UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA.

TEORÍA DE COLAS Y SU APLICACIÓN AL SISTEMA BANCARIO.

TRABAJO DE GRADUACIÓN PRESENTADO POR:
CARLOS LUIS FLORES GARCÍA.
CAROLINA LISSETTE LINARES ALVARENGA.
JUAN MIGUEL BONILLA IRAHETA.

PARA OPTAR AL GRADO DE:
LICENCIADO/A EN ESTADÍSTICA.

CIUDAD UNIVERSITARIA, JUNIO DE 2009.

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA.

TEORÍA DE COLAS Y SU APLICACIÓN AL SISTEMA BANCARIO.

TRABAJO DE GRADUACIÓN PRESENTADO POR:
CARLOS LUIS FLORES GARCÍA.
CAROLINA LISSETTE LINARES ALVARENGA.
JUAN MIGUEL BONILLA IRAHETA.

PARA OPTAR AL GRADO DE:
LICENCIADO/A EN ESTADÍSTICA.

ASESOR:

_____
DR. JOSÉ NERYS FUNES TORRES.

CIUDAD UNIVERSITARIA, JUNIO DE 2009.

UNIVERSIDAD DE EL SALVADOR

RECTOR : ING. RUFINO ANTONIO QUEZADA SÁNCHEZ

SECRETARIO GENERAL : LIC. DOUGLAS VLADIMIR ALFARO CHÁVEZ

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO : DR. RAFAEL ANTONIO GÓMEZ ESCOTO

SECRETARIA : MARÍA TRINIDAD TRIGUERROS DE CASTRO

ESCUELA DE MATEMÁTICA

DIRECTOR : ING. CARLOS MAURICIO CANJURA LINARES

TRABAJO DE GRADUACIÓN APROBADO POR:

ASESOR : Dr. JOSÉ NERYS FUNES TORRES.

DEDICATORIA Y AGRADECIMIENTO

Doy gracias a Dios por nunca soltarme de la mano y estar de principio a fin en esta carrera, a mis padres Juan Linares Mauriz y Maria Alvarenga que con su sencilles me ha enseñado a seguir adelante a pesar de los obstáculos, a mi tia Cecila Perdomo por ayudarme a formar un carácter desde pequeña enseñandome que el dolor es solo algo momentaneo, a mis amigos Altagracia Saraí Peraza, Alvaro Campos, Gonzalo Alvarenga y Walter Otoniel Campos por sus consejos, ánimos, paciencia, cariño y tantas cosas buenas que encontre en ellos las cuales me ayudarón a seguir adelante, a mi compañero de tesis Juan Miguel Iraheta por tolerarme, darme animo y apoyarme en este proyecto; a su familia por su hospitalidad y afecto durante estos meses, a el Dr. Nerys Funes por su confianza y tiempo.

Carolina Lissette Linares Alvarenga.

A Dios, mi familia y amigos.

Juan Miguel Bonilla Iraheta.

Son muchas las personas a quienes de manera desinteresada debo gran parte de este triunfo, lograr alcanzar mi culminación académica, la cual es el anhelo de todos los que así lo deseamos. Primeramente a Dios quien nos guió cuando más lo necesitamos, a mi asesor de tesis, Nerys Funes, gracias a él y a su sabiduría; por su inestimable contribución, y consecución de nuestro trabajo.

A mi fabuloso equipo de tesis; Carolina y Juan Miguel, y a sus respectivas familias que si bien no fue parte del grupo legalmente, fueron un pilar en la obtención de nuestro trabajo de graduación. Este logro que era NUESTRO SUEÑO, y ahora es una realidad.

A mis padres, mis hermanos, tíos/as, abuelos y demás familiares por darme la estabilidad emocional, económica, sentimental; para poder alcanzar MI SUEÑO definitivamente no hubiese podido ser realidad sin la ayuda de todos ustedes. Madre, serás siempre mi inspiración para alcanzar mis metas, por enseñarme que todo lo que se hace con esfuerzo es al final recompensado.

Carlos Luis Flores García.

Índice general

Introducción.

Las colas son un aspecto de la vida moderna que nos encontramos continuamente en nuestras actividades diarias. Esto suele ocurrir, cuando la demanda real de un servicio es superior a la capacidad que existe para dar dicho servicio. Ejemplos reales de esa situación son: el pago de servicios en el sistema bancario, los semáforos mal ajustados, el pago de peaje en una autopista, los cajeros automáticos, el tráfico aéreo de un aeropuerto, la atención a clientes en un establecimiento comercial, la espera de los electrodomésticos dañados para ser reparados por un servicio técnico, la espera de pacientes para ser atendidos en un hospital, etc.

La teoría de colas o línea de espera, es una colección de modelos matemáticos que describen sistemas en donde los clientes esperan en una cola para recibir un servicio, estos clientes son elegidos de acuerdo a ciertos criterios del sistema de elección. Entre muchos usos, los modelos de colas pueden ser empleados para encontrar un "estado estable" en el sistema que genere un consumo óptimo de recursos, también se puede determinar la longitud promedio de la línea de espera (cola) y el tiempo de espera promedio para un sistema dado. El problema radica en determinar que tamaño o tasa de servicio proporciona ese consumo óptimo, esto no es sencillo de determinar en sistemas de colas en bancos, dada la aleatoriedad en las llegadas de nuevos clientes; además el tiempo de servicio no es fijo, en algunos casos. Esta información, junto con los costos pertinentes, se usan para determinar la capacidad del sistema apropiada.

El análisis de teoría de colas es una de las herramientas utilizadas para la mejora de los diferentes sistemas de producción o servicios; particulamente en lo que concierne a los sistemas de servicios, el sistema bancario es una de las aplicaciones de está teoría. Puesto que puede usarse para proporcionar respuestas aproximadas a muchas preguntas como las siguientes: ¿Cómo cambia el tiempo de espera en la cola si se agrega o se retira un canal de servicio (cajero)?, ¿Cuántos clientes son atendidos por servidor y cuánto tiempo desocupado tendrán los servidores?, ¿Cuántos canales de servicio o cajeros deberá tener un banco si se conoce el número de servicios por hora, de manera que el tiempo de espera sea aceptable tanto por el cliente, como para el servidor?. A todas las preguntas formuladas anteriormente se les dará respuesta en el desarrollo de nuestra investigación.

Capítulo 1

Introducción a la teoría de colas.

Históricamente, los primeros trabajos que comenzaron a dar cuerpo a la Teoría de Colas son los debidos al matemático danés A.K. Erlang, quien en 1909 publicó *La teoría de probabilidades y las conversaciones telefónicas*. Erlang era para ese entonces empleado de la Compañía Telefónica Danesa en Copenhage y su trabajo fue una aplicación de técnicas existentes en teoría de probabilidad al problema de determinar el número óptimo de líneas telefónicas en una centralita, teniendo en cuenta la frecuencia de las llamadas y su duración. Las aplicaciones de la Teoría de Colas a la telefonía continuaron después de Erlang. En 1927, E.C. Molina publicó *Aplicación de la teoría de la probabilidad a problemas de líneas telefónicas*; en 1928 T.C. Fry publicó *Probabilidad y sus usos en Ingeniería*. A principios de los años 30, F. Pollaczek publicó trabajos innovadores sobre el caso de llegadas poissonianas y servicios arbitrarios. También, por esa época, los matemáticos de la escuela Rusa A.N. Kolmogorov y A.Y. Khintchine, así como C.D. Crommelin, en Francia, y C. Palm, en Suecia, realizaron importantes aportaciones a la teoría de colas. A pesar de que a comienzos del estudio de la Teoría de Colas, las aportaciones fueron muy escasas, esta situación cambió notablemente a partir de los años 50, comenzando a publicarse gran número de trabajos sobre el tema. En la actualidad las aplicaciones de la Teoría de Colas en los campos de la Administración de empresas, Informática, las Telecomunicaciones y, en general, las nuevas tecnologías, abren aún un mayor porvenir a esta teoría matemática.

Como bien sabemos la formación de colas o líneas de espera suele ocurrir cuando la demanda real de un servicio es superior a la capacidad que existe para dar dicho servicio. Ejemplos reales de esa situación son: los cruces de dos vías de circulación, los semáforos, el peaje de una autopista, los cajeros automáticos, la atención a los clientes en un banco, la avería de electrodomésticos u

otro tipo de aparatos que deben ser reparados por un servicio técnico, etc. Fenómenos como los citados anteriormente y muchísimos otros tienen ciertas características comunes que dan lugar al desarrollo de modelos de sistemas de colas.
La teoría de colas estudia el comportamiento de sistemas donde existe un conjunto limitado de recursos para atender las peticiones generadas por los clientes, de tal manera que cuando un cliente envía una tarea al sistema, ésta podrá tener que esperar para ser atendida por algún recurso del sistema, o, incluso, podrá ser rechazada si el sistema no tiene capacidad suficiente para que sea atendida. El estudio de estos sistemas implicará el modelado no sólo del sistema en sí, sino también el comportamiento aleatorio del tráfico ofrecido por los clientes al sistema. Este tráfico ofrecido por los usuarios se modela mediante dos procesos estocásticos: procesos de llegadas de las tareas al sistema y procesos de servicios en el sistema, usualmente considerados independientes entre sí.
Con frecuencia, las empresas deben tomar decisiones respecto al caudal de servicios que debe estar preparada para ofrecer. Sin embargo, muchas veces es imposible predecir con exactitud cuándo llegarán los clientes que demandan el servicio y/o cuanto tiempo será necesario para dar ese servicio; es por eso que esas decisiones implican dilemas que hay que resolver con información escasa. Estar preparados para ofrecer todo servicio que se nos solicite en cualquier momento puede implicar mantener recursos ociosos y costos de operación excesivos. Pero, por otro lado, carecer de la capacidad de servicio suficiente causa colas excesivamente largas en ciertos momentos. Cuando los clientes tienen que esperar en una cola para recibir algún servicio, están pagando un coste, en tiempo, más alto del que esperaban. Las líneas de espera largas también son costosas para la empresa; ya que produce pérdida de prestigio y pérdida de clientes.

1.1. Componentes de un sistema de colas.

Los elementos que intervienen en un sistema de colas son los siguientes:

a) *El cliente*: es todo individuo de la población potencial que solicita servicio. Suponiendo que los tiempos de llegada de clientes consecutivos son $0 < t_1 < t_2 < \cdots$, será importante conocer el patrón de probabilidad según el cual la fuente de entrada genera clientes. Lo más habitual es tomar como referencia los tiempos entre las llegadas de dos clientes consecutivos: $\tau_k = t_k - t_{k-1}$, fijando su distribución de probabilidad. Normalmente, cuando

la población potencial es infinita se supone que la distribución de probabilidad de los τ_k (que será la llamada distribución de los tiempos entre llegadas) no depende del número de clientes que estén en espera de completar su servicio, mientras que en el caso de que la fuente de entrada sea finita, la distribución de los τ_k variará según el número de clientes en proceso de ser atendidos.

b) *La fuente de entrada o población potencial*: son todos aquellos clientes que pueden solicitar el servicio que brinda el sistema. Podemos considerarla finita o infinita. Aunque el caso de infinitud no es realista, sí permite (por extraño que parezca) resolver de forma más sencilla muchas situaciones en las que, en realidad, la población es finita pero muy grande. Dicha suposición de infinitud no resulta restrictiva cuando, aún siendo finita la población potencial, su número de elementos es tan grande que el número de individuos que ya están solicitando el citado servicio prácticamente no afecta a la frecuencia con la que la población potencial genera nuevas peticiones de servicio.

c) *La cola o línea de espera*: es donde los clientes, una vez que han llegado a solicitar el servicio, aguardan para ser atendidos. Siempre que los servidores se encuentren ocupados.

d) *La capacidad de la cola*: es el máximo número de clientes que pueden estar haciendo cola (antes de comenzar a ser servidos). De nuevo, puede suponerse finita o infinita. Lo más sencillo, a efectos de simplicidad en los cálculos, es suponerla infinita. Aunque es obvio que en la mayor parte de los casos reales la capacidad de la cola es finita, no es una gran restricción el suponerla infinita si es extremadamente improbable que no puedan entrar clientes a la cola por haberse llegado a ese número límite en la cola.

e) *El mecanismo de servicio*: Para determinar totalmente el mecanismo de servicio debemos conocer el número de servidores de dicho mecanismo (si dicho número fuese aleatorio, la distribución de probabilidad del mismo) y la distribución de probabilidad del tiempo que le lleva a cada servidor en dar un servicio. En caso de que los servidores tengan distinta destreza para dar el servicio se debe especificar la distribución del tiempo de servicio para cada uno.

f) *El sistema de la cola*: es el conjunto formado por la cola y el mecanismo de servicio, junto con la disciplina de la cola, que es lo que nos indica el criterio de qué cliente de la cola elegir para pasar al mecanismo de servicio.

Para describir un sistema de colas hay que saber interpretar el rol que desempeña cada uno de los elementos que la componen, haciendo las distintas hipótesis sobre éstos que correspondan en cada caso. A continuación, se muestran algunas de las hipótesis que pueden hacerse de cada uno de los elementos. En general, un centro puede ser multiservicio, de modo que un cliente ha de pasar por varios servidores, sin embargo las hipótesis que se presentan a continuación deben hacerse para cada uno de los puntos de servicio.

- Sobre las llegadas:
  - *Una o varias fuentes*: los clientes pueden proceder de una o varias fuentes.
  - *Independencia entre llegadas*: el tiempo entre las llegadas de clientes es independiente.
  - *Intervalos entre llegadas*: los tiempos entre llegadas pueden ser considerados de dos tipos: deterministas o aleatorios, y si se trata de estos últimos habrá que dar su distribución.
- Sobre la fuente:
  - *Fuente finita o infinita*: el tamaño de la fuente puede ser considerado infinito o un valor limitado, denominándose sistema abierto si es infinito o sistema cerrado si es finito.
  - *Llegadas en bloque o unitarias*: los clientes pueden llegar de forma unitaria (aunque puedan coincidir dos llegadas) o en bloques de tamaño fijo o variable.
- Sobre el servicio:
  - *Uno o varios servidores*: ha de establecerse el número de servidores que hay en el sistema.
  - *Independencia entre servidores*: puede haber independencia en cómo atiende cada uno de los servidores o no, o incluso si varía la forma de trabajar de un servidor dependiendo del estado de los demás.

- *Independencia de los tiempos de servicio*: los tiempos de servicio en un mismo servidor pueden ser independientes o no.
- *Duración de los tiempos de servicio*: esta duración puede ser de tipo determinista si es conocida con seguridad, o de tipo aleatorio, en cuyo caso ha de establecerse una distribución para esos tiempos.
- *Homogeneidad de los servidores*: los servidores pueden ser homogéneos o no, es decir, pueden tener todos una misma tasa de servicio o tenerla diferente, pero, en cualquier caso hay que establecer cuál es la de cada uno de ellos, o por simplificación y dependiendo del modelo y los objetivos planteados, si se dispone de la tasa del sistema, aunque sean diferentes pueden ser considerados homogéneos con la tasa del sistema dividido por el número de servidores.

- Sobre el comportamiento de los clientes:
  - *Impaciencia*: la impaciencia en los clientes se entiende como la renuncia prematura de un cliente a la petición de servicio que había acudido a solicitar.

- Sobre la cola:
  - *Número de canales en la cola*: la cola puede estar formada por un único canal o por varios.
  - *Interferencia de canales*: en el caso en que hayan varios canales en la cola, pueden producirse interferencias entre ellos, es decir, movimientos de clientes de un canal a otro.
  - *Capacidad limitada*: el sistema puede tener una capacidad limitada o no. Se entiende por capacidad el número máximo de clientes que pueden haber en el sistema, es decir, en servicio y en espera.
  - *Disciplina de la cola*: los clientes son seleccionados de la línea de espera mediante algún mecanismo, pudiendo ser el orden de selección alguno de los siguientes, que son los más habituales, o algún otro:
    - FIFO: (first in first out): Primero en llegar primero en ser servido.
    - LIFO: (Last in first out): Último en llegar primero en ser servido.

- SIRO: (Service in random order): Los clientes son servidos en orden aleatorio.
- PRI: (Priority): los clientes tienen distinta prioridad.

1.2. Terminología y notación.

En lo sucesivo utilizaremos las herramientas probabilísticas de los procesos de nacimiento y muerte para el estudio de colas con distribución del tiempo entre llegadas y distribución del tiempo de servicio exponencial. Antes hemos de fijar la notación que vamos a usar.

- $N(t)$: Denota el número de clientes en el sistema en el instante t. $N(t)$ es un proceso estocástico en tiempo continuo y con espacios de estados discreto.
- $Z(t)$: Denota el número de clientes que han salido del sistema hasta el instante t.
- $N_q(t)$: Representa el número de clientes en la cola en el instante t.
- $P_n(t)$: Es la probabilidad de que en el instante t, se encuentren n clientes en el sistema. Supongamos conocido el número de clientes en el instante cero (usualmente dicho número es cero).
- c: Denota el número de servidores del mecanismo de servicio.
- λ_n: Representa el número medio de llegadas de clientes al sistema, por unidad de tiempo, cuando ya hay n clientes en él. También se denomina tasa de llegadas (que se correspondería con la tasa de nacimientos si $N(t)$ es un proceso de nacimiento y muerte). Cuando las tasas de llegada no dependen de n (es decir, todos los λ_n son constantes) suele denotarse como λ dicho valor constante.
- μ_n: Es el número medio de clientes a los que se les completa el servicio, por unidad de tiempo, cuando hay n clientes en el sistema. Es frecuente referirse a los μ_n como tasas de complección de servicio (o, simplemente, tasas de servicio). Si todos los servidores tienen la misma distribución del tiempo de servicio, suele denotarse por μ el número medio de clientes que puede atender cada servidor por unidad de tiempo. Como consecuencia se tiene que $\mu_n = n\mu$ si $n = 1, 2, \ldots, c$ y $\mu_n = c \cdot \mu$ para $n \geq c$.

- ρ: Es la llamada constante de utilización del sistema o intensidad de tráfico. Se define, como
$$\rho = \frac{\lambda}{c \cdot \mu}$$
Cuando los λ_n son constantes y todos los servidores tienen la misma distribución de tiempo de servicio, λ es el número medio de clientes que entran en el sistema y $c \cdot \mu$ es el número medio de clientes a los que pueden dar servicio los c servidores cuando todos están ocupados. En estas condiciones, ρ representa la fracción de recursos del sistema que es consumida por los clientes. Así, intuitivamente, parece necesario que se cumpla, en estos casos, que $\rho < 1$ y además cuanto más cercano a 1 sea ρ, más tráfico ha de soportar el sistema (o menos tiempo libre tendrán los servidores, o más espera habrán de sufrir los clientes, como se quiera expresar). Aunque es evidente que ρ no tiene unidades, es habitual medir la intensidad de tráfico en *Erlangs*, en honor a los trabajos pioneros de *Erlang* en la teoría de colas.

Los modelos de colas que estudiaremos en el siguiente capítulo son todos estacionarios. En ellos las distribuciones de probabilidad marginales de los procesos estocásticos $\{N(t)\}_{t\geq 0}$ y $\{N_q(t)\}_{t\geq 0}$ no cambian con el tiempo t. En tales condiciones tiene perfecto sentido definir los siguientes conceptos:

- N: Es la variable aleatoria que contabiliza el número de clientes en el sistema.
- N_q: Denota la variable aleatoria número de clientes en la cola.
- S: Representa el tiempo de servicio.
- T_q: Representa el tiempo que un cliente invierte en la cola.
- $T = T_q + S$: El tiempo total que un cliente invierte en el sistema.
- p_n: Es la probabilidad de que se encuentren n clientes en el sistema para $n = 0, 1, \ldots$.
- L: Representa el número medio de clientes en el sistema, es decir $L = E[N]$.
- L_q: Es el número medio de clientes en la cola, o lo que es lo mismo, $L_q = E[N_q]$.
- **W**: Es la variable aleatoria que describe el tiempo que un cliente pasa en el sistema.

- $\mathbf{W}_q$: Es la variable aleatoria que representa el tiempo que un cliente espera en la cola.
- W: Es el tiempo medio que un cliente está en el sistema, o simplemente, $W = E[\mathbf{W}]$.
- W_q: Denota el tiempo medio de espera en la cola para un cliente genérico. Matemáticamente, $W_q = E[\mathbf{W}_q]$.

Para clasificar los posible tipos de sistemas de colas debemos especificar las características que determinan los elementos que lo componen. Así, Kendall introdujo en 1953 la notación $A/B/c$ para indicar que la distribución del tiempo entre llegadas es de tipo A, que B es la distribución del tiempo de servicio y que c es el número de servidores. Posteriormente esta notación se extendió dando lugar a la más habitual en nuestros días, consistente en designar el sistema de una cola con la nomenclatura $A/B/c/k/m/D$, donde:

- A: Distribución del tiempo entre llegadas consecutivas.
- B: Alude al patrón del tiempo de servicio por parte de los servidores disponibles.
- c: Es el número de canales de servicio.
- k: Es la restricción en la capacidad de la cola.
- m: Es el tamaño de la población potencial.
- D: Es la disciplina de la cola.

Así, por ejemplo, la notación $M/D/2/\infty/\infty/FIFO$ indica que se trata del sistema de una cola con tiempo entre llegadas exponenciales, tiempo de servicio determinístico (i.e. siempre se tarda el mismo tiempo en darle servicio a cada cliente), hay 2 servidores en el mecanismo de servicio, no existe límite para el número de clientes que pueden estar en la cola de espera, la población potencial se supone con infinitos clientes y los clientes son atendidos según una disciplina FIFO. Como los tres últimos valores $\infty/\infty/FIFO$ son precisamente los asignados por defecto, la notación anterior podría abreviarse como $M/D/2$. Obsérvese que este tipo de abreviaturas sólo se pueden realizar si todos los parámetros a partir de uno dado son iguales a los valores por defecto, ya que en caso contrario se produciría ambigüedad. Así, el modelo $E_2/U/3/\infty/4/FIFO$, no podría abreviarse como $E_2/U/3/4/FIFO$ (aunque sí como $E_2/U/3/\infty/4$), ya que, aunque es claro que $A = E_2$, $B = U$, $c = 3$ y $Z = FIFO$, nunca sabríamos si con él pretendemos indicar $k = 4$ y $m = \infty$ ó bien $m = 4$ y $k = \infty$.

1.2.1. Fórmulas de Little

En los modelos con distribución entre llegadas y distribución del servicio exponencial (así como en muchos otros modelos más generales llamados ergódicos) se verifican ciertas fórmulas que relacionan los números medios de clientes en el sistema o en la cola, con los tiempos medios de un cliente en el sistema o en la cola. Estas son las llamadas fórmulas de Little. Cuando las tasas de llegada son constantes (es decir; $\lambda_n = \lambda$ para todo $n = 0, 1, \dots$), la primera fórmula de Little establece la siguiente igualdad:

$$L = \lambda \cdot W$$

mientras que la segunda se expresa mediante

$$L_q = \lambda \cdot W_q$$

Realmente sólo la primera de ellas fue obtenida por Little en 1961, pero es costumbre referirse a ambas con el término primera y segunda fórmula de Little.

Una forma intuitiva de entender el porqué de la validez de las fórmulas de Little es la siguiente: considérese un cliente que llega al sistema justo ahora. Después de un tiempo, cuya media es W, ese cliente saldrá servido del sistema.

Como el número medio de clientes que llegan al sistema por unidad de tiempo es λ, el número medio de clientes que habrán llegado desde que nuestro cliente en cuestión entró en el sistema hasta que salió de él es $\lambda \cdot W$. Por otra parte, es obvio que dicho número medio de clientes es precisamente el número medio de clientes que hay en el sistema justo en el momento que sale del sistema nuestro cliente particular, es decir; L. Un razonamiento análogo es válido para la segunda fórmula de Little.

Obviamente, las fórmulas de Little no pueden ser válidas si los λ_n no son constantes (¿qué sería λ entonces?), pero si pueden generalizarse a esa situación mediante:

$$\begin{aligned} L &= \bar{\lambda} \cdot W \\ L_q &= \bar{\lambda} \cdot W_q \end{aligned}$$

siendo

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

Otra relación importante (en este caso para relacionar W y W_q) es la dada por

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

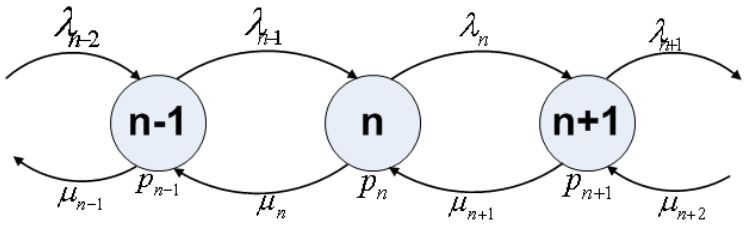
Su deducción es inmediata pues viene a decir que el tiempo medio que un cliente está en el sistema (W) coincide con la suma del tiempo medio en la cola (W_q) más el tiempo medio que tarda en ser servido $\frac{1}{\mu}$, ya que μ es el número medio de clientes que un servidor puede atender por unidad de tiempo.

1.3. Procesos de nacimiento y muerte.

El proceso de Poisson (y en general, los procesos de conteo) son útiles para modelar situaciones en las cuales el objetivo es contabilizar el número de ocurrencias de cierto fenómeno (nacimientos en una población) hasta un instante t. Sin embargo, existen otros procesos considerados más generales, llamados *procesos de nacimiento y muerte* los cuales contemplan la posibilidad de que el número de individuos en la población pueda disminuir. Por ejemplo: es posible considerar el caso de contabilizar el número de individuos en la población cada vez que se produzca una muerte. Entonces, los procesos de nacimiento y muerte, además de generalizar el proceso de Poisson, permite que las tasas de nacimientos y muertes puedan depender del número de individuos en la población.

Definición 1.3.1. *(Procesos de nacimiento y muerte).*
Considérese un proceso estocástico $\{N(t)/t \geq 0\}$ *con espacio de estados discretos:* $E = \{0, 1, 2, \ldots\}$ *y supóngase que el proceso describe un sistema que diremos que se encuentra en el estado* E_n *en el instante* t *cuando* $N_t = n$. *Se dirá que el proceso estocástico es de nacimiento y muerte si existen sucesiones de números no negativos* $\{\lambda_n;\ n = 0, 1, 2, \ldots\}$ *y* $\{\mu_n;\ n = 1, 2, \ldots\}$ *(llamadas tasas de nacimiento y muerte, respectivamente), tales que se verifican las siguientes propiedades*

1. *Los cambios de estado permitidos son de* E_0 *a* E_1 *y desde* E_n *a* E_{n-1} *o a* E_{n+1}, *para* $n \geq 1$ *(Ver Figura 1.1).*



2. *Si el sistema se encuentra en el estado E_n en el instante t, entonces, la probabilidad de que entre t y $t+\Delta t$ pase al estado E_{n+1} es $\lambda_n \Delta t + o(\Delta t)$, y si $n \geq 1$, la probabilidad de que pase a E_{n-1} es $\mu_n \Delta t + o(\Delta t)$.*

3. *La probabilidad de que ocurra más de un cambio en un intervalo de tiempo entre t y $t+\Delta t$ es $o(\Delta t)$.*

La función $o(\Delta t)$, es la que representa el error debido al tamaño del intervalo Δt, siendo tan pequeña, cuanto menor es el intervalo y además cumpliendo la propiedad:

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

De esta propiedad se deduce que la función de distribución de probabilidad de los tiempos entre llegadas será continua en cero, es decir; la probabilidad de que el tiempo entre llegadas toma el valor cero, es considerado infinitesimal. Entonces obtener el proceso de Poisson, a partir del proceso de nacimiento y muerte implica considerar tasas de muertes iguales a cero y tasas de nacimientos constantes e iguales a λ.
A continación se pretende determinar la distribución de probabilidades asociadas a un proceso de nacimiento y muerte. Sea $P_n(t)$, la probabilidad de que el sistema se encuentre en el estado E_n en el instante t, matemáticamente, $P_n(t) = P(N_t = n)$. Es importante considerar que en los procesos de nacimiento y muerte la posibilidad de transitar entre diferentes estados se puede explicar de la siguiente forma:
Teniendo en cuenta que la probabilidad de que se produzcan dos o más sucesos en un intervalo pequeño de tiempo es despreciable ($o(\Delta t)$), si en el instante $t + \Delta t$ hay n individuos, sólo deben considerarse tres posibilidades:

- En el instante t habían $n-1$ individuos y entre t y $t+\Delta t$ hubo un nacimiento y nínguna muerte.
- En el instante t habían n individuos y entre t y $t+\Delta t$ no hubo ni nacimientos, ni muertes.
- En el instante t habían $n+1$ individuos y entre t y $t+\Delta t$ hubo una muerte y ningún nacimiento.

De acuerdo con la regla de las probabilidades totales y con ayuda de la figura 1.1, se logra expresar de manera sencilla, la probabilidad de que el sistema esté en el estado E_n en el instante

$t + \Delta t$, dado los posibles estados que puede tomar el sistema en el instante t. Así, si $n \geq 1$, se tiene que:

$$\begin{aligned}
P_n(t+\Delta t) &= P(N_{t+\Delta t} = n/N_t = n-1)P(N_t = n-1) + P(N_{t+\Delta t} = n/N_t = n)P(N_t = n) \\
&+ P(N_{t+\Delta t} = n/N_t = n+1)P(N_t = n+1) + o(\Delta t) \\
&= P_{n-1}(t)\left[\lambda_{n-1}\Delta t + o(\Delta t)\right]\left[1 - \mu_{n-1}\Delta t + o(\Delta t)\right] \\
&+ P_n(t)\left[1 - \lambda_n\Delta t + o(\Delta t)\right]\left[1 - \mu_n\Delta t + o(\Delta t)\right] \\
&+ P_{n+1}(t)\left[1 - \lambda_{n+1}\Delta t + o(\Delta t)\right]\left[\mu_{n+1}\Delta t + o(\Delta t)\right] + o(\Delta t) \\
&= P_{n-1}(t)\left[\lambda_{n-1}\Delta t - \lambda_{n-1}\mu_{n-1}(\Delta t)^2\right] \\
&+ P_n(t)\left[1 - \mu_n\Delta t - \lambda_n\Delta t + \mu_n\lambda_n(\Delta t)^2\right] \\
&+ P_{n+1}(t)\left[\mu_{n+1}\Delta t - \mu_{n+1}\lambda_{n+1}(\Delta t)^2\right] + o(\Delta t) \\
&= P_{n-1}(t)\lambda_{n-1}\Delta t - P_{n-1}(t)\lambda_{n-1}\mu_{n-1}(\Delta t)^2 + P_n(t) - P_n(t)\mu_n\Delta t - P_n(t)\lambda_n\Delta t \\
&+ P_n(t)\mu_n\lambda_n(\Delta t)^2 + P_{n+1}(t)\mu_{n+1}\Delta t - P_{n+1}(t)\mu_{n+1}\lambda_{n+1}(\Delta t)^2 + o(\Delta t)
\end{aligned}$$

Si se resta $P_n(t)$ y se divide por Δt se tiene que,

$$\begin{aligned}
\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} &= P_{n+1}(t)(\mu_{n+1} - \lambda_{n+1}\mu_{n+1}\Delta t) + P_{n-1}(t)(\lambda_{n-1} - \lambda_{n-1}\mu_{n-1}\Delta t) \\
&- P_n(t)(\lambda_n + \mu_n - \lambda_n\mu_n\Delta t) + \frac{o(\Delta t)}{\Delta t}
\end{aligned}$$

aplicando límite cuando $\Delta t \to 0^+$

$$P_n'(t) = \lambda_{n-1}P_{n-1}(t) - (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t)$$

A partir de la última expresión se obtienen las llamadas **ecuaciones diferenciales de balance**:

$$P_n'(t) = \lambda_{n-1}P_{n-1}(t) - (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t),\ n \geq 1$$

y

$$P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t)$$

Supongamos que al inicio del tiempo el sistema se encuentra en el estado E_0 (es decir; en $t = 0$, no hay individuos en la población), se tienen condiciones iniciales $P_0(0) = 1$ y $P_n(0) = 0, \forall n \geq 1$. En general, las ecuaciones de balance son difíciles de resolver. De todas formas hay algunos casos

particulares en los que la resolución es más sencilla.
Así, si $\mu_n = 0\ \forall n = 1, 2, \ldots$ y $\lambda_n = \lambda\ \forall n = 0, 1, \ldots$ (Es decir; para el proceso de Poisson), las ecuaciones de balance resultan especialmente sencillas:

$$P_n'(t) = \lambda P_{n-1}(t) - \lambda P_n(t),\ n \geq 1$$

$$P_0'(t) = -\lambda P_0(t)$$

y puede probarse sin excesiva dificultad que la solución es

$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!},\ \forall t \geq 0.$$

De hecho, para $n = 0$, la ecuación diferencial $P_0'(t) = -\lambda P_0(t)$ esta expresión es muy fácil de resolver, siendo su solución general, de la forma $P_0(t) = Ce^{-\lambda t}$ pero bajo la condición inicial $P_0(0) = 1$ y procediendo, en el caso general por inducción en n.
Cuando el proceso estocástico de nacimiento y muerte es estacionario, las funciones $P_n(t)$ son constantes p_n, que no dependen de t, y por tanto, el sistema de ecuaciones diferencias de balance se convierte en un sistema de infinitas ecuaciones lineales:

$$0 = \lambda_{n-1}p_{n-1} - (\lambda_n + \mu_n)p_n + \mu_{n+1}p_{n+1} \quad si\ n \geq 1$$

y

$$0 = -\lambda_0 p_0 + \mu_1 p_1$$

que también pueden expresarse como:

$$(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} \quad si\ n \geq 1$$

y

$$\lambda_0 p_0 = \mu_1 p_1$$

Una forma intuitiva de interpretar estas ecuaciones es la siguiente: para cada posible estado, n, el miembro de la izquierda representa la probabilidad de dicho estado multiplicada por la suma de las tasas correspondientes a las formas de salir de este estado hacia otro distinto. Los términos de la derecha de cada ecuación de balance expresan la suma de las probabilidades de aquellos estados desde los cuales se puede llegar al estado n en una sola transición, multiplicadas por las tasas correspondientes a dicha transición.

Así, si $n \geq 1$, en el término de la izquierda se multiplica a p_n por la suma de las tasas λ_n, que corresponde al hecho de que se produzca un nacimiento cuando el sistema está en el estado n (pasando, por tanto a $n+1$) y para el caso de μ_n, corresponde a que se produzca una muerte cuando hay n individuos en la población (pasando por consiguiente, a una población con $n-1$ individuos). Cuando $n = 0$ el razonamiento anterior es válido salvo en lo referente a μ_0, el cual no aparece, porque no es lógico considerar de que se produzca una muerte cuando no hay individuos en la población. Para los términos de la derecha, cuando $n \geq 1$, se puede llegar a un estado n procedente del estado $n-1$ (en cuyo caso debe haber un nacimiento, con tasa λ_{n-1}) o bien del estado $n+1$ (siempre que haya una muerte, con tasa μ_{n+1}). Obviamente, si $n = 0$, no aparecerá el término correspondiente a $n-1$.

En este caso el proceso de nacimiento y muerte es estacionario, resulta bastante sencillo resolver las ecuaciones de balance. Así, tomando la ecuación correspondiente a $n = 0$ se puede despejar p_1, obteniendo

$$p_1 = \frac{\lambda_0}{\mu_1} p_0$$

Además, puede probarse por inducción una generalización de esta expresión para cualquier índice n:

$$p_n = \frac{\lambda_{n-1}}{\mu_n} p_{n-1}.$$

En efecto, la expresión es cierta para $n = 1$. Supongámos que también es cierta para n y probemos para $n+1$. Utilizando la ecuación de balance n-ésima, se tiene que

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

y gracias a las hipótesis de inducción, puede escribirse como:

$$\left(\frac{\lambda_n}{\mu_n} + 1\right) p_n = p_n + \frac{\mu_{n+1}p_{n+1}}{\mu_n},$$

Entonces,

$$\frac{\lambda_n}{\mu_n} p_n = \frac{\mu_{n+1}p_{n+1}}{\mu_n},$$

si simplificamos μ_n y despejamos p_{n+1}, llegamos a lo que se quería demostrar,

$$p_{n+1} = \frac{\lambda_n}{\mu_{n+1}} p_n.$$

De esta forma se tiene

$$p_1 = \frac{\lambda_0}{\mu_1} p_0$$

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

y en general

$$p_n = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} p_0, \ \forall n = 1, 2, \ldots$$

Si denotamos al cociente como c_n, entonces

$$c_n = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1}, \ \forall n = 1, 2, \ldots$$

se tiene que $p_n = c_n p_0$. Ahora, como las p_n son probabilidades, se puede verificar que

$$\begin{aligned} \sum_{n=0}^{\infty} p_n &= 1 \\ &= p_0 + \sum_{n=1}^{\infty} p_n \\ &= p_0 + \sum_{n=1}^{\infty} c_n p_0 \\ 1 &= \left(1 + \sum_{n=1}^{\infty} c_n\right) p_0 \end{aligned}$$

Por tanto podemos calcular p_0, de la siguiente manera:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} c_n},$$

Ahora, siempre que $\sum_{n=1}^{\infty} c_n < \infty$. Esta última expresión es llamada condición de estado estacionario, esto viene a significar que la existencia de un proceso de nacimiento y muerte estacionario con tasas de nacimiento $\{\lambda_n\}_{n \geq 0}$ y tasas de muerte $\{\mu_n\}_{n \geq 1}$ debe de cumplirse siempre que $\sum_{n=1}^{\infty} c_n < \infty$.

1.4. El modelo de colas determinista D/D/1.

Los problemas más simples en teoría de colas son aquellos que no necesitan de distribuciones de probabilidad para ser descritos. Este es el caso de la cola deterministica $D/D/1$, en la que las llegadas y los tiempos de servicio son constantes y existe un único servidor.
Dado el cliente que hace el número k de los que han entrado en el sistema definimos:

t_k = Instante de llegada del cliente k-ésimo.

τ_k = Tiempo trancurrido entre dos llegadas $(t_k - t_{k-1})$.

w_k = Tiempo de espera en cola.

s_k = Tiempo de servicio.

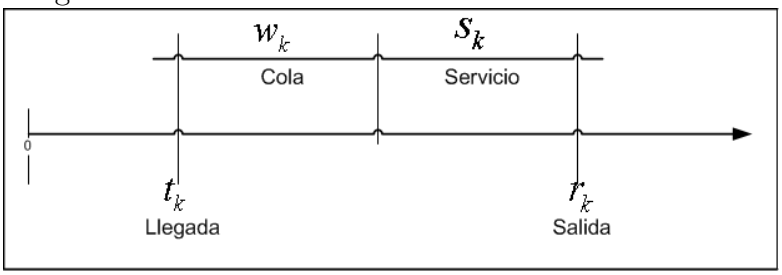
r_k = Tiempo de partida.

De lo anterior podemos verificar claramente que:

$$r_k = t_k + w_k + s_k \tag{1.1}$$

Lo cual podemos ver claramente en la figura 1.1.

Figura 1.1: Ilustración de la salida del k-ésimo cliente.



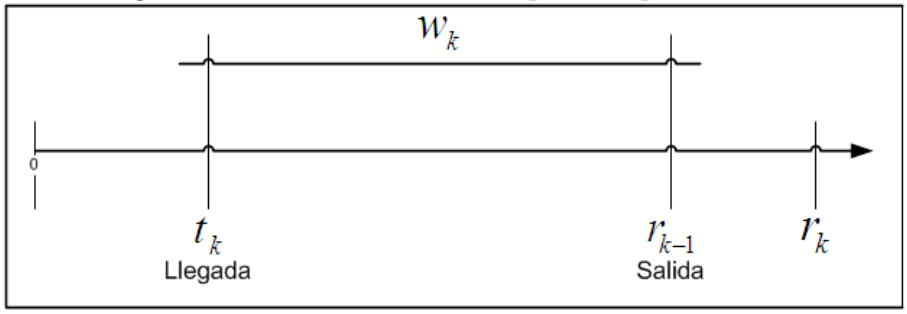
Si el sistema tiene un único canal $c = 1$, se tiene:

$$w_k = \begin{cases} r_{k-1} - t_k & , n_k > 0 \\ 0 & , n_k = 0 \end{cases} \tag{1.2}$$

donde n_k es el número de clientes en el sistema cuando llega el cliente k-ésimo, es decir si $n_k = 0$ no hay clientes en el sistema, el tiempo de espera para el k-ésimo cliente será 0, de lo contrario el tiempo en cola del k-ésimo será igual al tiempo que tarde el cliente $k-1$ en salir del sistema menos el instante de llegada del k-ésimo cliente.

Una observación importante es que conocidos los t_k y los s_k, podemos conocer los w_k por un proceso que reconstruye la historia de la cola, el cual veremos en el transcurso de esta sección. Supondremos que en el instante inicial $t = 0$ hay i clientes en cola. Si los clientes llegan a un único canal, a intervalos regulares de tiempo de longitud b (Tasa de servicio $1/b$), las variables que definen el problema son:

Figura 1.2: Ilustración del tiempo de espera en cola.



$\tau_k = a$
$s_k = b, \forall k$
$c = 1$

$$t_k = \begin{cases} 0 & ; k \leq i \\ (k-i)a & ; k > i \end{cases}$$

En primer lugar determinaremos el tiempo de espera en cola.

Proposición 1.4.1. *Si se tienen i clientes en cola cuando llega el cliente k-ésimo, y si el tiempo entre llegadas de dos clientes consecutivos es a y el tiempo de servicio de dichos clientes es b, entonces el tiempo en cola del cliente k-ésimo es:*

$$w_k = \begin{cases} (k-1)b & ; k \leq i \\ 0 & ; k > i, n_k = 0 \\ (k-1)b - (k-i)a & ; k > i, n_k > 0 \end{cases}$$

Demostración.
Sabemos que si $n_k > 0$ entonces $w_k = r_{k-1} - t_k$, siendo n_k el número de clientes en el sistema cuando llega el cliente k-ésimo, t_k el instante de llegada del cliente k y r_k el tiempo de partida del cliente $(k-1)$-ésimo.
Si $k \leq i$

$$r_{k-1} = \sum_{j=1}^{k-1} s_j = \sum_{j=1}^{k-1} b = (k-1)b, \ t_k = 0 \tag{1.3}$$

Si $k > i$

$$r_{k-1} = \sum_{j=1}^{k-1} s_j = \sum_{j=1}^{k-1} b = (k-1)b, \ t_k = (k-i)a \tag{1.4}$$

Proposición 1.4.2. *La relación entre los tiempos de espera en cola de los clientes k-ésimo y* $(k-1)$*-ésimo viene dada por:*

$$w_k = w_{k-1} + b - a,\ n_k > 0 \tag{1.5}$$

Demostración.

$$\begin{aligned} w_k &= r_{k-1} - t_k \\ &= (t_{k-1} + w_{k-1} + s_{k-1}) - t_k \\ &= -(t_k - t_{k-1}) + s_{k-1} + w_{k-1} \\ &= w_{k-1} + b - a \end{aligned} \tag{1.6}$$

ya que:

$$\begin{aligned} a &= t_k - t_{k-1} \\ b &= s_{k-1} \end{aligned}$$

□

A partir de aquí distinguiremos dos casos:

- **Primer caso** $b > a$**.**
- **Segundo caso** $b < a$**.**

Si $b = a$, no habrá espera si el sistema comienza con una cola vacía, si no es así, la cola se mantendrá con longitud constante.

Primer caso $b > a$**.**

En este caso el tiempo de espera es cada vez mayor y la cola crece indefinidamente.

Corolario 1.4.1. *El tiempo de espera en cola cuando el tiempo de servicio es mayor que el tiempo entre llegadas viene dado por:*

$$w_k = \begin{cases} (k-1)b & ;k \leq i \\ (k-1)(b-a) + (i-1)a & ;k > i \end{cases} \tag{1.7}$$

donde $(k-1)(b-a)$ *crece si* k *crece.*

□

Analicemos las variables asociadas a un instante [1].

Proposición 1.4.3. *Sean X, Y y Z variables asociadas a un instante t.*

i) $X(t) = i + \left[\frac{t}{a}\right]$ = *Número de clientes que han llegado al sistema hasta el instante t.*

ii) $Y(t) = \left[\frac{t}{b}\right] + 1$ = *Número de clientes que han entrado en servicio hasta el instante t.*

iii) $Z(t) = \left[\frac{t}{b}\right]$ = *Número de clientes que han salido del sistema hasta el instante t.*

Demostración.

i) El número de clientes que han entrado en el sistema hasta el instante t, será igual a los i que hay al principio, más los que han llegado hasta el instante t sabiendo que lo hacen a intervalos de tiempo de longitud constante a.

ii) El número de clientes que han entrado en servicio hasta el instante t será igual a la tasa de servicio $\left(\frac{1}{b}\right)$ por el tiempo t más el último cliente que ha entrado en servicio y todavía no ha salido.

iii) El número de clientes que han salido del sistema hasta el instante t será igual a los que han sido servidos $Y(t)$ menos el último que entró y está siendo servido: $Z(t) = Y(t) - 1$.

□

Proposición 1.4.4. *El número de clientes que se encuentran en el sistema en el instante t viene dado por la siguiente expresión:*

$$N(t) = i + \left[\frac{t}{a}\right] - \left[\frac{t}{b}\right], \tag{1.8}$$

que crece con t.

Demostración.

$$\begin{aligned} N(t) &= X(t) - Z(t) \\ &= i + \left[\frac{t}{a}\right] - \left[\frac{t}{b}\right] \end{aligned}$$

[1] Los corchetes $[\cdot]$ indican el mayor entero que es menor que la fracción.

Donde $N(t)$ es el número de clientes en el sistema en ese momento.

□

Segundo caso $b < a$.

En este caso el servicio es más rápido que la afluencia de clientes a la cola, con lo cual el tiempo de espera w_k, irá disminuyendo hasta hacerse $w_k = 0$. A partir de ese instante los clientes no tendrán que esperar en cola; es decir serán servidos directamente.

Proposición 1.4.5. *El último cliente tal que $w_k > 0$ (es decir, tiene que esperar en cola) viene dado por:*

$$K = \begin{cases} \left[\dfrac{ia-b}{a-b}\right] & ; ai - b \neq m(a-b),\ m \in \mathbb{N} \\ \left[\dfrac{ia-b}{a-b}\right] - 1 & ; caso\ contrario \end{cases} \tag{1.9}$$

Donde K es el k-ésimo cliente que tiene que esperar.

Demostración.

Retomando el resultado obtenido de la proposición 1.4.1 para la primera ecuación en (1.9), tenemos:

$$w_k = (k-1)b - (k-i)a > 0 \tag{1.10}$$

$$(k-1)b - (k-i)a > 0$$

$$kb - b - ka + ia > 0$$

$$k(b-a) - b + ia > 0$$

$$k(a-b) + b - ia < 0$$

$$k(a-b) < ai - b$$

$$k < \frac{ai-b}{a-b} \tag{1.11}$$

Tomando ahora:

$$w_{k+1} = kb - (k+1-i)a \leq 0 \tag{1.12}$$

$$kb - (k+1-i)a \leq 0$$

$$kb - ka - a + ia \leq 0$$

$$k(b-a)+(i-1)a \leq 0$$

$$k(a-b)-(i-1)a \geq 0$$

$$k(a-b) \geq a(i-1)$$

$$\begin{aligned} k \geq \frac{a(i-1)}{a-b} &= \frac{ai-a+b-b}{a-b} \\ &= \frac{ai-b}{a-b}-\frac{a-b}{a-b} \\ &= \frac{ai-b}{a-b}-1 \end{aligned}$$

$$k \geq \frac{ai-b}{a-b}-1 \tag{1.13}$$

Así pues, $K=\left[\frac{ai-b}{a-b}\right]$ *si* $ai-b \neq m(a-b)$ *con* $m \in \{1,2,3,...\}$
Si $ai-b=m(a-b)$ *con* $m \in \{1,2,3,...\}$ *entonces* $K=\left[\frac{ai-b}{a-b}\right]-1$

□

Corolario 1.4.2. *El tiempo de espera en cola será:*

$$w_k=\begin{cases} (k-1)b & ;k \leq i \\ (k-1)b-(k-i)a & ;i<k \leq K \\ 0 & ;k>K \end{cases} \tag{1.14}$$

□

A continuación calcularemos el instante T, en que desaparece la cola.

Proposición 1.4.6. *Así pues el instante en que la cola desaparece está dado por la siguiente expresión:*

$$T=(K-1)b=\left[\frac{ia-b}{a-b}-1\right] \; ;ai-b \neq m(a-b),\ m \in \mathbb{N} \tag{1.15}$$

Demostración.

$$\begin{aligned} T &= t_K+w_K \\ &= (K-i)a+(K-1)b-(K-i)a=(K-1)b \\ &= \left[\frac{ia-b}{a-b}-1\right]b \end{aligned}$$

□

Las variables asociadas a un instante verifican:

Proposición 1.4.7. *El número de clientes que han entrado al sistema, que estan en servicio y que han salido de él hasta el instante t son respectivamente:*

$$X(t) = i + \left[\frac{t}{a}\right] \tag{1.16}$$

$$Y(t) = \begin{cases} \left[\frac{t}{b}\right] + 1 & ; t \leq T \\ K + \left[\frac{t - t_k}{a}\right] & ; t > T \end{cases} \tag{1.17}$$

$$Z(t) = \begin{cases} \left[\frac{t}{b}\right] & ; t \leq T \\ K + \left[\frac{t - t_k}{a}\right] - 1 & ; t > T, t \in [t_k + ja, t_k + ja + b) \\ K + \left[\frac{t - t_k}{a}\right] & ; t > T, t \in [t_k + ja + b, t_k + (j+1)a) \end{cases} \tag{1.18}$$

Donde $j = 0, 1, 2, \ldots$

Demostración.

En el caso de la ecuación (1.16) se resuelve como en el caso i) de la proposición 1.4.3.

Si $t < T$*, (1.17) y (1.18) se razona de igual forma que la proposición 1.4.3, pero si* $t > T$*, razonamos sobre un ejemplo: Sea* $t > T$ *y* $K = 5 \Rightarrow T = (K-1)b = 4b$

En $T = 4b$ *han salido* $K - 1 = 4$ *clientes y el* K*-ésimo cliente entra en servicio. A partir de* T *los clientes entran en servicio según llegan y el sistema permanece vacío hasta que llega el siguiente cliente.*

Así pues, para $t > T$*, habrá entrado en servicio* $Y(t) = K + \left[\frac{t-t_k}{a}\right] = 5 + 3 = 8$*.*

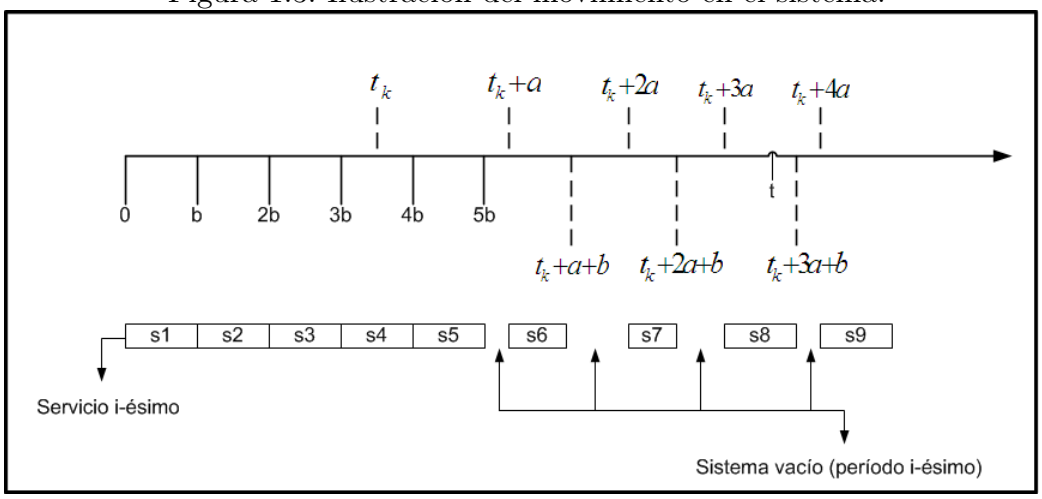
Además:

si $t \in [t_k + ja, t_k + ja + b) \Rightarrow Z(t) = Y(t) - 1 = 7$*, en el ejemplo.*

si $t \in [t_k + ja + b, t_k + (j+1)a) \Rightarrow Z(t) = Y(t)$*.*

Es decir, si t está en un período vacío no hay que restar 1 porque no hay nadie en servicio.

Figura 1.3: Ilustración del movimiento en el sistema.



□

Ejemplo 1.4.1. *En una cierta fábrica se encargan de la elaboración de piezas para carros, se trabaja en especial con una pieza de motor la cual es llevada por una banda transportadora hasta donde se le aplica su capa de revestimiento de un material especial; dicha banda transportadora recibe las piezas cada 2 minutos y siempre tiene que esperar 5 piezas antes de pasar al proceso de revestimiento, el cual tarda 8 minutos en ser acabado. Describir cuantitativamente el sistema cuando llega la sexta pieza en el tiempo* $t = 7$.

Solución:

Los datos del problema son los siguientes: Estamos en el caso en que $b > a$.

$$a = 2 \text{ minutos}$$

$$b = 8 \text{ minutos}$$

$$i = 5$$

Calculamos primero el tiempo medio en cola para la sexta pieza:

$$w_6 = (6-1)(8-2) + (5-1)2 = 30 + 8 = 38min.$$

Calculamos ahora el número de clientes que han llegado al sistema hasta el instante $t = 7$

$$X(7) = 5 + \left[\frac{7}{2}\right] = 5 + 3 = 8$$

El número de clientes que han entrado en servicio hasta el instante $t = 7$

$$Y(7) = \left[\frac{7}{8}\right] + 1 = 0 + 1 = 1$$

El número de clientes que han salido del sistema hasta $t = 7$

$$Z(7) = \left[\frac{7}{8}\right] = 0$$

El número de clientes en el sistema en el instante $t = 7$ *será*

$$N(7) = X(7) - Z(7) = 8 - 0 = 8$$

1.5. El modelo M/M/1

Este modelo indica que el tiempo entre llegadas consecutivas de clientes al sistema es una distribución exponencial con parámetro λ, independiente del número de clientes que haya dentro del sistema, además que los tiempos entre servicio de los clientes también están distribuidos exponencialmente con parámetros μ y un solo servidor. Los valores de los últimos tres parámetros según la notación Kendall son por defecto los siguientes:

- No hay restricciones respecto al número de clientes en cola.
- La población potencial es infinita.
- La disciplina de la cola es $FIFO$.

por lo tanto es un sistema de espera de un solo recurso y denotamos por λ la tasa de llegadas al sistema y por μ la velocidad de servicio del recurso.
Las tasas de llegada y de servicio son:

$$\lambda_n = \lambda, \forall n = 0, 1, 2, 3, ...$$

$$\mu_n = \mu, \forall n = 1, 2, 3, ...$$

Nos interesa evaluar el sistema en régimen permanente, así que calculamos la distribución de probabilidad. Esta distribución se obtiene sustituyendo la tasa de llegadas y de servicios en la función:

$$p_n = \frac{\lambda_{n-1}}{\mu_n} p_{n-1} \tag{1.19}$$

viene dada de la distribución de probabilidades en un proceso de nacimiento y muerte, al hacer iteraciones de p_n resulta:

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \tag{1.20}$$

Además, siendo el sistema estable, p_n es una distribución de probabilidad, es decir se satisface que:

$$\sum_{n=0}^{\infty} p_n = 1 \tag{1.21}$$

Sea:

$$c_n = \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \tag{1.22}$$

dado la igualdad de los parámetros se tiene:

$$c_n = \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = \frac{\lambda^n}{\mu^n} = \rho^n \tag{1.23}$$

Como $\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} \rho^n$ es una serie geométrica, será convergente si y sólo si $|\rho| < 1$, es decir; que $0 < \rho < 1$. Esta condición $0 < \rho < 1$, es por tanto equivalente a que el modelo sea estacionario. Otra forma de expresarla es $\lambda < \mu$, que tiene la interpretación adicional de que el número medio de clientes que entran en el sistema por unidad de tiempo sea menor que el número medio de clientes que podrían ser atendidos por el servidor por unidad de tiempo, en caso de que éste estuviese absolutamente todo el tiempo atendiendo a clientes (cosa que no ocurre siempre).

En lo que sigue supondremos que el sistema de la cola es estacionario (es decir; que $\rho < 1$). Lo primero que debemos calcular es la suma de la serie c_n:

$$\begin{aligned}
\sum_{n=1}^{\infty} c_n &= \sum_{n=1}^{\infty} \rho^n \\
\sum_{n=1}^{\infty} \rho^n &= \rho + \rho^2 + \rho^3 + \rho^4 + \cdots \\
&= \rho\left(1 + \rho + \rho^2 + \rho^3 + \rho^4 + \cdots\right) \\
\sum_{n=1}^{\infty} \rho^n &= \rho\left(1 + \sum_{n=1}^{\infty} \rho^n\right)
\end{aligned}$$

$$\begin{aligned}\sum_{n=1}^{\infty}\rho^n - \rho\sum_{n=1}^{\infty}\rho^n &= \rho \\ (1-\rho)\sum_{n=1}^{\infty}\rho^n &= \rho \\ \sum_{n=1}^{\infty}\rho^n &= \frac{\rho}{(1-\rho)}\end{aligned} \tag{1.24}$$

Así de las ecuaciones anteriores podemos obtener el valor de p_0, donde:

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

$$p_0 = \begin{cases} 0 & ,\rho > 1 \Rightarrow p_n = 0, \forall n \text{ caso desbordado} \\ 1-\rho & ,0 < \rho < 1 \Rightarrow p_n = \rho^n(1-\rho), \forall\ n \end{cases} \tag{1.25}$$

En realidad, para que $0 < \rho < 1$, lo que ha ocurrido es que $\lambda < \mu$ es decir, que el tiempo medio de llegadas sea menor que el tiempo medio de servicios. Esto es razonable si pensamos que para que exista una distribución estacionaria la longitud de la cola no puede tender a infinito, deduzcamos entonces p_0, para ello tenemos que:

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty} p_n = p_0 + \sum_{n=1}^{\infty} p_n \\ &= p_0 + \sum_{n=1}^{\infty} p_0\rho^n = p_0(1 + \sum_{n=1}^{\infty}\rho^n) \\ p_0 &= \frac{1}{1+\sum_{n=1}^{\infty}\rho^n} = \frac{1}{1+\frac{\rho}{1-\rho}} \\ p_0 &= \frac{1}{\frac{1-\rho+\rho}{1-\rho}} \\ p_0 &= 1-\rho \end{aligned} \tag{1.26}$$

De aquí se tiene la función de probabilidad:

$$\begin{aligned} p_n &= p_0\rho^n \\ p_n &= (1-\rho)\rho^n \end{aligned}$$

Resulta interesante observar que la probabilidad de encontrar k o más clientes en el sistema es[2]:

$$\begin{aligned} P(N \geq k) &= \sum_{i=k}^{\infty} p_i \\ &= \sum_{i=k}^{\infty} (1-\rho)\rho^i \\ &= (1-\rho)\sum_{i=k}^{\infty} \rho^i \\ &= (1-\rho)(\rho^k + \rho^{k+1} + \rho^{k+2} + \cdots) \\ &= (1-\rho)\rho^k(1 + \rho + \rho^2 + \cdots) \\ &= (\rho^k - \rho^{k+1})(1 + \rho + \rho^2 + \cdots) \\ &= \rho^k + \rho^{k+1} + \rho^{k+2} + \cdots - \rho^{k+1} - \rho^{k+2} \\ &= \rho^k \end{aligned} \tag{1.27}$$

Una vez calculada la distribución de probabilidad del sistema, calcularemos los parámetros de interés desde el punto de vista del usuario:

1. **Número medio de clientes en el sistema:**

$$\begin{aligned} L = E(N) &= \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\ &= (1-\rho)\sum_{n=0}^{\infty} n\rho^n \\ &= (1-\rho)\rho\sum_{n=0}^{\infty} n\rho^{n-1}; \quad \frac{d}{d\rho}(\rho^n) = n\rho^{n-1} \\ &= (1-\rho)\rho\sum_{n=0}^{\infty} \frac{d\rho^n}{d\rho} \\ &= (1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right) \\ &= (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) \end{aligned}$$

[2]Está probabilidad proporciona una cota superior de la probabilidad de desbordamiento en el sistema equivalente al tamaño k correspondiente a $p_k < \rho^k$.

$$\begin{aligned} L &= (1-\rho)\rho\frac{1}{(1-\rho)^2} \\ L &= \frac{\rho}{1-\rho} \end{aligned} \tag{1.28}$$

2. **Número medio de clientes en cola:**

Puede obtenerse a partir de N_q de la siguiente manera:

$$\begin{aligned} L_q = E[N_q] &= \sum_{n=1}^{\infty}(n-1)p_n \\ &= \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n = \sum_{n=0}^{\infty} np_n - (1-p_0) \\ &= L - \rho = \frac{\rho}{1-\rho} - \rho \\ L_q &= \frac{\rho^2}{1-\rho} \end{aligned} \tag{1.29}$$

3. **Tiempos medio de espera del cliente en el sistema:**

Aplicando directamente la fórmula de Little tenemos:

$$W = \frac{L}{\lambda} = \frac{\frac{\rho}{1-\rho}}{\lambda} = \frac{\frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}}{\lambda} = \left(\frac{1}{\mu-\lambda}\right) \tag{1.30}$$

4. **Tiempo medio de clientes en cola:**

Aplicando nuevamente la fórmula de Little encontramos:

$$W_q = \frac{L_q}{\lambda} = \frac{\frac{\rho^2}{1-\rho}}{\lambda} = \frac{\frac{(\frac{\lambda}{\mu})^2}{1-\frac{\lambda}{\mu}}}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{1.31}$$

Además, puede comprobarse fácilmente que se verifica la relación $W = W_q + \frac{1}{\mu}$. De hecho esta relación junto con las dos fórmulas de Little permite calcular el valor de cualesquiera tres de las cantidades L, L_q, W y W_q, dada la cuarta.
Si se desea tener más información sobre la espera de clientes en la cola o en el sistema, debe calcularse la distribución de probabilidad de las variables $\mathbf{W}$ y $\mathbf{W}_q$. Estas distribuciones permitirán calcular la probabilidad de cualquier suceso relativo al tiempo de estancia en la cola o en el sistema.

Primeramente abordaremos el cálculo de la función de distribución de $\mathbf{W}$, que denotaremos por $\mathbf{W(t)}$. Para ello aplicamos la regla de las probabilidades totales condicionando al número, N, de clientes que hay en el sistema cuando llega el cliente en cuestión y tenemos en cuenta que $\mathbf{W}|_{N=n} \stackrel{d}{=} \Gamma(\mu, n+1)$. De esta forma, para cada $t \geq 0$, se tiene

$$\begin{aligned}
\mathbf{W(t)} &= P[\mathbf{W} \leq t] = \sum_{n=0}^{\infty} P[\mathbf{W} \leq t|_{N=n}]P[N=n] \\
&= \sum_{n=0}^{\infty} \left(\int_0^t \frac{\mu^{n+1}}{n!} x^n e^{-\mu x} dx \right) p_n \\
&= \sum_{n=0}^{\infty} \left(\int_0^t \frac{\mu^{n+1}}{n!} x^n e^{-\mu x} dx \right) \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n \\
&= \int_0^t \mu \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu x} \left[\sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} \right] dx \\
&= \int_0^t \mu \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu x} e^{\lambda x} dx \\
&= \int_0^t (\mu - \lambda) e^{-(\mu-\lambda)x} dx \\
&= \left[-e^{-(\mu-\lambda)x} \right]_{x=0}^{x=t} \\
&= 1 - e^{-(\mu-\lambda)t}
\end{aligned} \tag{1.32}$$

que es la función de distribución de una exponencial de parámetro $\mu - \lambda$. Así pues,

$$\mathbf{W} \stackrel{d}{=} exp(\mu - \lambda)$$

Es obvio, por tanto, que como conclusión de esto también se puede volver a obtener $\mathbf{W} = \dfrac{1}{\mu - \lambda}$.
Dado que la distribución del tiempo de servicio para el sistema $M/M/1$ es exponencial y esta no tiene memoria, entonces el valor del tiempo que le queda por servir al cliente que está siendo atendido tendrá la misma distribución que el tiempo de servicio demandado.
Asi $\mathbf{W}_q$ está condicionado a que hayan n tareas en el sistema por lo tanto, es una suma de n variables exponenciales independientes e idénticamente distribuidas; es decir es una variable aleatoria $Erlang - n$ cuya función de distribución es:

$$\mathbf{W}_q|_n(t) = 1 - e^{-\mu t} \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!}, \quad \forall t \geq 0$$

Así la distribución de probabilidad marginal de $\mathbf{W}_q$ vendrá dada por:

$$\begin{aligned}
\mathbf{W}_q(t) &= P[\mathbf{W}_q \leq t] \\
&= \sum_{n=0}^{\infty} p_n P[\mathbf{W}_q \leq t|_{N=n}] \\
&= \sum_{n=0}^{\infty} p_n F_{\mathbf{W}_q|_n}(t)] \\
&= \sum_{n=0}^{\infty} p_n \left[1 - e^{-\mu t} \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!}\right] \\
&= p_0 \mathbf{W}_q|_0(t) + \sum_{n=1}^{\infty} p_n \left[1 - e^{-\mu t} \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!}\right] \\
&= (1-\rho) + \sum_{n=1}^{\infty} p_n - e^{-\mu t} \sum_{n=1}^{\infty} p_n \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!} \\
&= \sum_{n=0}^{\infty} p_n - e^{-\mu t} \sum_{n=1}^{\infty} (1-\rho)\rho^n \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!} \\
&= 1 - (1-\rho) e^{-\mu t} \sum_{n=1}^{\infty} \rho^n \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!} \qquad (1.33)
\end{aligned}$$

donde:

$$\sum_{n=1}^{\infty} \rho^n \sum_{j=0}^{n-1} \frac{(\mu t)^j}{j!} = \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \sum_{i=k+1}^{\infty} \rho^i$$

asi pues

$$\begin{aligned}
\sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \sum_{i=k+1}^{\infty} \rho^i &= \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \left(\rho^{k+1} + \rho^{k+2} + \cdots\right) \\
&= \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \rho^k \left(\rho + \rho^2 + \rho^3 + \cdots\right) \\
&= \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \rho^k \left[\sum_{i=1}^{\infty} \rho^i\right] \\
&= \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \rho^k \left[\frac{\rho}{1-\rho}\right]
\end{aligned}$$

$$= \left[\frac{\rho}{1-\rho}\right] \sum_{k=0}^{\infty} \frac{(\mu t \rho)^k}{k!}$$

$$\sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \sum_{i=k+1}^{\infty} \rho^i = \left[\frac{\rho}{1-\rho}\right] e^{\mu t \rho} \tag{1.34}$$

volviendo a (1.33) utilizando el resultado de (1.34) obtenemos:

$$\begin{aligned} \mathbf{W}_q(t) &= 1 - (1-\rho)e^{-\mu t}\left[\frac{\rho}{1-\rho}\right] e^{\mu t \rho} \\ &= 1 - \rho e^{-\mu t(1-\rho)} \\ &= 1 - \frac{\lambda}{\mu} e^{-t(\mu-\lambda)} \end{aligned} \tag{1.35}$$

De está forma se obtiene:

$$\mathbf{W}_q(t) = \begin{cases} 1 - \dfrac{\lambda}{\mu} e^{-(\mu-\lambda)t} & , t \geq 0 \\ 0 & , t \leq 0 \end{cases} \tag{1.36}$$

Esta función de distribución es discontinua en $t = 0$, valiendo su salto

$$\mathbf{W}_q(0) - \mathbf{W}_q(0^-) = 1 - \frac{\lambda}{\mu} = 1 - \rho = p_0$$

Se trata por tanto de una variable que es mezcla de continua y discreta; toma el valor de 0 con probabilidad p_0 y para todos los $t > 0$ tiene una componente continua con función de subdensidad dada por: $\dfrac{\lambda(\mu-\lambda)}{\mu} e^{-(\mu-\lambda)t}$

Ejemplo 1.5.1. *Al supercomputador de un centro de cálculo llegan usuarios según un proceso de Poisson de 5 usuarios cada hora. Sabiendo que éstos consumen un tiempo de cómputo aleatorio cuya distribución puede suponerse exponencial de media 1/6 de hora y que la disciplina de atención es FIFO. Se pide:*

1. *Calcular L y L_q.*

$$\rho = \frac{\lambda}{\mu} = \frac{5}{6} \approx 0.8333.$$

$$L = \frac{\rho}{1-\rho} = \frac{\frac{5}{6}}{1-\frac{5}{6}} = 5 \ usuarios$$

$$L_q = \frac{\rho^2}{1-\rho} = \frac{\frac{25}{36}}{\frac{1}{6}} = \frac{25}{6} \approx 4.16 \ usuarios$$

2. *Obtener el tamaño medio de la cola sabiendo que hay gente esperando. Puede formularse matemáticamente de la siguiente manera:*

$$L'_q = E[L_q|L_q > 0]$$

Sabemos que se satisface:

$$E[L_q] = E[L_q|L_q = 0]P[L_q = 0] + E[L_q|L_q > 0]P[L_q > 0]$$

donde, obviamente, $E[L_q|L_q = 0] = 0$ *por lo tanto la expresión puede escribirse como:*

$$E[L_q|L_q > 0] = \frac{E[L_q]}{P[L_q > 0]} \tag{1.37}$$

Sabiendo que:

$$P[L_q > 0] = P[N \geq 2] = 1 - P[N < 2] = 1 - p_0 - p_1,$$

y sustituyendo en la ecuación (1.37), de la ecuación (1.29) se tiene como solución:

$$E[L_q|L_q > 0] = \frac{\frac{\rho^2}{1-\rho}}{1 - p_0 - p_1} = \frac{1}{1-\rho} = 6 \text{ usuarios}$$

3. *¿Que porcentaje de usuarios llega al sistema y lo encuentra desocupado?*
La probabilidad de que el sistema esté ocupado viene determinado por el factor de ocupación, así p_0 *es la probabilidad de espera nula.*

$$p_0 = 1 - \rho = 1 - \frac{5}{6} \approx 0.1666 \approx 16.7\,\%$$

4. *Si en la sala de espera hay 4 sillas, ¿cuál es la probabilidad de que un usuario tenga que espera de pie? la probabilidad de que un usuario tenga que esperar de pie es igual de que hayan más usuarios que sillas* $N \geq 5$

$$P[N \geq 5] = 1 - P[N < 5]$$

también se puede hacer uso de la ecuación (1.27)con la cual se obtiene que:

$$\begin{aligned} P[L \geq 5] &= \sum_{n=5}^{\infty} p_n \\ &= \sum_{n=5}^{\infty} (1-\rho)\rho^n \\ &= \rho^5 \approx 0.4018 \end{aligned}$$

□

1.6. El modelo M/M/1/k

Se trata de un modelo como el M/M/1, ya estudiado, pero con limitación k para el tamaño de la cola. Donde, la distribución del tiempo entre dos intentos de llegadas al sistema de clientes consecutivos es una $exp(\lambda)$, la distribución del tiempo de servicio es $exp(\mu)$ y sólo hay un servidor. Además el número de clientes que pueden estar en la cola es como mucho k, la población potencial es infinita y la disciplina es FIFO. Obviamente, en este modelo se puede dar el caso de que un cliente que intente entrar en el sistema no lo consiga, por estar la cola llena. A partir de las especificaciones anteriores se deducen fácilmente las tasas de llegada:

$$\lambda_n = \begin{cases} \lambda & , n = 0, 1, 2, 3, 4, \ldots, k \\ 0 & , n = k+1, k+2, \ldots \end{cases} \tag{1.38}$$

mientras que las tasas de servicio son idénticas a las de un M/M/1,

$$\mu_n = \mu, \forall n = 1, 2, \ldots$$

Haciendo uso de λ_n y μ_n se obtienen inmediatamente los c_n:

$$c_n = \begin{cases} \rho^n & ; si\ n = 0, 1, 2, 3, 4, ...k+1 \\ 0 & ; si\ n = k+2, k+3, ... \end{cases} \tag{1.39}$$

Dado que la serie $\sum_{n=1}^{\infty} c_n$ tiene tan sólo un número finito de términos distintos de cero, es trivialmente convergente sin ninguna condición acerca de ρ. Esto puede interpretarse que por muy frecuente que sea la llegada de clientes al sistema en relación con la capacidad del servidor para dar servicio, la propia limitación en el tamaño de la cola fuerza a la estacionariedad, pues lo peor que podríamos imaginar es que prácticamente todo el tiempo estuviera el sistema saturado (es decir; $P(N = k+1) = 1$). La suma de la serie de los c_n es realmente una suma finita.

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{k+1} \rho^n = \begin{cases} \dfrac{\rho^{k+2} - \rho}{\rho - 1} & , \rho \neq 1 \\ k+1 & , \rho = 1 \end{cases} \tag{1.40}$$

En la ecuación (1.40), observamos que si el valor de $\rho = 1$ la expresión es inmediata, ya que ρ en la sumatoria sería constante y $k+1$ veces $\rho = 1$, ahora bien para el caso en que $\rho \neq 1$ vemos

lo siguiente:

$$\begin{aligned}
\sum_{n=1}^{k+1}\rho^n &= \rho+\rho^2+\cdots+\rho^{k+1}\\
\rho\sum_{n=1}^{k+1}\rho^n &= \rho(\rho+\rho^2+\cdots+\rho^{k+1})\\
\rho\sum_{n=1}^{k+1}\rho^n+\rho &= \rho+\rho^2+\rho^3+\cdots+\rho^{k+1}+\rho^{k+2}\\
&= \sum_{n=1}^{k+1}\rho^n+\rho^{k+2}\\
\sum_{n=1}^{k+1}\rho^n(\rho-1) &= \rho^{k+2}-\rho\\
\sum_{n=1}^{k+1}\rho^n &= \frac{\rho^{k+2}-\rho}{(\rho-1)} \qquad (1.41)
\end{aligned}$$

Esta distinción, $\rho\neq 1$ ó $\rho=1$ habrá que hacerla a lo largo de todos los cálculos sucesivos.

Caso $\rho\neq 1$: En primer lugar calculamos p_0:

$$p_0=\frac{1}{1+\sum_{n=1}^{\infty}c_n}=\frac{1}{1+\frac{\rho^{k+2}-\rho}{\rho-1}}=\frac{\rho-1}{\rho^{k+2}-1} \qquad (1.42)$$

$$p_n=\begin{cases}\dfrac{\rho-1}{\rho^{k+2}-1}, & n=0,1,2\ldots,k+1\\[2ex] 0, & n=k+2,k+3,\ldots\end{cases} \qquad (1.43)$$

El número medio de clientes en el sistema puede calcularse a partir de su definición:

$$\begin{aligned}
L &= \sum_{n=0}^{\infty}np_n\\
&= \sum_{n=0}^{k+1}n\frac{\rho-1}{\rho^{k+2}-1}\rho^n\\
&= \frac{(\rho-1)\rho}{\rho^{k+2}-1}\sum_{n=0}^{k+1}n\rho^{n-1} \qquad (1.44)
\end{aligned}$$

Ahora bien, siguiendo las mismas pautas que las usadas en el M/M/1 en 1.24 para calcular la suma de una serie convertible en geométrica, podemos ahora calcular la suma de un número

finito de términos de la misma. Así, definiendo:

$$\begin{aligned}
\sum_{n=0}^{k+1}\rho^n &= \sum_{n=0}^{\infty}\rho^n - \sum_{n=k+2}^{\infty}\rho^n \\
&= 1+\frac{\rho}{1-\rho} - \sum_{n=k+2}^{\infty}\rho^n \\
&= \frac{1}{1-\rho} - (\rho^{k+2}+\rho^{k+3}+\cdots) \\
&= \frac{1}{1-\rho} - \rho^{k+2}(1+\rho+\rho^2+\cdots) \\
&= \frac{1}{1-\rho} - \rho^{k+2}\sum_{n=0}^{\infty}\rho^n \\
&= \frac{1}{1-\rho}(1-\rho^{k+2})
\end{aligned}$$

Además sabemos que:

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

$$\begin{aligned}
\sum_{n=0}^{k+1}\frac{d}{d\rho}(\rho^n) &= \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)(1-\rho^{k+2}) \\
&= \frac{d}{d\rho}\left(\frac{1-\rho^{k+2}}{1-\rho}\right) \\
&= \frac{-(k+2)\rho^{k+1}+(k+2)\rho^{k+2}+1-\rho^{k+2}}{(1-\rho)^2} \\
&= \frac{(k+2-1)\rho^{k+2}-(k+2)\rho^{k+1}+1}{(1-\rho)^2} \\
&= \frac{(k+1)\rho^{k+2}-(k+2)\rho^{k+1}+1}{(1-\rho)^2} \qquad (1.45)
\end{aligned}$$

Por lo tanto, de la expresión 1.45 podemos calcular L, de la siguiente forma:

$$\begin{aligned}
L &= \frac{(\rho-1)\rho}{\rho^{k+2}-1}\cdot\left[\frac{(k+1)\rho^{k+2}-(k+2)\rho^{k+1}+1}{(1-\rho)^2}\right] \\
&= \frac{(\rho-1)}{\rho^{k+2}-1}\cdot\left[\frac{(k+1)\rho^{k+3}-(k+2)\rho^{k+2}+\rho}{(1-\rho)^2}\right] \\
&= \frac{(\rho-1)}{\rho^{k+2}-1}\cdot\left[\frac{k\rho^{k+3}+\rho^{k+3}-(k+2)\rho^{k+2}+\rho+\rho^{k+3}-\rho^{k+3}}{(1-\rho)^2}\right]
\end{aligned}$$

$$
\begin{aligned}
L &= \frac{(\rho-1)}{\rho^{k+2}-1}\cdot\left[\frac{k\rho^{k+3}+2\rho^{k+3}-(k+2)\rho^{k+2}+\rho-\rho^{k+3}}{(1-\rho)^2}\right] \\
&= \frac{(\rho-1)}{\rho^{k+2}-1}\cdot\left[\frac{(k+2)(\rho^{k+3}-\rho^{k+2})+\rho(1-\rho^{k+2})}{(1-\rho)^2}\right] \\
&= \frac{(\rho-1)}{\rho^{k+2}-1}\cdot\left[\frac{\rho(1-\rho^{k+2})-(k+2)(\rho^{k+2}-\rho^{k+3})}{(1-\rho)^2}\right] \\
&= \frac{(1-\rho)}{1-\rho^{k+2}}\cdot\left[\frac{\rho(1-\rho^{k+2})}{(1-\rho)^2}-\frac{(k+2)(\rho^{k+2}-\rho^{k+3})}{(1-\rho)^2}\right] \\
&= \frac{\rho}{1-\rho}-\frac{(k+2)\rho^{k+2}(1-\rho)^2}{(1-\rho)^2(1-\rho^{k+2})} \\
L &= \frac{\rho}{1-\rho}-\frac{(k+2)\rho^{k+2}}{(1-\rho^{k+2})} \qquad (1.46)
\end{aligned}
$$

El primer sumando de L es precisamente L del modelo M/M/1. Las fórmulas de Little y la relación entre tiempos medios pueden usarse para calcular las otras tres cantidades medias de interés. para ello será necesario calcular $\overline{\lambda}$, ya que ahora las λ_n no son constantes:

$$
\begin{aligned}
\overline{\lambda} &= \sum_{n=0}^{\infty}\lambda\cdot p_n \\
&= \sum_{n=0}^{k}\lambda\cdot p_n \\
&= \lambda(1-p_{k+1});\quad p_n=p_0\rho^n \\
&= \lambda(1-\frac{(\rho-1)}{\rho^{k+2}-1}\rho^{k+1}) \\
&= \lambda\frac{\rho^{k+2}-1-(\rho-1)\rho^{k+1}}{\rho^{k+2}-1} \\
&= \frac{\lambda(\rho^{k+1}-1)}{\rho^{k+2}-1} \qquad (1.47)
\end{aligned}
$$

Apartir de esta expresión se tiene:

$$
\begin{aligned}
W &= \frac{L}{\overline{\lambda}} \\
&= \frac{\frac{(1-\rho)}{1-\rho^{k+2}}\cdot\frac{(k+1)\rho^{k+3}-(k+2)\rho^{k+2}+\rho}{(1-\rho)^2}}{\frac{\lambda(\rho^{k+1}-1)}{\rho^{k+2}-1}}
\end{aligned}
$$

$$
\begin{aligned}
W &= \frac{1-\rho}{1-\rho^{k+1}} \cdot \frac{(k+1)\rho^{k+3} - (k+2)\rho^{k+2} + \rho}{\lambda(1-\rho)^2} \\
&= \frac{1-\rho}{1-\rho^{k+1}} \cdot \frac{\rho - \rho^{k+2} - (k+1)(1-\rho)\rho^{k+2}}{\lambda(1-\rho)^2} \\
&= \frac{1-\rho}{1-\rho^{k+1}} \cdot \left[\frac{\rho(1-\rho^{k+1})}{\lambda(1-\rho)^2} - \frac{(k+1)(1-\rho)\rho^{k+2}}{\lambda(1-\rho)^2} \right] \\
&= \frac{\rho}{\lambda(1-\rho)} - \frac{(k+1)\rho^{k+2}}{\lambda(1-\rho^{k+1})} \\
W &= \frac{1}{\mu-\lambda} - \frac{(k+1)\rho^{k+2}}{\lambda(1-\rho^{k+1})} \qquad (1.48)
\end{aligned}
$$

A partir de esta última expersión podemos obtener W_q de la misma forma que en el modelo M/M/1

$$
\begin{aligned}
W_q &= W - \frac{1}{\mu} \\
&= \frac{1}{\mu-\lambda} - \frac{(k+1)\rho^{k+2}}{\lambda(1-\rho^{k+1})} - \frac{1}{\mu} \\
&= \frac{\lambda}{\mu(\mu-\lambda)} - \frac{(k+1)\rho^{k+2}}{\lambda(1-\rho^{k+1})} \qquad (1.49)
\end{aligned}
$$

De igual manera utilizando las fórmulas de Little se obtiene:

$$
\begin{aligned}
L_q &= \overline{\lambda} W_q \\
&= \overline{\lambda}\left(W - \frac{1}{\mu} \right) \\
&= L - \frac{\overline{\lambda}}{\mu}; \quad \text{De las expresiones 1.46 y 1.47} \\
&= \frac{\rho}{1-\rho} - \frac{(k+2)\rho^{k+2}}{1-\rho^{k+2}} - \frac{\frac{\lambda(\rho^{k+1}-1)}{\rho^{k+2}-1}}{\mu} \\
&= \frac{\rho}{1-\rho} - \frac{(k+2)\rho^{k+2}}{1-\rho^{k+2}} - \frac{\rho(1-\rho^{k+1})}{1-\rho^{k+2}} \\
&= \frac{\rho}{1-\rho} - \frac{(k+2)\rho^{k+2}}{1-\rho^{k+2}} - \frac{\rho - \rho^{k+2} + \rho^{k+3} - \rho^{k+3}}{1-\rho^{k+2}} \\
&= \frac{\rho}{1-\rho} - \frac{(k+2)\rho^{k+2}}{1-\rho^{k+2}} - \frac{\rho^{k+2}(\rho-1) + \rho(1-\rho^{k+2})}{1-\rho^{k+2}} \\
&= \frac{\rho}{1-\rho} - \rho - \frac{(k+1+\rho)\rho^{k+2}}{1-\rho^{k+2}} \\
&= \frac{\rho^2}{1-\rho} - \frac{(k+1+\rho)\rho^{k+2}}{1-\rho^{k+2}} \qquad (1.50)
\end{aligned}
$$

Esta última expresión, cuyo primer sumando es precisamente la fórmula para L_q en un modelo M/M/1. Obviamente, si el objetivo es calcular las cuatro cantidades (L, Lq, W y Wq) es más eficiente obtener el valor de $\overline{\lambda}$ y, una vez calculada una de las cuatro, obtener las tres restantes directamente de las fórmulas de Little y la relación entre tiempos medios.

Aunque el modelo $M/M/1/k$ no contiene al $M/M/1$ como caso particular, en el caso $\rho < 1$ (para el cual el $M/M/1$ es estacionario) el modelo $M/M/1/k$ debe tender al $M/M/1$ cuando $k \longrightarrow \lambda$ (que es tanto como decir que el tamaño máximo permitido para la cola es más y más grande).

En efecto, los resultados que ofrecen las fórmulas anteriores para los p_n, L, L_q, W y W_q coinciden con los que aparecen en el $M/M/1$. Así por ejemplo:

$$\begin{aligned} \lim_{k\longrightarrow\infty} L_{M/M/1/k} &= \frac{\rho}{1-\rho} - \lim_{k\longrightarrow\infty} \frac{(k+2)\rho^{k+2}}{1-\rho^{k+2}} \\ &= \frac{\rho}{1-\rho} \\ &= L_{M/M/1} \end{aligned} \tag{1.51}$$

haciendo uso de la regla de *L'Hôpital* queda que el límite del numerador es cero y por lo tanto el límite del modelo $M/M/1/k$ coincide con el de un modelo $M/M/1$.

CASO ρ =1: En este caso los resultados son fáciles de obtener y vienen dados de la siguiente manera:

$$p_0 = \frac{1}{1+\sum_{n=1}^{\infty} c_n} = \frac{1}{1+k+1} = \frac{1}{k+2} \tag{1.52}$$

tomando encuenta que $c_n = \rho^n = 1$, para $n = 1, 2, 3, \ldots, k+1$, entonces

$$p_n = \frac{1}{k+2}, \forall n \ = 0, 1, \ldots, k+1 \tag{1.53}$$

se puede decir que N tiene una distribución uniforme discreta sobre su conjunto de valores posibles, y en este caso,

- Número medio de clientes en el sistema:

$$\begin{aligned} L &= \sum_{n=0}^{k+1} np_n = \sum_{n=0}^{k+1} n\frac{1}{k+2} \\ &= \frac{1}{(k+2)}\frac{(k+2)(k+1)}{2} \\ &= \frac{k+1}{2} \end{aligned} \tag{1.54}$$

- Tiempo medio de llegada de los clientes:

$$\begin{aligned}\overline{\lambda} &= \sum_{n=0}^{\infty} \lambda p_n \\ &= \lambda \sum_{n=0}^{k} p_n = \lambda \left[p_0 + p_1 + p_2 + \cdots + p_k\right] \\ &= \lambda \left[\frac{1}{k+2} + \frac{1}{k+2} + \cdots + \frac{1}{k+2}\right] \\ &= \lambda \left[\frac{k+1}{k+2}\right] \end{aligned} \tag{1.55}$$

- Tiempo medio de clientes en el sistema:

$$W = \frac{L}{\overline{\lambda}} = \frac{\frac{k+1}{2}}{\frac{\lambda(k+1)}{k+2}} = \frac{k+2}{2\lambda} \tag{1.56}$$

- Tiempo medio de clientes en cola:

$$W_q = W - \frac{1}{\mu} = \frac{k+2}{2\lambda} - \frac{1}{\lambda} = \frac{k}{2\lambda} \tag{1.57}$$

- Número medio de clientes en cola:

$$L_q = \overline{\lambda} W_q = \frac{\lambda(k+1)}{(k+2)} \frac{k}{2\lambda} = \frac{k(k+1)}{2(k+2)} \tag{1.58}$$

Para cualquier valor de λ (igual o distinto de 1), si se desea obtener las funciones de distribución del tiempo que un cliente está en el sistema y del tiempo que un cliente está en la cola, es necesario previamente definir q_n, como la probabilidad de que haya n clientes en el sistema justo cuando una nueva llegada se esta produciendo. Denotando, como siempre, por N el número de clientes en el sistema y por T el tiempo que falta para que se produzca la llegada del siguiente cliente, las q_n representan $P(N = n|T = 0)$. Usando la notación $f(t|N = n)$ para la función de densidad de la variable T condicionada a que $N = n$, se sigue que $f(t|N = n)$ corresponde a la densidad de una $exp(\lambda_n)$.

Así, aplicando la regla de Bayes, se tiene:

$$\begin{aligned} q_n &= P(N=n|_{T=0}) \\ &= \frac{f(0|_{N=n})p_n}{\sum_{m=0}^{k} f(0|_{N=m})p_m} \\ &= \frac{\lambda_n e^{-\lambda_n 0} p_n}{\sum_{m=0}^{k} \lambda_m e^{-\lambda_m 0} p_m} \\ &= \frac{\lambda \cdot p_n}{\lambda \sum_{m=0}^{k} p_m} \\ &= \frac{p_n}{1-p_{k+1}} \end{aligned} \tag{1.59}$$

para $n = 0, 1, \ldots, k$; mientras que $q_n = 0$ para $n = k+1, k+2, \ldots$ para el caso del modelo $M/M/c$ (incluyendo el caso de $c = 1$) debe de verificarse que las variables T y N son independientes y, por lo tanto:

$$q_n = P(N = n|_{T=0}) = P(N = n) = p_n$$

Las funciones de distribución de las variables W y W_q vienen dadas por

$$W = 1 - e^{-\mu t} \sum_{n=0}^{k} q_n \sum_{r=0}^{n} \frac{(\mu t)^r}{r!}, \ si\ t \geq 0\ (y\ W(t) = 0\ en\ otro\ caso) \tag{1.60}$$

$$W_q = 1 - e^{-\mu t} \sum_{n=1}^{k} q_n \sum_{r=0}^{n-1} \frac{(\mu t)^r}{r!}, \ si\ t \geq 0\ (y\ W_q(t) = 0\ si\ t < 0) \tag{1.61}$$

Este modelo (y en otros posteriores) el significado de ρ como intensidad de tráfico se desvirtúa. Aquí ρ no puede interpretarse como el cociente entre el número medio de llegadas de clientes al sistema por unidad de tiempo entre el número medio de clientes a los que el servidor tendría capacidad de dar servicio por unidad de tiempo, sino más bien como un cociente semejante, pero donde el numerador representa el número medio de intentos de llegada, más que de llegadas efectivas al sistema. De hecho, por este motivo ρ puede ser mayor o igual que 1, aún siendo el sistema estacionario.

el valor de $\overline{\lambda}$ sí representa el número medio de entradas efectivas de clientes en el sistema por unidad de tiempo y, así, la verdadera intensidad de tráfico podría medirse através de

$$\overline{\rho} = \frac{\overline{\lambda}}{\mu} = \begin{cases} \frac{\frac{\lambda(k+1)}{k+2}}{\mu} = \frac{k+1}{k+2} & ; si\ \rho = 1 \\ \frac{\frac{\lambda(\rho^{k+1}-1)}{\rho^{k+2}-1}}{\mu} = \frac{\rho^{k+2}-\rho}{\rho^{k+2}-1} & ; si\ \rho \neq 1 \end{cases}$$

que efectivamente sí es siempre menor que 1.

Ejemplo 1.6.1. *Los clientes llegan a una peluqueria con una media de 5 por hora y con los tiempos entre llegadas consecutivas distribuidos exponencialmente; hay un peluquero disponible en todo momento y cuatro sillas para los clientes que llegan cuando el peluquero está ocupado. El reglamento del local de prevención de incendios limita el número total de clientes en el servio a 5. EL tiempo de servicio se distribuye exponencialmente con media que cambia según el número de cliente, los datos son los siguientes:*

Número de clientes en fila	1	2	3	4	5
Tiempo medio de atención	9	10	10	13	20

a. Determinar el número medio de clientes en cola.

$$\begin{aligned} \lambda &= 5 \\ \mu_1 &= \frac{60}{9} \\ \mu_2 &= \frac{60}{10} \\ \mu_3 &= \frac{60}{10} \\ \mu_4 &= \frac{60}{13} \\ \mu_5 &= \frac{60}{20} \end{aligned}$$

$$\begin{aligned} c_1 &= \frac{\lambda}{\mu_1} = 0.75 \\ c_2 &= \frac{\lambda^2}{\mu_1\mu_2} = 0.625 \\ c_3 &= \frac{\lambda^3}{\mu_1\mu_2\mu_3} = \frac{25}{48} \\ c_4 &= \frac{\lambda^4}{\mu_1\mu_2\mu_3\mu_4} = \frac{325}{576} \\ c_5 &= \frac{\lambda^5}{\mu_1\mu_2\mu_3\mu_4\mu_5} = \frac{1625}{1728} \end{aligned}$$

$$p_0 = \frac{1}{1 + \sum_{i=1}^{5} c_i} = 0.2272$$

$$\begin{aligned} p_0 &= 0.2272 \\ p_1 &= 0.1704 \\ p_2 &= 0.1420 \\ p_3 &= 0.1184 \\ p_4 &= 0.1282 \\ p_5 &= 0.2137 \end{aligned}$$

$$L_q = \sum_{n=2}^{5} = (n-1)p_n = 0.1420 + 0.2 + 0.4 + 0.9 = 1.618$$

b. Tiempo estimado que el cliente espere en cola.

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

donde

$$\overline{\lambda} = \lambda(1 - p_5) = 5(0.7863) = 3.9315 \approx 4 \; clientes/hora$$

entonces

$$W_q = \frac{1.6182}{4} = 0.41 = 24.7 \; min$$

c. Porcentaje que permanece ocioso el peluquero.

$$p_0 = 0.2272$$

El 22 % del tiempo pasa desocupado.

□

1.7. El modelo M/M/c.

Este modelo estudia los tipos de colas en donde la distribución de los tiempos entre llegadas son exponenciales con parámetro λ, y además los tiempos de servicio se distribuyen exponencialmente con parámetro μ y en este caso existen c servidores, bajo la suposición de que cada servidor realiza la misma función o actividad con el mismo nivel de eficacia que los demás servidores. Con respecto a los demás parámetros del modelo tenemos que la población potencial y la

capacidad de la cola es infinita, la forma de elegir a los clientes en cola para que sean servidos es FIFO. Es por ello que este modelo es considerado como la generalización del modelo $M/M/1$. Por tanto, las tasas de llegada vienen dadas por:

$$\lambda_n = \lambda; \quad \forall n = 0, 1, \ldots$$

y las tasas de servicio son:

$$\mu_n = \begin{cases} n\mu; & \forall 1 \leq n \leq c \\ \\ c\mu; & \forall n > c \end{cases} \tag{1.62}$$

De forma similar al caso $M/M/1$ para obtener c_n:

$$c_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_1} = \prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i} = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & ;\text{si } 1 \leq n \leq c \\ \\ \dfrac{\lambda^n}{c!c^{n-c}\mu^n} & ;\text{si } \forall n > c \end{cases} \tag{1.63}$$

Es necesario también establecer la condición estacionaria para el sistema ($\rho < 1$), o lo que es lo mismo, siempre que se cumpla $\lambda < c\mu$. Para ello basta con determinar la convergencia de la serie siguiente:

$$\begin{aligned} \sum_{n=1}^{\infty} c_n &= \sum_{n=1}^{c-1}\frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{\infty}\frac{\lambda^n}{c!c^{n-c}\mu^n} \\ &= \sum_{n=1}^{c-1}\frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{\infty}\frac{\lambda^c\lambda^{n-c}}{c!c^{n-c}\mu^c\mu^{n-c}} \\ &= \sum_{n=1}^{c-1}\frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c!\mu^c}\sum_{n=c}^{\infty}\frac{\lambda^{n-c}}{c^{n-c}\mu^{n-c}}; \quad \rho = \frac{\lambda}{c\mu} \\ &= \sum_{n=1}^{c-1}\frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c!\mu^c}\sum_{n=c}^{\infty}\rho^{n-c} \end{aligned} \tag{1.64}$$

Entonces a partir del término c-ésimo, la serie es geométrica de razón ρ y por tanto convergente siempre que $\lambda < c\mu$ que puede ser interpretado como el número medio de clientes que entran al sistema por unidad de tiempo, debe ser menor que el número medio de clientes a los que se le completa su servicio, por unidad de tiempo; multiplicada por el número de servidores que hay

en el sistema. Probar la convergencia de la serie $\sum_{n=c}^{\infty} \rho^{n-c}$ resulta sencilla:

$$\begin{aligned}
\sum_{n=c}^{\infty} \rho^{n-c} &= 1 + \rho + \rho^2 + \rho^3 + \dots \\
&= 1 + \rho(1 + \rho + \rho^2 + \rho^3 + \dots) \\
&= 1 + \rho \left(\sum_{n=c}^{\infty} \rho^{n-c} \right) \\
(1-\rho) \sum_{n=c}^{\infty} \rho^{n-c} &= 1 \\
\sum_{n=c}^{\infty} \rho^{n-c} &= \frac{1}{1-\rho}
\end{aligned} \tag{1.65}$$

Entonces de 1.64, la suma de las c_n es:

$$\begin{aligned}
\sum_{n=1}^{\infty} c_n &= \sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c!\mu^c} \cdot \frac{1}{1-\rho} \\
&= \sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c(c-1)!\mu\mu^{c-1}(1-\rho)} \\
&= \sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)}
\end{aligned} \tag{1.66}$$

Análogamente al modelo $M/M/1$ se puede calcular p_0 de la siguiente manera:

$$\begin{aligned}
p_0 &= \frac{1}{1 + \sum_{n=1}^{\infty} c_n} \\
&= \frac{1}{1 + \sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)}}
\end{aligned} \tag{1.67}$$

A la hora de interpretar el cálculo para p_0 debemos tener en cuenta, que si el número de servidores del sistema es alto, los términos con $n!$ en el denominador de p_0, puede ocurrir errores del tipo "overflow" o perdidas de tiempo/datos por exceso de cálculos. Es importante hacer notar que en estos casos es ineficiente aplicar las fórmulas cada vez que se desee calcular los valores, es más útil reutilizar dichos datos. Por ejemplo: si $c = 100$, en el denominador de c_{98} aparecería 98!, que, en lugar de calcularlo directamente es más eficiente multiplicar por 98 al término 97! que aparece en c_{97}. En resumen, es más eficiente definir $c_0 = 1$ y al utilizar cálculos recursivos, tenemos:

$$c_n = c_{n-1} \frac{\lambda_{n-1}}{\mu_n} = c_{n-1} \frac{\lambda}{n\mu}, \quad \forall n = 1, 2, \dots, c-1$$

Esto se deduce de la definición de los c_n. Además para el caso del término $\dfrac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)}$ que representa $\displaystyle\sum_{n=c}^{\infty} c_n$ en (1.64), vemos que:

$$\frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)} = \frac{\lambda^{c-1}}{(c-1)!\mu^{c-1}}\left(\frac{\lambda}{c\mu-\lambda}\right) = d_{c-1}\frac{\lambda}{c\mu-\lambda}$$

y que denotaremos por $D_{\geq c}$, puede calcularse fácilmente a partir de d_{c-1} mediante

$$D_{\geq c} = d_{c-1}\frac{\lambda}{c\mu-\lambda}$$

Teniendo en cuenta todo lo anterior, se llega a la siguiente forma de implementar eficiente el cálculo de p_0:

$$p_0 = \frac{1}{\sum_{n=0}^{c-1} c_n + D_{\geq c}}$$

donde $c_0 = 1$, $c_n = c_{n-1}\dfrac{\lambda}{n\mu}$, $\forall n = 1, 2, \ldots, c-1$ y $D_{\geq c} = d_{c-1}\dfrac{\lambda}{c\mu-\lambda}$.
Ahora para determinar una expresión explícita para las p_n

$$p_n = c_n p_0 = \begin{cases} \dfrac{\lambda^n}{n!\mu^n}p_0; & \forall 1 \leq n < c \\ \dfrac{\lambda^c}{c!\mu^c}\rho^{n-c}p_0; & \forall n \geq c \end{cases} \tag{1.68}$$

Nuevamente, la aplicación directa de la fórmula anterior, resulta ser poco eficiente en términos computacionales. De manera similar al cálculo de p_0, también aquí podemos proceder más eficientemente. Una vez calculado p_0, y si se mantiene en memoria los valores de c_n el cálculo de los p_n sería muy sencillo. Así, si $n = 1, 2, \ldots, c-1$, se tiene $p_n = c_n p_0$. Cuando $n \geq c$ los p_n pueden ser calculados recursivamente ya que p_{c-1} se ha obtenido antes.

$$\begin{aligned} p_n &= c_n p_0 \\ &= c_{n-1}\frac{\lambda_{n-1}}{\mu_n}p_0 \\ &= \frac{\lambda_{n-1}}{\mu_n}p_{n-1} \\ &= \frac{\lambda}{c\mu}p_{n-1} \\ &= \rho.p_{n-1}, \forall n \geq c \end{aligned} \tag{1.69}$$

Entonces, la manera de proceder es la siguiente:
Calcular directamente $p_n = c_n p_0; \quad 1 \leq n < c$
Recursivamente $p_n = \rho.p_{n-1}; \quad n \geq c$
Utilizando las expresiones obtenidas para p_n, puede encontrarse fácilmente el valor de L_q.

$$\begin{aligned}
L_q = E(N_q) &= 0(p_0 + p_1 + \ldots + p_{c-1}) + \sum_{n=c}^{\infty}(n-c)p_n; \quad \text{donde } p_n = c_n p_0 \\
&= \sum_{n=c}^{\infty}(n-c)c_n p_0 \\
&= \sum_{n=c}^{\infty}(n-c)\frac{\lambda^n}{c!c^{n-c}\mu^n}p_0 \\
&= \sum_{n=c}^{\infty}(n-c)\frac{\lambda^{n-c}\lambda^c}{c!c^{n-c}\mu^{n-c}\mu^c}p_0 \\
&= \frac{\lambda^c}{c!\mu^c}p_0\sum_{n=c}^{\infty}(n-c)\rho^{(n-c)}; \quad \text{sea } k = n-c \\
&= \frac{\lambda^c}{c!\mu^c}p_0\sum_{k=0}^{\infty}k\rho^k \\
&= \frac{\lambda^c}{c!\mu^c}p_0\frac{\rho}{(1-\rho)^2}; \quad \rho = \frac{\lambda}{c\mu} \\
&= \frac{\lambda^{c+1}p_0}{(c\mu)^2(c-1)!\mu^{c-1}\left(\frac{c\mu-\lambda}{c\mu}\right)^2} \\
&= \frac{\lambda^{(c+1)}p_0}{(c-1)!\mu^{(c-1)}(c\mu-\lambda)^2} \qquad (1.70)
\end{aligned}$$

En la práctica, dado que $D_{\geq c} = \frac{\lambda^c}{(c-1)!\mu^{(c-1)}(c\mu-\lambda)}$, L_q puede calcularse de forma eficiente mediante la expresión:

$$L_q = D_{\geq c} \cdot \frac{\lambda p_0}{c\mu - \lambda}$$

Mediante la segunda fórmula de Little, se puede obtener W_q (Tiempo medio de espera en cola) a partir de L_q calculado anteriormente, entonces:

Según *Little*:

$$\begin{aligned} W_q &= \frac{1}{\lambda} L_q \\ &= \frac{1}{\lambda} \cdot \frac{\lambda^{c+1} p_0}{(c-1)!\mu^{c-1}(c\mu-\lambda)^2} \\ &= \frac{\lambda^c p_0}{(c-1)!\mu^{c-1}(c\mu-\lambda)^2} \end{aligned} \tag{1.71}$$

Ahora se puede obtener W (Tiempo medio de espera en el sistema) de la siguiente forma $W = W_q + \frac{1}{\mu}$ y luego el número medio de clientes en el sistema $L = \lambda W$. Entonces ya conocidas λ, μ, c y p_0. Es fácil obtener

$$\begin{aligned} W &= W_q + \frac{1}{\mu} \\ &= \frac{\lambda^c \cdot p_0}{(c-1)!\mu^{c-1}(c\mu-\lambda)^2} + \frac{1}{\mu} \end{aligned} \tag{1.72}$$

luego $L = \lambda W$

$$\begin{aligned} L &= \lambda W \\ &= \lambda \left[\frac{\lambda^c \cdot p_0}{(c-1)!\mu^{c-1}(c\mu-\lambda)^2} + \frac{1}{\mu} \right] \\ &= \frac{\lambda^{c+1} \cdot p_0}{(c-1)!\mu^{c-1}(c\mu-\lambda)^2} + \frac{\lambda}{\mu} \end{aligned} \tag{1.73}$$

También, si en algún momento deseamos conocer la función de distribución del tiempo que un cliente tarda en la cola de un sistema, ésta puede calcularse de la siguiente forma:

Lema 1.7.1. *Sea $F_{T_q}(t)$ la función de distribución del tiempo que un cliente pasa en cola, entonces*

$$F_{T_q}(t) = \begin{cases} 1 - \dfrac{c(\lambda/\mu)^c}{c!(c-\lambda/\mu)} p_0 & ; si\ t = 0 \\ \dfrac{\mu(\lambda/\mu)^c(1-e^{-(\mu c-\lambda)t})}{(c-1)!(\mu c-\lambda)} p_0 + F_{T_q}(0) & ; si\ t > 0 \end{cases} \tag{1.74}$$

Demostración:

$$F_{T_q}(0) = P(T_q = 0) = P(N \leq c-1) = \sum_{n=0}^{c-1} p_n = p_0 \sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n}$$

Pero sabemos que

$$\begin{aligned} p_0 &= \frac{1}{1+\sum_{n=1}^{c-1}\frac{\lambda^n}{n!\mu^n}+\frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)}} \\ &= \frac{1}{\sum_{n=0}^{c-1}\frac{\lambda^n}{n!\mu^n}+\frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu-\lambda)}} \end{aligned}$$

si y sólo si

$$\sum_{n=0}^{c-1}\frac{\lambda^n}{n!\mu^n}=\frac{1}{p_0}-\frac{c(\lambda/\mu)^c}{c!(c-\lambda/\mu)}$$

Entonces

$$F_{T_q}(0)=1-\frac{c(\lambda/\mu)^c}{(c-\lambda/\mu)c!}\cdot p_0$$

Ahora, si $n \geq c$, el proceso $Z(t)$ (el número de salidas hasta el instante t) se comporta como un proceso de Poisson $(c\mu)$. Esto significa que los tiempos entre llegadas consecutivas $\tau_1, \tau_2, \ldots$ son v.a.i.i.d. $exp(c\mu)$. Consecuentemente la variable T_k, representa el tiempo transcurrido hasta que terminen k servicios y se distribuye según una $\Gamma(p=k, a=c\mu)$.
Entonces;

$$\begin{aligned} F_{T_q}(t) &= P(T_q \leq t) = P(T_q = 0) + P(0 < T_q < t) \\ &= F_{T_q}(0) + \sum_{n=c}^{\infty} p_n \underbrace{\int_0^t f(x)dx}_{\Gamma(p=n-c+1, a=c\mu)} \\ &= F_{T_q}(0) + \sum_{n=c}^{\infty} c_n p_0 \int_0^t \frac{(c\mu)^{n-c+1}}{(n-c)!}x^{n-c}e^{-c\mu x}dx; \quad Sea\ r=\frac{\lambda}{\mu} \\ &= F_{T_q}(0) + p_0\sum_{n=c}^{\infty}\frac{r^n}{c^{n-c}c!}\int_0^t \frac{(c\mu)^{n-c+1}}{(n-c)!}x^{n-c}e^{-c\mu x}dx \\ &= F_{T_q}(0) + p_0\frac{r^c}{(c-1)!}\int_0^t \mu e^{-\mu c x}\left[\sum_{n=c}^{\infty}\frac{(\mu r x)^{n-c}}{(n-c)!}\right]dx \\ &= F_{T_q}(0) + p_0\frac{r^c}{(c-1)!}\int_0^t \mu e^{-\mu(c-r)x}dx \\ &= F_{T_q}(0) + p_0\frac{r^c}{(c-1)!}\cdot\frac{\mu}{\mu(c-r)}\left[-e^{-\mu(c-r)x}\right]_0^t \\ &= F_{T_q}(0) + p_0\frac{r^c}{(c-1)!}\cdot\frac{\left(1-e^{-\mu(c-r)t}\right)}{(c-r)} \\ F_{T_q}(t) &= 1-\frac{cr^c}{(c-r)c!}p_0+\frac{r^c}{(c-1)!}\cdot\frac{\left(1-e^{-\mu(c-r)t}\right)}{(c-r)}p_0 \end{aligned} \tag{1.75}$$

Ejemplo 1.7.1. *Una sucursal bancaria tiene dos cajeros igualmente eficientes y capaces de atender un promedio de 60 operaciones por hora, con los tiempos reales de servicio distribuidos exponencialmente. Los clientes llegan al banco siguiendo un procesos de Poisson con una tasa media de 100 por hora. Determinar:*

- *La probabilidad de que haya más de 3 clientes en el banco.*
- *La probabilidad de que uno de los cajeros esté ocioso.*
- *¿Cuál es el número medio de clientes que están en cola?*
- *¿Cuál es el número medio de clientes en la sucursal?*

Solución:
Como se puede apreciar se trata de un modelo $M/M/2$, donde $\lambda = 100$ y $\mu = 60$. Se puede comprobar que se trata de un sistema estacionario ya que $\rho = \lambda/c\mu < 1$.
Para determinar la probabilidad de que haya más de 3 clientes en el banco se puede proceder de la siguiente manera.

$$P(N > 3) = 1 - P(N \leq 3) = 1 - (p_0 + p_1 + p_2 + p_3)$$

Donde:
p_0 es la probabilidad de que no haya clientes en el banco.
p_i es la probabilidad de que no haya clientes en el banco para $i = 1, 2, \ldots$
Cuando $n = 0$, p_0 se calcula mediante la siguiente expresión

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{(c-1)!\mu^{c-1}(c\mu - \lambda)} \right)^{-1} \approx 0.0909$$

Para $n \geq 1$ sabemos que

$$p_n = c_n p_0 = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} \cdot p_0; & \forall 1 \leq n < c \\ \dfrac{\lambda^c}{c!\mu^c} \rho^{n-c} \cdot p_0; & \forall n \geq c \end{cases}$$

Se obtiene;

$$\begin{aligned} p_1 &= \frac{\lambda}{\mu} \cdot p_0 = \frac{100}{60}(0.0909) = 0.1515 \\ p_2 &= \frac{100^2}{2! \cdot 60^2}(0.0909) = \frac{1}{2!}\left(\frac{5}{3}\right)^2 (0.0909) = 0.12626 \\ p_3 &= \frac{1}{3! \cdot 2}\left(\frac{5}{3}\right)^3 (0.0909) = 0.03507 \end{aligned}$$

Entonces

$$\begin{aligned} P(N > 3) &= 1 - P(N \leq 3) \\ &= 1 - (p_0 + p_1 + p_2 + p_3) \\ &= 1 - (0.0909 + 0.1515 + 0.12626 + 0.03507) \\ &= 1 - 0.40373 \\ &= 0.59627 \end{aligned}$$

Por tanto; la probabilidad de que haya más de 3 clientes en la sucursal bancaria es de 0.59527*, esto también se puede expresar como la probabilidad de que haya exactamente un cliente en la cola. Para obtener la probabilidad de que un cajero esté ocioso, resulta fácilmente:*

$$\begin{aligned} P(\textit{Un cajero esté ocioso}) &= P(\textit{Ningún cliente en la sucursal}) \\ &+ P(\textit{Haya un sólo cliente en el banco}) \\ &= p_0 + p_1 \\ &= 0.0909 + 0.1515 \\ &= 0.2424 \end{aligned}$$

Para determinar el número medio de clientes que están en cola, utilizamos la segunda fórmula de Little:

$$\begin{aligned} L_q &= \frac{\lambda^{c+1} \cdot p_0}{(c-1)!\mu^{(c-1)}(c\mu - \lambda)^2} \\ &= \frac{100^3}{60(2(60) - 100)^2} \cdot (0.0909) \\ &= 3.7875 \end{aligned}$$

Entonces el número medio de clientes en la cola de la sucursal bancaria es aproximadamente 3 clientes. Por último necesitamos saber ¿Cuál es el número medio de clientes que están en el sistema?.

$$\begin{aligned} L &= \lambda W \\ &= \lambda\left(W_q + \frac{1}{\mu}\right) \\ &= L_q + \frac{\lambda}{\mu} \end{aligned}$$

$$\begin{aligned} L &= 3.787 + \frac{100}{60} \\ &= 5.4537 \end{aligned}$$

Entonces el número promedio de clientes en la sucursal bancaria es de 5.

□

1.8. El modelo M/M/c/k.

Este modelo es una extensión del modelo $M/M/1/k$, ya que en este caso existen c servidores y la capacidad de la cola es limitada a k clientes. Donde las tasas entre llegadas se comportan como las del modelo $M/M/1/k$, mientras que las tasas de servicio son iguales a las de un $M/M/c$:

$$\lambda_n = \begin{cases} \lambda; & \text{si } n = 0, 1, 2, \ldots, k+c-1 \\ 0; & \text{si } n = k+c, k+c+1, k+c+2, \ldots \end{cases}$$

$$\mu_n = \begin{cases} n\mu; & \text{si } n = 1, 2, \ldots, c \\ c\mu; & \text{si } n = c+1, c+2, \ldots \end{cases}$$

Como consecuencia de esto se obtiene:

$$c_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & ;\text{si } n = 1, 2, \ldots, c \\ \dfrac{\lambda^n}{c!c^{n-c}\mu^n} & ;\text{si } n = c+1, c+2, \ldots, k+c \\ 0 & ;\text{si } n = k+c+1, k+c+2, \ldots \end{cases} \tag{1.76}$$

Entonces la serie $\sum_{n=1}^{\infty} c_n$ tiene sólo un número finito de términos distintos de cero y, por tanto, siempre es convergente. Así pues el sistema es estacionario siempre (independientemente del valor de ρ). Para calcular p_0 necesitamos encontrar una expresión, lo más sencilla posible para la suma de la serie:

$$\begin{aligned}\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{k+c} c_n &= \sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{k+c} \frac{\lambda^n}{c!c^{n-c}\mu^n} \\ &= \sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c!\mu^c} \sum_{n=c}^{k+c} \rho^{n-c}\end{aligned}$$

Obteniendo como resultado:

$$\sum_{n=1}^{\infty} c_n = \begin{cases} \displaystyle\sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c!\mu^c} \cdot \frac{1-\rho^{k+1}}{1-\rho} & ;\text{si } \rho \neq 1 \\ \displaystyle\sum_{n=1}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c(k+1)}{c!\mu^c} & ;\text{ si } \rho = 1 \end{cases} \tag{1.77}$$

Como consecuencia,

$$p_0 = \begin{cases} \left(\displaystyle\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c}{c!\mu^c} \cdot \frac{1-\rho^{k+1}}{1-\rho} \right)^{-1} & ;\text{si } \rho \neq 1 \\ \left(\displaystyle\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^c(k+1)}{c!\mu^c} \right)^{-1} & ;\text{si } \rho = 1 \end{cases} \tag{1.78}$$

Al igual que en el caso de un sólo servidor usaremos la notación $\rho = \lambda/c\mu$. Tal como ya se comentó en el modelo $M/M/c$, la implementación directa de las fórmulas anteriores no es precisamente la manera más eficiente de calcular p_0. Entonces, si empleamos las fórmulas recursivas puede calcularse p_0 mediante

$$p_0 = \frac{1}{\sum_{n=0}^{c-1} c_n + D_{\geq c}}$$

donde: $c_0 = 1, c_n = c_{n-1}\frac{\lambda}{n\mu}$, $\forall n = 1, 2, \ldots$

Además

$$D_{\geq c} = \begin{cases} d_{c-1}\dfrac{\rho - \rho^{k+2}}{1-\rho} & ;\text{si } \rho \neq 1 \\ d_{c-1}(k+1) & ;\text{si } \rho = 1 \end{cases} \tag{1.79}$$

A partir de p_0, y los parámetros de entrada del modelo, pueden obtenerse expresiones explícitas para las p_n:

$$p_n = c_n \cdot p_0 = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & ;\text{si } n = 1, 2, \ldots, c \\ \dfrac{\lambda^c}{c!\mu^c} \rho^{n-c} p_0 & ;\text{si } n = c+1, c+2, \ldots, k+c \end{cases} \tag{1.80}$$

Nuevamente, las fórmulas implementadas directamente a partir de estas expresiones son muy poco eficientes y resulta preferible proceder de forma análoga al modelo $M/M/c$.
Es decir; reutilizando los valores de c_n, necesarios para la implementación del cálculo eficiente de p_0, los p_n se obtienen de manera sencilla: $p_n = c_n \cdot p_0, \forall n = 1, 2, \ldots, c-1$. Por otra parte, si $n \geq c$, el término p_n debe ser calculado recursivamente, comenzando en p_{c-1} que ya ha sido calculado:

$$p_n = \frac{\lambda_{n-1}}{\mu_n} p_{n-1} = \rho \cdot p_{n-1}, \forall n = c, c+1, \ldots, k+c$$

Fórmula que, aún siendo válida para todo valor de ρ, se trivializa si $\rho = 1$, dando como resultado

$$p_{k+c} = p_{k+c-1} = \cdots = p_c = p_{c-1} \quad \text{para } \rho = 1$$

Las cuatro medidas de interés (L, L_q, W, y W_q) pueden calcularse a partir de los valores de las p_n recien encontradas. Posiblemente el cálculo que resulte más sencillo sea el de L_q, por el cual comenzaremos. Si $\rho \neq 1$, a partir de la igualdad

$$\sum_{n=0}^{k+1} n x^{n-1} = \frac{(k+1)x^{k+2} - (k+2)x^{k+1} + 1}{(x-1)^2} \tag{1.81}$$

demostrada y utilizada ya para el modelo $M/M/1/k$, se tiene

$$\begin{aligned} L_q &= \sum_{n=c+1}^{c+k} (n-c) p_n = \sum_{n=c}^{c+k} (n-c) \frac{\lambda^n}{c! c^{n-c} \mu^n} p_0 \\ &= \frac{\lambda^c}{c!\mu^c} \cdot p_0 \cdot \rho \sum_{n=c}^{c+k} (n-c) \rho^{n-c-1} \\ &= \frac{\lambda^c}{c!\mu^c} \cdot p_0 \cdot \rho \sum_{j=0}^{k} j \cdot \rho^{j-1} \\ &= \frac{\lambda^c}{c!\mu^c} \cdot p_0 \cdot \rho \cdot \frac{k\rho^{k+1} - (k+1)\rho^k + 1}{(\rho-1)^2} \\ &= \frac{\lambda^c \cdot p_0 \cdot \rho \cdot [1 + k\rho^{k+1} - (k+1)\rho^k]}{c!\mu^c (\rho-1)^2} \end{aligned} \tag{1.82}$$

Sin embargo, si $\rho = 1$ entonces

$$\begin{aligned} L_q &= \sum_{n=c+1}^{c+k} (n-c) \cdot p_n \\ &= p_{c-1} \sum_{N=c+1}^{c+k} (n-c) \\ &= p_{c-1} \frac{k(k+1)}{2} \\ &= \frac{\lambda^{c-1} k(k+1)}{2(c-1)!\mu^{c-1}} \end{aligned} \tag{1.83}$$

En resumen, la expresión explícita para el número medio de clientes en la cola es

$$L_q = \begin{cases} \dfrac{\lambda^{c-1} k(k+1) \cdot p_0}{2(c-1)!\mu^{c-1}} & ;\text{si } \rho = 1 \\ \dfrac{\lambda^c \cdot p_0 \cdot \rho[1 + k\rho^{k+1} - (k+1)\rho^k]}{c!\mu^c (1-\rho)^2} & ;\text{si } \rho \neq 1 \end{cases} \tag{1.84}$$

sin embargo, para la obtención de un cálculo mucho más eficiente usaremos,

$$L_q = \begin{cases} \dfrac{k(k+1)}{2} p_{c-1} & ;\text{si } \rho = 1 \\ \dfrac{[1 + k\rho^{k+1} - (k+1)\rho^k]\rho^2}{(1-\rho)^2} p_{c-1} & ;\text{si } \rho \neq 1 \end{cases} \tag{1.85}$$

Para la obtención de las demás medidas debemos de usar las fórmulas de Little y calcular, primeramente, $\bar{\lambda}$:

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n \cdot p_n = \lambda \sum_{n=0}^{k+c+1} p_n = \lambda(1 - p_{k+c}),$$

que es la implementación a usar para un cálculo eficiente. Una fórmula más explícita que proporcionará este valor es:

$$\bar{\lambda} = \lambda \left(1 - \frac{\lambda^c}{c!\mu^c} \rho^k p_0\right),$$

que se simplifica a $\bar{\lambda} = \lambda \left(1 - \frac{\lambda^{c-1}}{(c-1)!\mu^{c-1}} p_0\right)$, si $\rho = 1$. De nuevo, en este modelo vuelve a ocurrir que ρ no representa la intensidad de tráfico efectiva.

Este valor puede calcularse mediante

$$\begin{aligned} \bar{\rho} = \frac{\bar{\lambda}}{c\mu} &= \rho\left(1 - \frac{\lambda^c}{c!\mu^c}\rho^k p_0\right) \\ \bar{\rho} &= \rho - \frac{\lambda^c}{c!\mu^c}\rho^{k+1} p_0 \end{aligned} \tag{1.86}$$

A partir de los valores de L_q y $\bar{\lambda}$, calculados de forma eficiente, pueden usarse las fórmulas de Little para obtener:

$$\begin{aligned} W_q &= \frac{L_q}{\bar{\lambda}}, \\ W &= W_q + \frac{1}{\mu}, \\ L &= \bar{\lambda}W. \end{aligned}$$

Es importante mencionar que en este modelo resulta bastante complicado obtener la distribución del tiempo que un cliente está en el sistema. Sin embargo es posible determinar el tiempo que un cliente pasa en cola a través de la siguiente expresión:

$$\mathbf{W}_q(t) = 1 - e^{-c\mu t}\sum_{n=c}^{k+c-1} q_n \sum_{r=0}^{n-c}\frac{(c\mu t)^r}{r!}, \quad \text{si } t \geq 0 \quad (\text{y} \quad \mathbf{W}_q(t) = 0 \text{ si } t < 0)$$

donde las q_n tienen el mismo significado que en el modelo $M/M/1/k$, es decir, la probabilidad de que haya n clientes en el sistema justo cuando una llegada se está produciendo, viene dada por

$$q_n = \frac{p_n}{1 - p_{k+c}}, n = 0, 1, 2, \ldots, k + c - 1$$

Ejemplo 1.8.1. *En una estación de trabajo constituida por* 3 *servidores, se ejecutan programas informáticos, (se supone que es la única carga de trabajo de la estación) con tiempo de CPU que se distribuyen como una exponencial de media 3 minutos. Los programas son servidos de acuerdo a una disciplina FIFO, la población potencial del sistema es infinita y el tamaño de la cola es limitado a un sólo programa. Sabiendo que las llegadas a la estación de trabajo se producen según un proceso de Poisson con una intensidad promedio de 15 programas cada hora, se pide:*

- *¿Cuál es la probabilidad de que haya más de dos programas en espera de ejecución (además de los c que se están ejecutando)?*

- *¿Cuál es el número medio de programas que están a la espera de comenzar a ejecutarse?*
- *Hallar el número medio total de procesos en la estación de trabajo.*

Solución:
El modelo pasa a ser un $M/M/c/k$ *con c servidores igual a* 3, $k = 1$, $\lambda = 15$ *y* $\mu = 20$. *Para dar respuesta al primer apartado resulta obvio considerar que si la capacidad de la cola esta limitada a un solo programa, entonces* $P(N_q > 2) = 0$. *Ahora, para determinar el número de programas que están a la espera de ejecutarse, se requiere calcular lo siguiente:*
$c_0 = 1$, $c_1 = c_0\frac{15}{20} = 0.75$, $c_2 = c_1\frac{15}{2(20)} = 0.28125$, $c_3 = c_2\frac{15}{3(20)} = 0.07031$
y $c_4 = c_3\frac{15}{4(20)} = 0.01318$, *entonces* $c_n = 0 \quad \forall n \geq 5$.
Para calcular p_0*:*

$$\begin{aligned} p_0 &= \frac{1}{\sum_{i=0}^{4} c_i} \\ &= \frac{1}{1 + 0.75 + 0.28125 + 0.07031 + 0.01318} \\ &= 0.4728 \end{aligned}$$

Dado que sólo hay 5 *estados posibles en el sistema* $(N = 0, 1, 2, 3, 4)$, *entonces procedemos a calcularse directamente:*

$$L_q = 0(p_0 + p_1 + p_2 + p_3) + 1 \cdot p_4 = c_4 \cdot p_0 = (0.01318)(0.4728) = 0.006231$$

Para determinar el número medio total de procesos en la estación de trabajo:

$$\begin{aligned} \sum_{n=0}^{4} np_n &= p_0(c_1 + 2c_2 + 3c_3 + 4c_4) \\ &= (0.4728)(0.75 + 2(0.28125) + 3(0.07031) + 4(0.01318)) \\ &= (0.4728)(1.57615) \\ &= 0.7452 \end{aligned}$$

□

1.8.1. El modelo $M/M/c/c$ (un caso especial del modelo $M/M/c/k$)

Un caso especial de este modelo es cuando la capacidad de la cola es igual al número de servidores que hay en el sistema, es decir; $k = c$. Entonces bajo estas condiciones nunca se producirán colas, es por ello que muchas veces este modelo no es considerado como un sistema de colas propiamente dicho. Sin embargo; resulta de especial importancia este modelo ya que es considerado cómo un modelo clásico de las centrales telefónicas, en la cuál el número de servidores es el número de líneas en el sistema. Además, en este tipo de modelos existe un valor que es especialmente relevante, nos referimos a la probabilidad de que el sistema esté saturado, es decir, haya c clientes en el sistema, frecuentemente identificado en las centrales de telefonía como "la probabilidad de no tener línea", y cuyo valor como se ha visto con anterioridad es:

$$p_c = \frac{(c\rho)^c/c!}{\sum_{i=0}^{c}(c\rho)^i/i!} \tag{1.87}$$

También se puede determinar la probabilidad de que haya n elementos en el sistema por:

$$p_n = \frac{\frac{(\lambda/\mu)^n}{n!}}{\sum_{i=0}^{c}\frac{(\lambda/\mu)^i}{i!}} \tag{1.88}$$

Demostrar que (1.88) es cierta resulta sencillo, para ello partimos de la expresión $p_n = c_n p_0$ utilizada anteriormente para denotar la probabilidad de que hayan n clientes en el sistema, entonces:

$$\begin{aligned}
p_n &= c_n p_0 \\
&= p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \\
&= p_0 \prod_{i=1}^{n} \frac{\lambda}{i\mu}; \quad \forall 1 \leq n \leq c \\
&= p_0 \left(\frac{\lambda^n}{n!\mu^n}\right) \\
p_n &= \frac{(\frac{\lambda}{\mu})^n}{n!} p_0
\end{aligned}$$

Como p_n representa probabilidades, entonces se cumple que

$$1 = \sum_{i=0}^{\infty} p_i = p_0 \sum_{i=0}^{c} \frac{(\frac{\lambda}{\mu})^i}{i!}$$

Entonces

$$p_0{}^{-1} = \sum_{i=0}^{c} \frac{(\frac{\lambda}{\mu})^i}{i!}$$

Por tanto

$$p_n = \frac{(\lambda/\mu)^n}{n!} p_0 = \frac{(\lambda/\mu)^n}{n! \sum_{i=0}^{c} \frac{(\lambda/\mu)^i}{i!}} \quad , \forall 1 \leq n \leq c$$

1.9. El modelo M/G/1.

Como su nombre indica, se trata del sistema de una cola con un único servidor, con tiempo entre llegadas de clientes consecutivos con distribución exponencial y con tiempo de servicio de distribución arbitraria, G. En este contexto, λ seguira siendo el parámetro de la exponencial que rige los tiempos entre dos llegadas de clientes consecutivos. Esto significa que el número de clientes que llegan al sistema por unidad de tiempo será también λ. Como la distribución G no tiene porque ser exponencial, el parámetro μ pasará ahora a significar el inverso del tiempo medio de servicio, es decir

$$\mu = \frac{1}{E[G]}$$

Esta cola viene caracterizada por las siguientes hipótesis:

1. Tiempos entre llegadas exponenciales independientes; es decir, $\tau_1, \tau_2, \ldots$ son variables aleatorias independientes e igualmente distribuidas (v.a.i.i.d) $exp(\lambda)$. Por tanto, $N(t)$: número de llegadas hasta el instante t sigue un proceso estocástico (P.E) de Poisson de parámetro λ.

2. Tiempos de servicio independientes; es decir, $s_1, s_2, \ldots$ son v.a.i.i.d según una distribución de probabilidad cualquiera $F(\cdot)$.

3. $\tau_1, \tau_2, \ldots$ y $s_1, s_2, \ldots$ son independientes.

4. hay un único canal de servicio.

Como el tipo de distribución para el tiempo de llegadas es desconocido, los cálculos de las medidas de efectividad para este tipo de modelos se pueden obtener mediante la fórmula que se definirá acontinuación.

1.9.1. La fórmula de Pollaczek-Khintchine.

Sea:

$$L = \rho + \frac{\rho^2 + \lambda^2 \sigma_S^2}{2(1-\rho)} \tag{1.89}$$

Donde:

$\rho = \frac{\lambda}{\mu}$, $E[S] = \frac{1}{\mu}$, $V[S] = \sigma_S^2$ y L que es el número medio de clientes en el sistema en estado estacionario.

Para demostrar (1.89) veamos la siguiente relación:

$$X_{n+1} = X_n - U(X_n) + A_{n+1}, \tag{1.90}$$

donde:

$$U(X_n) = \begin{cases} 1, & X_n > 0 \\ 0, & X_n = 0 \end{cases}$$

Si tomamos esperanzas en (1.90) tenemos:

$$E\left[X_{n+1}\right] = E\left[X_n\right] - E\left[U(X_n)\right] + E\left[A_{n+1}\right]$$

y tomando en cuenta que en estado estacionario se verifican las igualdades

$$L = E\left[X_n\right] = E\left[X_{n+1}\right]$$

entonces la ecuación (1.90) nos queda de la siguiente forma:

$$E\left[U(X_n)\right] = E\left[A_{n+1}\right]$$

Así pues

$$\begin{aligned} E\left[A_{n+1}\right] &= E\left[E\left[A_{n+1}/S_{n+1}\right]\right] \\ &= \int_0^\infty E\left[A_{n+1}/S_{n+1}\right] g(t)dt \\ &= \int_0^\infty \lambda t g(t)dt \\ &= \lambda \int_0^\infty t g(t)dt \\ &= \lambda E[S_{n+1}] \\ &= \frac{\lambda}{\mu} \\ &= \rho \end{aligned} \tag{1.91}$$

Elevando al cuadrado (1.90) y tomando esperanzas tenemos:

$$\begin{aligned}
X_{n+1}^2 &= [X_n - U(X_n) + A_{n+1}]^2 \\
&= [X_n - U(X_n) + A_{n+1}]\,[X_n - U(X_n) + A_{n+1}] \\
&= X_n^2 + X_n A_{n+1} - X_n U(X_n) + A_{n+1}X_n + A_{n+1}^2 - A_{n+1}U(X_n) \\
&\quad -U(X_n)X_n - U(X_n)A_{n+1} + U(X_n)^2 \\
&= X_n^2 + U(X_n)^2 + A_{n+1}^2 + 2X_nA_{n+1} - 2U(X_n)X_n - 2A_{n+1}U(X_n) \\
E[X_{n+1}^2] &= E[X_n^2] + E[U(X_n)^2] + E[A_{n+1}^2] + 2E[X_n]E[A_{n+1}] \\
&\quad -2E[U(X_n)]E[X_n] - 2E[U(X_n)]E[A_{n+1}] \\
0 &= E[U(X_n)^2] + E[A_{n+1}^2] - 2E[U(X_n)]E[X_n] \\
&\quad -2E[U(X_n)]E[A_{n+1}] + 2E[X_n]E[A_{n+1}] \\
&= \rho + E[A_{n+1}^2] - 2L - 2\rho^2 + 2L\rho \\
2L - 2L\rho &= \rho + E[A_{n+1}^2] - 2\rho^2 \\
L &= \frac{\rho - 2\rho^2 + E[A_{n+1}^2]}{2(1-\rho)} \qquad (1.92)
\end{aligned}$$

desarrollando $E[A_{n+1}^2]$

$$\begin{aligned}
E[A_{n+1}^2] &= V[A_{n+1}] + E[A_{n+1}]^2 \\
&= V[A_{n+1}] + \rho^2 \qquad (1.93) \\
V[A_{n+1}] &= E[V[A_{n+1}/S_{n+1}]] + V[E[A_{n+1}/S_{n+1}]] \\
&= E[\lambda S_{n+1}] + V[\lambda S_{n+1}] \\
&= \lambda E[S_{n+1}] + \lambda^2 V[S_{n+1}] \\
&= \lambda\frac{1}{\mu} + \lambda^2\sigma_S^2 \\
&= \rho + \lambda^2\sigma_S^2 \qquad (1.94)
\end{aligned}$$

Asi que:

$$\begin{aligned}
L &= \frac{\rho - 2\rho^2 + \rho + \lambda^2\sigma_S^2 + \rho^2}{2(1-\rho)} \\
&= \frac{2\rho(1-\rho) + \lambda^2\sigma_S^2 + \rho^2}{2(1-\rho)} \\
&= \rho + \frac{\lambda^2\sigma_S^2 + \rho^2}{2(1-\rho)} \qquad (1.95)
\end{aligned}$$

A partir de la ecuación anterior se pueden obtener W, W_q y L_q ya que siguen verificándose las fórmulas de Little y la relación entre tiempos medios en el sistema de la cola, vemos entonces los siguientes resultados: $W = \frac{L}{\lambda}$, $W_q = W - \frac{1}{\mu}$, $L_q = \lambda W_q$.

Capítulo 2

Simulación de Sistemas.

La primera vez en la historia que se habló de simulación fue en el año de 1949 cuando John Von Neumann y Stanislaw Ulam presentaron el denominado método de Montecarlo. Desde entonces la simulación ha sufrido un crecimiento muy fuerte y, especialmente en las dos últimas décadas, este crecimiento ha sido vertiginoso gracias al desarrollo de los ordenadores.
Dar una definición exacta de la simulación no es una tarea fácil dada la amplitud de las aplicaciones y sistemas a los que se aplica. Sin embargo, una buena definición sería la dada por Robert Shannon en 1975, según el cual, simulación es el proceso de diseñar un modelo de un sistema real y llevar a cabo experiencias con él, con la finalidad de aprender el comportamiento del sistema o de evaluar diversas estrategias para el funcionamiento del sistema.
En ocasiones, puede ser imposible o extremadamente costoso observar ciertos procesos en el mundo real. Por ejemplo previo a los primeros vuelos tripulados realizados por los Estados Unidos o la Unión Soviética, la NASA no tenía información de los efectos que tales vuelos tendrían sobre los seres humanos, pues nadie lo había experimentado antes. Era claro que una alternativa para obtener información inicial acerca de los vuelos espaciales con seres humanos, era la realización de un gran número de vuelos experimentales usando pilotos de prueba. Este método fue rechazado debido al alto valor que se da a la vida humana. Sin embargo, la NASA implantó con éxito un método: la simulación en computadora de los vuelos y sus efectos en los pilotos de prueba. Es así como cada vuelo orbital logrado por la NASA ha sido precedido por una experimentación durante meses y años, con vuelos simulados.
Otros ejemplos de procesos en los que puede ser imposible o muy oneroso obtener los datos, pueden ser: el reporte de ventas de una empresa para el año siguiente; el movimiento de tran-

sacciones bancarias en un período de tiempo dado; los efectos de una campaña publicitaria en las ventas totales de una empresa. En todos estos casos la simulación puede ser empleada como medio efectivo para generar datos numéricos que describen procesos que de otra manera implicaría un elevado costo en proporcionar la información.
El sistema observado puede ser tan complejo que sea imposible describirlo en términos de un sistema de ecuaciones matemáticas, del cual se puedan tener soluciones analíticas para ser usadas con propósitos predictivos. Aún cuando un modelo matemático logre formularse para describir algún sistema de interés, es posible no obtener una solución del modelo por medio de técnicas analíticas directas. La mayoría de los sistemas económicos se encuentran en está categoría. También, puede ser costoso o imposible realizar experimentos de validación en determinados modelos matemáticos. Se ha constatado que la simulación constituye un instrumento extremadamente efectivo para trabajar con problemas de este tipo. Otro tipo de problemas que presentan dificultades semejantes son los fenómenos de espera en gran escala, aquellos que implican canales múltiples, sean ellos en serie y/o en paralelo.

2.1. Introducción al modelado y simulación de sistemas

2.1.1. Conceptos básicos

Un **sistema** puede definirse como una colección de entes que actúan o interactúan para la consecución de un determinado fin. Los objetivos del estudio de un sistema suelen condicionar el conjunto de entidades consideradas, es decir, para un estudio dado puede ser suficiente considerar un subconjunto de las entidades que componen el sistema global.
El **Estado** de un sistema viene determinado por el conjunto de variables necesarias para describirlo en cualquier instante temporal, recibiendo cada una de estas variables el nombre de **variable de estado**.
La forma de evolución temporal de las variables de estado permite establecer una primera clasificación de los sistemas:

- *Sistemas discretos*: aquellos en los que sus variables de estado cambian en un conjunto de instantes de tiempo contable (infinito numerable).

- *Sistemas continuos*: aquellos en los que sus variables de estado cambian de manera conti-

nua a lo largo del tiempo.

En muchas ocasiones, se hace necesario estudiar un sistema con el objeto de establecer relaciones entre algunas de sus entidades, medir sus prestaciones, o bien predecir su comportamiento bajo ciertas condiciones nuevas. Este estudio se puede acometer de las siguientes formas:

- *Realizando experimentos sobre el sistema.* Estos experimentos sólo se pueden realizar cuando se dispone del sistema y cuando es posible alterar sus condiciones de funcionamiento. Por tanto, esta solución no es aplicable cuando los estudios se realizan antes de disponer del sistema, o bien cuando el coste de modificación de las condiciones de funcionamiento resulta muy elevado.

- *Realizando experimentos sobre un modelo del sistema.*

Un **modelo** es una representación de un sistema construido con el propósito de estudiarlo. Normalmente se clasifican en:

- *Modelos físicos*: Estos modelos son muy usados en las industrias aeronáutica y del automóvil. En esta última, por ejemplo, inicialmente se construye un modelo a escala del vehículo que se está diseñando y sobre este modelo se realizan las pruebas.
  En general, y dada la naturaleza del problema, este tipo de modelos tiene poco interés en la investigación de operaciones y en el análisis de sistemas.

- *Modelos matemáticos*: Estos modelos representan un sistema mediante un conjunto de relaciones cuantitativas y lógicas entre sus componentes, permitiendo estudiar cómo se comporta el modelo del sistema cuando cambia alguno de sus componentes.

Una vez definido un modelo matemático de un sistema, se debe realizar un primer estudio con el objetivo de determinar cómo usar este modelo para dar respuesta a las cuestiones de interés planteadas sobre el sistema que representa. Si el modelo es lo suficientemente sencillo, será posible obtener una **solución analítica** que relacione las magnitudes de interés. Si se puede obtener y si su coste computacional es asumible, se preferirán las soluciones analíticas. Sin embargo, la elevada complejidad de muchos sistemas reales imposibilita la obtención de modelos suficientemente ajustados con soluciones analíticas o, en caso de obtenerlas, la carga computacional que conllevan desaconseja su uso. En este último caso, el modelo debe estudiarse de forma aproximada recurriendo a su **simulación**. En principio existen, por tanto, dos posibilidades (Ver figura 2.1):

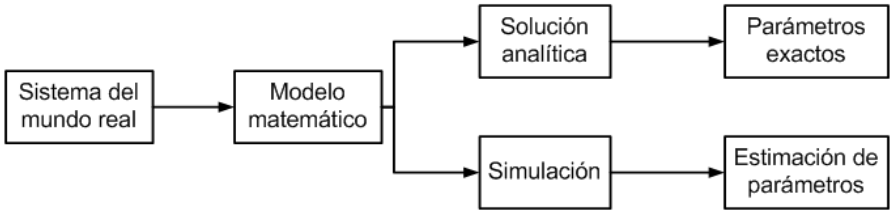


Figura 2.1: Vías de estudio de un sistema.

- *Solución analítica*: supone analizar totalmente el modelo del sistema y obtener una solución que valdrá para todo momento y para obtener cualquier parámetro de interés.
- *Simulación*: se recrea una o varias evoluciones temporales del modelo con el fin de estimar un conjunto de parámetros. Los modelos de simulación son modelos matemáticos que permiten obtener una estimación del comportamiento del sistema para una configuración determinada.

2.1.2. Tipos de modelos de simulación

Los modelos de simulación se pueden clasificar atendiendo a diferentes criterios:

1. Según el instante temporal que representan:
   - *Estáticos*: representan a un sistema en un instante determinado.
   - *Dinámicos*: representan a un sistema que evoluciona a lo largo del tiempo.
2. Según la aleatoriedad de sus variables de estado:
   - *Deterministas*: la representación del sistema no contiene ninguna variable de estado aleatoria.
   - *Estocásticos*: la representación del sistema contiene al menos una variable de estado no determinista.
3. Según el modo en que evolucionan sus variables de estado.
   - *Discretos o de eventos discretos*: Si las variables de estado del modelo varían en un conjunto contable de instantes de tiempo.

- *Continuos*: si las variables de estado varían de modo continuo en función del tiempo.

2.1.3. Formulación de un modelo para simulación

La elevada carga computacional que requieren la mayor parte de las simulaciones de eventos discretos (estudio de modelos discretos, estocásticos y dinámicos) hace imprescindible la utilización del computador como herramienta para la implementación y ejecución de las mismas. A la hora de diseñar un modelo de simulación de estos sistemas, podemos seguir los pasos siguientes:

1. *Identificación de eventos.* Un evento o suceso será cualquier acción (instantánea) susceptible a modificar el estado del sistema.

2. *Definición del mecanismo de control de tiempos.* Dada la naturaleza dinámica de la simulación de eventos discretos, durante el proceso de simulación es necesario mantener información sobre el tiempo de simulación, así como disponer de un mecanismo que permita decidir cuándo este tiempo evoluciona de un valor a otro. Llamaremos **reloj de simulación** a la variable que almacena el valor actual del tiempo de simulación, es decir, el instante temporal en el que está representado el sistema. En cuanto al mecanismo de avance del tiempo de simulación, pueden establecerse dos tipos de temporización:

   - Temporización síncrona. En este caso el tiempo simulado (reloj) avanza a incrementos fijos.
   - Temporización asíncrona. En este caso el tiempo simulado avanza sólo cuando ocurre un evento.

3. *Identificación de las estructuras de datos:*

   - Definir cuáles son las variables de estado.
   - Reloj de simulación. Obviamente crucial ya que llevará el cómputo del tiempo de simulación.
   - Listas o tablas de sucesos. En estas estructuras se almacenarán los instantes temporales en los que se ha planificado su ocurrencia, además de otra información relativa a dichos sucesos o eventos futuros.

- Contadores estadísticos. En estas variables se almacena la información necesaria para la medición de las magnitudes de interés. No son variables de estado sino auxiliares, es decir, no representan el estado del sistema.

4. *Definición del flujo de control y de datos del programa simulador.* Suponiendo temporización asíncrona adoptaremos el diagrama de la figura 2.2.
   La aleatoriedad de algunas de las variables de estado del sistema se conseguirá mediante el uso de un generador de números aleatorios.

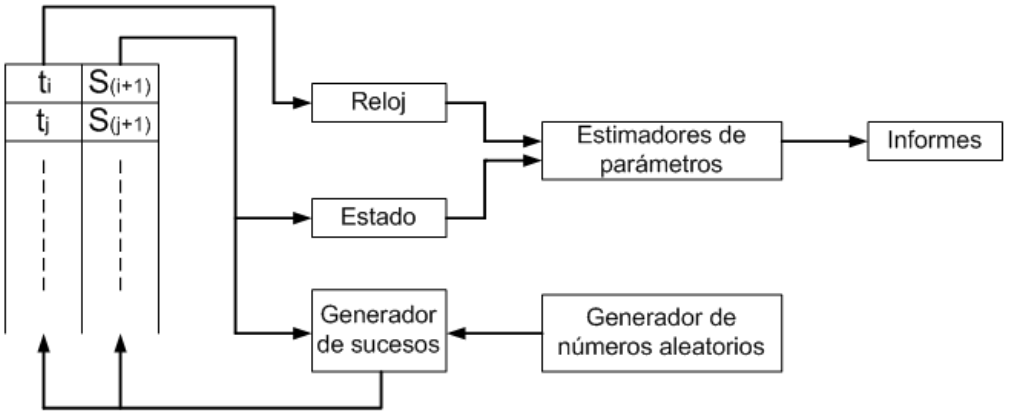


Figura 2.2: Flujo de datos de un modelo con temporización asíncrona.

5. *Definir la estructura del programa que realiza la simulación.* Éste estará formado por los siguientes módulos (ver figura 2.3).

   - Programa principal.
   - Rutina de inicialización.
   - Rutina de temporización (manejo del reloj).
   - Colección de rutinas de manejo de eventos y actualización de contadores estadísticos.
   - Rutinas de generación de informes.
   - Rutinas de generación de secuencias aleatorias, utilizadas para la generación de los diferentes eventos que se producen en el simulador.

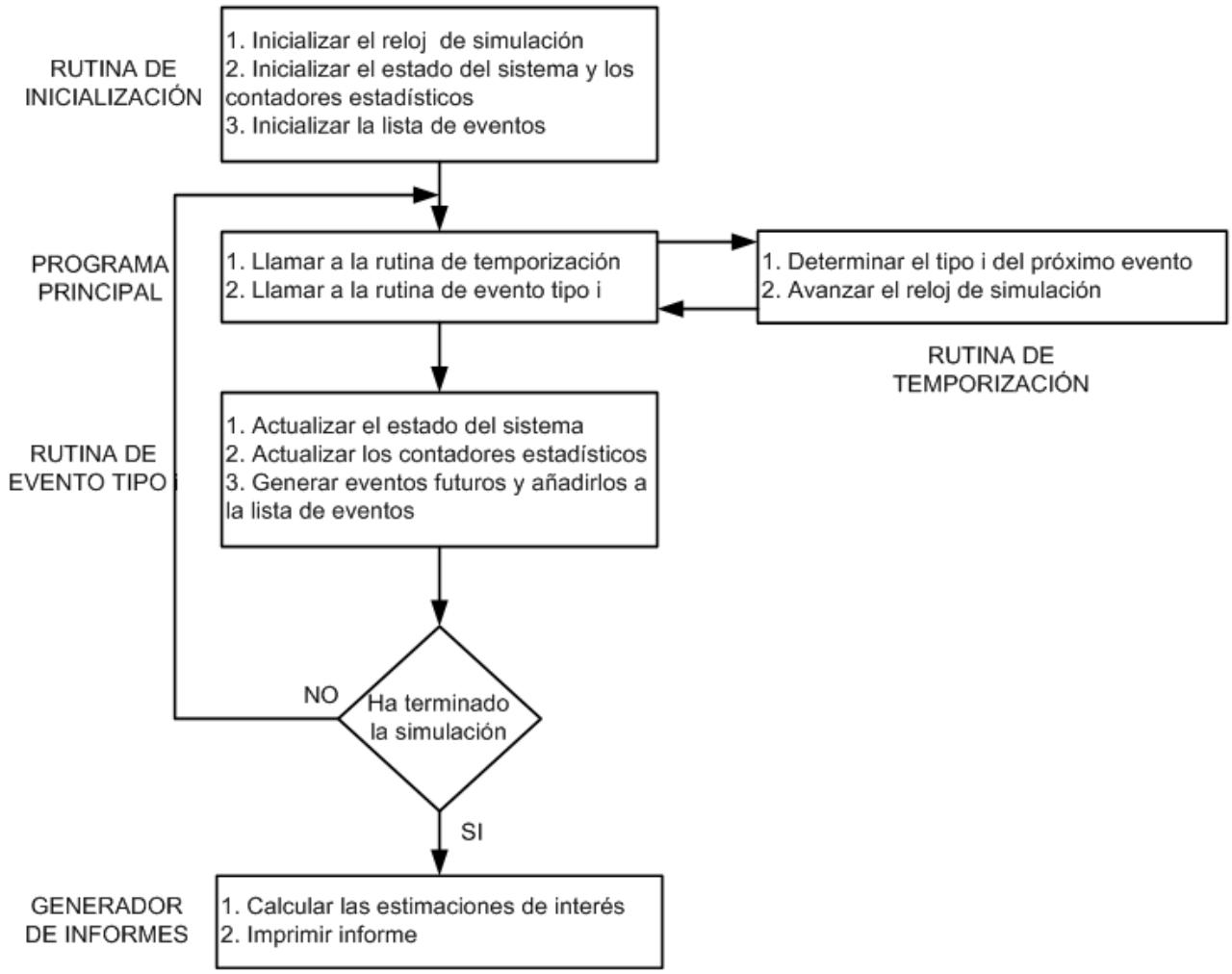


Figura 2.3: Estructura y evolución de un programa simulador.

6. *Selección del lenguaje de programación.* Tradicionalmente se pueden clasificar en dos grandes grupos de acuerdo a su estrategía: Programación de sucesos (*Event Scheduling* ES), Interacción de procesos (*Process Interantion* PI). Pero en la actualidad hay lenguajes que pueden admitir ambas estrategías; es por ello que una mejor clasificación de los lenguajes de simulación es la siguiente:

   - Lenguaje de propósito general.
     - FORTRAN, ALGOL, ASEMBLER, PL/1, C++, PASCAL, BASIC.
   - Lenguajes de simulación discreta.
     - Enfoque de flujo de transacciones: GPSS, WINQSB, BOSS.
     - Enfoque de eventos: GASPII, SIMSCRIPT, SIMCOM, SIMPAC.
     - Enfoque de procesos: SIMULA, OPL, SOL, SIMULATE.
     - Enfoque de actividades: CSL, ESP, FORSIM-IV, MILITRAN.
   - Lenguajes de simulación discreta y continua.
     - GASP-IV, C-SIMSCRIPT, SLAM.
   - Lenguaje de simulación continua.
     - Ecuaciones discretas: DSL-90, MINIC, GHSI, DYHYSYS.
     - Enfoque de bloques: MIDAS, DYNAMO, SCADS, MADBLOC, COBLOC.

   Entonces, la elección de alguno de estos software para el desarrollo de simulaciones, radica en que algunos de ellos son ampliamente utilizados para simular sistemas dinámicos y estocásticos, con un enfoque hacia la simulación de flujo de transacciones, el manejo y la compatibilidad en una amplia gama de ordenadores, resultando ideal para simular sistemas relacionados a líneas de espera.

Lenguajes Específicos de simulación.

En un principio las simulaciones se elaboran utilizando algún lenguaje de propósito general como FORTRAN, ALGOL, C++, PASCAL, etc. Para ello era necesario un gran trabajo de programación; pero con el paso del tiempo se fueron identificando diferentes situaciones, hasta llegar a estandarizarse ciertas instrucciones de programación en rutinas

bien definidas.

De este concepto nació el diseño de *lenguaje específico de simulación* con los cuales se pudo facilitar al usuario la programación de sus modelos. Es por ello que desarrollar un modelo mediante un lenguaje de simulación obviamente presenta algunas ventajas frente al uso de un lenguaje de propósito general. Algunas de estas ventajas son las siguientes:

a) El tiempo de desarrollo de la programación es muy corto porque se trata de lenguajes sintéticos basados en programación por bloques o subrutinas, e incluso algunos de ellos están encaminados a que los usuarios sólamente definan los parámetros y otros componentes que determinan el modelo a estudiar. Por eso no se requiere de usuarios con grandes conocimientos en programación.

b) Tienen una alta flexibilidad para hacer cambios. Esto nos permite realizar modificaciones en el modelo cuando se desarrollan distintos experimentos.

c) Integran funciones como generación de números aleatorios, análisis estadístico y gráficas.

d) Permite definir y entender el sistema a simular gracias a que se tiene una visibilidad superior de la estructura general del modelo y se aprecian más fácilmente las interrelaciones.

Sin embargo, el desarrollo de un modelo con un lenguaje específicos de simulación presenta ciertas desventajas en comparación a los lenguajes de propósito general como:

a) Es necesario invertir en la adquisión del software. Que por lo general no son lenguajes gratuitos, y por consiguiente representa un fuerte inversión.

b) Determinar la compatibilidad del equipo de computo con el software de simulación.

c) Los lenguajes de propósito general son más conocidos ya que los programadores conocen al menos uno. Pero no suelen conocer lenguajes de simulación.

7. *Programación.*

8. *Verificación del simulador.* Necesaria para corregir las divergencias de la implementación respecto al modelo de simulación.

9. *Validación del modelo de simulación.* Necesaria para evaluar la idoneidad del modelo de simulación como representación de estudio del sistema real.
   Evaluar un modelo significa desarrollar un nivel aceptable de confianza de modo que las inferencias obtenidas del comportamiento del modelo sean correctas y aplicables al sistema del mundo real. La validación y verificación es una de las tareas más importantes y difíciles que enfrenta la persona que desarrolla un modelo de simulación.
   La verificación se refiere a la comparación del modelo conceptual con el código computacional que se generó, para lo cual es necesario contestar preguntas como: ¿está correcta la codificación?, ¿son correctas las entradas de datos y la estructura lógica del programa?.
   Por otra parte, la validación consiste en demostrar que el modelo es una representación fiel de la realidad. Entonces un modelo validado es aquel que se ha probado ser una *abstracción* razonable del sistema real que intenta representar a través de éste.
   Al usar la simulación para estudiar un sistema complejo, existen varios tipos de errores que se cometen:

   - Errores de diseño.
   - Errores en la programación.
   - Errores en los datos utilizados.
   - Errores en el uso del modelo.
   - Errores en la interpretación de los resultados obtenidos.

   En el proceso de validación usualmente se emplean las pruebas estadísticas siguientes:

   - Prueba de estimaciones de los parámetros de la población asumiendo una distribución de probabilidad (pruebas F, t y z).
   - Pruebas de las estimaciones de los parámetros de la población que no son dependientes de la suposición de una distribución de población implícita (*prueba de medias Mann-Whitney*).

   - Pruebas para determinar la distribución de probabilidad de la cual proviene la muestra (*pruebas de bondad de ajuste de Kolmogorov-Smirnov o* χ^2).

Ejemplo 2.1.1. *La situación real de la empresa FATSA en cuanto a la producción de carburadores por día, de acuerdo con los datos de los últimos 8 días es la siguiente: 115, 105, 97, 96, 108, 104, 99 y 107. El modelo creado para la simulación de la planta arroja los siguientes 10 resultados de producción de carburadores por día: 110, 97, 100, 105, 108, 99, 118, 104, 105 y 103. ¿Son los resultados del modelo estadísticamente iguales a los reales?*

- *Hipótesis sobre la varianza.*
  $H_0 : V(Modelo) = V(real)$
  $H_1 : V(Modelo) \neq V(real)$
  $V(real) = 40.57$
  $V(Modelo) = 36.96$

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{40.57}{36.96} = 1.0977$$

  El F_c de tablas con 7 y 9 grados de libertad y con un nivel de rechazo de un 5 % es 3.29. Ya que F_0 es menor que F_c, se acepta que el modelo de simulación está arrojando resultados con la misma varianza que el sistema real.

- *Hipótesis sobre la media.*
  $H_0 : \mu(Modelo) = \mu(real)$
  $H_1 : \mu(Modelo) \neq \mu(real)$
  $E(real) = 103.87$
  $E(Modelo) = 104.9$
  El estadístico a utilizar es el correspondiente a varianzas iguales con poblaciones desconocidas y con media poblacional desconocida, puesto que solamente se tienen los datos de dos muestras.

$$t = \frac{\bar{x_1}}{\sqrt{\frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}}} - \frac{\bar{x_2}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

  Entonces

$$t = \frac{104.9}{\sqrt{\frac{8(14.57)+10(36.96)}{8+10}}} - \frac{103.87}{\sqrt{\frac{1}{8}+\frac{1}{10}}} = 0.3496$$

El estadístico t_c con $8 + 12 - 2 = 16$ grados de libertad y con el nivel de rechazo del 5 % es 1.746. Ya que t es mayor que t_c, se acepta que los resultados en cuanto a la producción de carburadores por día del simulador son estadísticamente iguales, en el caso de la media, a los de la producción real. En cuanto a la prueba de forma entre ambas muestras no se puede afirmar nada ya que la cantidad pequeña de datos que se está manejando imposibilita la formación de histogramas para realizarla.

Veamos un ejemplo de un sistema de una línea de espera con un servidor. El objetivo es estimar el número medio de clientes en el sistema.

Ejemplo 2.1.2. *Se suponen las siguientes hipótesis y datos:*

- *Tiempos entre llegadas de clientes según distribución F.*
- *Tiempos de servicio según distribución G.*
- *Distribución de tiempos independientes entre sí.*
- *T Tiempo máximo de simulación.*
- *La variable de estado es N el número de clientes en el sistema.*

Los eventos posibles son:

- *Llegada de un cliente al sistema.*
- *Final del servicio de un cliente. El mecanismo de transición se define como:*

$$N(t) = \begin{cases} N(t) + 1 & , si\ es\ llegada\ de\ un\ cliente \\ N(t) - 1 & , si\ es\ final\ de\ servicio \end{cases} \tag{2.1}$$

Otras variables auxiliares son:

- *TM reloj de simulación.*
- *DL tiempo entre llegadas $\stackrel{d}{=} F$.*
- *DS tiempo de servicio $\stackrel{d}{=} G$.*

- *TL instante de la próxima llegada.*
- *SUMA contador acumulando suma de productos de clientes en el sistema por tiempo de permanencia.*
- *TANT variable auxiliar (instante del último evento).*

El diagrama del modelo lo podemos ver en las figura 2.4 y 2.5.

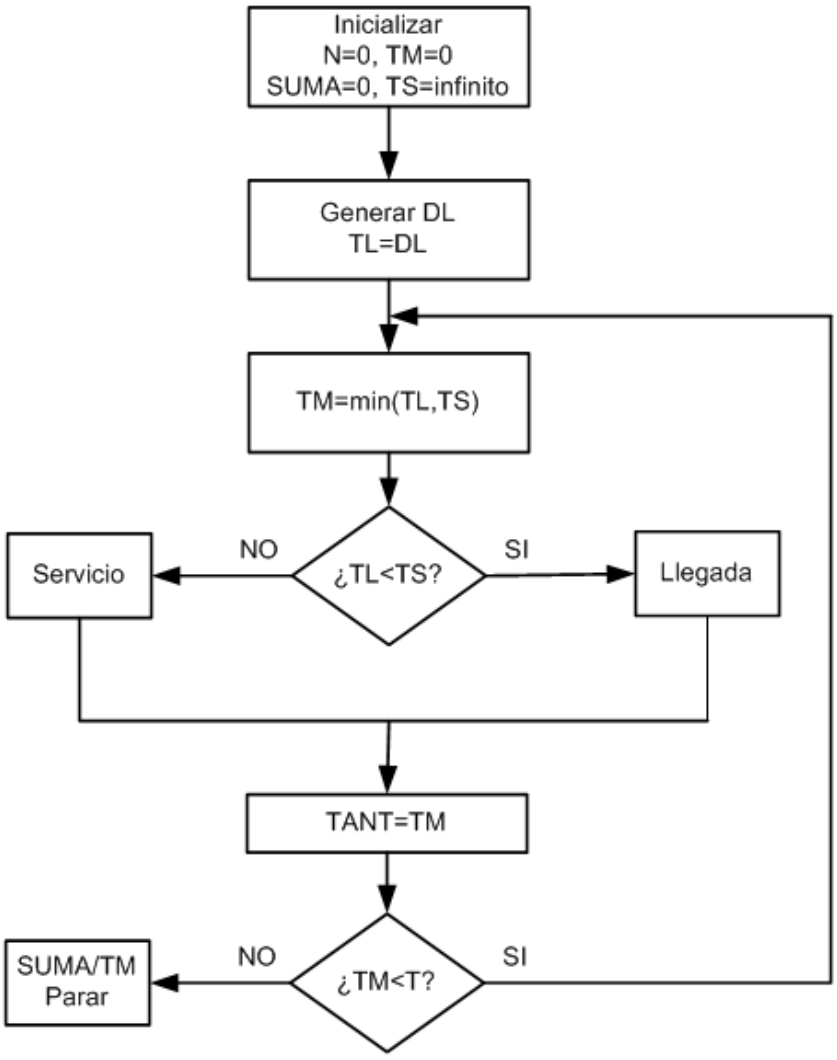


Figura 2.4: Diagrama general del modelo de simulación de una cola con un servidor.

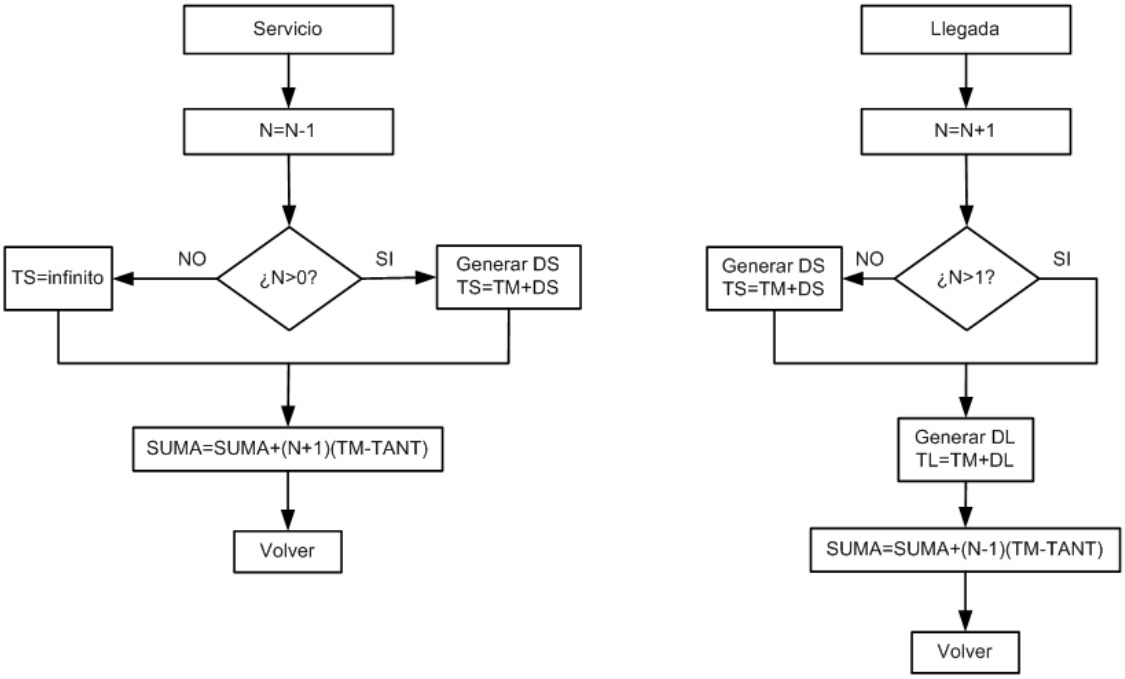


Figura 2.5: Diagrama detallado del modelo de simulación de una cola con un servidor.

A continuación (Ver cuadro 2.1), se presenta la traza del modelo. La traza es una tabla en la que se recogen los valores de las variables que intervienen en el modelo en varias iteraciones. Es útil obtener una traza manualmente para ayudar a obtener el modelo, de modo que se vean las variables que intervienen, los cálculos, etc. Gracias a ella es posible detectar la necesidad de variables auxiliares. Además, es una forma de verificar la programación posterior, de modo que una vez programado el modelo, con los mismos datos de entrada debe ejecutar lo que se recoge en la traza. En caso de que no sea así, puede ser por una programación errónea o incluso un modelado erróneo.

Veamos la ejecución del modelo de simulación. Se obtienen muestras de los tiempos entre llegadas DL dando lugar a estos valores: $3, 2, 5, 1, 2, 6, 6, 2, 8$ *y los tiempos de servicio DS:* $4, 1, 3, 1, 3, 2, 3, 5$. *Se simula durante* $T = 35$ *obteniendose la traza del cuadro 2.1.*

Gráficamente el número de clientes en el sistema a lo largo del tiempo se representa en la figura 1.6. Para el tiempo de simulación $T = 35$ *el número medio de clientes en el sistema resulta ser* $\hat{E}[N] = 31/35 = 0.89$. *Si el tiempo de simulación hubiera sido reducido a 18 en lugar de 35 el número medio de clientes en el sistema resultaría* $\hat{E}[N] = 20/18 = 1.11$, *lo cual pone de*

Cuadro 2.1: Traza del modelo.

Nº evento	Reloj simulación	Tipo evento	N	TL	TS	SUMA
0	0	Inicio	0	3	∞	0
1	3	Llegada	1	5	7	0+0*3=0
2	5	Llegada	2	10	7	0+1*2=2
3	7	Servicio	1	10	8	2+2*2=6
4	8	Servicio	0	10	∞	6+1*1=7
5	10	Llegada	1	11	14	7+0=7
6	11	Llegada	2	13	14	7+1*1=8
7	13	Llegada	3	19	14	8+2*2=12
8	14	Servicio	2	19	15	12+3*1=15
9	15	Servicio	1	19	18	15+2*1=17
10	18	Servicio	0	19	∞	17+3*1=20
11	19	Llegada	1	25	21	20+0=20
12	21	Servicio	0	25	∞	20+1*2=22
13	25	Llegada	1	27	28	22+0=22
14	27	Llegada	2	35	28	22+1*2=24
15	28	Servicio	1	35	33	24+2*1=26
16	33	Servicio	0	35	∞	26+1*5=31
17	35	final				31+0*2=31

manifiesto la importancia de la condición de finalización del proceso de simulación y su posible influencia en los resultados obtenidos.

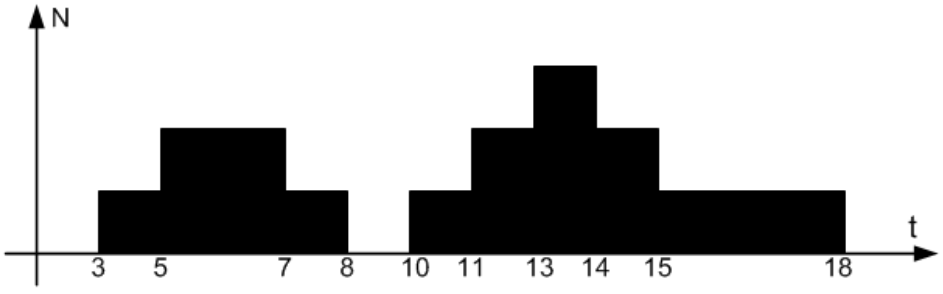


Figura 2.6: Número de clientes en el sistema a lo largo del tiempo.

2.2. Principios del modelado de la aleatoriedad en simulación

Como ya se ha comentado, los modelos de simulación habitualmente incluyen aleatoriedad, hasta el punto de que en muchos casos cuando el modelo es determinista se considera un caso particular de un modelo aleatorio más general. Por tanto, es necesario modelar correctamente la aleatoriedad incluida y disponer de procedimientos rápidos y eficientes para generar valores de variables aleatorias con una distribución determinada y conocida.

2.2.1. Variables aleatorias y sus propiedades

Un *experimento aleatorio* es un proceso cuyo resultado no se conoce con exactitud. El conjunto de resultados posibles de dicho experimento es lo que se conoce como *espacio muestral*, S. Los resultados concretos de realizar el experimento se denominan *muestras*. Un ejemplo puede ser el resultado de tirar un dado con seis caras. Su espacio muestral sería $S = \{1, 2, 3, 4, 5, 6\}$. Sin embargo, no todos los experimentos aleatorios tienen resultados numéricos; por ejemplo, al tirar una moneda el espacio muestral sería $S = \{cara, cruz\}$.

Una *variable aleatoria* X es una función que asigna un valor a cada resultado del experimento. Así por ejemplo, si al resultado del lanzamiento de una moneda le asignamos el valor de 1 si es cara y -1 si es cruz, tendremos una variable aleatoria cuyos posibles valores o *soporte* es el conjunto $S = \{-1, 1\}$. Las variables aleatorias son de gran importancia pues trasladan la probabilidad de un espacio no numérico a uno numérico con todas las ventajas que ello conlleva, entre otras las caracterizaciones y propiedades que se verán a continuación.

Las variables aleatorias que pueden ser de nuestro interés, se clasifican en: discretas, absolutamente continuas (en general, las llamaremos continuas) y mixtas.

Se dice que una variable aleatoria es *discreta* cuando puede tomar una cantidad numerable de valores.

Estas variables son caracterizadas, además de por la función de distribución, por la función de *masa de probabilidad*, que es una función que asigna a los posibles valores su probabilidad y al resto 0. Los valores de esta función han de estar entre 0 y 1 ya que son probabilidades y la suma de todas debe ser 1. La función de distribución $F(x)$ de este tipo de variables se calcula como la suma de la función de probabilidad para valores iguales o inferiores al valor x, y resulta por

lo tanto una función escalonada:

$$F(x) = \sum P(x = x_i); -\infty < x < \infty$$

Se dice que una variable aleatoria es *absolutamente continua* cuando existe una función, denominada *función de densidad*, *f(x)*, tal que su integral para un conjunto I determina la probabilidad de ocurrencia de dicho conjunto.

$$P(x \in I) = \int_I f(x)dx$$

Se ha de tener en cuenta que la integral de la función de densidad en todo el espacio debe ser la unidad.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

La función de distribución en este caso se calcula como la integral de la función de densidad en el intervalo $[-\infty, x]$, y de aquí que su derivada constituye la función de densidad asociada a la variable x. La función de distribución en este caso es continua en todos los puntos, y es derivable en todos excepto a lo sumo en una cantidad numerable de ellos.

$$F(x) = P(x \in [-\infty, x]) = \int_{-\infty}^{x} f(y)dy; -\infty \leq x \leq \infty$$

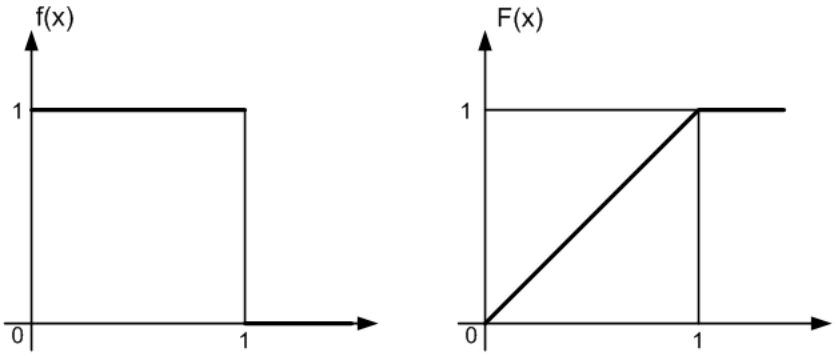
$$F'(x) = f(x)$$

Así por ejemplo para una variable aleatoria uniforme para el intervalo $[0, 1]$ se tiene que:

$$f(x) = \begin{cases} 1 & , 0 \leq x \leq 1 \\ 0 & , \text{resto} \end{cases} \tag{2.2}$$

$$F(x) = \int_0^x f(x)dx = \int_0^x 1dx = x$$

Las variables aleatorias *mixtas* son variables que en algunos intervalos se comportan como variables absolutamente continuas, pero que además concentran probabilidad en algún punto. Para ellas no existe otra caracterización teórica válida que no sea la función de distribución, que es continua en todos los puntos excepto en los que acumulan probabilidad que tiene un salto. Dentro del tratamiento conjunto de variables aleatorias, existe un concepto que es de máxima importancia: el concepto de *independencia*. Se dice que un conjunto de variables aleatorias es independiente si y sólo si la función de distribución conjunta es producto de sus distribuciones



marginales. Esta caracterización para variables absolutamente continuas puede darse diciendo que la función de densidad conjunta ha de ser el producto de las funciones de densidad marginales, y para variables discretas que la función de masa conjunta es el producto de las funciones de masa marginales.

Obsérvese que el concepto de independencia estadística que se acaba de definir no implica que no haya relación alguna entre las variables, sino que el conocimiento de unas no modifica la distribución de probabilidad de las otras. Por ejemplo, sea el experimento: sacar una carta de una baraja española y observar el palo y número de la carta que ha salido. Así sea la variable X el palo (numerando 1 a copas, 2 a oros, 3 a espadas y 4 a bastos), y la variable Y el número (sota es 8, caballo 9 y rey 10). La función de masa de ambas variables conjuntamente es $p(x, y) = 1/40, \forall x \in \{1, 2, 3, 4\}; \forall y \in \{1, \ldots, 10\}$. Las funciones de masa marginales son $p(x) = 1/4, \forall x \in \{1, 2, 3, 4\}$ y $p(y) = 1/10, \forall y \in \{1, \ldots, 10\}$.

Evidentemente, la función conjunta es producto de las marginales, luego, se puede decir que ambas variables son estadísticamente independientes, es decir, si dada una carta se sabe cuál es el palo no modifica la probabilidad de cuál será su número.

Es importante tener presente este concepto de independencia pues en simulación, y en general en probabilidad, se maneja muy habitualmente.

A continuación se presenta una revisión de las características que se consideran más relevantes de una variable aleatoria.

La *esperanza* o *valor esperado* o *media* de una variable X, denotada por μ o $E[X]$, es una medida central de comportamiento de la variable que representa su centro de masa. Se calcula de manera distinta dependiendo de si se trata de una variable discreta o continua.

$$E[X] = \mu = \begin{cases} \sum_{x_i} x_i P[x_i] & , X \text{ variable discreta} \\ \\ \int_{\mathbb{R}} x f(x) dx & , X \text{ variable continua} \end{cases} \tag{2.3}$$

La esperanza por definición es un operador lineal, es decir, $E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i E\left[X_i\right]$. Por otra parte, obsérvese que la esperanza de una variable puede ser un valor que no pertenezca al soporte, especialmente en variables aleatorias discretas, aunque siempre se encontrará entre el mínimo y el máximo de los posibles valores.

Otra medida central de una variable aleatoria es la *mediana*. Se denota por $x_{0.5}$ y se define como el menor valor de la variable para el cual $F_X(x) \geq 0.5$. Cuando la variable tiene asimetría es habitual dar este valor de medida central, ya que la media se suele ver muy afectada por los valores más extremos.

La *varianza* de una variable aleatoria es una medida para determinar el nivel de dispersión que tienen los datos de la variable con respecto a su media. Se denota por $var(X)$ o σ^2, y se definde como la media de las desviaciones a la esperanza al cuadrado. Así su definición y un método alternativo de cálculo son:

$$\sigma^2 = E[(X_i - \mu_i)^2] = E(X_i^2) - \mu_i^2$$

Por definición, la varianza es siempre no negativa y es un operador cuadrático, de modo que $var(cX) = c^2 var(X)$. Por otra parte, para transformaciones lineales, sólo en el caso en que las variables sean independientes o incorreladas (término que se explica acontinuación) se cumple que $var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i)$.

Obsérvese que la varianza va en unidades al cuadrado; para dar una medida relativa a las unidades que están manejando se utiliza la *desviación típica* o *estándar*, que se define cómo la raíz cuadrada de la varianza de la variable y se denota por σ.

Una medida de la dependencia lineal entre dos variables X_i y X_j viene dada por la *covarianza* denotada por C_{ij} o $cov\left(X_i, X_j\right)$. Se define como la esperanza o media de la multiplicación de las desviaciones de la variable X_i con respecto a su media μ_i por las desviaciones de la variable X_j con respecto a su media μ_j. Así la definición y un método alternativo de cálculo son las siguientes:

$$cov\left(X_i, Xj\right) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = E(X_i X_j) - \mu_i \mu_j$$

Las covarianzas son medidas simétricas, es decir, $C_{ij} = C_{ji}$. Si $i = j$, entonces la covarianza es igual a la varianza de la variable. Si el valor de la covarianza es positivo, ello implica que X_i y X_j están correladas positivamente, es decir, si el valor que se tiene de X_i es mayor que su media μ_i, entonces el valor que se espera para X_j será mayor que su media μ_j. Por el contrario si el valor de la covarianza es negativo, ello implica que X_i y X_j están correladas negativamente, es decir, si el valor que se tiene de X_i es mayor que su media μ_i, entonces el valor que se espera para X_j será menor que su media μ_j.

La covarianza tiene el inconveniente de que es una medida dimensional que no nos permite cuantificar la correlación entre variables de una forma directa. Para obtener una medida adimensional de dicha correlación se utiliza el *coeficiente de correlación lineal*, ρ_{ij}. Dicho factor adquiere un valor entre -1 y 1 y se calcula como cociente entre la covarianza entre dos variables y el producto de sus desviaciones típicas respectivas.

$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j}, -1 \leq \rho_{ij} \leq 1$$

Como puede observarse lleva el signo de la covarianza, y su interpretación es por lo tanto la misma.

Aún así, se suele utilizar otro factor para medir el grado de relación lineal entre dos variables: el *coeficiente de determinación*, ρ_{ij}^2. Este coeficiente representa la proporción de la variabilidad (Varianza) de una variable que es explicada por la variabilidad de la otra. Se calcula como el cuadrado del coeficiente de correlación lineal, con lo cual lo que nos indica es el signo de la relación pero su interpretación es mucho más intuitiva.

Dos variables se dicen *incorreladas* si el coeficiente de correlación lineal (o el de determinación o la covarianza) es cero. Si dos variables son independientes entonces son incorrelada, al revés no es siempre cierto, excepto en distribuciones normales.

Un *proceso estocástico*, $\{X_t\}_{t \in T}$, es un conjunto de variables aleatorias definidas todas sobre un mismo espacio de probabilidad (intuitivamente se diría que miden lo mismo pero en distinto instante de tiempo o punto espacial). Se denomina *espacio de estados* al conjunto de posibles valores de las variables.

Al conjunto T se le denomina *conjunto de índices*, y si se trata de un conjunto numerable de $\mathbb{R}$, el proceso es denominado *proceso en tiempo discreto*, mientras que si no es un conjunto numerable se denomina *proceso en tiempo continua*.

2.2.2. Números aleatorios

Las metodologías de generación de números aleatorios tienen una larga e interesante historia, desde los métodos manuales más primitivos *lanzamiento de dados o extracción de cartas* hasta los actuales (basados en la utilización de computadores) se ha recorrido un largo camino. Sin embargo, algunos de aquellos métodos antiguos todavía siguen vigentes; sin ir más lejos la selección de ganadores de algunas loterías siguen realizándose mediante la extracción de una o más bolas de entre un conjunto de ellas introducidas en un bombo.

La evolución en la generación de números aleatorios vino marcada por la aplicación de sistemas mecánicos y electrónicos hasta llegar a las más recientes generaciones algorítmicas. A finales de los años 30 Kendall y Babington-Smith (1938) obtenían tablas de 100,000 números aleatorios a partir de un disco giratorio y, unos años después, surgían los primeros métodos electrónicos. Estos métodos se basaban en tubos de vacío aleatoriamente pulsantes y conseguían generar hasta cincuenta números aleatorios por segundo. Una de estas máquinas pioneras, ERNIE (*Electronic Random Number Indicator Equipment*), fue utilizada por la oficina de correos Británica para la selección de ganadores en uno de sus juegos de lotería (Thomson, 1959). Otra máquina basada en la misma filosofía, fue utilizada por la Rand Corporation (Corporation, 1955) para generar la primera tabla de un millón de números aleatorios.

Con el rápido crecimiento de las tecnologías asociadas a los computadores, y consecuentemente de la simulación, cobraron mayor interés aquellos métodos de generación compatibles con la forma de trabajar de un computador: la utilización de tablas donde se almacena una secuencia de números o bien la generación algorítmica de una secuencia pseudoaleatoria (secuencia perfectamente reproducible). La utilización de tablas presenta dos graves inconvenientes que provoca que su uso se vea relegado en favor de los generadores algorítmicos: por un lado la gran cantidad de memoria necesaria para el almacenamiento de la tabla y, por otro, la segura repetición de la secuencia en el caso de trabajar con simulaciones largas.

Uno de los primeros algorítmos de generación de secuencias de números pseudoaleatorios del que se tiene noticia fue el propuesto por Von Neumann y Metropolis en los años cuarenta. El método consistia en lo siguiente:

1. Se establece un número de cuatro dígitos, Z_0, que llamaremos semilla.

2. Cada valor Z_{i+1} se obtendrá a partir del valor Z_i tomando los cuatro dígitos centrales del número Z_i^2. Si el número Z_i^2 no tiene ocho dígitos se completará con ceros por su izquierda

hasta que sí los tenga y entonces es cuando se selecciona el valor Z_{i+1}.

Aunque intuitivamente parece ser un buen método, la secuencia presenta una fuerte tendencia a cero, de la que, además, no se recupera (Von Neumann, 1951).
En la actualidad debido a su buen comportamiento, los *generadores congruenciales lineales* y los *múltiplemente recursivos*, como su extensión lógica, son los más habituales en la generación de secuencias de números pseudoaleatorios.

Requisitos de un generador de números pseudoaleatorios.

Sobre la definición de aleatoriedad (Niederreiter, 1978; Ripley 1987) y que las características de un generador de números pseudoaleatorios ha de satisfacer para poder considerarlo un generador adecuado (Park y Miller, 1988) se ha escrito y debatido abundantemente; sin embargo, parece ampliamente aceptado que las cuatro características siguientes son deseables en un generador de números pseudoaleatorios:

- *Aleatoriedad.* Deberá generar números distribuidos de forma aproximadamente uniforme en el intervalo $[0, 1]$. Además no debe haber correlación entre las muestras.

- *Eficiencia.* La eficiencia se mide en tiempo y en cantidad de memoria; será deseable que las muestras se generen lo más rápido posible y que el generador precise de poca memoria para hacerlo.

- *Período máximo.* Es útil que la secuencia que se genere sea lo mayor posible, para reducir la posibilidad de que se produzca una repetición durante la simulación. El período máximo establece el número de muestras que se puede obtener antes de repetir la secuencia.

- *Secuencia producible.* El tener una secuencia que se pueda reproducir tiene como principal ventaja el poder repetir exactamente la misma simulación para facilitar las tareas de depuración y verificación de los programas. Además, podremos utilizar la misma secuencia para varias simulaciones y mejorar la precisión de los resultados mediante algunas técnicas de reducción de la varianza, expuestas más adelante.

2.2.3. Generadores congruenciales lineales de números pseudoaleatorios

Estos generadores fueron introducidos por Lehmer (1951) y todos ellos tienen en común que la secuencia que generan satisface la expresión recursiva:

$$Z_i = (a \cdot Z_{i-1} + c) \mod \text{m} \qquad 0 \leq Z_i \leq m-1,$$

donde la selección de a, c y m caracteriza unívocamente al generador, todos ellos incluido el Z_0, son números enteros no negativos:

- a: multiplicador que deberá satisfacer $a < m$
- c: incremento que cumpla que $c < m$. La elección de este parámetro implicará la inclusión del generador dentro de uno de los dos siguientes subconjuntos:
  - $c = 0 \Longrightarrow$ **Generador congruencial multiplicativo.**
  - $c \neq 0 \Longrightarrow$ **Generador congruencial mixto.**
- m: se denomina módulo y satisface que $m > 0$. Era usual su elección como potencia de 2 para una mayor eficiencia en los cálculos.
- Z_0: recibe el nombre de semilla o valor inicial de la secuencia y deberá ser tal que $Z_0 < m$. Los números pseudoaleatorios que se obtienen serán de la forma:

$$U_i = \frac{Z_i}{m}, \ 0 \leq U_i \leq 1.$$

La secuencia se repetirá con período $p \leq m$, por lo que el generador alcanza el **período máximo** si $p = m$.
Su carácter de repetividad, común a todos los generadores de números pseudoaleatorios, que establece que cualquier número generado Z_i está totalmente determinado por los valores que caracterizan el generador $-m, a, c-$ y por la semilla escogida Z_0:

$$Z_i = \left(a^i Z_0 + \frac{c \cdot (a^i - 1)}{a - 1} \right) \mod \text{m}$$

Es precisamente el origen de los inconvenientes del método.
De todas formas, es necesario notar que, si se eligen los parámetros de forma adecuada, la

secuencia obtenida podrá superar diferentes test estadísticos que valoren positivamente el comportamiento de los U_i como variables aleatorias independientes e idénticamente distribuidas (Fishman, 1978; Knuth, 1997).
Otro inconveniente que se le puede achacar a esta metodología de obtención de números pseudoaleatorios es que los U_i que se obtienen son de la forma i/m, lo que podría llevar a creer que estamos ante una secuencia de valores discretos. Sin embargo, dado que el valor de m suele ser de orden 10^9, en realidad se puede considerar que los U_i son densos en $\mathbb{R}$; es más, de considerarse necesario, se pueden utilizar los bits de varios Z_i para construir eficientemente una muestra $U(0,1)$ con la presición dada por la presentación de los números reales del computador[1].

Generadores congruenciales mixtos

Las condiciones siguientes, teorema de Hull y Dobell (1962), aseguran que el generador lineal congruencial mixto que las satisfaga tendrá período máximo:

a) El único entero positivo que exactamente divide a m y a c es el 1, es decir, son primos entre sí.

b) Si q es un número primo que divide a m, entonces q también divide a $(a-1)$.

c) Si 4 divide a m entonces 4 también divide a $(a-1)$.

Sin embargo, además del período máximo, es necesario tener en cuenta otros factores para la elección de los parámetros del generador, como sus propiedades estadísticas y una mayor sencillez de los cálculos necesarios para obtener la secuencia.
En la tabla siguiente se indican los generadores congruenciales lineales mixtos propuestos por Coveyou y MacPherson (1967) y por Kobayashi (1978);

Parámetros	**Kobayashi**	**Coveyou y MacPherson**
a	314,159,269	5^{15}
b	453,806,245	1
m	2^{31}	2^{31}

[1]La técnica consiste en introducir los bits en la variable en coma flotante, forzar exponente entre 1 y 2, y restar 1.

Generadores congruenciales multiplicativos

Los generadores congruenciales multiplicativos son anteriores a los mixtos y, por este motivo, han sido más estudiados. La ausencia del parámetro c puede considerarse una ventaja, pues permite búsquedas exhaustivas completas del mejor parámetro a para un m dado, aunque no pueden tener período máximo ya que nunca tendrán $Z_i = 0$ y, obviamente, no se satisface la condición (a) del teorema de Hull y Dobell (1962); sin embargo, sí pueden llegar a alcanzar el período $m - 1$ si se seleccionan m y a de forma adecuada:

- m ha de ser un número primo.
- a ha de ser raíz primitiva de m, es decir, ha de satisfacer:

$$a^n \mod m \neq 1 \quad n = 1, \ldots, m - 2.$$

Seleccionando $m = 2^{31} - 1$, número primo ya propuesto en su día por Lehmer (1951), existen 534,600,000 raíces primitivas de m y, por lo tanto, posibles candidatas para parametizar un generador multiplicativo. En este caso, a será raíz primitiva si:

$$a = 7^b \mod m \quad \text{con } b \text{ y } m - 1 \text{ son primitivos entre sí.}$$

Fijándose ahora en la eficiencia de la implementación se puede exponer el problema como sigue:

- $m = a \cdot q + r \Longrightarrow q = m \div a$ donde $r = m \bmod a$.
- Si $r < q$ -Satisfecho por 23,093 raíces primitivas- entonces podemos hallar Z_i utilizando el siguiente algoritmo:

Algoritmo 1. *Generador congruencial lineal multiplicativo.*

▷ $h = Z_{i-1} \div q;$

▷ $\iota = Z_{i-1} \ mod \ q;$

▷ $t = a \cdot \iota - r \cdot h;$

▷ ***if*** $t > 0$ ***then***

▷ $Z_i = t;$

▷ ***else***

▷ $Z_i = t + m;$

▷ ***end if***

Fishman y Moore (1986) sometieron a diferentes test estadísticos a los 23,093 posibles valores de a, concluyeron que el generador congruencial lineal multiplicativo de mejor comportamiento es el siguiente[2]:

Parámetros	**Valor**
a	48,271
q	44,488
r	3,399

2.2.4. Generadores múltiplemente recursivos

La secuencia de números pseudoaleatorios obtenida tras la aplicación de un generador **MRG** (*Múltiple Recursive Generators*) satisface la siguiente relación recursiva (Knuth, 1997; L'Ecuyer, 1994):

$$Z_n = (a_1 Z_{n-1} + \cdots + a_K Z_{n-K}) \mod m \tag{2.4}$$

de forma que m y K son enteros positivos y los coeficientes $a_i \in \{0, 1, \ldots, m-1\}$. El máximo período conseguido con este tipo de generadores es $P = m^k - 1$, y éste sólo se alcanza si m es primo y el polinomio característico de la relación (2.4):

$$P(z) = z^K - a_1 z^{K-1} - \cdots - a_K$$

es primitivo. Para obtener un polinomio primitivo se necesitan al menos que dos de los coeficientes a_i sean no nulos, de esta forma la relación (2.4) mínima será:

$$Z_n = (a_r Z_{n-r} + a_K Z_{n-K}) \mod m$$

Vemos entonces que este tipo de generadores permiten obtener secuencias de números aleatorios mucho mayores que las que se obtenían con los generadores congruenciales lineales. De todas formas, no basta con tener un período elevado, también es preciso que la secuencia generada

[2]En este caso, para un valor de semilla $Z_0 = 1$, el valor 10,000 será $Z_{10,000} = 399,268,537$

sea de calidad (buen comportamiento estadístico) y que permita una implementación eficiente. Tanto este tipo de generadores como los generadores congruenciales lineales son susceptibles de mejorarse en cuanto a su comportamiento estadístico mediante la composición de varios generadores.

Composición de generadores

Con el objetivo de mejorar el comportamiento estadístico de los generadores, se han venido realizando diferentes estudios que tratan de combinar diferentes generadores de números aleatorios: tanto generadores congruenciales, como generadores múltiplemente recursivos.
L'Ecuyer (1996) propuso y analizó uno de estos generadores, basándose en la combinación de J $MRGs$, todos ellos con el mismo orden K y con diferentes módulos m_j. Cada uno de los $MRGs$ tendrá la forma siguiente:

$$Z_{j,n} = (a_{j,1}Z_{j,n-1} + \cdots + a_{j,K}Z_{j,n-K}) \mod m_j \tag{2.5}$$

Suponiendo que los parámetros de cada una de las recurrencias de la Ecuación (2.5) tenga período máximo $P_J = m_j^K - 1$, entonces el período máximo que se puede conseguir para el generador resultante de combinar los J $MRGs$ viene dado por:

$$P = \frac{P_1 \cdot P_2 \cdots P_j}{2^{j-1}}.$$

La combinación de los $Z_{j,n}$ para obtener un elemento de la secuencia a generar Z_n viene dada por la siguiente relación:

$$Z_n = \left(\sum_{j=1}^{j} \delta_j Z_{j,n} \right) \mod m_1,$$

donde los δ_j son valores enteros satisfaciendo que cada uno de ellos es primo relativo con su correspondiente m_j.

2.2.5. Método de los cuadrados medios

Se toma un número al azar, x_0, de $2n$ cifras. Se eleva al cuadrado y se toman de este resultado las $2n$ cifras centrales. Se repite el proceso.

Ejemplo 2.2.1.

$$\begin{aligned} x_0 &= 4122 \qquad x_0^2 = 16|9908|84 \\ x_1 &= 9908 \qquad x_1^2 = 98|1684|64 \\ x_2 &= 1684 \qquad x_2^2 = 02|8358|56 \end{aligned}$$

La secuencia 4122, 9908, 1684, ... puede ser considerada, al menos a partir de un cierto intervalo, como una secuencia de número pseudoaleatorios.

El principal problema de este método es que los números pueden repetirse a partir de una secuencia muy corta. Por ejemplo, si $x_0 = 3708$ se tiene que $x_4 = 6100 = x_8$, lo cual no es controlable.

2.2.6. Método de Lehmer

Sea x_0 un número al azar de n cifras. Se multiplica por otro k' (fijo del generador) de k cifras, dando lugar a uno de $n + k$. Se quitan las k cifras de la izquierda, obteniendo uno de n cifras al que se resta el de k cifras que se había separado.

Ejemplo 2.2.2.

$$\begin{aligned} x_0 = 4122 \quad k' = 76 \quad 4122 \cdot 76 = 31|3272 \quad 3272 - 31 = 3241 \\ x_1 = 3241 \quad k' = 76 \quad 3241 \cdot 76 = 24|6316 \quad 6316 - 24 = 6292 \end{aligned}$$

Este método acaba degenerando a 0.

2.3. Generación de muestras de variables aleatorias

Una vez que se ha seleccionado una distribución para modelar la aleatoriedad de una variable de entrada, es necesario establecer procedimientos para obtener valores de esta variable durante la simulación del modelo. Así pues esta sección se dedica a la simulación de variables aleatorias con una distribución determinada.
Se empieza por generar valores de variables aleatorias con distribución uniforme en el intervalo $(0, 1)$. La razón es que todos los métodos para generar variables aleatorias no son más que transformaciones de variables aleatorias con esta distribución y por ello el primer paso es saber

obtener ésta.

Los mejores métodos para generar números aleatorios son los físicos y de entre ellos el mejor es el de la ruleta. Este método consiste en una ruleta dividida en 10 partes iguales, a las que se les asignan los valores del 0 al 9, y una flecha fija en un punto fuera de la ruleta. Si se la hace girar y posteriormente se la detiene bruscamente, se puede anotar el número que señala la flecha. Repitiendo esta operación n veces, se obtiene una secuencia de n números que obviamente constituye una secuencia de números aleatorios.

Estos valores se pueden sonsiderar valores de una variable cuya distribución sea una uniforme discreta que toma valores de 0 a 9 con probabilidad 1/10 cada uno de ellos. Pero también los podemos agrupar de k en k y considerar que son valores de una uniforme discreta que toma valores $0, \ldots, 10^k - 1$.

Si se desean generar valores de una uniforme en el intervalo de $(0, 1)$, podemos agrupar los valores de k en k y considerar que cada grupo son las cifras decimales de una realización de una variable aleatoria con distribución $U(0, 1)$.

Obviamente, para considerar informativos los resultados obtenidos mediante la simulación de un modelo es necesario simularlo más de una vez. De hecho, se debe hacer una gran cantidad de veces, en general, cuanto más complicado es el modelo, lo que hace ver la necesidad del uso del ordenador.

Existen tablas de números aleatorios obtenidos por el método de la ruleta y otros métodos físicos, pero no es un buen método para su uso en ordenador.

Ésta fue la razón por la que se crearon métodos aritméticos particulares adaptados al ordenador, aunque con un cierto deterioro de la aleatoriedad, denominandose *números pseudoaleatorios*.

Un generador ideal de números pseudoaleatorios debe proporcionar secuencia de números con las siguientes propiedades:

- Tener distribución uniforme (son realizaciones de uniformes)
- Ser estadísticamente independientes
- Han de ser reproducibles
- Capaces de producir diferentes secuencias de números
- Deben tener un ciclo no repetitivo tan largo como se desee

- Ser generados rápidamente
- Ocupar poca memoria o almacenamiento en el ordenador

En las secciones prescedentes vemos algunos métodos para generar números pseudoaleatorios, mostrando los diferentes tipos de variables que podemos encontrar en un estudio en particular.

2.3.1. Generación de variables aleatorias discretas

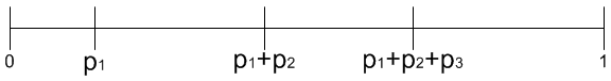
Sea una variable aleatoria con una distribución discreta que toma ciertos valores con determinada probabillidad

$$X = \begin{cases} x_1 & \text{con prob } p_1 \\ x_2 & \text{con prob } p_2 \\ x_3 & \text{con prob } p_3 \\ \vdots & \end{cases}$$

siendo

$$\sum_k p_k = 1.$$

La idea intuitiva del procedimiento es dividir el intervalo (0,1) en tantos subintervalos como valores puede tomar la variable y de tamaño las probabilidades de éstos. Generar un valor uniformemente distribuido en (0,1) y observar en qué subintervalo se encuentra y asignar a la variable el valor correspondiente a ese subintervalo.

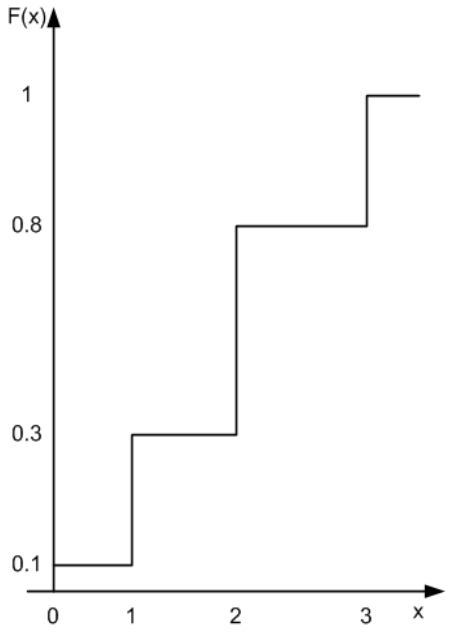


Más formalmente, consiste en generar un número aleatorio uniformemente distribuido $u \in U(0,1)$, entonces $X = x$, si $\sum_{k=1}^{i-1} p_k \leq u < \sum_{k=1}^{i} p_k$, es decir, si $F_x(x_{i-1}) \leq u < F_x(x_i)$.

Ejemplo 2.3.1. *Sea*

$$X = \begin{cases} 0 & con\ prob \quad p_1 = 0.1 \\ 1 & con\ prob \quad p_2 = 0.2 \\ 2 & con\ prob \quad p_3 = 0.5 \\ 3 & con\ prob \quad p_4 = 0.2 \end{cases}$$

con función de distribución



Para la secuencia de números aleatorios 0.27, 0.54, 0.06, 0.89 y 0.15, la variable tomará los valores siguientes: 1,2,0,3 y 1.

Vamos a ver métodos concretos para algunas distribuciones, binomial y geométrica, sin olvidar que el método anterior sirve para cualquier distribución discreta. Para las demás distribuciones conocidas también hay métodos basados en su definición o en algunas propiedades.

Binomial(n, p)

Se basa en la propiedad según la cual la distribución de la suma de v.a.i.i.d. con distribución Bernoulli de parámetro p es Binomial(n, p). Luego, para generar valores de una variable con

distribución Binomial(n, p), se pueden generar n Bernoulli de parámetro p y sumarlas; es decir:

$$X \stackrel{d}{=} Bin(n,p) \Longleftrightarrow X = \sum_{i=1}^{n} X_i$$

donde cada X_i son v.a.i.i.d. $Ber(p)$.
Algoritmo.

1. $x \longleftarrow 0$
2. Hacer n veces
   Generar $u \in U(0,1)$
   Si $u \leq p$ $x \longleftarrow x + 1$
3. Salida: X se distribuye según $Binomial(n,p)$

Geométrica(p)

La distribución geométrica correspondiente al número de ensayo en que aparece el primer éxito al repetir un experimento de Bernoulli de parámetro p. Así que para generar valores con esa distribución es posible hacerlo con el método tradicional o aprovechando esta propiedad. Así, sabiendo que:

$$P(X = x) = (1-p)^{x-1}p \quad x = 1, 2, 3, \ldots$$

Podemos obtener su función de distribución:

$$\begin{aligned} F_X(x) &= \sum_{j=1}^{x} P(x=j) = p\sum_{j=1}^{x}(1-p)^{j-1} = p\frac{1-(1-p)^{x-1}(1-p)}{1-(1-p)} \\ &= p\frac{1-(1-p)^x}{p} \\ &= 1-(1-p)^x \quad x = 1, 2, \ldots \end{aligned}$$

Siguiendo el método general,

$$\begin{aligned} X \longleftarrow x &\iff F_X(x-1) \leq u < F_X(x) \\ &\iff 1-(1-p)^{x-1} \leq u < 1-(1-p)^x \\ &\iff (1-p)^{x-1} \geq 1-u > (1-p)^x \\ &\iff (x-1)ln(1-p) \geq ln(1-u) > xln(1-p) \\ &\iff (x-1) \leq \frac{ln(1-u)}{ln(1-p)} < x \\ x &= 1+\left\lceil \frac{ln(1-u)}{ln(1-p)} \right\rceil \end{aligned} \tag{2.6}$$

para ver que la expresión sería

$$x = 1+\left\lceil \frac{ln(1-u)}{ln(1-p)} \right\rceil \stackrel{d}{=} 1+\left\lceil \frac{ln(u)}{ln(1-p)} \right\rceil,$$

Utilizando la propiedad que la relaciona con los experimentos de Bernoulli el algoritmo sería:

1. $x \longleftarrow 0$
2. Hacer hasta que $u \leq p$
   $x \longleftarrow x+1$
   Generar $u \in U(0,1)$
3. Salida: X se distribuye según Geométrica(p)

Este método es bueno cuando el valor de p es grande.

2.3.2. Generación de variables aleatorias absolutamente continuas

Método de la transformada inversa

Sea X La variable aleatoria cuya función de distribución es $F(x) = P(X \leq x)$. Se genera un número aleatorio uniforme entre 0 y 1, u, y luego se determina x tal que $F(x) = u$.
Supongamos que la variable tiene *distribución exponencial* con $F(x) = 1 - e^{-\lambda x}$ para $x \geq 0$ siendo $1/\lambda$ la media de la distribución.

Dado un número aleatorio u tal que $F(x) = u$, luego

$$\begin{aligned} X &= F_X^{-1}(U) \\ u &= 1 - e^{-\lambda x} \\ 1 - u &= e^{-\lambda x}, \quad u \stackrel{d}{=} 1 - u \\ -\lambda x &= ln(u) \\ X &= -\frac{ln(U)}{\lambda} \end{aligned}$$

Otra aplicación directa de este procedimiento es para la *distribución uniforme* en un intervalo cualquiera $(0, 1)$. La función de densidad de probabilidad para este caso es $f(x) = \frac{1}{b-a}$, para $a < x < b$, por tanto su función de distribución es: $F(x) = \frac{x-a}{b-a}$, para $x \in (a, b)$, (0 para valores menores y 1 para valores mayores) y dado un número aleatorio u tal que $F(x) = u$, se tiene que:

$$\begin{aligned} X &= F_X^{-1}(U) \\ u &= \frac{x-a}{b-a} \\ X &= U(b-a) + a \end{aligned}$$

La *distribución de Weibull*(α, β) es otra distribución para la que se puede aplicar este procedimiento directamente. La función de densidad de una distribución de $Weibull(\alpha, \beta)$ de media $\left(\frac{1}{\alpha\beta}\right)\Gamma(\frac{1}{\alpha})$, es

$$f(x) = \alpha\beta^\alpha x^{\alpha-1} e^{-(\beta x)^\alpha}, \quad x \geq 0,$$

con lo que la función de distribución se obtiene de forma inmediata

$$F(x) = 1 - e^{-(\beta x)^\alpha} \quad x \geq 0.$$

Así, dado un valor aleatorio uniforme en $(0, 1)$, u, el valor generado sería:

$$\begin{aligned} X &= F_X^{-1}(U) \\ u &= 1 - e^{-(\beta x)^\alpha} \\ ln(u) &= -(\beta x)^\alpha \\ [-ln(u)]^{\frac{1}{\alpha}} &= \beta x \\ X &= \frac{1}{\beta}[-ln(U)]^{\frac{1}{\alpha}} \end{aligned}$$

Aunque éste es el procedimiento más extendido, sin embargo, muestra una dificultad fundamental para su aplicación, la necesidad de conocer explícitamente la función de distribución. La forma habitual de caracterizar una distribución absolutamente continua es mediante su función de densidad, de ahí que se hayan diseñado otros procedimientos basados en esta función.

Método de Aceptación-Rechazo.

Se trata de un método general para variables absolutamente continuas. Existen dos versiones, la primera es más sencilla y con más limitaciones, pero tambien más utilizada precisamente por su sencillez.

Método simple de rechazo.

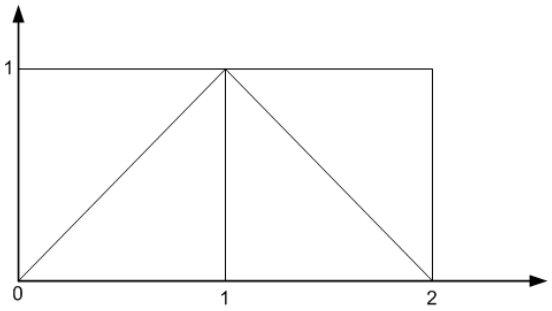
Se quieren obtener valores de una variable aleatoria con función de densidad $f(x)$ cuyo soporte es un intervalo acotado (a_1, a_2). Sea c un valor tal que $c \geq max\{f(x) : x \in (a_1, a_2)\}$.
La idea básica es generar puntos uniformemente en el rectángulo de base (a_1, a_2) y altura $(0, c)$. Si el punto está por encima de la curva el punto es rechazado y habrá que generar otro y si está por debajo se acepta su coordenada x como valor de una variable aleatoria con función de densidad $f(x)$.
El procedimiento es:

1. Generar $u_1, u_2 \in U(0,1)$
   Calcular $x = a_1 + (a_2 - a_1)u_1$
   Calcular $y = cu_2$
2. Calcular $f(x)$. Si $y > f(x)$ ir a 1.
3. Salia: X toma el valor x que se distribuye según $f(x)$

Obsérvese que P(aceptar un valor dado por (x, y)) $= \dfrac{1}{c(a_2 - a_1)}$, por lo tanto el valor de c es deseable que sea lo más pequeño posible. En particular, siempre que sea fácil de obtener se toma $c = max\{f(x) : x \in (a_1, a_2)\}$.
Este procedimento es especialmente relevante para distribuciones triangulares y trapezoidales. Veamos el siguiente ejemplo para una distribución triangular.

Ejemplo 2.3.2. *Sea*

$$X = \begin{cases} x & ,0 \leq x \leq 1 \\ 2-x & ,1 \leq x \leq 2 \\ 0 & ,\textit{fuera de } [0,2] \end{cases}$$



El procedimiento es el siguiente:

1. *Generar* $u_1 \overset{d}{=} U(0,1)$ *y* $u_2 \overset{d}{=} U(0,1)$
   Calcular $x = 2u_1$ *e* $y = u_2$
2. *Aceptar* x *si* $y \leq f(x)$, *Rechazar si* $y > f(x)$ *y volver al paso 1*

Un algoritmo general para este método es el siguiente:

1. Generar $u_1, u_2 \in U(0,1)$
   Calcular $x = a_1 + u_1(a_2 - a_1)$
   $y = c \cdot u_2$
2. Calcular$f(x)$
   Si $y > f(x)$ ir a 1.
3. Salida $x, f(x)$

Observación:

$$\left.\begin{array}{l} X \overset{d}{=} U(a_1, a_2) \\ Y \overset{d}{=} U(0,c) \end{array}\right\} \Longrightarrow (X,Y) \overset{d}{=} U((a_1,a_2) \times (0,c)).$$

Veamos que efectivamente, la variable de salida se distribuye según la función de densidad de la que pretendemos simular valores.

Ejemplo 2.3.3. *Generar valores de una variable con distribución* $Beta(\alpha = 2, \beta = 2)$.

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} = 6x(1-x)$$

$$c = max\left\{f(x) : x \in (0,1)\right\}$$

Pares generados de una $U(0,1) \longrightarrow (0.2034, 0.9952); (0.8849, 0.7862); (0.8915, 0.0319)$
Si $x = 0.2034,\ y = 1.5 * 0.9952 = 1.4928$
$f(x) = 0.9721$*; Cómo* $f(x) < y$*, lo rechazamos.*
Así sucesivamente tomando en cuenta el criterio de decisión (Aceptación o Rechazo) seguimos con los demás puntos que hemos generado.

Este método tiene dos limitaciones principales:

1. El soporte tiene que ser un intervalo acotado

2. Se utiliza una envoltura rectangular, cuando claramente puede haberlas mejores.

Para solventar estas dificultades, se introduce el método generalizado.

Método generalizado de rechazo.

Sea $f(\cdot)$ una función de densidad con soporte no necesariamente finito de la que se desean obtener valores simulados. Sea $g(\cdot)$ una función de densidad elegida por nosotros tal que $\exists\ a > 1$ tal que $f(x) \leq a \cdot g(x) \forall x \in \mathbb{R}$ (en particular, esto implica que el soporte de $g(\cdot)$ ha de contener el soporte de $f(\cdot)$)[3].
El procedimiento es el siguiente:

1. Generar $x \overset{d}{=} g(\cdot)$
   Generar $y \overset{d}{=} U(0, a \cdot g(x))$
2. Calcular $f(x)$; Si $y > f(x)$ir a 1
3. Salida: X se distribuye según $f(x)$

[3] $g(\cdot)$ se toma para que se puedan simular valores sin dificultad y previamente al aplicar el algoritmo es necesario calcular el valor de la constante a que haga que se verifique la relación anterior.

Respecto a la elección de la envolvente $g(\cdot)$ no hay nada establecido de cuál es la mejor, pero se debe elegir lo más parecido posible a la función a generar pero que sea sencillo obtener valores para ella. En cuanto al valor de la constante a, hay que tener en cuenta que $P(\text{aceptar } x) = P(y \leq f(x)) = \dfrac{1}{a}$ y, por lo tanto, a debe ser lo menor posible pero manteniendo la hipótesis inicial, es decir, su valor óptimo sería $a = sup\left\{\dfrac{f(x)}{g(x)} : g(x) > 0\right\}$.
Obsérvese también que el método simple visto anteriormente es un caso particular del método generalizado (con una distribución uniforme como envolvente).

Ejemplo 2.3.4. *Supongamos que se desean generar valores de una distribución* $Normal(0,1)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

utilizando como envolvente la distribución logística cuya función de densidad es

$$g(x) = \frac{e^{-x}}{(1+e^{-x})^2},$$

y de aquí se deduce su función de distribución de forma inmediata

$$G(x) = \int_{-\infty}^{x} \frac{e^{-y}}{(1+e^{-y})^2} dy = \frac{1}{(1+e^{-x})}$$

El primer paso es determinar la constante a del algoritmo.

<u>*Paso 1.*</u> *Encontrar* $a > 1$ *tal que* $f(x) \leq a \cdot g(x)\ \forall x$

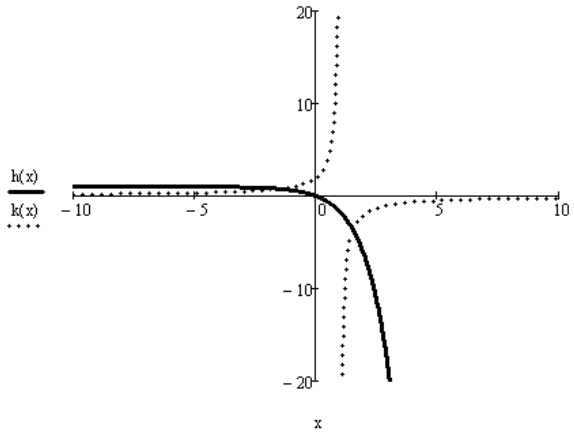
$$\begin{aligned} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} &\leq a \cdot \frac{e^{-x}}{(1+e^{-x})^2} \\ e^{-\frac{1}{2}x^2} \cdot \frac{(1+e^{-x})^2}{e^{-x}} &\leq a \cdot \sqrt{2\pi} \\ \underbrace{-\frac{1}{2}x^2 + 2 \cdot ln\left(1+e^{-x}\right) + x}_{t(x)} &\leq ln(a \cdot \sqrt{2\pi}) \end{aligned}$$

Donde se puede ver que

$$\begin{aligned} t`(x) &= -x + 1 - 2 \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= -x + 1 - \frac{2}{1+e^{x}} = 0 \\ \Longleftrightarrow\ & \frac{2}{1-x} = 1 + e^{x} \end{aligned}$$

Si observamos las gráficas (Ver figura 2.7) de las funciones que quedan a cada lado de la igualdad se tiene que la única solución está en $x = 0$,

Figura 2.7: $f(x) = \dfrac{2}{1-x}$ y $g(x) = 1 + e^x$



Ahora bien, tomando la segunda derivada vemos

$$t\text{“}(x) \;\; = \;\; -1 + \frac{2e^x}{(1+e^x)^2}$$

$t\text{“}(0) = 1/2$, *luego 0 es máximo y como*

$$\lim_{x \longrightarrow \infty} t(x) = -\infty \quad y \quad \lim_{x \longrightarrow -\infty} t(x) = -\infty$$

se trata de un máximo absoluto.

$$t(0) = 2ln(2) = ln(4) \Longrightarrow ln(4) = ln(a \cdot \sqrt{2\pi}) \Longrightarrow 4 = a\sqrt{2\pi} \Longrightarrow a = \frac{4}{\sqrt{2\pi}} \approx 1.595769$$

<u>Paso 2.</u> *Generar un valor de $X \stackrel{d}{=} g$. Lo haremos por el método de la transformada inversa.*

$$u_1 = G(X) \Longrightarrow u_1 = \frac{1}{1+e^{-x}} \Longleftrightarrow 1 + e^{-x} = \frac{1}{u_1} \Longleftrightarrow e^{-x} = \frac{1}{u_1} - 1 = \frac{1-u_1}{u_1}$$

$$x = -ln\left[\frac{1-u_1}{u_1}\right]$$

Generar un valor $U(0, a \cdot g(x))$.

$$y = a \cdot g(x) \cdot u_2 = a\frac{e^{-x}}{(1+e^{-x})^2}u_2 = a\frac{\frac{1-u_1}{u_1}}{(1+\frac{1-u_1}{u_1})^2}u_2 = \frac{a\frac{1-u_1}{u_1}}{\frac{1}{u_1^2}}u_2 = a \cdot u_1 \cdot (1-u_1) \cdot u_2$$

$$y = 1.5958 \cdot u_1 \cdot (1 - u_1) \cdot u_2.$$

Paso 3. *Si* $Y > f(X)$ *ir a paso 2.*
Si $u_1 = 0.2034 \qquad u_2 = 0.5952$
$x = 1.365 \qquad f(x) = 0.157$
$y = 1.5958 * 0.2034 * (1 - 0.2034) * 0.5952 = 0.1378$
Como $y < f(x)$*, aceptamos* $x = 1.365$ *con* $N(0, 1)$.

El principal inconveniente para utilizar este método es que no tiene una forma general ya que para cada distribución hay que elegir la envolvente más apropiada. Sin embargo, hay casos en que es muy sencilla su aplicación, las distribuciones truncadas. En este caso, es claro que la mejor envolvente es la propia distribución sin truncar y que la aplicación del algoritmo se reduce a rechazar los valores obtenidos en la región que al truncar ya no forma parte del soporte y aceptar los demás.

Generación de variables aleatorias mixtas

Hasta ahora hemos visto como generar variables aleatorias cuya distribución es discreta o absolutamete continua. En esta sección vamos a ver un método general, que en particular, permitirá generar valores de variables cuya distribución no sea ninguno de los casos anteriores.

Teorema 2.3.1. *Si* U *es una variable aleatoria* $U(0, 1)$ *y* $F(\cdot)$ *una función de distribución arbitraria, la variable aleatoria definida por*

$$Y = inf\,\{z : U \leq F(z)\}$$

tiene por función de distribución $F(\cdot)$.

Algunos ejemplos de cómo se aplica este resultado son los siguientes:

- Variable absolutamente continua: el método aplicado a este caso coincide con el método de la transformada inversa.
- Variable discreta: en este caso es el mismo procedimiento que el del método general o estándar que vimos para distribuciones discretas, en el cual se dividía el intervalo en subintervalos de longitud las probabilidades.

- Variable mixta: Sea la variable aleatoria mixta X cuya distribución viene definida por

$$F(x) = \begin{cases} 0 & , x < 3/2 \\ x - 1 & , 3/2 \leq x \leq 2 \\ 1 & , x \geq 2 \end{cases}$$

Entonces la variable aleatoria $Y = inf\{z : U \leq F(z)\}$ resulta ser

$$Y(u) = \begin{cases} 3/2 & , x < 3/2 \\ u + 1 & , u \geq 1/2 \end{cases}$$

que tendrá la misma distribución que X.

2.4. Experimentación y análisis de resultados.

Es uno de los últimos pasos dentro del proceso de simulación y que puede efectuarse antes o durante la implantación de las soluciones en el proceso real. Consiste en *jugar o experimentar* con el modelo ante situaciones nuevas o imprevistas, que tengan cierta probabilidad de ocurrencia, con el objeto de encontrar una solución óptima ante ese posible escenario. Otra de las situaciones donde se puede experimentar, es cuando queremos conseguir una configuración del sistema lo más óptima posible, que nos permita brinda el mejor servicio, con el menor costo de operación posible. La experimentación resulta muy útil, ya que los sistemas reales por lo general son dinámicos y de esta forma podemos hacerles frente con anticipación a los cambios que se produzcan. Por otra parte, el análisis de sensibilidad se enfoca principalmente en estudiar las variables no controlables dentro del proceso real cambiando progresivamente ciertos valores y determinar como esos cambios afectan las medidas de eficiencia del sistema.

2.4.1. Evaluación del número óptimo de simulaciones.

Debido a la naturaleza aleatoria del sistema en estudio, resulta impredecible crear un modelo cuyos resultados sean estadísticamente iguales a los sistemas reales. Uno de los factores que afecta de forma directa es el número de corridas del modelo simulado para encontrar resultados confiables.

El tamaño de una corrida de simulación depende principalmente del tipo de distribución que se

intenta simular y por decirlo de alguna forma; de la bondad del generador de números U(0,1) que se esta utilizando y de las condiciones iniciales con que comenzamos a simular del sistema. En forma general. Para calcular el número de simulaciones es la siguiente:

$$n = \frac{\sigma^2 (Z_{\alpha/2})^2}{K^2}$$

Donde:

Z : Estadístico normal estandar para cierta α.

$K = 0.16674\sigma$: Es la desviación absoluta máxima permitida sobre la media de la distribución a simular.

σ^2 : Varianza de la distribución a simular.

Pueden usarse esta fórmula siempre y cuando la información de donde se obtienen los estimadores sigan, estadísticamente, una distribución normal. En caso de que los datos analizados sigan otra distribución se debe hacer uso del teorema de Tchebycheff de tal suerte que el cálculo se ve reducido a:

$$n = \frac{m^2}{\alpha}$$

Donde:

α : Es la probabilidad de error permitida.

m^2 : Número de desviaciones estándar máximo permitido sobre la media de la distribución a simular.

El cálculo del número óptimo de corridas, para modelos de simulación en donde se tengan varias variables probabilísticas, se realiza ejecutando el cálculo de n para cada una de ellas y se selecciona la mayor de todas las n; éste será el número de simulaciones del modelo computacional. Ahora bien; para obtener resultados independientes hay que repetir r veces la simulación para cierto intervalo de tiempo, con diferentes semillas de números aleatorios.

Teniendo los resultados de cada una de las réplicas, es necesario tomar estos resultados para calcular los estimadores de media, varianza e intervalo de confianza de acuerdo con el siguiente procedimiento. Encuentre la media y varianza entre réplicas con las fórmulas siguientes:

$$\bar{x} = \frac{1}{r} \sum_{j=1}^{r} x_j$$

$$s^2 = \frac{1}{r-1} \sum_{i=1}^{n} (x_j - \bar{x})^2$$

Debido a la naturaleza probabilística de los resultados, es indispensable que para cada variable de respuesta se calcule el intervalo de confianza de acuerdo con:

$$Ic = \bar{x} \pm \frac{(s)^t}{x\sqrt{r_{r-1,\alpha/2}}}$$

2.4.2. Análisis del comportamiento de los datos.

En la simulación de sistemas se buscan respuestas a preguntas sobre el sistema en estudio a través de la información que proporcionan los experimentos con el modelo del sistema. A su vez los experimentos buscan, en general, respuestas a preguntas del tipo:*¿Que pasaría sí?* que se plantean en distintas fases del ciclo de vida: diseño, modificaciones de sistemas ya existentes. Las respuestas que buscamos mediante los experimentos servirán de soporte para tomar decisiones razonables sobre el sistema.

Comportamiento transitorio y estacionario de un proceso estocástico.

En los modelos de simulación dinámicos y aleatorios las variables respuesta varían con el tiempo, y conviene decir que durante este período de tiempo no se debe realizar trabajo de campo.

Así pues comenzaremos explicando brevemente los dos tipos de comportamiento que puede tener un proceso estocástico: *el transitorio y el estacionario o permanente*. Considérese el proceso estocástico respuesta $Y_1, Y_2, \ldots$ y sea $F_i(y/I) = P(Y_i \leq y/I)$, $y \in \mathbb{R}$ la distribución de la variable i-ésima del proceso dada ciertas condiciones iniciales I. $F_i(y/I)$ es denominada *distribución transitoria del proceso en el instante i para las condiciones iniciales I*.

Si cuando el tiempo se hace tender a ∞, la distribución ya no depende del tiempo y de las condiciones iniciales, es decir, $F_i(y/I) \underset{i\to\infty}{\to} F(y)$, $\forall y$, $\forall I$, entonces se dice que $F(y)$ es la *distribución estacionaria*.

La figura 2.8 muestra la evolución de un cierto proceso estocástico que alcanza un estado estacionario, muestra la variable respuesta y como va cambiando con el tiempo. A la hora de desarrollar simulaciones, cabe distingir varios tipos de simulación según sea el tipo de análisis que se pretenda hacer, según su horizonte temporal. Así los tiempos de simulación que se pueden plantear son:

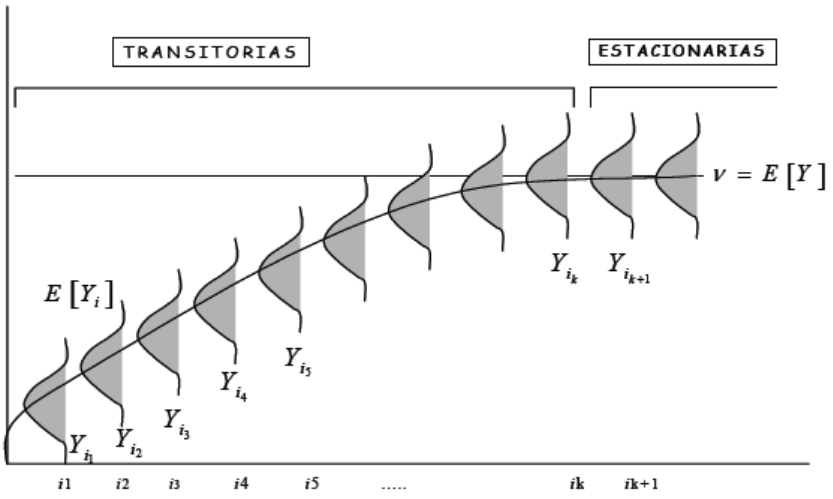


Figura 2.8: Evolución de la distribución de la variable respuesta.

- *Simulación con horizonte finito*: es la que se lleva a cabo cuando existe un evento natural E que específica la longitud de cada simulación o replicación. En ese evento el sistema se inicializa, obteniendo una muestra aleatoria simple de variables respuesta. Las condiciones iniciales generalmente afectan a las medidas de desarrollo, por lo que representan el sistema real, para lo cuál se puede plantear un período arranque (*warm up*) en el que se alcancen esas condiciones o se puede modelar las condiciones iniciales y aleatorizar en cada replica. En ocasiones, cuando el horizonte es muy lejano respecto al período de arranque, siendo el comportameinto estacionario el que rige la mayor parte del tiempo, la simulación con horizonte finito se puede asimilar a una simulación con horizonte infinito.

- *Simulación con horizonte infinito*: no existe tal evento que indique el final de la replicación, y nuestro interés se centra en el comportamiento a largo de replicación, donde pueden darse varias posibilidades:

  1. *Existe distribución estacionaria:* el objetivo es estimar los parámetros estacionarios del modelo.

  2. *No existe distribución estacionaria, pero sí por ciclos:* entonces hay que estimar los parámetros estacionarios de cada ciclo y la duración de éstos.
  3. *No existe distribución estacionaria, pues los datos de entrada varían en el tiempo:*

entonces hay que considerar que cada vez que cambian es un final de horizonte, y tratarlo así, como si fueran sucesiones de simulaciones con horizonte finito.

2.5. Técnicas de reducción de la varianza en simulación

Dado que la simulación de sistemas complejos suele consumir un elevado tiempo de procesado, es conveniente el empleo de métodos que permitan mejorar la eficiencia estadística del programa de simulación, es decir, conseguir que el intervalo de confianza de las estimaciones sea mejor para un mismo tamaño de muestra n, $\{X_i, i = 1, 2, \ldots, n\}$ o, de forma equivalente, que para alcanzar un intervalo de confianza especificado se reduce el tamaño muestral necesario n. En ambos casos habremos conseguido mejorar la eficiencia de la simulación, ya que habremos conseguido una mejoría en los resultados obtenidos en un tiempo de ejecución no superior al de partida. Para poder alcanzar cualquiera de los dos objetivos previos deberemos conseguir estimadores de menor varianza de los estadísticos de los procesos estocásticos bajo estudio.
Con este objetivo se han desarrollado diversas técnicas que se conocen bajo la denominación genérica de *técnicas de reducción de varianza*. Estas técnicas explotan las propiedades estadísticas de los métodos de simulación para intentar reducir la incertidumbre en los datos de salida y la aplicación de una u otra técnica dependerá fundamentalmente de las características del modelo bajo estudio. Es necesario notar, además, que usualmente es imposible conocer con antelación si el método tendrá éxito o no, es decir, si se va a producir un aumento o disminución de la varianza y, en este último caso, conocer la magnitud de la reducción.
En este apartado explicaremos tres de las técnicas más conocidas: el método de la variación antitética, el método de la variable de control y el método de los números aleatorios comunes, si bien un estudio más detallado sobre este tema puede desarrollar otros métodos existentes (los cuales implican técnicas de muestreo avanzadas y un conocimiento previo de diseños experimentales), nosotros solo trabajaremos los tres métodos mencionados anteriormente.

2.5.1. Método de la variación antitética

El método de la variación antitética (Hammersley y Morton, 1956) se basa en la ejecución de dos simulaciones en paralelo con secuencias de números aleatorios complementarias. De esta forma, se intenta que una observación pequeña en una de las simulaciones se corresponda con

una grande en la otra, es decir, que ambas simulaciones estén correladas negativamente. Como resultado se adoptará el promedio de las dos observaciones, ya que éste tenderá a estar más próximo de la esperanza μ que cada una de las observaciones por separado.
Supongamos que X_1 y X_2 es una muestra de tamaño 2 de la variable respuesta. El estimador de la media de la variable será la media muestral, $\overline{X} = \frac{X_1+X_2}{2}$ y su varianza será,

$$V(\overline{X}) = \frac{1}{4}\left(V(X_1) + V(X_2) + 2 \cdot Cov(X_1, X_2)\right)$$

Obsérvese que si las variables son independientes, la covarianza es cero, pero el valor de la varianza mejoraría si la covarianza fuera negativa. por lo tanto, el objetivo de esta técnica es intentar lograr una covarianza negativa entre dos replicaciones del mismo modelo. Para ello, el método propone que en una replicación se utilicen unos valores de la uniforme y en la otra los valores complementarios, es decir, 1 menos los anteriores.
Así si $U_1, \ldots, U_n$ son variables aleatorias con distribución uniforme en $(0, 1)$, las variables $1-U_1, \ldots, 1-U_n$ tambien son uniformes en $(0, 1)$. Y es más, si $X_1, \ldots, X_n$ son variables obtenidas mediante la transformada inversa a partir de $U_1, \ldots, U_n$, $X_1', \ldots, X_n'$ son obtenidas a partir de $1 - U_1, \ldots, 1 - U_n$, entonces las variables X_j y X_j' están correladas negativamente.
De aquí se deduce que si en una simulación se utilizan los valores de una uniforme y en la siguiente sus complementarios las variables estarán correladas negativamente, mejorando el valor de la varianza del estimador.
Sin embargo, no siempre es cierto ya que aunque las variables de entrada estén correladas negativamente, si el sistema es complejo las de salida no pueden estarlo. Además, el resultado está demostrado para variables de entrada generadas por el método de la transformada inversa, pero no siempre éste es el método que se utiliza para generar una variable aleatoria.

2.5.2. Método de la variable de control

Este método también intenta aprovechar la correlación entre varias variables aleatorias para obtener una reducción de la varianza del estimador de la media en una de ellas.
Interesa estimar la media μ de una variable aleatoria, X, de salida de la simulación; conocemos otra variable aleatoria, Y, que aparece también en la simulación, pero cuya esperanza $\nu = E[Y]$ es conocida; y, por último, sabemos que X e Y están correlacionadas, bien positiva o bien negativamente. Suponiendo que la correlación es positiva, entonces valores de Y mayores que su

media ($Y > \nu$) tenderán a ir acompañados de valores de X mayores que su media ($X > \mu$), y viceversa. En este caso la aplicación del método consiste básicamente en lo siguiente: cuando en la simulación se observa que $Y > \nu$, entonces podemos sospechar que $X > \mu$ y con esta información es posible corregir el valor de X, disminuyéndolo en cierta cantidad; cuando se produzca el efecto contrario, se aumentará el valor de X. De esta forma se utiliza el conocimiento que se tiene sobre la desviación de Y respecto a su media para acercar X hacia la suya μ, reduciendo así su variabilidad. Se dice, entonces, que Y es la *variable de control* de X.
Es necesario entonces determinar cuál es la cantidad que se necesita para ajustar el valor de X, así se define el estimador controlado X_C como

$$X_C = X - a * (Y - \nu),$$

donde a es, en principio, un valor constante que tiene el mismo signo que la correlación entre X e Y.
Las principales dificultades de este método se centran en dos problemas: (1) ¿Cuál es el valor óptimo para a? y (2) ¿Cuál es la variable que debe utilizarse como variable de control?.
La respuesta a la primer interrogante parece inmediata, el mejor valor de a será aquel que minimice la varianza del estimador X_C; Loh(1997) repasa los resultados al respecto, y establece las condiciones para conjugar esta técnica con el método de bloques.
En el caso de la segunda pregunta, una buena variable de control debe estar fuertemente correlacionada con X, para que ofrezca la mayor información posible sobre las variaciones de X y , además, debería de presentar una varianza reducida. Para hallar la variable de control más adecuada en cada simulación se deberá analizar la estructura del sistema o bien reducir la experimentación sobre el modelo.

2.5.3. Método de los números aleatorios comunes

Este método es diferente a los otros en la medida en que se aplica cuando estamos comparando dos o más configuraciones alternativas para un mismo sistema en lugar de analizar el comportamiento de una única configuración. La idea básica es la comparación de ambas alternativas bajo condiciones experimentales similares. En un modelo de simulación, estas condiciones experimentales son generadas por las variables aleatorias que modelan las circunstancias ambientales bajo las que se encuentra el sistema.

Visto de un modo general, podemos considerar que nos encontramos ante dos alternativas diferentes, donde $X_i^{(1)}$ y $X_i^{(2)}$ son las observaciones i-ésimas de la primera y segunda configuración respectivamente. Si lo que deseamos estimar es la esperanza matemática $\xi = \mu_1 - \mu_2 = E[X_i^{(1)}] - E[X_i^{(2)}]$ y tenemos dos muestras, una por configuración de tamaño n, entonces podemos obtener la muestra $\{Z_i; i = 1, 2, \ldots, n\}$ donde $Z_i = X_i^{(1)} - X_i^{(2)}$. Satisfaciéndose que $\xi = E[Z_i]$. El estimador insesgado de ξ será entonces

$$\overline{Z}(n) = \frac{\sum_{i=1}^{n} Z_i}{n},$$

y como las muestras Z_i son variables aleatorias independientes e idénticamente distribuidas, entonces

$$Var[\overline{Z}(n)] = \frac{Var[Z_i]}{n} = \frac{Var[X_i^{(1)}] + Var[X_i^{(2)}] - 2Cov[X_i^{(1)}, X_i^{(2)}]}{n}$$

En caso de que las simulaciones de las dos configuraciones alternativas sean independientes, es decir, con diferentes secuencias de números aleatorios, $X_i^{(1)}$ y $X_i^{(2)}$ serán independientes, por lo que $Cov[X_i^{(1)} X_i^{(2)}] = 0$. En otro caso, si se realizan ambas simulaciones introduciendo una correlación positiva entre $X_i^{(1)}$ y $X_i^{(2)}$, entonces se conseguirá que $Cov[X_i^{(1)}, X_i^{(2)}] > 0$, con lo que la varianza del estimador $Var[\overline{Z}(n)]$ se reducirá.

Con está técnica se pretende, entonces, introducir correlación positiva utilizando la misma secuencia de números aleatorios en ambas simulaciones. Es decir, si se usa una muestra de la variable aleatoria $U \sim U(0, 1)$ para un propósito en una de las simulaciones, deberá utilizarse la misma muestra para el mismo propósito en la otra. Como en el caso del método de la variación antitética, es vital mantener la sincronización para que esta técnica realmente funcione y se logre la reducción de varianza deseada.

2.6. Test de hipótesis

Es test de hipótesis permite, a partir de la información contenida en un conjunto de muestras de una variable aleatoria, aceptar o rechazar una determinada hipótesis sobre dicha variable. La hipótesis puede referirse a:

- Uno o más parámetros de la variable aleatoria.
- La función de distribución de probabilidad de la variable.

2.6.1. Procedimiento del test de hipótesis referido a un parámetro de la variable aleatoria

- Determinar la hipótesis nula H_o y la hipótesis alternativa H_1.
- Seleccionar el estadístico del test calculable a partir de las muestras; por ejemplo $\overline{X}(n)$ o $S^2(n)$.
- Elegir la *región de rechazo* (también llamada región crítica), es decir, el conjunto de valores tal que si H_0 es cierta, la probabilidad de que el estadístico del test pertenezca a dicho conjunto sea α, siendo α un valor preseleccionado llamado *nivel de significación del test* (normalmente del 5 % ó 1 %).
- Calcular el *estadístico del test* sobre las muestras de la variable aleatoria.
  - Si el valor calculado está en la región de rechazo, entonces procede rechazar H_0 y aceptar H_1.
  - En otro caso no se consigue rechazar H_0.

  Normalmente esta última opción no significará que la hipótesis nula sea cierta.

Errores en la decisión final del test de hipótesis

- **Error tipo I:** rechazar la hipótesis H_0 cuando es cierta. Este error se produce con probabilidad α.
- **Error tipo II:** aceptar la hipótesis H_0 cuando es falsa. Este error se produce con probabilidad β.

La potencia de un determinado test de hipótesis viene dada por $1 - \beta$, valor que indica la capacidad del test para rechazar H_0 cuando es falsa. Es más, si el test no consigue rechazar H_0 no pondrá en evidencia la verosimilitud de dicha hipótesis, sino la incapacidad del test para rechazarla: pudiera ser que un test más potente si fuese capaz de conseguirlo.

2.6.2. Test de bondad de ajuste

Son contrastes de hipótesis donde el planteamiento es:
Sean $X_1, \ldots, X_n$ una muestra aleatoria simple (v.a.i.i.d.) con distribución F desconocida.
Sea F_0 una distribución particular.
El contraste que se plantea es:

$$\left.\begin{array}{ll} H_0 & : F = F_0 \\ H_1 & : F \neq F_0 \end{array}\right\}$$

Observaciones:

1. Por la naturaleza de los contrastes en general, si el resultado es rechazar la hipótesis nula se hará con bastante seguridad; si el resultado es aceptarla se debe decir "no rechazar H_0 ya que no hay evidencia estadística para ello". En general, los test tienden a aceptar que los datos pueden provenir de la distribución propuesta.

2. Para una cantidad grande de datos es habitual que el resultado sea rechazar H_0, ya que algunos contrastes son muy sensibles a pequeñas variaciones.

3. Se debe utilizar el p-valor como medida del ajuste: El p-valor es el nivel de significación para el que se rechazaría H_0 con los datos utilizados para realizar el contraste, es una cota de la probabilidad de cometer el error de tipo I (rechazar H_0 siendo cierta). Cuanto mayor sea el p-valor mejor es el ajuste.
   Sin embargo, no se ha de ser ciego a la hora de seleccionar una distribución, ya que puede ser preferible una distribución más sencilla aunque tenga un p-valor peor (siempre que sea razonable) que otra que tenga un mejor valor pero sea muy compleja.

2.6.3. Test de la χ^2 (Pearson, 1900)

La idea básica de este test es agrupar los datos en intervalos y comparar las frecuencias de estos intervalos con las probabilidades que la distribución teórica les asigna, midiendo la distancia entre ambas.
Este test es aplicable tanto a distribuciones discretas como a continuas, aunque requiere tener una muestra suficientemente grande ya que se basa en un resultado asintótico.
El procedimiento sería el siguiente:

a) Dividir el rango de la distribución ajustada en k intervalos adyacentes:

$$[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$$

b) Sea N_j = número de X_i en intervalo j-ésimo = frecuencia absoluta observada de intervalo $[a_{j-1}, a_j)$.

c) Calcular de la distribución teórica la probabilidad de cada intervalo:

$$p_j = P([a_{j-1}, a_j)),$$

y a partir de ella la frecuencia esperada como $n \cdot p_j$.

d) Calcular:

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n \cdot p_j)^2}{n \cdot p_j}$$

Si H_0 es cierta se verifica que

$$X^2 \xrightarrow[n \longrightarrow \infty]{D} \chi^2_{k-m-1},$$

donde m es el número de parámetros estimados.

e) Rechazar H_0 si $X^2 \geq \chi^2_{k-m-1,\alpha}$

Una consideración a hacer es que el valor del estadístico X^2 varía según los intervalos elegidos. No hay normas claras al respecto, excepto que no puede haber intervalos de frecuencia nula, pero, una idea es que sean de igual amplitud, que haya al menos tres intervalos y que la frecuencia esperada sea al menos 5, es decir, $n \cdot p_j \geq 5$.

2.6.4. Test de Kolmogorov-Smirnov

Este test compara la función de distribución empírica de los datos con la función de distribución teórica.
Su aplicación se presenta sólo para distribuciones continuas (existe una versión para distribuciones discretas), y es válido para cualquier tamaño de muestra, incluso aunque se disponga de pocos datos.
El procedimiento para aplicarlo es el siguiente:

a) Calcular la función de distribución empírica de los datos sin interpolación, es decir:

$$F_n(x) = \frac{\text{número de } X_i'^s \leq x}{n}$$

b) Obtener la máxima diferencia entre la función de distribución empírica y la teórica: esta diferencia se alcanzará en los puntos observados, ya que son los puntos de salto de la función empírica, pudiendo ser en el propio punto o justo antes. Por ello, se calcula

$$D_n^+ = \max_{1\leq i\leq n} \left| \frac{i}{n} - F(X_{(i)}) \right|$$

que es la máxima diferencia en los puntos observados con el valor de la distribución empírica justo en el punto, y se calcula

$$D_n^- = \max_{1\leq i\leq n} \left| \frac{i-1}{n} - F(X_{(i)}) \right|$$

que es la máxima diferencia tomando el valor de la distribución empírica justo antes (por la izquierda) del punto. Así la máxima diferencia entre la función de distribución empírica y la teórica será el máximo de ambos valores, es decir, $D_n = \max\{D_n^+, D_n^-\}$.

c) Rechazar H_0 si $D_n > d_{n,1-\alpha}$, donde $d_{n,1-\alpha}$ es el valor que se recoge para la simplificación y tamaño de muestra en las tablas de *Kolmogorov-Smirnov*.

Capítulo 3

Aplicación de teoría de colas al sistema bancario.

En el presente capítulo se muestra como llevar a cabo un estudio de teoría de colas en una agencia bancaria real, haciendo uso de todas las herramientas que intervenen desde el punto de vista teórico como práctico, entre ellas se tienen: modelos probabilísticos de teoría de colas, análisis de sistemas, trabajo de campo, manejo de base de datos y manipulación de sistemas informáticos específicos para estudios de colas.

Al poner en práctica todas estas técnicas se llega a determinar el funcionamiento de un sistema en el cual intervienen colas, para nuestro caso se hace un análisis dentro de una sucursal bancaria y se siguen los siguientes pasos: en primer lugar se trabajó con base a lo descrito en el capítulo I, conociendo de primera mano las componentes y funcionamientos básicos del sistema, luego se hizo el levantamiento y organización de los datos extraidos de campo, los cuales una vez recolectados se procede a la descripción y análisis exploratorio de los mismos; inmediatamente después se busca la distribución de probabilidad para cada una de las variables consideradas en este estudio (tiempos entre llegadas y tiempos de servicio) y el tipo de modelo de colas apropiado a las distribuciones que presenten los datos.

Si las distribuciones de probabilidad para las variables consideradas en el estudio son de tipo general, es decir, sus medidas de eficiencia son difíciles de obtener de forma analítica debido a que los datos recolectados no se ajustaron a los modelos desarrollados en el capítulo I, se hará entonces una simulación del comportamiento de las variables, para ello se aplicará la teoría de simulación de eventos discretos orientada al funcionamiento del sistema bancario, la cual se

encuentra expuesta en el capítulo II.

Para obtener las medidas de eficiencia para el sistema en estudio y hacer la simulación del funcionamiento de dicho sistema bancario en estudio, se utilizará el software *winQSB*. Este lenguaje es muy completo ya que calcula la mayor cantidad de medidas de eficiencia necesarias para describir cuantitativamente el funcionamiento del sistema; se trata entonces de un lenguaje eminentemente para experimentación enfocado a: análisis de teoría de colas, programación lineal, control estadístico de la calidad entre otros.

Uniendo todos los pasos en un estudio de colas, logramos hacer una interpretación de los resultados obtenidos del modelo encontrado, así como también, darnos otros modelos que ayudarán a la realización de comparaciones para poder emitir conclusiones adecuadas en torno al funcionamiento óptimo del sistema en estudio.

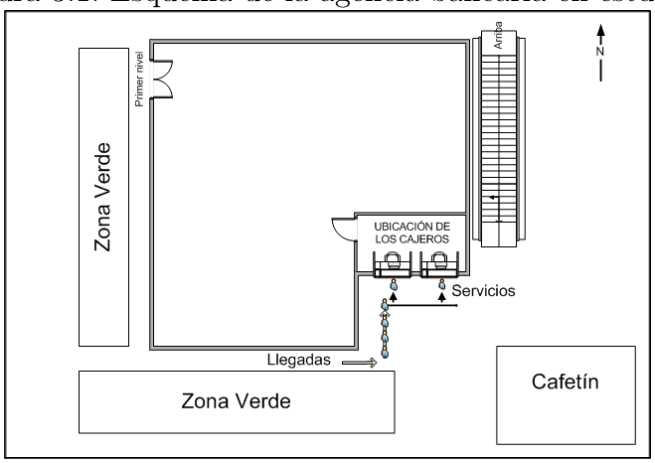
Es de mencionar que de aqui en adelante nos referiremos al lugar donde se realizó el estudio como *agencia bancaria*, debido al secreto estaístico no revelaremos el nombre específico y dirección de dicha agencia bancaria.

3.1. Metodología de aplicación de la Teoría de Colas.

3.1.1. Análisis del sistema de colas.

Uno de los primeros pasos en la elaboración de un análisis de colas es la descripción general de cada uno de los elementos que intervienen en él; uno de ellos es la infraestructura donde se llevan a cabo los procesos de servicio, la cual se describe en la siguiente figura.

Figura 3.1: Esquema de la agencia bancaria en estudio.



Apreciando el esquema de las instalaciones podemos describir de forma preliminar que es una agencia con dos ventanillas externas, la cual cuenta con dos cajeros a disposición de los clientes en los horarios mostrados en el cuadro 3.1.

Cuadro 3.1: Horarios de servicio al cliente.

Día	**Horario de atención**
De lunes a Viernes	De 9:00 a.m - 12:00 m.d De 1:00 p.m - 4:00 p.m
Sábado	De 9:00 a.m - 12:00 m.d

Otro de los aspectos a considerar en el anásis de colas son los componentes; éstos son los elementos ya sea tangibles o intangibles que aportan al sistema para evacuar una tarea o servicio, y es básicamente lo que hemos tratato en el capítulo I, ahora bien, para el caso actual de la agencia bancaria obtenemos los siguientes componentes:

- **Tipo de clientes:** éstos provienen de una población infinita, debido a la ubicación geográfica de la *agencia bancaria* ya que todos los clientes del banco y usuarios en general

tienen acceso a ella.

- **Fuente de entrada o población potencial:** es infinita; cualquier persona natural o jurídica puede ingresar al sistema para que se les de servicio.
- **La cola o línea de espera:** se cuenta con una línea de espera para la atención de los clientes; es decir, los clientes que llegan al área de espera deberán hacer una única cola para poder ser servidos por alguno de los cajeros que se encuentre disponible.
- **La capacidad de la cola:** se considera infinita, ya que el sistema no está restringido a un número específico de clientes.
- **El mecanismo de servicio:** básicamente se cuenta con dos cajeros a disposición de los clientes para atender la cola que se va generando, considerándose también un cierto comportamiento homogéneo entre los tiempos de servicio empleados por los dos cajeros.
- **El sistema de la cola:** es de tipo FIFO, es decir el primero en llegar es el primero en ser atendido.

El funcionamiento de servicio al cliente en la *agencia bancaria* es de ventanilla externa, en las cuales se admiten todo tipo de transacciones, excepto apertura de cuentas y servicio personalizado al cliente o a grandes clientes; además para los días de pago de salarios y de pago de impuestos (recibos de agua, luz, teléfono, entre otros), las dos cajas restringen todo tipo de transacción y sólo se dedican a cobros de éstos, también durante el período de tiempo del día 29 de un mes determinado al día 5 del siguientes mes se encuentra un poco recortado el servicio.

3.1.2. Formulación del problema

Una vez determinados los elementos que intervienen en el funcionamiento del sistema nos interesa conocer si se encuentra trabajando adecuadamente o no, es decir, si está apunto de desbordarse o se está haciendo un mal uso de alguno de los recursos que intervienen en él.
Para derterminar un funcionamiento óptimo es necesario analizar las componentes del sistema en conjunto y por separado, el objetivo es determinar si alguno de los pasos que se siguen se encuentra mal definido, se tarda mucho tiempo en completarse o simplemente es innecesario.
Subsanada esta depuración preliminar, pasamos a la determinación de las variables utilizadas

en el estudio de campo, lo cual sirve para determinar la forma en que éstos datos se distribuyen y así poder determinar el modelo de colas adecuado; las variables son las siguientes:

- Tiempos entre dos llegadas consecutivas al sistema (en minutos).
- Tiempos de servicio por parte de los cajeros (en minutos).

Luego de haber definido las variables que intervienen en el estudio de colas, se procede a realizar el análisis estadístico para determinar las distribuciones de probabilidad de los datos, entre éstos análsis podemos mencionar los siguientes:

- Análisis de las medidas descriptivas básicas.
- Histograma, gráfico Probability-Plot y gráfico de contraste.
- Contrastes de hipótesis (Test Chi-Cuadrado y Kolmogorov-Smirnov).

Determinadas las distribuciones de probabilidad de las variables en estudio, se procede a la definición del modelo de colas y obtención de las medidas de eficiencia, entre las que mencionamos,

- La cantidad de usuarios por unidad de tiempo que llegan al sistema.
- El tiempo promedio entre dos llegadas consecutivas.
- La cantidad de usuarios atendidos por unidad de tiempo.
- El tiempo promedio de atención para los usuarios.
- El tiempo en cola del usuario.
- El tiempo en el sistema del usuario.
- La cantidad promedio de usuarios en cola.
- La cantidad promedio de usuarios en el sistema.
- El factor de utilización del sistema.

3.1.3. Descripción del trabajo de campo.

Una vez elegida la sucursal bancaria y analizando su funcionamiento, lo primordial ahora es tener información que nos permita definir el modelo de colas adecuado; para ello es necesario recoger la mayor cantidad de datos de las variables principales en el estudio, las cuales son: los tiempos entre dos llegadas consecutivas y los tiempos entre servicios.
Para desarrollar esta investigación se inicio haciendo gestiones con el personal ejecutivo del Scotiabank, a fin de obtener la autorización para la implementación de un análisis de colas en una agencia bancaria relativamente grande y con un número considerable de servidores, la autorización solicitada fue concedida verbalmente y se nos asignó la agencia bancaria ubicada en el paseo general escalón, nos presentamos a dicha agencia y no se nos permitió recoger la información de las variables en estudio. en este sentido, volvimos a hablar con los ejecutivos del Banco y se nos denegó rotundamente el levantamiento de la información, por lo tanto se tomó la decisión de escoger libremente un lugar que nos permitiera sin mayor complicación la recolección de la información necesaria para el estudio de colas, es así como se determinó escoger la *agencia bancaria* descrita en el apartado 3.1.1 y realizar el trabajo de campo pertinente.
Dicho trabajo de campo consta de las siguientes etapas:

1. Determinación de las variables (Tiempo entre llegadas y tiempos de servicio).
2. Determinación de las unidades de medida (en minutos).
3. Preparación de dos personas para la toma de los datos (uno para los tiempos de llegada y otro para los tiempos de servicio).
4. Determinación de los horarios para el trabajo de campo (Ver cuadro 3.1).
5. Determinación de los días de trabajo de campo (1 mes).
6. Llenado de formulario en la recolección de los datos.
7. Revisión de la información (Control de calidad en los datos).

Cabe mencionar que la toma de los tiempos se realizó en los horarios ya mencionados y con la ayuda de un cronómetro para reducir los errores de captura de información, con lo cual se obtuvo la siguiente tabla de datos (Ver cuadro 3.2):

Cuadro 3.2: Muestra de los tiempos recolectados en minutos.

Nombre de la sucursal: *Agencia Bancaria.*							
Fecha: Marzo de 2009				**Hora:** 9:30 a.m - 11:30 a.m			
				1:30 p.m - 03:30 p.m			
Tipo de conteo: Llegadas				**Tipo de conteo:** Servicios			
Min.	**Min.**	**Min.**	**Min.**	**Min.**	**Min.**	**Min.**	**Min.**
1.9	3.8	2.6	2.0	3.0	3.1	5.2	4.1
4.1	2.1	2.9	1.9	1.6	5.3	4.0	3.3
4.6	3.1	3.2	3.1	3.3	4.2	3.0	2.0
5.1	4.3	5.7	4.0	5.1	3.0	1.6	3.2
3.9	3.2	4.0	3.2	4.0	2.6	2.1	4.0
3.8	5.3	3.1	5.2	3.8	2.2	3.3	1.6
2.1	3.8	4.3	4.3	3.6	1.8	4.1	2.1
3.1	3.1	5.1	3.0	2.0	2.9	4.2	3.3
5.1	2.0	4.7	1.8	3.3	3.1	3.0	4.0
4.1	1.8	4.3	2.0	2.1	4.0	2.6	3.0
5.1	4.3	5.7	4.0	5.1	3.0	1.6	3.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.9	3.2	4.0	3.2	4.0	2.6	2.1	4.0
3.1	5.4	3.1	5.2	3.5	3.2	3.3	1.6

3.2. Análisis del conjunto de datos.

En el siguiente apartado trabajaremos analizando el comportamiento de los datos recolectados en campo, primeramente se analizan los tiempos entre dos llegadas consecutivas al sistema y luego los tiempos de servicio de los cajeros.

Tiempos entre llegadas (TELLmin).

Los estadísticos descriptivos obtenidos para los tiempos entre llegadas de clientes al banco se muestran en el cuadro 3.3; tenemos que la media de los tiempos entre dos llegadas consecutivas al sistema (la cual llamaremos TELLmin) es de 3.7 minutos, la asimetría es cercana a cero, el coeficiente de curtosis es aproximadamente -1, la media es muy similar a la mediana, asemejándose entonces a las distribuciones insesgadas como: Normal ó T-Student.

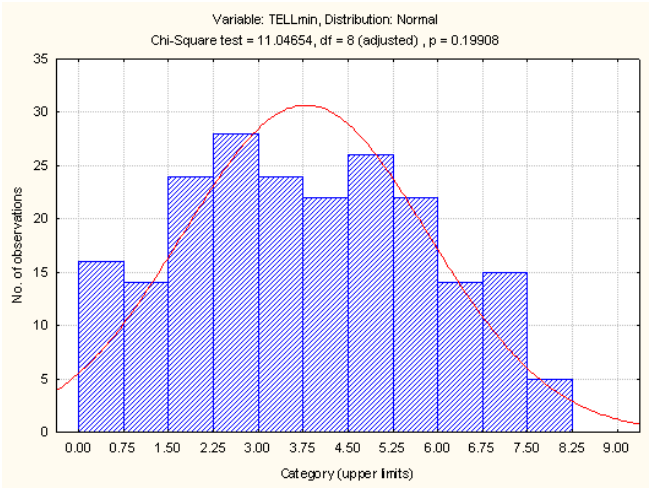
Cuadro 3.3: Estadísticos descriptivos de la variable tiempo entre llegadas al sistema.

	N	Media	Mediana	Mín	Máx	Desv. Tip.	Asimetría	Curtosis
TELLmin	210	3.7	3.7	0.02	8.1	2.0	0.1	-0.9

Análisis gráfico de los tiempos entre llegadas (TELLmin).

Analizando la figura correspondiente se puede apreciar de manera preliminar que la distribución de los datos se comporta como una Normal (Ver figura 3.2).

Figura 3.2: Gráfico de contraste de los tiempos entre llegadas (en minutos).

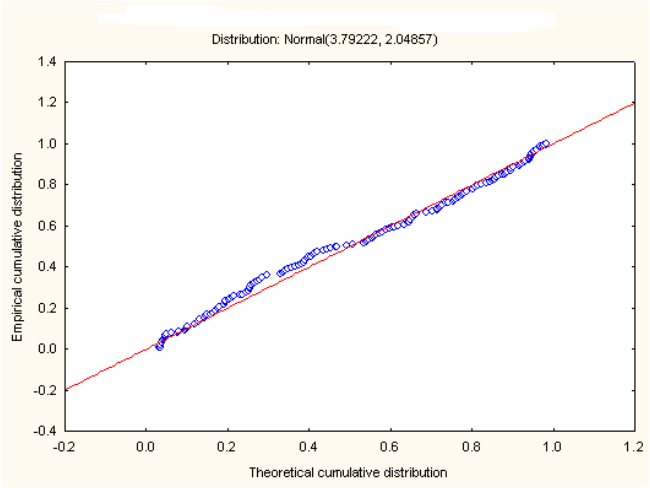


Ahora bien, para justificar mejor lo anterior veamos el contraste que está representado en la curva que se sobrepone al histograma; y en la que se observa un p-valor de la prueba de bondad de ajuste de la Chi-Cuadrado, el cual realiza la siguiente prueba de hipótesis:

- H_0: Los datos de la variable que mide el tiempo entre llegadas se distribuye como una Normal.
- H_1: Los datos de la variable que mide el tiempo entre llegadas no se distribuye como una Normal.

El p-valor obtenido en esta prueba es de 0.2, es decir; aceptamos la hipótesis de que los datos se distribuyen como una normal; si vemos el gráfico P-P Normal (Ver figura 3.3) tambien se aprecia que los datos se ajustan a dicha distribución.

Figura 3.3: Gráfico Probability Plot de los tiempos entre llegadas (en minutos).



Realizando la prueba de Kolmogorov-Smirnov obtuvimos un p-valor de 0.309, con lo que podemos concluir que la distribución de probabilidad de los tiempos entre llegadas (TELLmin) es una Normal de media 3.7 y desviación típica de 2.0.

Tiempos de servicio (TSmin).

En el caso de los tiempos de servicio (TSmin) vemos el cuadro 3.4, dicho cuadro muestra que el promedio de tiempo que se tardará el cliente en ser servido es aproximadamente 5 minutos, además, se observa una Asimetría de 0.3 y la Curtosis cercana a -1, en este caso la media es mayor que la mediana, por tanto, se trata entonces de una distribución ligeramente sesgada a la izquierda, con un pico relativamente bajo asemejándose a las distribuciones: Gamma, Chi-Cuadrado o F.
Para verificar estos supuestos, es necesario realizar el análisis gráfico y de bondad de ajuste; al igual que en el caso de los tiempos de llegada y ver cuál es la distribución de probabilidad a la que se ajustan los datos.

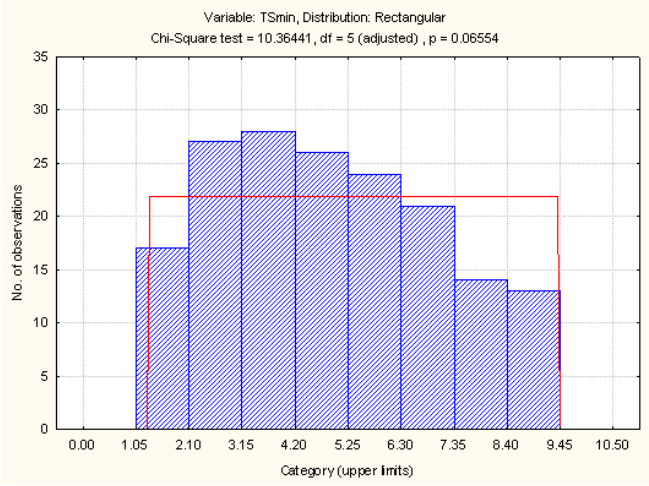
Cuadro 3.4: Estadísticos descriptivos de los tiempos de servicio.

	N	Media	Mediana	Mín	Máx	Desv. Tip.	Asimetría	Curtosis
TSmin	170	4.9	4.6	1.3	9.5	2.2	0.3	-0.9

Análisis gráfico de los tiempos de servicio (TSmin).

Al revisar el histograma de los tiempos de servicio se puede notar cierta uniformidad en las frecuencias de cada uno de los intervalos (Ver figura 3.4); tambien se aprecia que las frecuencias disminuyen muy despacio en cada una de las categorias.
Al revisar el contraste realizado se puede apreciar de manera preliminar que la distribución de los datos puede ser una uniforme.
Las hipótesis que se deben probar en relación a la distribución de los tiempos de servicio son:

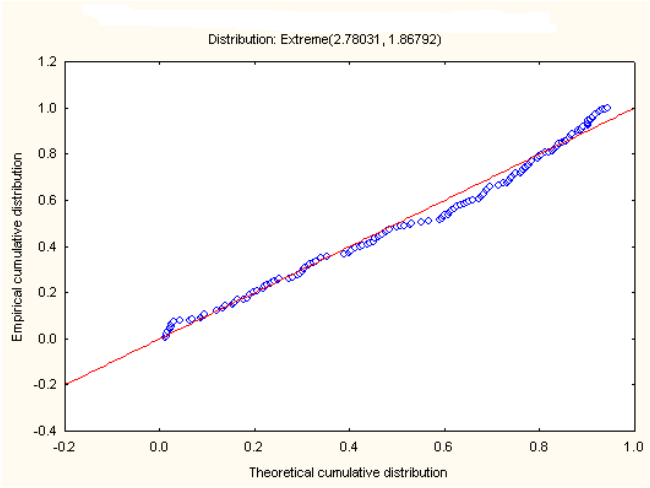
Figura 3.4: Histograma y gráfico de contraste para los tiempos de servicio (TSmin).



- H_0: Los datos de la variable que mide el tiempo de servicio se distribuye como una Uniforme.
- H_1: Los datos de la variable que mide el tiempo de servicio no se distribuye como una Uniforme.

En el gráfico de contraste se obtiene el p-valor de la bondad de ajuste de los datos; el cual nos muestra un valor de 0.065, por tanto, podemos aceptar la hipótesis nula relacionada a que la distribución de los tiempos de servicio es una uniforme.
Podemos visualizar en el gráfico Probability Plot de los tiempos de servicio (Ver figura 3.5), que la mayor parte de los datos se encuentran entorno a la recta de ajuste, por lo que se concluye que los datos provienen de una distribución uniforme; además realizando el contraste de Kolmogorov-Smirnov nos da un p-valor de 0.05 que nos afirma nuevamente el hecho de

Figura 3.5: Gráfico Probability Plot para tiempos de servicio (TSmin).



aceptar la hipótesis nula de que los datos se distribuyen uniformemente entre 1.05 y 9.45.
Ahora que sabemos la distribución de probabilidad asociada a las variables TELLmin y TSmin, se obtiene el cuadro 3.5 que nos muestra el resumen del análisis anterior:

Cuadro 3.5: Distribuciones de probabilidad de las variables en estudio.

Variable	**Distribución**
TELLmin	Normal(3.7,2.0)
TSmin	Uniforme(1.05,9.45)

3.2.1. Aplicación de la teoría de colas mediante simulación.

En esta parte del estudio se dejan de utilizar los datos extraidos de campo y se utiliza solamente el resultado mostrado en el cuadro 3.5, se establece el tipo de modelo de colas adecuado para comenzar a realizar el trabajo de simulación.
Podemos decir que estamos en presencia de un modelo de colas con dos servidores, tiempos de llegada distribuidos como una Normal y tiempos de servicio distribuidos uniformemente; su representación sería:

$$G/G/2.$$

En el desarrollo de la aplicación de teoría de colas a una agencia bancaria, optamos por utilizar la simulación de dicho sistema mediante el uso del software winQSB (para el cálculo de las medidas

de eficiencia, cálculo del número de réplicas, experimentación y análisis de sensibilidad).
Naturalmente el desarrollo de simulaciones implica considerar una serie de pasos de forma secuencial, aunque en realidad es un proceso en el cual se puede regresar pasos atrás en el desarrollo de la simulación.

- **Inicio:** Para generar simulaciones de modelos, se debe identificar primero, el tipo de distribución de probabilidad que mejor se ajusta a los tiempos entre llegadas y a los tiempos de servicios (donde ambos tiempos fueron recolectados en minutos). Las componentes del sistema especificadas anteriormente para un modelo de colas son: la capacidad de la cola (infinita), el número de servidores en el sistema (dos cajeros), identificar el tamaño de población fuente (infinita), la disciplina de la cola (FIFO), en términos generales estamos en presencia de un modelo G/G/2, utilizando este modelo se pretende calcular las medidas de eficiencia que presenta la sucursal bancaria que se ha considerado en el estudio. Si las simulaciones del modelo determinan que la configuración del sistema presenta ciertas irregularidades, debemos experimentar con otra configuración que nos permita optimizar el funcionamiento del sistema.

- **Identificamos los eventos:** son tres los principales eventos o sucesos que se desarrollan en este sistema: las llegadas de los clientes, la espera en cola de los clientes y el servicio que se le presta al cliente en la sucursal bancaria.

- **Definición del mecanismo del control de tiempo:** este mecanismo se llevará a través del tiempo simulado y avanzará sólo cuando ocurra un evento de llegada o servicio.

- **Identificación de la estructura de los datos:** la cual contiene las variables de estado de los tiempos entre llegadas y de servicio, además, una tabla de sucesos en la cual se describen los estadísticos relacionados al tiempo de conteo del reloj de simulación.

- **Definición del flujo de control de datos del programa simulador:** es la forma en la que se van comportando los diferentes eventos en base a como se mueve el reloj de simulación.

- **Definir la estructura del programa que realiza la simulación:** se encuentra formada por todas las partes del programa de simulación el cual sirve para la extracción de las

medidas de eficiencia utilizadas en los anásis posteriores.

- **Selección del lenguaje de simulación:** se utilizará el software winQSB por su sencillez, fácil manejo y porque proporciona resultados confiables.

- **Programación:** se analiza el diagrama de flujo de la información y este mecanismo se introduce al software WinQSB.

- **Verificador del simulador:** introduciendo los paramétros del modelo al simulador y utilizando una semilla de aleatorizacion de 39541 de manera que las muestras generadas por dicho lenguaje arrojen parámetros en promedio similares a los datos obtenidos en éste reporte; luego de esto validar el modelo mediante la obtención de las medidas de eficiencia, las cuales se comparan con los parametros dados de la simulación.

- **Validación del modelo de simulación:** para ello se generaron muestras de acuerdo con la fórmula expuesta en la sección 2.4.1; con una probabilidad de error permitida del 5 % y 1.5 desviaciones estándar permitidas sobre la media de la distribución de los datos, con lo que se tiene:

$$n = \frac{1.5^2}{0.05} = 45$$

lo cual dio como resultado 45 simulaciones; de las cuales se obtuvieron las medidas de eficiencia del sistema, con esto se garantiza además que el modelo generado arrojará medidas de eficiencias confiable.

3.3. Interpretación de los resultados obtenidos.

La experimentación resulta ser una fase muy importante en el estudio de simulación, ya que permite obtener información de las medidas de eficiencia del sistema que determinan el comportamiento del mismo y con ellas tomar las mejores decisiones para su configuración, así también, podemos experimentar con la base de los tiempos obtenidos en el sistema real para que nos permita cambiar los parámetros de sus distribuciones logrando la configuración óptima del sistema.

Como podemos observar en el cuadro 3.6 que el modelo G/G/2 posee una porcentaje de utilización del sistema relativamente bajo, el cual es de 71 %, esto quiere decir que el 29 % del

Cuadro 3.6: Medidas de eficiencia para el modelo G/G/2.

Medida	**Valor**
Distribución de los tiempos de llegada	Normal(3.7,2.0)
Distribución de los tiempos de servicio	Uniforme(1.05,9.45)
Tasa de llegadas de clientes al sistema (por minuto)	0.3
Tasa de servicio de los cajeros en el sistema (por minuto)	0.2
Porcentaje de utilización del sistema	71 %
Número medio de clientes en el sistema	1.7
Número medio de clientes en cola	0.4
Tiempo medio que un cliente espera en el sistema	6.6
Tiempo medio que un cliente espera en cola	1.3
Probabilidad que al llegar un cliente al sistema lo encuentre ocupado	59 %

tiempo de trabajo los cajeros se encuentran desocupados. Además se puede mencionar que el número medio de clientes en el sistema es cercano a 2, es decir, un cliente con cada cajero, vemos tambien que el tiempo medio de espera en cola es aproximadamente 1 minuto y medio, por tanto podemos pensar de manera preliminar que el servicio que los cajeros proporcionan no es el adecuado o la tasa de llegadas al sistema es relativamente baja como para tener 2 cajeros en él; para brindar una mejor explicación de lo anterior se realizará una simulación del funcionamiento de la agencia bancaria, bajo los siguientes supuestos:

- Quitar un cajero al sistema.
- Disminuir el tiempo de servicio.
- Dejar solamente un cajero y disminuir el tiempo de servicio.

En estos casos es posible regular el factor de utilización del sistema logrando un servicio más eficiente a sus cliente. Es por ello que nos preguntamos *¿Que pasaría si le quitamos un cajero al sistema?*, entonces para responder esta pregunta se realizó una simulación la cual nos permite ver que el nuevo modelo sería un G/G/1 y para el cual se obtienen las siguientes medidas de eficiencia (Ver cuadro 3.7). Al quitarle un canal de servicio al sistema vemos que se desborda a tal grado de tener en cola a 48 clientes y un tiempo medio de espera en cola de 176 minutos. Para este caso vemos que al eliminar un canal de servicio el sistema se vuelve un completo caos. Ahora bien, si nosotros decidimos reducir un minuto al tiempo de servicio en cada una de las transacciones que se llevan a cabo (Tiempo de servicio - 1 minuto), resulta entonces que obtenemos una nueva media para dicho tiempo; la cual será de $\mu = 3.9\ minutos$, manteniéndose

Cuadro 3.7: Medidas de eficiencia para el modelo G/G/1.

Medida	**Valor**
Tasa de llegadas de clientes al sistema (por minuto)	0.3
Tasa de servicio de los cajeros en el sistema (por minuto)	0.2
Porcentaje de utilización del sistema	99.3 %
Número medio de clientes en el sistema	49
Número medio de clientes en cola	48
Tiempo medio que un cliente espera en el sistema	182
Tiempo medio que un cliente espera en cola	176
Probabilidad que al llegar un cliente al sistema lo encuentre ocupado	99.3 %

la distribución uniforme a diferencia que ahora se mueve en el intervalo de $[0.3, 8.45]$, realizando la simulación con los cambios respectivos en los parámetros de los tiempos de servicio se obtienen los datos del cuadro 3.8. Puede verse en el cuadro 3.8 que el factor de utilización del sistema es

Cuadro 3.8: Medidas de eficiencia para el modelo G/G/2 reduciendo los tiempos de servicio.

Medida	**Valor**
Tasa de llegadas de clientes al sistema (por minuto)	0.3
Tasa de servicio de los cajeros en el sistema (por minuto)	0.2
Porcentaje de utilización del sistema	60 %
Número medio de clientes en el sistema	1.2
Número medio de clientes en cola	0.1
Tiempo medio que un cliente espera en el sistema	4.4
Tiempo medio que un cliente espera en cola	0.3
Probabilidad que al llegar un cliente al sistema lo encuentre ocupado	33 %

más bajo que lo mostrado en el cuadro 3.6; entonces, si a esta nueva variante de reducción del tiempo de servicio le quitamos un servidor obtenemos lo mostrado en el cuadro 3.9.

El porcentaje de utilización del sistema para este caso se encuentra en un 98 %, los tiempos de espera en cola son demasiado altos al igual que el número de clientes esperando a ser atendidos.

Cuadro 3.9: Medidas de eficiencia para el modelo G/G/1 reduciendo los tiempos de servicio.

Medida	**Valor**
Tasa de llegadas de clientes al sistema (por minuto)	0.3
Tasa de servicio de los cajeros en el sistema (por minuto)	0.2
Porcentaje de utilización del sistema	98 %
Número medio de clientes en el sistema	27
Número medio de clientes en cola	26
Tiempo medio que un cliente espera en el sistema	111
Tiempo medio que un cliente espera en cola	107
Probabilidad que al llegar un cliente al sistema lo encuentre ocupado	98 %

Analizando entonces las posibles combinaciones de modelos podemos llegar a la conclusión que la afluencia de clientes a la *Agencia Bancaria* es muy bajo; la ubicación es una de las principales desventajas, por lo tanto, podemos concluir que el modelo se encuentra bien como esta y que lo único que debería de realizarse es algun tipo de publicidad de dicha sucursal bancaria para que así la afluencia de clientes sea un poco mayor y evitar tiempo ocioso de los cajeros (para una mejor referencia de las salidas del programa WinQSB recomendamos ver la sección de apéndices).

Apéndice A

Apéndice

Figura A.1: Anexo 1. Corrida en el WinQSB del modelo G/G/2 mostrado en el cuadro 3.6

07-09-2009	Performance Measure	Result
1	System: G/G/2	From Approximation
2	Customer arrival rate (lambda) per min =	0.2703
3	Service rate per server (mu) per min =	0.1905
4	Overall system effective arrival rate per min =	0.2703
5	Overall system effective service rate per min =	0.2703
6	Overall system utilization =	70.9459 %
7	Average number of customers in the system (L) =	1.7824
8	Average number of customers in the queue (Lq) =	0.3635
9	Average number of customers in the queue for a busy system (Lb) =	0.6172
10	Average time customer spends in the system (W) =	6.5948 mins
11	Average time customer spends in the queue (Wq) =	1.3448 mins
12	Average time customer spends in the queue for a busy system (Wb) =	2.2836 mins
13	The probability that all servers are idle (Po) =	16.9960 %
14	The probability an arriving customer waits (Pw) or system is busy (Pb) =	58.8879 %
15	Average number of customers being balked per min =	0
16	Total cost of busy server per min =	$0
17	Total cost of idle server per min =	$0
18	Total cost of customer waiting per min =	$0
19	Total cost of customer being served per min =	$0
20	Total cost of customer being balked per min =	$0
21	Total queue space cost per min =	$0
22	Total system cost per min =	$0

Figura A.2: Anexo 2. Corrida en el WinQSB del modelo G/G/1 mostrado en el cuadro 3.7

07-09-2009	Performance Measure	Result
1	System: G/G/1	From Simulation
2	Customer arrival rate (lambda) per min =	0.2703
3	Service rate per server (mu) per min =	0.1905
4	Overall system effective arrival rate per min =	0.2676
5	Overall system effective service rate per min =	0.1817
6	Overall system utilization =	99.4275 %
7	Average number of customers in the system (L) =	43.6806
8	Average number of customers in the queue (Lq) =	42.6863
9	Average number of customers in the queue for a busy system (Lb) =	42.9321
10	Average time customer spends in the system (W) =	161.0865 mins
11	Average time customer spends in the queue (Wq) =	155.6150 mins
12	Average time customer spends in the queue for a busy system (Wb) =	156.5111 mins
13	The probability that all servers are idle (Po) =	0.5725 %
14	The probability an arriving customer waits (Pw) or system is busy (Pb) =	99.4275 %
15	Average number of customers being balked per min =	0
16	Total cost of busy server per min =	$0
17	Total cost of idle server per min =	$0
18	Total cost of customer waiting per min =	$0
19	Total cost of customer being served per min =	$0
20	Total cost of customer being balked per min =	$0
21	Total queue space cost per min =	$0
22	Total system cost per min =	$0
23	Simulation time in min =	1000.0000
24	Starting data collection time in min =	0
25	Number of observations collected =	182
26	Maximum number of customers in the queue =	86
27	Total simulation CPU time in second =	0.0470

Figura A.3: Anexo 3. Corrida en el WinQSB del modelo G/G/2 reduciendo los tiempos de servicio, mostrado en el cuadro 3.8

07-09-2009	Performance Measure	Result
1	System: G/G/2	From Approximation
2	Customer arrival rate (lambda) per min =	0.2703
3	Service rate per server (mu) per min =	0.2210
4	Overall system effective arrival rate per min =	0.2703
5	Overall system effective service rate per min =	0.2703
6	Overall system utilization =	61.1487 %
7	Average number of customers in the system (L) =	1.4015
8	Average number of customers in the queue (Lq) =	0.1785
9	Average number of customers in the queue for a busy system (Lb) =	0.3846
10	Average time customer spends in the system (W) =	5.1854 mins
11	Average time customer spends in the queue (Wq) =	0.6604 mins
12	Average time customer spends in the queue for a busy system (Wb) =	1.4232 mins
13	The probability that all servers are idle (Po) =	24.1090 %
14	The probability an arriving customer waits (Pw) or system is busy (Pb) =	46.4063 %
15	Average number of customers being balked per min =	0
16	Total cost of busy server per min =	$0
17	Total cost of idle server per min =	$0
18	Total cost of customer waiting per min =	$0
19	Total cost of customer being served per min =	$0
20	Total cost of customer being balked per min =	$0
21	Total queue space cost per min =	$0
22	Total system cost per min =	$0

Figura A.4: Anexo 4. Corrida en el WinQSB del modelo G/G/1 reduciendo los tiempos de servicio, mostrado en el cuadro 3.9

07-09-2009	Performance Measure	Result
1	System: G/G/1	From Simulation
2	Customer arrival rate (lambda) per min =	0.2703
3	Service rate per server (mu) per min =	0.2210
4	Overall system effective arrival rate per min =	0.2766
5	Overall system effective service rate per min =	0.2197
6	Overall system utilization =	99.1872 %
7	Average number of customers in the system (L) =	28.6968
8	Average number of customers in the queue (Lq) =	27.7049
9	Average number of customers in the queue for a busy system (Lb) =	27.9320
10	Average time customer spends in the system (W) =	105.0675 mins
11	Average time customer spends in the queue (Wq) =	100.5523 mins
12	Average time customer spends in the queue for a busy system (Wb) =	101.3763 mins
13	The probability that all servers are idle (Po) =	0.8128 %
14	The probability an arriving customer waits (Pw) or system is busy (Pb) =	99.1872 %
15	Average number of customers being balked per min =	0
16	Total cost of busy server per min =	$0
17	Total cost of idle server per min =	$0
18	Total cost of customer waiting per min =	$0
19	Total cost of customer being served per min =	$0
20	Total cost of customer being balked per min =	$0
21	Total queue space cost per min =	$0
22	Total system cost per min =	$0
23	Simulation time in min =	1000.0000
24	Starting data collection time in min =	0
25	Number of observations collected =	220
26	Maximum number of customers in the queue =	58
27	Total simulation CPU time in second =	0.0630

Figura A.5: Anexo 5.

07-09-2009	Result	Clientes
1	Total Number of Arrival	264
2	Total Number of Balking	13
3	Average Number in the System (L)	1.4668
4	Maximum Number in the System	3
5	Current Number in the System	2
6	Number Finished	249
7	Average Process Time	5.4484
8	Std. Dev. of Process Time	2.5307
9	Average Waiting Time (Wq)	0.4221
10	Std. Dev. of Waiting Time	1.0030
11	Average Transfer Time	0
12	Std. Dev. of Transfer Time	0
13	Average Flow Time (W)	5.8705
14	Std. Dev. of Flow Time	2.7141
15	Maximum Flow Time	13.6414
	Data Collection: 0 to	1000 mins
	CPU Seconds =	0.7030

Figura A.6: Anexo 6.

07-09-2009	Server Name	Server Utilization	Average Process Time	Std. Dev. Process Time	Maximum Process Time	Blocked Percentage	# Customers Processed
1	Cajero1	66.84%	5.9151	2.5475	9.4207	0.00%	113
2	Cajero2	68.82%	5.0606	2.4500	9.2700	0.00%	136
	Overall	67.83%	5.4484	2.5307	9.4207	0.00%	249
Data	Collection:	0 to	1000	mins	CPU	Seconds =	0.7030

Figura A.7: Anexo 7.

07-09-2009	Queue Name	Average Q. Length (Lq)	Current Q. Length	Maximum Q. Length	Average Waiting (Wq)	Std. Dev. of Wq	Maximum of Wq
1	Cola	0.1075	0	1	0.4285	1.0073	5.6608
Data	Collection:	0 to	1000	mins	CPU	Seconds =	0.7030

Bibliografía

[1] Adan, Ivo and Jacques Resing. Queueing Theory. Department of Mathematics and Computing Science Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, February 2002.

[2] Allen, A. O. (1978): Probability, Statistics and Queueing Theory With Computer Science Applications. Academic Press.

[3] Cao Abad, Ricardo et. al. (2007-2008): Teoría de colas. Departamento de matemáticas (Universidad de A Coruña).

[4] Domingo Morales González (2003), Teoría de Colas. ISBN 84-605-1045-X.

[5] García Sabater J. P. (2001): Métodos Cuantitativos de Organización Industrial.

[6] Goddard L. S. (1969): Técnicas Matemáticas de la Investigación Operacional. Editorial Alhambra, S. A.

[7] Gross D.and Harris C. M. (1974): Fundamentals of Queueing Theoty, John Wiley.

[8] Gross, Donald y Carl Harris (1998). Resumen traducido del libro Fundamentals of Queueing Theory por. Dpto. de Organización de Empresas, E.F. y C., Año 2001.

[9] Gómez, Andrés (1997): Introducción a la Teoría de Colas. Universidad Nacional de Colombia Sede Medellín.

[10] Hillier, F. S. y Liebenmen, G. J. (1980): Introducción al Investigación de Operaciones, McGraw-Hill.

[11] Leandro,Gabriel (2003). Línea de Espera: Teoría de Colas, Cursos Métodos Cualitativos. http://www.auladeeconomia.com año 2002.

[12] Linares, Pedro. i.e.Modelos matemáticos de simulacion. Universidad Pontífica de Comillas. Octubre 2005.

[13] Ortuño, Maria Teresa. Notas de clase maestria en estadística. Universidad de El Salvador. Septiembre de 2008.

[14] Parzen, Emanuel. Procesos Estocásticos, Departamento de estadística de la universidad de Stanford. Madrid 1972.

[15] Pazos, Arias et. al. Teoría de colas y simulación de eventos discretos. Pearson Prentice Hall. España 2003.

[16] Ross S. M. (1985): Introduction to Probability Models, Academic Press.

[17] Saaty T. L. (1967): Elementos de la teoría de colas / Thomas L. Saaty; TR. Rafael Pro Bermejo. Madrid. 1967.

[18] Vitoriano, Begoña. Teoría de colas o líneas de espera. Universidad Complutense de Madrid. Julio de 2008.