \lambda P_0(t) z + \lambda P_1(t) z^2 + \cdots \qquad 1.3.3.9$$

Si se saca factor común λz en el segundo miembro, este puede escribirse como $\lambda z P(z,t)$. Así, el sistema se reduce a una ecuación diferencial lineal en la función generatriz dada por:

$$\frac{\partial P(z,t)}{\partial t} - \lambda(z-1)P(z,t) = 0 \qquad 1.3.3.10$$

Esta ecuación es muy similar a los tipos más elementales de ecuaciones diferenciales ordinarias que se estudian en los cursos de cálculo elemental. Su solución, tratando a z como a una constante por ser independiente de t, está dada por:

$$P(z,t) = Ce^{\lambda(z-1)t} \qquad 1.3.3.11$$

Esto puede comprobarse por sustitución, pudiendo C depender de z.

Supongamos que para $t = 0$ no ha habido ninguna llegada; entonces $P(z, 0) = 1$, ya que $i = 0$. Así, $C = 1$ y se tiene:

$$P(z,t) = Ce^{\lambda(z-1)t} \qquad 1.3.3.12$$

Pero, como ya se ha indicado, $P_n(t)$ puede ser obtenido por derivación; por tanto,

$$P_n(t) = \frac{1}{n!} \frac{\partial^n P(z,t)}{\partial z^n}\Big|_{z=0} \qquad 1.3.3.13$$

Así:

$$P_0(t) = e^{-\lambda t}$$
$$P_1(t) = \lambda t e^{-\lambda t} \qquad 1.3.3.14$$

y

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \qquad 1.3.3.15$$

que da el deseado proceso de Poisson. Quede claro que t es la longitud del intervalo de tiempo en que ocurren los sucesos y no tiempo absoluto.

El método anterior de deducción es típico para muchos problemas de colas. Las ecuaciones pueden ser más complicadas, pero el procedimiento básico es el mismo. No tendrá dificultad en escribir para $i \neq 0$ y $n \geq i$:

$$P_n(t) = \frac{(\lambda t)^{n-i} e^{-\lambda t}}{(n-i)!} \qquad 1.3.3.18$$

Ya que en este caso $P(z, 0) = z^i$, C tiene el valor z^i y :

$$P(z,t) = z^i \sum_{n=i}^{\infty} P_n(t) z^{n-i} \qquad 1.3.3.19$$

Calcular de $P_n(t)$ es muy útil para obtener medidas de efectividad para estudiar las líneas de espera, no siempre es posible e incluso, deseable estudiar colas en términos de $P_n(t)$.

Habiendo obtenido $P_n(t)$, puede desearse calcular el número medio de unidades que llegan durante el tiempo t, que es:

$$L = \sum_{n=0}^{\infty} nP_n(t) \qquad 1.3.3.20$$

que para el proceso de Poisson es igual a λt. La varianza:

$$\sum_{n=0}^{\infty} (n-L)^2 P_n(t) \qquad 1.3.3.21$$

es también igual a λt.

Los intervalos de tiempo entre llegadas de un proceso de Poisson con parámetro λ tienen la distribución de frecuencia exponencial $\lambda e^{-\lambda t}$.

Para ver esto, obsérvese que, si ha ocurrido ahora una llegada, el tiempo hasta la próxima llegada es menor que t si, y solamente si, hay una o más llegadas en este intervalo. La probabilidad de una o más llegadas y, por tanto, la probabilidad de que el tiempo entre llegadas sea menor o igual que t, es:

$$\sum_{n=1}^{\infty} \frac{(\lambda t)^n e^{-\lambda t}}{n!} = 1 - e^{-\lambda t} \qquad 1.3.3.22$$

La cantidad de la derecha es una distribución acumulativa de la que por diferenciación, se obtiene la función de densidad exponencial $\lambda e^{-\lambda t}$ como distribución de los tiempos entre llegadas. Por tanto, cuando los tiempos de llegadas a una línea de espera vienen dados por un proceso de Poisson, los tiempos entre llegadas tienen una distribución exponencial afín. Recíprocamente, si, por ejemplo, los tiempos de servicio tienen una distribución exponencial, los clientes son admitidos al servicio según un proceso de Poisson.

De nuevo, si $\mu\Delta t$ es la probabilidad de completar el servicio durante Δt, y, por tanto, $1-\mu\Delta t$ es la probabilidad de que no acabe el servicio durante Δt, y si $P(t)$ es la probabilidad de que el servicio no haya terminado durante el tiempo t, resulta.:

$$P(t+\Delta t)=(1-\mu\Delta t)P(t) \qquad 1.3.3.23$$

Así se tiene:

$$\frac{dP(t)}{dt}=\lim_{\Delta t\to 0}\frac{P(t+\Delta t)-P(t)}{\Delta t}=-\mu P(t) \qquad 1.3.3.24$$

Esta ecuación diferencial admite la solución:

$$P(t)=ce^{-\mu t} \qquad 1.3.3.25$$

puesto que el servicio terminará más pronto o más tarde, debemos, tener:

$$\int_0^{\infty} ce^{-\mu t}dt=1 \qquad 1.3.3.26$$

Por tanto, $c=\mu$. Por consiguiente, el servicio está descrito, por la distribución exponencial negativa con densidad $\mu e^{-\mu t}$. Obsérvese que la distribución exponencial tiene la propiedad de "ser distraída": si, por ejemplo, es aplicada a los, tiempos de servicio, la distribución exponencial se aplica para el tiempo de completar un servicio en cierto tiempo lo mismo que si fuese el comienzo del servicio.

En muchos problemas, la dependencia del tiempo no es la situación típica que se estudia. Después de un largo período de operación, el sistema adquiere características de comportamiento que, aunque sean describibles en términos de probabilidad, no dependen del tiempo. Así, cuando $t \to \infty$, puede esperarse que ocurra esta situación en muchos sistemas. La cuestión de si existen tales $p_n = \lim_{t\to\infty} P_n(t)$ es asunto del análisis ergódico donde de hecho se buscan las probabilidades p_n que describen el estado estacionario o en equilibrio. Pero nótese también que el que las probabilidades $P_n(t)$ no cambien con el tiempo es la definición de estado estacionario.

La variación de $P_n(t)$ con respecto a t, está descrita por la derivada $P_n'(t)$ y la condición de estado estacionario requiere que $P_n'(t) = 0$.

Por tanto, se tienen esencialmente dos formas de obtener las probabilidades del estado estacionario:

1) a partir de $P_n'(t) = 0$, que da probabilidades p_n independientes de t, y

2) $\lim_{t\to\infty} P_n(t)$, que también da las p_n independientes de t.

1.3.4 Modelo de Erlang.

Si suponemos que una operación empieza sin ninguna unidad esperando en línea, las ecuaciones siguientes describen una cola ordenada, primero en llegar primer en ser servido, con canal único, ingreso de Poisson con parámetro λ y tiempo de duración exponencial con parámetro μ:

$$\begin{aligned} P_n(t+\Delta t) &= P_n(t)[1-(\lambda+\mu)\Delta t] + P_{n-1}(t)\lambda\Delta t + P_{n+1}(t)\mu\Delta t && n \geq 1 \\ P_0(t+\Delta t) &= P_0(t)(1-\lambda\Delta t) + P_1(t)\mu\Delta t && n = 0 \end{aligned} \qquad 1.3.4.1$$

Estas ecuaciones expresan que la probabilidad de n unidades en el sistema en el instante $t + \Delta t$ es igual a la probabilidad de n unidades en el instante t multiplicada por la probabilidad de que no haya llegadas ni partidas, más la probabilidad de $n-1$ unidades en el sistema en el instante t multiplicada por la probabilidad de una llegada y ninguna partida, más la probabilidad de $n+1$ unidades en el sistema en el instante t multiplicada por la probabilidad de una única partida y ninguna llegada.

Obsérvese que la probabilidad de que no haya ni llegadas ni partidas está dada por $(1 - \lambda\Delta t)(1 - \mu\Delta t)$. El término en $(\Delta t)^2$ desaparece al formar la ecuación diferencial. Por ello, se puede utilizar $1 - (\lambda + \mu)\Delta t$. Para los restantes dos términos, obsérvese que también $\lambda\Delta t(1 - \mu\Delta t) \approx \lambda\Delta t$ y $\mu\Delta t(1 - \lambda\Delta t) \approx \mu\Delta t$.

Por transposición y aplicando el límite con respecto a Δt, estas ecuaciones se convierten en:

$$\begin{aligned} \frac{dP_n(t)}{dt} &= -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) && n \geq 1 \\ \frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \mu P_1(t) && n = 0 \end{aligned} \qquad 1.3.4.2$$

Como ya se ha indicado, la solución de estado estacionario, independiente del tiempo, se obtiene, o bien resolviendo las ecuaciones del estado de transición, dependientes del tiempo, dadas anteriormente y haciendo $t \to \infty$ en la solución, o bien haciendo iguales a cero las derivadas con respecto al tiempo y resolviendo las ecuaciones, del estado estacionario que resultan. Las soluciones para el estado de transición son particularmente útiles cuando la intensidad de tráfico o factor de utilización $\rho = \lambda/\mu \geq 1$, puesto que en este caso no se alcanza el estado estacionario. Aquí se deducirá una expresión para el número esperado de unidades en espera y para el tiempo esperado de espera, en estado estacionario con $\lambda/\mu < 1$. Si $\lambda/\mu \geq 1$, el número de unidades esperando podría ser infinito. Compárense estas condiciones con las de cola determinista considerada anteriormente.

Igualando a cero las derivadas y eliminando el tiempo en las ecuaciones de arriba, resulta, después de transponer términos:

$$\begin{aligned} (\lambda+\mu)p_n &= \lambda p_{n-1} + \mu p_{n+1} && n \geq 1 \\ \lambda p_0 &= \mu p_1 && n = 0 \end{aligned} \qquad 1.3.4.3$$

Sea $\rho = \lambda/\mu$, entonces estas ecuaciones, se convierten en

$$\begin{aligned} (1+\rho)p_n &= p_{n+1} + \rho p_{n-1} && n \geq 1 \\ p_1 &= \rho p_0 && n = 0 \end{aligned} \qquad 1.3.4.4$$

Sea $n=1$ en la primera ecuación. Entonces $(1+\rho)\, p_1 = p_2 + \rho p_0$. Sustituyendo p_1 por su valor dado en la segunda ecuación, se tiene $p_2 = \rho^2 p_0$. La repetición de este proceso lleva a $p_n = \rho^n p_0$. Ahora bien: $\sum_{n=0}^{\infty} p_n = 1$, ya que la suma da la probabilidad total de que no haya ninguna unidad en el sistema, de que haya una, de que haya dos, etc.

Esta probabilidad total debe ser la de la certeza, puesto que se cuentan todos los posibles estados del sistema. Por tanto,

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

$$o$$

$$\sum_{n=0}^{\infty} p_0 \rho^n = p_0 \sum_{n=0}^{\infty} \rho^n = \frac{p_0}{1-\rho} = 1 \qquad 1.3.4.5$$

$$o$$

$$p_0 = 1 - \rho$$

Por tanto,

$$p_n = \rho^n (1-\rho) \qquad 1.3.4.6$$

que es un tipo de distribución de probabilidad que se conoce como distribución geométrica.

El número esperado de unidades dentro del sistema está dado ahora por:

$$L = \sum_{n=0}^{\infty} n p_n = (1-\rho) \sum_{n=0}^{\infty} n \rho^n = \frac{\rho}{1-\rho} \qquad 1.3.4.7$$

como puede fácilmente comprobarse. Nótese que L es un valor esperado y, naturalmente, ocurrirán fluctuaciones en el número de unidades que esperan. Esto puede verse mejor calculando la varianza:

$$\sum_{n=0}^{\infty} (n-L)^2 p_n = \sum_{n=0}^{\infty} n^2 p_n - \left(\frac{\rho}{1-\rho} \right)^0 \qquad 1.3.4.8$$

Pero

$$\sum_{n=0}^{\infty} n^2 p_n = (1-\rho)\sum_{n=0}^{\infty} n^2 \rho^n = (1-\rho)\rho \frac{d}{d\rho}\rho \frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n = \frac{\rho}{1-\rho} + \frac{2\rho^2}{(1-\rho)^2} \qquad 1.3.4.9$$

Por tanto, la varianza está dada por:

$$\frac{\rho}{1-\rho} + \frac{\rho^2}{(1-\rho)^2} = L + L^2 \qquad 1.3.4.10$$

El número esperado de unidades en la línea está dado por:

$$L_q = \sum_{n=1}^{\infty}(n-1)p_n = L - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho} \qquad 1.3.4.11$$

Obsérvese que hay momentos en que el canal no está ocupado. El tiempo medio esperando en la línea, W_q, es igual a L_q/λ o L/μ.

Una práctica útil para descubrir posibles incorrecciones en una fórmula es el análisis de dimensiones. Así, por ejemplo, en la última igualdad de arriba, tanto la expresión de la izquierda como la de la derecha carecen de dimensiones.

1.3.5 LA FÓRMULA DE POLLACZEK - KHINTEHINE.

Supongamos ahora que las llegadas ocurren al azar, según un proceso de Poisson a razón de λ por unidad de tiempo, a una línea de espera, en equilibrio estadístico, ante una única instalación de servicio. Supongamos también que son servidas, según una distribución arbitraria de tiempos de servicio, a razón de μ por unidad de tiempo, primera en llegar, primera servida. Como en el caso determinista, supongamos que $\lambda/\mu < 1$. Supongamos que un cliente que parte deja q en la línea, incluyendo al que está siendo servido, cuyo tiempo de servicio es t. Sean r los clientes que llegan durante este tiempo t. Si el próximo cliente que parte deja tras de sí a q' clientes, se pueden relacionar q y q' como sigue:

$$q' = max(q-1,0) + r = q - 1 + \delta + r \qquad 1.3.5.1$$

donde:

$$\delta(q) = \begin{cases} 0 & \text{si} \quad q > 0 \\ 1 & \text{si} \quad q = 0 \end{cases} \qquad 1.3.5.2$$

Obsérvese que introduciendo δ se ha evitado utilizar la expresión máximo.

Supongamos que existen valores para la cola en equilibrio de los momentos primero y segundo $E[q]$ y $E[q']$. Nótese que q es tratada como una variable aleatoria. Por definición, $\delta^2 = \delta$ y $q(1-\delta) = q$. También, $E[q] = E[q']$ y $E[q^2] = E[q'^2]$, ya que se supone que q y q' tienen la misma distribución en equilibrio. Obsérvese que, por estar en equilibrio el sistema, no existe diferencia entre los tipos de cola que deja tras de sí cualquier cliente; es decir, todos tienen la misma distribución de probabilidad, independiente del tiempo.

Así, tomando el valor esperado de las diferentes variables en la ecuación, se tiene:

$$E[q'] E[q] - E[l] + E[\delta] + E[r] \qquad 1.3.5.3$$

de donde:

$$E[\delta] = 1 - E[r] \qquad 1.3.5.4$$

Pero durante un tiempo de servicio de longitud t, se tiene:

$$E[r] = \sum_{r=0}^{\infty} r \frac{(\lambda t)^r}{r!} e^{-\lambda t} = \lambda t$$

$$E[r^2] = \sum_{r=0}^{\infty} r^2 \frac{(\lambda t)^r}{r!} e^{-\lambda t} = (\lambda t)^2 + \lambda t \qquad 1.3.5.5$$

Por tanto, al tomar esperanzas con respecto al tiempo de servicio t, se tiene $E[r] = \lambda/\mu \equiv \rho$. Así, por ejemplo, si la distribución del servicio es exponencial,

$$E[r] = \mu \int_0^{\infty} (\lambda t) e^{-\mu t} dt = \frac{\lambda}{\mu} \equiv \rho \qquad 1.3.5.6$$

puesto que el valor medio de la distribución del tiempo de servicio es $1/\mu$.

Obsérvese que $E[r]$ es un número que no es afectado al tomar medias. Esto da $E[\delta] = 1 - \rho$. Ahora bien: la probabilidad de r llegadas es independiente de q, longitud de la cola, y de δ, cuyo valor se supone dependiente sólo de q, que es independiente de r. En consecuencia, si se toma el valor esperado sobre las variables r, q y δ, se debe tomar el valor esperado de r y de q por separado dondequiera que encontremos su producto. Esto es cierto también para r y δ.

También promediando r^2 sobre el tiempo:

$$E[r^2] = \lambda^2 \operatorname{var}(t) + \rho^2 + \rho \qquad 1.3.5.7$$

Elevando al cuadrado la ecuación que relaciona q y q' y utilizando el que $\delta^2 = \delta$, $\delta q \equiv 0$,

$$q'^2 = q^2 - 2q(1-r) + (r-1)^2 + \delta(2r-1) \qquad 1.3.5.8$$

Por tanto, a causa del equilibrio,

$$0 = E[q'^2] - E[q^2] = 2E[q]E[r-1] + E[(r-1)^2] + E[\delta]E[2r-1] \qquad 1.3.5.9$$

simplificando y teniendo en cuenta las anteriores relaciones, se tiene la fórmula de Pollaczek - Khintchine:

$$E[q] = \frac{E[(r-1)^2] + E[\delta]E[2r-1]}{2E[1-r]}$$

$$= \frac{E[r^2] - 2E[r] + 1 + E[\delta](2E[r]-1)}{2(1-E[r])}$$

$$= \frac{\lambda^2 \operatorname{var}(t) + \rho^2 + \rho - 2\rho + 1 + (1-\rho)(2\rho-1)}{2(1-\rho)} \qquad 1.3.5.10$$

$$= \rho + \frac{\rho^2 + \lambda^2 \operatorname{var}(t)}{2(1-\rho)}$$

Por tanto, una vez conocida la varianza del tiempo de servicio *t* a partir de su distribución, el número medio de unidades en el sistema está determinado. Es importante observar que la anterior media se toma en instantes exactos que siguen a las partidas y no es el valor medio, en el tiempo del número de unidades en el sistema. Si $E_t(q)$ es la media en el tiempo, todo lo que se puede saber sin otro razonamiento es que $E[q] \leq E_t(q) < E[q] + 1$.

Nótese el hecho, que es válido en general aun si se tienen varios canales en paralelo, de que el número medio de unidades en el sistema es igual a la suma del número medio de canales ocupados (en este caso es ρ, intensidad de tráfico) más el número medio de unidades en la línea.

Para obtener ahora el tiempo medio de espera, razonaremos como sigue: Si designamos por $E[w]$ el tiempo medio esperando en la cola, excluido el servicio. $\lambda[E(w) + 1/\mu]$ es el número esperado de llegadas durante el tiempo total de espera más el tiempo de servicio de un cliente, es decir, durante su estancia dentro del sistema. Pero este es exactamente el número dentro del sistema inmediatamente después de su partida, designado por $E[q]$. Por tanto,

$$W_q \equiv E[w] = \frac{\rho^2 + \lambda^2 \operatorname{var}(t)}{2\lambda(1-\rho)} = \frac{L_q}{\lambda} \qquad 1.3.5.11$$

fórmula a la que podría darse el nombre de Pollaczek - Khintchine.

Una medida útil que introduciremos ahora es la razón del tiempo medio de espera de un cliente a su tiempo de servicio. Naturalmente puede no ser deseable esperar mucho tiempo cuando se requiere un servicio de corta duración. Para esta medida se tiene:

$$\mu E[w] = \frac{\rho}{2(1-\rho)}\left(1 + C_t^2\right) \qquad 1.3.5.12$$

donde C_t es el coeficiente de variación del tiempo de servicio, definido por:

$$C_t^2 = \frac{\text{var}(t)}{E^2[t]} = \frac{\text{var}(t)}{1/\mu^2} \qquad 1.3.5.13$$

Evidentemente, la medida de arriba $\mu E[w]$ depende de ρ y de C_t^2.

Cualquier variación en la tasa de ingreso λ o en la de servicio μ se refleja en su razón ρ. Por tanto, para contrastar la eficacia, o sea la sensibilidad a los cambios de valor de ρ a partir de un valor medio $\bar{\rho}$ dado (llamado valor presente) de ρ se desarrolla el segundo miembro, en serie de Taylor alrededor del punto $\bar{\rho}$ para variaciones pequeñas en torno, a él. Se tiene:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^n(a)}{n!}(x-a)^n + \cdots \qquad 1.3.5.14$$

donde el número de comillas indica el orden de la derivada que debe tomarse.

Aplicando esta fórmula al segundo miembro de ***1.3.5.14***, se tiene:

$$\mu E[w] = \frac{\bar{\rho}}{2(1-\bar{\rho})}(1+C_t^2) + \frac{1+C_t^2}{2(1-\bar{\rho})^2}(\rho-\rho^2) + \frac{1+C_t^2}{2!(1-\bar{\rho})^3}(\rho-\bar{\rho})^2 + \cdots \qquad 1.3.5.15$$

Si tomamos, medias con respecto a ρ, entonces el segundo término se hace cero, ya que $E[\rho-\bar{\rho}] = \bar{\rho}-\bar{\rho}$, y despreciando términos de grado superior al segundo, tenemos:

$$\frac{\bar{\rho}}{2(1-\bar{\rho})}(1+C_t^2)\left[1+\frac{\sigma_\rho^2}{\bar{\rho}(1-\bar{\rho})^2}\right] \qquad 1.3.5.16$$

donde:

$$\sigma^2 = E\left[\left(\rho - \overline{\rho}\right)^2\right]$$

Este análisis es de utilidad cuando puede considerarse una operación de colas como un conjunto de operaciones cada una en estado estacionario y con sus parámetros. Entonces se puede determinar la amplitud del recorrido de fluctuación permitida en los valores del parámetro para tener una variación tolerable en la razón del tiempo de espera al tiempo de servicio, antes que buscar una nueva solución en estado estacionario.

1.4 MODELOS DE LLEGADAS

La forma en que llegan las unidades al sistema es aleatoria si no puede predecirse exactamente cuándo llegará cada una. El tiempo de llegada es una variable aleatoria que puede describirse matemáticamente con una distribución de probabilidad. La que vemos, depende de la forma de las llegadas, como lo muestran los datos observados y la naturaleza de las operaciones. Una de las distribuciones que se encuentran más comúnmente en los problemas de líneas de espera, es la distribución Poisson, que se emplea en la modelación de problemas de líneas de espera concerniente a los arribos aleatorios de unidades que demandan el servicio.

La distribución Poisson es una distribución de probabilidad discreta que determina la probabilidad del número de llegadas en un tiempo dado, donde el proceso de llegadas es independiente de lo que haya ocurrido en las observaciones precedentes.

La suposición de una variable aleatoria de tipo Poisson indica que las llegadas ocurren en forma aleatoria, y su tasa se representara con la constate λ. Esta constante representa el número de llegadas por unidad de tiempo, mientras que $1/\lambda$ es la longitud del intervalo de tiempo entre dos llegadas consecutivas (t y $t+\Delta t$).

La distribución Poisson (curva con parámetro λT), donde n es el número de llegadas dentro del intervalo **T**, el parámetro λ es la razón del número de llegadas por unidad de tiempo, y **T** es el tiempo total que se considera, que es:

$$f(n,\lambda,T)=\frac{(\lambda T)^n e^{-(\lambda T)}}{n!} \qquad 1.4.1$$

Utilizando la ecuación 1.4.1 y asignando el valor de 1 a **T**, la ecuación se convierte en:

$$f(n,\lambda)=\frac{\lambda^n e^{-\lambda}}{n!} \qquad 1.4.2$$

Si asignamos 1 a n tenemos la expresión de la distribución exponencial (función de densidad de **T**), es:

$$f(T,\lambda)=\lambda e^{-\lambda T} \qquad 1.4.3$$

En problemas prácticos de líneas de espera se supone que la distribución de arribos es una distribución Poisson, se puede probar la suposición de que las llegadas siguen una distribución Poisson, lo que se podría lograr buscando un intervalo fijo de tiempo y contando el número de unidades que llegan en ese intervalo, y después realizar una prueba de bondad de ajuste con la ji cuadrada. Esta está basada en una comparación entre datos observados y datos teóricos, donde los teóricos se obtienen a partir de la distribución teórica que se prueba.

1.4.1 CLASIFICACIÓN DE KENDALL Y LEE

En 1953 Kendall y Lee propusieron en sistema de clasificación de las líneas de espera, ampliamente utilizado en la actualidad. Esta clasificación considera seis de las características de la estructura de los modelos de líneas de espera, expresándolas en el formato ***(a/b/c)(d/e/f)***, donde :

A	Distribución de probabilidad del tiempo entre llegadas de las transacciones.	
B	Distribución de probabilidad del tiempo de servicio	
	Los símbolos utilizados en estos dos primeros campos son:	
	D	Constante
	E_k	Distribución Erlang con parámetro k
	G	Cualquier tipo de distribución
	GI	Distribución general independiente
	H	Distribución hiperexponencial
	M	Distribución exponencial
C	Número de servidores	
D	Orden de atención a los clientes	
	Los símbolos utilizados en este campo son:	
	FIFO	Primeras entradas, primeros en ser servidos
	LIFO	Últimas entradas, primeros en ser servidos
	RANDOM	Aleatorio
	PR	En base a prioridades
	GD	En forma general
E	Número máximo de clientes que soporta el sistema en un mismo instante de tiempo	
F	Número de clientes potenciales del sistema de líneas de espera.	

Por ejemplo, un modelo (M/D/3)(FIFO/20/20) representa la clasificación de un sistema donde existen 3 servidores en paralelo atendiendo de acuerdo con un orden de primeras entradas, primeras salidas, con un tiempo de servicio constante . El sistema tiene solo 20 clientes potenciales, los cuales podrían encontrarse dentro del sistema en un mismo instante. El tiempo entre llegadas de los clientes siguen una distribución exponencial y, en caso de llegar y encontrar todos los servidores ocupados, pasan a incorporarse a una fila común.

1.4.2 MODELO DE TIEMPO DE SERVICIO

El tiempo de servicio es el intervalo entre el principio del servicio y su terminación. La tasa media de servicio (μ) es el número de clientes a los que se da servicio por unidad de tiempo, mientras que el promedio de tiempo de servicio ($1/\mu$) es el de las unidades de tiempo por cliente. El tiempo de servicio suministrado se da con una distribución exponencial (que muchos autores llaman distribución exponencial negativa), cuando el servicio dado a un cliente ocurre entre el tiempo t y $t+\Delta t$. Hay que notar que la distribución Poisson (que se aproxima a una distribución normal y esta sesgada a un lado) no puede aplicarse el servicio.

Ordinariamente hay algún tiempo ocioso de parte del encargado. La distribución Poisson se aplica a un intervalo de tiempo fijo de servicio continuo, pero nunca se podrá estar seguro de que esto ocurra en cualquier situación, y por esa razón se usa la distribución exponencial (negativa).

Substituyendo a μ por λ en la ecuación 1.4.1 y si n es el número de servicios potenciales que pueden suministrarse en el intervalo T, la fórmula de Poisson para tasa de servicio es la siguiente:

$$g(n,\mu,T)=\frac{(\mu T)^n e^{-(\mu T)}}{n!} \qquad 1.4.2.1$$

Remplazando una μ por λ en la ecuación 1.4.3, la probabilidad de que se complete el servicio de una unidad en el tiempo T para la distribución exponencial, es de:

$$g(T,\mu)=\mu e^{-\mu T} \qquad 1.4.2.2$$

A fin de comprobar la suposición de que los tiempos de servicio se distribuyen en forma exponencial, se puede realizar una prueba como la que se propuso con anterioridad.

1.4.3 LLEGADAS POISSON UNA SOLA LINEA CON SERVICIO EXPONENCIAL: M/M/1

Se considera que las entradas (clientes, tareas, etc.), llegan en forma aleatoria. La tasa de servicio es independiente del número de elementos en línea. Cuando tanto las llegadas como el tiempo de servicio son aleatorias, el problema se llama de líneas de espera, con una distribución de tipo Poisson para las llegadas, y distribución exponencial para el tiempo de servicio.

A fin de determinar las propiedades del sistema, es necesario encontrar una expresión para la probabilidad de que haya n unidades en el sistema en el tiempo t, o $P_n(t)$, y desarrollaremos para $P_n(t+\Delta t)$. Donde Δt es un pequeño intervalo de tiempo.

Para esto se observará que existen cuatro eventos mutuamente excluyentes y exhaustivos (sólo uno de los eventos puede y debe ocurrir), donde n es mayor que cero.

Hay que notar que cada uno de esos elementos es un evento compuesto, formado por otros tres eventos sencillos. (Tabla 1)

Tabla 1

Evento	Probabilidad de que haya n unidades en la línea en el tiempo t	Llegadas en el intervalo t a $t+\Delta t$	Unidades que recibieron servicio en el intervalo t a $t+\Delta t$	Unidades en la línea al tiempo $t+\Delta t$
1	P_n	0	0	n
2	P_{n+1}	0	1	n
3	P_{n-1}	1	0	n
4	P_n	1	1	n

Analicemos cada uno de los eventos:

P_n es la probabilidad de que hayan n elementos en el sistema al tiempo t, pero lo que nos interesa es la probabilidad P_n al tiempo $t+\Delta t$. Por tanto se analizará lo que puede pasar en ese pequeño intervalo de tiempo (Δt). Si al inicio existen n elementos en el sistema y al finalizar existen n en él, tuvo que suceder cualquiera de estos eventos:

- que no haya llegado nadie al sistema y que no haya salido nadie de él
- por el contrario que haya llegado un elemento y que se haya atendido a un elemento.

Estos dos eventos están representados en la Tabla 1 como los eventos 1 y 4 respectivamente.

Por otra parte el evento 2 nos indica que existen al inicio $n+1$ elementos en el sistema y para que al finalizar existan n elementos. En este caso, forzosamente tendría que salir un elemento del sistema y que no llegara ningún elemento al sistema. Finalmente tenemos el último evento, en el cual existen n-1 elementos en el sistema y al finalizar tenemos n. Esto sólo se podrá lograr cuando haya llegado un elemento al sistema y que no haya salido ninguno de él (evento 3).

Como sólo debe ocurrir uno de los cuatro eventos, podemos obtener $P_n(t+\Delta t)$, donde n es mayor que cero, sumando las probabilidades de cada uno de los eventos separados y compuestos que se encuentran en la Tabla 1, para lo que se tendrá:

$$P_n(t+\Delta t) = P_n(t)(1-\lambda\Delta t-\mu\Delta t) + P_{n+1}(t)(\mu\Delta t) + P_{n-1}(t)(\lambda\Delta t) \qquad 1.4.3.1$$

Si expandemos el lado derecho de la ecuación 1.4.3.1 y restamos $P_n(t)$ a ambos lados de la ecuación y finalmente se divide por Δt tenemos:

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \frac{(P_{n+1}(t)\mu t + P_{n-1}(t)\lambda - P_n(t)(\lambda+\mu))\Delta t}{\Delta t} \qquad 1.4.3.2$$

Por definición, la derivada de P_n con respecto a t, es:

$$\frac{dP_n(t)}{dt} = \lim_{\Delta t \to \infty} \frac{P_n(t+\Delta t) - P(t)}{\Delta t} \qquad 1.4.3.3$$

Cuando Δt se aproxima a cero, puede darse la siguiente ecuación diferencial, para expresar la relación entre las probabilidades (P_n, P_{n+1}, y P_{n-1} en el tiempo t) y la tasa media de llegadas (λ) y la tasa media de servicio (μ) son:

$$\frac{dP_n(t)}{dt} = P_{n+1}(t)\mu + P_{n-1}(t)\lambda - P_n(t)(\lambda + \mu) \qquad 1.4.3.4$$

Ahora es necesario resolver para cuando n es igual con cero, sólo pueden ocurrir dos eventos mutuamente excluyentes y colectivamente exhaustivos, como lo podemos ver en la Tabla 2

Tabla 2

Evento	Probabilidad de que haya n unidades en la línea en el tiempo t	Llegadas en el intervalo t a $t+\Delta t$	Unidades que recibieron servicio en el intervalo t a $t+\Delta t$	Unidades en la línea al tiempo $t+\Delta t$
1	P_0	0	-	0
2	P_1	0	1	0

La probabilidad de que no haya unidades en la línea en el tiempo $t+\Delta t$, se obtiene sumando las probabilidades de cada uno de los eventos separados que se muestran en la Tabla 2, como sigue:

$$P_0(t+\Delta t) = P_0(t)(1-\lambda\Delta t) + P_1(t)(\mu\Delta t) \qquad 1.4.3.5$$

Cuando Δt se aproxima a cero, la ecuación diferencial que indica la relación entre las probabilidades (P_0 y P_1), la tasa media de llegadas (λ) y la tasa media de servicio (μ) son:

$$\frac{dP_0(t)}{dt} = P_1(t)\mu - P_0(t)\lambda \qquad 1.4.3.6$$

Las ecuaciones 1.4.3.4 y 1.4.3.6 son ecuaciones diferenciales con respecto a ***t*** y ecuaciones en diferencias de primer orden lineales con respecto a ***n***, generalmente llamadas ecuaciones diferenciales en diferencias. Proporcionan relaciones que incluyen la función de densidad de probabilidades $P_n(t)$ para todos lo valores de ***n***. Sin embargo, estas ecuaciones no nos permiten resolver la función de densidad de probabilidades. Nos interesa lo que ocurre cuando se estabiliza la línea de espera, o sea cuando llega al estado estacionario.

Para poder resolver estas ecuaciones hay que igualar con cero y realizaremos un cambio de variable entonces la ecuación 1.4.3.4 se convierte en:

$$\frac{dP_n(t)}{dt} = 0 \qquad 1.4.3.7$$

Y $P_n(t)$ cambiará por P_n en las siguientes ecuaciones.:

$$\lambda P_{n-1} + \mu P_{n+1} - (\lambda + \mu) P_n = 0 \qquad 1.4.3.8$$

$$\mu P_1 - \lambda P_0 = 0 \qquad 1.4.3.9$$

de la ecuación 1.4.3.9 se tiene que:

$$P_1 = \frac{\lambda}{\mu} P_0 \qquad 1.4.3.10$$

Donde P_1 es la probabilidad de que haya una unidad en la línea excluyendo la que está recibiendo servicio.

Con la ecuación 1.4.3.10 se obtienen las siguientes relaciones, substituyendo P_n con P_{n-1} precedente hasta llegar a P_0:

$$P_n = \frac{\lambda}{\mu} P_{n-1} \qquad \text{donde } n = 1,2,\cdots$$

$$P_2 = \frac{\lambda}{\mu} P_1 = \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} P_0\right) = \left(\frac{\lambda}{\mu}\right)^2 P_0 \qquad \text{donde } n = 2$$

$$P_3 = \frac{\lambda}{\mu} P_2 = \left(\frac{\lambda}{\mu}\right)^3 P_0 \qquad \text{donde } n = 3$$

Ahora es evidente que está apareciendo cierta forma:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 \qquad n \geq 0 \qquad \text{1.4.3.11}$$

La ecuación 1.4.3.11 es casi lo que se ha buscado, en términos de una expresión de la función de densidad de probabilidad. La única dificultad es que la expresión contiene P_0, y que n en este punto es una incógnita. Se conoce una probabilidad de una función de densidad de probabilidades, que puede permitir determinar P_0. Si x es una variable aleatoria con una función de probabilidad, $P(x)$ puede expresarse así:

$$\sum P(x) = 1$$

Esto está de acuerdo con el concepto fundamental de que la suma de todas las probabilidades es igual a 1.

De esa propiedad de una función de distribución de probabilidad, se sigue que 1 es igual a:

$$P_0 + P_1 + P_2 + \cdots , \qquad \text{ó}$$
$$\sum_{n=0}^{\infty} P_n = 1$$

Substituyendo la ecuación 1.4.3.11 en la suma anterior por el término P_n, la nueva expresión es la siguiente:

$$P_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1 \qquad 1.4.3.12$$

En todas las situaciones en que la línea de espera no crece sin límites, λ/μ será menor que uno. En esos casos, los términos de la suma de la ecuación 1.4.3.12 forman una serie geométrica convergente. La suma de una serie infinita se obtiene con la fórmula:

$$S_\infty = \frac{a}{1-r} \qquad 1.4.3.13$$

Donde a es el primer término y r es la razón común. Por lo tanto la ecuación 1.4.3.12 se reduce a:

$$P_0 = 1 - \frac{\lambda}{\mu} \qquad 1.4.3.14$$

Substituyendo la ecuación 1.4.3.14 en la ecuación 1.4.3.11 se obtendrá finalmente una expresión matemática completamente definida para P_n, donde n es igual a cero o mayor que él. La ecuación P_n es:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \qquad n \geq 0 \qquad 1.4.3.15$$

Como ya se ha resuelto P_n, se puede escribir una expresión para el número promedio o esperado de los que están recibiendo servicio y esperando en el sistema, $E(n)$:

$$E(n)=\sum_{n=0}^{\infty} nP_n$$

Substituyendo la ecuación 1.3.4.15 por P_n, en la ecuación anterior resultará:

$$E(n)=\left(1-\frac{\lambda}{\mu}\right)\sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n \qquad 1.4.3.16$$

La secuencia de términos en la ecuación 1.4.3.16 tiene la forma $0, a, 2a^2, 3a^3, \cdots, xa^x, \cdots$. Se utilizará la serie :

$$S_\infty = \frac{a}{(1-a)^2} \qquad 1.4.3.17$$

para resolver la ecuación 1.4.3.16 siendo $a=(\lambda/\mu)$ por lo tanto tenemos

$$E(n)=\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda/\mu}{\left(1-\lambda/\mu\right)^2}\right)$$

$$E(n)=\frac{\lambda/\mu}{1-\lambda/\mu}$$

$$E(n)=\frac{\lambda}{\mu-\lambda} \qquad 1.4.3.18$$

A fin de determinar el promedio de clientes que esperan servicio, es necesario emplear una notación distinta de la anterior ***E(n)***. Usaremos ***E(w)*** para designar el tiempo esperado que se pasa un cliente en el sistema antes de entrar a la instalación de servicio. Como sólo puede haber una unidad en cualquier tiempo en la instalación de servicio, el número promedio de unidades (tanto en espera como en servicio), o en ***E(n)***, menos la que está recibiendo servicio (λ/μ) debe ser el promedio de clientes que esperan servicio. Por lo tanto la ecuación es:

$$E(w) = E(n) - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu - \lambda)} \qquad 1.4.3.19$$

Con las ecuaciones 1.4.3.18 y 1.4.3.19 se pueden derivar otros modelos de líneas de espera. Esta vez puede determinarse el tiempo promedio que pasa un cliente en el sistema (***E(v)***).

$$E(v) = \frac{1}{\mu - \lambda} \qquad 1.4.3.20$$

Otra podría ser el tiempo promedio que espera un cliente antes de recibir servicio ***E(y)***.

$$E(y) = \frac{\lambda}{\mu(\mu - \lambda)} \qquad 1.4.3.21$$

1.5 COLAS EN MASA

Existen muchas actividades de colas en las que las llegadas y el servicio pueden ocurrir en grupos, o sea, en masa o en tandas. Varias personas pueden ir juntas a un restaurante y obtener servicio en una tanda.

Cierto número de llamadas telefónicas interurbanas pueden presentarse simultáneamente ante una telefonista. Un ascensor sirve a un grupo de personas al mismo tiempo. Aunque un avión llegue en solitario a un aeropuerto, si se consideran las operaciones de colas por las que tienen que pasar sus pasajeros, es evidente que para algunos objetivos se tiene que aplicar el supuesto de llegada en masa, como por ejemplo disponer suficientes funcionarios de aduanas que atiendan a los grupos que llegan. Obsérvese que los pasajeros no se presentarán todos a la vez, pero tienden a moverse en pequeños grupos.

Es evidente que estamos tratando de un tipo de problema de colas más general que comprende, como casos particulares, a los de llegada y servicio de uno en uno. En muchas aplicaciones prácticas, las llegadas y el servicio en masa son las únicas hipótesis realistas.

Se desarrollarán las fórmulas que dan el número esperado de espera, en el estado estacionario, para una cola con canal de servicio único, ingreso de Poisson y tiempos de duración del servicio arbitrario, en la que los ítems son servidos en tandas que no exceden de un cierto número s. Así, cada vez que se desocupa la instalación de servicio, s ítems son servidos si la longitud de la cola es mayor que s. En otro caso, la cola entera es servida en una tanda.

Kendall ha utilizado el hecho de que se puede estudiar una cola de canal único con ingreso de Poisson e instantes de salida que sean puntos de regeneración (concepto debido a Palm[3]). De acuerdo con él, un instante es un punto de regeneración si el conocimiento del estado del proceso en un instante particular tiene la característica de Markov: que una información sobre el comportamiento en el pasado del proceso no tiene valor de predicción. Por tanto, un proceso de Markov es un proceso para el cual cada instante es un punto de regeneración.

La ventaja de esto es que el problema se reduce a uno con una cadena de Markov con tiempo discreto a pesar de que el proceso de congestión no sea en sí mismo de Markov.

El concepto de puntos de regeneración puede aplicarse a los instantes de llegada si los tiempos de servicio se distribuyen exponencialmente.

Para ingreso y tiempos de servicio arbitrarios, los únicos puntos de regeneración posibles son instantes en los que ocurren simultáneamente una llegada y una partida, lo que es extremadamente raro, e instantes en los que llega un nuevo cliente y encuentra libre la instalación (lo que ocurre en ocasiones), pero aquí no existe una forma sencilla de tratamiento con puntos de regeneración.

Obsérvese, como hemos visto en la deducción de la fórmula Pollaczek-Khintchine, la necesidad de ingreso de Poisson, que implica la independencia de la extensión de la cola en el momento de partida de un cliente y del número de ítems que llegan durante el período de servicio siguiente. Para ingresos más generales, puede esperarse la dependencia entre dos cantidades.

1.5.1 LLEGADAS DE POISSON, SERVICIO EN MASA.

Se estudiará el caso de un ingreso de Poisson y distribución arbitrario *B(t)* del tiempo de servicio.

En primer lugar, introduciremos las probabilidades de cambio r_{ij} que dan la probabilidad condicionada de que el estado siguiente sea E_j (o sea, que haya j ítems en la línea) cuando el estado anterior era E_i. La longitud de la cola se mide en aquellos instantes (puntos de regeneración) exactamente antes que tenga lugar el servicio de una tanda o, lo que es equivalente, exactamente después que una tanda haya completado el servicio. Si se designa por π_j *(j =0,1,2,...)* la probabilidad de estar en el estado E_j antes de empezar el servicio, π_j se obtiene, evidentemente, a partir de π_i multiplicando π_i por r_{ij} y sumando sobre i para tener en cuenta todas las posibles maneras de pasar al estado E_j desde el E_i.

Luego cuando el sistema está en equilibrio estadístico se tiene:

$$\pi_j = \sum_{i=0}^{\infty} \pi_i r_{ij} \qquad 1.5.1.1$$

Obsérvese que en le caso de transición tendríamos π_j^{n+1} en el primer miembro y π_i^n en el segundo.

Puesto que el objetivo es hallar π_j para todo j, tendremos que encontrar en primer lugar un método r_{ij}. Obsérvese que utilizaremos unas π_j que no son necesariamente las probabilidades verdaderas de estado estacionario en un instante.

El proceso de Poisson $\left((\lambda t)^n e^{-\lambda t}\right)/_{n!}$ $(n = 0,1,2,...)$ define la distribución de llegada de n ítems durante un intervalo de servicio, t, dado. Sea $B(t)$ la distribución acumulativa del tiempo de servicio $\left[dB(t)/_{dt} = b(t), \text{ si existe}\right]$. Supongamos que los tiempos de servicio, que son también los tiempos entre instantes sucesivos de partida, se distribuyen independientemente. Entonces la probabilidad de n llegadas durante un tiempo de servicio elegido al azar es:

$$k_n = \frac{1}{n!}\int_0^\infty e^{-\lambda t}(\lambda t)^n \, dB(t), \qquad n = 0,1,2,... \qquad 1.5.1.2$$

Donde se ha tomado la integral sobre todos los intervalos de longitud t, cuya distribución es $B(t)$.

$$K(z) = \sum_{n=0}^{\infty} k_n z^n = \int_0^\infty e^{-(1-z)\lambda t} \, dB(t) \equiv \beta[\lambda(1-z)] \qquad 1.5.1.3$$

$K(z)$ se relaciona fácilmente con la transformación de Laplace - Stieltjets $\beta(z)$ de la distribución del tiempo de servicio.

Multiplicamos ahora la ecuación *(1.5.1.1)* por z^j y sumemos sobre j. Se tiene así

$$P(z) \equiv \sum_{j=0}^{\infty} \pi_j z^j = \sum_{j=0}^{\infty} z^j \sum_{i=0}^{\infty} \pi_i r_{ij} \qquad 1.5.1.4$$

Puesto que las unidades son servidas en tandas de s o menos, una variación en el sistema desde i a j unidades se obtiene o bien sirviendo a todas las i unidades en una tanda (si $i \leq s-1$) y llegando j unidades, o bien (si $i \geq s$) sirviendo a s unidades y contemplando las $i-s$ restantes con $j-(i-s)$ nuevas llegadas el total j. Por último, si $i > j+s$, $r_{ij} = 0$, ya que si son servidas s unidades es imposible reducir el número a j. Por tanto,

$$r_{ij} = \begin{cases} k_j & para \quad 0 \leq i \leq s-1 \\ k_{j-(i-s)} & para \quad j+s \geq i \geq s \\ 0 & para \quad i > j+s \end{cases} \qquad 1.5.1.5$$

y $k_j = 0$ para $j < 0$

Tenemos ahora:

$$P(z) = \sum_{j=0}^{\infty} z^j \left(\sum_{i=0}^{s-1} \pi_i k_j + \sum_{i=s}^{j+s} \pi_i k_{j-i+s} \right) =$$

$$= K(z) \sum_{i=0}^{s-1} \pi_i \div \sum_{j=0}^{\infty} z^j \left(\pi_s k_j + \pi_{s-1} k_{j-1} + \ldots + \pi_{s+j} k_0 \right) \qquad 1.5.1.6$$

Obsérvese que $\pi_s k_j + \pi_{s-1} k_{j-1} + \ldots + \pi_{s+j} k_0$ es la convolución de dos sucesiones.

Donde la convolución esta definida de la siguiente manera:

Intuitivamente podemos mirar a la convolución de dos funciones $g(z)$ y $f(x)$ como la función resultante que aparece después de efectuar los siguientes pasos:

a) Girar respecto del origen los valores de una de ellas, es decir $g(z) = g(-z)$ para toda z desde - ∞ a ∞ Ir trasladando la función girada sobre la otra $f(z) * g(x - z)$
b) En cada punto x calculamos el valor que resulta de sumar los productos obtenidos de multiplicar para todas las z los correspondientes valores de las funciones $f(z)$ y $g(x - z)$.

En esencia estamos calculando para cada valor de *x* una especie de valor ponderado de una de las funciones *f(x)* con los valores de la otra *g(x)*. En el caso de que el área encerrada por la curva de *g(x)* fuese igual

1 entonces estaríamos calculando para x una media ponderada. Matemáticamente la expresión para esta operación es:

$$f(x)*g(x)=h(x)=\int_{-\infty}^{\infty} f(z)\,g(x-z)\,dz \qquad 1.5.1.7$$

De la expresión anterior puede verse como para un valor fijo de x los origines de las funciones *f* y *g* están desplazados justamente en ese valor *x*. Los valores de *f* para *z* crecientes van siendo multiplicados por valores de *g* para *(x - z)*.

Utilicemos el hecho de que la función generatriz de la convolución es el producto de las funciones generatrices de las sucesiones por separado. La función generatriz de k_j es *K(z)* y la de π_{s+j} es

$$\sum_{t=0}^{\infty}\pi_{j+s}z^{n}=\sum_{t=s}^{\infty}\pi_{t}z^{t-s}$$

También

$$z^{-s}\sum_{t=s}^{\infty}\pi_{t}z^{t}=\left[P(z)-\sum_{t=0}^{s-1}\pi_{t}z^{t}\right]z^{-s} \qquad 1.5.1.8$$

y, por último tenemos

$$P(z)=K(z)z^{-s}\left[P(z)+\sum_{t=0}^{s-1}\pi_{t}\left(z^{s}-z^{t}\right)\right] \qquad 1.5.1.9$$

de donde despejando *P(z)*, resulta:

$$P(z)=\frac{\sum_{t=0}^{s-1}\pi_{t}\left(z^{s}-z^{t}\right)}{\left(z^{s},\ K(z)\right)-1} \qquad 1.5.1.10$$

Las probabilidades desconocidas $\pi_0, \pi_1, \ldots, \pi_{s-1}$ se determinan considerando que, para que *P(z)* converja sobre y en el interior del círculo unitario (como se desprende de su definición), los ceros del numerador tienen que coincidir con los del denominador dentro del círculo $z = 1 + \delta$ para $\delta > 0$ y pequeño. Esta condición se satisface en el ejemplo que se considera más adelante.

Supóngase que los intervalos de servicio se distribuyen según la función ji- cuadrada con *2k* grados de libertad, es decir, tienen la distribución de Erlang.

$$b(t) = \frac{\mu^k}{\Gamma(k)} t^{k-1} e^{-\mu t}, \qquad 0 \le t < \infty \qquad 1.5.1.11$$

Donde el tiempo de servicio esperado k/μ define la intensidad de tráfico ρ por la relación:

$$\rho \equiv \frac{\lambda k}{\mu s} = \frac{\lambda}{s}\left(\sum_{j=0}^{s} j\pi_j + s\sum_{j=s+1}^{\infty}\pi_j\right)^{-1} \qquad 1.5.1.12$$

Es decir, el número medio de llegadas en un intervalo de servicio es igual a la extensión media de la tanda servida. Obsérvese que tal cola puede alcanzar un estado estacionario si $\rho < 1$, lo que supondremos que se verifica.

Utilizando las ecuaciones (1.5.1.2) y la (1.5.1.3), se tiene:

$$K(z) = \left[1 + \frac{\rho s(1-z)}{k}\right]^{-k} \qquad 1.5.1.13$$

Si se considera el denominador de la ecuación (1.4.1.10), puede demostrarse que tiene *s-1* ceros (raíces) simples z en $z \le 1$ distintos de *z=1*.

Observación: Para determinar el número de ceros, se aplica el teorema de Rouché a z^s y a $z^s - K(z)$ en el interior y sobre el círculo $|z| = 1 + \delta$. Cada uno de los ceros (raíces), cuando se sustituye z en el numerador, anula a este. Sobre y en el interior del círculo unitario no puede haber ceros múltiples; de otra forma se anularían el denominador y su derivada, y al dividir las dos expresiones se obtiene que $\rho > 1$, lo que hemos excluido.

Evidentemente, cuando $z \to 1$, $P(z) \to 1$. Por tanto, aplicando la regla de L'Hôpital a la ecuación (1.4.1.11), se tiene, cuando $z \to 1$.

$$\sum_{i=0}^{s-1} (s-i)\pi_i = s - s\rho \qquad 1.5.1.14$$

Esto, juntamente con las $s-1$ ecuaciones obtenidas del numerador al sustituir los ceros del denominador e igualar a cero, forman un conjunto consistente (es decir, el determinante de los coeficientes no se anula) a partir del cual pueden determinarse las π_j. Si se escribe en una columna el conjunto de coeficientes de cada π_j se obtiene un nuevo determinante al restar de cada columna la columna de su izquierda y sacar como factores los factores comunes de los elementos de cada fila, determinante que no se anula, puesto que todas las z_i son distintas y ninguna es igual a la unidad:

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ 1 & z_1 & \cdots & z_1^{s-1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & z_{s-1} & \cdots & z_{s-1}^{s-1} \end{vmatrix} \prod_{i=1}^{s'1} (z_i - 1) \qquad 1.5.1.15$$

Observando la ecuación (1.4.1.13), se concluye que el denominador de la ecuación (1.4.1.10) tiene k ceros z_j fuera del círculo unitario, ya que son $s+k$ ceros en total; en consecuencia, después de reducir los s factores comunes del numerador y del denominador, o sea, $z-1$ y $z-z_i$ $(i = 1,\ldots,s'1)$ [de otra manera no existiría $P(z)$ en el círculo unitario para esos valores], se tiene:

$$P(z) = \frac{A}{\prod_{j=s}^{s+k-1} (z_j - z)} \qquad 1.4.1.16$$

Donde A es una constante de proporcionalidad. Poniendo $z=1$ y teniendo en cuenta que $P(1)=1$, resulta:

$$P(z)=\prod_{j=s}^{s+k-1}\frac{z_j-1}{z_j-z} \qquad 1.4.1.17$$

de la que se pueden obtener las π_j por desarrollo en fracciones parciales.

La función generatriz de momentos $M(\theta)$, resulta cuando se hace $z = e^{\theta}$ en $P(z)$, y la función generatriz de cumulantes $\Psi(\theta)$ se obtiene tomando el logaritmo de la función generatriz de momentos, como de costumbre. el número esperado en la cola se obtiene hallando la primera derivada de la función generatriz de cumulantes con respecto a θ y haciendo $\theta = 0$. Esto da:

$$\sum_{j=s}^{s+k-1}(z_j-1)^{-1} \qquad 1.4.1.18$$

La segunda derivada de la varianza:

$$\sum_{j=s}^{s+k-1}z_j(z_j-1)^{-2} \qquad 1.4.1.19$$

Como se puede ver es sumamente difícil encontrar los ceros de la ecuación (1.4.1.10), para con ellos poder determinar las probabilidades del tamaño de la cola, en cualquier instante, solo se puede encontrar en forma analítica la esperanza y la varianza de la cola en el estado estable.

Capitulo 2

CAPITULO 2

2 MODELO DE SIMULACIÓN

Aunque la teoría de líneas de espera se ha venido desarrollando desde principio de siglo, a la fecha existen aún modelos que no han podido ser resueltos en su forma analítica. Uno de estos modelos es el de líneas de espera con servicio en masa. La ecuación para este modelo:

$$P(z)=\frac{\sum_{i=0}^{s-1}\pi_i\left(z^s-z^i\right)}{z^s-K(z)^{-1}} \qquad 1.5.1.10$$

presenta la siguiente dificultad para su solución:

- Para determinar las s raíces del polinomio, se aplica el teorema de Rouché a z^s y a z^s - $K(z)$ en el interior y sobre el círculo $z = 1+\delta$. Cada uno de los ceros (raíces), cuando se sustituye z en el numerador, anula a éste. Sobre y en el interior del círculo unitario no puede haber ceros múltiples; de otra forma se anularían el denominador y su derivada, y al dividir las dos expresiones se obtiene que $\rho > 1$, lo que hemos excluido, por que esto implica que la línea de espera crece infinitamente. Para $\rho \leq 1$ se propone simular el proceso de la línea de espera.

- Para entender el proceso de simulación, considérense los intervalos de tiempo entre llegadas de un proceso de Poisson con parámetro λ el cual tiene una distribución de frecuencia exponencial $\lambda e^{-\lambda t}$. Obsérvese que, si ha ocurrido una llegada, el tiempo hasta la próxima llegada es menor que t si, y solamente si, hay una o más llegadas en este intervalo. La probabilidad de una o más llegadas y, por tanto, la probabilidad de que el tiempo entre llegadas sea menor o igual que t, es:

$$\sum_{n=1}^{\infty}\frac{(\lambda t)^n e^{-\lambda t}}{n!}=1-e^{-\lambda t} \qquad 2.1$$

La cantidad del lado derecho de la ecuación (*2.1*) es una distribución acumulativa. Por lo tanto, se obtiene por diferenciación la función de densidad exponencial $\lambda e^{-\lambda t}$ como distribución de los tiempos entre llegadas que se aplicará cuando los tiempos entre llegadas tienen una distribución exponencial afín. Reciprocamente, si, por ejemplo los tiempos de servicio tienen una distribución exponencial, los clientes son admitidos al servicio con un proceso de Poisson.

Con esto vemos que el número de llegadas al servidor dependen del tiempo de servicio.

Ahora resolviendo la integral:

$$P(k)=\int_0^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!}\,\mu e^{-\mu t}\,dt \qquad 2.2$$

se tiene:

$$P(k)=\int_0^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!}\,\mu e^{-\mu t}\,dt \quad =\left(\frac{\rho}{1+\rho}\right)^k\left(\frac{1}{1+\rho}\right) \qquad 2.3$$

donde:

$$\rho=\left(\frac{\lambda}{\mu c}\right) \qquad 2.4$$

Se puede demostrar fácilmente que es una función de distribución de probabilidad ya que:

$$\left(\frac{\rho}{1+\rho}\right)+\left(\frac{1}{1+\rho}\right)=1 \qquad 2.5$$

La ecuación (2.3) es la probabilidad de llegadas o arribos al sistema.

Ahora proponemos a U_n^t como la probabilidad de que se hallan quedado n personas en espera o en línea al tiempo t, justo después de que el servidor cerró las puertas de servicio.

Esto es:

$$U_n^t = \left\{ \sum_{i=0}^{s} D_i^t , D_{s+1}^t , D_{s+2}^t , \cdots \right\} \qquad 2.6$$

donde:

$$D_n^t = Conv\left(P(k), U_n^{t-1}\right) \qquad 2.7$$

Y se define como D_n^t que es la convolución (definida anteriormente) entre la función de distribución $P(k)$ y la función de distribución U_n^{t-1}.

Se puede ver con facilidad que $U_0^0 = 1$ y $U_n^0 = 0$ para $n > 0$. Esto es, no existe ninguna persona esperando al servidor al tiempo cero. Esto es lógico debido a que no puede haber personas en espera del servidor antes de que empiece a brindarse el servicio.

Por lo tanto para la primera etapa de servicio se tendrá:

$$D_n^1 = Conv\left(P(k), U_n^0\right) = \mathrm{P(k)} \qquad 2.8$$

y se puede representar a la convolución con la siguiente matriz

	U_0	U_1	U_2	...
P_0	a_{00}	a_{01}	a_{02}	...
P_1	a_{10}	a_{11}	a_{12}	...
P_2	a_{20}	a_{21}	a_{22}	...
...	...	...	...	...

Por lo tanto si $a_{ij} = P_i U_j$, entonces la convolución es $D_k = \sum_{j=0}^{k} U_j P_{k-j}$.

Con los resultados obtenidos en la convolución, se obtiene una nueva función de distribución para la probabilidad de que se encuentren k unidades en el sistema.

Con esto se obtiene el modelo para la primera etapa de servicio. Para las siguientes etapas se utilizaran las ecuaciones (2.7) y (2.8).

El vector de probabilidad *P(k)* nunca cambia a lo largo del tiempo, debido a que las probabilidades de llegadas o arribos al sistema nunca cambian, más sin embargo, el vector D_n^t se va expandiendo a lo largo del tiempo.

Como se puede ver, este es un sistema iterativo y se considera estable cuando la diferencia entre las medias de D_n^t y D_n^{t-1} sea menor que un ε previamente determinado.

Capitulo 3

CAPITULO 3

3 Análisis de los resultados

En este capítulo nos centraremos en el análisis de los resultados obtenidos en base al sistema de simulación propuesto. Una de las características del modelo es que, es posible encontrar la distribución de probabilidad para la línea de espera en cualquier momento del proceso lo que no se puede hacer mediante el resultado analítico, debido a que éste sólo puede encontrar el promedio del número de elementos en la fila y únicamente cuando el sistema a alcanzado el estado estable.

Para realizar esta simulación fue necesario desarrollar un programa en lenguaje C, debido a su gran flexibilidad para el manejo de funciones, apuntadores y de memoria, que se requerían para implementar el programa.

El algoritmo para esta simulación parece muy sencillo pero no lo es del todo cierto:

Primero se tiene que generar el vector P(k) de probabilidades a partir de la ecuación 2.3 y con él se generarán los diferentes vectores D y U, como se mencionó en el capítulo anterior.

Debido a que estamos desarrollando un programa de computo debemos cuidar que se cumpla la condición de normalidad para los diferentes vectores probabilísticos; esto es, que la suma de todas las probabilidades sea uno. Para realizar esta tarea es necesario proponer un ε pequeño para seguir generando probabilidades hasta que se cumpla la siguiente diferencia $(1 - P(K)) < \varepsilon$, siendo P(K) el vector de probabilidad acumulada. Con el resultado obtenido se calculará el siguiente cociente P(k) / P(K), asegurando el cumplimiento de la condición antes mencionada. Esto se aplica a los demás vectores antes mencionados.

También se debe de cuidar dentro del proceso de convolución que no se estén realizando operaciones innecesarias como: realizar productos con ceros, ya que en el proceso de convolución existen algunos elementos tanto de P(k) como de U que son ceros y no se necesita realizar este producto.

Este proceso es iterativo y se considera estable cuando la diferencia entre las medias de D_n^t y D_n^{t-1} sea menor que un ε previamente determinado.

Una vez desarrollado y probado el programa computacional, se aplicó obteniéndose los siguientes resultados para diferentes capacidades del servidor:

Capacidad del servidor	Con ρ = .9		Con ρ = .8	
	Media	Varianza	Media	Varianza
1	8.9999	80.9999	4.0	16.0
2	6.8297	46.5389	3.0789	9.440
3	6.1065	37.2893	2.7689	7.6352
6	5.3210	28.3130	2.4355	5.9316
12	5.018	25.180324	2.3034	5.3056
24	4.8295	23.3240	2.1995	4.8215

Tabla I

Dawton resolvió la ecuación 1.5.1.10 encontrando solamente la media y la varianza para el estado estable, sin resultados para estados intermedios (en cualquier instante). En la siguiente tabla se presentan los resultados obtenidos por Dawton[4] en 1955. Siendo los únicos disponibles a la fecha.

Capacidad del servidor	Con $\rho = .9$		Con $\rho = .8$	
	Media	Varianza	Media	Varianza
1	9.0	81.0	4.0	16.0
2	6.83	46.64	3.08	9.46
3	6.11	37.28	2.77	7.66
6	5.39	29.0	2.46	6.05
12	5.02	25.23	2.31	5.32
24	4.84	23.44	2.23	4.97

Tabla 2

El estado estable, aplicando el programa computacional de simulación diseñado, se encuentra después de 2500 iteraciones, y como se puede apreciar los resultados son muy parecidos a los que encontró Dawton (Tabla 1 y Tabla 2 respectivamente), para cada una de las diferentes capacidades.

Una ventaja que se tiene sobre Dawton es que sólo se tiene que ejecutar el programa de simulación y alimentarlo para obtener los resultados deseados, (las probabilidades del tamaño de la fila). en cualquier instante del tiempo, mientras que para poderlo resolverlo analíticamente es necesario resolver varios procesos algebraicos, cuyos resultados (media y variancia) son puntuales en el horizonte del tiempo del estado estable.

Capitulo 4

CAPITULO 4

4 Conclusiones y Recomendaciones

El desarrollo de este trabajo fue de gran importancia personal debido a que me permitió adentrarme con mayor detalle a los diferentes campos de estudio relacionados con mi carrera y con ello reafirmar conocimientos obtenidos a lo largo de ésta.

Además, es posible que este trabajo le pueda servir a alumnos de diferentes carreras como una guía para adentrarse al estudio de simulación y sobre todo al de los modelos de las líneas de espera, y podría ser útil para las carreras tanto del área de matemáticas aplicadas como a las de ingeniería y no solo a los alumnos, si no también podría ser una ayuda para todas personas que de alguna forma están relacionadas con los sistemas de líneas de espera.

Un punto importante que me gustaría hacer énfasis en este trabajo, fue que se tuvo que realizar un programa de cómputo que fue de gran ayuda para mi desarrollo, debido a que tuve la necesidad de aprender un nuevo lenguaje de programación que en este caso fue C++, gracias es este lenguaje he conseguido oportunidades de trabajo y de desarrollo profesional.

Conclusiones:

Gracias a este tipo de simulación podemos analizar diferentes tipos de modelos de líneas de espera, que son difíciles de resolver analíticamente, por lo general éstos son cuando la capacidad del servidor es en masa.

Es posible determinar el número de servicios que se tienen que realizar y el número de elementos que forman la línea de espera antes de encontrar el estado estable.

Este modelo se podría proponer a la industria como solución a sus problemas de líneas de espera, debido a que son éstas las que por lo general dan inicio a los cuellos de botella dentro de la línea de producción o se podría utilizar para optimizar los tiempos entre llegadas en las líneas del metro para dar mejor servicio, etc.

Como ventaja de este método se tiene que por medio de la simulación es posible encontrar la **distribución de probabilidad de la línea de espera en cualquier instante** y no solamente en el estado estable como Dowton lo encontró.

Recomendaciones

Este es sólo el principio de un estudio que puede ampliarse:

Como por ejemplo en este trabajo se propone que el servicio consta de una sola etapa y se podría modificar haciendo que constara de varias etapas, como las diferentes estaciones de una línea del metro, o modificar las distribuciones de probabilidad para las llegadas, etc.

Bibliografía

Referencias bibliográficas

[1] Erlang "Probability and Telephone Exchanges", Nyt Tidsskr. Vol 20 1909
[2] Molina "The theory of probabilities Applied to Telephone Trunking Problems", Bell Systems Vol 1 1922

[3] Palm, "Calculus exact de la perte dans le groups de circuits èchelonnes" Ericson Tech vol 4 1936

[4] Downton "Waiting Time in Bulk Service Queue" Articulo de la Universidad de Liverpool 1955

Bibiografía

Artículos

Kendall "Some problems in the theory of queue" Roy. Stat. Soc. vol 13 1951
Kendall "Stochastic Processes Occurring in the Theory of Queue and their Analysis by the Method of the Imbedded Markov Chain" Oxford University 1953
Bailey "On Queuing Processes whit Bulk Service" Nuffield Lodge Regent's Park London 1953
Downton "Waiting Time in Bulk Service Queue" Articulo de la Universidad de Liverpool 1955

Libros

Saaty "Elements of Queuing Theory " McGraw Hill Book Company 1961
Hiller Liberman "Introducción a la Investigación de Operaciones" McGraw Hill 1997
Taha "Investigación de Operaciones" Alfaomega 1991
Thierauf y Grosse "Toma de Decisiones por medio de Investigación de Operaciones" Limusa 1975
Brunk "Introducción a la Estadística Matemática" Trillas 1979
Feller "An Introduction to Probability Theory and Its Applicattions" Wiley 1968
Leon Garcia "Probability and Random Processes for Electrical Engineering" Addison Wesley 1994
Canavos "Probabilidad y Estadística" Mcgraw Hill 1988

Anexo A

```
/* programa para la simulación de un sistema de líneas de espera con servicio en masa
Los datos que hay que  proporcionar son los de Lamda , Miu la capacidad el servidor y el
error deseado para la diferencia entre las medias deseadas  */


#include <iostream.h>
#include <conio.h>
#include <math.h>
#include <stdio.h>

double media1(double *,int);
double varianza(double *,int);
double suma(double *,int);
void obtener_U(double *, double *,int, int);
int obtencion_probabilidades(double *,double *,double,int);
void muestra_vec(double *, double*,int , char *);
void primerad(double *, double *,int);
double convolucion(double *,double *,int,int,int);
double Diferencia(double *,double *,int,int);
void asigna(double *,double *,int);
void main ()
{
clrscr();
double *SD1,*D0,*D1,*U,*PK,*SPK,ro,lamda,miu;
double error=0,sumad=0,sumad1=0;
double med1=0,med2=0,var1=0,var2=0;
int paro=0,contafin,num_servi,r=500,cont=0,multi=0,u=0;
cout<<" Da el valor de lamda \n";
cin>>lamda;
cout<<" Da el valor de miu \n";
cin>>miu;
cout<<" Da la capacidad del servidor \n";
cin>>num_servi;
cout<<" Da el error aceptable \n";
cin>>error;
U= new double [r];
PK=new double [r];
SPK=new double [r];
D0=new double [r];
D1=new double [r];
SD1=new double [r];
for(int i=0;i<=r;i++)
{
        U[i]=0;
        PK[i]=0;
        SPK[i]=0;
        D0[i]=0;
```

```
        D1[i]=0;
        SD1[i]=0;
}
ro=lamda/(miu*num_servi);
cont= obtencion_probabilidades(PK,SPK,ro,r);
muestra_vec(PK,SPK,cont,"Vectores\n");
cout<<"el valor de cont es: "<<cont<<"\n";
u=cont-num_servi;
multi=cont+cont-num_servi;
primerad(PK,D0,r);
int parar=0;
for(int ii=0;ii<=3000;ii++)
{
parar++;
obtener_U(D0,U,num_servi,r);
//muestra_vec(U,D0,cont,"Vectores U y D0\n");
//med1=media1(D0,r);
//var1=varianza(D0,r);
//med2=media1(U,r);
//var2=varianza(U,r);
//cout<<"media de D0  "<<med1<<"  varianza  "<<var1<<"\n";
//cout<<"media de U   "<<med2<<"  varianza  "<<var2<<"\n";
sumad=0;
for(i=0;i<=multi;i++)
{
   D1[i]=convolucion(U,PK,i,cont,u);
   sumad+=D1[i];
   SD1[i]=1-sumad;
}
//muestra_vec(D1,SD1,multi," **************\\\\  Vectores ////*****\n");
//cout<<"la suma de las d's es   "<<sumad<<"\n";
//getch();
for(i=multi;i<=r;i++)
{
   D1[i]=0;
}
//int paro=0;
sumad1=0;
for (i=0;i<r;i++)
{
//paro++;
D1[i]=D1[i]/sumad;
sumad1+=D1[i];
if(sumad1>=1.0-(1e-15))
{
paro=i;
i=r;
}
```

```
}
med1=media1(D0,r);
var1=varianza(D0,r);
med2=media1(D1,r);
var2=varianza(D1,r);
cout<<"media de D1   "<<med2<<"  varianza  "<<var2<<"\n";
if(fabs(med1-med2)<=error && ii!=0)
{
int final=ii;
ii=3000;
cout<<"se encontro el estado estable en  "<<final<<"\n";
cout<<"media de D1   "<<med2<<"  varianza  "<<var2<<"\n";
med1=media1(U,r);
var1=varianza(U,r);
cout<<"media de U   "<<med1<<"  varianza  "<<var1<<"\n";
}
cont=paro;
u=cont-num_servi;
multi=cont+cont-num_servi;
asigna(D0,D1,r);
}//fin ciclado
getch();
}//fin main

double media1(double *pk,int r)
{
        double media=0;
        for(int i=0;i<r;i++)
    {
        media=media+i*pk[i];
    }
 return(media);
}
double varianza(double *pk,int r)
{
    double media=media1(pk,r),prod=0,vari1=0;
    for(int i=0;i<r;i++)
    {
        prod=(i-media);
        vari1=vari1+(prod*prod*pk[i]);
    }
return(vari1);
}

void asigna(double *d0,double *d1, int j)
{
    for(int i=0;i<=j;i++)
    {
```

```
                d0[i]=d1[i];
        }
}
double Diferencia(double *d0,double *d1,int j)
{       double resta=0;
        for(int i=0;i<=j;i++)
        {resta=resta+(fabs(d0[i]-d1[i]));}
        return(resta);
}

double convolucion(double *uk,double *pk, int i, int p, int u)
{
        double suma1=0;
        int m=0;
        if(i>u)
        {
                if(i>p)
                {
                for(m=0;m<=(p+u-i);m++)
                        {
                        suma1+=uk[u-m]*pk[i-u+m];  //las dos son<i
                        }
                return(suma1);
                }    //if de i>p
                else
                {
                        for(m=0;m<u+1;m++)
                        {
                        suma1+=uk[u-m]*pk[i-u+m];     //i>u , i<p
                        }
                return(suma1);
                }//else de i>p
        } // fin de i>u
        else          //    i<u   //
        {
                if(i>p)
                {
                        for(m=0;m<=p;m++)
                        {
                        suma1+=uk[i-m]*pk[m];     //i<u , i>p
                        }
                return(suma1);
                }//if de i>p
                else
                {
                        for(m=0;m<=i;m++)
                        {
                        suma1+=uk[m]*pk[i-m];   // i<<u,  i<<p
```

```
                    }
                return(suma1);
                }// else de i>p
            //fin de else de  i<u
        } //if de i>u

}

void obtener_U(double *d,double *u.int num_servi,int r)
{ double a=0;
        for(int j=0;j<r-num_servi-1;j++)
        {
                if(j==0)
                        {a=suma(d.num_servi);}
                u[0]=a;
                u[j]=d[num_servi+j];
        }
        for(j=r-num_servi-1;j<=r;j++)
        {
        u[j]=0;
        }
}
double suma(double *d,int num_servi)
{
double sum=0;
        for(int k=0;k<=num_servi;k++)
        {
                sum=sum+d[k];
        }
        return(sum);
}

void primerad(double *PK,double*D0,int r)
{
        for(int i=0;i<=r;i++)
        D0[i]=PK[i];
}

int obtencion_probabilidades(double *pk,double *spk,double ro,int r)
{
        int contador=-1;
        double suma1=0;
        double auxi=0,p=0,q=0,aux1=1e-10;
        q=(ro/(1+ro));
        p=(1/(1+ro));
        for(int i=0;i<=r;i++)
        {
         contador=contador+1;
```

```
        pk[i]=(pow(q,i)*p);
        spk[i+1]=spk[i]+pk[i];
        if(spk[contador]>=1-aux1)
                {
                auxi=spk[contador];
                i=r;
                }
        }
        for(i=contador;i<=r;i++)
        {spk[i]=0;pk[i]=0;}
        for(i=0;i<=contador;i++)
        {
                pk[i]=pk[i]/auxi;
                suma1=suma1+pk[i];
        }

        cout<<"suma de pk = "<<suma1<<"\n";

        return(contador-1);
}
void muestra_vec(double *pk, double *spk,int r, char *p)
 {
 int con=1;
 cout<<p<<"\n";
 cout<<"\tk"<<"\t probabilidad "<<"\t\t\t"<<"suma de probabilidades\n";
 for(int i=0;i<r;i++){
  if( i==(10*con)){con++; }
   cout<<"\t"<<i<<"\t\t"<<pk[i]<<"\t\t\t"<<spk[i]<< "\n";}

 }
```